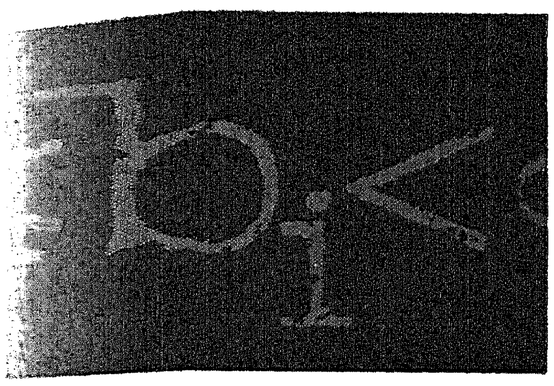
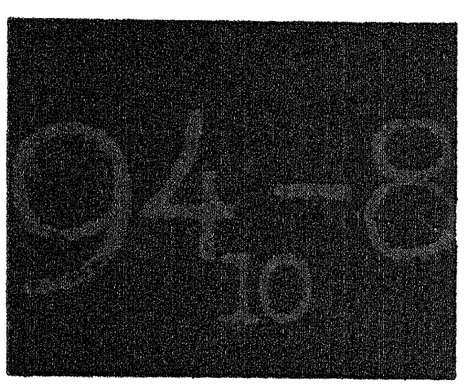
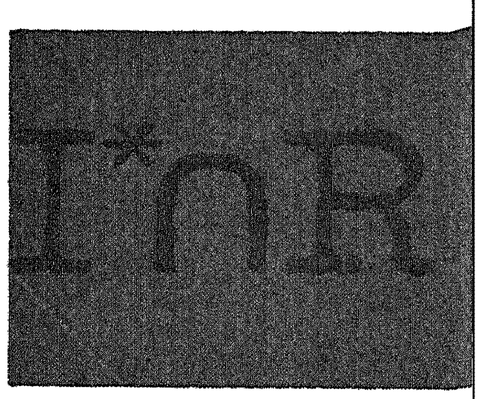
MARKOV DECISION THEORY

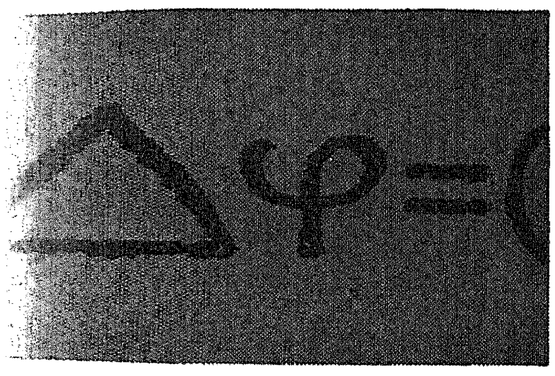
PROCEEDINGS OF THE ADVANCED SEMINAR
ON MARKOV DECISION THEORY HELD AT
AMSTERDAM, THE NETHERLANDS, SEPTEMBER 13-17, 1976

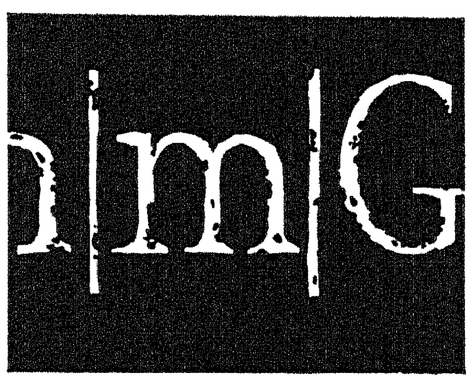
Edited by
H.C. TIJMS & J. WESSELS

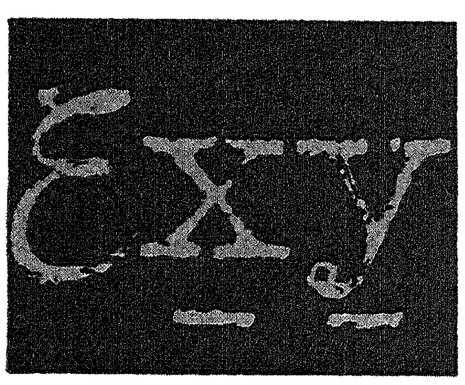


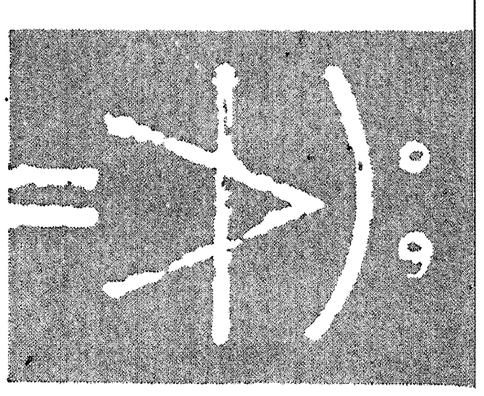


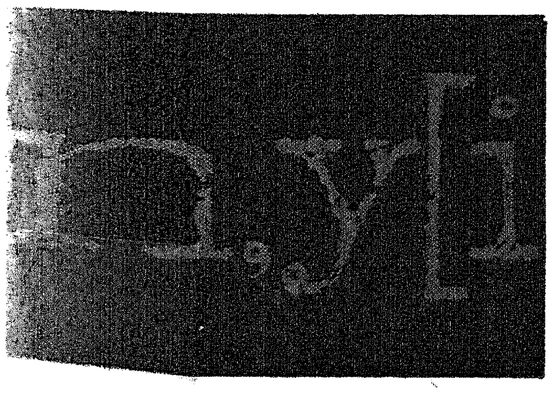


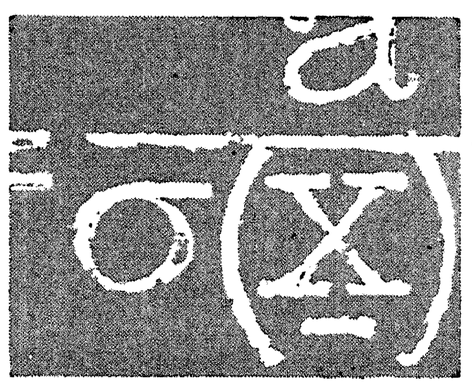


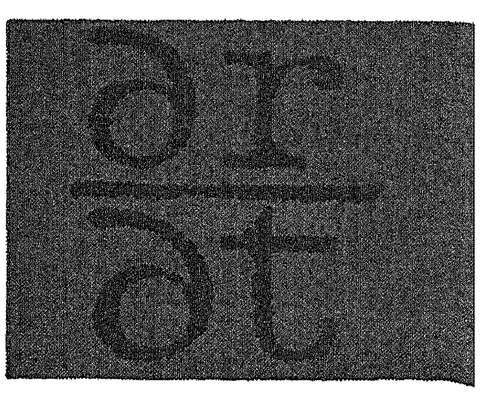












MATHEMATICAL CENTRE TRACTS 93

MATHEMATICAL CENTRE TRACTS 93

Edited by
H.C. TIJMS & J. WESSELS

MARKOV DECISION THEORY

PROCEEDINGS OF THE ADVANCED SEMINAR
ON MARKOV DECISION THEORY HELD AT
AMSTERDAM, THE NETHERLANDS, SEPTEMBER 13-17, 1976

MATHEMATISCH CENTRUM AMSTERDAM 1977

AMS(MOS) subject classification scheme (1970): 90C40, 90C45, 90D15

ISBN 90 6196 160 2

PREFACE

The Advanced Seminar on Markov Decision Theory was organized by the Mathematical Centre in cooperation with the Eindhoven University of Technology. The Seminar was held at the University of Amsterdam, September 13-17, 1976.

The main theme of this meeting was successive approximations in Markov decision theory. The Seminar was arranged as a series of lectures in which the state of affairs and recent developments were presented.

In these proceedings the reader will find written versions of most of the lectures. Some of the papers contain new developments which were stimulated by the discussions during the meeting. We hope that this volume will be as stimulating for further research as the Seminar has proved to be.

We are grateful to the Mathematical Centre who provided the financial support and in the person of A. Federgruen took charge of the organization. We are indebted to the Econometric Institute of the University of Amsterdam for their hospitality. We further thank Professor G. de Leve and Dr. J.A.E.E. van Nunen who contributed in an excellent way to the success of the meeting.

The editors

CONTENTS

MARKOV DECISION PROCESSES WITH UNBOUNDED REWARDS

J.A.E.E. van Nunen
Graduate School of Management, Delft, The Netherlands

J.Wessels
Eindhoven University of Technology, Eindhoven, The Netherlands

1. INTRODUCTION

We consider a Markov decision system with a countable state space S. So the states in S may be labelled by the natural numbers S := {1,2,3,...}. The system can be controlled at discrete points in time t = 0,1,2,... by choosing an action a from an arbitrary nonempty action space A. Let $\mathcal{A}$ be a σ-field on A, such that $\{a\} \in \mathcal{A}$ for all $a \in A$.

The chosen action $a \in A$ and the current state $i \in S$ at time t exclusively determine the probability of occurence of state $j \in S$ at time t + 1. This probability is denoted by $p^a(i,j)$. If state i has been observed at time t and action $a \in A$ has been chosen, the (expected) reward r(i,a) is earned. The objective is to find a decision rule for which the total expected reward over an infinite time horizon is maximal. For the determination of such a decision rule and for the computation of the total expected reward we have in fact to solve a functional equation of the following form

$$v(i) = \sup_{a \in A} \{r(i,a) + \sum_j p^a(i,j)v(j)\}, \qquad i \in S.$$

The more sophisticated methods for solving these functional equations, if they have a unique solution, are linear programming (D'EPENOUX [3], DE GHELLINCK & EPPEN [4]) and policy iteration (HOWARD [13]), which is a

very beautiful and elegant method. Actually, linear programming and policy iteration are in a sense equivalent (MINE & OSAKI [18], WESSELS & VAN NUNEN [29]).

However, for large scaled problems, successive approximation methods tend to be more efficient than the known sophisticated methods (e.g. VAN NUNEN [19]).

It appears that successive approximation methods allow for elegant and relatively good extrapolation and error analysis. Moreover, the incorporation of suboptimality tests can improve those methods considerably. Finally, it appears that policy iteration methods (there are many versions with differences in the policy improvement procedures, see e.g. HASTINGS [6], VAN NUNEN [21]) are essentially successive approximation methods. These methods happen to converge in finitely many iterations if state and action space are finite.

For these reasons it is still interesting to investigate successive approximation methods for Markov decision processes and likewise for Markov games (see VAN DER WAL [27]). Here we will mainly be concerned with the conditions which allow successive approximations with guaranteed convergence in some strong sense allowing the construction of upper and lower bounds. For convergence in a weaker sense, of course, weaker conditions can be used we refer to SCHÄL [25] and VAN HEE & VAN DER WAL [12].

After the introduction of the model and the underlying assumptions we will develop some properties.

Moreover, we will indicate the specific successive appproximation algorithm. Finally we will analyse the assumptions and compare them with those in literature.

Most of the assertions can be extended to nondenumerable state spaces in the obvious way.

2. THE MODEL AND THE ASSUMPTIONS

We will first introduce our assumptions on the transition probabilities and the rewards. The assumptions will be somewhat weaker than those proposed in [21].

ASSUMPTION 2.1

a) $p^a(i,j) \geq 0, \sum_j p^a(i,j) \leq 1,$ for all $i,j \in S$ and all $a \in A$.

b) $p^a(i,j)$ is measurable for all $i,j \in S$ as a function of a.

c) $r(i,a)$ is measurable for all $i \in S$ as a function of a.

REMARK 2.1. We allow substochastic behaviour. Defectiveness of transition probabilities may be interpreted as a positive probability of leaving the system, which results in the stopping of all earnings. In a more formal set-up this may be handled by introducing an extra state which is absorbing for all actions and does not give any earnings. This has been executed e.g. in [21] by VAN NUNEN and in [11] by HINDERER. Without such a device quite a lot can be achieved in a correct formal way as has been done by WESSELS [28]. Actually, as long as the outcomes in which one is interested may be expressed in terms of bounded order histories, there is no serious problem. In this paper we will suppose that there is such an extra state, without giving it a name or mentioning it explicitly. Compare section 5 for the meaning of substochasticity.

DEFINITION 2.1.

(i) A *decision rule* π is a sequence of transition probabilities $\pi := (q_0, q_1, \ldots)$, where q_t is a transition probability of $(H_t, \mathcal{H}_t)$ into $(A, \mathcal{A})$, with $H_t := S \times A \times S \times \ldots \times S$ (t+1 times S) and $\mathcal{H}_t$ is the corresponding product σ-field.
The class of all decision rules is denoted by $\mathcal{D}$.

(ii) A decision rule π will be called *nonrandomized* or *a strategy* if q_t is degenerated for all t and all $h_t \in H_t$. So a strategy is a non-randomized decision rule.

(iii) A decision rule π is called *Markov* if q_t only depends on the last component of $h_t \in H_t$.
The class of (randomized) Markov decision rules is denoted by RM.

(iv) A Markov decision rule is called *stationary* if q_t does not depend on t.
A *policy* f is a function of S into A. By F we denote the set of all policies. Stationary strategies correspond (one to one) to policies and Markov strategies correspond to sequences of policies. We will apply these correspondences deliberately.
The class of Markov strategies is denoted by M.

In an obvious way - see e.g. VAN NUNEN [21] - any starting state $i \in S$ and any decision rule $\pi \in \mathcal{D}$ determine a stochastic process $\{(X_t,Z_t)\}_{t=0}$ on $S \times A$, where X_t denotes the state of the system at time t, and Z_t denotes the action at time t. The relevant probability measure on $(S\times A)^\infty$ will be denoted by $\mathbb{P}_i^\pi$. Expectations with respect to this measure will be denoted by $\mathbb{E}_i^\pi$. By $\mathbb{E}^\pi X$ we denote the columnvector with i-th component $\mathbb{E}_i^\pi X$, where X is any random variable.

ASSUMPTION 2.2. We assume a positive function μ on S to be given. Let W be the Banach space of vectors w (real valued functions on S) which satisfy

$$\|w\| := \sup_{i\in S} |w(i)| \cdot \mu^{-1}(i) < \infty.$$

For matrices (real valued functions on $S \times S$) we introduce the operator-norm

$$\|B\| := \sup_{\|w\|=1} \|Bw\|.$$

Note that

$$\|B\| = \sup_{i\in S} \mu^{-1}(i) \sum_j |B(i,j)| \cdot \mu(j).$$

ASSUMPTION 2.3.

(i) $$\sup_{\pi\in M} \mathbb{E}_i^\pi \sum_{n=0}^{\infty} r^+(X_n,Z_n) < \infty \quad \text{for all } i \in S,$$

where $r^+(a,b) := \max\{0,r(a,b)\}$.

(ii) $$\sup_{f\in F} \|P(f)\| =: \rho_* < 1,$$

where P(f) is the matrix with $P(f)(i,j) := p^{f(i)}(i,j)$.

(iii) $$\sup_{f\in F} \|P(f)\bar{r} - \rho\bar{r}\| =: M_1 < \infty \quad \text{for some } \rho \text{ with } 0 < \rho < 1,$$

and $\bar{r}$ is the vector with i-th component $\bar{r}(i) := \sup_{a\in A} r(i,a)$.

REMARK 2.3. Note that $P(f)\bar{r}^+ < \infty$ (componentwise) since $\sup_{g\in F} P(f)r^+(g) < \infty$. Moreover, $P(f)\bar{r}^- < \infty$ as is implicitly stated in assumption 2.2. iii. The model in fact combines the main features of the models introduced by HARRISON [5], WESSELS [28] and VAN HEE [9], and yields a slight extension with respect to the model considered by VAN NUNEN [21].

Since we will prove similar results as HARRISON [5], WESSELS [28], VAN NUNEN [21], this paper generalizes their results.

We will first show that under assumption 2.3.i the restriction to Markov strategies is allowed if one is interested in the criterion of total expected rewards.

Given that assumption 2.3.i is satisfied it will be clear that for any $\pi \in M$

$$v(\pi) := \mathbb{E}^{\pi} \sum_{n=0}^{\infty} r(X_n, Z_n)$$

is properly defined and that all manipulations with integration and summation are allowed. However, $v_i(\pi)$ may be $-\infty$ for some $i \in S$. Furthermore $\sup_{\pi \in M} v_i(\pi) < \infty$. In [9] VAN HEE shows that under assumption 2.3.i $v_i(\pi)$ is properly defined for all $\pi \in RM$ since

$$\sup_{\pi \in RM} \mathbb{E}_i^{\pi} \sum_{n=0}^{\infty} r^+(X_n, Z_n) = \sup_{\pi \in M} \mathbb{E}_i^{\pi} \sum_{n=0}^{\infty} r^+(X_n, Z_n).$$

Moreover, he proves that

$$\sup_{\pi \in RM} v_i(\pi) = \sup_{\pi \in M} v_i(\pi).$$

It then follows straightforwardly from the generalisation of a result of DERMAN and STRAUCH [2] that $v_i(\pi)$ is defined properly for all $\pi \in D$ and $i \in S$, viz. for any $i \in S$ and any $\pi \in D$ there exists a $\pi^* \in RM$, such that

$$\mathbb{P}_i^{\pi}[X_n = j, Z_n \in A_0] = \mathbb{P}_i^{\pi^*}[X_n = j, Z_n \in A_0]$$

for all $j \in S$, $A_0 \in A$, $n = 0,1,\ldots$.

Hence

$$\mathbb{E}_i^{\pi} \sum_{n=0}^{\infty} r^+(X_n, Z_n) = \mathbb{E}_i^{\pi^*} \sum_{n=0}^{\infty} r^+(X_n, Z_n) < \infty,$$

so $v_i(\pi)$ is properly defined and equal to $v_i(\pi^*)$.

This implies

$$\sup_{\pi\in\mathcal{D}} v_i(\pi) = \sup_{\pi\in M} v_i(\pi).$$

This actually means that one can restrict oneself to strategies which only depend on the starting state, on the time instant t and on the state at that time. Such strategies are sometimes called semi-Markov strategies. The starting state and the time instant will be proved to be superfluous later on.

3. SOME PROPERTIES

Let $\overline{\mathbb{R}}$ denote the set of real numbers with $+\infty$ and $-\infty$ included. Let W^- contain those $w \in \overline{\mathbb{R}}^\infty$, such that $w \le w_0$ for some $w_0 \in W$, (w_0 is not fixed, but may depend on w, so $W \subset W^-$). P(f) is properly defined as an operator on W and on W^- as well. P(f) maps each of these sets into itself. Here "properly defined" means that $(P(f)w)(i)$ is independent of the order of summations. It is straightforward that P(f) is monotone on W and W^-. Moreover P(f) is contracting on W with contraction radius $\|P(f)\| \le \rho_* < 1$. The set V is defined as the set of vectors v in $\mathbb{R}^\infty$ such that $v - (1-\rho)^{-1}\bar{r} \in W$. Since W is a Banach space the set V is a complete metric space with respect to the metric $\|v_1-v_2\|$. The set V^- contains those $v \in \overline{\mathbb{R}}^\infty$ such that for some $v_0 \in V$ we have $v \le v_0$.

LEMMA 3.1.

$$\|P(f_n)\ldots P(f_1)\bar{r} - \rho^{n-}\bar{r}\| \le n\rho_0^{n-1} M_1, \quad n \ge 1$$

with $\rho_0 := \max\{\rho,\rho_*\}$.

PROOF.

$$P(f_2)P(f_1)\bar{r} \le P(f_2)(\rho\bar{r} + M_1\mu)$$

$$\le \rho^{2-}\bar{r} + \rho M_1\mu + \rho_* M_1\mu$$

$$\le \rho^{2-}\bar{r} + 2\rho_0 M_1\mu$$

similarly

$$P(f_2)P(f_1)\bar{r} \geq P(f_2)(\rho\bar{r} - M_1\mu)$$

$$\geq \rho^2\bar{r} - \rho M_1\mu - \rho_* M_1$$

$$\geq \rho^2\bar{r} - 2\rho_0 M_1$$

The proof proceeds further in an inductive way. □

Corollary 3.1.

(i) $$\mathbb{E}^\pi \sum_{n=0}^{\infty} \bar{r}(X_n) \in V \qquad \text{for all } \pi \in M$$

(ii) $$\mathbb{E}^\pi \sum_{n=0}^{\infty} r(X_n, Z_n) \leq (1-\rho)^{-1}\bar{r} + \sum_{n=1}^{\infty} n\rho_0^{n-1} M_1\mu$$

$$= (1-\rho)^{-1}\bar{r} + (1-\rho_0)^{-2} M_1\mu \in V$$

for all $\pi \in \mathcal{D}$.

PROOF. For $\pi \in M$ part (ii) follows straightforwardly from the foregoing lemma. Because of the results of section 2 this may be extended to $\pi \in \mathcal{D}$. □

DEFINITION 3.1. $L(f)$ is a mapping of V^- into V^- defined by $L(f)v := r(f) + P(f)v$ where $r(f)$ is the vector with i-th component equal to $r(i,f(i))$. $L(f)$ maps V^- into V^- viz. $r(f) \leq \bar{r}$; $v \leq v_0$ for some $v_0 \in V$, therefore

$$\| v_0 - (1-\rho)^{-1}\bar{r} \| = M_2 < \infty,$$

hence

$$r(f) + P(f)v \leq \bar{r} + P(f)(1-\rho)^{-1}\bar{r} + P(f)M_2\mu$$

$$\leq \bar{r} + (1-\rho)^{-1}(\rho\bar{r} + M_1\mu) + \rho M_2\mu$$

$$= (1-\rho)^{-1}\bar{r} + (M_1(1-\rho)^{-1} + \rho M_2)\mu \in V.$$

LEMMA 3.2.

(i) *If* $r(f) - \bar{r} \in W$, *then* $L(f)$ *maps* V *into* V *and* $L(f)$ *is contracting on* V *with contraction radius* $\|P(f)\| \leq \rho_* < 1$. *The fixed point of* $L(f)$ *in* V *is* $v(f) := f((f,f,f,\ldots))$.

(ii) $L(f)$ *is monotone on* V^-.

(iii) *If* $v \in V$, *then* $L^n(f)v \to v(f)$ *for* $n \to \infty$.

PROOF. Part (i) can be found in [28], part (ii) of the lemma is trivial. The final part is straightforward if $r(f) - \bar{r} \in W$, since in that case the assertion is implied by the Banach fixed point theorem and the convergence is in norm. If $r(f) - \bar{r} \notin W$ we have

$$L^n(f)v = \sum_{k=0}^{n-1} P^k(f)r(f) + P^n(f)v.$$

Since v can be written as

$$v = (1-\rho)^{-1}\bar{r} + w \qquad \text{with } w \in W$$

we have $P^n(f)v = (1-\rho)^{-1}P(f)\bar{r} + P^n(f)w$.
However, $P^n(f)w$ tends to zero for $n \to \infty$ since $P(f)$ is contracting on W (assumption 2.3 ii) and $P^n(f)\bar{r}$ tends to zero for $n \to \infty$ as follows from lemma 3.1. This implies

$$\lim_{n\to\infty} L^n(f)v = \sum_{k=0}^{\infty} P^k(f)r(f) = v(f). \qquad \square$$

DEFINITION 3.2. U is a mapping of V into V defined by

$$Uv := \sup_{f\in F} L(f)v \qquad \text{(componentwise)}.$$

U maps V into V, viz.

$$Uv = \sup_{f\in F} \{r(f) + P(f)[(1-\rho)^{-1}\bar{r} + w]\}$$

$$\leq \bar{r} + \sup_{f\in F} \{(1-\rho)^{-1}P(f)\bar{r}\} + \sup_{f\in F} P(f)w$$

$$\leq (1-\rho)^{-1}\bar{r} + (1-\rho)^{-1}M_1\mu + \rho_*\|w\|\mu \in V$$

and

$$Uv \geq \bar{r} + \inf_{f\in F} (1-\rho)^{-1}P(f)\bar{r} + \inf_{f\in F} P(f)w$$

$$\geq \bar{r} + (1-\rho)^{-1}\rho\bar{r} - M_1\mu(1-\rho)^{-1} - \rho_*\|w\|\mu$$

$$= (1-\rho)^{-1}\bar{r} - M_1(1-\rho)^{-1}\mu - \rho_*\|w\|\mu \in V.$$

LEMMA 3.3.

(i) U *is monotone on* V;

(ii) U *maps* $B := \{v \in V \mid \|v - (1-\rho)^{-1}\bar{r}\| \leq M_1 (1-\rho)^{-1}(1-\rho_*)^{-1}\}$ *into itself;*

(iii) U *is contracting on* V *with contraction radius* γ: $\gamma \leq \rho_* < 1$.

The proof proceeds in a similar way as the proof of theorem 4.3.3. in VAN NUNEN [21]. □

REMARK 3.1. Suppose the supremum in Uv for $v \in V$ is attained for certain f then

$$r(f) + P(f)v \in V$$

hence

$$r(f) + P(f)(1-\rho)^{-1}\bar{r} + P(f)w \in V$$

and

$$r(f) + (1-\rho)^{-1}\bar{r} \in V$$

so

$$r(f) - \bar{r} + \bar{r} + (1-\rho)^{-1}\bar{r} = r(f) - \bar{r} + (1-\rho)^{-1}\bar{r} \in V$$

consequently $r(f) - \bar{r} \in W$.

The same holds if L(f)v approximates Uv in norm. Then $L(f)v \in V$ as well. Hence $r(f) - \bar{r} \in W$ so the use of a successive approximation method (even without computing the supremum exactly) leads to a sequence of policies $f_n \in F$ with $r(f_n) - \bar{r} \in W$.

Since U is contracting in V there exists a unique fixed point v^* of U in V. This fixed point is the unique solution of the optimality equation in V

$$v = \sup_{f \in F} \{r(f) + P(f)v\}.$$

Furthermore $\|U^n v - v^*\| \to 0$ for $n \to \infty$ and any $v \in V$. In the sequel we will prove that

$$v^* = \sup_{\pi \in \mathcal{D}} \mathbb{E}^{\pi} \sum_{n=0}^{\infty} r(X_n, Z_n) = \sup_{\pi \in \mathcal{D}} v(\pi).$$

THEOREM 3.1.

(i) $v(\pi) \leq v^*$ *for all* $\pi \in \mathcal{D}$

(ii) *For any* $\varepsilon > 0$ *there exists a policy* f *such that*

$$\|v(f) - v^*\| \leq \varepsilon$$

hence

$$\sup_{\pi \in \mathcal{D}} v(\pi) = \sup_{f \in M} v(f) = v^*.$$

Moreover, if for some f *holds that*

$$v^* = r(f) + P(f)v^*$$

Then

$$v(f) = v^*.$$

PROOF. The proof of this theorem proceeds exactly along the same lines as the proof of theorem 4.3.4 in [21]. In [21] part (i) has been proved by

showing first that the assertion is true for $\pi \in M$ and then using the results of section 2. Part (ii) follows directly if we choose $f \in F$ such that

$$v^* - \delta\mu \leq L(f)v^* \leq v^*$$

then

$$L(f)[v^* - \delta\mu] \leq L^2(f)v^* \leq v^*$$

hence

$$v^* + \delta(1+\rho)\mu \leq L^2(f)v \leq v^*$$

iterating this inequality gives

$$v^* - \frac{\delta}{1-\rho}\mu \leq v(f) \leq v^*$$

so by choosing $\delta = \varepsilon(1-\rho)$ the statement will be clear. □

4. SUCCESSIVE APPROXIMATIONS

In the previous section we showed that the unique fixed point v^* of the contraction operator U in V is the optimal value vector of the Markov decision problem. Hence, v^* can be approximated by

$$v_n = U^n v_0 \qquad (v_0 \in V \text{ and } n = 1,2,\ldots).$$

Furthermore, we proved the existence of stationary Markov strategies with value functions that approximate v^* (in norm).

Usually one not only wishes to find v^* but one is also interested in good (stationary Markov) strategies. It may occur that the supremum in Uv cannot be computed exactly. Nevertheless, there are several successive approximation methods for the computation of v^* and the determination of an (ε-) optimal stationary Markov strategy. We refer to [22] in this volume. Here, as an example, we describe a method which uses monotonicity of the v_n. Consequently the convergence of the algorithm can be shown by relatively simple proofs.

LEMMA 4.1. *Let* $\delta > 0$, *suppose* v , $v' \in V$, *such that* $Uv' - \delta\mu \leq v$ *then*

$$v^* \leq v + \frac{\delta+\rho_*\|v-v'\|}{1-\rho_*}\mu$$

PROOF. The proof can also be found in [28] and proceeds as follows.

$$Uv = U(v'+v-v').$$

Hence, since $Uv' \leq v + \delta\mu$ we have

$$Uv \leq Uv' + \rho_*\|v - v'\|\ \mu \leq v + \delta\mu + \rho_*\|v - v'\|\mu$$

or

$$Uv \leq v + \varepsilon\mu \qquad \text{with } \varepsilon = \delta + \rho_*\|v - v'\|.$$

Similarly

$$U^2v \leq U(v+\varepsilon\mu) = U(v'+v-v'+\varepsilon\mu)$$

$$\leq Uv' + \rho_*\|v - v'\|\mu + \rho_*\varepsilon\mu$$

$$\leq v + \delta\mu + \rho_*\|v - v'\|\mu + \rho_*\varepsilon\mu = v + \varepsilon(1+\rho_*)\mu.$$

Iterating in the same way gives

$$U^nv \leq v + \varepsilon(1+\rho_*+\ldots\rho_*^{n-1})\mu \leq v + \frac{\varepsilon}{1-\rho_*}\mu.$$

This implies

$$\lim_{n\to\infty} U^nv = v^* \leq v + \frac{\varepsilon}{1-\rho_*}\mu. \qquad \square$$

LEMMA 4.2. *If* v, $v' \in V$ *with* $L(f)v' = v$, *then*

$$r(f) - \bar{r} \in W$$

and

$$v + \frac{\rho_f \|v-v'\|_-}{1-\rho_f}\mu \le v(f) \le v + \frac{\rho_* \|v-v'\|}{1-\rho_*}\mu,$$

where

$$\|v-v'\|_- := \inf_{i\in S} \mu^{-1}(i)(v(i)-v'(i))$$

and

$$\rho_f := \inf_{i\in S} \mu^{-1}(i) \sum_j p^{f(i)}(i,j)\ \mu(j).$$

PROOF. The proof of this lemma proceeds along the same lines as the proof of the foregoing lemma. □

The convergence of the following successive approximation algorithm will be clear as a consequence of the foregoing two lemmas.

ALGORITHM 4.1.

STEP 0. Choose $\alpha > 0$; choose $\delta > 0$ such that $\delta(1-\rho_*)^{-1} < \alpha$; choose $v_0 \in V$ such that $v_0 < Uv_0$; $n := 1$;

STEP 1. Determine f_n such that

$$v_n := L(f_n)v_{n-1} \ge \max\{v_{n-1}, Uv_{n-1}-\delta\mu\};$$

STEP 2. If

$$\frac{\delta+\rho_* \|v_n-v_{n-1}\|}{1-\rho_*} - \frac{\rho_{f_n} \|v_n-v_{n-1}\|_-}{1-\rho_{f_n}} < \alpha$$

then go to step 3 else go to step 1 with $n := n + 1$;

STEP 3. End of the algorithm.

Lemma 4.1 and 4.2 provide that the algorithm stops after a finite number of iterations and that in the n-th iteration step of the algorithm,

we have

$$v_n + \frac{\rho_{f_n}\|v_n - v_{n-1}\|}{1-\rho_{f_n}} \leq v(f_n) \leq v^* \leq v_n + \frac{\delta+\rho_*\|v_n - v_{n-1}\|}{1-\rho_*}$$

If the algorithm ends at iteration step n_0 with policy f_{n_0} then the distance between $v^* - v(f_{n_0})$ is at most α and the distance between upper and lowerbound for $v(f_n)$ is less than $\alpha - \delta(1-\rho_*)^{-1}$.

Note that the choice of v_0 and the way in which v_n is computed assure that v_n converges monotonically from below to v^* i.e.

$$v_{n-1} \leq v_n \leq v(f_n) \leq v^*$$

and

$$\lim_{n\to\infty} v_n = v^*.$$

For proofs we refer to [21], [28].

If we release the monotonicity assumptions and choose $v_0 \in V$ arbitrary it remains possible to give adequate successive approximation algorithms, see [22] in this volume.

In all these methods a main role is played by the concept of upper and lowerbound. In fact the fast convergence of the algorithms is caused by the use of this concept, see e.g. MACQUEEN [16], PORTEUS [23], VAN NUNEN [11]. Moreover, upper and lowerbounds can be used to formulate sub-optimality tests which may even improve the efficiency of the algorithms considerably, see e.g. MACQUEEN [17], HASTINGS and VAN NUNEN [8], HASTINGS and MELLO [7], HÜBNER [14].

5. ANALYSIS OF THE ASSUMPTIONS

Let us first make some remarks on the assumptions.

REMARK 5.1.

(i) $\bar{r}$ may be replaced by any vector b with $b - \bar{r} \in W$, so it is not

necessary to compute $\bar{r}$ exactly. Such an approach is applied in VAN NUNEN [21].

(ii) In the model semi-Markov decision processes, discounted Markov decision processes and discounted semi-Markov decision processes are contained as well.

(a) Semi-Markov decision processes (without discounting) are covered by taking the number of the decision instant as decision time and the expected reward until the next decision instant as reward. Alternatively spoken one considers the embedded process, see e.g. MINE and OSAKI [18].

(b) Discounted Markov decision processes are included by incorporating the decision factor β (if $\beta \leq 1$) in the transition probabilities i.e. $\tilde{p}^a(i,j) := \beta p^a(i,j)$. If $\beta > 1$ the theory should be slightly adapted.

However

$$\sup_{\pi \in M} \mathbb{E}_i^{\pi} \sum_{n=0}^{\infty} \beta^n r^+(X_n, Z_n) < \infty$$

remains a sufficient condition for restriction to stationary Markov strategies. (See VAN HEE [9]).

(c) For discounted semi-Markov decision processes with discount rate $\alpha \geq 0$ again incorporation in the transition probabilities is appropriate, for $\alpha < 0$ the theory needs slight modifications.

We now relate the use of the translation function $(1-\rho)^{-1}\bar{r}$, as introduced in a slightly different way by HARRISON [5], to an approach of PORTEUS [24].

PORTEUS proposed, for the finite state-finite action case, that the use of a translation function might be replaced by a transformation of the data.

He therefore introduced the return transformation

$$\tilde{r}(i,a) := r(i,a) - (1-\rho)^{-1}\{\bar{r}(i) - \sum_{j \in S} p^a(i,j)\bar{r}(j)\}$$

$$\tilde{p}^a(i,j) := p^a(i,j).$$

For the transformed problem we have

$$\tilde{\bar{r}}(i) \le \bar{r}(i) - (1-\rho)^{-1}\bar{r}(i) + (1-\rho)^{-1}\rho\bar{r}(i) + (1-\rho)^{-1}M_1\mu(i)$$

$$= (1-\rho)^{-1}M_1\mu(i) \quad \text{for all } i \in S$$

similarly

$$\tilde{\bar{r}}(i) \ge \bar{r}(i) - (1-\rho)^{-1}\bar{r}(i) - (1-\rho)^{-1}\rho M_1\mu(i)$$

$$= -(1-\rho)^{-1}M_1\mu(i) \quad \text{for all } i \in S.$$

Hence, we have

(1) $\quad \tilde{\bar{r}} \in W$

(2) $\quad \|\tilde{P}(f)\| = \|P(f)\| \le \rho_* < 1.$

This implies that the transformed problem can be handled without using a translation and fits into the model in WESSELS [28] (see also VAN NUNEN [21]). The question remains whether for all $i \in S$ and $\pi \in \mathcal{D}$ one has $\tilde{v}_i(\pi) = v_i(\pi) + u(i)$ for some function u on S which is independent of π. As a consequence of (1) and (2) we have that

$$\tilde{v}_i(\pi) = \mathbb{E}_i^\pi \sum_{n=0}^{\infty} \tilde{r}(X_n,Z_n) = \sum_{n=0}^{\infty} \mathbb{E}_i^\pi \tilde{r}(X_n,Z_n),$$

and that any π may be replaced by a randomized Markov decision rule, without any effect on $\tilde{v}_i(\pi)$.

$$\tilde{v}_i(\pi) = \sum_{n=0}^{\infty} \mathbb{E}_i^\pi [r(X_n,Z_n) - (1-\rho)^{-1}\bar{r}(X_n) + (1-\rho)^{-1} \sum_j p^{Z_n}(X_n,j)\bar{r}(j)]$$

$$= \sum_{n=0}^{\infty} \mathbb{E}_i^\pi \mathbb{E}_i^\pi [r(X_n,Z_n) - (1-\rho)^{-1}\bar{r}(X_n) + (1-\rho)^{-1}\bar{r}(X_{n+1}) \,|\, X_n,Z_n]$$

$$= \lim_{N\to\infty} \sum_{n=0}^{N} \{\mathbb{E}_i^\pi (r(X_n,Z_n) - (1-\rho)^{-1}\bar{r}(X_n) + (1-\rho)^{-1}\bar{r}(X_{n+1})\}$$

$$= \lim_{N\to\infty} \{ \sum_{n=0}^{N} \mathbb{E}_i^{\pi} r(X_n,Z_n) - (1-\rho)^{-1}\bar{r}(i) + (1-\rho)^{-1} \mathbb{E}_i^{\pi} \bar{r}(X_{N+1}) \}$$

$$= v_i(\pi) - (1-\rho)^{-1}\bar{r}(i),$$

where the third equality is allowed since

$$\mathbb{E}_i^{\pi} \{r^+(X_n,Z_n) + (1-\rho)^{-1}\bar{r}^-(X_n) + (1-\rho)^{-1} r^+(X_{n+1})\} < \infty,$$

and the final equality is achieved since

$$\lim_{N\to\infty} \mathbb{E}_i^{\pi} \bar{r}(X_{n+1}) = 0.$$

We will illustrate now how the results of LIPPMAN [15] can be embedded in our theory (see also VAN NUNEN and WESSELS [20]). Lippman proves the convergence of successive approximations at a geometric rate under the following conditions which are given in our notations.

CONDITIONS OF LIPPMAN. There exists a function $u : S \to [1,\infty)$, an integer $m \geq 1$, and constants $0 \leq \beta < 1$, $b > 0$ such that for all $i \in S$, $a \in A$

$$|r(i,a)|u^{-m}(i) \leq M$$

$$\sum_{j\in S} u^n(j)p^a(i,j) \leq \beta[u(i) + b]^m \qquad \text{for } n = 1,\ldots,m.$$

However, we then have for any $\rho_* \geq \beta$ and any

$$c \geq b[(\frac{\rho_*}{\beta})^{1/m} - 1]^{-1},$$

that for $\mu(i) := [u(i) + c]^m$

the following holds:

a) $\qquad \|P(f)\| \leq \rho_*$

and

b) $\|r(f)\| \leq M.$

So we can use for Markov decision processes as described by Lippman the latter simpler and more general conditions a and b.

The assumption 2.3.ii requires some transient behaviour of the processes involved. This may be characterized as *strong excessiveness*, i.e.

$$P(f)\mu \leq \rho_* \mu, \qquad \text{for all } f \in F$$

with $\rho_* < 1$ and μ a positive function on S.

For strong excessiveness several sufficient and necessary conditions can be given. In order to make assumption 2.3.ii more transparent and to relate the latter assumption to the assumptions of other authors we will give those conditions.

LEMMA 5.1. (VAN HEE and WESSELS [10]). *The process is strongly excessive with* $\mu(i) \geq \delta > 0$ *if and only if the lifetimes of the process are exponentially bounded, i.e.*

$$\mathbb{P}_i^{\pi}(X_n \in S) \leq a(i)\gamma^n$$

for all $i \in S$, $\pi \in M$, *where* $\gamma < 1$ *and* a *is a positive function on* S.

PROOF. "if" choose $\mu(i) := \sup_{\pi \in M} \sum_{n=0}^{\infty} \nu^n \mathbb{P}_i^{\pi}(X_n \in S, X_{n+1} \notin S)$ with $1 < \nu < \gamma^{-1}$ and $\rho_* := \nu^{-1}$, now it is straightforwardly verified that $P(f)\mu \leq \rho_* \mu$.
"only if" Note that for $\pi := (f_0, f_1 \ldots)$

$$\rho_*^m \mu \geq P(f_0)\ldots P(f_{n-1})\mu \geq \delta P(f_0)\ldots P(f_{n-1})e = \delta \mathbb{P}^{\pi}(X_n \in S)$$

with $e := \{1,1,\ldots\}$. □

LEMMA 5.2. (VAN HEE and WESSELS [10]). *The process is strongly excessive with* $\Delta \geq \mu(i) \geq \delta > 0$ *for some constants, if and only if the lifetimes of the process are exponentially bounded, uniformly in* $i \in S$, *i.e.*

$$\mathbb{P}_i^{\pi}(X_n \in S) \leq a\gamma^n \qquad (\textit{with } a > 0,\ 0 < \gamma < 1).$$

PROOF. The "if" part of the lemma follows straightforward, the "only if" part can be achieved by choosing e.g. $a(i) = \Delta\delta^{-1}$. □

LEMMA 5.3. (See VEINOTT [26], DENARDO [1], VAN HEE and WESSELS [10]).
The process is strongly excessive with $\Delta \geq \mu(i) \geq \delta > 0$ *for some constants* $\Delta \geq \delta > 0$ *if and only if the maximum expected lifetime is uniformly bounded in* $i \in S$, *i.e.*

$$\sup_{\pi\in M} \sum_{n=0}^{\infty} \mathbb{P}_i^{\pi}(X_n \in S) < M \quad \textit{for some} \quad M > 0, \textit{ and all } i \in S.$$

PROOF. Let $\mu(i)$ be the maximum expected lifetime if the process starts in state $i \in S$. So

$$\mu(i) := \sup_{\pi\in M} \sum_{n=0}^{\infty} \mathbb{P}_i^{\pi}(X_n \in S).$$

Clearly

$$\mu \geq e + P(f)\mu,$$

and

$$\mu \geq \frac{1}{M}\mu + P(f)\mu.$$

This yields

$$P(f)\mu \leq (1-\frac{1}{M})\mu.$$

So for $\rho_* = (1-\frac{1}{M})$, $\delta := 1$ and $\Delta := M$ the "if"-part will be clear. On the other hand if the process is strongly excessive with $\delta \leq \mu(i) \leq \Delta$, then the lifetimes are uniformly exponentially bounded and hence the maximum expected lifetimes are bounded. □

COROLLARY 5.1. *The following three assertions are equivalent.*

1) *The process is strongly excessive with* $0 < \delta \leq \mu(i) \leq \Delta$.
2) *The lifetimes of the process are uniformly exponentially bounded.*
3) *The maximum expected lifetimes of the process are bounded as function of the starting state.*

Note that the maximum expected lifetime $\ell(i)$ if the process starts in state $i \in S$ can be found as the smallest positive solution to

$$\ell \geq \sup_{f \in F} [e + P(f)\ell].$$

There is a close relation between strong excessivity and so called "N-stage" contraction. This relation is given in the following lemma.

LEMMA 5.4. (See VAN HEE and WESSELS [10]). *Let* u *be a positive function on* S *such that* $P(f)u \leq Mu$ *for some* $M > 0$ *and all* $f \in F$ *and suppose* $P(f_0)\ldots P(f_{N-1})u \leq \rho' u$, *with* $0 < \rho' < 1$ *(N-stage contraction) for all* $f_0,\ldots,f_{N-1} \in F$, *then there exists a positive function* μ *on* S *and* ρ_* *with* $0 < \rho_* < 1$, *such that*

$$P(f)\mu \leq \rho_*\mu \quad \textit{for all } f \in F.$$

PROOF. Choose ρ_* such that $\rho' < \rho_*^N < 1$ and choose

$$\mu := \sup_{\pi \in M} \sum_{n=0}^{\infty} \frac{1}{\rho_*^n} \mathbb{E}^{\pi} u(X_n). \qquad \square$$

As a consequence of the foregoing lemma we see that "N-stage" contraction in one norm (the u-norm) implies one-stage contraction in another norm (the μ-norm). A final characterization of strongly excessive processes is given in the following lemma which can again be found in VAN HEE and WESSELS [10]. This lemma gives a probabilistic characterization of the transient behaviour of the process.

LEMMA 5.5. *A process is strongly excessive if and only if there exists a partition* $\{S_k | k \text{ integer}\}$ *of* S *and numbers* $\alpha > 1$, $\beta \geq 1$, *such that for all* $\pi \in M$

$$\sum_{n=0}^{\infty} \mathbb{P}_i^{\pi}(X_n \in S_k) \leq \beta \min\{1, \alpha^{\ell-k}\} \quad \textit{for} \quad i \in S_\ell.$$

PROOF. First note that the lemma states that there is necessarily a drift to lower S_k or a drift out of the system.
The "if" part follows by defining

$$\mu := \sup_{\pi \in M} \mathbb{E}^{\pi} \sum_{n=0}^{\infty} u(X_n)$$

where $u(i) := (\alpha\varepsilon)^k$ if $i \in S_k$ with $0 < \varepsilon < 1$ and $\alpha\varepsilon > 1$. The "only if" part follows since

$$i \in S_\ell \iff \alpha^{\ell-1} < \mu(i) \leq \alpha^{\ell} \quad \text{with } 1 < \alpha < \rho_*^{-1}. \qquad \square$$

We conclude this section on the analysis of the basic assumptions by giving the relation between the use of weighted supremum norms (μ-norm) and the use of the "similarity transformation" as described by PORTEUS [24]. For the finite state space-finite action space situation Porteus proposed the following transformation of the original process. Let Q be a diagonal matrix with positive diagonal elements

$$Q := \begin{pmatrix} \mu^{-1}(1) & & \bigcirc \\ & \mu^{-1}(2) & \\ & & \ddots \\ \bigcirc & & \end{pmatrix}$$

Define

$$\tilde{r}(f) := Qr(f) \ ,$$

and

$$\tilde{P}(f) := QP(f)Q^{-1}.$$

Then the optimal return vector $\tilde{v}^*$ of the transformed problem is just equal to Qv^*.

Viz.

$$\begin{aligned} \tilde{v}^* &= \sup_{f \in F} (I-\tilde{P}(f))^{-1}\tilde{r}(f) = \sup_{f \in F} (I-QP(f)Q^{-1})^{-1}Qr(f) \\ &= \sup_{f \in F} [Q(I-P(f))Q^{-1}]^{-1}Q = r(f) = \sup_{f \in F} Q(I-P(f))^{-1}r(f) \\ &= Q \sup_{f \in F} (I-P(f))^{-1}r(f) = Qv^*. \end{aligned}$$

So the assumptions 2.3 can be replaced by the same assumptions with $\mu(i) = 1$ for the transformed problem.

REFERENCES

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

[2] DERMAN, C. & R.E. STRAUCH, *A note on memoryless rules for controlling sequential control processes*, Ann. Math. Statist. 37 (1966), 276-278.

[3] EPENOUX, F.D., *Sur un problème de production et de stochage dans l'aléatoire*, Rev. Tranc. Rech. Opère 14 1960, 3-16.

[4] GHELLINCK DE, G.T. & G.D. EPPEN, *Linear programming solutions for separable Markovian decision problems*, Management Sci. 13 (1967), 371-394.

[5] HARRISON, J., *Discrete dynamic programming with unbounded rewards*, Ann. Math. Statist. 43 (1972), 636-644.

[6] HASTINGS, N.A.J., *Some notes on dynamic programming and replacement*, Oper. Res. Quart. 19 (1968), 453-464.

[7] HASTINGS, N.A.J. & J. MELLO, *Test for nonoptimal actions in discounted Markov programming*, Management Sci. 19 (1973), 1019-1022.

[8] HASTINGS, N.A.J. & J.A.E.E. VAN NUNEN, *The action elimination algorithm for Markov decision processes*, In this volume.

[9] HEE VAN, K.M., *Markov strategies in dynamic programming*, Univ. of Technology Eindhoven, Dept. of Math. 1975 (Memorandum COSOR 75-20).

[10] HEE VAN, K.M. & J. WESSELS, *Markov decision processes and strongly excessive functions*, Univ. of Technology Eindhoven, Dept. of Math. 1975 (COSOR Memorandum 75-22).

[11] HINDERER, K., *Bounds for stationary finite stage dynamic programs with unbounded reward functions*, Hamburg, Institut für Math. Stochastik der Univ. Hamburg, June 1975, Report.

[12] HEE VAN, K.M. & J. VAN DER WAL, *Strongly convergent dynamic programming: some results*, Univ. of Technology Eindhoven, Dept. of Math. 1976 (COSOR Memorandum 76-26).

[13] HOWARD, R.A., *Dynamic programming and Markov processes,* Cambridge (Mass.) M.I.T. press, 1960.

[14] HÜBNER, G., *Improved procedures for eliminating suboptimal actions in Markov programming by the use of contraction properties,* Transactions of the 7th Prague Conference on Information theory, statistical decision functions, Random processes (including 1974 European Meeting of Statisticians) Academia Prague (To appear).

[15] LIPPMAN, S.A., *On dynamic programming with unbounded rewards,* Management Sci. 21 (1975), 1225-1233.

[16] MACQUEEN, J., *A modified dynamic programming method for Markovian decision problems,* J. Math. Anal. Appl. 14 (1966), 38-43.

[17] MACQUEEN, J., *A test for suboptimal actions in Markovian decision problems,* Operations Res. 15 (1967) 559-561.

[18] MINE, H. & S. OSAKI, *Markovian decision processes,* New York etc. Elsevier 1965.

[19] NUNEN VAN, J.A.E.E., *A set of successive approximation methods for discounted Markovian decision problems,* Zeitschrift für Operations Res. 20 (1976), 203-208.

[20] NUNEN VAN, J.A.E.E. & J. WESSELS, *A note on dynamic programming with unbounded rewards,* Eindhoven, Univ. of Technology, Dept. of Math. 1975, (Memorandum COSOR 75-13).

[21] NUNEN VAN, J.A.E.E., *Contracting Markov decision processes,* Amsterdam, Mathematisch Centrum, 1976 (Mathematical Centre Tract no. 71).

[22] NUNEN VAN, J.A.E.E. & J. WESSELS, *The generation of successive approximation methods for Markov decision processes by using stopping times,* In this volume.

[23] PORTEUS, E.L., *Some bounds for discounted sequential decision processes,* Management Sci. 18 (1971).

[24] PORTEUS, E.L., *Bounds and transformations for discounted finite Markov decision chains,* Operations Res. 23 (1975), 761-784.

[25] SCHÄL, M., *Conditions for optimality in dynamic programming and for the limit if N-stage optimal policies to be optimal,* Zeitschrift für Wahrscheinlichkeits Rechnung 32 (1975) 179-196.

[26] VEINOTT, A.F., *Discrete dynamic programming with sensitive discount optimality criteria,* Ann. Math. Statist. 40 1635-1660.

[27] WAL VAN DER, J. & J. WESSELS, *Successive approximation methods for Markov games,* In this volume.

[28] WESSELS, J., *Markov programming by successive approximations with respect to weighted supremum norms,* J. Math. Anal. Appl. 58 (1977).

[29] WESSELS, J. & J.A.E.E. VAN NUNEN, *Discounted semi-Markov decision processes:Linear programming and policy iteration,* Statistica Neerlandica 29 (1975), 1-7.

THE GENERATION OF SUCCESSIVE APPROXIMATIONS FOR MARKOV DECISION PROCESSES BY USING STOPPING TIMES

J.A.E.E. van Nunen
Graduate School of Management, Delft, The Netherlands

J.Wessels
Eindhoven University of Technology, Eindhoven, The Netherlands

1. INTRODUCTION

In [5] we introduced the standard successive approximation method for Markov decision processes with respect to the total expected reward criterion. In fact there exist some variants of this method. These variants differ in the policy improvement procedure: the standard procedure may be replaced by a Gauss-Seidel procedure (see e.g. HASTINGS [1], Kushner and KLEINMAN [4], an overrelaxation procedure (see REETZ [11] and SCHELLHAAS [13]) or some other variants (see VAN NUNEN [8]). In [14] it has been shown that such variants can be generated by stopping times. This approach has been generalized in [6]. In section 2 we will introduce the main idea of this approach.

Policy interation -with its several variants- as introduced by HOWARD [3] is usually not viewed upon as a successive approximation technique. However, in [7] it has been shown to be an extreme element of a class of extended successive approximation techniques, the so-called value-oriented methods. This approach has been combined in [9] with the stopping time approach. In [6] a further generalization has been given (mainly with respect to the conditions). Value-oriented methods will be treated in section 3.

Section 4 will be devoted to upper and lower bounds for the techniques presented in the earlier section. Furthermore some remarks on numerical aspects will be made.

In this paper we will use the same notations as in [5], however, in order to keep the proofs simple, we will work under somewhat stronger assumptions. In fact, our assumptions are the same as those in [6]. For details we will refer repeatedly to [6].

ASSUMPTIONS. Our assumptions are the same as the assumptions in [5], with assumption 2.3 (i) replaced by

(a) $$\exists_{M>0} \forall_{f\in F} \|r(f) - \bar{r}\| \leq M$$

(b) $$\sup_{\pi\in M} \mathbb{E}_i^{\pi} \sum_{n=0}^{\infty} |\bar{r}(X_n)| < \infty \qquad \text{for all } i \in S.$$

These stronger assumptions make the spaces V^- and W^- superfluous.

As remarked in [5] (remark 5.1) one may replace $\bar{r}$ in the assumptions (and definition of V) by a vector b with $b - \bar{r} \in W$. We will do so in this paper in order to facilitate referring to [6].

2. STOPPING TIMES AND SUCCESSIVE APPROXIMATIONS

In this section we will show that each stopping time characterized by a go ahead function δ for the sequence $\{X_n\}_{n=0}^{\infty}$ induces an operator U_δ on V, such that U_δ is monotone and (usually) contracting.

Furthermore all these contracting operators on V have the same unique fixed point v^*. So we have for any $v_0 \in V$ and any δ:

$$v_n := U_\delta v_{n-1} \in V \qquad \text{for } n = 1,2,\ldots$$

and $v_n \to v^*$.

DEFINIITON 2.1. A (randomized) *go ahead function* δ is a function which maps

$$G_\infty := \bigcup_{k=1}^{\infty} S^k \qquad \text{into } [0,1].$$

By Δ we denote the set of all go ahead functions.

$1 - \delta(s_0,s_1,\ldots,s_n)$ will be interpreted as the probability to stop the process at time n, given that $X_0 = s_0,\ldots,X_n = s_n$ and the process has not been stopped earlier.

DEFINITION 2.2.

(a) $\delta \in \Delta$ is said to be *nonrandomized* if $\delta(\alpha) \in \{0,1\}$ for all $\alpha \in G_\infty$;

(b) $\delta \in \Delta$ is said to be *nonzero* if $\delta(i) > \varepsilon > 0$ for some ε and all $i \in S$;

(c) $\delta \in \Delta$ is said to be *transition memoryless* if $\delta(\alpha)$ only depends on the the last two entries of α, for those α with at least two entries and satisfying $\delta(s_0,\ldots,s_k) \neq 0$ for all $k < n$, if $\alpha = s_0,\ldots,s_n$.

So for a transition memoryless go ahead function the stopping probability only depends on the most recent transition. The relevance of this notion will become clear in the course of this section.

EXAMPLES 2.1. Below some examples of nonzero go ahead functions will be given. These examples will be used repeatedly in this paper.

(a) Define the go ahead function δ_n $(n = 1,2,\ldots)$ by $\delta_n(\alpha) := 1$ if α contains less than $n + 1$ entries, otherwise $\delta_n(\alpha) := 0$. The go ahead function δ_n are nonrandomized, δ_n is only transition memoryless if $n = 1$.

(b) define δ_R by $\delta_R(s,s,\ldots,s) := 1$ for all s and all sequences of finite length, $\delta_R(\alpha) := 0$ otherwise.
δ_R is nonrandomized and transition memoryless.

(c) δ_H is defined by $\delta_H(s_0,\ldots,s_n) := 1$ if $s_0 < s_1 < \ldots < s_n$ (any n), otherwise $\delta_H(\alpha) := 0$.
δ_H is nonrandomized and transition memoryless.

(d) $\delta_r(i) = 1/2$ for all $i \in S$, $\delta_r(\alpha) := 0$ elsewhere.
δ_r is transition memoryless.

Since we introduced a probabilistic go ahead concept, we have to incorporate it in the probability space and measure. Therefore we extend the space $(S \times A)^\infty$ (see [5] section 2) to $(S \times E \times A)^\infty$, with $E := \{0,1\}$.

Furthermore the stochastic process $\{X_t,Z_t\}_{t=0}^{\infty}$ is extended to $\{X_t,Y_t,Z_t\}_{t=0}^{\infty}$, where $Y_t = 0$ as long as the process may go ahead.

Now any starting state i, go ahead function δ, and any decision rule π determine a probability measure on $(S \times E \times A)^{\infty}$ with the required properties in an obvious way (see [6] for details). This probability measure will be denoted by $\mathbb{P}_i^{\pi,\delta}$. Expectations will be denoted by $\mathbb{E}_i^{\pi,\delta}$. Note that $\mathbb{P}_i^{\pi,\delta}$ and $\mathbb{P}_i^{\pi}$ are equal for events which do not depend on the variables Y_t.

In fact the go ahead concept induces a stopping time.

DEFINITION 2.3. The random variable τ taking values in $\{0,1,\ldots,\infty\}$ is defined by

$$\tau = n : \Leftrightarrow Y_0 = \ldots = Y_{n-1} = 0 \text{ and } Y_n = 1$$

$$\tau = \infty : \Leftrightarrow Y_t = 0 \text{ for all } t = 0,1,\ldots .$$

τ is a randomized stopping time with respect to $X_0,X_1,\ldots$.
Now we will introduce our operators.

DEFINITION 2.4. For each $\delta \in \Delta$ and each strategy (= nonrandomized decision rule) π the operator L_δ^π on V is defined by

$$L_\delta^\pi v := \mathbb{E}^{\pi,\delta}\Big[\sum_{k=0}^{\tau-1} r(X_k,Z_k) + v(X_\tau)\Big] \qquad \text{for } v \in V, \text{ with } v(X_\tau) := 0$$

if $\tau < \infty$.

LEMMA 2.1. *L_δ^π is monotone and (for nonzero δ) strictly contracting on V. Therefore L_δ^π possesses a unique fixed point v_δ^π in V.*

PROOF. The contraction factor of L_δ^π is $\|P_\delta(\pi)\| := \rho_\delta^\pi$, where $P_\delta(\pi)$ is the matrix with (i,j) entry $\mathbb{P}_i^{\pi,\delta}(\tau < \infty, X_\tau = j)$. $\rho_\delta^\pi < 1$ if and only if δ is nonzero.

EXAMPLES. Take for π an arbitrary stationary strategy (f,f,...).

(a) $$(L_{\delta_1}^\pi v)(i) = r(i,f(i)) + \sum_j p^{f(i)}(i,j)v(j) \ ;$$

(b) $$(L^{\pi}_{\delta_R} v)(i) = [1 - p^{f(i)}(i,i)]^{-1}[r(i,f(i)) + \sum_{j\neq 1} p^{f(i)}(i,j)v(j)];$$

(c) $$(L^{\pi}_{\delta_H} v)(i) = r(i,f(i)) + \sum_{j<i} p^{f(i)}(i,j)(L^{\pi}_{\delta_H} v)(j) + \sum_{j\geq i} p^{f(i)}(i,j)v(j);$$

(d) $$(L^{\pi}_{\delta_r} v)(i) = \frac{1}{2}v(i) + \frac{1}{2}[r(i,f(i)) + \sum_{j} p^{f(i)}(i,j)v(j)];$$

(e) let δ be nonzero, then $v^{\pi}_{\delta} = v(\pi)$, independent of δ.

REMARK 2.1. If π is a nonstationary strategy then there exist values for $\{p^a(i,j),\ r(i,a)\}$ and go ahead functions δ' and δ'' such that $v^{\pi}_{\delta'} \neq v^{\pi}_{\delta''}$ (see lemma 5.1.7 in [6]).

We now come to the operators U_δ.

DEFINITION 2.5. The operator U_δ on V is defined by

$$U_\delta v := \sup_{\pi} L^{\pi}_{\delta} v \ ,$$

where the supermum is taken componentwise.

Note that L^{π}_{δ} has only been defined for strategies π, so the supermum is only taken over the strategies (= nonrandomized decision rules). Extension to the randomized decision rules would not affect the value of $U_\delta v$.

THEOREM 2.1. *Let $\delta \in \Delta$, then U_δ is monotone and (only for nonzero δ) strictly contracting with contraction radius $\nu_\delta := \sup_{\pi} \rho^{\pi}_{\delta}$. Therefore U_δ possesses (for nonzero δ) a unique fixed point. v^* is fixed point for all U_δ with δ nonzero.*

PROOF. For details we refer to the proof of theorem 5.2.1 in [6]. With respect to the last statement we remark:

$$U_\delta v^* \geq L^{\pi}_{\delta} v^* \geq L^{\pi}_{\delta} v(\pi) = v(\pi) \qquad \text{if } \pi = (f,f,\ldots) \ .$$

Since f may be chosen such that $v(\pi) \geq v^* - \varepsilon\mu$ ([5] theorem 3.1 (ii)), we obtain $U_\delta v^* \geq v^*$. If we had $U_\delta v^* > v^*$, then it would be possible to construct a strategy π' with $v(\pi') > v^*$.

This theorem serves as the basis for a δ-based successive approximation algorithm, since $v_n := U_\delta v_{n-1}$ converges in norm to v^* if $v_0 \in V$. In the definition of U_δ we take the supermum over all strategies. One would naturally prefer to restrict oneself to Markov strategies and even use the algorithm for constructing ε-optimal stationary strategies. The following theorem (for the proof we refer to [6] theorem 5.2.2 and 5.2.3) shows that the concept of transition memoryless go ahead functions plays a crucial role in this problem.

THEOREM 2.2.

(a) *Let δ be transition memoryless, $\varepsilon > 0$, $v \in V$.*
Then there exists a policy f, such that

$$L_\delta^f v \geq U_\delta v - \varepsilon\mu \quad .$$

(b) *Let δ be not transition memoryless, then there exist values for the parameters $\{p^a(i,j), r(i,a)\}$, such that for some $v \in V$ and some $\varepsilon > 0$ there is no $f \in F$ with*

$$L_\delta^f v \geq U_\delta v - \varepsilon\mu \quad .$$

Hence, if δ is transition memoryless we have

$$U_\delta v = \sup_f L_\delta^f v,$$

where the sup is not necessarily componentwise. Whereas if δ is not transition memoryless

$$\sup_f L_\delta^f v$$

may only be defined componentwise and may not be equal to $U_\delta v$. For non-zero and transition memoryless go ahead functions we now obtain the following iteration procedures

(a) *(if* $\sup_f L_\delta^f v$ *is attained for some* f*).*

Choose $v_0 \in V$, *define* $v_n := U_\delta v_{n-1}$ *and choose* f_n *such that* $v_n = L_\delta^{f_n} v_{n-1}$, then

(i) $\|v_n - v^*\| \leq \nu_\delta^n \|v_0 - v^*\|$

(ii) $\|v_n - v(f_n)\| \leq (1 - \nu_\delta)^{-1} \nu_\delta \|v_n - v_{n-1}\|$

(iii) *if* v_0 *satisfies* $U_\delta v_0 \geq v_0$, *then* $v_{n-1} \leq v_n \leq v(f_n) \leq v^*$.

(b) *Choose* $\varepsilon > 0$ *and* $v_0 \in V$ *with* $v_0 \leq U_\delta v_0 - \varepsilon\mu$.

Choose f_n $(n = 1,\ldots)$ *such that*

$$L_\delta^{f_n} v_{n-1} \geq \max\{v_{n-1}, U_\delta v_{n-1} - \varepsilon(1 - \nu_\delta)\mu\}$$

define

$$v_n := L_\delta^{f_n} v_{n-1},$$

then

(i) $\|v_n - v^*\| < \varepsilon$ *for* n *sufficiently large*

(ii) $v_{n-1} \leq v_n \leq v(f_n) \leq v^*$.

In fact, as in the case of δ_1, *more efficient lower and upperbounds can be obtained (see section 4).*

EXAMPLES 2.3. The examples 2.2 (a)-(b) induce numerically well-executable policy improvement procedures. In fact δ_1 induces the standard successive approximation technique based on Gauss-Jordan-iteration; δ_R induces Jacobi iteration (compare Porteus [10]); δ_H yields Gauss-Seidel iteration; other choices of δ yields overrelaxation and combinations of overrelaxation and Gauss-Seidel iteration (in this respect lemma 7.2.3 in [6] has interesting consequences).

3. VALUE ORIENTED METHODS

In the foregoing section we developed a whole class of policy improvement procedures or successive approximations techniques. As we saw in section 2, at the n-th stage of any policy improvement procedure the best estimate for the optimal strategy is the stationary strategy f_n. This makes the next policy improvement more efficient if the value v_n is nearer to $v(f_n)$. In fact the policy iteration techniques owe their high efficiency in the policy improvement part to the fact that they have $v_n = v(f_n)$. A disadvantage of policy iteration is in fact the computation of these v_n. However, there is an alternative in combining the advantages of policy iteration and successive approximations. Namely suppose f_n is chosen such that

$$L_\delta^{f_n} v_{n-1} = U_\delta v_{n-1} ,$$

then define

$$v_n := (L_\delta^{f_n})^\lambda v_{n-1} \qquad (\lambda \in \{1,2,\ldots,\infty\}) .$$

Note that

$$\lim_{\lambda\to\infty} (L_\delta^{f_n})^\lambda v_{n-1} = v(f_n) ,$$

so by the choice of λ we in fact determine how good v_n approximates $v(f_n)$. The choice $\lambda = 1$ gives the successive approximation of section 2, whereas the choice $\lambda = \infty$ gives for any transition memoryless and nonzero go ahead function a variant of the policy iteration technique.

Below we give a more formal treatment.

DEFINITION 3.1 Let δ be nonzero and transition memoryless and suppose that the $\sup_f L_\delta^f v$ is attained for some policy if $v \in V$. Furthermore we assume that we have a unique way of designating such a policy. We define the operators $U_\delta^{(\lambda)}$ on V for $\lambda = 1,2,\ldots,\infty$ by

$$U_\delta^{(\lambda)} v = (L_\delta^f)^\lambda v ,$$

if the sup in $U_\delta v$ is attained for f.

Note that

$$U_\delta^{(\infty)} v = \lim_{n\to\infty} (L_\delta^f)^n v = v(f) \ .$$

It does not seem revolutionary to conjecture that $v_n := U_\delta^{(\lambda)} v_{n-1}$ converges to v^* if $v_0 \in V$. However, one becomes somewhat more prudent as soon as one realizes that $U_\delta^{(\lambda)}$ is neither necessarily monotone, nor necessarily contracting as one can see in the following simple example for $\delta = \delta_1$, $S = \{1,2\}$, $\mu \equiv 1$, $A = \{1,2\}$: $p^1(i,2) = p^2(i,1) = 0.99$, $r(i,1) = 1$, other probabilities and rewards being zero.

Now one obtains for $v := (0,0)^T$, $w := (10,1)^T$ $\lim_{\lambda\to\infty} U_\delta^{(\lambda)} v = (100,100)^T$, whereas $\lim_{\lambda\to\infty} U_\delta^{(\lambda)} w = (0,0)^T$.

We will now prove that the proposed iteration step leads to a converging algorithm.

THEOREM 3.1. *Let the situation be such that* $U_\delta^{(\lambda)}$ *is defined and choose* $v_0 \in V$ *with* $U_\delta v_0 \geq v_0$.
Then $v_n := U_\delta^{(\lambda)} v_{n-1}$ *converges in norm to* v^*
and

$$\|v_n - v^*\| \leq \nu_\delta^n \|v_0 - v^*\|$$

$$v_{n-1} \leq v_n \leq v(f_n) \leq v^* \ ,$$

where f_n *is the policy (unique, possibly after tie breaking) which maximizes* $L_\delta^f v_{n-1}$.

PROOF. By assumption we have

$$L_\delta(f_1) v_0 = U_\delta v_0 \geq v_0 .$$

Hence

$$v_0 \leq L_\delta(f_1) v_0 \leq \ldots \leq [L_\delta(f_1)]^\lambda v_0 = v_1 \leq \lim_{n\to\infty} [L_\delta(f_1)]^n v_0 = v(f_1) .$$

Since $U_\delta v_1 = L_\delta(f_2)v_1 \geq L_\delta(f_1)v_1$, one obtains $v_1 \leq v_2 \leq v(f_2)$. By induction this gives $v_{n-1} \leq v_n \leq v(f_n) \leq v^*$. On the other hand $v_n \geq U_\delta^n v_0$, which tends to v^* for $n \to \infty$. Therefore $v_n \to v^*$ and

$$\|v_n - v^*\| \leq \|U_\delta^n v_0 - U_\delta^n v^*\| \leq v_\delta^n \|v_0 - v^*\|.$$

In the same way as in [5] for the standard algorithm one may obtain more sophisticated bounds (see section 4). Futhermore the assumption that the sup in $U_\delta v$ is attained can be weakened as in [5] by introducing approximations (in norm) of the sup. This can be extended in several ways. For a detailed description of these possibilities see [6].

As already stated, the case $\lambda = \infty$ represents a variety of policy iteration procedures. In fact the procedures (for any nonzero transition memoryless δ) generate sequences of policies with increasing value. Hence an optimal policy is obtained after a finite number of iterations if the state and action spaces are finite.

If $\delta = \delta_1$, then we have the standard policy iteration algorithm as introduced by HOWARD in [3] for the finite state, finite action discounted case. If $\delta = \delta_H$, then we have the Gauss-Seidel variant as introduced by HASTINGS [1].

4. SOME REMARKS ON NUMERICAL AND OTHER ASPECTS

For the algorithms based on the operators U_δ (section 2) and $U_\delta^{(\lambda)}$ (section 3) we proved goemetric convergence. However, the extrapolation based on the convergence rate only are usually not very good. As in the case of U_δ (see [5]) one can obtain better bounds rather easily. For the case the sup in $U_\delta v_n$ is attained and exactly computed in the algorithm based on $U_\delta^{(\lambda)}$ ($\lambda = 1,2,\ldots,$) we obtain, if $U_\delta v_0 \geq v_0$:

$$v_n + (1-\rho_{f_{n+1}})^{-1}\|L_\delta(f_{n+1})v_n - v_n\|_- \leq v(f_{n+1}) \leq v^* \leq$$

$$v_n + (1 - \nu_\delta)^{-1}\|L_\delta(f_{n+1})v_n - v_n\|,$$

where

$$\rho_f := \inf_f \mu^{-1}(i) \sum_j p^{f(i)}(i,j)\mu(j), \quad \|v\|_- := \inf_f \mu^{-1}(i)v(i) \quad .$$

For a more detailed description we refer to [6]. The proof in this case is completely similar to the proof in the case $\delta = \delta_1$.

For numerical experience it appears that value oriented methods can give a considerable gain in computational efficiency. This is especially true if the policy improvement requires many operations. Generally speaking one may say that U_δ-based successive approximations methods only need a smaller number of iterations to reach a near-optimal policy, however, the proof of this near-optimality requires relatively many additional iterations. So in quite a lot of iterations f_n does not change substantially. Therefore it is efficient to choose λ greater than one. In fact it is still more profitable to increase the value of λ in subsequent iterations. To give an idea of the gain in computational efficiency we mention that we found in a number of examples with $\delta = \delta_1$ a saving in computing time of 20 - 40% when we took $\lambda = 5$ instead of $\lambda = 1$ (in both situations we used a suboptimality test; the numbers of states ranged between 40 and 1000), see [8].

In all procedures (all δ and all λ) the standard suboptimality test is allowed and also the more sophisticated and more efficient suboptimality test which is described in the paper by HASTINGS and VAN NUNEN [2] in this volume.

Instead of defining δ-based operators U_δ one may transform the data in the problem and solve the transformed problem by the standard successive approximation methods. This approach has been presented by PORTEUS [10]. In our notation the transformation is

$$\tilde{r}(f) := \mathbb{E}^{f,\delta} \sum_{n=0}^{\tau-1} r(X_n, Z_n),$$

$$\tilde{P}(f) := P_\delta(f) \qquad \text{(see proof of lemma 2.1)} \quad .$$

By introducing the matrices $Q(f)$ with $Q(f)(i,j) := p^{f}(i,j)\delta(i,j)$ we obtian

$$\tilde{P}(f) = \sum_{k=0}^{\infty} Q^{k}(f)[P(f) - Q(f)] ,$$

$$\tilde{r}(f) = \sum_{k=0}^{\infty} Q^{k}(f)r(f) ,$$

being exactly Porteus' preinverse transformation. In fact we showed in section 2, that the transformed problem possesses the same optimal value vector as the original problem.

In fact some extension is possible with respect to the conditions under which the U_{δ}- and $U_{\delta}^{(\lambda)}$-based procedures converge. We mentioned already the kind of conditions of [5]. Another approach is in considering a fixed δ and require strict or N-stage contraction for U_{δ} on V or W. In [11] Reetz chooses such an approach for $\delta = \delta_{H}$. One might conjecture that -as in the case of δ_{1} (see [5]) -N-stage contraction implies 1-stage contraction with respect to a different norm.

REFERENCES

[1] HASTINGS, N.A.J., *Some notes on dynamic programming and replacement*, Oper. Res. Q. 19 (1968) 453-464.

[2] HASTINGS, N.A.J., J.A.E.E. VAN NUNEN, *The action elimination algorithm for Markov decision processes*, in this volume.

[3] HOWARD, R.A., *Dynamic programming and Markov decision processes*, Cambridge (Mass.), M.I.T.-Press, 1960.

[4] KUSHNER, H.J., A.J. KLEINMANN, *Accelerated procedures for the solution of discrete Markov control problems*, IEEE-Trans. on Aut. Contr. A.C. 16 (1971) 147-152.

[5] NUNEN VAN, J.A.E.E., J. WESSELS, *Markov decision processes with unbounded rewards*, in this volume.

[6] NUNEN VAN, J.A.E.E., *Contracting Markov decision processes*, Mathematical Centre Tract 71, Amsterdam 1976.

[7] NUNEN VAN, J.A.E.E., *A set of successive approximation methods for discounted Markovian decision problems,* Zeitschrift für Operations Res. 20 (1976) 203-208.

[8] NUNEN VAN, J.A.E.E., *Improved successive approximation methods for discounted Markov decision processes,* pp. 667-682 in A. Prékopa (ed.), Process in Operations Research, Amsterdam, North-Holland Publ. Comp. 1976.

[9] NUNEN VAN, J.A.E.E., J. WESSELS, *A principle for generating optimization procedures for discounted Markov decision processes,* pp. 683-695 in the same volume as [8].

[10] PORTEUS, E.L., *Bounds and transformations for discounted finite Markov decision chains,* Oper. Res. 23 (1975) 761-784.

[11] REETZ, D., *Solution of a Markovian decision problem by overrelaxation,* Zeitschrift für Operations Res. 17 (1973) pp. 29-32.

[12] REETZ, D., *A decision exclusion algorithm for a class of Markovian decision processes,* Zeitschrift für Operations Res. Vol. 20, (1976), pp. 125-131.

[13] SCHELLHAAS, H., *Zur Extrapolation in Markoffschen Entscheidungsmodellen mit Diskontierung,* Z.f.O.R. 18 (1974) pp. 91-104.

[14] WESSELS, J., *Stopping times and Markov programming, Transactions of the seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes* (including 1974 European Meeting of Statisticians), Academia, Praque (to appear).

SUCCESSIVE APPROXIMATION METHODS FOR MARKOV-GAMES

J. van der Wal
Eindhoven University of Technology, Eindhoven, The Netherlands

J.Wessels
Eindhoven University of Technology, Eindhoven, The Netherlands

1. INTRODUCTION

The main purpose of this paper is to investigate the following question: can the theory of successive approximations for Markov decision processes be extended to Markov games?

A preliminary answer to this question can be very short, since SHAPLEY [14] introduced already in 1953 successive approximations for Markov games, which were only introduced in 1957 for Markov decision processes [BELLMAN, 1]. However, for Markov decision processes, under relatively weak conditions, several types of successive approximation methods have been derived, together with sophisticated extrapolation procedures, see e.g. [8] and [9] in this volume. So the present paper will be mainly concerned with the question of generalizing this theory to Markov games. For an elementary treatment of dynamic programming in Markov games we refer to [20]. For other aspects of the theory of Markov games we refer to the recent bibliography and survey by PARTHASARATHY and STERN [10].

In section 2 the model will be introduced, also the finite stage case will be treated. We will allow unbounded rewards, but as in [8] and [9] contraction will be assumed. In section 3 the infinite stage case will be treated. In that section we will show that the standard successive approximations technique (extrapolations included) for the expected total reward criterion may be extended to contracting Markov games. In section 4

it will be shown that this is less easy for the policy iteration and other value oriented methods. However, a suitable extension will be presented. In section 5 positive Markov games with stopping actions for the second player will be considered. These games are not necessarily contracting. Section 6 is devoted to Markov games with the average reward criterion, and section 7 to the nonzero-sum case.

2. THE MODEL

As in [8] and [9], we consider a system, which is observed at descrete points in time $t = 0,1,2,\ldots$. The system can be in one of a countable number of states: $S = \{1,2,\ldots\}$. In each state i and at each time t the proceedings of the system may be influenced. This may be done by two players P_1 and P_2. Except in section 7 these players are supposed to have completely opposite aims. In each state i there are two finite (nonempty) sets K_i and L_i of allowed actions for P_1 and P_2 respectively. If at some time t the system is in state i and the players choose actions k and ℓ from K_i and L_i respectively, then this results in an immediate reward $r(i,k,\ell)$ for P_1 (to be paid by P_2) and it further results in a transition of the system to state j with probability $p(j|i,k,\ell)$. We suppose $\sum_j p(j|i,k,\ell) \leq 1$.

A *strategy* π for P_1 specifies for all times t and all possible histories h_t the probability $\pi_t(k|h_t)$ of choosing action k. Here the history h_t equals the sequence of states and actions in the past:

$$h_t = (s_0,k_0,\ell_0,\ldots,s_{t-1},k_{t-1},\ell_{t-1},s_t)$$

where s_τ is the state of the system at time τ and k_τ, ℓ_τ are the actions chosen at time τ by P_1 and P_2. If these probabilities only depend on s_t instead of h_t, then π is called a *Markov strategy*. If, moreover, π_t does not depend on t explicitly then π is called a *stationary strategy*.

Stationary strategies correspond to *policies*, where a policy f for P_1 is any function on S such that f(i) is a probability distribution on K_i, we will use the notation f both for policies and stationary strategies, by f(i,k) we denote the probability of $k \in K_i$. The set of policies for P_i is denoted by F, the set of strategies by π. Similarly one defines strategies $\gamma \in \Gamma$ and policies $g \in G$ for the player P_2.

NOTATIONS

$r(f,g)$ will denote for $f \in F$, $g \in G$ a real valued function on S with $r(f,g)(i) := \sum_{k \in K_i} \sum_{\ell \in L_i} f(i,k)g(i,\ell)r(i,k,\ell)$;

$P(f,g)$ will denote a nonnegative function on $S \times S$ with $P(f,g)(i,j) := \sum_{k \in K_i} \sum_{\ell \in L_i} f(i,k)g(i,\ell)p(j|i,k,\ell)$.

Functions on S and $S \times S$ respectively will be treated as columnvectors and matrices, with matrix products and matrix-vector products defined in the obvious way.

We will work under the following assumptions in this section and in sections 3-4.

ASSUMPTIONS

It is supposed that there is a positive function μ on S, which defines (as in [8] and [9]) a Banach space W_μ of vectors w with the norm $\|w\|_\mu = \sup_i |w(i)|\mu^{-1}(i)$, such that

(a) $\quad \|r(f,g)\|_\mu \leq M \quad$ for some M and all $\quad f \in F, \; g \in G.$

(b) $\quad \|P(f,g)\|_\mu \leq \rho < 1 \quad$ for some ρ and all $\quad f \in F, \; g \in G.$

In order to simplify our notations we will write W and $\| \; \|$ instead of W_μ and $\| \; \|_\mu$ whenever this is possible.

REMARK

These assumptions are somewhat more restrictive than those in [8] and even those in [9]. However, as shown in [21], assumption (a) may be weakened to

$$\bar{r} := \sup_{f \in F} \inf_{g \in G} r(f,g) \in W.$$

Furthermore the use of the Harrison translation does not present essential difficulties. To avoid technical details we will stick in this and the following two sections to the assumptions stated before.

As in [8] and [9] the transition probabilities may be defective, i.e. $\sum_j p(j|i,k,\ell) \leq 1$. This may be repaired by the introduction of an absorbing state. We will not do this explicitly.

Also the condition $\mathbf{K}_i$, $\mathbf{L}_i$ finite is not very essential. It may be replaced by $\mathbf{K}_i$, $\mathbf{L}_i$ compact, $p(j|i,k,\ell), r(i,k,\ell)$ continuous in k,ℓ.

A starting state i and strategies $\pi \in \Pi$, $\gamma \in \Gamma$ determine a stochastic process $\{(S_t,K_t,L_t)\}_{t=0}^{\infty}$ in an obvious way, where S_t is the state at time t and K_t, L_t the actions chosen at time t by P_1 and P_2 respectively. Probabilities referring to this process are denoted by $\mathbf{P}_i^{\pi,\gamma}$, expectations by $\mathbf{E}_i^{\pi,\gamma}$. If the index i is deleted, a column vector of probabilities or expectations is meant.

The assumptions guarantee (compare [8])

$$\mathbf{E}_i^{\pi,\gamma} \sum_{t=0}^{\infty} |r(S_t,K_t,L_t)| < \infty$$

and even

$$\| \mathbf{E}^{\pi,\gamma} \sum_{t=N}^{\infty} |r(S_t,K_t,L_t)| \| \leq M(1-\rho)^{-1}\rho^N \ .$$

Therefore the total expected reward (for P_1) is properly defined for any pair of strategies:

$$V(\pi,\gamma) := \mathbf{E}^{\pi,\gamma} \sum_{t=0}^{\infty} r(S_t,K_t,L_t) \ .$$

Strategies π^*, γ^* are said to be *optimal* if

$$V(\pi,\gamma^*) \leq V(\pi^*,\gamma^*) =: V^* \leq V(\pi^*,\gamma) \quad \text{for all } \pi \in \Pi,\ \gamma \in \Gamma \ .$$

V^* will be called the *value* of the game.

Analogous to [8] we introduce the following operators in W.

$$L(f,g)w := r(f,g) + P(f,g)w$$

$$Uw := \max_{f\in F} \min_{g\in G} L(f,g)w \ ,$$

with max-min taken componentwise.

Note that (Uw)(i) is the value of the matrixgame with entries

$$r(i,k,\ell) + \sum_j p(j|i,k,\ell)w(j) \ .$$

Now define $w_n := Uw_{n-1}$ $(n = 1,\ldots,T)$ for some $w_0 \in W$ and find policies f_n and g_n which satisfy for $n = 1,\ldots,T$

$$L(f,g_n)w_{n-1} \leq L(f_n,g_n)w_{n-1} \leq L(f_n,g)w_{n-1} \quad \text{for all } f,g \ .$$

Then we get the following result for the T-stage Markov game with terminal reward w (actually for this result the assumption $\rho < 1$ may be replaced by $\rho < \infty$):

THEOREM 1 *The T-stage Markov game with terminal reward* $w_0 \in W$, *i.e. the game with criterion function*

$$V_T(\pi,\gamma) := \mathbb{E}^{\pi,\gamma}\left[\sum_{t=0}^{T-1} r(S_t,K_t,L_t) + w_0(S_T)\right]$$

has the value w_T *and the strategies* π_T *and* γ_T, *which might be denoted by* $(f_T,\ldots,f_1)$, $(g_T,\ldots,g_1)$, *are optimal.*

The proof proceeds by induction. For details see [20]. □

This shows that in the finite-stage case optimal strategies may be found by dynamic programming or successive approximation. In the following section we will extend this result of the infinite-stage case. For that case our methods of proof bear more heavily on the assumptions. Expecially (compare e.g. [21], [17]) it is very essential that the assumptions imply

LEMMA 1 $L(f,g)$ and U are contracting with contraction radii $\|P(f,g)\|$ and ν respectively, with

$$\nu \leq \max_{f,g}\|P(f,g)\| \leq \rho < 1$$

As a consequence of this lemma the operators $L(f,g)$ and U possess unique fixed points. For $L(f,g)$ this fixed point is exactly $V(f,g)$, the criterion value for the stationary strategies f,g in the infinite-stage Markov game. For U this fixed point will be shown to be equal to the value V^* of the game.

3. THE ∞-STAGE MARKOV GAME

Let $w^* \in W$ be the unique fixed point of U in W, so $Uw^* = w^*$. Let f^*, g^* satisfy

$$L(f,g^*)w^* \leq L(f^*,g^*)w^* \leq L(f^*,g)w^* \qquad \text{for all } f \in F,\ g \in G\ .$$

We will prove the following result, which has been proved already in 1953 by SHAPLEY [14] for the finite state case with

$$\sum_j p(j|i,k,\ell) \leq \beta < 1:$$

THEOREM 2 *The stationary strategies* f^* *and* g^* *are optimal in the* ∞*-stage Markov game and* w^* *is the value of the game, i.e.* $V^* = w^*$.

PROOF. Obviously (theorem 1) the T-stage game with terminal reward w^* has value w^* and f^*, g^* are optimal stationary for that game.
Suppose P_1 plays f^* and P_2 an arbitrary strategy γ. Then for all T,

$$V(f^*,\gamma) \geq w^* - P^{(T)}w^* - \frac{M\rho^T}{1-\rho}\mu\ ,$$

where $P^{(T)}$ is the matrix with (i,j) entry $\mathbb{P}_i^{f^*,\gamma}(S_T = j)$.
One may show

$$P^{(T)}w^* \leq P^{(T)}\|w^*\|\mu \leq \rho^T\|w^*\|\mu\ .$$

Hence

$$V(f^*,\gamma) \geq w^* = V(f^*,g^*)\ .$$

Similarly one shows

$$V(\pi,g^*) \leq w^* \qquad \text{for all } \pi \in \Pi.$$

Hence, $w^* = V^*$. □

So the ∞-stage game possesses a value and optimal stationary strategies. It will now be investigated whether successive approximations produce ε-optimal stationary strategies and bounds for V^* which are arbitrarily close.

DEFINITION

$\pi_\varepsilon \in \Pi$ is called ε-*optimal* if $V(\pi_\varepsilon,\gamma) \geq V^* - \varepsilon\mu$ for all $\gamma \in \Gamma$.

$\gamma_\varepsilon \in \Gamma$ is called ε-*optimal* if $V(\pi,\gamma_\varepsilon) \leq V^* + \varepsilon\mu$ for all $\pi \in \Pi$.

An obvious way of approximating V^* is suggested by the fixed point property of U:

THEOREM 3 *Choose* $w_0 \in W$. *Then* $w_n := Uw_{n-1}$ $(n = 1,2,\ldots)$ *converges (in* μ*-norm) to* V^* *and one actually gets the following bounds*

$$w_n - \nu(1-\nu)^{-1}\|w_n - w_{n-1}\|\mu \leq V^* \leq w_n + \nu(1-\nu)^{-1}\|w_n - w_{n-1}\|\mu. \quad \square$$

However, somewhat better estimates can be given and one may simultaneously give bounds for the policies f_n and g_n (see section 2) found in the n-th iteration.

We first introduce some notations:

$$\lambda_n := \inf_i (w_n(i) - w_{n-1}(i))\mu^{-1}(i) \quad ,$$

$$\nu_n := \sup_i (w_n(i) - w_{n-1}(i))\mu^{-1}(i) \quad ,$$

$$a_n := \begin{cases} \sup\limits_{i,g} \mu^{-1}(i) \sum\limits_j P(f_n,g)(i,j)\mu(j) & \text{if } \lambda_n < 0 \quad , \\ \inf\limits_{i,g} \mu^{-1}(i) \sum\limits_j P(f_n,g)(i,j)\mu(j) & \text{if } \lambda_n \geq 0 \quad , \end{cases}$$

$$b_n := \begin{cases} \inf\limits_{i,f} \mu^{-1}(i) \sum\limits_j P(f,g_n)(i,j)\mu(j) & \text{if } \nu_n < 0 \quad , \\ \sup\limits_{i,f} \mu^{-1}(i) \sum\limits_j P(f,g_n)(i,j)\mu(j) & \text{if } \nu_n \geq 0 \quad . \end{cases}$$

THEOREM 4 *Choose* $w_0 \in W$, *define* $w_n := Uw_{n-1}$ $(n = 1,2,\ldots)$. *Let* $f_n \in F$, $g_n \in G$ *satisfy*

$$L(f,g_n)w_{n-1} \leq L(f_n,g_n)w_{n-1} = w_n \leq L(f_n,g)w_{n-1} \quad .$$

Then we have the following bounds for V^*, $V(f_n,g_n)$, $V(f_n,\gamma)$, *and* $V(\pi,g_n)$:

(a) $$w_n + a_n\lambda_n(1 - a_n)^{-1}\mu \le V^* \le w_n + b_n\nu_n(1 - b_n)^{-1}\mu \quad ,$$

(b) $$V(f_n,\gamma) \ge w_n + a_n\lambda_n(1 - a_n)^{-1}\mu \qquad \textit{for all } \gamma \in \Gamma,$$

(c) $$V(\pi,g_n) \le w_n + b_n\nu_n(1 - b_n)^{-1}\mu \qquad \textit{for all } \pi \in \Pi,$$

(d) $$w_n + a_n\lambda_n(1 - a_n)^{-1}\mu \le V(f_n,g_n) \le w_n + b_n\nu_n(1 - b_n)^{-1}\mu \ .$$

PROOF. a, d are direct consequences of b and c. The proof of c will be sketched (for more details proofs in somewhat different situations, see [17], [21]).

It suffices to prove c for stationary strategies π (compare [8]). Consider a policy (or stationary strategy) f.

$$L(f,g_n)w_{n-1} \le w_n \le w_{n-1} + \nu_n\mu \quad \text{(by definition)}$$

Hence

$$L^2(f,g_n)w_{n-1} \le L(f,g_n)[w_{n-1} + \nu_n\mu]$$

$$= L(f,g_n)w_{n-1} + \nu_n P(f,g_n)\mu \le w_n + \nu_n b_n\mu \ .$$

In this way one obtains

$$L^N(f,g_n)w_{n-1} \le w_n + \nu_n(b_n + \ldots + b_n^{N-1})\mu \ .$$

Hence

$$V(f,g_n) = \lim_{N\to\infty} L^N(f,g_n)w_{n-1} \le w_n + b_n\nu_n(1 - b_n)^{-1}\mu \ . \qquad \square$$

In this way the standard successive approximations technique may be extended to Markov games. On the upper- and lowerbounds of theorem 4 one may base tests for suboptimality (see [16] and for a more detailed treatment [12]).

In [9] it has been shown that an extensive class of successive approximation techniques may be generated by using stopping times. This also holds for Markov games. This will not be worked out in this paper, since the concepts and proofs are rather straightforward (for finite state discounted Markov games this has been worked out in [16] and [17]).

4. VALUE ORIENTED METHODS

In this volume VAN NUNEN and WESSELS [9] consider a set of value oriented methods for MDP which can be viewed upon as a special type of successive approximations method. One of these methods being Howard's policy iteration method. A straightforward generalization of Howard's method to Markov games has been proposed by POLLATSCHEK and AVI-ITZHAK [11]. This generalization may be formulated as follows

ALGORITHM

step 1 $v_0(i) := 0$ for all $i \in S$.

step 2 (Policy iteration). Determine policies f_n and g_n, such that

$$L(f,g_n)v_{n-1} \leq L(f_n,g_n)v_{n-1} \leq L(f_n,g)v_{n-1} \quad .$$

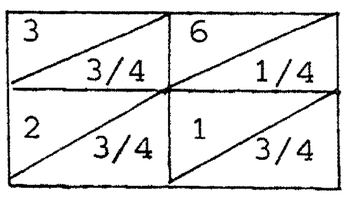
step 3 (Value determination) $v_n := V(f_n,g_n)$.

Pollatschek and Avi-Itzhak proved in the finite state case that the algorithm converges under the following condition

$$\max_i \sum_j \{\max_{k,\ell} p(j|i,k,\ell) - \min_{k,\ell} p(j|i,k,\ell)\} < 1 - \max_{i,k,\ell} \sum_j p(j|i,k,\ell).$$

In [18] essentially the following example has been given which proves that this algorithm does not converge in general for finite state discounted Markov games.

EXAMPLE



There is but one state. In this state both players have two actions. If P_1 picks action 2 and P_2 action 1 then P_2 pays P_1 2 units and the system stays in state 1 with probability 1/4, etc.

A policy f is completely determined by the probability f(1,1). If we apply the algorithm we find $f_1(1,1) = g_1(1,1) = 1$, $v_1 = 12$, $f_2(1,1) = g_2(1,1) = 0$, $v_2 = 4$, $f_3(1,1) = g_3(1,1) = 1$, $v_3 = 12$, etc. So the algorithm cycles without ever finding an optimal pair of strategies.

A somewhat more refined extension of Howard's method is the following. This extension has been inspired by Hoffman and Karp's method [5] for the average reward Markov game, and was presented by POLLATSCHEK and AVI-ITZHAK [11].

ALGORITHM (H,K)

step 1: Choose v_0 such that $Uv_0 \leq v_0$.
step 2: Determine Uv_n and a policy g_{n+1} with $L(f,g_{n+1})v_n \leq Uv_n$ for all f.
step 3: Determine $v_{n+1} := \max_f V(f,g_{n+1})$.

As in the case of MDP one may consider this algorithm as an extreme element of the following set of value oriented methods:

ALGORITHM (λ)

step 1: Choose v_0 such that $Uv_0 \leq v_0$.
step 2: Determine Uv_n and a policy g_{n+1} with $L(f,g_{n+1})v_n \leq Uv_n$ for all f.
step 3: Determine $v_{n+1} := U_{g_{n+1}}^{\lambda} v_n$, where the operator U_g is defined by $U_g v := \max_f L(f,g)v$.

For $\lambda = 1$ we have again the standard successive approximations method treated in section 3. For $\lambda = \infty$ we have Hoffman and Karp's algorithm. One may prove, using the monotonicity of the operators and $Uv_0 \leq v_0$, that v_n converges monotonically to V^*.

For the finite state case the proof is given in [18]. The extension of this proof to the case we deal with here is straightforward. One just has to prove by induction $V^* \leq v_n \leq Uv_{n-1} \leq v_{n-1}$, and $v_n \leq U^n v_0$. Since $\|U^n v_0 - V^*\| \leq \nu^n \|v_0 - V^*\|$ we also have $\|v_n - V^*\| \leq \nu^n \|v_0 - V^*\|$.

A possible extension is again the introduction of the stoppingtime-based L_δ and U_δ operators as has been executed in [17]. Another extension is that instead of using a fixed λ one may use a different λ_n in each iteration step. Note also, that if the first player has only one action in each state we get the set of value oriented methods presented by VAN NUNEN and WESSELS [9] for MDP.

5. STRICTLY POSITIVE MARKOV GAMES WITH STOPPING ACTIONS

In this section we will consider a type Markov game for which successive approximations still converge but where the U and L(f,g) operators are no longer strictly contracting. We release the assumptions of section 2 and replace them by: S, $\mathbb{K}_i$, $\mathbb{L}_i$ all finite

$$\sum_{j\in S} p(j|i,k,\ell) \le 1,\ r(i,k,\ell) > 0 \quad \text{for all } i,\ k \text{ and } \ell,$$

and moreover

$$\mathbb{L}_i^{STOP} := \{\ell \in \mathbb{L}_i \mid \sum_{j\in S} p(j|i,k,\ell) = 0 \quad \text{for all } k \in \mathbb{K}_i\} \neq \emptyset \quad \text{for all } i \in S.$$

By $\|v\|$ we mean standard maximum norm, $\|v\| = \max_{i\in S} |v(i)|$.
So all rewards are strictly positive and -since S, $\mathbb{K}_i$, $\mathbb{L}_i$ are finite- also bounded away from zero. The assumptions allow $V(\pi,\gamma)(i) = \infty$ for some π,γ and i. But since $\mathbb{L}_i^{stop}$ is nonempty P_1 can stop playing immediately in each state and thus restrict his loss to some finite amount.

As in section 2 we have the following lemma.

LEMMA. The n-stage game with terminal reward w_0 has the value $U^n w_0$ with optimal strategies $(f_n,\ldots,f_1)$ and $(g_n,\ldots,g_1)$ satisfying $L(f,g_k)w_{k-1} \le L(f_k,g_k)w_{k-1} =: w_k \le L(f_k,g)w_{k-1}$ for all f and g.

The problem remains to investigate how w_n behaves as n tends to infinity. Let $r^{STOP}(i)$ be defined as the value of the matrix game with entries $r(i,k,\ell)$, $k \in \mathbb{K}_i$, $\ell \in \mathbb{L}_i^{STOP}$. Then for any w_0 we obviously have $w_n \le r^{STOP}$, $n \in \mathbb{N}$, since in state i the second player may restrict his loss to $r^{STOP}(i)$ by choosing a good randomized action in $\mathbb{L}_i^{STOP}$.

We also have $0 \le U^{n-1}0 \le U^n 0$, $\le n = 2,3,\ldots$, hence $\lim_{n\to\infty} U^n 0$ exists. Call it w^*. Hence w^* is a fixed point of U : $w^* = Uw^*$.

THEOREM 5. *w^* is the unique fixed point of U and $U^n v \to w^*$ $(n \to \infty)$ for any $v \in \mathbb{R}^N$.*

PROOF. First we prove the uniqueness. Let u and v be fixed points of U which have (f_u, g_u) and (f_v, g_v) as optimal strategies in the one-stage game with terminal payoff u and v, respectively. Then

$$u = U^n u = L^n(f_u, g_u)u \geq L^n(f_v, g_u)u \geq L^n(f_v, g_u)v +$$

$$- I\!P^{f_v, g_u}(S_n \in S) \, \|u - v\| \geq v - I\!P^{f_v, g_u}(S_n \in S) \, \|u - v\| .$$

Obviously for all $i \in S$ $I\!P_i^{f_v, g_u}(S_n \in S) \to 0$ $(n \to \infty)$ since otherwise $V(f_v, g_u)(i) = \infty$ contracting $V(f_v, g_u) \leq u$. Hence $u \geq v$.
Similarly $u \leq v$ and thus $u = v$.
So it remains to show $U^n v \to w^*$ for any v. This follows from

$$U^n v \geq L(f_w^*, g_n) \ldots L(f_w^*, g_1)v \geq L(f_w^*, g_n) \ldots L(f_w^*, g_1)w^* +$$

$$- I\!P^{f_w^*, (g_n, \ldots, g_1)}(S_n \in S) \, \|v - w^*\|$$

$$\geq w^* - I\!P^{f_w^*, (g_n, \ldots, g_1)}(S_n \in S) \, \|v - w^*\| .$$

Again it is obvious that $I\!P^{f^*, (g_n, \ldots, g_1)}(S_n \in S) \to 0$ $(n \to \infty)$.
Therefore $\liminf_{n\to\infty} U^n v \geq w^*$. Similarly one may show $\limsup_{n\to\infty} U^n v \leq w^*$.
Hence $\lim_{n\to\infty} U^n v = w^*$. □

Here it is again possible to determine bounds for w^* using that $I\!P_i^{\pi, (g_n, \ldots, g_1)}(S_n \in S)$ converges to zero geometrically. This has been worked out in [19].

It is not necessary to assume that P_2 can quit playing in any state. It is sufficient to assume that P_2 can restrict his loss to some finite amount. This, more general case, has been treated by KUSHNER and CHAMBERLAIN [6].

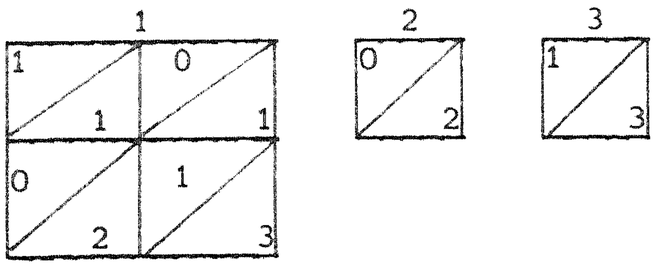
6. AVERAGE REWARD MARKOV GAMES

In this section the state space will be assumed to be finite. In the previous sections we have seen that it is possible to extend many of the results with respect to successive approximations in MDP to Markov games. In the average reward case however, we encounter substantial difficulties. This is illustrated by the following example called the big match. It is

due to GILLETTE [4] and studied by BLACKWELL and FERGUSON [3].

EXAMPLE.

$$S_1 = \{1,2,3\}, \quad \mathbb{K}_1 = \mathbb{L}_1 = \{1,2\}, \quad \mathbb{K}_2 = \mathbb{L}_2 = \mathbb{K}_3 = \mathbb{L}_3 = \{1\}.$$



If in state 1 P_1 picks action 2 and P_2 action 1 the payoff will be zero and the system moves to state 2, etc. So states 2 and 3 are absorbing.

One easily argues that, if P_2 takes in state 1 action 1 with probability 1/2, the average reward for P_1 will be 1/2, whatever, strategy P_1 uses. But it is not very clear how P_1 can guarantee himself an average payoff of 1/2. Any Markov strategy guarantees only 0. This is seen as follows: Let p(n) denote the probability that P_1 has picked action 2 before or on time n, and define $p := \lim_{n\to\infty} p(n)$. Now let $\varepsilon > 0$ be given arbitrarily and let N_ε be such that

$$p - p(N_\varepsilon) \leq \varepsilon .$$

Then P_2's strategy "play action 1 until time N_ε and action 2 thereafter" gives an average payoff of at most ε.

Blackwell and Ferguson show that P_1 can guarantee himself the average payoff $N/2(N+1)$ by playing strategy π_N defined as follows: Let P_2's first n choices be $\ell_1,\ldots,\ell_n$, $\ell_k \in \{1,2\}$, and let c_n be the excess of 1's over 2's among $\ell_1,\ldots,\ell_n$. Then take action 2 with probability $(N + c_n + 1)^{-2}$.

The difficulties here arise from the fact that there are strategies with more than one recurrent subchain.

Under the assumption that all pairs of stationary strategies induce an irreducible Markov chain (one recurrent subchain and no transient states) HOFFMAN and KARP [5] show that the game has a value and that their algorithm (H,K) from section 4 yields ε-optimal stationary strategies.

RIOS and YANEZ [13] consider the game with for all i,j,k and ℓ $p(j|i,k,\ell) \geq \geq \rho > 0$. (Then obviously Hoffman and Karp's irreducibility assumption is satisfied.) They show that in this case the standard successive approximations method converges. Recently TANAKA and WAKUTA [15], dealing with compact state and action spaces under appropriate continuity assumptions,

consider the following condition: $\mathbb{P}_i^{\pi,\gamma}(S_n = s_0) \geq \alpha > 0$ for some $s_0 \in S$ and all i, π and γ. And show that in this case the game has a value and that successive approximations converge.

7. NONZERO-SUM TWO-PERSON MARKOV GAMES

This section show that finite-stage two-person-nonzero-sum Markov games do have at least one Nash equilibrium point [7] which may be determined by successive approximations.

The main difference with the zero-sum games of the previous sections is that now we have two reward functions $r_1(x,k,\ell)$ and $r_2(x,k,\ell)$, where r_i denotes the reward for P_i, i = 1,2. Furthermore we have two terminal reward functions w_1 and w_2. As a result we have to define two total expected reward functions $V_1(\pi,\gamma)$ and $V_2(\pi,\gamma)$ for P_1 and P_2 respectively. Now we are looking for a Nash equilibrium pair (cf. [7]) for this game; that is a pair of strategies π^*, γ^* satisfying $V_1(\pi,\gamma^*) \leq V_1(\pi^*,\gamma^*)$ and $V_2(\pi^*,\gamma) \leq V_2(\pi^*,\gamma^*)$ for all π and γ. In bimatrix games (1-stage games) there can in general be more than one equilibrium pair.

The assumptions in this section are the following:

(i) S is countable, $\mathbb{K}_i$, $\mathbb{L}_i$ finite

(ii) There exist two positive vectors, μ_1 and μ_2 such that

$$\|r_1(f,g)\|_{\mu_1} \leq M_1 \quad \text{and} \quad \|P(f,g)\|_{\mu_1} \leq \rho_1 \quad \text{for all f and g}$$

$$\|r_2(f,g)\|_{\mu_2} \leq M_2 \quad \text{and} \quad \|P(f,g)\|_{\mu_2} \leq \rho_2 \quad \text{for all f and g}$$

where M_1, M_2, ρ_1, ρ_2 are real numbers.

Analogous to section 2 we define the operators L_1 and L_2 on W_{μ_1} and W_{μ_2} respectively by

$$L_i(f,g)w(x) := \sum_{k\in\mathbb{K}_x} \sum_{\ell\in\mathbb{L}_x} f(x,k)g(x,\ell)[r_i(x,k,\ell) + \sum_{j\in S} p(j|x,k,\ell)\, w(j)] , \quad i = 1,2.$$

Now for all $x \in S$, $w_1 \in W_{\mu_1}$, $w_2 \in W_{\mu_2}$ $L_1(f,g)w_1$ and $L_2(f,g)w_2$ determine a bimatrix game. Note that the assumption (ii) guarantees that $L_i(f,g)w_i$ lies again in W_{μ_i}.

Let us consider the n-stage game with terminal payoffs w_1 and w_2 for P_1 and P_2 respectively, with $w_i \in W_{\mu_i}$.
Now define $w_1^0 := w_1$, $w_2^0 := w_2$. Let f_n and g_n be a pair of policies satisfying $L_1(f,g_n)w_1^{n-1} \leq L_1(f_n,g_n)w_1^{n-1}$ and $L_2(f_n,g)w_2^{n-1} \leq L_2(f_n,g_n)w_2^{n-1}$ for all f and g and define $w_i^n := L_i(f_n,g_n)w_1^{n-1}$, $i = 1,2$. Then we have the following result: The pair of strategies $\pi_n := (f_n,\ldots,f_1)$, $\gamma_n := (g_n,\ldots,g_1)$ is a Nash equilibrium pair of strategies for the n-stage game under consideration. The proof of this statement goes along the same lines as the proof in [20] for zero-sum games, essentially using the monotonicity of the L operators

For infinite stage games there are a number of theorems about the existence of a pair of equilibrium strategies. See for example the survey paper by PARTHASARATHY and STERN [10]. BENIEST [2] considers a game with S finite, and

$$\sum_{j \in S} p(j|i,k,\ell) < 1 \quad \text{for all } i, k \text{ and } \ell,$$

under two different cooperation schemes and shows that in both cases there exist a unique pair of value vectors v_1^*, v_2^* which may be determined by successive approximations.

For the case of noncooperation the following example shows one of the problems we encounter when considering infinite stage games.

EXAMPLE.

5,5,3/4	1,7,3/4
7,1,3/4	2,2,3/4

There is only one state. If P_1 picks action 1 and P_2 action 2 then P_1 receives 1, P_2 7 and the system vanishes with probability 1/4, etc.
For each finite horizon game there is only one equilibrium pair of strategies, namely pick always action 2. In the infinite horizon game however there is still another equilibrium pair consisting of non-Markov strategies. Namely pick action 1 until your opponent has picked action 2, then continue to play action 2. One easily argues that if both players use this strategy this is indeed an equilibrium pair.

ACKNOWLEDGEMENT

With respect to this section, the authors gratefully acknowledge the contribution of their student Mr. Pulskens.

REFERENCES

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

[2] BENIEST, W., *Jeux stochastiques totalement cooperatifs arbitres*, Cahiers du Centre d'Etude de Recherche Operationelle 5 (1963), pp. 124-138.

[3] BLACKWELL, D., T.S. FERGUSON, *The big match*, Ann. Math. Statist. 39 (1968), pp. 159-163.

[4] GILLETTE, D., *Stochastic games with zero stop probabilities*, Contributions to the theory of games III, eds. M. Dresher, A.W. Tucker and D. Wolfe, Princeton University Press, Princeton New Jersey, 1957, pp. 179-187.

[5] HOFFMAN, A.J., R.M. KARP, *On nonterminating stochastic games*, Management Science 12 (1966), pp. 359-370.

[6] KUSHNER, H.J., S.G. CHAMBERLAIN, *Finite state stochastic games: Existence theorems and computational procedures*, IEEE, Trans. Automatic Control AC-14 (1969), pp. 248-255.

[7] NASH, J., *Non-cooperative games*, Ann. of Math. 54 (1951), pp. 286-295.

[8] NUNEN, J. VAN, J. WESSELS, *Markov decision processes with unbounded rewards*, In this volume.

[9] NUNEN, J. VAN, J. WESSELS, *The generation of successive approximation methods for Markov decision processes by using stopping times*, In this volume.

[10] PARTHASARATHY, T., M. STERN, *Markov games - a survey report* (1976), University of Illinois at Chicago Circle, Chicago Illinois.

[11] POLLATSCHEK, M.A., B. AVI-ITZHAK, *Algorithms for stochastic games*, Management Science 15 (1969), pp. 399-415.

[12] REETZ, D., J. VAN DER WAL, *On suboptimality in two-person zero-sum Markov games*, Memorandum COSOR 76-19, December 1976, Eindhoven University of Technology, Dept. of Math.

[13] RIOS, S., I. YANEZ, *Programmation sequentielle en concurrence*, Research papers in statistics, Ed. F.N. David, John Wiley and Sons, London - New York - Sydney 1966, pp. 289-299.

[14] SHAPLEY, L.S., *Stochastic games*, Proc. Nat. Acad. Sci. 39 (1953), pp. 1095-1100.

[15] TANAKA, K., K. WAKUTA, *On Markov games with the expected average reward criterion (II)*, Sci. Rep. Niigata Univ. Ser. A. 13 (1976), pp. 49-54.

[16] WAL, J. VAN DER, *The solution of Markov games by successive approximation*, Master's thesis, February 1975, Eindhoven University of technology, Dept. of Math.

[17] WAL, J. VAN DER, *Discounted Markov games;successive approximations and stopping times*, Intern. J. Game Theory, 6 (1977), pp. 11-22.

[18] WAL, J. VAN DER, *Discounted Markov games; the generalized policy iteration method*, J. Optim. Theory Appl., to appear.

[19] WAL, J. VAN DER, *Positive Markov games with stopping actions*, Memorandum 131, May 1976, Twente University of Technology, Dept. of Appl. Math.

[20] WAL, J. VAN DER, J. WESSELS, *On Markov games*, Statistica Neerlandica 30 (1976), pp. 51-71.

[21] WESSELS, J., *Markov games with unbounded rewards*, Memorandum COSOR 76-05, March 1976, Eindhoven University of Technology, Dept. of Math.

ON APPROXIMATE AND EXACT SOLUTIONS FOR FINITE STAGE DYNAMIC PROGRAMS

K.Hinderer
University of Karlsruhe, Karlsruhe, West Germany

G.Hübner
University of Hamburg, Hamburg, West Germany

1. INTRODUCTION

In this paper we are concerned a) with general thoughts on the appropriateness of finite-stage dynamic programs (DP's) for the description of real-world processes, b) with recently developed solution procedures for such DP's. In order to make the paper easily accessible, we restrict ourselves under b) to basic ideas and the simplest kind of procedures. For details and more sophisticated algorithms we shall refer to the pertinent literature. We shall also omit all questions of measurability; everything can be established in rigour if state and action spaces are standard Borel spaces.

2. NOTATION

The DP's to be considered are defined by the following data: the state space S; the action space A; the constraint set $D \subset S \times A$, whose s-section $D(s)$ is the set of admissible actions when being in state s; the transition law $P(s,a,B)$, being the probability of moving from s into $B \subset S$ under the influence of action a; the one-stage reward function $r : D \to \mathbb{R}$; the terminal reward function $v^0 : S \to \mathbb{R}$; the discount factor $\beta > 0$, not necessarily smaller than 1 (the case $\beta > 1$ is interpreted as "inflation").

The finite horizon will usually be denoted by N, and the corresponding DP will be denoted by DP_N. In order to give also the non-expert easy access to the paper, we assume for sections 2 to 5 that P is stochastic (rather than substochastic) and that r and V^0 are bounded. Extensions are indicated in the final section.

A (measurable) map f from S into A such that $f(s) \in D(s)$ for all $s \in S$ is called a *decision rule*. Let F denote the set of all decision rules; then F^N is the set of all N-stage (deterministic Markovian) *policies* $\pi = (f^N, f^{N-1}, \ldots, f^1)$. The expected N-stage reward under π when starting in s is defined as usual, and is denoted by $V_\pi^N(s)$; if $\pi = (f,f,\ldots$ $\pi = (f,f,\ldots,f)$, then we write V_f^N instead of V_π^N. Then

$$s \to V^N(s) := \sup_{\pi \in F^N} V_\pi^N(s)$$

is the *N-stage value function*; under our assumption it exists and is bounde bounded. The notions of optimality and ε-optimality of a policy are defined as usual. For $f \in F$ we use the abbreviations

$$r_f(s) := r(s,f(s)), \quad P_f(s,B) := p_f^s(B) := P(s,f(s),B).$$

Moreover, we put

$$(Pv)\ (s,a) := \int P(s,a,dt)v(t),$$
$$(P_f v)\ (s) := \int P_f(s,dt)v(t),$$

whenever the integrals exist.

We shall use the following three well-known operators L, U_f, U, defined on the set M of measurable bounded functions v on S: by

$$Lv := r + \beta Pv,$$
$$U_f v := r_f + \beta P_f v,$$
$$Uv := \sup_{f \in F} U_f v.$$

DEFINITION. The decision rule f is called an ε-*maximizer of the function* *Lv*, if

(1) $\quad Lv(s,f(s)) \geq \sup_{a \in D(s)} Lv(s,a) - \varepsilon,$

i.e. if

(1') $\quad L_f v \geq Uv - \varepsilon.$

In the special case that f is an ε-maximizer of LV^{n-1}, we call it an *ε-maximizer at stage n*.

Two basic results of dynamic programming may now be formulated as follows:

a) The *value iteration* holds:

$$V^n = U\,V^{n-1}, \quad n \in \mathbb{N}.$$

b) If f^n is an ε_n-maximizer at stage n, $1 \leq n \leq N$, then the policy $(f^N,\ldots,f^1)$ is $\sum_{n=1}^{N} \beta^{N-n}\varepsilon_n$-optimal.

It is well-known that both results hold under the assumption stated above.

3. SOME GENERAL THOUGHTS ON THE APPROPRIATENESS OF FINITE-STAGE DP's

This section consists of an amplification of ideas first expounded in HINDERER [7] and [9].

The time span over which real-world processes run is certainly finite, ever if measured e.g., in microseconds. Hence it seems that the a priori conceptual setting for *all* real-world sequential decision processes should be DP's with *finite horizon*. Nevertheless, there is a strong bias in the research literature (both theoretical and applied) towards infinite stage DP's, denoted in the sequel by DP_∞. The *main arguments usually presented in favor of* DP_∞ seem to be the following ones:

(i) Often the horizon is not known.

(ii) If N is "large", then DP_∞ is a good approximation for DP_N.

(iii) In many situations one should take in DP_N for V^0 the maximal

expected reward for the indefinite future (beginning at stage N), which implies that DP_N is (in general) formally equivalent to DP_∞.

(iv) With the exception of the inefficient plain value iteration many solution procedures developed for DP_∞ (such as policy improvement, linear programming, etc.) are not applicable to DP_N.

(v) The time-dependence of DP_N makes the theory less elegant and less interesting than that one for DP_∞.

We are going to make several *comments* on these arguments.

(α) Comment on (i): There are a fair number of problems, where the horizon is either known exactly or where there are available reasonable bounds for it. To the examples mentioned in HINDERER [7] and [9] we add the following ones: In allocation problems with a given number of objects the horizon is exactly known; in problems of replacement (e.g. of a medical electronic equipment) one usually will know upper and lower bounds N',N" for the time point N at which a technically improved version of the equipment will be installed.

(β) Comment on (ii): Behind the common usage of DP_∞ is hidden the expectation (often not stated explicitely), that using the decision rule f* from an optimal (or ε-optimal) stationary infinite stage policy (f*,f*,...) as long as our process really runs will be "nearly" optimal - at least if N is "large".

Two critical remarks have to be made here:

(β_1) One should not use DP_∞ as an approximation for DP_N unless one has some information about the goodness of the approximation; e.g. one should have an upper bound for the relative "error" $(V^N - V_{f^*})/|V^N|$. It is not sufficient to know that N is large, say 1000 or 10000, since the error will also depend substantially on the difference between V^0 and V^∞.

(β_2) There are well-known examples of the optimal stopping type, where both V^∞ and $V := \lim_{n\to\infty} V^n$ exist, but are not identical. In such cases one must expect (though one cannot be sure) that for given "large" horizon N, V is closer to V^N than V^∞, and if g is a maximizer of LV, we can hope that it is better to use g N times than to use the above-mentioned f* N times. Hence we propose to *abandon* (for practical purposes) DP_∞ *completely* and to use instead V and maximizers (or ε-maximizers) of LV; this latter procedure will be abbreviated by "$\lim_{n\to\infty} DP_n$".

The change from DP_∞ to lim DP_n (or more generally to span-fixed-points v* of U and maximizers of Lv*; cf. below) implies that some of the classical problems for DP_∞ - such as existence of stationary optimal policies - become irrelevant for applications, and we need not care about the well-known "paradoxes". However, the consequences of this change of view are much less radical as they may appear at first sight; cf. our comment (ζ) below.

(γ) DP_N makes sense for all values of the discount factor β, but lim DP_n need not exist. This may constitute a serious problem for the following reason: there is always (in particular if N is "large") some uncertainty about the numerical data one should use for the description of a given problem. (Are you sure, e.g., which of the discount factors 0.999, 1, 1.0005 is the appropriate one?) It would be desirable to find an N-stage policy π that behaves uniformly well for "small perturbations" of the data. Most easiest seems to be to keep the uncertainty for β under control by "solving" DP_N for a whole interval of possible values of β; but for some or all of them lim DP may not exist. In a typical situation β is very near to 1 and lim V^n exists only for $\beta < 1$.

(δ) Comment on (iii): There are quite a number of situations where the choice $V^0 := V^\infty$ (or rather $V^0 :=$ lim V^n) is not appropriate. If e.g. horizon N means that the production of some item is discontinued, then the scrap value V^0 of the production equipment will usually be much smaller than V^∞.

(ε) Comment on (iv): This argument seems to be valid for most of the classical algorithms developed for the solution of DP_∞. Note that most iterative algorithms for DP_∞ generate in the n-th step approximations v_n for V^∞ (starting often with $v_0 := 0$, irrespective of the value of V^0), which in general are quite different from V^n. Moreover, the suggestion of DERMAN/KLEIN [3] and DERMAN [2], p. 61, to convert DP_N into a DP_∞ (with an absorbing state) is useful only with additional constraints or for theoretical purposes. Solving the resulting DP_∞ by any of the well-known algorithms needs more computational effort than solving DP_N by plain value iteration. However, argument (iv) is not generally true since recently procedures for solving DP_N have been developed which are more efficient than plain value iteration. We shall report on those in sections 4 to 6 below.

(ζ) Comment on (v): This argument is only partially true. Instead of considering problems for DP_∞, for some methods (cf. sections 4B and 5A below) one has to consider similar problems for lim DP_n (or span-fixed-points of U), and quite often the elegant methods developed for the former model are also useful for the latter one. Moreover, new interesting problems arise, e.g. the existence and computation of fixed points of U for arbitrary discount factors; cf. section 4B below.

4. EXACT SOLUTIONS

It is well-known that some DP_N's with special structure (e.g. the classical investment problem of PHELPS [18] admit "explicit" exact solutions for arbitrary horizon N and arbitrary discount factor. There is no hope that equally simple results hold for "unstructured" problems, even if state and action space are very small; cf. the complicated explicit solution of Howard's simple toymaker problem given in HINDERER [9], p.235.

A. Decision exclusion

For DP_∞, the decision exclusion algorithm (DEA) - introduced by MACQUEEN [17] and developed further by several authors - constitutes a valuable exact method. Here we are going to report on results of HÜBNER [11], [12], who constructed (improved) versions of the DEA for DP_N which may considerably accelerate plain value iteration. At first we shall explain how general DEA's work.

DEFINITION. A subset D' of D is called *non-optimal**) *at stage n*, if

$$LV^{n-1}(s,a) < UV^{n-1}(s) \text{ for all } (s,a) \in D'. \tag{2}$$

The following property of such sets D' is obvious: All maximizers at stage n, that might be contained in a set $D'' \subset D$ are also contained in the (hopefully small) set $D'' - D'$.

Now, assuming that there exists for each stage n, $1 \le n \le N$, some maximizer, and that V^1 already has been computed, DEA works as follows:

---

* We gladly agree with Prof. Hasting's proposal to replace the usual (but misleading) "sub-optimal" by "non-optimal".

step 1: Look for a "large" set $D_1' \subset D$, that is non-optimal for *all* stages k, $1 < k \leq N$. Then $D_1 := D - D_1'$ contains all maximizers for stages k, $1 < k \leq N$, and V^2 may be computed by $V^2(s) = \sup_{a \in D_1(s)} LV^1(s,a)$

step n: Look for a "large" set $D_n' \subset D_{n-1}$, that is non-optimal for all stages k, $n < k \leq N$. Then $D_n := D_{n-1} - D_n'$ contains all maximizers for stages k, $n < k \leq N$, and V^{n+1} may be computed by $V^{n+1}(s) = \sup_{a \in D_n(s)} LV^n(s,a)$.

At step $N - 1$ we arrive at V^N.

Of course, the usefulness of DEA depends in a crucial way on efficient methods of finding "large" non-optimal sets. We need several preparations.

An important role will play the functional "*span*" defined on the linear space M by

$$(3) \qquad \operatorname{sp} v := \sup v - \inf v.$$

It is easily seen, that sp(.) is a seminorm on M, and sp v = 0 iff v is constant. BATHER [1] made ingenious use of the span in his investigation of the average cost criterion. From the well-known inequalities

$$(4) \qquad \beta \inf(v-w) \leq Uv - Uw \leq \beta \sup(v-w), \quad v,w \in M$$

(which possibly were first used - implicitly - by MACQUEEN [16] and which have been generalized and improved since then; cf. e.g. PORTEUS [19], lemma 3, and HÜBNER [12]) follows immediately

$$(5) \quad \operatorname{sp}(Uv-Uw) \leq \beta \operatorname{sp}(v-w); \qquad v,w \in M;$$

cf. BATHER [1], lemma 2.1. If e.g., S and A are finite, and the matrix $P((s,a),\{t\})$ has a positive column, inequality (5) may be improved as in (10) below.

Let G be a non-empty set of decision rules and put

$$(6) \qquad U_G v := \sup_{g \in G} U_g v.$$

LEMMA 1. (Cf. HÜBNER [12], Theorem 4.6). *For* $v,w \in M$ *and* $s,s' \in S$ *holds:*

$$(7) \qquad Uv(s) - Uw(s) - [U_G v(s') - U_G w(s')] \le \beta\gamma_G(s,s') \cdot sp(v-w),$$

where

$$(8) \qquad \gamma_G(s,s') := \tfrac{1}{2} \sup_{\substack{f\in F \\ g\in G}} \sup_{\substack{x\in M \\ |x|\le 1}} |p_f^s x - p_g^{s'} x| \le 1.$$

REMARKS. 1. There are other representations of $\gamma_G(s,s')$. It follows e.g. from HALMOS [5], p.124, problem 7, that

$$\gamma_G(s,s') = \tfrac{1}{2} \sup_{\substack{f\in F \\ g\in G}} |p_f^s - p_g^{s'}| \ (S),$$

where $|p_f^s - p_g^{s'}|$ is the total variation of the finite signed measure $p_f^s - p_g^{s'}$.

2. In applications, P will have a density p with respect to some measure μ. Then $\gamma_G(s,s')$ has the simpler form

$$(9) \qquad \gamma_G(s,s') = 1 - \inf_{\substack{f\in F \\ g\in G}} \int \min(p_f^s, p_g^{s'}) \, d\mu.$$

The special cases $G := F$ and $(G := \{g\}) \wedge (s=s')$ yield.

COROLLARY 2. *For* $v,w \in M$ *and* $g \in F$ *holds:*

$$(10) \qquad sp(Uv-Uw) \le \beta\gamma \cdot sp(v-w),$$

where

$$(11) \qquad \gamma := \sup_{s,s'\in S} \gamma_F(s,s') \le 1,$$

and

$$(12) \qquad Uv - U_g v - (Uw - U_g w) \le \beta\gamma_g \cdot sp(v-w),$$

(13) $$\gamma_g := \sup_{s\in S} \gamma_{\{g\}}(s,s) \le \gamma.$$

REMARKS. 1. If $(\mu_i, i\in I)$ is an arbitrary family of finite measures on some σ-algebra $\mathcal{F}$, then the set of measures μ on $\mathcal{F}$ for which $\mu \le \mu_i$ for all $i\in I$ has a (unique) largest element, which will be denoted by $\inf_{i\in I} \mu_i$; cf. SCHAEFER [23], II 8.3, and HÜBNER [12], p.81.
Now an upper bound γ' of γ may be found (HÜBNER [12],p. 34):

(14) $$\gamma' := 1 - \left(\inf_{(s,a)\in D} P(s,a,.)\right)(S) \le 1.$$

2. If P has a density p with respect to a σ-finite measure μ, then one gets the useful formulas

(15) $$\gamma = 1 - \inf_{\substack{(s,a)\in D \\ (s',a')\in D}} \int \min(p(s,a,\cdot),p(s',a',\cdot))\,d\mu$$

and

(16) $$\gamma' = 1 - \int \operatorname*{ess\,inf}_{(s,a)\in D} p(s,a,\cdot)\,d\mu.$$

It is not difficult to infer from (15) and (16), that e.g. in Howard's toymaker example we have $\gamma = \gamma' = 0.4$, whereas in Howard's automobile replacement example one gets only $\gamma = \gamma' = 1$.
3. The number γ plays an important role as so-called "coefficient of ergodicity" in limit theorems for Markov processes; cf. SENETA [26] for a review on this and related topics.

Making essential use of corollary 2, one may obtain e.g. the following (sharpened) finite-stage version of MACQUEEN's [17] DEA.

THEOREM 3 (cf. HÜBNER [12], Theorem 4.7). *Assume* $1 \le n < N < \infty$.
The set D'_n *of those* $(s,a) \in D$ *for which*

(17) $$LV^{n-1}(s,a) < V^n(s) - \operatorname{sp}(V^n - V^{n-1}) \cdot \sum_{\nu=1}^{N-n} (\beta\gamma)^\nu,$$

is non-optimal for all stages k, $n < k \le N$.

More sophisticated DEA's are available, and also the modified DEA's of HASTINGS/MELLO [6] and PORTEUS [20] have finite-stage counterparts. For more details we must refer to HÜBNER [12].

B. Turnpike horizons

If state and action spaces are finite, it may happen that a DEA stops after n steps, namely when the remaining constraint set D_n consists of a single decision rule f*, which must then be a maximizer for all stages $k > n$. It follows, that for the N-stage problem (N>n) we may use f* for the first N - n steps, and afterwards some n-stage (non-stationary) policy. This phenomenon was observed a long time ago in deterministic DP's. The theoretical investigation of the stochastic case was initiated by SHAPIRO [27], who called f* a "turnpike".

DEFINITION. Assume that there exists for each $\varepsilon > 0$ and each $n \in \mathbb{N}$ an ε-maximizer at stage n, and let G be a non-empty set of decision rules. Then

(18) $$N^*(G) := \inf\{k \in \mathbb{N} : \text{for all } \varepsilon > 0 \text{ and all } n \geq k \text{ the set } G \text{ contains an } \varepsilon\text{-maximizer at stage } n\}$$

is called the *turnpike horizon of G*.
If G is finite (e.g. if S and A are finite), then

$$N^*(G) = \inf\{k \in \mathbb{N} : \text{for all } n \geq k \text{ the set } G \text{ contains a maximizer at stage } n\}.$$

The aim is to find a "small" set G which has finite turnpike horizon and to find in addition a good upper bound n_0 for $N^*(G)$. If this has been achieved, we know that the value iteration simplifies for all stages $n \geq n_0$ to $V^n = U_G V^{n-1}$; and if G happens to consist of a single decision rule g, then we know that g is a maximizer at all stages $n \geq n_0$. Combining (12) with ideas of SHAPIRO [27], we get the following slight extension of theorem 2.3 in HINDERER/HÜBNER [10].

THEOREM 4. *Let* $w \in M$, *and assume that the set* G *of maximizers of* Lw *is not empty. Then*

$$(19)\qquad N^*(G) \le n_0(G) := \inf\{k \in \mathbb{N} : \beta\gamma \cdot sp(w-V^{n-1}) < \rho(G) \text{ for all } n \ge k\}$$

where

$$(20)\qquad \rho(G) := \inf_{g \notin G} \sup (Uw - U_g w).$$

REMARKS 1. The number $\rho(G)$ is a kind of measure how well the decision rules outside G behave with respect to the maximization of the function $g \to U_g w$.2. Theorem 4 does not make an assertion about finiteness of $N^*(G)$.

For the application of theorem 4 one needs efficient upper estimates for $sp(w-V^{n-1})$. If $\beta < 1$, S and A finite, $w := V^{\infty}$ (which is the unique fixed point of U), and if we insert into (19) the rough estimate

$$\gamma \cdot sp(V^{n-1}-w) \le 2 \|V^{n-1}-w\| \le 2\beta^{n-1}\|V^1-V^0\|/(1-\beta)$$

(where $\|\cdot\|$ is the sup-norm), then we obtain the bound $n_1(G)$ of SHAPIRO [27]. However sometimes considerable extension and improvement is possible: Following BATHER [1], we call $v^* \in M$ a *span-fixed-point* of U, if $sp(Uv^*-v^*) = 0$, i.e. if Uv^*-v^* is constant; note that v^* is an ordinary fixed point of U iff $\|Uv^*-v^*\| = 0$ for some norm on M, and then v^* is also a span-fixed-point of U. If v^* is a span-fixed-point of U, then obviously $sp(v^*-V^{n-1}) = sp(U^{n-1}v^*-U^{n-1}V^0)$, and hence theorem 4 and (10) lead to

THEOREM 5. *Assume*

(i) *There exists a span-fixed-point* v^* *of* U *such that the set* G *of maximizers of* Lv^* *is not empty.*

(ii) $\rho(G) > 0$ and $\beta < \gamma^{-1}$.

Then

$$N^*(G) \le n_2(G) := \inf\{n \in \mathbb{N} : (\beta\gamma)^n < \rho(G)/sp(v^*-V^0)\} =$$

$$= 1^+ \left[\log \frac{\rho(G)}{sp(v^*-V^0)} / \log(\beta\gamma)\right]^+ < \infty$$

where $[\cdot]$ *is the greatest integer function.*

It is important for the application of theorem 5, that U may have span-fixed-points even if V^{∞} does not exist. Thus e.g. in HOWARD's toy-maker example there exists a span-fixed-point of U for each $\beta > 0$ with the exception of $\beta = 10$. Table 1 shows values of N*(g), SHAPIRO's bound $n_1(g)$ and our bound $n_2(g)$ for the unique maximizer g of the unique span-fixed-point of U for

a) HOWARD's toymaker problem with $V^0 = (105,100)$,

b) HOWARD's automobile replacement problem with V^0 equal to the "trade-in value".

	toymaker			automobile replacement
	$\beta = 0.98$	$\beta = 1$	$\beta = 1.1$	$\beta = 0.97$
$n_1(g)$	320	∞	∞	345
$n_2(g)$	3	3	3	185
$N^+(g)$	2	2	2	29

Table 1.

It should be noted, that sometimes the bound $n_2(G)$ in theorem 5 can be considerably improved with little additional computation: We may first compute by value iteration (and DEA) V^k for some "small k" and formulate our DP_N as DP_{N-k}, where V^0 is replaced by V^k. This practical device is also useful for the approximate solutions of section 5 below, and we shall call it *"solving the tail of length k"*.

In HINDERER/HÜBNER [10] there is also shown, how the bounds for turnpike horizons may be improved if the DP has an absorbing set, i.e. a set $B \subset S$ such that $P(s,a,B^C) = r(s,a) = V^0(s) = 0$ for $s \in B$; application is made there to a lot size problem. REETZ [21] deduces turnpike theorems (for finite S) using the semi-norm

$$V \to \min_{\alpha \in \mathbb{R}} \sum_{s \in S} |v(s) - \alpha| \cdot$$

instead of the span. DIRICKX [4] considers turnpike theorems for deterministic DP's and $\beta > 1$. SHAPIRO [27] has also a discussion of the case $\beta = 1$. LEMBERSKY [14] studied turnpike-theorems for continuous-time DP's.

For the applications of theorem 5 as well as of the approximation

method of section 5.A below, one must know span-fixed-points of U. If $\beta < 1$, then $V = V^{\infty}$ is the unique fixed-point of U, and many solution procedures are known. For $\beta = 1$ methods developed for the average reward criterion are applicable. In HÜBNER [13] a procedure similar to that of MACQUEEN [16] is developed for finding the unique fixed point of U for values $1 < \beta < \gamma^{-1}$ (provided that $\gamma < 1$).

5. APPROXIMATE SOLUTIONS

Solving DP_N approximately will mean the following: (i) One must find a "simple" N-stage policy π^* which has a chance to be "good". (ii) In order to judge the goodness of π^*, one needs upper bounds for the error $V^N - V^N_{\pi^*}$ or - more appropriate - for the relative error

$$(21) \qquad (V^N - V^N_{\pi^*}) / |V^N| .$$

Upper and lower bounds for V^N may be of less direct importance for practice, but they are essential for our method of getting bounds for (21), cf. subsection A below. (iii) If one has decided to use π^*, one should know (e.g. for financial planning beyond stage N) bounds for $V^N_{\pi^*}$.

A. Approximation of DP_N by span-fixed-points v* of U and maximizers of Lv*

Here we assume that we know some span-fixed-point v* of U, the constant $c := Uv^* - v^*$ and some maximizer f* of Lv*. The bounds we derive will be in terms of v*, V^0 (or V^k for "small" k).

If $\beta < 1$ (and therefore V^{∞} is a span-fixed-point of U with $c = 0$) then one may derive a bound for $V^{\infty} - V^N$ from the fixed point theorem for contractions:

$$(22) \qquad \|V^{\infty} - V^N\| \leq (1-\beta)^{-1} \beta^N \|V^1 - V^0\| ,$$

which is greatly improved, using (4), to

$$(23) \qquad (1-\beta)^{-1} \beta^N \inf(V^1 - V^0) \leq V^{\infty} - V^N \leq (1-\beta)^{-1} \beta^N \sup(V^1 - V^0) .$$

(cp. section 5B below). Both bounds may be improved further by "solving

the tail of length k" and introducing $v^k - v^{k-1}$ and β^{N-k+1} in (22) and (23) instead of $v^1 - v^0$ and β^N (resp.). Moreover, (22) and (23) do not take into account that v^∞ is assumed to be known and therefore may be used in estimates for $v^\infty - v^N$, as it is the case in the sequel.

At first, (22) may be easily improved and extended to the case of arbitrary β by observing that U is Lipschitz-continuous with module β; putting

$$c_N := c \sum_{\nu=0}^{N-1} \beta^\nu \tag{24}$$

we get

$$\|v^*+c_N-v^N\| \le \beta^N \|v^*-v^0\| . \tag{25}$$

It is not difficult to obtain a similar bound for $v^N-v^N_{f^*}$, when f^* is a maximizer of Lv^*. Since $U_{f^*}v^* = Uv^*$, v^* is also a span-fixed-point of U_{f^*} and $U_{f^*}v^* - v^* = c$. Since $v^N_{f^*} = U^N_{f^*}v^0$, we get in analogy to (25) bounds for $v^N_{f^*}$ in the form

$$\|v^*+c_N-v^N_{f^*}\| \le \beta^N \|v^*-v^0\| , \tag{26}$$

and combining (25) and (26) we have

$$0 \le v^N - v^N_{f^*} \le 2\beta^N \|v^*-v^0\| . \tag{27}$$

Finally, using the fact that $x \le y \le z$ implies $|y| \ge \max(x^+,z^-)$ for $x,y,z \in \mathbb{R}$, we may combine (27) and (25) to obtain an upper bound for $(v^N-v^N_{f^*})/|v^N|$, but we omit the somewhat complicated expression; numerically it is easiest evaluated by first computing (27) and (25) separately.

The bounds (25) to (27) may be improved once more by using inequality (4) to obtain

THEOREM 6. (HINDERER [9], theorems 5.4 and 6.2). *If v^* is a span-fixed-point of U and if f^* is a maximizer of Lv^*, then both $w := v^N$ and $w := v^N_{f^*}$ satisfy*

$$\inf(v^0-v^*)\beta^N \le w - v^* - c_N \le \sup(v^0-v^*)\cdot\beta^N, \tag{28}$$

and moreover

$$(29) \qquad 0 \le v^N - v^N_{f*} \le \beta^N \, sp\,(v^0 - v*).$$

The usefulness of theorem 6 (and similarly of the subsequent theorems) is enlarged by the abovementioned device of "solving the tail of length k". Thus, if $\sigma \in F^k$ is optimal and $\pi := (f*,\dots,f*,\sigma)$ then both $w := v^N$ and $w := v^N_\pi$ satisfy

$$(30) \qquad \inf(v^k - v*)\beta^{N-k} \le w - v* - c_N \le \sup(v^k - v*)\beta^{N-k};$$

and

$$(31) \qquad 0 \le v^N - v^N_\pi \le \beta^{N-k} \, sp(v^k - v*);$$

and the bounds (30) and (31) are improving with increasing k.

The quality of (30) and (31) has been tested in HINDERER [9] for Howard's toymaker with $v^0 = (105,100)$, and the results are very satisfying, with respect to relative errors in a whole two-sided neighbourhood of $\beta = 1$.

For Howard's automobile replacement problem with $\beta = 0.97$ and v^0 equal to the trade-in value (cp. table 1) the right hand bound of (31) may be found in table 2 for $k = 0,5,10,15,20$ and $N = 40,200$ (note that $v^0 - v^\infty \approx 5300$).

k	N = 40	N = 200
0	37.3	0.285
5	27.8	0.213
10	13.6	0.104
15	5.86	0.045
20	2.97	0.023

Table 2

B. Extrapolation from DP_k to DP_N ($k << N$)

In real problems it will often happen, that span-fixed-points of U are not available (they may even not exist). Then one may try to carry

through value iteration for a "small" number, say k, of steps and try to extrapolate from V^k to V^N.

If f^n turns out to be a maximizer at stage n for $1 \le n \le k$, then one may find again with the help of (4) an estimate for the performance of the "simple" N-stage policy $\pi^* := (f^k,\ldots,f^k,f^{k-1},\ldots,f^1)$; in the case $k = 1$ π^* is myopic.

THEOREM 7. (HINDERER [9], theorems 4.5 and 6.1). *If* $1 \le k \le N$, *then both* $w := V^N$ *and* $w := V^N_{\pi^*}$ *satisfy*

$$(32) \qquad \inf(V^k-V^{k-1}) \cdot \sum_1^{N-k} \beta^\nu \le w - V^k \le \sup(V^k-V^{k-1}) \cdot \sum_1^{N-k} \beta^\nu,$$

and moreover

$$(33) \qquad 0 \le V^N - V^N_{\pi^*} \le \mathrm{sp}(V^k-V^{k-1}) \cdot \sum_1^{N-k} \beta^\nu,$$

and the bounds are improving with increasing k.

We remark, that (32) is easily derived from PORTEUS [19], lemma 5, which was used there to obtain bounds for V^∞. Also (32) and (33) have been tested for Howard's toymaker. The bounds were still good, but less precise than those obtained from (28) and (29); but it should be noted, that the latter bounds require the computation of v*.

The proof of (32) is easily accomplished by applying (4) e.g. to the representation

$$(34) \qquad V^N - V^k = \sum_{n=1}^{N-k} (U^n V^k - U^{n-1} V^k).$$

6. EXTENSIONS, IMPROVEMENTS AND FURTHER RESULTS

(i) Using representations of $V^N - V^k$ similar to (35), one may derive an infinite set of bounds for $V^N - V^k$, from which, however, all but a finite number may be discarded. Among several classes of computationally simple bounds of that type "best" elements can be identified, and finally the asymptotic behaviour of error bounds can be studied. Details are given in HINDERER [9].

(ii) If the DP has a non-empty absorbing set, then the bounds of section 5 may be improved; cf. HINDERER [7].

(iii) Our assumption that r and v^0 are bounded may be considerably relaxed. One of the more easier yet important generalizations requires only that the DP has a *bounding function*, i.e. a (measurable) function $b : S \to \mathbb{R}_+$ such that for some constants c_1 and c_2

$$\alpha) \qquad |r| \leq c_1 b,$$

$$\beta) \qquad |v^0| \leq c_2 b,$$

$$\gamma) \qquad \delta := \sup_{f \in F} \sup_{s \in S} P_f b(s)/b(s) < \infty.$$

It seems that (special) bounding functions were first used in LIPPMAN [15], whereas a systematic study of them (for DP_∞) is made in WESSELS [29]. It was shown in HINDERER [8], that the approximation methods may be extended to DP's with a bounding function. Let us just mention the following sample result: If v* is a fixed point of U and if $\inf(v^0 - v^*) \leq 0 \leq \sup(v^0 - v^*)$, then

$$(35) \qquad b(s)(\beta\delta)^N \inf \frac{v^0 - v^*}{b} \leq v^N(s) - v^*(s) \leq b(s)(\beta\delta)^N \sup \frac{v^0 - v^*}{b}.$$

(iv) The recent work of HÜBNER [12] constitutes a systematic study of the ideas developed in sections 4 and 5 above, in particular on decision exclusion and extrapolation. Let us mention but a few of the problems dealt with there:

a) "Generalized DP's" are considered: The transition law P need not be stochastic but only bounded, i.e.

$$(36) \qquad \bar{\alpha} := \sup_{(s,a) \in D} P(s,a,S) < \infty,$$

and also boundedness of r and v^0 are weakened. By this set-up it is possible to handle DP's with a bounding function by applying to it the "similarity transformation" introduced by VEINOTT [28] and used by PORTEUS [20] for DP_∞. This is another way to prove the bounds in HINDERER [8].

We like to point out that DP's with bounding functions could also be treated by DP's with state- and action-dependent discountfactor (and bounded reward functions), a model that has been introduced by

SCHÄL [24], and which is in a sense equivalent to the "generalized DP's" cited above.

b) More generally, the results of sections 4 and 5 hold for generalized DP's, if - roughly spoken - β is replaced by $\alpha\beta$.

c) Many refined - though - sometimes more complicated - bounds are derived, and applied for DEA.

d) Error-bounds $\varepsilon(N)$ depending on the horizon N are proposed, and e.g. a forecast is derived for the length of the "tail" in order to get an extrapolation of preassigned quality.

e) Following an idea of SCHELLHAAS [25] for DP_∞, another extrapolation method for DP_N is developed, where the bounds for $v^N - v^k$ are not constants, but multiplies of the functions $v^k - v^{k-1}$ or $v^{k-1} - v^{k-2}$, which may yield - in particular if P is not stochastic - much better bounds.

(v) Corollary 2 above was derived in HÜBNER [12] under the assumption that P is stochastic, yet very recently an extension to generalized DP's was discovered.

(vi) Recently RIEDER [22] showed that the bound obtained for DP's with bounding functions may sometimes be improved by the use of "one-sided" bounding functions.

REFERENCES

[1] BATHER, J. (1973), *Optimal decision procedures for finite Markov chains*, Part II: Communicating systems. Adv. Appl. Prob. 5, 521-540.

[2] DERMAN, C. (1970), *Finite state Markovian decision processes*. Academic Press, New York-London.

[3] DERMAN, C., M. KLEIN (1965), *Some remarks on finite horizon Markovian decision models*, Operations Res. 13, 272-278.

[4] DIRICKX, Y.M.I. (1973), *Turnpike theory in deterministic discrete dynamic programming with discount factor greater than one*, Siam J. Appl. Math. 24, 467-472.

[5] HALMOS, P.R. (1950), *Measure theory*, Van Nostrand, Princeton N.J.

[6] HASTINGS, N.A.J., J.C.M. MELLO (1973), *Tests for suboptimal actions in discounted Markov programming*, Management Sci. 19, 1019-1022.

[7] HINDERER, K. (1975), *Neuere Resultate in der stochastischen dynamischen Optimierung*, Z. Ang. Math. Mech. 55, T16-T26.

[8] HINDERER, K. (1975), *Bounds for stationary finite-stage dynamic programs with unbounded reward functions*, Preprint.

[9] HINDERER, K. (1976), *Estimates for finite-stage dynamic programs*, J. Math. Anal. Appl. 55, 207-238.

[10] HINDERER, K., G. HÜBNER (1974), *An improvement of J.F. Shapiro's turnpike-theorem for the horizon of finite stage, discrete dynamic programs*, to appear in "Trans. 7th Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, Prague 1974."

[11] HÜBNER, G. (1974), *Improved procedures for eliminating suboptimal actions in Markov programming by the use of contraction properties*, to appear in "Trans. 7th Prague Conference on Information Theory, Statistical Decision Functions, and Random Processes, Prague 1974."

[12] HÜBNER, G. (1975), *Extrapolation und Ausschliessung suboptimaler Aktionen in endlich-stufigen stationären Markoffsche Entscheidungsmodellen*. Habilitationsschrift, Univ. Hamburg.

[13] HÜBNER, G. (1976), *On the fixed points of the optimal reward operator in stochastic dynamic programming with discount factor greater than one*. To be published. A short version appeared in Z. Ang. Math. Mech. 56, T348-T349.

[14] LEMBERSKY, M.R. (1974), *Preferred rules in continuous time Markov decision processes*. Management Sci. 21, 348-357.

[15] LIPPMAN, S.A. (1973), *Semi-Markov decision processes with unbounded rewards*. Management Sci. 19, 717-731.

[16] MacQUEEN, J. (1966), *A modified dynamic programming method for Markovian decision problems*, J. Math. Anal. Appl. 14, 38-43.

[17] MacQUEEN, J. (1967), *A test for suboptimal actions in Markovian decision problems*, Operations Res. 15, 559-561.

[18] PHELPS, E. (1962), *The accumulation of risk capital: A sequential utility analysis*. Econometrica 30, 729-743.

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

[20] PORTEUS, E.L. (1975), *Bounds and transformations for discounted finite Markov decision chains*, Operations Res. 23, 761-784.

[21] REETZ, D. (1974), *Eine Klasse von expandierenden Markoffschen Entscheidungsprozessen mit Turnpike-Eigenschaften*, Proceedings in Operations Res. 4, Hrsg. J.-H. Zimmermann et al. 323-332.

[22] RIEDER, U. (1976), *Estimates for dynamic programs with lower and upper bounding functions*. To be published.

[23] SCHAEFER, H.H. (1974), *Banach lattices and positive operators*, Springer, Berlin-Heidelberg-New York.

[24] SCHÄL, M. (1971), *Ein verallgemeinertes stationäres Entscheidungsmodell der dynamischen Optimierung*, Oper. Res. Verfahren X, 145-162,

[25] SCHELLHAAS, H. (1974), *Zur Extrapolation in Markoffschen Entscheidungsmodellen mit Diskontierung*, Z. Operations Res. 18, 91-104.

[26] SENETA, E. (1973), *On the historical development of the theory of finite inhomogeneous Markov chains*, Proc. Cambr. Phil. Soc. 74, 507-513.

[27] SHAPIRO, J.F. (1968), *Turnpike planning horizons for a Markovian decision model*, Management Sci. 14, 292-300.

[28] VEINOTT, A.F. (1969), *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist. 40, 1635-1660.

[29] WESSELS, J. (1975), *Markov programming by successive approximations with respect to weighted supremum norms*. Preprint. To appear in J. Math. Anal. Appl. 58 (1977).

MARTINGALES AND DYNAMIC PROGRAMMING

R.Boel
University of Gent, Gent, Belgium

1. INTRODUCTION

The large difference in models, used in applying dynamic programming to various fields, makes it difficult to compare the results. The purpose of this paper is to give a model, encompassing both the Markov decision model of operations research and the state space descriptions usual in control engineering. An important class of stochastic processes, for this purpose, are the semimartingales, sums of martingales and predictable processes. Another important tool is the transformation of measures, using a Girsanov type translation theorem.

To avoid overburdening the paper with technical details, all proofs have been omitted. Except for some remarks at the end of the paper, only discrete time is treated. Proofs are usually straightforward calculations in this case. For details see [1] or [2].

Section 2 contains the mathematical preliminaries for the measure transformation model. In particular, it is shown that every discrete time stochastic process is a semimartingale. Section 3 introduces the measure transformation model of stochastic optimal control, and its relation to the Markov decision problem and state space descriptions. Optimality conditions are derived in section 4. Several martingale interpretations are discussed. In the final section the advantages of using martingales are

discussed, in particular the possibility of extending the results to the continuous-time case.

The author would like to thank dr. J. van Schuppen and dr. P. Brémaud for helpful discussions, and for making available to him the results of [2], prior to its publication. Sincere thanks are due to Prof. Varaiya for his inspiring help.

2. MATHEMATICAL PRELIMINARIES

Throughout this paper a measure space (Ω,F), a set of time indices $t \in T = N_+$ and an increasing family of σ-algebra's F^t $(\subset F)$ are held fixed. Various probability measures P_u will be defined on (Ω,F). The corresponding expected values will be denoted E_u. Stochastic processes (and random functions) are always at least F^t-adapted (i.e. $X_t(\omega)$ is F^t-measurable for all t). The σ-algebras generated by a stochastic process such as (X_t) will be denoted by superscripts, i.e. $X^t = \sigma(X_s, s \le t) \subset F^t$.

A random function $A_t(\omega)$ is F^t-predictable if $A_t(\omega)$ is F^{t-1}-measurable for all t. A stochastic process $M_t(\omega)$ over (Ω,F,P_u) is an F^t-martingale if

$$(1) \qquad E_u|M_t| < \infty,\ \forall\ t,\ E_u\,(M_s|F^t) = M_t \ \forall\ s \ge t \ge 0$$

(or equivalently $E_u(M_{t+1}-M_t|F^t) = 0, \forall t$). If (1) is replaced by $E_u(M_s|F^t) \ge$ $\ge M_t$, (M_t) is called a submartingale. Reversing the inequality gives a supermartingale. Sub- and supermartingales are special cases of semimartingales. A stochastic process $Z_t(\omega)$ over (Ω,F,P_u) is a semimartingale if it can be written as

$$Z_t = Z_0 + A_t + M_t$$

where M_t is an F^t-martingale and A_t is F^t-predictable, and Z_0 is an F^0-measurable random variable. By the *Doob decomposition theorem* every discrete-time stochastic process (Z_t) over (Ω,F,P_u) can be written in a unique way as a semimartingale [3]:

$$\begin{aligned} Z_t &= Z_{t-1} + E_u(\Delta Z_t|F^{t-1}) + \Delta Z_t - E_u(\Delta Z_t|F^{t-1}) \\ &= Z_{t-1} + \Delta A_t^u + \Delta M_t^u = Z_0 + A_t^u + M_t^u \end{aligned}$$

where

$$(2) \qquad \Delta Z_t = Z_t - Z_{t-1},\ \Delta A^u_t = A^u_t - A^u_{t-1} \triangleq E_u(\Delta Z_t \,|\, F^{t-1})$$

and

$$\Delta M^u_t = M^u_t - M^u_{t-1} = \Delta Z_t - E_u(\Delta Z_t \,|\, F^{t-1})$$

are the increments of the semi-martingale Z_t, of the F^t-predictable process A^u_t and of the (F^t)-martingale M^u_t.

The predictable covariation of two F^t-martingale (M_t) and (N_t) over (Ω,F,P_u) is defined as:

$$(3) \qquad < M,\ N >_t = \sum_{\ell=1}^{t} E_u(M_\ell N_\ell - M_{\ell-1}N_{\ell-1} \,|\, F^{\ell-1})$$

$$= \sum_{\ell=1}^{t} E_u[(M_\ell - M_{\ell-1})\ (N_\ell - N_{\ell-1}) \,|\, F^{\ell-1}].$$

The dependence of the Doob decomposition on the probability measure P_u can be expressed by the *martingale translation theorem* (see VAN SCHUPPEN & WONG [5]):

Given the F^t-martingales (X_t) and (M_t) on (Ω,F,P_0), with $E_0\ \Delta\ M_t^2 < \infty$, $E_0\ \Delta\ X_t^2 < \infty$, $\Delta\ M_t > -1$ a.s. $\forall\ t$, define a probability measure P on (Ω,F) by its restriction P_t to (Ω,F^t) as:

$$(4) \qquad \frac{dP_t}{dP_0} = \Lambda_t = 1 + \sum_{\ell=1}^{t} \Lambda_{\ell-1} \cdot \Delta M_\ell = \prod_{\ell=1}^{t} (1+\Delta M_\ell).$$

Then Λ_t is a positive martingale and

$$Z_t = X_t - < X,\ M >_t \text{ is an } F^t\text{-martingale on } (\Omega,F,P)$$

where

$$\Delta < X,\ M >_t = E_0(X_t M_t - X_{t-1}M_{t-1} \,|\, F^{t-1}).$$

3. STOCHASTIC OPTIMIZATION MODELS

Assume given a random function $X_t(\omega)$ (the state) on (Ω,F), with values in R^n, the family of σ-algebras X^t (the complete history) and an X^t-adapted process Y_t (the observations) generating the family of σ-algebras Y^t (the observed history). Also given is a measurable space $(V,\mathcal{V})$ of controle values $v \in V$.

A control law $u = u(t,\omega) \in V$ is a Y^t-predictable process (i.e. at each time t a control value u_t is chosen based on the information Y^{t-1}). The set of admissible control laws, U, is assumed relatively complete*, i.e. if $u \in U$, $w \in U$, then $(u^n,w) \in U$ where

$$(5) \qquad \begin{cases} (u^n,w)\ (t,\omega) = u(t,\omega) & t \le n \\ \qquad\qquad\ = w(t,\omega) & t > n \end{cases}$$

To each admissible control law u there corresponds a probability measure P_u on (Ω,F), such that P_u restricted to X^t depends only on $u_0,u_1,\ldots,u_t$ while $E_u(Z|X^t)$ for $Z \in X^s$ depends only on $u_{t+1}, u_{t+2},\ldots,u_s$.

The cost to be minimized in this problem is given by:

$$(6) \qquad J(u) = E_u \sum_{t\in T} \alpha^t\, c(t,u_t,X_t)$$

where $c : T \times V \times R^n \longrightarrow [0,K]$ for simplicity (generalizations to integrable X^t-measurable functions, depending on u_t, are possible).

The connection of this model with Markov decision problems is obvious. Assume $Y_t = X_t$, $Y^t = X^t$ and the probability measure P_u is defined by an initial condition $P_{0,u_0}(X_0 \in A)$ and transition probabilities $P_{t,u_t}(X_t \in A|X_{t-1})$. The probability measure P_u is constructed iteratively, starting with

$$P_u(X_2\in A_2, X_1\in A_1, X_0\in A_0) = \int_{A_1}\int_{A_0} P_{2,u_2}(X_2\in A_2|X_1)\cdot P_{1,u_1}(X_1\in dX_1|X_0)\cdot$$

$$\cdot P_{o,u_0}(X_0\in dX_0).$$

---

* This is related to the product property introduced by HORDIJK [4].

Many non-Markovian problems can be defined similarly by defining X^{t-1}-measurable transition probabilities $P_{t,u_t}(A,\omega)$. Randomized controls are automatically included in the above model.

The relation of the abstract model with the state space description, is obtained by applying Doob decomposition:

$$(7) \qquad X_t = X_{t-1} + E_u(\Delta X_t | X^{t-1}) + \Delta X_t - E_u(\Delta X_t | X^{t-1})$$

$$= X_{t-1} + \Delta F_t^u + \Delta M_t^u$$

$$= X_{t-1} + f(t,u_t,X^{t-1}) + m(t,u_t,\Delta X_t,X^{t-1})$$

and similarly for the observation equation

$$(8) \qquad Y_t = Y_{t-1} + g(t,u_t,X^{t-1}) + \tilde{m}(t,u_t,\Delta Y_t,X^{t-1}).$$

A true state description is obtained when $X_t = Y_t$ is a Markov process, which holds if and only if $f(t,u_t,X^{t-1})$ does no longer depend on X^{t-2}, while the distribution of the noise term $m(t,u_t,\Delta X_t)$ depends only on X_{t-1}. Conditions under which one can go from a state model to a measure transformation model, for $X^t = F^t$, (i.e. derive a unique P_u from $f(t,u_t,X^{t-1})$ and the noise distribution) will be investigated in a later paper.

4. OPTIMALITY CONDITIONS

For each $u \in U$ let $u^t \triangleq \{u_0,u_1,\ldots,u_t\}$, then the *value function*

$$(9) \qquad V(t,u^t,Y^t) = \inf E_{u^t,v}\left[\sum_{\ell=t+1}^{\infty} \alpha^\ell c(\ell,v_\ell,X_\ell) \,|\, Y^t\right]$$

is the expected minimal future cost, given control law u has been used up to the present time t, and given the information Y^t.

PRINCIPLE OF OPTIMALITY

For all $0 \le s \le t < \infty$ and for all $u \in U^*$

---

* Note that $P_{u^t} \triangleq P_u$ restricted to X^t.

$$V(s,u^s,Y^s) \leq E_{u^t}\left[\sum_{\ell=s+1}^{t} \alpha^\ell c(\ell,u_\ell,X_\ell) + V(t,u^t,Y^t)\,|Y^s\right] P_{u^t} - a.s. \tag{10}$$

with u being optimal if and only if (10) is an equality.

This statement can be interpreted in martingale terms, by defining the minimal expected total cost, at time t:

$$H(t,u^t,Y^t) = E_{u^t}\left[\sum_{\ell=1}^{t} \alpha^\ell c(\ell,u_\ell,X_\ell) + V(t,u^t,Y^t)\,|Y^t\right].$$

The principle of optimality is equivalent to the statement that $H(t,u^t,Y^t)$ is, for each $u \in U$, a Y^t-submartingale, and u is optimal if and only if $H(t,u^t,Y^t)$ is a Y^t-martingale on (Ω,F,P_u). This is intuitively obvious: the longer we wait before switching to an optimal control, the higher the expected total cost .

Another interpretation uses,

$$R(t,u,Y^t) = E_u\left[\sum_{\ell=t+1}^{\infty} \alpha^\ell c(\ell,u_\ell,X_\ell)\right] - V(t,u^t,Y^t),$$

the excess expected future cost if we do not switch to the optimal control in the future. This is a Y^t-potential (i.e. a positive Y^t-supermartingale tending to 0 a.s. P_u for $t \to \infty$) and u is optimal if and only if $R(t,u,Y^t) = 0$ a.s. P_{u^t}.

Applying the Doob decomposition to either of the processes $V(t,u^t,Y^t)$, $H(t,u^t,Y^t)$ or $R(t,u,Y^t)$ leads to the standard Bellman equation:

$$0 = \inf_{v\in V} E_{u^t,u_{t+1}=v}\left[\alpha^{t+1} c(t+1,v,X_{t+1}) + V(t+1;u^t,v,Y^{t+1}) - V(t,u^t,Y^t)\,|Y^t\right]. \tag{11}$$

The difference between (11) and (10) is that the martingale increment $\Delta V(t,u^t) - E_{u^{t-1}}(\Delta V(t,u^t)/Y^{t-1})$ has disappeared. It is intuitively obvious that only the predictable part of the value function should be taken into account, not the martingale part.

Combination of equations (8), (9) and (11) shows that a value iteration method could be applied to control problems. This, and the even more difficult problem of explicit solutions, will be greatly simplified if the following conditions are met:

(i) there exists a sufficient statistic Z_t of low dimension, recursively defined by $Z_{t+1} = f(t+1, Z_t, Y_{t+1}, u_{t+1})$ such that $V(t,u^t,Y^t) = V(t,u^t,Z_t)$;

(ii) the value function is independent of the past controls, i.e. $V(t,u^t,Z_t) = V(t,Z_t)$. This happens if $X_t = Y_t$ or if u influences only the cost, but not the probability measure.

5. CONCLUSIONS

It has been shown that martingales allow interesting interpretations and generalizations of the optimality conditions. This advantage becomes even more pronounced in the continuous-time case. For example, in the proofs of [6,7] existence of an optimal control is not required, leading to proofs independent of selection theorems, or of unrealistic continuity assumptions. One can also hope that application of the powerful martingale inequalities and martingale convergence theorems [3] will lead to new results in dynamic programming. A first step in this direction is reported in [8]. It is also interesting to note that ROCKAFELLAR and WETS [9] have interpreted the increments of the predictable part and the martingale part of the value function as shadow prices (dual variables) in a convex programming context.

REFERENCES

[1] BOEL, R., *Discrete-time martingales in filtering and stochastic control*, report, NTH Trondheim, Division of engineering cybernetics, 1976.

[2] BREMAUD, P. & J. VAN SCHUPPEN, *Discrete time processes I: martingale calculus and innovations kernels*, preprint, 1976.

[3] NEVEU, J., *Discrete-time martingales*, North-Holland Publishing Cy, Amsterdam, 1975.

[4] HORDIJK, A., *Convergent dynamic programming*, Mathematical Center, Amsterdam, 1975.

[5] SCHUPPEN, J. VAN & E. WONG, *Transformation of local martingales under a change of law*, Annals of Probability, 2, pp. 879-888, 1974.

[6] DAVIS, M. & P. VARAIYA, *Dynamic programming conditions for partially observable stochastic systems*, SIAM J. Control, 11, pp. 226-261, 1973.

[7] BOEL, R. & P. VARAIYA, *Optimal control of jump processes*, to appear in SIAM J. Control and Optimization.

[8] GROENEWEGEN, L. & K. VAN HEE, *Markov decision processes and quasi-martingales*, Memorandum COSOR 76-04, Dept. Math., Eindhoven, University of Technology, 1976. To appear in Proceedings of the 8th EMS (Grenoble 1976).

[9] ROCKAFELLAR, R. & R. WETS, *Nonanticipativity and L^1-martingales in stochastic optimization problems*, in "Stochastic Systems: Modeling, Identification and Optimization", ed. R. Wets, North-Holland Publ. Co., 1976.

SENSITIVE OPTIMALITY IN STATIONARY MARKOVIAN DECISION PROBLEMS ON A GENERAL STATE SPACE

J.Wijngaard
Eindhoven University of Technology, Eindhoven, The Netherlands

1. INTRODUCTION

In considering Markovian decision problems with no discounting the first interest is in general in the average costs. But if there are more average optimal strategies one can distinguish between these by considering the bias, the limit of the difference of the n-period costs and n times the average costs. An average optimal strategy which, among all average optimal strategies, minimizes the bias, is called sensitive optimal. Sensitive optimality is equivalent with 1-optimality (BLACKWELL [2]).

Sensitive optimality and extensions are considered by VEINOTT [10], [11], MILLER & VEINOTT [8] for a finite state space and by HORDIJK & SLADKY [7] for a countable state space.

In this paper we consider the existence of sensitive optimal strategies for problems on a general state space. Compactness of the space of strategies and continuity of the transition probability and the one-period costs on the space of strategies are used to derive sufficient conditions for the existence of sensitive optimal strategies.

2. PRELIMINARIES

Let (V,Σ) be a measurable space. The linear space $B(V,\Sigma)$ is defined as the space of all complex valued bounded measurable functions on V. Let

$\|f\| := \sup_{u\in V} |f(u)|$ for all $f \in B(V,\Sigma)$, then $\|\cdot\|$ is a norm on $B(V,\Sigma)$ and with this norm $B(V,\Sigma)$ is a Banach space.

A Markov process on (V,Σ) with transition probability P defines a bounded linear operator in $B(V,\Sigma)$ by

$$(Pf)(u) = \int_V f(v)P(u,dv), \qquad f \in B(V,\Sigma).$$

The norm of this operator in $B(V,\Sigma)$ is denoted by $\|P\|$ and its spectrum by $\sigma(P)$. Since P is a Markov process, $1 \in \sigma(P)$ and $\sigma(P)$ contains no points outside the unit circle.

For $A \in \Sigma$ the sub-Markov process P_A is defined by

$$P_A(u,E) := P(u,A\cap E), \qquad u \in V,\ F \in \Sigma.$$

Let $A \in \Sigma$, $B = V\backslash A$ and let Q be the embedded sub-Markov process of P on A, then

$$Q(u,E) = \sum_{n=0}^{\infty} (P_B^n P_A 1_E)(u), \qquad u \in V,\ E \in \Sigma.$$

If $\lim_{n\to\infty} (P_B^n 1_V)(u) = o$ for all $u \in V$ then Q is a Markov process.

Let c be a nonnegative measurable function. The pair (P,c) is called a *Markov process with costs*. If P is *quasi-compact* (satisfies the *Doeblin condition*) and c is bounded, the *average costs* $g := \lim_{n\to\infty} \frac{1}{n} \sum_{\ell=0}^{n-1} P^\ell c$ exist and the functions $w_m := \lim_{k\to\infty} \sum_{\ell=0}^{kd+m} P^\ell(c-g)$ exist for all $m = 0,1,2,\ldots$ and for d equal to the period of P.

Let $v := \frac{1}{d} \sum_{m=0}^{d-1} w_m$, then v is a solution of $y = c-g + Py$ and if P has only one ergodic set this solution is unique upto a constant. The function v is called the *bias* of (P,c).

A *stationary Markovian decision problem* (SMD) is a set of Markov processes with costs $\{(P_\alpha,c_\alpha)\}$, $\alpha \in A$. The elements $\alpha \in A$ are called *strategies*. It is clear that if in a Markovian decision process only stationary policies are allowed, it can be interpreted as an SMD. An important property of an SMD is the *product property*. An SMD satisfies the product property if for each $\alpha_1,\alpha_2 \in A$ and for each $F \in \Sigma$ there exists an $\alpha \in A$ such that

$$P_\alpha(u,E) = P_{\alpha_1}(u,E) \quad \text{and} \quad c_\alpha(u) = c_{\alpha_1}(u) \quad \text{for } u \in F,$$
$$P_\alpha(u,E) = P_{\alpha_2}(u,E) \quad \text{and} \quad c_\alpha(u) = c_{\alpha_2}(u) \quad \text{for } u \in V\backslash F.$$

This product property is always satisfied in Markovian decision processes, the actions in the different states may be chosen independently of each other.

If the product property holds it is possible to prove that for two arbitrary strategies, $\alpha_1,\alpha_2 \in A$ there exists a third strategy $\alpha \in A$ which is better than both. This is worked out in the next lemma.

LEMMA 1. *Let* $\{(P_\alpha,c_\alpha)\}$, $\alpha \in A$ *be an SMD with* P_α *quasi-compact and* c_α *bounded on* V, *uniform in* α. *Assume that the product property is satisfied. Let* $\alpha_1,\alpha_2 \in A$ *and* $g_{\alpha_1},g_{\alpha_2}$ *and* $v_{\alpha_1},v_{\alpha_2}$ *the corresponding average costs and bias. Then*

(i) *there exists a strategy* $\alpha_0 \in A$ *such that*

$$g_{\alpha_0}(u) \leq \min\{g_{\alpha_1}(u),g_{\alpha_2}(u)\} \quad \text{for all } u \in V;$$

(ii) *if* α_1,α_2 *are both average optimal then there exists a strategy* $\alpha_0 \in A$ *such that*

$$v_{\alpha_0}(u) \leq \min\{v_{\alpha_1}(u),v_{\alpha_2}(u)\} \quad \text{for all } u \in V.$$

PROOF. For the proof of the first part we refer to [12], section 4.1.3. Now let α_1,α_2 be two average optimal strategies, $g_{\alpha_1} = g_{\alpha_2} = g$. Let $F := \{u \mid v_{\alpha_1}(u) < v_{\alpha_2}(u)\}$ and $G := V\backslash F$. Let Q_{α_2} be the embedded sub-Markov process of P_{α_2} on F and Q_{α_1} the embedded sub-Markov process of P_{α_1} on G. The strategy α_0 is chosen such that

$$P_{\alpha_0}(u,E) = P_{\alpha_1}(u,E), \quad c_{\alpha_0}(u) = c_{\alpha_1}(u) \quad \text{for } u \in F,$$

$$P_{\alpha_0}(u,E) = P_{\alpha_2}(u,E), \quad c_{\alpha_0}(u) = c_{\alpha_2}(u) \quad \text{for } u \in G.$$

The product property implies that there is such a strategy α_0 in *A*. Let R_{α_0} be the entry process of P_{α_0} on F, that means that R_{α_0} is the sub-Markov process which describes the state of the system each time the set F is entered,

$$R_{\alpha_0}(u,E) = Q_{\alpha_2}(u,E), \quad u \in G$$

$$R_{\alpha_0}(u,E) = (Q_{\alpha_1}Q_{\alpha_2})(u,E), \quad u \in F.$$

Define $v_{\alpha_1 n \alpha_2}$ as the bias of the (non-stationary) strategy which applies α_0 until the set F is entered for the n^{th} time and from then on the strategy α_1.

First consider the case that α_0 has only one invariant probability π_{α_0}. If $\pi_{\alpha_0}(F) > 0$ and $\pi_{\alpha_0}(G) > 0$ then Q_{α_2} and Q_{α_1} are Markov processes and

$$v_{\alpha_1 1 \alpha_2}(u) = \sum_{n=0}^{\infty} P_{\alpha_2 G}^n (c_{\alpha_2} - g)(u) + (Q_{\alpha_2} v_{\alpha_1})(u), \qquad u \in G,$$

$$v_{\alpha_1 1 \alpha_2}(u) = \sum_{n=0}^{\infty} P_{\alpha_1 F}^n (c_{\alpha_1} - g)(u) + (Q_{\alpha_1} v_{\alpha_1 1 \alpha_2})(u), \quad u \in G,$$

and for n = 2,3,4,...

$$v_{\alpha_1 n \alpha_2}(u) = \sum_{n=0}^{\infty} P_{\alpha_2 G}^n (c_{\alpha_2} - g) + (Q_{\alpha_2} v_{\alpha_1 n-1 \alpha_2})(u), \quad u \in G,$$

$$v_{\alpha_1 n \alpha_2}(u) = \sum_{n=0}^{\infty} P_{\alpha_1 F}^n (c_{\alpha_1} - g)(u) + (Q_{\alpha_1} v_{\alpha_1 n \alpha_2})(u), \quad u \in F.$$

If $\pi_{\alpha_0}(F) = 0$ the sum $\sum_{n=0}^{\infty} P_{\alpha_2 G}^n (c_{\alpha_2} - g)(u)$ in these expressions has to be replaced by $\sum_{n=0}^{\infty} P_{\alpha_2 G'}^n (c_{\alpha_2} - g)(u) + Q'_E v_{\alpha_2}$, where $E \subset G$ is a maximal invariant set of P_{α_2}, $G' := G \backslash E$ and Q' is the embedded Markov process of P_{α_2} on $F \cup E$. Notice that $Q'_F = Q_{\alpha_2}$. If $\pi_{\alpha_0}(G) = 0$ the sum $\sum_{n=0}^{\infty} P_{\alpha_1 F}^n (c_{\alpha_1} - g)(u)$ has to be replaced in the same way. But in each of these cases ($\pi_{\alpha_0}(F) > 0$, $\pi_{\alpha_0}(G) > 0$; $\pi_{\alpha_0}(F) = 0$, $\pi_{\alpha_0}(G) = 1$; $\pi_{\alpha_0}(F) = 1$, $\pi_{\alpha_0}(G) = 0$) it is easy to verify that

$$(*) \qquad \min\{v_{\alpha_1}(u), v_{\alpha_2}(u)\} - v_{\alpha_1 n \alpha_2}(u) \geq \sum_{\ell=1}^{n} R_{\alpha_0}^{\ell} (v_{\alpha_2} - v_{\alpha_1})(u), \quad u \in V.$$

Let g_{α_0} be the average costs of the strategy α_0. Using $v_{\alpha_0} = c_{\alpha_0} - g_{\alpha_0} + P_{\alpha_0} v_{\alpha_0}$ we get, for the case that $\pi_{\alpha_0}(F) > 0$, $\pi_{\alpha_0}(G) > 0$,

$$v_{\alpha_0}(u) = \sum_{n=0}^{\infty} P_{\alpha_2 G}^n (c_{\alpha_2} - g_{\alpha_0})(u) + (Q_{\alpha_2} v_{\alpha_0})(u), \qquad u \in G,$$

$$v_{\alpha_0}(u) = \sum_{n=0}^{\infty} P_{\alpha_1 F}^n (c_{\alpha_1} - g_{\alpha_0})(u) + (Q_{\alpha_1} v_{\alpha_0})(u), \qquad u \in F.$$

If $g_{\alpha_0} = g$ then $v_{\alpha_1 n \alpha_2} = v_{\alpha_0} + R_{\alpha_0}^n (v_{\alpha_1} - v_{\alpha_0})$ and if $g_{\alpha_0} > g$ then $v_{\alpha_1 n \alpha_2} \to +\infty$ for $n \to \infty$, but this is impossible by (*) since $\sum_{\ell=1}^{n} R_{\alpha_0} (v_{\alpha_2} - v_{\alpha_1}) \geq 0$. Hence $g_{\alpha_0} = g$ and $v_{\alpha_1 n \alpha_2} = v_{\alpha 0} + R_{\alpha 0}^n (v_{\alpha 1} - v_{\alpha 0})$. This

holds also for the cases $\pi_{\alpha_0}(F) = 1$, $\pi_{\alpha_0}(G) = 0$ and $\pi_{\alpha_0}(F) = 0$, $\pi_{\alpha_0}(G) = 1$. Therefore

$$\min\{v_{\alpha_1}(u), v_{\alpha_2}(u)\} - v_{\alpha_0}(u) - R_{\alpha_0}^n(v_{\alpha_1} - v_{\alpha_0}) \geq$$
$$\geq \sum_{\ell=1}^{n} R_{\alpha_0}^{\ell}(v_{\alpha_2} - v_{\alpha_1})(u).$$

The boundedness of the sequence $R_{\alpha_0}^n(v_{\alpha_1}-v_{\alpha_0})(u)$ in n implies the convergence of the sum $\sum_1^\infty R_{\alpha_0}^{\ell}(v_{\alpha 2}-v_{\alpha 1})(u)$. But since $v_{\alpha 2} > v_{\alpha 1}$ everywhere on F this implies that the entry process R_{α_0} is absorbing, that means $\pi_{\alpha_0}(F) = 0$ or $\pi_{\alpha_0}(G) = 0$.

Hence $R_{\alpha_0}^n(v_{\alpha_1}-v_{\alpha_0})(u) \to 0$ and

$$v_{\alpha_0}(u) \leq \min\{v_{\alpha_1}(u), v_{\alpha_2}(u)\} - \sum_{\ell=1}^{\infty} R_{\alpha_0}^{\ell}(v_{\alpha_2} - v_{\alpha_1})(u).$$

This completes the proof of (ii) for the case that P_{α_0} has only one ergodic set. If P_{α_0} has more disjoint ergodic sets the proof can be given in the same way by considering the process on each of these sets. □

3. EXISTENCE OF AVERAGE OPTIMAL AND SENSITIVE OPTIMAL STRATEGIES

In this section an SMD $\{(P_\alpha, c_\alpha)\}$, $\alpha \in A$ is considered such that

(i) P_α is quasi-compact for all $\alpha \in A$;

(ii) c_α is bounded on V, uniform in α;

(iii) A is a metric space, metric ρ, such that

$$\lim_{\rho(\alpha,\alpha_0) \to 0} \|P_\alpha - P_{\alpha_0}\| \to 0 \quad \text{for all } \alpha_0 \in A,$$
$$\lim_{\rho(\alpha,\alpha_0) \to 0} \|c_\alpha - c_{\alpha_0}\| \to 0 \quad \text{for all } \alpha_0 \in A.$$

Let g_α, v_α be the average costs and the bias of (P_α, c_α). The strategy $\alpha_0 \in A$ is called *sensitive optimal* if α_0 is average optimal and if $v_{\alpha_0}(u) \leq v_\alpha(u)$ for all $u \in V$ and all average optimal strategies α. We will derive conditions for the existence of sensitive optimal strategies using the compactness of A and the continuity of P_α and c_α. Define A_n, n = 1,2,... as the set of all $\alpha \in A$ such that P_α has n disjoint ergodic

sets.

In the following lemma the continuity of g_α and v_α on A_n is stated. The proof is analogous to the proof of lemma 1.15 in [12] and uses operator valued functions and perturbation theory of linear operators (see DUNFORD-SCHWARTZ [3], VII).

LEMMA 2. *Let* $\{\alpha_i\}$ *be a sequence in* A_n *converging to* $\alpha_0 \in A_n$. *Then*

$$\lim_{i\to\infty} \|g_{\alpha_0} - g_{\alpha_i}\| = 0 \quad \textit{and} \quad \lim_{i\to\infty} \|v_{\alpha_0} - v_{\alpha_i}\| = 0.$$

The following example shows that the continuity of v_α does not hold on the whole space A.

EXAMPLE. Let $\{(P_\alpha, c_\alpha)\}$, $\alpha \in A$ be a problem with two states given by

$$P_\alpha = \begin{pmatrix} 1-\alpha & \alpha \\ 0 & 1 \end{pmatrix}, \quad c_\alpha = \begin{pmatrix} -\sqrt{\alpha} \\ 0 \end{pmatrix}, \quad A = \{\alpha \mid 0 \le \alpha \le \tfrac{1}{2}\}.$$

Then

$$g_\alpha = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad \text{for all } \alpha \in [0, \tfrac{1}{2}],$$

$$v_\alpha = \begin{pmatrix} -\sqrt{\alpha}/\alpha \\ 0 \end{pmatrix} \qquad \text{for all } \alpha > 0,$$

and

$$v_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Hence $v_\alpha(1)$ has a discontinuity in $\alpha = 0$. This discontinuity is due to the fact that for $\alpha > 0$ there is only one ergodic set and for $\alpha = 0$ two.

If in general $\{\alpha_\ell\}$ is a sequence in A_1 converging to $\alpha_0 \in A_n$ then in each neighbourhood of 1 (in the complex plane) there are eigenvalues of P_{α_ℓ} for ℓ large enough. Assume that the spectrum of the operators P_{α_ℓ} is of the following structure, $\sigma(P_{\alpha_\ell}) = 1 \cup \{\lambda_\ell\} \cup \sigma_\ell$ where $\lambda_\ell \to 1$ for $\ell \to \infty$ and σ_ℓ is for all ℓ a set within a circle with radius $\rho < 1$ (ρ independent of ℓ).

Let g_{λ_ℓ} be the projection of $c_{\alpha_\ell} - g_{\alpha_\ell}$ on $N((\lambda_\ell - P_{\alpha_\ell})^{\nu_\ell})$, where ν_ℓ is the index of λ_ℓ as eigenvalue of P_{α_ℓ}. Then

$$\lim_{\ell\to\infty}(v_{\alpha_\ell} - (1/(1-\lambda_\ell))\, g_{\alpha_\ell}) = v_{\alpha_0}$$

and

$$\lim_{\ell\to\infty}(g_{\lambda_\ell} + g_{\alpha_\ell}) = g_{\alpha_0}.$$

In the example $g_{\lambda_\ell} = -\sqrt{\alpha}_\ell$, $\lambda_\ell = 1 - \alpha_\ell$.

REMARK. The average costs g_α have as function of α the same sort of discontinuities, but it is possible to define a rather general class of problems (communicating systems) where the set of all strategies A is dominated by the set of all strategies with a unique invariant probability. The communicativeness is introduced by BATHER [1] for a finite state space and used by HORDIJK [5] for a countable state space and WIJNGAARD [12] for a general state space.

To investigate the existence of sensitive optimal strategies we have to consider the existence of average optimal strategies first. This is done in the next theorem.

THEOREM 3. *Let A be compact, A_n closed in A for all* $n = 1,2,3,\ldots$ *and the number of ergodic sets of* P_α bounded in α. *Assume that the product property is satisfied. Then an average optimal strategy exists.*

PROOF. From lemma 2 and the assumption it follows immediately that for each $u \in V$ there is a strategy $\alpha_u \in A$ such that $g_{\alpha_u}(u) \le g_\alpha(u)$ for all $\alpha \in A$ (the strategy α_u is u-optimal). Since A is a compact metric space it is separable. Let $\{\alpha_n\}_1^\infty$ be a countable subset of A which is dense in A. Then $\inf_n g_{\alpha_n}(u) = g_{\alpha_u}(u)$ for all $u \in V$. Let the strategies γ_n, $n = 1,2,\ldots$ be such that $g_{\gamma_1} = g_{\alpha_1}$ and $g_{\gamma_n} \le \min\{g_{\gamma_{n-1}}, g_{\alpha_n}\}$ for all $n = 2,3,4,\ldots$. The existence of such strategies g_{γ_n} is guaranteed by lemma 1. The sequence $g_{\gamma_n}(u)$ is then monotonically non-increasing for each $u \in V$ and $g_{\gamma_n}(u) \le g_{\alpha_n}(u)$. Hence $\lim_{n\to\infty} g_{\gamma_n}(u) = g_{\alpha_u}(u)$, $u \in V$. The boundedness of the number of ergodic sets, the compactness of A and the closedness of A_n for each n implies the existence of an integer ℓ and a subsequence $\{\gamma_n\}$ in A_ℓ converging to some γ in A_ℓ. This strategy γ is average optimal. □

A condition for closedness of A_n for all $n = 1,2,3,\ldots$ is given in the next lemma. For the proof we refer to [12].

LEMMA 4. *If there is a* ρ, $0 < \rho < 1$ *such that for all* $\alpha \in A$ *the spectrum of*

P_α *has no points* λ *with* $\rho < |\lambda| < 1$, *then* A_n *is closed in* A *for all* $n = 1,2,3,\ldots$.

If the conditions of theorem 3 are satisfied the existence of a sensitive optimal strategy can be proved in the same way as the existence of an average optimal strategy. The continuity of g_α in α implies the closedness and hence compactness of the set of all average optimal strategies. We have the following result.

THEOREM 5. *If the conditions of theorem* 3 *are satisfied, a sensitive optimal strategy exists.*

If α_0 is a sensitive optimal strategy, it is easy to prove that

$$v_{\alpha_0} = \min_{\alpha \in A'} \{c_\alpha - g + P_\alpha v_{\alpha_0}\},$$

where A' is the set of all α such that $P_\alpha g = g$. But even in the finite state space the converse is not true (see BLACKWELL [2]). That means that the sensitive optimal strategy cannot be approximated in general by policy improvement. If successive approximations can be applied depends on the question of $V_n - ng$ converges to v_{α_0} (V_n are the minimal expected n-period costs). For a treatment of this problem, see for instance HORDIJK, SCHWEITZER & TIJMS [6], TIJMS [9] and FEDERGRUEN & SCHWEITZER [4].

REFERENCES

[1] BATHER, J. (1973): *Optimal decision procedures for finite Markov chains, part II: Communicating systems,* Adv. in Appl. Prob. 5, 521-540.

[2] BLACKWELL, D. (1962): *Discrete dynamic programming,* Ann. Math. Statist. 33, 719-729.

[3] DUNFORD, N. & J.T. SCHWARTZ (1958): *Linear Operators, Part I,* Interscience publishers, New York.

[4] FEDERGRUEN, A. & P.J. SCHWEITZER (1976): *Asymptotic behaviour of undiscounted value iteration in Markov decision problems,* Report BW 44/76, Math. Centre, Amsterdam.

[5] HORDIJK, A. (1974): *Dynamic programming and Markov potential theory,* Math. Centre Tracts, no. 51, Amsterdam.

[6] HORDIJK, A., SCHWEITZER, P.J. & H. TIJMS (1975): *The asymptotic behaviour of the minimal total expected costs for the denumerable state Markovian decision model,* Jnl. Appl. Prob. 12, 298-305.

[7] HORDIJK, A. & K. SLADKY (1975): *Sensitive optimality criteria in countable state dynamic programming,* Report BW 48/75, Math. Centre, Amsterdam.

[8] MILLER, B.L. & A.F. VEINOTT Jr. (1969): *Discrete dynamic programming with a small interest rate,* Ann. Math. Statist. 40, 366-370.

[9] TIJMS, H. (1975): *On dynamic programming with arbitrary state space, compact action space and the average return as criterion,* Report BW 55/75, Math. Centre, Amsterdam.

[10] VEINOTT, A.F. Jr. (1966): *On finding optimal policies in discrete dynamic programming with* no *discounting,* Ann. Math. Statist. 37, 1284-1294.

[11] VEINOTT, A.F. Jr. (1969): *Discrete dynamic programming with sensitive discount optimality criteria,* Ann. Math. Statist. 40, 1635-1660.

[12] WIJNGAARD, J. (1975): *Stationary Markovian decision problems, discrete time, general state space,* Dissertation, Eindhoven University of Technology.

AVERAGE PAYOFF CRITERIA FOR ρ-RECURRENT MARKOV DECISION PROCESSES

D.Reetz
Free University Berlin, Berlin, West Germany

1. INTRODUCTION

In the following investigation we consider an infinite stage Markovian Decision Process (MDP) with finite state space $S = \{1,\ldots,i,j,\ell,m,\ldots,M\}$ and finite feasible decision spaces K_i. A decision $k \in K_i$ in state i at stage $t = 0,1,2,\ldots$ determines a payoff $r_i(k)$ per time unit to be received over an interval of $\rho^t > 0$ time units. At the end of this interval a transition to a new state j occurs with probability $p_{ij}(k) \geq 0$, where

$$(1) \qquad \sum_{j\in S} p_{ij}(k) \leq 1 \ . \qquad (i\epsilon S, k\epsilon K_i)$$

A vector $(k_1,\ldots,k_i,\ldots,k_M)$ of decisions $k_i \in K_i$ defines a decision function f. A sequence of decision functions $\{f_0,f_1,\ldots,f_t,\ldots\}$ is called a policy x. If $f_t = f$ for all $t = 0,1,2,\ldots$ the policy $x = f^\infty$ is *stationary*. The set of all decision functions is denoted by F, and that of all policies by X.

Setting

$$(2) \qquad p_{i\ell}(k) = 0 \qquad (i\epsilon S, k\epsilon K_i)$$

for some state ℓ in the above MDP determines an ℓ-*punctuated* MDP. Multistage transition probabilities under a policy x are given by the taboo probabilities ${}_\ell p_{ij}^{(t)}(x)$. Letting S_ℓ denote the set $S - \{\ell\}$ we may obtain the *mean first passage times* from i to ℓ by the possibly divergent series

$$\mu_{i\ell}(x) := \sum_{t=0}^{\infty} \sum_{j\in S_\ell} \rho^t {}_\ell p_{ij}^{(t)}(x) > 0 \quad . \qquad (i\in S) \tag{3}$$

If the *maximal first passage times*

$$\mu_{i\ell} := \sup_{x\in X} \mu_{i\ell}(x) \qquad (i\in S) \tag{4}$$

are finite the state ℓ is called *positively* ρ-*recurrent*; otherwise ℓ is *null* ρ-*recurrent*, cf. FERSCHL [2] and HORDIJK [4] for related results. For a positively ρ-recurrent state ℓ the maximal first passage times to ℓ are the unique solution of

$$\mu_{i\ell} = \max_{k\in K_i} \left\{ 1 + \rho \sum_{j\in S_\ell} p_{ij}(k)\ \mu_{j\ell} \right\} \quad . \qquad (i\in S) \tag{5}$$

Furthermore it is possible to define *present payoff values* for the ℓ-punc tuated MDP under an arbitrary policy x by the absolutely convergent series

$$\mu_{i\ell}(x) := \sum_{t=0}^{\infty} \sum_{j\in S_\ell} \rho^t {}_\ell p_{ij}^{(t)}(x) r_j^{(t)}(x) \quad . \qquad (i\in S) \tag{6}$$

The expression $r_j^{(t)}(x)$ represents the payoff per unit time in state j and stage t under policy x.

For each positively ρ-recurrent state ℓ and policy x an average payoff criterion, the ℓ-*annuity of* x may be defined by

$$a_\ell(x) := \frac{u_{\ell\ell}(x)}{u_{\ell\ell}(x)} \quad . \tag{7}$$

A policy x_ℓ^* is called ℓ-*optimal* if

$$a_\ell := \sup_{x\in X} a_\ell(x) = a_\ell(x_\ell^*) \, . \tag{8}$$

We classify a MDP as *positively* ρ-*recurrent* if all of its states are positively ρ-recurrent. For such a MDP a policy x^* is *optimal* if $x^* = x_\ell$ for all $\ell \in S$.

RESULTS

Using a corresponding theorem of transient MDP (BLACKWELL [1]) we prove the existence of an ℓ-optimal policy which is stationary under the assumption that ℓ is positively ρ-recurrent. For an arbitrary real parameter a define

the parametric *present payoff values* of policy x by the absolutely convergent series

$$(9) \qquad v_{i\ell}(a,x) := \sum_{t=0}^{\infty} \sum_{j\in S_\ell} \rho^t {}_\ell p_{ij}^{(t)}(x)[r_j^{(t)}(x)-a] \,. \qquad (i\in S)$$

Obviously, $v_{\ell\ell}(a,x)$ is a continuous, monotonically decreasing function in a with

$$(10) \qquad \lim_{a\to\pm\infty} v_{\ell\ell}(a,x) = \bar{+}\,\infty \,.$$

Hence there exists a unique zero $a_\ell^*(x)$ of $v_{\ell\ell}(a,x)$:

$$(11) \qquad v_{\ell\ell}(a_\ell^*(x),x) = 0.$$

Using (9), the equation (11) can be solved for $a_\ell^*(x)$ yielding the ℓ-annuity $a_\ell(x)$ of (7).

The positive ρ-recurrence of ℓ implies the transience of the ℓ-punctuated MDP. Thus for each parameter a there exists a stationary policy $f_\ell^\infty(a)$ yielding maximal *parametric present payoff values*

$$(12) \qquad v_{i\ell}(a) = v_{i\ell}(a,f^\infty(a)) \geq v_{i\ell}(a,x) \qquad (x\in X, i\in S)$$

Such a policy $f_\ell^\infty(a)$ along with the associated present values may be determinded by solving the system

$$(13) \qquad v_{i\ell}(a) = \max_{k\in K_i} \left\{[r_i(k)-a] + \rho \sum_{j\in S_\ell} p_{ij}(k) v_{j\ell}(a)\right\} \,. \qquad (i\in S)$$

It can be shown that $v_{\ell\ell}(a)$ is a continuous, monotonically decreasing function in a with

$$(14) \qquad \lim_{a\to\pm\infty} v_{\ell\ell}(a) = \bar{+}\,\infty.$$

Consequently we may infer the existence of a unique zero a_ℓ^* of $v_{\ell\ell}(a)$. Substitution of a_ℓ^* in (13) yields a decision function $f_\ell = f_\ell(a_\ell^*)$. Using a previous result we see that the ℓ-annuity of f_ℓ^∞ coincides with the above zero:

$$(15) \qquad a_\ell(f_\ell^\infty) = a_\ell^* \,.$$

Let $x = \{f_0, y\}$ with $y = \{f_1, f_2, f_3, \ldots\}$ be an arbitrary policy. Using (13), (12), (9) we may write

$$(16) \qquad v_{\ell\ell}(a) = \max_{k \in K_\ell} \left\{ [r_\ell(k) - a] + \rho \sum_{j \in S_\ell} p_{\ell j}(k) v_{j\ell}(a) \right\}$$

$$\geq \max_{k \in K_\ell} \left\{ [r_\ell(k) - a] + \rho \sum_{j \in S_\ell} p_{\ell j}(k) v_{j\ell}(a, y) \right\}$$

$$\geq [r_\ell(f_0) - a] + \rho \sum_{j \in S_\ell} p_{\ell j}(f_0) v_{j\ell}(a, y)$$

$$= v_{\ell\ell}(a, x).$$

As an immediate consequence of (16) we obtain

$$(17) \qquad a_\ell(x) \leq a_\ell(f_\ell^\infty) = a_\ell . \qquad (x \in X)$$

The above results are summarized in the following theorem.

THEOREM 1. *If ℓ is positively ρ-recurrent, then there exists an ℓ-optimal policy f_ℓ^∞ which is stationary. The maximal ℓ-annuity a_ℓ can be obtained by solving the system*

$$(18) \qquad v_{i\ell} + a_\ell = \max_{k \in K_i} \left\{ r_i(k) + \rho \sum_{j \in S} p_{ij}(k) v_{j\ell} \right\} \qquad (i \in S)$$

$$(19) \qquad v_{\ell\ell} = 0.$$

The maximization procedure in (18) *yields an ℓ-optimal policy f_ℓ^∞.*

As can be observed, the positive ρ-recurrence of ℓ is a fairly general condition guaranteeing the existence and uniqueness of the solution to the functional equation (18) - (19). Furthermore under this condition positive payoffs r are sufficient for a positive maximal ℓ-annuity a_ℓ.

We compare an ℓ-punctuated MDP with an m-punctuated MDP, both of which are assumed to be ρ-recurrent with row sums (1) identically equal to one. Under the last assumption the solution set of (18) consists of all expressions $(v+c, a-(1-\rho)c, f)$ where c is an arbitrary constant parameter. This result may be used in proving Theorem 2.

THEOREM 2. *If ℓ and m are positively* ρ*-recurrent in a* MDP *with row sums* (1) *identically equal to one, then*

$$v_{im} = v_{i\ell} - v_{m\ell} \tag{20}$$

$$a_m = a_\ell + (1-\rho)v_{m\ell} \tag{21}$$

$$f_m = f_\ell \,. \tag{22}$$

COROLLARY 1. *In a positively* ρ*-recurrent* MDP *with row sums* (1) *identically equal to one, there exists an optimal policy which is stationary.*

A single-chain ergodic MDP is obviously positively 1-recurrent. Hence an optimal stationary policy may be obtained by solving (18) (19) with $\rho = 1$, cf. HOWARD [5].

In the following we show that three different successive approximation schemes may be applied to obtain the maximal ℓ-annuity and associated ℓ-optimal policy. First we investigate the policy iteration procedure of Howard.

Let a be an arbitrary real member which serves as a parameter in the policy iteration procedure:

Value-Determination. Given $f_\ell^{(n-1)}$ calculate $v_{i\ell}(a,f_\ell^{(n-1)})$ according to

$$v_{i\ell}(a,f_\ell^{(n-1)}) = [r_i(f_\ell^{(n-1)})-a] + \rho \sum_{j\in S_\ell} p_{ij}(f_\ell^{(n-1)})v_{j\ell}(a,f_\ell^{(n-1)}) \,. \qquad (i\in S) \tag{23}$$

Policy-Improvement. Using the values $v_{i\ell}(a,f_\ell^{(n-1)})$ obtained in (23) determine $f_\ell^{(n)}$ according to

$$[r_i(f_\ell^{(n)})-a] + \rho \sum_{j\in S_\ell} p_{ij}(f_\ell^{(n)})v_{j\ell}(a,f_\ell^{(n-1)}) = \max_{k\in K_i} \left\{[r_i(k)-a] + \rho \sum_{j\in S_\ell} p_{ij}(k)v_{j\ell}(a,f_\ell^{(n-1)})\right\} \,. \qquad (i\in S) \tag{24}$$

A result on p.87 in HOWARD [5] yields

$$v_{\ell\ell}(a,f_\ell^{(n-1)}) \le v_{\ell\ell}(a,f_\ell^{(n)}) \,. \tag{25}$$

Letting $a_\ell^{(n)}$ be the unique zero of $v_{\ell\ell}(a,f^{(n)})$, we see that inequality (25) implies

$$a_\ell^{(n-1)} \leq a_\ell^{(n)}. \tag{26}$$

Therefore the bounded sequence $\{a_\ell^{(n)}\}$ converges to a limit, which is reached after some $N < \infty$ iterations in case of finite state and decision spaces: $a_\ell^{(n-1)} = a_\ell^{(n)} = a_\ell (n \geq N)$. Observing that the policy improvement routine is independent of a we may set $a = 0$ in (24). Furthermore if we select $a = a_\ell^{(n-1)}$ in the value determination operation (23) by setting $v_{\ell\ell}(a_\ell^{(n-1)}, f_\ell^{(n-1)}) = 0$ then Howard's policy iteration procedure is obtained, cf. p.38 in HOWARD [5].

THEOREM 3. *If ℓ is positively ρ-recurrent, then the system* (18) - (19) *may be solved using Howard/s policy iteration procedure.*

A second successive approximation method can be obtained by applying White's algorithm to a transformed system (18) - (19) with invariant a_ℓ and f_ℓ. Let

$$\tilde{r}_i(k) := \frac{r_i(k)}{\mu_{i\ell}} \qquad (i \in S, k \in K_i) \tag{27}$$

$$\tilde{p}_{ij}(k) := \frac{p_{ij}(k)\mu_{j\ell}}{\mu_{i\ell}} \qquad (i \in S, j \in S_\ell, k \in K_i) \tag{28}$$

$$\tilde{p}_{i\ell}(k) := \alpha - \alpha_i(k) \geq 0 \qquad (i \in S, k \in K_i) \tag{29}$$

where

$$\alpha_i(k) := \sum_{j \in S_\ell} \tilde{p}_{ij}(k) \qquad (i \in S, k \in K_i) \tag{30}$$

$$\alpha := \max_{k \in K_i} \alpha_i(k). \tag{31}$$

A positively ρ-recurrent ℓ remains positively ρ-recurrent under the above similarity transformation. Hence we may define mean first passage times μ and present payoff values $\tilde{u}$ for the transformed MDP. The annuities of the transformed MDP coincide with those of the original MDP:

$$(32)\qquad \tilde{a}_\ell(x) = \frac{\tilde{u}_{\ell\ell}(x)}{\tilde{u}_{\ell\ell}(x)} = \frac{u_{\ell\ell}(x)/\mu_{\ell\ell}}{u_{\ell\ell}(x)/\mu_{\ell\ell}} = \frac{u_{\ell\ell}(x)}{u_{\ell\ell}(x)} = a_\ell(x) \quad . \qquad (x\in X)$$

Hence the maximal ℓ-annuity a_ℓ and an ℓ-optimal policy f_ℓ^∞ may be obtained by solving (18) - (19) with (r,p) replaced by $(\tilde{r},\tilde{p})$.

Let Θ: $v \to \Theta v$ be an operator defined by

$$(33)\qquad (\Theta v)_{i\ell} = \max_{k\in K_i}\left\{\tilde{r}_i(k) + \rho \sum_{j\in S} \tilde{p}_{ij}(k) v_{j\ell}\right\} \qquad (i\in S)$$

$$- \max_{k\in K_\ell}\left\{\tilde{r}_\ell(k) + \rho \sum_{j\in S} \tilde{p}_{\ell j}(k) v_{j\ell}\right\}.$$

Using the seminorm $sp(v) := \max(v) - \min(v)$ it can be shown that

$$(34)\qquad sp(\Theta v - \Theta w) \le \gamma_H sp(v-w) \qquad (v,w\in \mathbb{R}^M)$$

with

$$(35)\qquad \gamma_H = \rho\Big[\alpha - \min_{\substack{i,i'\\k,k'}} \sum_{j\in S}\{\min\{\tilde{p}_{ij}(k), \tilde{p}_{i'j}(k')\}\Big] ,$$

cf. WHITE [10] and HÜBNER [6]. Dividing the system of equations (5) by $\mu_{i\ell} > 0$ one obtains $\rho\alpha < 1$, so that $\gamma_H < 1$. Therefore, since $(\Theta v)_{\ell\ell} = 0$ for all $v \in \mathbb{R}^M$, any vector sequence generated by Θ converges to the solution of the transformed system.

THEOREM 4. *If ℓ is positively ρ-recurrent, then the maximal ℓ-annuity and an ℓ-optimal policy can be obtained by applying White/s algorithm to the transformed system* (18) - (19) *where* (r,p) *is to be replaced by* $(\tilde{r},\tilde{p})$.

As a third alternative a modified value iteration procedure may be used if certain criteria are satisfied. In analogy to Theorem 2 we obtain the following Theorem 5.

THEOREM 5. *Let ℓ be positively ρ-recurrent in a* MDP *with row sums* (1) *identically equal to* 1. *If there exists a vector* b *guaranteeing the convergence of the value iteration algorithm* $w^{(n)} = \phi_b w^{(n-1)}$:

$$(36)\qquad w_i^{(n)} = \max_{k\in K_i}\left\{r_i(k) + \rho \sum_{j\in S} [p_{ij}(k) - b_j] w_j^{(n-1)}\right\} \qquad (i\in S)$$

to a fixed point vector w, then

$$(37)\qquad (1-\rho)w_\ell^{(n)} + \rho \sum_{j\in S} b_j w_j^{(n)} \xrightarrow[(n\to\infty)]{} a_\ell$$

and an ℓ-optimal policy f_ℓ^∞ is eventually generated by (36).

If we select $b_j = \min_{i,k} p_{ij}(k)$, then Doeblin's criterion

$$(38)\qquad \rho\Big[1 - \sum_{j\in S} \min_{i,k} p_{ij}(k)\Big] < 1$$

is sufficient for the convergence of (36). In this case the unique fixed point w may also be calculated by applying Howard's policy iteration procedure to the functional equation $w = \Phi_b w$.

Letting

$$(39)\qquad \underline{p}_{ij} := \min_{k\in K_i} p_{ij}(k)$$

$$(40)\qquad \bar{p}_{ij} := \max_{k\in K_i} p_{ij}(k)$$

$$(41)\qquad b_{ij} := \frac{1}{2}\,[\underline{p}_{ij} + \bar{p}_{ij}]$$

select

$$(42)\qquad b_j = \operatorname*{median}_{1\le i\le M} b_{ij} = b_{i^*(j),j} \qquad (M \text{ odd}).$$

Another sufficient condition for the convergence of (36) is given by

$$(43)\qquad \rho \max_{j\in S}\Big\{-\sum_{i\in S_1(j)} \underline{p}_{ij} - \frac{1}{2}\,\underline{p}_{i^*(j),j} + \frac{1}{2}\,\bar{p}_{i^*(j),j} + \sum_{i\in S_2(j)} \bar{p}_{ij}\Big\} < 1$$

with $S_1(j)$ and $S_2(j)$ defined by

$$(44)\qquad S_1(j) = \{i \,|\, b_{ij} < b_j, \quad i \in S\}$$

$$(45)\qquad S_2(j) = \{i \,|\, b_{ij} > b_j, \quad i \in S\},$$

cf. REETZ [7].

We conjecture that many properties of transient MDP with unbounded rewards can be extended to the class of ρ-recurrent MDP with unbounded rewards, cf. HINDERER [3] and WESSELS [9]. Furthermore it would be desirable to investigate the connection of the concept of annuity with the usual definitions of average payoff. Finally, multiple punctuations should make

possible the extension of some results to multiple-chain MDP, cf. SCHWEITZER & FEDERGRUEN [8].

The author has developed a third equivalent characterization of annuities using a generalized concept of steady-state probabilities. Details are omitted for lack of space.

REFERENCES

1. BLACKWELL, D., *Discrete dynamic programming,* Ann. Math. Statist. 33, 719-726 (1962).

2. FERSCHL, F., *Markovketten,* Springer-Verlag, Berlin, 1970.

3. HINDERER, K., *Bounds for stationary finite-stage dynamic programs with unbounded reward functions,* Preprint 1975.

4. HORDIJK, A., *Regenerative Markov decision models,* Mathematical Centre Report BW 49/75, Amsterdam. To appear in Math. Programming Systems.

5. HOWARD, R.A., *Dynamic programming and Markov processes,* M.I.T. Press, Cambridge, Mass. 1960.

6. HÜBNER, G., *On the fixed points of the optimal reward operator in stochastic dynamic programming with discount factor greater than one,* ZAMM 56, T 348-T 349 (1976).

7. REETZ, D., *Eine Klasse von expandierenden Markoffschen Entscheidungsprozessen mit Turnpike-Eigenschaft,* Proc. Operations Research 4, 323-332, 1974.

8. SCHWEITZER, P.J. & A. FEDERGRUEN, *The functional equations of undiscounted Markov renewal programming,* Mathematical Centre Report BW 60/76, Amsterdam.

9. WESSELS, J., *Markov programming by successive approximations with respect to weighted supremum norms,* to appear in J. Math. Anal. Appl.

10. WHITE, D.J., *Dynamic programming, Markov chains, and the method of successive approximations,* J. Math. Anal. Appl. 6, 373-376 (1963).

SMOOTHING SYSTEM DESIGN AND PARAMETRIC MARKOVIAN PROGRAMMING

B.S.Verkhovsky
Princeton University, Princeton, USA

1. INTRODUCTION

Great number of examples of smoothing systems can be found in [1]. A bibliography about Markovian decision processes now has more than a thousand papers and books. See, for example, [2], [5] and [14]. Probabilistic analysis of reservoir as a smoothing system, a reader can find in [6]. Mathematical modelling and analysis of optimal policy for operating with reservoir has been described in [15], [16], and [10]. The main goal of this paper is to describe the smoothing system as a Markovian decision process, *depending on a parameter*. In the case of water resource system, this parameter is the size of the reservoir. More generally, to decrease an influence of uncertainty of dynamics, one can use the smoothing systems (reservoir for water resources system, some amount of real money for banking, national reserves for a country, etc.). The more the capacity of the smoothing system, the better but how much is worthwhile to spend on it. A mathematical model of such systems will be given. Optimal design of smoothing systems is *three-level-optimization* (first level is the optimal usage of resource or optimization of singletime efforts; second level is the optimization of the total expected incomes for whole time of operating; and third level is the optimal choice of the parameter or parameters of smoothing). This three-level-optimization needs lots of calculations and the main goal of the paper is to represent some *regular* properties of the problem and to show

how one can exploit them to decrease essentially the necessary amount of computation. Special algorithms have been developed by the author for the case, [8, 9, 11 and 17]. The computer experiment results are given.

2. DESCRIPTION OF A SMOOTHING SYSTEMS

Let: w_t be the volume of water in a reservoir at moment t.

U_t be the volume of water taken from the reservoir for use in a water resource system (for irrigation, water supply, industrial use, etc.)

Q_t be the inflow to the reservoir at moment t. Q is a stochastic value with given distribution of probabilities:

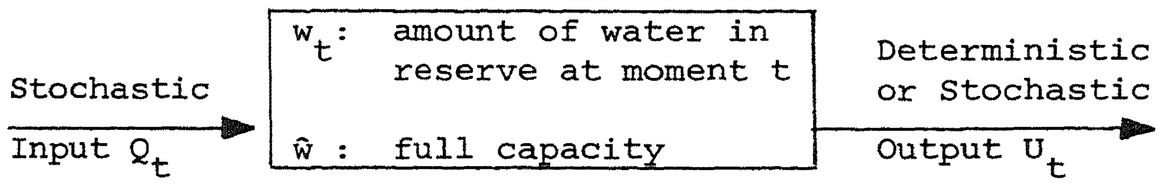
$$
\begin{aligned}
&Q^{(1)} \sim P_1 \\
&Q^{(2)} \sim P_2 \\
&\cdots\cdots\cdots \\
&Q^{(n)} \sim P_n \\
&\sum_{i=1}^{n} P_i = 1, \qquad P_i \geq 0 \ .
\end{aligned}
\tag{1}
$$

Let $\hat{w}$ be the volume of the reservoir (i.e.: ful capacity). For such a system, the equation of transformation is:

$$
w_{t+1} = \min\{\hat{w}, w_t + Q_t - U_t\} \tag{2}
$$

(The possibility of overflow has to be taken into account)

Smoothing System: Reservoir



3. DYNAMIC PROGRAMMING EQUATION

The system is controlled by the choice of U_t. The optimal pattern of control is to choose such $U_1, U_2, \ldots, U_{t-1}, U_t, \ldots$, which maximizes the

expected effect of operating the system for all time of operation. Let $\omega(U)$ be the direct income if one uses a volume of water U in the water resources system (a single time effect).

NOTE. $\omega(U)$ may be a very complicated function and very expensive to calculate. But it is very important to stress that to obtain function $\omega(U)$, we have to consider our water resources system *statically* only.

Let v_i be the optimal expected income for the whole time of operation of the system if at the moment $t = 0$, the volume of water in the reservoir is i. $\{v_i\}$ are unknowns for all $i = 1,2,\ldots,n$, and to find them is our initial problem.

Let $P^{(k)}$ be the transition probability matrices: $\{p_{ij}^{(k)}\}$ is defined as the porbability that after state i, the system will be in state j if we use control k. All matrices $\{p_{ij}^{(k)}\}$ can be found from the equation of transformation (2) and by using the distribution of probabilities (1). All $\{p_{ij}^{(k)}\}$ implicitly depend on the capacity of the smoothing system, $\hat{w}$ (see the table).

From the definition of v_i and from the principle of optimality [2], we may write the equations:

$$(3) \qquad v_i = \max_{0 \le k \le i} \{\omega_{ik} + \beta \sum_{j=1}^{n} p_{ij}^{(k)} v_j\}, \qquad i = 1,2,\ldots,n$$

or

$$(4) \qquad \bar{v} = \max_{k} \{\bar{\omega}_k + \beta P^{(k)} \bar{v}\}$$

where β with $0 < \beta < 1$ is the discount factor.

The equations (3) and (4) describe how to find the optimal control k as a function of state i *for a given full capacity* $\hat{w}$. Later, we will consider the case of how to find the optimal $\hat{w}$ - this is the main subject of this paper.

4. ALGORITHMS FOR EQUATION (3)

1. The problem can be reduced to a linear programming problem, [5].
2. Another approach has been suggested by R.A. HOWARD [2]. This algorithm converges to the solution of (3) in a finite number of iterations. However, the algorithm needs to find the solution of the system of equations

(5) *at every iterative step*. This means that many computations are required for systems with many states i (i = 1,2,...,n).

3. An iterative algorithm (B.S. VERKHOVSKY, V.A. SPIVAK), [16].

$$v^{r+1} = \max_k \{\omega_k + \beta P^{(k)} v^r\}. \tag{5}$$

The scheme (6) converges to the solution of (3) if

$$0 < \beta < 1. \tag{6}$$

In this case

$$\text{if } \tilde{v}^0 = \frac{c}{1-\beta} \quad \text{where} \quad c = \max_{i,k} \omega_{ik} \quad \text{and if} \quad \underset{\sim}{v}^0 = 0$$

then, by induction

$$\tilde{v}^r = \frac{c\beta^n}{1-\beta} + \underset{\sim}{v}^r. \tag{7}$$

On the other hand

$$\underset{\sim}{v}^0 \leq \underset{\sim}{v}^1 < \ldots < \underset{\sim}{v}^r \leq \ldots \leq v \leq \ldots \leq \tilde{v}^r \leq \ldots \leq \tilde{v}^1 \leq \tilde{v}^0.$$

Since

$$\lim_{r\to\infty} (\tilde{v}^r - \underset{\sim}{v}^r) = \frac{c}{1-\beta} \cdot \lim_{r\to\infty} \beta^r = 0,$$

then

$$\lim_{r\to\infty} (v^r - v) = 0.$$

After r iterations, the absolute error ε_r is given by

$$\varepsilon_r = \tilde{v}^r - v \leq \tilde{v}^r - \underset{\sim}{v}^r = \frac{c\beta^r}{1-\beta}.$$

As an example, for $\beta = 0.90$ and $c = 1$, approximately 240 iterations are needed to decrease ε_r to a value of 10^{-10}. Twice as many iterations are needed if $\beta = 0.95$. -

5. DESIGN OF SMOOTHING SYSTEMS

Consider a time scale:

where T_d is defined as the moment to make the decision about design of the proposed smoothing system and T_s is defined as the moment when smoothing system starts to operate. The peculiarity is that at time T_d one does not know from which state i the system will start to operate at time T_s.

If ρ_i is the probability that the system will start operating from state i, and $f(\hat{w})$ is the expenditure required to construct the system to capacity $\hat{w}$, the expected optimal profit is:

$$\max_{\hat{w}} \{\sum_i \rho_i v_i(\hat{w}) - f(\hat{w})\} \tag{8}$$

where $v_i(\hat{w})$ is the solution, for given $\hat{w}$, of the system of functional equations (3).

6. PROPERTY OF THE FUNCTIONS $v_i(\hat{w})$

THEOREM. *For every* i, $v_i(\hat{w})$ *is a concave function of* $\hat{w}$.

PROOF. See Appendix 1. We here outline the proof.

1. The equation (3) can be rewritten in the following form:

$$v_i(\hat{w}) = \max_{0\le k\le i} \{\omega_k + \beta \sum_{j=1}^{m} p_j v_{\min(\hat{w},i-k+j)}(\hat{w})\} \tag{9}$$

2. $v_i(\hat{w})$ is a concave function of i.
3. $v_i(\hat{w})$ is an increasing function of i and $\hat{w}$.
4. $\min(\hat{w},i-k+j)$ is a concave function of $\hat{w}$.
5. The systems of equations

$$v_i^{(r+1)}(\hat{w}) = \max_{0\le k\le i} \{\omega_k + \beta \sum_{j=1}^{m} p_j v_{\min(\hat{w},i-k+j)}^{(r)}(\hat{w})\}$$

converge to the solution of (9).

$\{\cdot\}^{(r)}$ are concave functions of $\hat{w}$ for all r

$(r = 0,1,2,....)$

6. *Then* for every i, $v_i(\hat{w})$ is a concave function of $\hat{w}$. $(i = 1,2,...,n)$

In this case, the function

$$\psi(\hat{w}) = \sum_i \rho_i v_i(\hat{w}) - f(\hat{w})$$

is concave if $f(\hat{w})$ is a linear or convex function. This is possible if $\psi(\hat{w})$ satisfies the inequality:

$$\sum_i \rho_i \frac{d^2 v_i(\hat{w})}{d\hat{w}^2} - \frac{d^2 f(\hat{w})}{d\hat{w}^2} < 0, \quad \text{for all } \hat{w} .$$

However, to optimize the one-dimensional search for the optimum of the function $\psi(\hat{w})$, the more important property is the unimodality of the function $\psi(\hat{w})$, [7].

7. ALGORITHM FOR THE MARKOVIAN PROCESS DECISIONS (3)

Step 1: For given v^r, find control k:

$$\max_k \{\omega_k + \beta P^{(k)} v^r\} = \omega_{k_r} + \beta P^{(k_r)} v^r . \tag{10}$$

Let for simplicity, the notations

$$\omega_{k_r} \equiv \omega_* \text{ and } P^{(k_r)} \equiv P_*$$

Step 2: Iterate:

$$x_r^{s+1} = \omega_* + \frac{\beta P_* x_r^s}{\phi(I-\beta P_*)x_r^s} \quad \text{for } s = 0,1,..., \text{ r is fixed} \tag{11}$$

where x_r^s is a non-negative m-dimensional column vector, $x_r^0 \equiv v^r$, ϕ of m-dimensional row vector and

$$\phi \equiv \frac{c}{c\omega_*}$$

where c is an m-dimensional non-negative row vector, and $c\omega_*$ is the inner product, I is identical matrix m × m.

Step 3: Stop iterative process when $\|x_r^{s+1} - x_r^s\| \leq \varepsilon$ where ε is a given accuracy.

Step 4: $v^{r+1} \equiv x_r^{s+1}$.

Step 5: If $\|v^{r+1} - v^r\| \leq \delta$ where δ is a given accuracy, then v^r is the solution of the Markovian dynamic programming equations (3) and k_r^r is the corresponding optimal controls. Otherwise, (if $\|v^{r+1} - v^r\| > \delta$), return to the Step 1, [18].

8. PROPERTIES OF THE ITERATIVE PROCESS (11)

1. It converges extremely fast even if β is very close to unity (for instance $\beta = .999$). For example, for $m = 60$ to get the given accuracy $\varepsilon = 10^{-10}$ one needs just 8 iterations.
2. The process (11) converges even if β is greater than unity.
3. The larger the dimension m, the faster the rate of convergence.
4. The number of additional operations for the process (11) negligible when compared with the simplest iterative method, the Jacobi algorithm (for $m > 10$ the number of additional operations are less than 1%).

For more details about process (11), see Appendix II.

9. GENERALIZED DELINEARIZATION ALGORITHM

More general algorithm can be used for solution a system of linear equations $x = b + Ax$. Let us consider the process:

$$v^{r+1} = b + \sum_{s=1}^{3} \theta_s \gamma_s A v^r, \quad \text{where} \quad \sum_s \gamma_s = 1, \theta_s \geq 0 \quad \text{and}$$

$$\gamma_1^r = [\phi(I-A)v^r]^{-1},$$

$$\gamma_2^r = \left(\frac{\phi A v^r}{\phi v^r_{-1}}\right)^a, \quad a = \frac{\lambda_1(A)}{1-\lambda_1(A)}$$

$$\gamma_3^r = \left(\frac{1+\phi A v^r}{\phi v^r}\right)^d, \quad d = [1-\lambda_1(A)]^{-1},$$

where $\lambda_1(A)$ is largest eigenvalue of matrix A. If for some r one of the denominators equal zero, then corresponding θ_s also must be equal zero. At least, one of the three denominators will not be equal zero. For more details about γ_s^r see [9,11,17]. This algorithm can be used with other algorithms. For example, successive overrelaxation [19], Abramov's algorithm [20].

APPENDIX I

PROPERTY OF FUNCTIONS $V_i(\hat{w})$

LEMMA. *If* V(u) *is a concave function of* u *and an increasing function of* u, *and* u(R) *is a concave function of* R, *then* V(R) *is concave function of* R.

PROOF.

1. $V(\lambda_1 u_1+\lambda_2 u_2) \geq \lambda_1 V(u_1) + \lambda_2 V(u_2), \lambda_1 + \lambda_2 = 1, \lambda_i \geq 0$

2. $u(\gamma_1 R_1+\gamma_2 R_2) > \gamma_1 u(R_1) + \gamma_2 u(R_2), \gamma_1 + \gamma_2 = 1, \gamma_i \geq 0.$

Let us consider

$$V[u(\gamma_1 R_1+\gamma_2 R_2)] \geq V[\gamma_1 u(R_1) + \gamma_2 u(R_2)]$$

since V(u) is an increasing function of u.
On the other hand,

$$V[\gamma_1 u(R_1) + \gamma_2 u(R_2)] \geq \gamma_1 V\, u(R_1) \quad + \gamma_2 V[u(R_2)].$$

Thus, V(R) is a concave function of R. For convenience, let us rewrite equation (3) in the following form:

$$V(x) = \max_{0\leq y\leq x} \{\omega(y) + \beta \sum_{z=1}^{m} p(z)\cdot v[\min(R,x-y+z)]\}, \qquad 0 \leq x \leq R.$$

Let

$$V(R,x) \equiv \beta \sum_{z=1}^{m} p(z)\cdot v[\min(R,x)]$$

the V(•) is a concave function of R and x because

$$\beta > 0 \quad \text{and} \quad p(z) \geq 0 \quad \text{for all} \quad z.$$

Consider

$$v_*(x_1) = \max_{0\leq y\leq x_1} \{\omega(y) + V(R,x_1-y) \quad = \omega(y_1) + V(R,x_1-y_1)$$

$$v_*(x_2) = \max_{0\leq y\leq x_2} \{\omega(y) + V(R,x_2-y)\} = \omega(y_2) + V(R,x_2-y_2)$$

$$v_*(\ _1x_1+\ _2x_2) = \max_{0\le y\le\lambda_1x_1+\lambda_2x_2} \{\omega(y) + V(R,\lambda_1x_1+\lambda_2x_2-y)\}.$$

Consider

$$\lambda_1v_*(x_1) + \lambda_2v_*(x_2) = \lambda_1\omega(y_1) + \lambda_2\omega(y_2) + \lambda_1V(R,x_1-y_1) + \lambda_2V(R,x_2-y_2) \le$$

$$\le \omega(\lambda_1y_1+\lambda_2y_2) + V[R,\lambda_1(x_1-y_1) + \lambda_2(x_2-y_2)] \le$$

$$\le \max_{0\le y\le\lambda_1x_1+\lambda_2x_2} \{\omega(y) + V(R,\lambda_1x_1+\lambda_2x_2-y)\} = v_*(\lambda_1x_1+\lambda_2x_2).$$

Thus, $v_*(x)$ is a concave function of x. Let us consider the concavity of function $v_*(x,R)$ on R.

$$v_*(x,R,) = \max_{0\le y\le x} \{\omega(y) + V(R_1,x-y)\} = \omega(y_1) + V(R_1,x-y_1), \quad (y_1\le x)$$

$$v_*(x,R_2) = \max_{0\le y\le x} \{\omega(y) + V(R_2,x-y)\} = \omega'y_2) + V(R_2,x-y_2), \quad (y_2\le x)$$

$$v_*(x,\lambda_1R_1+\lambda_2R_2) = \max_{0\le y\le x} \{\omega(y) + V(\lambda_1R_1+\lambda_2R_2,x-y)\}$$

$$\lambda_1v_*(x,R_1) + \lambda_2v\ (x,R_2) = \lambda_1\omega(y_1) + y_2\omega(y_2) + \lambda_1V(R_1,x-y_1)+\lambda_2V(R_2,x-y_2)$$

$$\le \omega(\lambda_1y_1+\lambda_2y_2) + V[\lambda_1R_1 + \lambda_2R_2,\lambda_1(x-y_1) + \lambda_2(x-y_2)] \le$$

$$\le \max_{0\le y\le x} \{\omega(y) + V[\lambda_1R_1 + \lambda_2R_2,x-y)\} = v_*(x,\lambda_1R_1+\lambda_2R_2).$$

We have used the property that V [•] is a concave function of R and x.

APPENDIX II

ANALYSIS OF CONVERGENCE

To consider the convergence of the process (II), let is simplify the notations:

$$v^{(n)} \equiv x^{s}_{r}, \quad A \equiv \beta P_{*}, \quad b \equiv \omega_{*}$$

LEMMA. *The process* (II) *can be rewritten in the form*

$$v^{(n)} = \frac{D^{n}v(0)}{hD^{n-1}v(0)}$$

where

$$D \equiv b\phi(I-A) + A, \qquad h \equiv \phi(I-A).$$

PROOF. (by induction)

THEOREM. *If* $\phi \geq 0$, $\phi b = 1$, $D \geq 0$, $\phi v^{(0)} \neq 0$, *and matrix* $(I-A)^{-1}$ *exists, then* $\lim_{n\to\infty} v^{(n)} = \tilde{v}$, *where* $\tilde{v}$ *is the solution of the equation* $v = b+Av$.

PROOF.

1. ϕ is the eigenvector of matrix D^{T}, corresponding to eigenvalue 1. Indeed, $\phi D = \phi\, b\phi(I-A) + \phi A = \phi$. Hence, there exists another vector u_0 orthogonal to ϕ and such that $Du_0 = u_0$. Let u_0 and ϕ be normalized in such a way that $\phi u_0 = 1$. Then, $\lim_{n\to\infty} D^{n} = u_0\phi$ [12,13].
2. Thus

$$\lim_{n\to\infty} v^{(n)} = \frac{u_0\phi v^{(0)}}{hu_0\phi v^{(0)}} \equiv \tilde{v} \quad \text{since } \phi v^{(0)} \neq 0.$$

It must be shown that u_0/hu_0 is the solution of (1) and does not depend on the vector ϕ. If $hu_0 \neq 0$, then from $Du_0 = u_0$, we have

$$u_0 = bhu_0 + Au_0 \quad \text{or} \quad \frac{u_0}{hu_0} = b + A\left(\frac{u_0}{hu_0}\right).$$

Thus $u_0/hu_0 = \tilde{v}$ is the solution of (1). To prove that $hu_0 \neq 0$, assume the opposite: $hu_0 = 0$. Then $u_0 = bhu_0 + Au_0 = Au_0$. Hence u_0 is an eigenvector of matrix A corresponding to a given eigenvalue 1. This means that the

characteristic equation $|A-YI| = 0$ has a root $\lambda = 1$ or the determinant $|A-I| = 0$. However, this contradicts with the assumption that matrix (I-A) exists. Thus $hu_0 \neq 0$.

RATE OF CONVERGENCE

Consider $\varepsilon_n \equiv v^{(n)} - v$, where v is the solution of the system $v = b + Av$.

THEOREM. *In the neighborhood of the solution*

$$\varepsilon_{n+1} = (D-vh)\varepsilon_n + o(\varepsilon_n).$$

PROOF.

$$\varepsilon_{n+1} = \frac{(D-vh)(\varepsilon_n+v)}{h(\varepsilon_n+v)} = (D-vh)\varepsilon_n/(1+h\varepsilon_n).$$

Evidently, $(D-vh) = 0$ and $hv = 1$. If vector ε_n is small, then $h\varepsilon_n$ is also small compared with 1. Thus

$$\varepsilon_{n+1} \approx (D-vh)\varepsilon_n.$$

Taking into account that

$$D - vh = A(I-vh) = A[I-(I-A)^{-1}b\phi(I-A)],$$

one has

$$\varepsilon_{n+1} \approx A[I-(I-A)^{-1}b\phi(I-A)]\varepsilon_n = AB\varepsilon_n,$$

where

$$B \equiv [I-(I-A)^{-1}b\ (I-A)].$$

Matrices AB and $A(I-b\phi)$ have the same spectrum.

THEOREM. *If ϕ is eigenvector of matrix A^T corresponding to positive eigenvalue λ_0, and $\phi \geq 0$, $A \geq 0$, and exists the inverse matrix A^{-1}, then every nonzero eigenvalue of matrix AB is at the same time the eigenvalue of matrix A, however λ_0 is not eigenvalue of matrix AB; i.e. $r(AB) < r(A)$.*

REFERENCES

1. BELLMAN, R., *Dynamic programming*, Princeton University Press, Princeton, 1957, pp. 29, 70, 215.

2. HOWARD, R., *Dynamic ogramming and Markov processes*, The Technology Press of M.I.T., New York, 1960.

3. BELLMAN, R., *Introduction to matrix analysis*, McGraw-Hill Book Company, Inc., New York 1960.

4. GANTMACHER, F.R., *The theory of matrices*, V.2, Chelsea Publishing Company, New York, 1959.

5. WOLFE, P. & G.B. DANTZIG, *Linear programming in a Markov chain*, Operations Research 10 (1962) pp. 702-710.

6. MORAN, P.A.P., *The theory of storage*, 1959, London.

7. WILDE, D.J., *Optimum seeking methods*, Prentice Hall, 1964.

8. VERKHOVSKY, B.S., *Algorithm with non-linear acceleration for a system of linear equations*, Department of Civil Engineering Technical Report 76-WR-1, Princeton University, 1976.

9. VERKHOVSKY, B.S., *Algorithm with controlled feedback for system of equations with stochastic matrix*, IBM Technical Disclosure Bulletin, V.18, N10, March 1976, pp. 3466-3467.

10. BULINSKAYA, E.V., *Steady-state solutions in problems of optimum inventory control*, Theory Probability Appl. 9(1964), 502-507, Akademia Nank USSR.

11. VERKHOVSKY, B.S., *Algorithm for system of equations with stochastic matrix*, IBM Technical Disclosure Bulletin V.18, N10 March 1976, pp. 3464-3465.

12. KREIN, M.G. & M.A. ROOTMAN, *Linear Operators*, Uspechi Mathematicheskich Nauk, 3, N1, 1948 (Russian).

13. KARLIN, S., *Positive operators*, J. Math.Mech., V.8, N6, 1959 pp. 901-937.

14. DERMAN, C., *Finite state Markovian decision processes*, V.67, Mathématics in Science and Engineering, Academic Press, N.Y. 1970.

15. RUSSEL, C.B., *An optimal policy for operating a multi-purpose reservoir*, Oper. Res., V.20, N6, pp. 1181-1189 (1972).

16. VERKHOVSKY, B.S. & V.A. SPIVAK, *Water systems optimal design and controlled stochastic processes*, Ekonomika 1 Mathematicheskie Metody, Vol. VIII, N6, pp. 966-972, 1972.

17. VERKHOVSKY, B.S., *Freedback algorithm for system of equations*, IBM Technical Disclosure Bulletin, Vol. 18,10, march 1976, pp. 3468-69.

18. VERKHOVSKY, B.S., *Optimal complex use of controlled water resources of a basin*, Mathematical Models in Hydrology. Proceedings of the International Symposium, Warsaw, July, 1971.

19. YOUNG, D.M. & R.T. GREGORY, *A survey of numerical mathematics*, Vol. 2, 1973, Addisson-Wesley Publishing Co., pp. 1026-1039.

20. ABRAMOV, A.A., *Concerning a procedure of acceleration of iteration processes*, Doklady Akad. Nauk. SSSR, Vol. 74, 1950, pp. 1051-1052.

VALUE-ITERATION IN UNDISCOUNTED MARKOV DECISION PROBLEMS
PART I: ASYMPTOTIC BEHAVIOUR

A.Federgrün
Mathematical Centre, Amsterdam, The Netherlands

P.J.Schweitzer
IBM Thomas J. Watson Research Center, Yorktown Heights, USA

H.C.Tijms
Mathematical Centre / Free University, Amsterdam, The Netherlands

0. INTRODUCTION AND SUMMARY

We consider undiscounted Markov Decision Problems (MDP's) with finite state and action spaces.

$\Omega = \{1,\ldots,N\}$ denotes the state space, $K(i)$ the finite set of alternatives in state i, q_i^k the one-step expected reward and $P_{ij}^k \geq 0$ the transition probability to state j, when alternative $k \in K(i)$ is chosen in state i $(i = 1,\ldots,N)$.

Both the Policy Iteration Algorithm (cf. HOWARD [9]) and various Linear Programming formulations (cf. e.g. DENARDO & FOX [4]) provide exact and finite algorithms to find maximal gain policies as well as the maximal gain rate vector. However, when N, the number of states in the system becomes very large, both methods become infeasible since requiring the solution of large systems of equations, at each step of the procedure. As a consequence, the only practical way of locating maximal gain policies in large scale problems, is by using some successive approximation technique (cf. part III of this paper) which is based upon the value iteration equations:

$$(0.1) \qquad v(n+1)_i = \max_{k \in K(i)} \left\{ q_i^k + \sum_{j=1}^{N} P_{ij}^k \, v(n)_j \right\}, \quad i \in \Omega; \quad n = 1,2,\ldots$$

where $v(0)$ is a given N-vector.

Note that $v(n)_i$ may be interpreted as the maximal total expected reward in a planning horizon of n epochs, when starting in state i, and given an

amount $v(0)_j$ is given when ending up at state j.

This paper gives a survey of the state of affairs with respect to the characterization of the asymptotic behaviour of the sequence $\{v(n)\}_{n=1}^{\infty}$.

The study of the latter is motivated by at least the following two considerations:

a) the asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$ has important implications for the (convergence) properties of the above mentioned successive approximation techniques, and of the sequences of policies generated by these methods.

b) in most practical situations, the decisionmaker faces a planning horizon which is finite though large, and the exact length of which is often unknown in advance. Characterizing the behaviour of $\{v(n)\}_{n=1}^{\infty}$ is essential when considering the infinite horizon model with the average return per unit time criterion as an approximation for these finite planning horizon models.

In section 1, we give some notation and preliminary results. In section 2, a short historic review is given of the literature on this subject.
Next we discuss some recent developments on this topic as set down inter alia in [21],[22] and [23]. The methods used in these papers involve the set of all randomized policies, and especially its chain- and periodicity structure. The latter are discussed in section 3, whereas in section 4 we give a number of properties of the solution set of the optimality equation, which are needed for the remainder.

In section 5 we finally summarize what is known on the asymptotic behaviour of the sequence $\{v(n)\}_{n=1}^{\infty}$, and in section 6 we discuss some of its implications with respect to the working of the value-iteration method and with respect to some turnpike results.

1. NOTATION AND PRELIMINARIES

A (stationary) randomized policy f is a tableau $[f_{ik}]$ satisfying $f_{ik} \geq 0$ and $\sum_{k \in K(i)} f_{ik} = 1$, where f_{ik} is the probability that the k-th alternative is chosen when entering state i. We let S_R denote the set of all randomized policies, and S_P the set of all pure (non-randomized) policies (i.e. each $f_{ik} = 0$ or 1).

Associated with each $f \in S_R$, are a N-component reward vector q(f) and $N \times N$-matrix P(f):

$$(1.1) \qquad q(f)_i = \sum_{k\in K(i)} f_{ik}\, q_i^k;\ P(f)_{ij} = \sum_{k\in K(i)} f_{ik}\, P_{ij}^k,\ 1 \le i,\ j \le N.$$

Note that $P(f)$ is a stochastic matrix ($P(f)_{ij} \ge 0$, $\sum_{j=1}^{N} P(f)_{ij} = 1$; $1 \le i, j \le N$). For any $f \in S_R$, we define the stochastic matrix $\Pi(f)$ as the Cesaro limit of the sequence $\{P^n(f)\}_{n=1}^{\infty}$ and the *fundamental* matrix $Z(f)$ as $[I-P(f) + \Pi(f)]^{-1}$. These matrices always exist (cf. [10]).

Denote by $n(f)$ the number of subchains (*closed, irreducible sets of states*) for $P(f)$. Let $R(f) = \{j \mid \Pi(f)_{jj} > 0\}$, i.e. $R(f)$ is the set of recurrent states for $P(f)$. For all $m = 1,\ldots,n(f)$ let $d^m(f) \ge 1$ denote the period on $C^m(f)$, the m-th subchain of $P(f)$. In addition let $\{C^{m,\beta}(f) \mid \beta = 1,\ldots,d^m(f)\}$ indicate the set of cyclically moving subsets (c.m.s.) of $C^m(f)$ numbered such that for any $m = 1,\ldots,n(f)$ and $\beta = 1,\ldots,d^m(f)$ (cf. [10]):

$$(1.2) \qquad i \in C^{m,\beta}(f) \Rightarrow P(f)_{ij} > 0 \text{ only if } j \in C^{m,\beta+1}(f)$$

with the convention that hereafter β in $C^{m,\beta}(f)$ is taken modulo $d^m(f)$, e.g. $C^{m,\beta+1}(f) = C^{m,1}(f)$ if $\beta = d^m(f)$. We recall that, for all $i \in C^m(f)$, $m = 1,\ldots,n(f)$:

$$(1.3) \qquad d^m(f) = \text{greatest common divisor (g.c.d.) of } \{n \mid P(f)_{ii}^n > 0\}$$

For each $f \in S_R$, we define the gain rate vector $g(f) = \Pi(f)q(f)$, such that $g(f)_i$ represents the long run average expected return per unit time, when the initial state is i, and policy f is used. Next, define the maximal gain rate vector g^* by

$$(1.4) \qquad g_i^* = \sup_{f\in S_R} g(f)_i, \qquad i = 1,\ldots,N.$$

We know from DERMAN [5] that there exist pure policies f which attain the N suprema in (1.4) simultaneously. As a consequence we define:

$$(1.5) \qquad S_{PMG} = \{f \in S_P \mid g(f) = g^*\};\ S_{RMG} = \{f \in S_R \mid g(f) = g^*\}$$

as the set of all pure and the set of all randomized maximal gain policies.

Finally we consider the well-known pair of optimality equations for the average return per unit time criterion:

$$(1.6) \qquad g_i = \max_{k\in K(i)} \sum_j P_{ij}^k\, g_j, \qquad i \in \Omega$$

$$(1.7) \qquad v_i + g_i = \max_{k \in L(i)} \left\{ q_i^k + \sum_j P_{ij}^k v_j \right\}, \quad i \in \Omega$$

where

$$L(i) = \{k \in K(i) \mid g_i = \sum_j P_{ij}^k g_j\}.$$

We recall (cf.e.g. [21], th.3.1) that there always exists a solution pair (g,v) to (1.6) and (1.7). In addition any solution pair (g,v) to (1.6) and (1.7) has $g = g^*$, which implies that the g-part of the solution and hence each of the sets L(i) are uniquely determined. Finally let

$$(1.8) \qquad V = \{v \in E^N \mid v \text{ satisfies } (1.7)\}.$$

For any $v \in E^N$, define

$$b(v)_i^k = q_i^k - g_i^* + \sum_{j=1}^N P_{ij}^k v_j - v_i,$$

$i \in \Omega$, $k \in K(i)$ and note that

$$(1.9) \qquad v \in V \iff \max_{k \in L(i)} b(v)_i^k.$$

2. HISTORICAL REVIEW

The first asymptotic property of the sequence $\{v(n)\}_{n=1}^{\infty}$, is due to BELLMAN [2] who showed that if every one-step transition probability P_{ij}^k is strictly positive:

$$(2.1) \qquad \lim_{n \to \infty} \frac{v(n)_i}{n} = g^*, \quad \text{for all } i \in \Omega$$

where g^* is the maximal gain rate. Note that Bellman's assumption is the strongest possible one can make with respect to the chain- and periodicity structure of the problem. HOWARD [9] conjectured that there generally exist two N-vectors g^* and v^* such that

$$(2.2) \qquad \lim_{n \to \infty} v(n) - ng^* - v^* = 0.$$

However, (2.2) may clearly fail to hold, if some of the transition probability matrices (tpm's) are periodic, as is illustrated by the two-state Markov process which has $P_{12} = P_{21} = 1$ and $q_1 = q_2 = 0$ (Take e.g. v(0) = [1,0] and note that $\{v(n)\}_{n=1}^{\infty}$ alternates between the two limit points [1,0] and [0,1]).

BROWN ([3], th.4.3) succeeded in showing that $\{v(n) - ng^*\}_{n=1}^{\infty}$ is bounded, provided g^* is taken as the maximal gain rate vector. In addition he claimed the existence of some integer $J \geq 1$ such that

$$\lim_{n\to\infty} v(nJ+r) - (nJ+r)g^* \text{ exists for all } v(0) \in E^N; \quad r = 0,\ldots,J-1. \tag{2.3}$$

Unfortunately his proof contained an error as was pointed out in LEMBERSKY [13]. Sufficient conditions for the convergence result in (2.2) were established inter alia in WHITE [25] and SCHWEITZER ([18], [20]). The former used the assumption: there exists a state s and an integer $\nu \geq 1$, such that

$$P(f^1) \ldots P(f^\nu)_{is} > 0 \quad \text{for all } f^1,f^2,\ldots,f^\nu \in S\,; \quad i \in \Omega \tag{2.4}$$

and the latter obtained convergence for the case where all of the policies are unichained and aperiodic, which encompasses (2.4) as a special case. In addition, WHITE finds that (under his condition (2.4)) the approach to the limit in (2.2) is geometric (cf. part II of this paper).

The result stated in (2.3) was next studied in all generality in LANERY [11]. Although [11] contains most of the elements needed in order to establish the convergence result, the proof given is very lengthy and seems to be complete only for the case where every state is recurrent for some maximal gain policy. A correct proof is obtained in BATHER [1] for the case where the maximal gain rate is independent of the initial state of the system, i.e. $g_i^* = g^*$; however the proof in [1] does not clarify how the minimal integer J for which (2.3) holds, depends upon the structure of the problem. Related convergence results for MDP's with denumerable and general state spaces and for continuous time Markov Decision Processes were obtained in respectively HORDIJK, SCHWEITZER and TIJMS [8], TIJMS [24] and LEMBERSKY [13].

Finally, the first approach to establish (2.3) in all generality, as well as the weakest sufficient condition for the convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ (cf. (2.2)) is to our knowledge due to [22]; the authors prove that the limit in (2.2) exists *for every* $v(0) \in E^N$, *if and only if* there exists a randomized maximal gain policy whose tpm is aperiodic (but not necessarily unichained) and has $R^* = \{i \in \Omega \mid i \text{ is recurrent for some maximal gain policy}\}$ as its set of recurrent states.

In addition, it is shown that there exists an integer $d^* \geq 1$ such that (2.3) holds if and only if J is a multiple of d^*, and a full characteriza-

tion of this asymptotic period d^* is given in terms of the chain and periodicity structure of the problem. While the above result settles the issue if one demands convergence of $\{v(n) - ng^*\}_{n=1}^{\infty}$ (or some subsequence $\{v(nJ+r)-(nJ+r)g^*\}_{n=1}^{\infty}$; $J \geq 1$; $r = 0,\ldots,J-1$) for *every* $v(0) \in E^N$, it should be noted that $\{v(n) - ng^*\}_{n=1}^{\infty}$ always converges for $v(0)$ belonging to some non-empty subset $W \subseteq E^N$.

In a subsequent paper, SCHWEITZER and FEDERGRUEN [23] returned to the issue of the *rate* of convergence. As their main result they obtain the fact that if $\lim_{n\to\infty} v(nJ+r)-(nJ+r)g^*$ exists for some $v(0) \in E^N$, $J \geq 1$ and $r = 0,\ldots,J-1$, then the approach to the limit is geometric.

Consequently, this result shows that successive approximation methods which are based on value-iteration and which locate maximal gain policies in MDP's exhibit a geometric rate of convergence.

3. THE SET OF MAXIMAL GAIN POLICIES; ITS CHAIN- AND PERIODITY STRUCTURE

The methods used in [21] and [22] make an essential use of the entire set of all *randomized* maximal gain policies.

As a consequence, we use this section in order to provide a characterization of this set, and especially of its chain- and periodicity structure.

The following lemma characterizes the correspondence between the set of maximal gain policies and the solutions to the optimality equations (1.6) and (1.7).

LEMMA 3.1. (Properties of maximal gain policies)

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

b) *Let* $f \in S_R$:

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

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

As to the proof of this lemma, we refer to [21], th.3.1, part (a) and (e).

The following example shows that the set S_{RMG} is non-convex, i.e. *arbitrary* randomization of two maximal gain policies may fail to preserve the maximal gain property:

EXAMPLE 1.

i	k	p_{i1}^k	p_{i2}^k	p_{i3}^k	p_{i4}^k	p_{i5}^k	p_{i6}^k	q_i^k
1	1	0	1	0	0	0	0	-1
	2	1	0	0	0	0	0	0
2	1	0	1	0	0	0	0	0
	2	1	0	0	0	0	0	-1
3	1	0	0	0	1	0	0	0
4	1	0	0	1	0	0	0	0
	2	0	0	0	0	1	0	0
5	1	0	0	0	1	0	0	0
6	1	0	.2	.4	.2	.2	0	0
	2	0	0	.4	.4	.2	0	-1

$g^* = (0,0,0,0,0,0) \Rightarrow$ $K(i) = L(i)$ for all $i = 1,\ldots,6$

The set V is described by the following set of equations:

$$
\begin{aligned}
(3.1)\qquad v_1^* &= \max\{-1+v_2^*; v_1^*\} \\
v_2^* &= \max\{-1+v_1^*; v_2^*\} \\
v_3^* &= v_4^* \\
v_4^* &= \max\{v_3^*; v_5^*\} \\
v_5^* &= v_4^* \\
v_6^* &= \max\{-1+.4v_3^*+.4v_4^*+.2v_5^*; .2v_2^*+.4v_3^*+.2v_4^*+.2v_5^*\}
\end{aligned}
$$

which is equivalent to:

$$
\begin{aligned}
(3.2)\qquad & -1 \le v_1^* - v_2^* \le 1: \\
& v_3^* = v_4^* = v_5^* = c \\
& v_6^* = \max\{-1+c; .2v_2^*+.8c\}.
\end{aligned}
$$

Take $f^1 = (1,1,1,1,1,1)$: $f^2 = (2,2,1,2,1,2)$ and observe that $f^1, f^2 \in S_{PMG}$. Next, let $f = \frac{1}{2}f^1 + \frac{1}{2}f^2$, and note that

$$P(f) = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & .1 & .4 & .3 & .2 & 0 \end{bmatrix} \quad ; \quad q(f) = \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ 0 \\ 0 \\ 0 \\ -\frac{1}{2} \end{bmatrix}$$

such that $-\frac{1}{2} = g(f)_1 = g(f)_2 < g_1^* = g_2^* = 0$.

An action $k \in K(i)$ is called *suboptimal* for some solution $v \in V$, if $k \notin L(i)$ or $b(v)_i^k < 0$. Note that for every $v \in V$, either action 1 in state 2 or action 2 in state 2 is suboptimal. Lemma 3.1 part (b) exhibits that the suboptimality of action 1 in state 2 (or action 2 in state 2) does not jeopardize the maximal gain-property of policy f^1 or f^2 since state 1 and state 2 are transient under $P(f^1)$ and $P(f^2)$ resp. Observe however, that by randomizing the two policies *recklessly* both state 1 and state 2 become recurrent under $P(f)$. This implies that for any $v \in V$, policy f uses in at least one recurrent state a suboptimal action, which prohibits its optimality as a consequence of lemma 3.1, part (b).

We next describe a *particular* randomization procedure which preserves the maximal gain property when applied to policies in S_{PMG}, and which in addition "simplifies" both the chain- and periodicity structure by coalescing subchains and by *reducing* the periods. Note that both the chain- and periodicity structure of a policy f merely depend upon the set of positive entries in the tpm $P(f)$, rather than upon the actual magnitudes of the transition probabilities themselves. This implies that both the chain- and periodicity structure, as well as the fact whether a policy f is maximal gain or not, merely depend upon the *set* of *positive* entries in the tableau $[f_{ik}]_{i\in\Omega, k\in K(i)}$, i.e. upon the *set* of actions that are used with positive probability, rather than upon the actual weights themselves.

LEMMA 3.2. *Let* $f^1, f^2 \in S_R$ *define a randomization* $f^* \in S_R$ by

$$(3.3) \qquad \{k|f_{ik}^* > 0\} = \begin{cases} \{k|f_{ik}^1 > 0\} & \text{for all } i \in R(f^1)\backslash R(f^2) \\ \{k|f_{ik}^2 > 0\} & \text{for all } i \in R(f^2)\backslash R(f^1) \\ \{k|f_{ik}^1 > 0\} \cup \{k|f_{ik}^2 > 0\} & \text{otherwise.} \end{cases}$$

Then

(a) $f^1, f^2 \in S_{RMG} \Rightarrow f^* \in S_{RMG}$

(b) *Let* $C(f^1,f^2) = \{C^m(f^1) \mid m = 1,\ldots,n(f^1)\} \cup \{C^m(f^2) \mid m = 1,\ldots,n(f^2)\}$

The subchains of $P(f^*)$ *are given by the equivalence classes on* $C(f^1,f^2)$ *as generated by the following equivalence relation:*

(3.4) $C \sim C' \iff$ *there exists* $\{C^{(1)} = C, C^{(2)},\ldots,C^{(n)} = C'\}$ *with* $C^{(r)} \in C(f^1,f^2)$ and $C^{(r)} \cap C^{(r+1)} \neq \emptyset$, for $r = 1,\ldots,n-1$.

In particular we have

(3.5) $R(f^*) = R(f^1) \cup R(f^2)$

(c) *For each* $m = 1,\ldots,n(f^*)$: $d^m(f^*)$ *is a common divisor (c.d.) of*

(3.6) $\{d^r(f^1) \mid c^r(f^1) \subseteq c^m(f^*),\ 1 \leq r \leq n(f^1)\} \cup \{d^r(f^2) \mid c^r(f^2) \subseteq c^m(f^*),\ 1 \leq r \leq n(f^2)\}.$

REMARK 1. We observe that, when randomizing two policies f^1, f^2 as in (3.3), the period of a subchain C of the randomized policy may fail to be the *greatest* common divisor of the periods of the subchains of $P(f^1)$ and $P(f^2)$ that are contained within C. (cf. example 1 in [22]).

Let $R^* = \{i \in R(f) \mid f \in S_{RMG}\}$ denote the set of all states that are recurrent under some maximal gain policy. The following theorem which was proven in ([21], th.3.2) gives a characterization of the chain structure of the set S_{RMG}, by showing inter alia that the set R^* may be partitioned into n^* so-called *maximal subchains* $\{R^{*\alpha} \mid \alpha = 1,\ldots,n^*\}$ such that

(1) each subchain of each maximal gain policy is contained within one of the sets $\{R^{*\alpha} \mid \alpha = 1,\ldots,n^*\}$

(2) there exists a randomized maximal gain policy which has $\{R^{*\alpha} \mid \alpha = 1,\ldots,n^*\}$ as its set of subchains.

Moreover, there is *only one* partitioning of the set R^* which satisfies the properties (1) and (2) (cf. th.3.2 of [21], part (f)).

THEOREM 3.2.

(a) $R^* = \{i \in \Omega \mid i \in R(f), \textit{ for some } f \in S_{PMG}\}$

(b) $\{f \in S_{RMG} \mid R(f) = R^*\}$ *is non-empty*

(c) *Let* $n^* = \min\{n(f) \mid f \in S_{RMG}, R(f) = R^*\}$ *and define* $S^*_{RMG} = \{f \in S_{RMG} \mid R(f) = R^*, n(f) = n^*\}$. *Then, all* $f^* \in S^*_{RMG}$ *have the same collection of subchains* $\{R^{*\alpha} \mid \alpha = 1,\ldots,n^*\}$

(d) *Any subchain of any* $f \in S_{RMG}$ *is contained within one of the sets* $\{R^{*\alpha} \mid \alpha = 1,\ldots,n^*\}$

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

REMARK. Let $f^1,\ldots,f^M$ be an enumeration of the pure maximal gain policies (in S_{PMG}). We note that the partitioning of R^* into the sets $\{R^{*\alpha} \mid \alpha = 1,\ldots,n^*\}$ may be obtained by determining the equivalence classes, generated by the relation ~, as defined in (3.4) with $C(f^1,f^2)$ replaced by:

$$C = \{c^m(f^r) \mid r = 1,\ldots,M; \quad 1 \leq m \leq n(f^r)\}. \tag{3.7}$$

Moreover, by applying the randomization procedure defined in (3.3) repeatedly, i.e. first to f^1 and f^2, and next to the resulting randomized policy and f^3, etc. we end up with a policy f^* which uses the following sets of alternatives in the states 1,...,N:

$$\{k \mid f^*_{ik} > 0\} = \begin{cases} K^*(i), & i \in R^* \\ L(i), & i \in \Omega \backslash R^* \end{cases} \tag{3.8}$$

where the sets $\{K^*(i),\ i \in R^*\}$ have the following characterizations (cf.[22], p.11 and lemma 2.2) (for any $v \in V$):

$$\begin{aligned} K^*(i) &= \{k \in K(i) \mid \text{there exists a } f \in S_{PMG}, \text{ with } i \in R(f) \\ &\qquad \text{and } f_{ik} = 1\}, \\ &= \{k \in L(i) \mid b(v)^k_i = 0,\ \textstyle\sum_{j \in R^{*\alpha}} P^k_{ij} = 1\}, \quad i \in R^{*\alpha},\ \alpha = 1,\ldots,n^*. \end{aligned} \tag{3.9}$$

In addition, it was shown that any policy f^* which satisfies (3.8) belongs to S^*_{RMG}.

We illustrate the above characterization, of the chain structure of S_{RMG}, with the help of example 1:

Note that in example 1:

$$S_{PMG} = S_P \backslash \{f \in S_P \mid f_{11} = f_{22} = 1\},$$

and

$$S_{RMG} = \{f \in S_R \mid f_{11}.f_{22} = 0\}.$$

Observe that $C = \{\{1\},\{2\},\{3,4\},\{4,5\}\}$ denotes the set of subchains of the maximal gain policies, and conclude that

$$R^* = \{1,2,3,4,5\} \text{ with } n^* = 3, \quad \text{and } R^{*1} = \{1\},\ R^{*2} = \{2\},\ R^{*3} = \{3,4,5\}.$$

In addition,

$$K^*(1) = \{2\};\ K^*(2) = \{1\};\ K^*(3) = \{1\};\ K^*(4) = \{1,2\} \text{ and } K^*(5) = \{1\}.$$

We finally remark that *randomization* plays the indispensible role of coalescing subchains: there is no pure policy which has R^* as its set of subchains and there is no pure policy which has $\{3,4,5\}$ in *one* subchain.

Next, th.3.3 below describes the *periodicity* structure of the set of maximal gain policies; the theorem was proven in [22], th.3.2 and it shows inter alia that the partitioning of R^* into the class of so-called *maximal subchains* $\{R^{*\alpha} \mid \alpha = 1,\ldots,n^*\}$, may be pursued so as to obtain the class of so-called maximal cyclical moving subsets (c.m.s.) $\{R^{*\alpha,t} \mid \alpha = 1,\ldots,n^*;\ t = 1,\ldots,d(\alpha)\}$ such that

(1) each subchain of each maximal gain policy is contained within one of the sets $\{R^{*\alpha} \mid \alpha = 1,\ldots,n^*\}$

(2) each c.m.s. of each maximal gain policy is contained within one of the sets $\{R^{*\alpha,t} \mid t = 1,\ldots,d(\alpha),\ \alpha = 1,\ldots,n^*\}$

(3) there exists a policy $f \in S_{RMG}$, which has $\{R^{*\alpha,t} \mid t = 1,\ldots,d(\alpha);\ \alpha = 1,\ldots,n^*\}$ as its set of c.m.s.

In addition there is only one partition of R^* which satisfies the properties (2) and (3) (cf.th.3.2, part (k) in [22]).

We first define:

$$(3.10) \qquad d(\alpha) = \min\{d^m(f) \mid f \in S_{RMG},\ 1 \le m \le n(f),\ C^m(f) \subseteq R^{*\alpha}\};\ \alpha = 1,\ldots,n^*$$

(3.11) $\quad d_i = \min\{d^m(f) \mid f \in S_{RMG},\ 1 \le m \le n(f),\ i \in C^m(f)\},\ i \in R^*$

i.e. $d(\alpha)$ [d_i] denotes the minimum of the periods of the subchains of the maximal gain policies that lie within $R^{*\alpha}$ [that contain the state i].

THEOREM 3.3. (Periodicity structure)

Let f^* *be defined as in* (3.8); *and recall that* $f^* \in S^*_{RMG}$:

(a) $d^\alpha(f^*) = d(\alpha),\ \alpha = 1,\ldots,n^*$.

(b) *Fix* $\alpha \in \{1,\ldots,n^*\}$. *Let* $h \in S_{RMG}$ *and* $C^m(h) \subseteq R^{*\alpha}$. *Then* $d^m(h)$ *is a multiple of* $d(\alpha)$.

(c) $d(\alpha) = \text{g.c.d.}\ \{d^m(f) \mid f \in S_{PMG},\ 1 \le m \le n(f),\ C^m(f) \subseteq R^{*\alpha}\},\ \alpha = 1,\ldots,n^*$.

(d) $d_i = d(\alpha)$ *for all* $i \in R^{*\alpha},\ \alpha = 1,\ldots,n^*$.

(e) *The set* $S^{**}_{RMG} = \{f \in S^*_{RMG} \mid d^\alpha(f) = d(\alpha),\ \alpha = 1,\ldots,n^*\}$ *is non-empty*.

(f) *All* $f \in S^*_{RMG}$ *have the same collection of c.m.s.* $\{R^{*\alpha,t} \mid \alpha = 1,\ldots,n^*;\ t = 1,\ldots,d(\alpha)\}$. □

Part (a) of the above theorem shows that the minimal periods $d(\alpha)$, $\alpha = 1,\ldots,n^*$ and d_i, $i \in R^*$ are attained simultaneously by any policy f^* that satisfies (3.8). The policy that was constructed by a repeated application of the randomization procedure in (3.3) to the set S_{PMG} (remark 2), and which was shown to belong to the set S^*_{RMG}, has as a consequence the additional property of minimizing the periods of each of the states in R^*.

The intuitive foundation of this result is of course contained in part (b) of lemma 3.1. The fact that for all $\alpha = 1,\ldots,n^*$ $d(\alpha)$ may be calculated as the *greatest* common divisor of the periods of the subchains of the pure maximal gain policies that are contained within $R^{*\alpha}$, (cf. part (a)) is remarkable in view of remark 1.

For an illustration of this periodicity structure, we refer to example 1 in [22].

4. THE SOLUTION SET V

In this section, we give a number of properties of the solution set to the optimality equation (1.7) which are needed in order to characterize the asymptotic behaviour of the sequence $\{v(n)\}_{n=1}^{\infty}$.

For a complete description of the properties of V, in the more general context of Markov Renewal Programs, we have to refer to [21].

THEOREM 4.1. (Basic Properties of V)

(a) V *is closed and unbounded as* $v \in V$ *implies* $v + a_1\underline{1} + a_2 g^* \in V$ *for any scalars* a_1, a_2 (*where* $\underline{1}$ *is the* N-*vector of ones*).

(b) (cf.[11],[12], [2]), $v \in V$, *if and only if*

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

If addition, if $v \in V$, *then a policy* $f \in S_{PMG}$ *achieves all* N *maxima in* (4.1) *if and only if it achieves the* 2N *maxima in* (1.6) *and* (1.7).

Using (4.1) it may be verified that the difference between two solutions $v^0, v^1 \in V$ of the optimality equation (1.7) is a constant (say y_α) on each of the sets $R^{*\alpha}$, $\alpha = 1,\ldots,n^*$.

THEOREM 4.2. *Let* $v \in V$. *The following statements are equivalent*

(a) $v + x \in V$

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

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

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

$$x_i = \begin{cases} y_\alpha & , \quad i \in R^{*\alpha} \\ \max_{f \in S_{PMG}} \left[Z(f)b(v,f)_i + \sum_{\beta=1}^{n^*} \left(\sum_{j \in R^{*\beta}} \Pi(f)_{ij} \right) y_\beta \right], & i \in \Omega \backslash R^* \end{cases} \tag{4.2}$$

$$y_\alpha \geq Z(f)b(v,f)_i + \sum_{\beta=1}^{n^*} \left(\sum_{j \in R^{*\beta}} \Pi(f)_{ij} \right) y_\beta, \quad \alpha = 1,\ldots,n^*; \tag{4.3}$$

$$i \in R^{*\alpha}, \ f \in S_{PMG}.$$

Fix $v^0 \in V$. The above theorem shows that any particular solution $v \in V$, is specified by choosing n^* parameters $(y_1,\ldots,y_{n^*})$. In other words, by sweeping out all permitted combinations of n^*-tuples $(y_1,\ldots,y_{n^*})$ we sweep out all vectors v in V. Fix $v \in V$. Define the set of allowed parameters

$$Y(v) = \{y \in E^{n^*} \mid y \text{ satisfies (4.3)}\}. \tag{4.4}$$

We note that Y(v) is a closed, convex and unbounded polyhedral set containing $y = 0$.

After thorough analysis, we establish in addition that the set Y(v)

has an interior. This implies that the parameters $(y_1,\ldots,y_{n^*})$ may be selected within a n^*-*dimensional* set, *such that* V *is a* n^*-*dimensional set as well.*

Note as a consequence that the n^* parameters $(y_1,\ldots,y_{n^*})$ may be chosen independently over some (finite) region, and that V has n^* degrees of freedom some of which may only be *locally* independent.

5. THE ASYMPTOTIC BEHAVIOUR OF v(n)

In this section we analyze the asymptotic behaviour of the sequence $\{v(n)\}_{n=1}^{\infty}$.

First it was pointed out by BROWN [3] that v(n) grows linearly with n, i.e.

(5.1) $\quad v(n) - ng^*$ is bounded.

Next define d^* as the *least common multiple* (l.c.m.) of the integers $\{d(\alpha) \mid \alpha = 1,\ldots,n^*\}$. As one of our major results we established that certain subsequences of the type $\{v(nJ+r)-(nJ+r)g^*\}_{n=1}^{\infty}$ converge, as well as the necessary and sufficient condition for J,r to guarantee convergence for *every* possible choice of the scrap-value vector $v(0) \in E^n$:

THEOREM 5.1. (cf.th.5.4 part (b) in [22]). *Fix* $J \geq 1$ *and* $r = 0,\ldots,J-1$: $\lim_{n\to\infty} v(nJ+r) - (nJ+r)g^*$ *exists for all* $v(0) \in E^N$, *if and only if* J *is a multiple of* d^*. □

First the authors of [22] obtained convergence of the subsequences $\{v(nd+r)-(nd+r)g^*\}_{n=1}^{\infty}$ for all $r = 1,\ldots,d$ on any subchain C of a policy $f \in S_{RMG}$, where d represents the period of P(f) on C, i.e.

(5.2) If $f \in S_{RMG}$ and C is a subchain of P(f), with period d, then

$$\lim_{n\to\infty} v(nd+r)_i-(nd+r)g_i^* \text{ exists for all } i \in C;\ r = 1,\ldots,d$$

(5.2) was also established in LANERY [11], proposition 7.

Next, fix $f^* \in S_{RMG}^{**}$ and recall that f^* has $\{R^{*\alpha} \mid \alpha = 1,\ldots,n^*\}$ as its set of subchains, with $d(\alpha)$ denoting the period of $P(f^*)$ on $R^{*\alpha}, \alpha = 1,\ldots,n^*$. Hence, applying (5.2) to f^*, we obtain:

(5.3) $\lim_{n\to\infty} v(nd(\alpha)+r)_i - (nd(\alpha)+r)g_i^*$ exists, for all $i \in R^{*\alpha}$;

$r = 1,\ldots,d(\alpha)$; $\alpha = 1,\ldots,n^*$.

Using the definition of d^* we conclude that

(5.4) $\lim_{n\to\infty} v(nd^*+r)_i - (nd^*+r)g_i^*$ exists for all $i \in R^*$; $r = 1,\ldots,d^*$.

As far as the *sufficiency* part of th.5.1. is concerned, this leaves us with the need to extend the convergence result in (5.4) from R^* to Ω. In fact this extension constitutes the *hard* part of the proof; a first attempt was made by LANERY ([11],p.23-p.52); however the proof in [11] is lengthy and the arguments as stated seem incomplete or incorrect (cf. also note 1 in [22]).

The method of proof, as given in [22] involves both the characterizations with respect to the chain structure, the periodicity structure and the solution set V of the optimality equation (1.7), as described in sections 3 and 4.

In addition, our approach makes an essential use of the so-called "d^*-step" MDP, where for any $J \geq 1$, the "J-step" MDP is related in the following way to Q^J, the J-fold application of the operator Q:
Note that

(5.5) $$Q^J x_i = \max_{\xi\in\tilde{K}(i)} \{\tilde{q}_i^{\xi} + \sum_j \tilde{P}_{ij}^{\xi} x_j\}, \text{ where}$$

$$\tilde{K}(i) = \{(f^1,\ldots,f^J) \mid f^1,\ldots,f^J \in S_P\}$$

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

$$\tilde{P}_{ij}^{\xi} = P(f^1)\ldots P(f^J)_{ij};\ 1 \leq i,j \leq N \text{ and } \xi = (f^1,\ldots,f^J) \in \tilde{K}(i).$$

Conclude that $\tilde{Q} = Q^J$ may be interpreted as the value iteration operator in a related "J-step" MDP, denoted by a tilde, with Ω as its state space, $\tilde{K}(i)$ as the (finite) set of alternatives in state $i \in \Omega$, $\tilde{q}_i^{\xi}$ as the one-step expected reward and $\tilde{P}_{ij}^{\xi}$ as the transition probability to state j when alternative $\xi \in \tilde{K}(i)$ is chosen when entering state i.

The extension of the convergence result in (5.4) from R^* to Ω is established by exhibiting the correspondence between the chain - and periodicity

structure of the "d^*-step" MDP and the structure of the original MDP.

The latter is also essential when establishing the necessity part of th. 5.1.: Let $\tilde{V}$ denote the set of solutions to the optimality equation (1.7) in the "d^*-step" MDP.

It is verified that $\tilde{V}$ is a subset of E^N with dimension $\tilde{n} = \sum_{\alpha=1}^{n^*} d(\alpha)$ which contains V as a n^*-dimensional subset (cf. section 4). Note that when $d^* \geq 2$ at least one of the integers $d(\alpha)$, $\alpha = 1,\ldots,n^*$ must be greater than or equal to 2 such that:

$$(5.6) \qquad d^* \geq 2 \Rightarrow \tilde{n}^* = \dim \tilde{V} > n^* = \dim V.$$

We recall from section 4 that a solution $\tilde{v} \in \tilde{V}$ is determined by choosing $\tilde{n}^*$ parameters $y_1,\ldots,y_{\tilde{n}^*}$ within some $\tilde{n}^*$-dimensional polyhedral set. Finally it is shown that for some specific choice of the scrap-value vector $v(0)$ within $\tilde{V}\backslash V$ i.e. when choosing the parameters $y_1,\ldots,y_{\tilde{n}^*}$ in a special way:

$$(5.7) \qquad \lim_{n\to\infty} v(nJ+r)-(nJ+r)g^* \text{ converges only if } J \text{ is a multiple of } d^*.$$

Due to (5.1) we obtain the necessary and sufficient condition for the convergence of $\{v(n) - ng^*\}_{n=1}^{\infty}$ for all $v(0) \in E^N$ as a simple corollary:

COROLLARY 5.2. *The following four statements are (equivalent) necessary and sufficient conditions for the convergence of* $\{v(n) - ng^*\}_{n=1}^{\infty}$ *for all* $v(0) \in E^N$:

(I) $d^* = 1$

(II) *There exists an aperiodic randomized maximal gain policy* f, *with* $R(f) = R^*$

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

(IV) *For each* $\alpha \in \{1,\ldots,n^*\}$ *there exists a randomized maximal gain policy which has an aperiodic subchain within* $R^{*\alpha}$.

Observe that (I) $\Rightarrow$ (II) as a result of th.3.3. part (a); (II) $\Rightarrow$ (III) and (III) $\Rightarrow$ (IV) are immediate whereas (IV) $\Rightarrow$ (I) is immediate from the definitions of $d(\alpha)$, $\alpha = 1,\ldots,n^*$ and d^* (cf. (3.10)).

Example 2 below emphasizes the fact that the adjective "randomized" in conditions (II), (III) and (IV) cannot be replaced by (the more restrictive) "pure".

EXAMPLE 2.

N = 4; K(1) = K(3) = K(4) = {1}; K(2) = {1,2};
$P_{12}^1 = P_{34}^1 = P_{42}^1 = P_{21}^1 = P_{23}^2 = 1$; all $q_i^k = 0$, i.e.

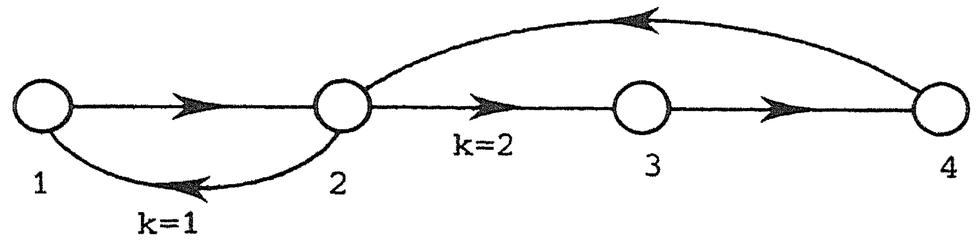


Figure 1.

Note that the two policies in S_P (and $S_{PMG} = S_P$) 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 $d^* = 1$. Note that none of the conditions (II), (III) and (IV) hold when replacing "randomized" by "pure" (cf. also the examples in [22])

Example 2 shows that conditions (I) - (IV) contain the possibility that all of the *pure* policies are *periodic*; on the other hand, the existence of an aperiodic maximal gain policy f is only sufficient for the convergence of $\{v(n) - ng^*\}_{n=1}^{\infty}$ for all $v(0) \in E^N$, when this policy has a subchain in each one of the sets $R^{*\alpha}$ ($\alpha = 1,\ldots,n^*$), e.g., when $R(f) = R^*$.

We conclude this section by enumerating a number of conditions that are sufficient for the existence of $\lim_{n\to\infty} v(n) - ng^*$ for *all* possible choices of $v(0) \in E^N$.
We have seen that for arbitrary $J \geq 1$ and some *fixed* v(o) the sequences $\{v(nJ+r)-(nJ+r)g^*\}_{n=1}^{\infty}$ may fail to converge for some (or all) $i \in \Omega$ and for some (or all) $r \in \{0,1,\ldots,J-1\}$. We refer to section 5 of [22] for an investigation of the various ways in which the convergence of these sequences interdepends.

THEOREM 5.3. (cf. th.5.5 of [22])
The following conditions are sufficient for the existence of $\lim_{n\to\infty} v(n)-ng^*$ *for all* $v(0) \in E^N$:

(I) *All of the transition probabilities are strictly positive:*
$P_{ij}^k > 0$ *for all* $i,j \in \Omega$ *and* $k \in K(i)$ (cf. BELLMAN [2], BROWN [3])

(II) *For all* $v(0) \in E^N$ *there exists an aperiodic* $f \in S_P$ *and an integer* n_0 *such that*

$v(n+1) = q(f)+P(f)v(n)$, *for all* $n \geq n_0$ (cf. MORTON [16])

(III) *There exists a state* s *and an integer* $\nu \geq 1$, *such that* $P(f^1)\ldots P(f^\nu)_{is} > 0$ *for all* $f^1, f^2, \ldots, f^\nu \in S_P$; $i \in \Omega$ (cf. WHITE [25]).

(IV) *Every* $f \in S_P$ *is aperiodic* (cf. SCHWEITZER [18] & [20]).

(V) *Every* $f \in S_{PMG}$ *is aperiodic* (cf. SCHWEITZER [18] & [20]).

(VI) *For each* $i \in R^*$ *there exists a pure maximal gain policy* f, *such that state* i *is recurrent and aperiodic for* P(f).

(VII) *Every pure maximal gain policy has a unichained* tpm *and at least one of them is aperiodic.*

6. SOME IMPLICATIONS OF THE ASYMPTOTIC BEHAVIOUR OF THE SEQUENCE $\{v(n)\}_{n=1}^{\infty}$.

In this section we briefly enumerate a number of topics on which the asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$ has a decisive impact:

(a) The properties of optimal or ε-optimal strategies in MDP's with a *finite* planning horizon.

(b) The use of the value-iteration method for locating maximal-gain policies; as pointed out before, this is especially important when the state space is large in which case *exact* solution methods like the Policy Iteration Algorithm (cf. [9]) and Linear Programming approaches (cf. [4]) become infeasible.

(c) The tightness of lower and upper bounds for the maximal gain rate g^* as developed by ODONI [17] and HASTINGS [7].

We first need the following notation:
A (Markov)-strategy $\pi = (..,f^n,\ldots,f^1)$ is an (infinite) sequence of policies $f \in S_P$; $n = 1,2,\ldots$. Applying strategy π to the n-stage model (for some $n \geq 1$) means using action $f^{\ell}(i)$ when the system is in state i, and when there are ℓ periods to go ($1 \leq \ell \leq n$). Fix a scrap-value vector $v(0) \in E^N$ and for any strategy π, let $v(n;\pi)$ denote the vector the i-th component of which denotes the total expected reward in the n-stage model when starting in state i and when applying strategy π.

A strategy π is called *optimal* if

$$(6.1) \qquad v(n;\pi)_i = v(n)_i \quad \text{for all } i \in \Omega \text{ and } n = 1,2,\ldots$$

and for any $\varepsilon > 0$ a strategy π is called ε-*optimal* if

$$(6.2) \qquad v(n;\pi)_i \geq v(n)_i - \varepsilon \quad \text{for all } i \in \Omega \text{ and } n = 1,2,\ldots .$$

One easily verifies that a strategy $\pi = (\ldots,f^n,\ldots,f^1)$ is optimal if and only if for all $\ell = 1,2,\ldots$ f^ℓ attains the N maxima in the value-iteration equation (0.1) for $n = \ell-1$. Now it follows from the multichain generalization of ODONI [17] that any policy achieving the maxima in (0.1) for large n, is maximal gain provided $v^* = \lim_{n\to\infty} v(n)-ng^*$ exists. More specifically we have:

(6.3) If $v^* = \lim_{n\to\infty} v(n)-ng^*$ exists,

then there exists an integer n_0 such that for all $n \geq n_0$:
$v(n+1) = q(f)+P(f)v(n) \Rightarrow f \in S_p(v^*)$

where for any $v \in V$, $S_p(v) = X_i\{k \in L(i) \,|\, b(v)_i^k = 0\} \subseteq S_{PMG}$ the last inclusion following from lemma 3.1. In case $\{v(n)-ng^*\}_{n=1}^\infty$ converges this implies that optimal strategies only use maximal gain policies, possibly with the exception of a *finite* number of final stages, and it follows that this is guaranteed for every $v(0) \in E^N$, if and only if $d^* = 1$.

Conversely it follows from example 4 in LANERY [11] that in case $\{v(n)-ng^*\}_{n=1}^\infty$ *fails to converge, non-maximal gain* policies may appear infinitely often in the sequence $\pi = (\ldots,f^n,\ldots,f^1)$ of an optimal strategy π. In [6] an example was even built where *every* optimal strategy uses *exclusively non-maximal gain* policies. For a more detailed investigation of the properties of optimal strategies both in the case of convergence and of oscillation of $\{v(n)-ng^*\}_{n=1}^\infty$ we refer to [6].

A strategy π is called asymptotically stationary if it uses the same policy f, at each stage of the problem, with the possible exception of a finite number of final stages. Example 1 in BATHER [1] shows that in general there is no policy convergence, i.e. in general an asymptotically stationary optimal policy may fail to exist.

As a much stronger result the same example, which has every policy unichained and aperiodic, shows that in general no optimal strategv $\pi = (\ldots,f^n,\ldots,f^1)$ exists which is *asymptotically J-periodic* for some $J \geq 1$, i.e. for which a J-tuple of policies $(f^{*J},\ldots,f^{*1})$ and an integer $n_0 \geq 1$, exists such that

(6.4) $f^{nJ+r} = f^{*r}$ for all $n \geq n_0$, $r = 1,\ldots,J$.

Bather's example thus falsifies the conjecture in BROWN [3] of the same tenor.

However, it has been proven in [6] that in all generality and for all $\varepsilon > 0$, a ε-optimal strategy exists which is asymptotically J-periodic for some $J = 1,2,\ldots$. To be more specific we recall the following result from [22], th. 5.8:

LEMMA 6.1. *Fix a scrap value vector* $x \in E^N$. *There exists an integer* $J^0(x) \geq 1$ *such that* $\lim_{n\to\infty} v(nJ+r)-(nJ+r)g^*$ *exists if and only if* J *is a multiple of* $J^0(x)$. *Moreover we have*

$$d^* = \max_{x\in E^N} J^0(x) = \text{l.c.m.}\ \{J^0(x) \mid x \in E^N\}. \quad \square$$

Using lemma 6.1, one obtains (cf. [6]) that for all $\varepsilon > 0$ ε-optimal strategies can be constructed with $J^0(v(0))$ as the asymptotic period. In particular we can conclude that for all $\varepsilon > 0$, asymptotically stationary ε-optimal strategies exist whenever $\lim_{n\to\infty} v(n)-ng^* = v^*$ exists and in [6] it was verified that in this case $S_p(v^*)$ is the set of policies that may be used in the initially stationary part of a ε-optimal strategy. In addition there exists for *every* scrap value vector $x \in E^N$ an asymptotically d^*-periodic ε-optimal strategy (for every $\varepsilon > 0$). Analogous results were obtained in LEMBERSKY [13], [14] and LEMBERSKY & OTT [15] for the *continuous time* Markov Decision Problems where no periodicity problems arise.

For the consequence of the asymptotic behaviour of $\{v(n)\}_{n=1}^{\infty}$ with respect to the working of several successive approximation schemes we refer to Part III of this paper. Finally we recall the bounds on g^* as obtained by a straightforward generalization of ODONI [17] & HASTINGS [7]:

$$(6.5) \qquad [v(n+1)-v(n)]_{min} \leq g(f^n)_i \leq g_i^* \leq [v(n+1)-v(n)]_{max},\ n = 1,2,\ldots$$

where f^{ℓ}, $\ell = 1,2,\ldots$ is any policy which attains the N maxima in the value-iteration equation with $n = \ell-1$. It is quite obvious that the sequences of lower and upper bounds converge (to g^*_{min} and g^*_{max} resp.) if and only if $\lim_{n\to\infty} v(n)-ng^*$ exists and will be oscillating otherwise. This implies that if the maximal gain rate is independent of the initial state of the system i.e. if $g^* = \langle g^* \rangle \underline{1}$, the lower and upper bounds will ultimately come within arbitrary precision of $\langle g^* \rangle$. (6.5) can easily be extended to:

$$(6.6) \qquad \frac{1}{J}\ v(n+J)-v(n)\ _{min} \leq g_i^* \leq \frac{1}{J}\,[v(n+J)-v(n)]_{max},\ n = 1,2,\ldots$$

where it follows from th.5.1 that *for all* $v(0) \in E^N$ the lower and upper bounds in (6.6) converge to g^*_{min} and g^*_{max} resp. as n tends to infinity, provided that J is a multiple of d^*.

REFERENCES

[1] BATHER, J., *Optimal decision procedures for finite Markov Chains*, Part I, II, Adv. Appl. Prob. 5 (1973), 328-339, 521-540.

[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 state space discrete time Markov Process*, Ann. Math. Stat. 36 (1965), 1279-1285.

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

[5] DERMAN, C., *Finite State Markovian Decision Process*, Academic Press, New York (1970).

[6] FEDERGRUEN, A. & P.J. SCHWEITZER, *Turnpike results in undiscounted Markov Decision Problems* (forthcoming).

[7] HASTINGS, N., *Bounds on the gain of a Markov Decision Process*, Op. Res. 19 (1971), p.240-244.

[8] HORDIJK, A., P.J. SCHWEITZER and H. TIJMS, *The asymptotic behaviour of the minimal total expected cost for the denumerable state Markov Decision Model*, J. Appl. Prob. 12 (1975). 298-305.

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

[10] KEMENY, J. & J. SNELL, *Finite Markov Chains*, Van Nostrand, Princeton (1961).

[11] LANERY, E., *Etude asymptotique des systèmes Markoviens à commande*, R.I.R.O. 1 (1967), 3-56.

[12] ———— ' *Complements à l' ètude asymptotique des systèmes Markoviens à commande*, I.R.I.A., Rocquencourt, France (1968).

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

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

[15] ———— & M. OTT, *A counterexample in continuous Markov Decision Chains* Man. Sci 21 (1974), 358-359.

[16] MORTON, T., *On the asymptotic convergence rate of cost differences for Markovian Decision Processes*, O.R. 19 (1971), 244-248.

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

[18] SCHWEITZER, P.J., *Perturbation theory and Markovian Decision Processes*, Ph.D. dissertation, MIT (1965) MIT Operations Research Center Report 15.

[19] ———— , *Perturbation theory and finite Markov Chains*, J. Appl. Prob. 5 (1968), 401-413.

[20] ———— , *A turnpike theorem for undiscounted Markovian Decision Processes*, presented at ORSA/TIMS, national meeting, May 1968.

[21] ———— , & A. FEDERGRUEN, *Functional equations of undiscounted Markov Renewal Programming*, Math. Center Report BW 71/77 (1976) (to appear in Math. of O.R.).

[22] ———— & ———— , *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.).

[23] ———— & ———— , *Geometric convergence of value-iteration in multichain Markov Decision Problems*, (unpublished manuscript) (1976).

[24] TIJMS, H., *On Dynamic Programming with arbitrary state space, compact action space and the average return criterion*, Math. Center Report BW 55/75 (1975)

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

VALUE-ITERATION IN UNDISCOUNTED MARKOV DECISION PROBLEMS
PART II: GEOMETRIC CONVERGENCE

A.Federgrün
Mathematical Centre, Amsterdam, The Netherlands

P.J.Schweitzer
IBM Thomas J. Watson Research Center, Yorktown Heights, USA

H.C.Tijms
Mathematical Centre / Free University, Amsterdam, The Netherlands

0. SUMMARY

In part I, a survey was given of what has become known with respect to both necessary and sufficient conditions for convergence of

$$(0.1) \qquad \{v(n) - ng^*\}_{n=1}^{\infty}$$

for every possible choice of the scrap value vector $v(0) \in E^N$.

In (0.1), $v(n)_i$ represents the total expected maximal reward for a planning horizon of n epochs, when starting in state i and given an amount $v(0)_j$ is obtained when ending up in state j. Moreover, g^* denotes the maximal gain rate vector and it is known from BROWN [2] that the sequence in (0.1) is always bounded. In this context, th.5.1 of part I represents the main result by stating the existence of an integer $d^* \geq 1$ such that

$$(0.2) \qquad \lim_{n\to\infty} v(nJ+r)-(nJ+r)g^* \quad \text{exists for all } v(0) \in E^N$$

if and only if J is a multiple of d^*.

In addition a characterization of d^* was given in terms of the chain- and periodicity structure of the MDP, from which the necessary and sufficient condition for $d^* = 1$, i.e. for *global* convergence was obtained as a corollary (cf. corollary 5.2).

Whereas part I settles the issue if one demands global convergence, i.e. convergence of $\{v(n)-ng^*\}_{n=1}^{\infty}$ for every $v(0) \in E^N$, one should observe that convergence always occurs for the special choice $v(0) = v + Rg^*$, where $v \in V$ and R is large (cf. section 4 and lemma 2.2 in [8]). In other words, there always exists a non-empty (closed) subset $W \subseteq E^N$ of scrap-value vectors for which $\{v(n)-ng^*\}_{n=1}^{\infty}$ converges.

In this part we turn to the topic of the *rate* of convergence. The notation is identical to the one used in part I. As our principal result we obtained in [8] the fact that if a subsequence of the type $\{v(nJ+r) - (nJ+r)g^*\}_{n=1}^{\infty}$ converges to a limit v^* (for some $J \geq 1$ and $r = 0,\ldots,J-1$) then the approach to the limit v^* is *geometric*, i.e. there exist numbers $K(v(0)) > 0$ and $0 \leq \lambda < 1$ such that

$$(0.3) \qquad \left|v(nJ+r)-(nJ+r)g^* - v^*\right| \leq K\lambda^n, \qquad n = 1,2,\ldots$$

As a consequence various successive approximation methods which are based on the value-iteration scheme (0.1) of part I exhibit a geometric rate of convergence as well (cf. part III). We observe that this generalization of

(1) what is known to be the case in a simple Markov Process, i.e. in a MDP with single policy (cf. [9]), and
(2) White's result [12]

holds in all generality with no restrictions imposed on either the chain-, periodicity- or reward structure of the problem. In addition, the result is to some extent surprising, since the value-iteration operator Q, defined by

$$(0.4) \qquad Qx_i = \max_{k \in K(i)} \{q_i^k + \Sigma_j P_{ij}^k x_j\}, \qquad i \in \Omega;\ x \in E^N$$

(which takes $v(n)$ into $v(n+1)$) is not a (J step) contraction mapping for any $J = 1,2,\ldots$ on E^N (cf. DENARDO [3] and [4], section 1): nor is there in general an obvious way of reducing it to such a mapping on some subspace of E^N. To be more specific, note that the Q-operator has the properties:

$$(0.5) \qquad Q(x+c\underline{1}) = Qx + c\underline{1};\ x \in E^N \quad \text{and all scalars } c$$

which excludes the possibility of Q having a unique fixed point and hence

of Q being a (J-step) contraction mapping (for some $J \geq 1$) as such (cf. [4] section 1). Even when considering the quasi-norm (cf. BATHER [1])

$$(0.6) \qquad \|x\|_d = x_{max} - x_{min}, \qquad x \in E^N$$

the usefulness of which is suggested by (0.5) it can be shown (cf. [4]) that Q is (J-step) contracting with respect to this quasi-norm only under very restrictive conditions on the chain-and periodicity structure of the problem. Example 1 below indicates e.g. that the uniqueness of the fixed point $v \in V$ (cf. part I) up to a multiple of $\underline{1}$, i.e. the uniqueness of $v \in V$ in $\| \ \|_d$- norm, which is equivalent to

(H1) $n^* = 1$, or the existence of a randomized maximal gain policy which has R^* as its single subchain

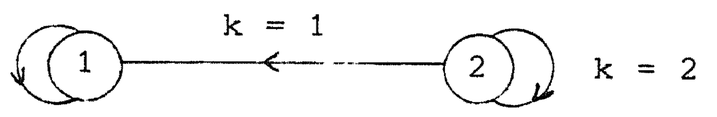
in combination with global convergence of $\{Q^n x - ng^*\}_{n=1}^{\infty}$ for all $x \in E^N$ which is equivalent to (cf. cor. 5.2 in part I)

(H2) $d^* = 1$, or the existence of an *aperiodic* randomized maximal gain policy with $R(f) = R^*$

is in itself an insufficient condition.

EXAMPLE 1.

$$\Omega = \{1,2\};\ K(1) = \{1\};\ K(2) = \{1,2\};\ q_1^1 = q_2^1 = 0;\ q_2^2 = -1;\ p_{ij}^k = \delta_{jk}.$$



Note that $g^* = 0$ and $V = \{c\underline{1} \mid c \in E^1\}$ and that every policy is aperiodic which guarantees (0.8). Take $x = [0,X]$ and $y = 0$. Verify that $Q^n x = [0, \max(X-n,0)]$ and $Q^n y = 0$ and conclude that for all $n = 1,2,\ldots$

$$1 = \sup \left\{ \frac{\|Q^n u - Q^n v\|_d}{\|u-v\|_d} \;\middle|\; \|u-v\|_d = 0 \right\} \geq$$

$$\lim_{X\to\infty} \frac{\|Q^n x - Q^n y\|_d}{\|x-y\|_d} = \lim_{X\to\infty} \frac{\max(X-n,0)}{X} = 1.$$

Thus for no $n = 1,2,\ldots$ is Q (n-step) contracting with respect to the $\| \ \|_d$-norm. For necessary and sufficient conditions for Q to reduce to a contraction mapping, we refer to [4].

The geometric convergence result in (0.3) is obtained by analyzing the

evolution of the Q-operator in $\{Q^n x\}_{n=1}^{\infty}$ for any $x \in W$. In fact when trying to establish (0.3) it suffices to consider the case $J = 1$. For when $J \geq 2$, the same analysis, applied to the J-step MDP as defined in section 5 of part I, establishes the geometric convergence result in (0.3) for all $r = 0,\ldots,J-1$.

In section 1, we show for all $x \in W$, convergence of $\{Q^n x - ng\}_{n=1}^{\infty}$ occurs in three phases, and we discuss the behaviour of the Q-operator during the first phase.

For all $x \in W$, let $L(x) = \lim_{n\to\infty} \{Q^n x - ng^*\}_{n=1}^{\infty}$. In the second and third phase $\|Q^n x - ng^* - L(x)\|_d$ is monotonically non-increasing and in section 2 we point out that the number of steps needed for strict contraction, i.e. for a strict decreasing of $\|Q^n x - ng^* - L(x)\|_d$ is bounded in $x \in W$.

Next we explain how this was used in [8] to establish the geometric convergence result in (0.3).

In addition we give a *sharp* upperbound for the convergence rate which is independent of the starting point $x \in W$. As a contrast, a *uniform* (n-step) contraction factor (i.e. a n-step contraction factor which is independent of $x \in W$) does not need to exist for any $n = 1,2,\ldots$ and we recall the necessary and sufficient condition for the existence of such a uniform contraction factor for MDP's satisfying (0.7).

Finally we discuss upperbounds for the number of steps needed for contraction.

1. THE EVOLUTION OF THE Q-OPERATOR

First of all we recall from lemma 2.2 in [8] or from BROWN [2] that for all $x \in E^N$ there exists an integer $n_1(x)$ such that

$$(1.1) \qquad Q^n x = T(X^{n-1} x) = T^{n-n_1}(Q^{n_1} x) \qquad \text{for all } n \geq n_1(x)$$

where the T-operator is defined by:

$$(1.2) \qquad Tx_i = \max_{k \in L(i)} \{q_i^k + \Sigma_j P_{ij}^k x_j\}; \qquad x \in E^N.$$

This is due to the fact that, after a finite number of iterations, only alternatives $k \in L(i)$ attain the maximum in the value-iteration equation

(0.1) in part I. Note that the T-operator has the additional properties:

$$(1.3) \qquad T(x+cg^*) = Tx + cg^* \quad \text{for all } x \in E^N; \qquad c \in E^1$$

and

$$(1.4) \qquad \|Tx-g^*-v\|_d = \|Tx-Tv\|_d \leq \|x-v\|_d \quad \text{for all } x \in E^N \text{ and all } v \in V.$$

In other words, after $n_1(x)$ iterations the "distance" between $Q^n x - ng^*$ and any $v \in V$, as measured by the $\| \ \|_d$-norm is monotonically *non-increasing*. Next, define for $x \in W$:

$$e(n,x) = Q^n x - ng^* - L(x)$$

and note that $\{e(n,x)\}_{n=1}^{\infty}$ satisfies the recursion equation:

$$(1.5) \qquad e(n+1,x)_i = \max_{k \in L(i)} \{b(L(x))_i^k + \Sigma_j P_{ij}^k e(n,x)_j\}, \quad n \geq n_1(x).$$

Since $\lim_{n\to\infty} e(n,x) = 0$ for all $x \in W$, it follows that after a still larger number of (say after $n_2(x)$) iterations, only alternatives $k \in L(i)$ attain the maximum in the value-iteration equation (0.1) of part I, for which $b(L(x))_i^k = 0$. More specifically, for any $v \in V$ let

$$(1.6) \qquad \delta(v) = \min\{|b(v)_i^k| \mid i \in \Omega, k \in L(i), b(v)_i^k < 0\}.$$

Next, for any $x \in W$, let $n_2(x) = \inf\{n \mid n \geq n_1(x); \|e(m,x)\|_d < \delta(L(x))$ for all $m \geq n\} < \infty$. Then for all $x \in W$ and $n \geq n_2(x)$:

$$(1.7) \qquad e(n+1,x) = U(L(x))e(n,x)$$

where for any $v \in V$, the $U(v)$ - operator is defined by

$$(1.8) \qquad U(v)x_i = \max_{k \in L(i,v)} [\Sigma_j P_{ij}^k x_j], \qquad i \in \Omega; \ x \in E^N$$

with

$$L(i,v) = \{k \in L(i) \mid b(v)_i^k = 0\}, \qquad i \in \Omega.$$

Observe that in spite of V being an infinite subset of E^N, only finitely many *distinct* $U(v)$-operators occur since there are only finitely many

subsets of $X_i L(i)$.

Note in addition that the U(v)-operators have all of the properties the Q- and T-operator have (property (1.3) included) but in addition the U(v)-operators have the extremely useful characteristic of bein *positively homogeneous* i.e.:

$$(1.9) \qquad U(v)[\,ax\,] = aU(v)x \quad \text{for all } x \in E^N \quad \text{and } a \geq 0.$$

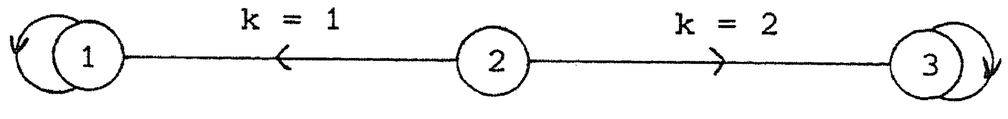
As a consequence the convergence of $\{Q^n x - ng^*\}_{n=1}^{\infty}$ for any $x \in W$ occurs in three phases. The first $n_1(x)$ iterations constitute the *first* phase and the second phase terminates after the $n_2(x)$-th iteration, and is followed by the third phase from there on.

We conclude this section by a short description of the behaviour of the Q-operator during the first phase. We first observe that this phase is void if, $K(i) = L(i)$ for all $i \in \Omega$, which is e.g. the case when $g_i^* = \langle g^* \rangle$, $i \in \Omega$, i.e. when the maximal gain rate is independent of the initial state of the system. On the other hand, $n_1(x)$ may be unbounded in $x \in E^N$ or $x \in W$. In fact in the worst case the length of the first phase may be linear in $\|x\|_d$ as is proven in [8], th.3.1. This is why the first phase is said to have a *finite though linear* type of convergence.
The following example illustrates this:

EXAMPLE 2.

$\Omega = \{1,2,3\}$; $K(1) = K(3) = \{1\}$; $K(2) = \{1,2\}$; $q_1^1 = q_2^1 = q_2^2 = 0$; $q_3^1 = 1$

$P_{11}^1 = P_{21}^1 = P_{23}^2 = P_{33}^1 = 1$.



Note that $g^* = (0,0,-1)$ and that $L(2) = \{1\}$

Let $x = [0,0,X]$ with $X >> 1$ and verify that $Q^n x = [0,\ \max(0,X-n+1),X-n]$ such that $n_1(x) = \|x\|_d = |x| = X$.

The behaviour of $\|Q^n x - ng^* - L(x)\|_d = \|e(n,x)\|_d$ during the first phase may be very capricious. E.g. $\{\|e(n,x)\|_d\}_{n=1}^{\infty}$ may be alternatingly increasing and decreasing such that the first phase is not necessarily terminated as soon as $[Q^n x - ng^*]$ starts coming closer in $\|\ \|_d$-norm to the limit $L(x)$ (cf. example 1 in [8]). In the second phase the Q-operator essentially reduces to the T-operator. Let $\tilde{W} = \{x \in E^N \mid \tilde{L}(x) = \lim_{n\to\infty} T^n x - ng^* \text{ exists}\}$

and note that $V \subseteq \tilde{W}$. In analogy to $e(n,x)$ define for $n = 1,2,\ldots$ and $x \in \tilde{W}$:

$$\tilde{e}(n,x) = T^n x - ng^* - \tilde{L}(x). \tag{1.10}$$

It follows from (1.1) that for all $x \in W$:

$$\begin{aligned} \tilde{L}(Q^{n_1}x) &= \lim_{n\to\infty} T^n(Q^{n_1}x) - ng^* = \\ &\lim_{n\to\infty} Q^{n+n_1(x)}x - (n+n_1(x))g^* + n_1(x)g^* = L(x) + n_1(x)g^*. \end{aligned} \tag{1.11}$$

In other words for all $x \in W$, $Q^{n_1}x \in \tilde{W}$. As a consequence studying the convergence of $\{Q^n x - ng^*\}_{n=1}^{\infty}$ in the second and third phase amounts to characterizing the behaviour of T on $\tilde{W}$.

2. THE SECOND AND THIRD PHASE; GEOMETRIC CONVERGENCE

First we define for all $x \in E^N$ and all $n = 1,2,\ldots$ the n-step contraction factor $f_n(x)$ by:

$$f_n(x) = \begin{cases} \dfrac{\|\tilde{e}(n,x)\|_d}{\|\tilde{e}(0,x)\|_d} = \dfrac{\|T^n x - ng^* - \tilde{L}(x)\|_d}{\|x - \tilde{L}(x)\|_d} = \dfrac{\|T^n x - T^n\tilde{L}(x)\|_d}{\|x - \tilde{L}(x)\|_d}, & \text{if } x \notin V \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

since $\|x - L(x)\|_d = 0$ can be shown to occur only if $x \in V$ and where the equality in (2.1) follows from a repeated application of (1.3).
Note that for all $x \in \tilde{W}$, and $n = 1,2,\ldots$, $f_n(x) \leq 1$ and that $\{f_n(x)\}_{n=1}^{\infty}$ is monotonically non-increasing towards 0, such that there exists an integer $M(x) \geq 1$ with:

$$f_n(x) < 1 \quad \text{for all } n \geq M(x). \tag{2.2}$$

Next the *key* result in the geometric convergence proof is provided by

THEOREM 1. *There exists an integer* M^* *such that for all* $x \in \tilde{W}$:

$$f_{M^*}(x) < 1. \quad \square \tag{2.3}$$

Thus, th.1 expresses that M(x) the number of steps needed for contraction is bounded in $x \in \tilde{W}$.

For each $m = 1,2,\ldots$ and $x \in \tilde{W}$, let

$$(2.4) \qquad h_m(x) = \sup_{n=1,2,\ldots} f_m(T^n x - ng^*) \leq 1$$

where the inequality follows from (2.3). The second part of the geometric convergence proof consists of showing that for all $x \in \tilde{W}$:

$$(2.5) \qquad h_{M^*}(x) < 1$$

(2.5) is obtained by a detailed analysis of the U(v)-operator appearing in the third phase of the process, and leads to:

$$(2.6) \qquad \|\tilde{e}(nM^*+r,x)\|_d \leq \|\tilde{e}(nM^*,x)\|_d \leq f_{M^*}(T^{(n-1)M^*}x-(n-1)M^*g^*)\|\tilde{e}((n-1)M^*,x)\|_d$$

$$\leq h_{M^*}(x)\|\tilde{e}((n-1)M^*,x)\|_d.$$

Finally, some further analysis leads to the main result:

THEOREM 2. (Geometric convergence)

For all $x \in W$, *there exists a number* K(x) *such that*

$$(2.7) \qquad |Q^n x - ng^* - L(x)| \leq K(x)[h_{M^*}(x)]^{[n/M^*]}$$

where [x] *indicates the largest integer less than or equal to* x.

We observe that $h_{M^*}(x)$ does *not* represent the *ultimate convergence rate* or *ultimate average contraction factor per step*, which is defined by:

$$(2.8) \qquad \begin{cases} \lim_{n\to\infty} f_n(x)^{1/n} = \lim_{n\to\infty} \left\{ \dfrac{\|\tilde{e}(n,x)\|_d}{\|\tilde{e}(0,x)\|_d} \right\}^{1/n} & \text{for } x \notin V \\ 0 & \text{for } x \notin V \,. \end{cases}$$

It can be shown that for all $x \in W$, the ultimate convergence rate may be bounded by

$$(2.9) \qquad \lambda^* \stackrel{\text{def}}{=} \max_{v\in V} \sup \left\{ \frac{\|U(v)^{M^*}y\|_d}{\|y\|_d} \;\middle|\; \lim_{n\to\infty} U(v)^n y = 0 \right\} < 1$$

Observe that on the right hand side of (2.9) the maximum is taken over a *finite* number of distinct U(v)-operators. Note in addition that in the case of a single policy this reduces to the well-known fact that the convergence rate is bounded by the subdominant eigenvalue of the associated transition probability matrix (cf. also MORTON [6], who found the same result in the special case of policy convergence, i.e. when there exists an integer $n_0(x)$ and a policy $f \in S_p$ such that:

$$(2.10) \qquad Q^n x = q(f) + P(f)Q^{n-1}x \qquad \text{for all } n \geq n_0(x).$$

Whereas the ultimate convergence rate is bounded on $\tilde{W}$, the same does not necessarily hold for the *n-step contraction factor* $f_n(x)$ whatever the choice for $n = 1,2,\ldots$. That is, we may have:

$$(2.11) \qquad \sup_{x \in \tilde{W}} f_n(x) = 1 \qquad \text{for all } n = 1,2,\ldots$$

as is illustrated by example 1.

The problem of finding conditions which in all generality are both necessary and sufficient for the existence of a uniform n-step contraction factor for *some* $n = 1,2,\ldots$, has not been solved yet. However, under (H1), the following necessary and sufficient condition was obtained in [8]:

(H3) There exists a *randomized* policy $f \in S_R$ which has $\hat{R}$ as its single subchain.

where $\hat{R} = \{i \in \Omega \mid i \in R(f),\ f \in S_p\}$.

Another topic of interest is the dependence of M^* on the size of the problem. Again, under (H1) it was shown that

$$(2.12) \qquad M^* \leq N^2 - 2N + 2.$$

The upperbound was obtained by a combinatorial proof and is sharp up to a term of O(N) (cf. example 2 in [8]). The quadratic upperbound obviously represents the worst case behaviour, and contrasts with the fact that computational experience as reported e.g. in SU and DEININGER [10] and TIJMS [11] shows that (in most cases) $M^* = 1$ or 2.

Obviously the bounds on g^* as obtained by ODONI [7] and HASTINGS [5] exhibit the same geometric rate of convergence.

With the above described geometric convergence rate, the state of the art with respect to the undiscounted model has become paralel to the discounted model, where geometric convergence of the value-iteration method follows immediately from the theory of contraction mappings.

However, as long as no upperbounds for λ^* in (2.9) can be computed it seems unlikely that

(1) bounds on $L(x)$, or
(2) bounds on the number of iterations needed to come within arbitrary precision of $L(x)$, or
(3) tests for *permanent* elimination of non-optimal actions

will be obtained.

REFERENCES

[1] BATHER, J., *Optimal decision procedures for finite Markov Chains,* Adv. in Appl. Prob. 5 (1973), p. 521-540.

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

[3] DENARDO, E., *Contraction Mappings in the theory underlying Dynamic Programming,* SIAM Review 9 (1967), p. 165-177.

[4] FEDERGRUEN, A. & P.J. SCHWEITZER & H.C. TIJMS, *Contraction Mappings underlying undiscounted Markov Decision Problems,* Math. Center Report BW 72/77 (1977, (to appear in J. Math. Anal. & Appl.).

[5] HASTINGS, N., *Bounds on the gain of a Markov Decision Process,* Op. Res. 19 (1971), p. 240-241.

[6] MORTON, T. & W. WECKER, *Discounting, Ergodicity and Convergence for Markov Decision Processes* (to appear in Man. Sc.).

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

[8] SCHWEITZER, P.J. & A. FEDERGRUEN, *Geometric Convergence of value-iteration in multichain Markov Decision Problems.*

[9] SENETA, E., *Non-ngeative matrices,* Allen & Unwin, London (1973).

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

[11] TIJMS, H., *An iterative method of approximating average cost optimal (s,S) inventory policies,* Zeitschrift für O.R. 18 (1974), p. 215-233.

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

VALUE-ITERATION IN UNDISCOUNTED MARKOV DECISION PROBLEMS

PART III: ALGORITHMS

A.Federgrün
Mathematical Centre, Amsterdam, The Netherlands

P.J.Schweitzer
IBM Thomas J. Watson Research Center, Yorktown Heights, USA

H.C.Tijms
Mathematical Centre / Free University, Amsterdam, The Netherlands

In this final part we show which successive approximation procedures can be used in order to find maximal gain policies and the maximal gain rate vector. For the schemes which are based upon pure value-iteration the convergence results obviously follow from the study of the asymptotic behaviour of the total n-stage maximal expected reward as n tends to infinity and as described in parts I and II.

In part I we observed that only in case $\{v(n)-ng^*\}_{n=1}^{\infty}$ 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 the case where $\{v(n)-ng^*\}_{n=1}^{\infty}$ may fail to converge for some $v(0) \in E^N$ i.e. whenever $d^* = 1$ is not guaranteed by the structure of the problem, the following alternatives can be used:

A) The modified value-iteration technique by HORDIJK and TIJMS [8]:
This scheme is essentially a discounted value-iteration scheme with a discountfactor β depending upon the index of the iteration stage, and tending to one as the index tends to infinity:

$$(1.1) \qquad w(n+1)_i = \max_{k \in K(i)} \{q_i^k + \beta_n \Sigma_j P_{ij}^k w(n)_j\}; \qquad i \in \Omega$$

where w(0) is a given N-vector.

The scheme can only be used when

(1.2) $\qquad g^* = \langle g^* \rangle \underline{1}.$

In this case

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

where $\{\gamma_n\}_{n=1}^{\infty}$ is obtained recursively by

$$\gamma_{n+1} = 1 + \beta_n \gamma_n \qquad \text{for } n \geq 0 \qquad \text{with } \gamma_0 = 0$$

provided that

(a) $\qquad \beta_n \beta_{n-1} \cdots \beta_1 \to 0$

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

(a) and (b) essentially express that $\{\beta_n\}_{n=1}^{\infty}$ should increase to one at a low enough rate, and a computationally tractible choice is provided by

(1.4) $\qquad \beta_n = 1 - n^{-b} \quad \text{with} \quad 0 < b \leq 1.$

The analysis of the behaviour of this scheme uses the Laurent series expansion of the total maximal discounted return vector for discountfactors that are close enough to one (cf. MILLER and VEINOTT [9]).

The scheme eventually settles upon maximal gain policies, and with the choice (1.4) it can be shown that the ultimate convergence rate is $0(n^{-b} ln n)$ which is substantially slower than the geometric convergence rate we obtained for the ordinary value-iteration scheme.
However the scheme has two very nice characteristics:

(1) convergence occurs regardless of the chain- and periodicity structure of the problem.
(2) For every starting point $w(0) \in E^N$ the scheme converges to the same limit vector w^* which has the following very important interpretation:

(1.5) $\qquad w = \max_{f \in S_{PMG}} w(f)_i = \max_{f \in S_{PMG}} Z(f)[q(f) - g^*]_i, \qquad i \in \Omega.$

That is, w^* is the *optimal bias-vector*, where the biasvector w(f) of a policy $f \in S_p$ is the *second* term in the Laurent series expansion of the total discounted return vector $V(f,\beta)$:

$$(1.6) \qquad V(f,\beta) = \frac{g(f)}{1-\beta} + w(f) + o(1-\beta), \qquad \beta \to 1$$

(cf. BLACKWELL [1] and MILLER and VEINOTT [9]).

The HORDIJK-TIJMS scheme, however, does not necessarily settle upon bias-optimal policies i.e. policies which attain the N maxima in (1.5) simultaneously.

The bounds on g^* as obtained by ODONI [11] and HASTINGS [6] for ordinary value-iteration (cf. also part I) have to be altered as follows. For $\ell = 1,2,\ldots$

$$(1.7) \qquad \min_i \{w(\ell)_i - \beta_\ell w(\ell-1)_i\} \leq g(f_\ell)_i \leq g_i^* < \max_i w(\ell)_i - \beta_\ell w(\ell-1)_i\}$$

where f_ℓ is any policy which attains the N maxima at the ℓ-th iteration stage of (1.1). Again, whenever $g_i^* = \langle g^* \rangle$, $i \in \Omega$, will the outer bounds in (1.7) converge to $\langle g^* \rangle$.

B) A second way to deal with the periodicity-problems mentioned in part I, is obtained by eliminating the periodicities using the following data-transformation (cf. SCHWEITZER [11]):

$$(1.8) \qquad \tilde{P}^k_{ij} = \tau(P^k_{ij} - \delta_{ij}) + \delta_{ij}: \quad 1 \leq i,j \leq N \text{ and } k \in K(i)$$

where $0 < \tau < 1$.

This transformation makes all of the diagonal elements of all of the tpm's strictly positive, such that in the transformed model all of the policies are aperiodic.

Moreover, the transformation turns the MDP into an *equivalent* one, in the sense that it has the same state- and policy space and that each policy has the same gain rate vector.

Due to the obtained aperiodicities, $\{v(n)-ng^*\}_{n=1}^{\infty}$ converges (geometrically fast) in the transformed model for whatever choice of $v(0) \in E^N$. In addition, the following simple relationship exists between V and $\tilde{V}$, the solution set to the optimality equation in the transformed model.

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

A second problem arises in both approaches due to the fact that the sequences generated ($\{w(n)\}_{n=1}^{\infty}$ and $\{v(n)\}_{n=1}^{\infty}$) diverge linearly with n. That is, one has to do computations with numbers that grow linearly with the number of stages needed to come within the required precision.

In case $g^* = \langle g^* \rangle 1$ the problem can be eliminated using White's procedure: e.g. in approach A) we generate

(1.10
$$\tilde{w}(n)_i = w(n)_i - w(n)_N = \max\{q_i^k + \beta_n \sum_{j=1}^{N} P_{ij}^k \tilde{w}(n-1)_j\} -$$
$$\max\{q_N^k + \beta_n \sum_{j=1}^{N} P_{Nj}^k \tilde{w}(n-1)_j\}.$$

Then

$$\tilde{w}(n) \to w^* - \langle w_N^* \rangle \underline{1}, \quad \text{and}$$

$$\max_{k \in K(i)} \{q_N^k + \beta_n \sum_j P_{Nj}^k \tilde{w}(n)_j\} \to \langle g^* \rangle.$$

In the general multichain case where (1.2) fails to hold, only approach B) needs to be considered. The only thing that comes to mind when trying to eliminate the above mentioned difficulty is the following:
Write $v(n) = ng(n) + y(n)$, with

(1.11) $$g(n) = v(n) - n(n-1)$$

(1.12) $$y(n) = nv(n-1) - (n-1)v(n).$$

Observe that the sequence $\{g(n)\}_{n=1}^{\infty}$ and $\{y(n)\}_{n=1}^{\infty}$ converge to g^* and $L(v(0))$ whenever $L(v(0)) = \lim_{n\to\infty} v(n) - ng^*$ exists. Note in addition that g(n) and y(n) can be generated from the schemes:

(1.13) $$g(n+1)_i = \max_{k \in K(i)} \{q_i^k + n \sum_j (P_{ij}^k - \delta_{ij}) g(n)_j + \sum_j (P_{ij}^k - \delta_{ij}) y(n)_j\}$$

(1.14) $$y(n+1)_i = y(n)_i + n[g(n)_i -$$
$$\max_{k \in K(i)} \{q_i^k + n \sum_j (P_{ij}^k - \delta_{ij})(y(n)_j + ng(n)_j)\}], \quad i \in \Omega.$$

By generating (1.13) and (1.14) only two *bounded* sequences of numbers have to be *stored*. Unfortunately, however, this solves our numerical difficulty only partially, since it is still necessary to do *computations* with unbounded terms when determining the right hand sides of (1.13) and (1.14).

In some cases one may be interested in obtaining (as large as possible a subset of) the entire set S_{PMG}, so as to make further selections on the basis of additional criteria.

In section 6 of part I we discussed the irregulatities that may appear in the sequences of policies generated by the value iteration method (and which are identical both in approach A) and B)). Nevertheless the following procedure may be used to get (cf. [3]):

X_i $L(i,w^*)$, when using approach A)

X_i $L(i,L(V(0))$, when using approach B).

In both approaches let $K(i,n,\varepsilon)$ be the set of actions that come within ε of attaining the maximum at the n-th stage of the iteration scheme ($i \in \Omega$; $n = 1,2,\ldots$ and $\varepsilon > 0$). By letting ε decrease to zero, as n tends to infinity, we get the existence of an integer $n_0 \geq 1$ such that for all $n \geq n_0$:

$$K(i,n,\varepsilon_n) = \begin{cases} L(i,w^*) & \text{in approach A)} \\ L(i,L(v(0))) & \text{in approach B)} \end{cases}$$

provided the sequence of positive numbers $\{\varepsilon_n\}_{n=1}^{\infty}$ is chosen to decrease to 0 at a rate which is slower than the convergence rate of the particular value-iteration scheme. That is, choose:

$$(1.15) \qquad \lim_{n\to\infty} (\ln n)^{-1} n^b \varepsilon_n \to \infty$$

in approach A) with the choice (1.4) for $\{\beta_n\}_{n=1}^{\infty}$, and

$$(1.16) \qquad \lim_{n\to\infty} \varepsilon_n \lambda^{-n} \to \infty, \qquad \text{in approach B)}.$$

To satisfy (1.15) take e.g. $\varepsilon_n = n^{-b/2}$ and to satisfy (1.16) any positive polynomial in n, may be chosen for ε_n^{-1}.

We mentioned before (cf. part II, section 2) that no tests have been derived so far for eliminating nonoptimal actions on a *permanent* basis, since for the multi-chain case no bounds on $L(v(0))$ have been derived. Only for the unichain case where $v \in V$ is unique up to an additive constant have upper and lower bounds been obtained, (cf. [5]). However, HASTINGS [7] recently suggested a device for eliminating actions on a *temporary* basis which reduces the amount of required computations considerably. Using the geometric convergence result of (ordinary) value-iteration can show that any action $k \in K(i)$ which does not lie within $L(i,L(v(0)))$ will eventually be eliminated (cf. remark {4} in [5]).

We finally turn to the wider class of "Markov Renewal Programs" (MRP's) in which the state of the system is not necessarily observed at equally spaced epochs, but in which the transition time between two successive observations of state is a random variable the distribution of which depends both upon the last state observed and the action chosen.

For all $i \in \Omega$, $k \in K(i)$ let $T_i^k > 0$ denote the expected holding time in state i when choosing action $k \in K(i)$.

Both the Policy Iteration Algorithm and the Linear Programming Approaches which were originally developed for MDP's have been adapted for the more general MRP-model (cf. e.g. [2]). To obtain a successive approximation method for undiscounted MRP's we recall the following generalization of the data-transformation (1.8) which turns every MRP into an *equivalent MDP* (the equivalence notion being defined above) (cf. [4] and [10]):

$$
\begin{aligned}
&\tilde{q}_i^k = q_i^k / T_i^k && i \in \Omega;\ k \in K(i) \\
&\tilde{P}_{ij}^k = \delta_{ij} + \tau(P_{ij}^k - \delta_{ij}) / T_i^k && i,j \in \Omega;\ k \in K(i) \\
&\tilde{T}_i^k = 1 && i \in \Omega;\ k \in K(i)
\end{aligned}
\tag{1.17}
$$

where τ has to be chosen such that

$$
0 < \tau \leq \min \{ T_i^k / (1-P_{ii}^k) \mid (i,k) \text{ with } P_{ii}^k < 1 \}. \tag{1.18}
$$

By choosing τ strictly less than the upperbound in (1.18) the *same* transformation ensures, that every policy in the transformed MDP is aperiodic, such that the value-iteration method is guaranteed to converge for any starting point, with all of the nice consequences that were exhibited

above.

In addition, the relationship between the solution set of the optimality equation in the MRP-model and the corresponding set in the transformed MDP-model is similar to (1.9). As a consequence, applying value-iteration to the transformed model will even yield us a solution to the optimality equation associated with the original MRP-model.

REFERENCES

1 BLACKWELL, D., *Discrete Dynamic Programming*, Ann. Math. Stat. 33 (1962), p. 719-726.

2 DENARDO, E., & B. FOX, *Multichain Markov Renewal Programs*, SIAM, J. Appl. Math. 16 (1968), 468-487.

3 FEDERGRUEN, A. & P.J. SCHWEITZER, *Turnpike properties in undiscounted Markov Decision Problems*, (forthcoming).

4 _____ & _____, *Data-transformations in Markov Decision Problems (forthcoming)*.

5 _____ & _____ & H.C. TIJMS, *Contraction Mappings underlying undiscounted Markov Decision Problems*, Math. Center Report BW 72/77 (to appear in J. Math. Anal. Appl.).

6 HASTINGS, N., *Bounds on the gain of a Markov Decision Process*, Op. Res. 19 (1971), p. 240-244.

7 _____ , *A test for non-optimal actions in undiscounted finite Markov Decision Chains*, Man. Sci. 23 (1976), p. 87-92.

8 HORDIJK, A. & H.C. TIJMS, *A modified form of the iterative method of dynamic programming*, An. of Stat. 3, (1975), p. 203-208.

9 MILLER, B. & A.F. VEINOTT, *Discrete Dynamic Programming with a small interest rate*, Ann. Math. Stat. 40, (1969), p. 366-370.

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

11 SCHWEITZER, P.J., *Iterative solution of the Functional Equations of undiscounted Markov Renewal Programming*, J.M.A.A. 34 (1971), p. 495-501.

THE ACTION ELIMINATION ALGORITHM FOR MARKOV DECISION PROCESSES

N.A.J.Hastings
Monash University, Melbourne, Australia

J.A.E.E. van Nunen
Graduate School of Management, Delft, The Netherlands

1. INTRODUCTION

We consider a finite Markov decision chain with or without discounting. The state space is S, where the states are labeled i = 1,2,...,N. If the system is in state $i \in S$ at time n an action k has to be selected from a nonempty finite set K_i. As a consequence of this action $k \in K_i$ we earn a(n) (expected) reward r(i,k) and the system moves to state $j \in S$ at time n + 1 with probability p(i,j,k). We assume $\sum_j p(i,j,k) = 1$.

The Cartesian product of all sets K_i is the policy space Λ. For any policy $\delta \in \Delta$ we denote by $P(\delta)$ the transition probability matrix and by $r(\delta)$ the column vector of rewards. Rewards earned in the n-th period are discounted by a factor $\beta > 0$ (eventually $\beta \geq 1$). Our goal is to find a strategy that maximizes the total expected reward over a time horizon $T \in \mathbb{N} \cup \{\infty\}$, and to determine the corresponding optimal reward vector v_T. Here, a strategy π_T for a T-horizon problem is a sequence of policies $\pi_T := (\delta_1, \delta_2, \ldots, \delta_T)$. Note that we restrict the considerations, as is allowed, to nonrandomized strategies. For $T = \infty$ it is even permitted to consider only stationary strategies i.e. $\pi := \pi_\infty := (\delta,\delta,\delta,\ldots)$. The optimal value vector v_T^* can be computed by the value iteration algorithm of dynamic programming. For finite horizon problems we refer to HINDERER [4] and HÜBNER [6]. For $T = \infty$ we refer to e.g. HASTINGS [1] or VAN NUNEN [9].

In the latter situation dynamic programming yields in the limit policies which can be used to constitute stationary strategies that are optimal.

As indicated in e.g. [1], [4], [6], [10] convergence is monitored by using upper and lowerbounds on the optimal return vector v_T^*. These bounds are used to construct sub-optimality tests, see for example references [8], [3], [2], [10]. The test proposed here increases the efficiency of the dynamic programming method considerably. A nonoptimal action for a given stage (iteration) is one which does not form part of an optimal policy for that stage. Until now, in the discounted case, tests have been devised whereby only those actions which can be identified as being non-optimal for all subsequent stages are eliminated. For the average reward situation HASTINGS [3] proposed to eliminate actions for one or more stages after which they may reenter the action space. Here we extend this idea to Markov decision processes which may be undiscounted or discounted, may have a finite or infinite time horizon and in the finite horizon case may have a discount factor that is allowed to be greater than one.

2. THE TEST

Let $f(n,i)$ be the maximum total expected return generated when the system starts in state $i \in S$ and continues for n-stages. Then

$$(1) \qquad f(n,i) := \max_{k \in K_i} [r(i,k) + \beta \sum_{j \in S} p(i,j,k) f(n-1,j)]$$

where $f(0)$ is given and $\beta > 0$. The value iteration algorithm computes $f(n,i)$ for $i \in S$ and $n = 1,2,\ldots,T$.

Define

$$(2) \qquad f(n,i,k) := r(i,k) + \beta \sum_{j \in S} p(i,j,k) f(n-1,j)$$

$$(3) \qquad y(n,i,k) := f(n,i) - f(n,i,k) \geq 0$$

$$(4) \qquad \theta_u(n) := \max_{i \in S} [f(n,i) - f(n-1,i)]$$

$$(5) \qquad \theta_\ell(n) := \min_{i \in S} [f(n,i) - f(n-1,i)]$$

(6) $\phi(n) := \beta(\theta_u(n) - \theta_\ell(n))$

(7) $H(m,n,i,k) := y(n,i,k) - \sum_{\ell=n}^{m-1} \phi(\ell) \qquad m > n.$

Note that

(8) $H(m+1,n,i,k) \leq H(m,n,i,k).$

In the test we will use, any action $k \in K_i$ is nonoptimal for state $i \in S$ at value iteration stage m if

$$H(m,n,i,k) > 0.$$

3. BASIC PROPERTIES

LEMMA 1

a) $\phi(m) \leq \beta\phi(m-1)$

b) $f(n+1,i,k) - f(n,i,k) \leq \beta\theta_u(n)$

c) $f(n+1,i) - f(n,i) \geq \beta\theta_\ell(n)$

d) $y(m,i,k) \geq H(n,i,k) \qquad$ for $m > n$

e) $H(m,n,i,k) \geq y(n,i,k) - \frac{1-\beta^{m-n}}{1-\beta}\phi(n), \qquad m > n$

PROOF. Part a is a direct consequence of HÜBNER [6] theorem 1.1. The second part of the lemma follows from

$$f(n+1,i,k) - f(n,i,k) = r(i,k) + \beta \sum_{j\in S} p(i,j,k)f(n,j) - r(i,k) - \beta \sum_{j\in S} p(i,j,k)f(n-1,j).$$

$$= \beta \sum_{j\in S} p(i,j,k)[f(n,j) - f(n-1,j)] \leq \beta\theta_u(n)$$

consider

$$f(n+1,i) - f(n,i) \geq r(i,k_0) + \beta \sum_{j\in S} p(i,j,k_0) f(n,j) - r(i,k_0) +$$

$$- \beta \sum_{j\in S} p(i,j,k_0) f(n-1,j)$$

$$= \beta \sum_{j\in S} p(i,j,k_0)[f(n,j) - f(n-1,j)] \geq \beta\theta_{\ell}(n),$$

with k_0 that action in K_i for which the maximum in $f(n,i)$ is attained.
This proves part c).
Since

$$y(m,i,k) = f(m,i) - f(m,i,k) \geq f(m-1,i) + \beta\theta_{\ell}(m-1) - f(m-1,i,k) +$$

$$- \beta\theta_{u}(m-1) = y(m-1,i,k) - \phi(m-1)$$

the result d) follows by iterating stagewise.

The final statement of the lemma is a direct consequence of part a of this lemma and the definition of $H(m,n,i,k)$. □

THEOREM 1.

a) *Action* k *at state* i *is nonoptimal at stage* $m > n$ *if* $H(m,n,i,k) > 0$.

b) *Action* k *at state* i *is nonoptimal at stage* $m > n$ *if*

$$y(n,i,k) - \frac{1-\beta^{m-n}}{1-\beta} \phi(n) > 0$$

c) *Action* k *at state* i *is nonoptimal for all subsequent stages if*

$$y(n,i,k) - \frac{1-\beta^{T-n}}{1-\beta} \phi(n) > 0, \qquad T > n.$$

PROOF. The proof follows from the foregoing lemma. Part b) and c) can also be found in HÜBNER [6].

REMARK. For $\beta = 1$ the term $\frac{1-\beta^{m-n}}{1-\beta}$ has to be replaced by $(m-n)$. For $T = \infty$ the theorem makes sense only if $\beta < 1$. However, the condition can be weakened see HÜBNER [6] or Porteus [11].

Since in our test actions are eliminated which are nonoptimal for perhaps only one stage, it will be clear that the first stage at which our test eliminates an action for the first time will in general be much earlier than the first stage at which e.g. the MACQUEEN test [8] or the HASTINGS and MELLO test [2] eliminates that action.

This follows directly from the foregoing theorem.

COROLLARY 1. *For $0 < \beta < 1$ our test is tighter than MacQueen's test and the Hastings and Mello test for eliminating optimal actions.*

PROOF. MacQueen based his test on part c) of theorem 1, with $T = \infty$. So in his test an action k is nonoptimal in state i if

$$y(n,i,k) - \frac{1}{1-\beta}\,\phi(n) > 0.$$

In our test an action is eliminated for the first time if $y(n,i,k) > \phi(n)$.

Clearly

$$\phi(n) < \frac{\phi(n)}{1-\beta} \qquad \text{for } 0 < \beta < 1.$$

Since the MacQueen test is tighter than the Hastings and Mello test the corollary is proved.

REMARK: From the foregoing analysis it will be clear that any action that fails the MacQueen-test at stage n cannot be optimal at any stage $m > n$. In our test such an action will be eliminated permanently the first time it passes the test after stage n.

REMARK. Note that for $\beta \to 1$ the relative power of our test will be greater since $(1-\beta)^{-1} \to \infty$ as $\beta \to 1$.

4. COMPUTATIONAL METHOD

To illustrate the computational method we give a flow chart of the test. Before drawing such a flow chart we have to give some more pre-

liminaries. We assume the terminal values f(0) = 0 and apply the test from stage two onwards. We set the test quantity T(n,i,k) at zero at stage 1. An action fails the test if its test quantity (called flag) is positive or if its flag is "nonoptimal". If the action fails at stage n its trial value f(n,i,k) is then not evaluated at that stage. For an action which passed the test at stage n, the flag T(n,i,k) could be reset to

$$T(n,i,k) := \begin{cases} \text{"nonoptimal" if } y(n,i,k) - \dfrac{1-\beta^{T-n}}{1-\beta}\,\phi(n) > 0, \\ \\ y(n,i,k) \quad \text{elsw.} \end{cases}$$

For an action which fails the test at stage n - 1, the flag T(n,i,k) is given by

$$T(n,i,k) := \begin{cases} \text{"nonoptimal" if } T(n-1,i,k) = \text{"nonoptimal"}, \\ \\ T(n-1,i,k) - \phi(n-1). \end{cases}$$

However as in [3], to avoid the making of a second pass it is preferable to use by resetting the "flag" after an action passed the test

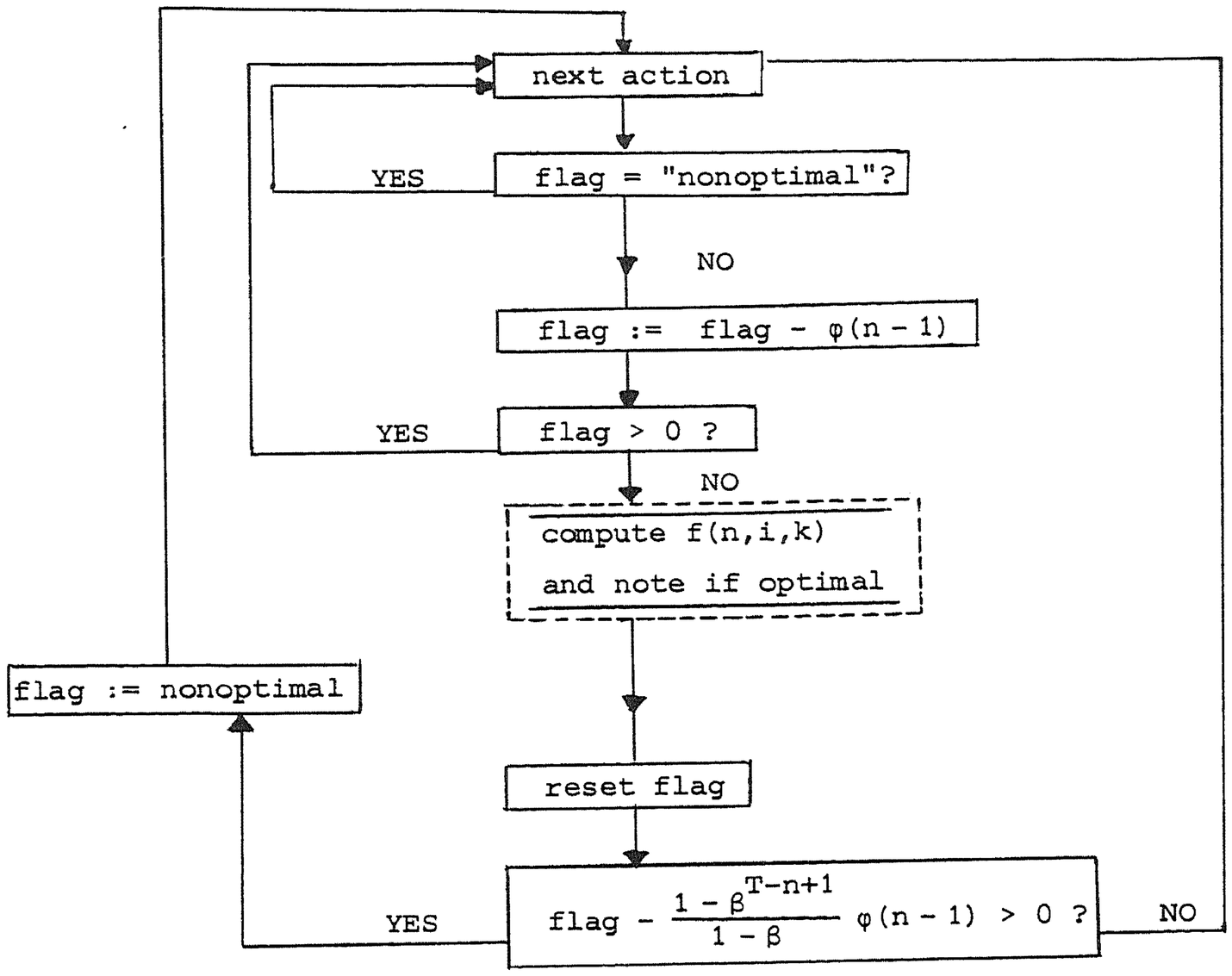
$$f(n-1,i,k) + \beta\theta_{\ell}(n-1) - f(n,i,k)$$

instead of

$$y(n,i,k) := f(n,i) - f(n,i,k).$$

The effect of the test is to reduce the number of times that the time consuming step of evaluating f(n,i,k) is carried out (this step is marked by a dotted line).

The flow chart of the action elimination algorithm has the following structure.



5. NUMERICAL EXAMPLE*

The extreme efficiency of the test will be shown by applying it to Howard's automobile replacement problem [5 p.p. 54-59] with discountfactor $\beta = 0.97$. We use the dynamic programming algorithm of MACQUEEN [7]. We compare the number of actions eliminated by the HASTINGS and MELLO test [2] with the number of eliminated actions by the test proposed in this paper. In the first test only actions which are nonoptimal for the whole future are eliminated. We start the dynamic programming algorithm with a starting vector with all components equal to zero i.e. $f(0,i) = 0$ for all $i \in S$.

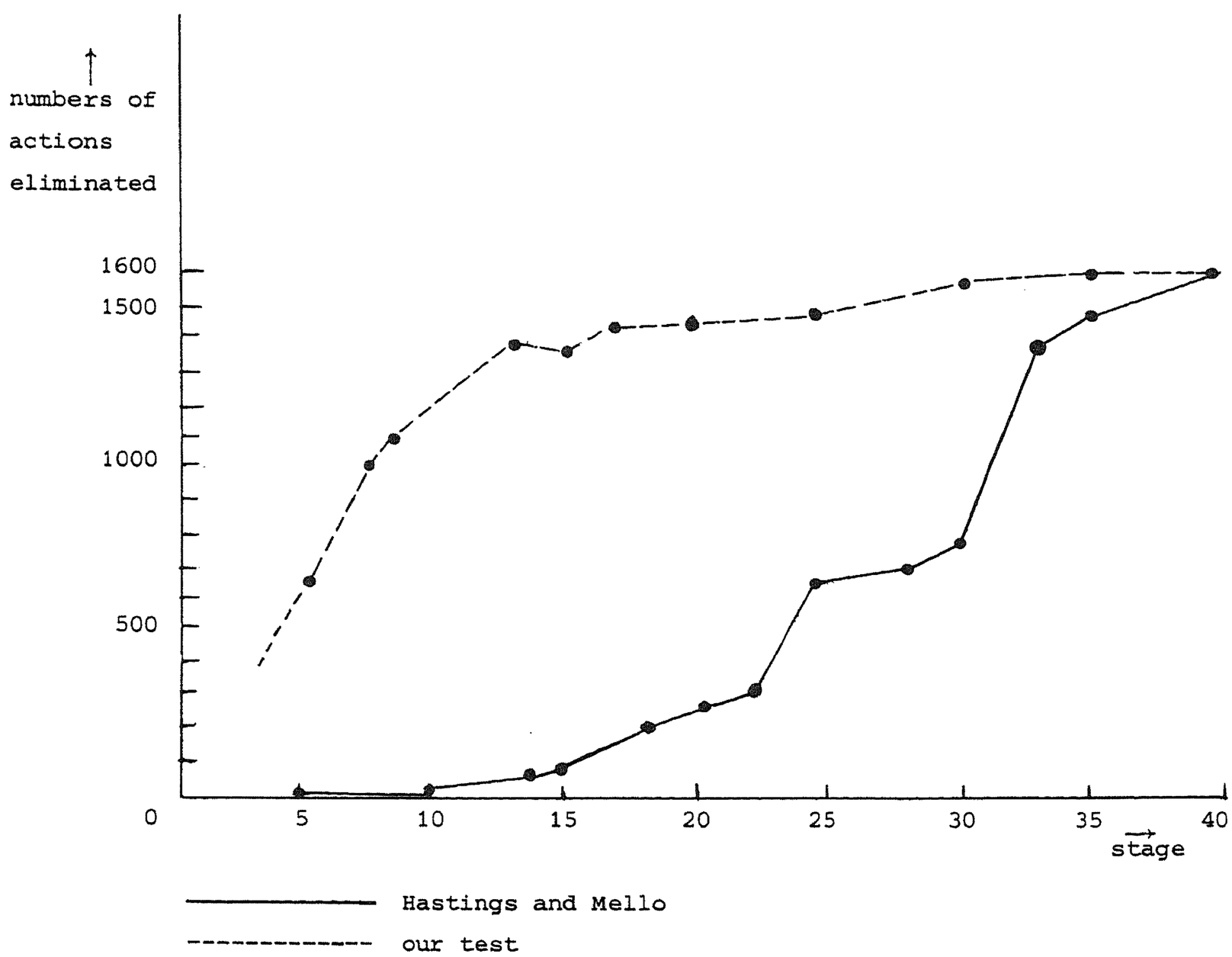
In figure 1 we see that the difference between the number of actions that are eliminated is significant. From iteration 8 until iteration 22 this difference is even over 1000 actions.

---

* The authors are grateful to mr. K. van der Hoeven for computational support.

Figure 1

Application to the automobile Replacement problem



7. SOME EXTENSIONS AND REMARKS

In this note we have assumed the equal row sum property. However, the same ideas can be used for a nonoptimality test, in the case that this assumption is released. We then have to exploit more sophisticated bounds for the values f(n+m,i). These bounds are described for example in PORTEUS [11] or VAN NUNEN [10]. In [10] it is shown that discounted semi-

Markov decision processes are covered by the model without the "equal row sum"-property, so also for such processes our test can be used.

In [9] and [10] a whole set of successive approximation algorithms for Markov decision problems containing the Jacobi-, the Gauss Seidel- and overrelaxation algorithms is developed. The nonoptimality test can be incorporated in those algorithms as well.

It is known see e.g. HÜBNER [6], PORTEUS [11] that the contraction factor is sometimes even smaller than the discount factor β. In that case the nonoptimality test can be refined by using the more sophisticated contraction factor.

For infinite horizon problems (in the equal row sum case) with respect to the total reward criterion convergence of $f(n,i)$ is only guaranteed if $\beta < 1$. However, for finite horizon problems β is allowed to be greater then or equal to one. If the equal row sum property is not satisfied, convergence of the total expected reward may occur for $\beta \geq 1$, see PORTEUS [11] or VAN NUNEN [10].

REFERENCES

[1] HASTINGS, N.A.J.,*"Dynamic Programming with Management Applications"*, Butterworths, London and Crane-Russak, New York, 1973.

[2] HASTINGS, N.A.J. and J.M.C. MELLO, *"Test for suboptimal actions in discounted Markov Programming"*, Management Sci. 19, 1973, pp. 1019-1022.

[3] HASTINGS, N.A.J., *"A test for nonoptimal actions in undiscounted finite Markov decision chains"*, Management Sci. 23, 1976, pp.87-91.

[4] HINDERER, K., *"Estimates for finite state dynamic programs"*, (preprint 1974) to appear in J. Math. Anal. Appl.

[5] HOWARD, R.A., *"Dynamic programming and Markov processes"*, Wiley, New York-London, 1960.

[6] HÜBNER, G., *"Improved procedures for eliminating suboptimal actions in Markov programming by the use of contraction properties"*, to appear in: Transactions of the seventh Prague Conf. 1974.

[7] MACQUEEN, J., *"A modified dynamic programming method for Markovian decision problems"*, J. Math. Anal. Appl. 14, 1966, pp. 38-43.

[8] MACQUEEN, J., *"A test for suboptimal actions in Markovian decision problems"*, Oper. Res. 15, 1967, pp. 559-561.

[9] NUNEN, J.A.E.E. VAN and J. WESSELS, *"A principle for generating optimization procedures for discounted Markov decision processes"*. In: "Progress in Operations Research" ed. by A. Prékopa. North-Holland Publ. Company 1976, pp. 683-695.

[10] NUNEN, J.A.E.E. VAN, *"Contracting Markov Decision Processes"*, Math. Centre Tract no. 71, Math. Centre, Amsterdam, 1976

[11] PORTEUS, E.L., *"Bounds and transformations for discounted finite Markov decision chains"*. Oper. Res. 23. 1975, pp. 761-784.

APPROXIMATIONS IN BAYESIAN CONTROLLED MARKOV CHAINS

K.M. van Hee
Eindhoven University of Technology, Eindhoven, The Netherlands

1. INTRODUCTION AND PRELIMINARIES

For a detailed description of the model we refer to VAN HEE (1976A), here we only give a sketch. For statements without proof see also VAN HEE (1976A). Consider a Markov decision process with a finite *state space* S and a finite *action space* A. Let $r: S \times A \to \mathbb{R}$ be the *reward function*.

Let X_n be the state of the system at time n. There is a subset $B \subset S$ such that if $X_n \in B$ the next state is partially determined by the outcome of a random variable Y_{n+1}, where $\{Y_n, n = 1,2,3,\ldots\}$ is a sequence of i.i.d. random variables not controllable by the decisionmaker. The process $\{Y_n, n = 1,2,3,\ldots\}$ is called the *external process* and has a finite state space E. If and only if $X_n \in B$ then Y_{n+1} becomes *visible* to the decisionmaker. Let P be a transition probability from $S \times A \times E$ to S such that

$$\mathbb{P}[X_{n+1} = t \mid X_n = s, A_n = a, Y_{n+1} = y] = P(t|s,a,y),$$

where A_n is the action at time n. (For $s \in S\backslash B$ $P(t|s,a,\cdot)$ is constant and we omit the dependence on y in this case). Only the distribution of the external process; i.e. $p(y|\theta) := \mathbb{P}_\theta[Y_{n+1} = y]$, depends on an unknown parameter $\theta \in \Theta$ where Θ is a finite *parameter* space. Note that for each fixed $\theta \in \Theta$ the process forms an ordinary Markov decision process with transition probability:

$$\mathbb{P}_\theta[X_{n+1} = t \mid X_n = s,\ A_n = a] = \sum_{y\in E} P(t|s,a,y)\cdot p(y|\theta).$$

Examples of such a model can be found in inventory control, where Y_n is the demand in period $(n-1,n]$ and also in queueing models, where Y_n is the number of newcomers.

Let Π be the set of all strategies based on the visible histories (i.e. for each $\pi \in \Pi$ the action A_n may depend on $X_0,\ldots,X_n$, $A_0,\ldots,A_{n-1}$ and on Y_k if $X_{k-1} \in B$, $k = 1,2,\ldots,n$).

For each starting state $s \in S$, each $\pi \in \Pi$ and $\theta \in \Theta$ we have a random process $\{(X_n,A_n,Y_{n+1}),\ n = 0,1,2,\ldots\}$ and a probability $\mathbb{P}^\pi_{s,\theta}[\ \ ]$ on the sample space. (The expectation w.r.t. this probability is denoted by $\mathbb{E}^\pi_{s,\theta}$) The *Bayesian expected discounted total return* $v(s,q,\pi)$ w.r.t. a *prior distribution* q on Θ is defined by

$$v(s,q,\pi) := \sum_{\theta\in\Theta} \mathbb{E}^\pi_{s,\theta}\Big[\sum_{n=0}^{\infty} \beta^n r(X_n,A_n)\Big]\cdot q(\theta),\quad s \in S,\ \pi \in \Pi,$$

where $\beta \in [0,1)$ is the discount factor.

The set of all distributions on Θ is denoted by W, and the function $v\colon S \times W \to \mathbb{R}$, defined by $v(s,q) := \sup_{\pi\in\Pi} v(s,q,\pi)$, is called the *value function*. We define a sequence of stopping times:

$$\sigma_1 := \inf\{n \geq 0;\ X_n \in B\}$$

$$\sigma_k := \inf\{n > \sigma_{k-1};\ X_n \in B\},\quad k = 2,3,4,\ldots$$

$$\tau_k := \sigma_k + 1,\quad k = 1,2,3,\ldots\ .$$

The Bayes criterion allows us to consider the parameter $\theta \in \Theta$ as a random variable Z with distribution q on Θ.

Given $q \in W$, $s \in S$ and $\pi \in S$ we have a probability $\mathbb{P}^\pi_{s,q}$ on the sample space of the process $\{Z,(X_n,A_n,Y_{n+1}),\ n = 0,1,2,\ldots\}$ and for each event C defined by

$$C := \{X_0 = s_0,A_0 = a_0,Y_1 = y_1,\ldots,X_n = s_n,A_n = a_n,Y_{n+1} = y_{n+1}\}$$

we have

$$\mathbb{P}^\pi_{s,q}[C] = \sum_{\theta\in\Theta} \mathbb{P}^\pi_{s,\theta}[C]q(\theta)$$

(the expectation w.r.t. $\mathbb{P}^{\pi}_{s,q}$ is denoted by $\mathbb{E}^{\pi}_{s,q}$).

We define on the event $\{\tau_n < \infty\}$ for $s \in S$, $q \in W$, $\pi \in \Pi$:

$$\gamma_n(\theta) := \mathbb{P}^{\pi}_{s,q}[Z = \theta \mid Y_{\tau_1},\ldots,Y_{\tau_n}]$$

and

$$Q_n(\theta) := \gamma_k(\theta) \quad \text{on} \quad \{\tau_k \le n < \tau_{k+1}\}.$$

The vector valued process $\{Q_n\}$ with $Q_n := \{Q_n(\theta),\ \theta \in \Theta\}$ is called the *Bayes process*.

Note that, since the values of the external process are not influenced by the starting state s and the strategy π, we have that $\gamma_n(\theta)$ does not depend on s and π on $\{\tau_n < \infty\}$, and likewise for $Q_n(\theta)$ if $B = S$.

If we are in the situation that expectations or conditional expectations do not depend on s and π we omit these sub- and superscript.

We sometimes need the following conditions:

(A) for all $s \in S$, $\pi \in \Pi$ and $\theta \in \Theta$

$$\mathbb{P}^{\pi}_{s,\theta}\left[\bigcap_{n=1}^{\infty} \{\tau_n < \infty\}\right] = 1.$$

(Note that $B = S$ implies (A).)

(B) For each pair $\theta,\tilde{\theta} \in \Theta$ there is a $y \in E$ such that

$$p(y|\theta) \neq p(y|\tilde{\theta}).$$

(The only place where (B) is used is in the proof of the following theorem.)

THEOREM 1. *Assume* (A,B). *Then for all* $s \in S$, $q \in W$ and $\pi \in \Pi$ *it holds that*

$$\lim_{n\to\infty} Q_n(\theta) = \delta_{Z,\theta} \quad \mathbb{P}^{\pi}_{s,q}\text{-a.s.}$$

We need some notations $p: E \times W \to [0,1]$ *such that*

$$p(y,q) := \sum_{\theta\in\Theta} q(\theta)\cdot p(y|\theta),$$

$T: W \times E \to W$ *such that*

$$T_y(q)(\theta) := \frac{p(y|\theta)q(\theta)}{p(y,q)} \quad \textit{if} \quad p(y,q) > 0, := q(\theta) \ \textit{otherwise.}$$

We may reduce our Bayesian decision problem to a *discounted dynamic program* with state space $S \times W$, action space A and reward function r, as stated in theorem 2.

THEOREM 2. *The value function* v *is the unique solution of the functional equation*

$$v(s,q) = \max_{a\in A} \{r(s,a) + \beta \sum_{y\in E} \sum_{t\in S} P(t|s,a,y)p(y,q)v(t,T_y(q))\}, \ s \in B$$

$$= \max_{a\in A} \{r(s,a) + \beta \sum_{t\in S} P(t|s,a)v(t,q)\}, \quad s \in S\backslash B.$$

COROLLARY 1. *There is an optimal strategy* π^* *which is stationary, i.e. there is a function* $g: S \times W \to A$ *such that* π^* *chooses action* $g(s,q)$ *in* $(s,q) \in S \times W$.

2. APPROXIMATIONS

In this section we shall give some approximations for $v(s,q)$, $s \in S$ and a fixed prior $q \in W$. In section 3 we consider the computational aspects. We identify each $\theta \in \Theta$ with the degenerated distribution at θ. Hence $v(s,\theta)$ is the optimal value of the Markov decision process if s is the starting state and θ is known. Let

$$(2.1) \qquad M := \{f \mid f: S \to A\}$$

be the set of *Markov policies* and identify the strategy $\pi \in \Pi$ that chooses action $f(s)$ in state (s,q) with f.

Further let

$$F_\theta := \{f \in M \mid v(s,\theta) = v(s,\theta,f) \ \text{ for all } s \in S\}, \ \theta \in \Theta$$

and $c: \Theta \to M$ be such that $c(\theta) \in F_\theta$, $\theta \in \Theta$. We define

$$(2.2) \qquad \text{i)} \quad F := \bigcup_{\theta\in\Theta} F_\theta$$

$$\text{ii)} \quad \bar{F} := \{f \in M \mid f = c(\theta) \text{ for some } \theta \in \Theta\}.$$

On $S \times W$ we define the following functions:

(2.3) i) $w(s,q) := \sum_{\theta\in\Theta} v(s,\theta)q(\theta)$

ii) $\underline{\ell}(s,q) := \max_{f\in F} \sum_{\theta\in\Theta} v(s,\theta,f)q(\theta)$

iii) $\bar{\ell}(s,q) := \max_{f\in \bar{F}} \sum_{\theta\in\Theta} v(s,\theta,f)q(\theta)$.

LEMMA 3.

i) $\bar{\ell}(s,q) \le \underline{\ell}(s,q) \le v(s,q) \le w(s,q)$ *for all* $s \in S$, $q \in W$.

ii) *Let* (A,B) *hold, then for all* $t \in S$, $\pi \in \Pi$ *and* $q \in W$

$$\lim_{n\to\infty} \max_{s\in S} \{w(s,Q_n) - \bar{\ell}(s,Q_n)\} = 0, \ \mathbb{P}^{\pi}_{t,q}\text{-a.s.}$$

PROOF.

i) $$\bar{\ell}(s,q) \le \underline{\ell}(s,q) = \max_{f\in F} v(s,q,f) \le \sum_{\theta\in\Theta} q(\theta) \sup_{\pi\in\Pi} v(s,\theta,\pi) = w(s,q).$$

ii) By theorem 1 we have

$$\lim_{n\to\infty} w(s,Q_n) = \lim_{n\to\infty} \sum_{\theta\in\Theta} v(s,\theta)Q_n(\theta) = v(s,Z), \ \mathbb{P}^{\pi}_{t,q}\text{-a.s.} .$$

Note that

$$\left|\max_{f\in\bar{F}} \sum_{\theta} Q_n(\theta)v(s,\theta,f) - \max_{f\in\bar{F}} v(s,Z,f)\right| \le \max_{f\in\bar{F}} \left|\sum_{\theta} Q_n(\theta)v(s,\theta,f) - v(s,Z,f)\right|.$$

Hence

$$\lim_{n\to\infty} \bar{\ell}(s,Q_n) = v(s,Z), \ \mathbb{P}^{\pi}_{t,q}\text{-a.s.} \quad \square$$

Define two functions:

i) $$\phi(s,a,\theta) := r(s,a) + \beta \sum_{t\in S} \sum_{y\in E} P(t|s,a,y)p(y|\theta)v(t,\theta) - v(s,\theta),$$

(2.4) $s \in S$, $a \in A$, $\theta \in \Theta$.

ii) $$\phi(s,q) := \max_{a\in A} \sum_{\theta\in\Theta} \phi(s,a,\theta)q(\theta), \ s \in S, \ q \in W.$$

Note that $\phi(s,a,\theta) \le 0$ for all $s \in S$, $a \in A$ and $\theta \in \Theta$ and note also that $\phi(s,q) = 0$ if q is a degenerated distribution.

LEMMA 4.

i) $$v(s,q) \geq w(s,q) + \frac{1}{1-\beta} \max_{a\in A} \sum_{\theta\in\Theta} \min_{x\in S} \phi(x,a,\theta)q(\theta).$$

ii) $$v(s,q) \geq w(s,q) + \frac{1}{1-\beta} \max_{f\in F} \sum_{\theta\in\Theta} \min_{x\in S} \phi(x,f(x),\theta)q(\theta).$$

iii) *If* B = S *then*

$$\operatorname*{span}_{s} \{w(s,q) - v(s,q)\} \leq \mathbb{E}_q\Big[\sum_{n=0}^{\infty} \beta^n \operatorname*{span}_{s} \phi(s,Q_n)\Big]. \quad *$$

PROOF. By (2.4) we have for a $\in$ A:

$$\sum_{\theta\in\Theta} q(\theta)v(s,\theta)+\phi(s,q) \geq r(s,a) + \beta \sum_{\theta} \sum_{t} \sum_{y} P(t|s,a,y)p(y|\theta)q(\theta)v(t,\theta).$$

Note that $p(y|\theta)q(\theta) = T_y(q))(\theta)p(y,q)$. Hence by substituting 2.3i) we have

$$w(s,q) + \phi(s,q) \geq r(s,a) + \beta \sum_{t} \sum_{y} P(t\ s,a,y)p(y,q)w(t,T_y(q)), \text{ if } s \in B$$

$$\geq r(s,a) + \beta \sum_{t} P(t|s,a)w(t,q), \quad \text{if } s \in S\backslash B.$$

Let π be a stationary strategy and define

$$f(s,q) := \mathbb{E}^{\pi}_{s,q}[w(X_1,Q_1)], \quad s,q \in S \times W$$

then, if a is the action strategy π in (s,q):

$$r(s,a) \leq \phi(s,q) + w(s,q) - \beta f(s,q).$$

By the Markov property we have for all (s,q) $\in$ S $\times$ W:

$$f(X_n,Q_n) = \mathbb{E}^{\pi}_{s,q}[w(X_{n+1},Q_{n+1}) \mid X_n,Q_n], \ \mathbb{P}^{\pi}_{s,q}\text{-a.s.}$$

Hence

---

* $\operatorname*{span}_{x} f(x) := \sup_{x} f(x) - \inf_{x} f(x)$.

$$(2.5)\qquad v(s,q,\pi) \le \mathbb{E}^{\pi}_{s,q}\Big[\sum_{n=0}^{\infty}\beta^n\phi(X_n,Q_n)\Big] + \mathbb{E}^{\pi}_{s,q}\Big[\sum_{n=0}^{\infty}\beta^n w(X_n,Q_n)\Big] + $$
$$- \mathbb{E}^{\pi}_{s,q}\Big[\beta\sum_{n=0}^{\infty}\beta^n w(X_{n+1},Q_{n+1})\Big] = w(s,q) + \mathbb{E}^{\pi}_{s,q}\Big[\sum_{n=0}^{\infty}\beta^n\phi(X_n,Q_n)\Big].$$

Let $\tilde{\pi}$ be the strategy that chooses in $(s,q) \in S \times W$ a fixed action a, maximizing $\sum_\theta q(\theta)\phi(s,a,\theta)$. Note that $\tilde{\pi}$ is stationary and note also that equality holds in (2.5) if $\pi = \tilde{\pi}$.

We first prove iii).

Let π^* be a stationary optimal strategy, then

$$(2.6)\qquad v(s,q) = v(s,q,\pi^*) \le w(s,q) + \mathbb{E}^{\pi^*}_{s,q}\Big[\sum_{n=0}^{\infty}\beta^n \max_{x\in S}\phi(x,Q_n)\Big].$$

But

$$(2.7)\qquad v(s,q) \ge v(s,q,\tilde{\pi}) \ge w(s,q) + \mathbb{E}^{\tilde{\pi}}_{s,q}\Big[\sum_{n=0}^{\infty}\beta^n \min_{x\in S}\phi(x,Q_n)\Big].$$

Remark that under the condition B = S the distribution of Q_n is independent of $s \in S$ and $\pi \in \Pi$, hence iii) is a direct consequence of (2.6) and (2.7).

To prove i) and ii) note that

$$\min_{x\in S}\phi(x,q) = \min_{x\in S}\max_{f\in F}\sum_\theta q(\theta)\phi(x,f(x),\theta) \ge \max_{f\in\bar{F}}\min_{x\in S}\sum_\theta q(\theta)\phi(x,f(x),\theta) \ge$$
$$\ge \max_{f\in\bar{F}}\sum_\theta q(\theta)\min_{x\in S}\phi(x,f(x),\theta) \ge \max_{a\in A}\sum_\theta q(\theta)\min_{x\in S}\phi(x,a,\theta).$$

Further note that the last two expressions are convex functions on W so by Jensen's inequality applied to the right hand side of (2.7) we have the desired result. □

REMARK. By the proof of lemma 4 we see that the lowerbound given in ii) is greater than or equal to the lowerbound of i), but it requires more work to compute it. Further note that, if (A,B) holds

$$(2.8)\qquad \lim_{n\to\infty}\max_{f\in\bar{F}}\sum_\theta \min_{x\in S}\phi(x,f(x),\theta)Q_n(\theta) = 0,\ \mathbb{P}^{\pi}_{s,q}\text{-a.s.}$$

since

$$Q_n(\theta) \to \delta_{Z,\theta},\ \mathbb{P}^{\pi}_{s,q}\text{-a.s.}$$

We introduce now an operator U working on the space G of bounded continuous functions on S × W (measurable w.r.t. the Borel σ-field on S × W):
Let f ∈ G:

$$(2.9) \qquad (Uf)(s,q) := \sup_{\pi\in\Pi} \mathbb{E}^{\pi}_{s,q}\left[\sum_{n=0}^{\tau_1-1} \beta^n r(X_n,A_n) + \beta^{\tau_1} f(X_{\tau_1},Q_{\tau_1})\right].$$

It is shown in VAN HEE (1976A) that if f ∈ g then also Uf ∈ g.

Note further that G is a Banach space w.r.t. the supremum norm.

In WESSELS (1974) and VAN NUNEN (1976) a class of operators of this type is studied for models with a finite respectively countable state space. They both prove the following theorem. For our situation it is proved in van HEE (1976A).

THEOREM 5. *The operator* U *(defined in 2.9) is monotone and contracting. The value function* v *is the unique fixed point of* U *in* G.

The next theorem is important for successive approximations. Let us assume that $\tilde{v}$ is an approximation of v and that the difference $|\tilde{v}-v|$ is bounded by a function ε.

THEOREM 6. *Let* v *be the value function and let* $\tilde{v}$ *and* ε ∈ G, *such that*

$$|v(s,q) - \tilde{v}(s,q)| \le \varepsilon(s,q) \qquad \textit{for all } s \in S,\ q \in W$$

then it holds that

$$|v(s,q) - (U^n\tilde{v})(s,q)| \le \sup_{\pi\in\Pi} \mathbb{E}^{\pi}_{s,q}\left[\beta^{\tau_n}\varepsilon(X_{\tau_n},Q_{\tau_n})\right].$$

PROOF. First we define the operator L: G → G by

$$(Lf)(s,q) := \sup_{\pi\in\Pi} \mathbb{E}^{\pi}_{s,q}\left[\beta^{\tau_1} f(X_{\tau_1},Q_{\tau_1})\right], \qquad f \in G,\ s \in S,\ q \in W$$

(it is easy to verify that Lf is continuous on W, so Lf ∈ G). It holds that

$$(U(v+\varepsilon))(s,q) \le (Uv)(s,q) + (L\varepsilon)(s,q) \le v(s,q) + (L\varepsilon)(s,q)$$

and therefore

$$(U^n(v+\varepsilon))(s,q) \leq v(s,q) + (L^n\varepsilon)(s,q)$$

and in the same way

$$(U^n(v-\varepsilon))(s,q) \geq v(s,q) - (L^n\varepsilon)(s,q).$$

So, again by the monotonicity of U, we have

$$\left|(U^n\tilde{v})(s,q) - v(s,q)\right| \leq (L^n\varepsilon)(s,q).$$

To complete the proof we have to verify that

$$(L^n\varepsilon)(s,q) \leq \sup_{\pi\in\Pi} \mathbb{E}^{\pi}_{s,q}[\beta^{\tau_n}\varepsilon(X_{\tau_n}, Q_{\tau_n})]$$

for the rather technical proof of this statement we refer to VAN HEE (1976A). □

COROLLARY 7. *Suppose that* B = S. *Let* $\tilde{v} \in G$ *and let* $\varepsilon: W \to \mathbb{R}$ *be a bounded continuous function. If*

$$\left|v(s,q) - \tilde{v}(s,q)\right| \leq \varepsilon(q)$$

then

$$\left|v(s,q) - (U^n\tilde{v})(s,q)\right| \leq \mathbb{E}_q[\beta^n\varepsilon(Q_n)].$$

To prove this statement note that B = S implies $\tau_n = n$ and that the distribution of Q_n is independent of the starting state and the strategy.

COROLLARY 8. *Suppose that* B = S. *Let* $\tilde{v}(s,q) := \frac{1}{2}\{w(s,q) + \ell(s,q)\}$ *and*

$$\varepsilon(q) := \tfrac{1}{2} \min_{f\in\bar{F}} \sum_{\theta\in\Theta} \max_{x\in S} \{v(x,\theta) - v(x,\theta,f(x))\}q(\theta).$$

Then:

i) $$\left|v(s,q) - (U^n\tilde{v})(s,q)\right| \leq \mathbb{E}_q[\beta^n\varepsilon(Q_n)],$$

ii) $\mathbb{E}_q[\varepsilon(Q_n)] \geq \mathbb{E}_q[\varepsilon(Q_{n+1})]$,

iii) $\lim_{n\to\infty} \mathbb{E}_q[\varepsilon(Q_n)] = 0$.

PROOF.

i) Note that

$$|v(s,q) - \tilde{v}(s,q)| \leq \tfrac{1}{2}\{w(s,q) - \ell(s,q)\} \leq \varepsilon(q).$$

ii) Note that $\varepsilon(q)$ is a concave function on W. Since $\{Q_n, n \in \mathbb{N}\}$ forms a martingale (see VAN HEE (1976A)) we have that $\{\varepsilon(Q_n), n \in \mathbb{N}\}$ forms a supermartingale.

iii) By theorem 1 we have $\mathbb{P}_q$-a.s.

$$\lim_{n\to\infty} \varepsilon(Q_n) = \tfrac{1}{2} \min_{f\in\bar{F}} \max_{x\in S} \{v(x,Z) - v(x,Z,f(x))\} = 0. \quad \square$$

REMARK 2.10. Let B = S and define

i) $\varepsilon(q) := \tfrac{1}{2} \frac{1}{1-\beta} \max_{f\in\bar{F}} \sum_{\theta\in\Theta} \min_{x\in S} \phi(x,f(x),\theta)q(\theta)$,

ii) $\tilde{v}(s,q) := w(s,q) + \varepsilon(q)$.

Then the three statements of corollary 8 hold also if we replace ε by $|\varepsilon|$. The proof proceeds along the same lines, using lemma 4 and 2.8.

3. COMPUTATIONAL ASPECTS AND ADDITIONAL REMARKS

The approximations given in section 2 are of interest for computations if we are prepared to determine the sets F and $\{v(s,f,\theta) \mid s \in S, \theta \in \Theta, f \in F\}$ (or F replaced by $\bar{F}$). Let $k := \#(\Theta)$ then the determination of F requires the solution of k ordinary Markov decision problems with a finite state and action space and the determination of all optimal policies. If $n := \#(F)$ (or $\#(\bar{F})$) then we have to solve (k-1)n systems of linear equations to determine the second set.

If there is a $f \in M$ which is optimal for all $\theta \in \Theta$ then $v(s,q) = w(s,q)$ for all $s \in S$, $q \in W$.

In VAN HEE (1976B) a class of problems, including some inventory control models, is considered with the property that $\operatorname{span}_s \phi(s,q) = 0$ for all $q \in W$ hence by lemma 4 iii) we have that $\operatorname{span}_s \{v(s,q) - w(s,q)\} = 0$.

For each $q \in W$ we define

$$W_n(q) := \{\phi \in W \mid \phi = T_{y_n}(T_{y_{n-1}}(\ldots(T_{y_1}(q))\ldots), y_1,\ldots,y_n \in E\}$$

hence $W_n(q)$ is the set of all n-stage posterior distributions of q. The sets $W_n(q)$ and $W_m(q)$, $n \neq m$ are in general not disjoint (see VAN HEE (1976A)). For a fixed $q \in W$ it follows from section 2, that, loosely speaking, the approximations of $v(s,\phi)$ for $\phi \in W_n(q)$ are better if n is large.

Since $(U^n\tilde{v})(s,q)$ requires only the values of $\tilde{v}(s,\phi)$ for $\phi \in W_n(q)$, $s \in S$ we may approximate $v(s,\phi)$ by $\tilde{v}(s,q)$ on $W_n(q)$ and then by backward induction we can determine $(U^n\tilde{v})(s,q)$. The only problem is the determination of n, the horizon.

For models with B = S corollary 8i) shows that the error determination is rather easy: we have only to compute $\beta^n \mathbb{E}_q[\varepsilon(Q_n)]$, which requires the determination of $\varepsilon(\phi)$ for all $\phi \in W_n(q)$, to check whether horizon n is sufficiently accurate or not. If B is a proper subset of S and if (A,B) holds a similar result is true since

$$\sup_{\pi\in\Pi} \mathbb{E}^{\pi}_{s,q}[\beta^{\tau_n}\varepsilon(Q_{\tau_n})] \leq \beta^n \mathbb{E}_q[\varepsilon(Q_{\tau_n})],$$

viz. the distribution of Q_{τ_n} depends only on q.

In VAN HEE (1976A) two algorithms are presented based on these arguments, for models with B = S and for models where B consists of only one state. Also numerical results are given there and attention is paid to the determination of optimal actions.

In MARTIN (1967) the usual method of successive approximations is proposed with a terminal function $t: S \to \mathbb{R}$. In our terminology Martin approximates $v(s,q)$ by $(U^n t)(s,q)$. The difficulty of this method is that the choice of the horizon must be made on the error estimate $\frac{1}{2}\beta^n \frac{\bar{M}-\underline{M}}{1-\beta}$, where

$$\bar{M} := \max_{s,a} r(s,a), \quad \underline{M} := \min_{s,a} r(s,a).$$

SATIA and LAVE (1973) also suggest the use of upper and lower bounds for $v(s,\phi)$, $\phi \in W_n(q)$. It is easy to see that their bounds are worse than ours.

REFERENCES

VAN HEE, K.M., (1967A), *Bayesian control of Markov chains*, Dept. of Math., Eindhoven University of Technology, Memorandum COSOR 76-29 (May 1977).

VAN HEE, K.M., (1976B), *Adaptive control of specially structured Markov chains*, Dept. of Math., Eindhoven University of Technology, Memorandum COSOR 76-28 (December 1976).

MARTIN, J.J., (1976), *Bayesian decision problems and Markov chains*, Wiley, New York.

VAN NUNEN, J.A.E.E., (1976), *Contracting Markov decision processes*, MC-tract, 71, Amsterdam.

SATIA, J.R. & R.E. LAVE (1973), *Markov decision processes with uncertain transition probabilities*, Operations Research 21.

WESSELS, J., (1974), *Stopping times and Markov Programming*, Transactions of the 1974 E.M.S. and 7-th Prague Conference on Information Theory, Statistical decision functions and Random processes.

SUCCESSIVE APPROXIMATIONS FOR CONVERGENT DYNAMIC PROGRAMMING

K.M. van Hee
Eindhoven University of Technology, Eindhoven, The Netherlands

A.Hordijk
University of Leiden, Leiden, The Netherlands

J. van der Wal
Eindhoven University of Technology, Eindhoven, The Netherlands

1. INTRODUCTION AND PRELIMINARIES

The main topic of this paper is the convergence of the method of successive approximations for dynamic programming with the expected total return criterion. We first sketch the framework of the dynamic programming model we are dealing with. Consider a countable set E, the *state space*, and an arbitrary set A, the *action space*, endowed with some σ-field containing all one-point sets. Let p be a *transition probability* from $E \times A$ to E (notation: $p(j|i,a)$, $i,j \in E$, $a \in A$). Let $H_n := (E\times A)^n \times E$ be the set of *histories* until time n ($n \geq 1$) and $H_0 := E$.

In all generality a *strategy* π is a sequence $(\pi_0,\pi_1,\ldots)$ where π_n is a transition probability from H_n to A. The set of all strategies is denoted by Π. The subset M of Π consists of all *Markovstrategies*; i.e. $\pi = (\pi_0,\pi_1,\ldots) \in M$ if and only if there is a sequence of functions $f_0,f_1,\ldots,f_n$: $E \to A$, $n = 0,1,\ldots$, such that

$$\pi_0(\{f_0(i)\}|i) = 1,\ \pi_n(\{f_n(i)\}|h_{n-1},a_{n-1},i) = 1$$

for all $h_{n-1} \in H_{n-1}$, $a_{n-1} \in A$, and $i \in E$. Each $i \in E$ and $\pi \in \Pi$ determine a probability $\mathbb{P}_{i,\pi}$ on $(E\times A)^\infty$ and a stochastic process $\{(X_n,Y_n),\ n = 0,1,\ldots\}$ where X_n is the state and Y_n the action at time n. The expectation with respect to $\mathbb{P}_{i,\pi}$ is denoted by $\mathbb{E}_{i,\pi}$.

The *reward function* r is a real measurable function on E × A. Throughout this paper we assume

$$(1.1)\qquad \sup_{\pi\in M} \mathbb{E}_{i,\pi}\left[\sum_{n=0}^{\infty} r^{+}(X_n,Y_n)\right] < \infty \qquad \text{for all } i \in E$$

(note that $x^{+} := \max(x,0)$). This assumption guarantees that the *expected total return* $v(i,\pi) := \mathbb{E}_{i,\pi}[\sum_{n=0}^{\infty} r(X_n,Y_n)]$ is defined for all $i \in E$ and $\pi \in \Pi$, and in [9] it is proved, using a well-known theorem of DERMAN and STRAUCH [4], that

$$(1.2)\qquad \sup_{\pi\in M} v(i,\pi) = \sup_{\pi\in\Pi} v(i,\pi) \qquad \text{for all } i \in E.$$

As a consequence of 1.2 we are mainly interested in Markov strategies and for that reason we introduce some notations which are especially useful for this class. First we define the set $\mathcal{P}$ of transition probabilities P from E to E for which there is a function f: S → A such that $P(i,\cdot) = p(\cdot\,|\,i,f(i))$ for all $i \in S$; and further a function r: $E \times \mathcal{P} \to \mathbb{R}$ (= the set of reals)

$$r_p(i) := \sup\{r(i,a) \mid P(i,\cdot) = p(i,a,\cdot),\ a \in A\}.$$

Note that each $\pi \in M$ is completely determined by a sequence $R = (P_0,P_1,\ldots)$, $P_n \in \mathcal{P}$, $n = 0,1,\ldots$. Hence we may identify each $\pi \in M$ with such a sequence R, and express

$$\mathbb{E}_{i,R}\, r(X_n,Y_n) = P_0 \ldots P_{n-1} r_{P_n}(i), \qquad \text{for } R = (P_0,P_1,\ldots),\ i \in E.$$

(By convention the empty product of elements of $\mathcal{P}$ is the identity operator, and if we omit the subscript i in $\mathbb{E}_{i,R}$ we mean the function on E). On E we define the functions:

$$(1.3)\qquad v := \sup_{R\in M} \mathbb{E}_R\left[\sum_{n=0}^{\infty} r(X_n,Y_n)\right], \qquad \text{the } \textit{value function} \ ;$$

for a function s: E → $\mathbb{R}$ with $\sup_R \mathbb{E}_R[s^{+}(X_k)] < \infty$, $\quad k = 0,1,\ldots$

$$(1.4)\qquad v_n^s := \sup_{R\in M} \mathbb{E}_R\left[\sum_{k=0}^{n-1} r(X_k,Y_k) + s(X_n)\right], \qquad v_n := v_n^0 \ ;$$

for a sequence $a = (a_0,a_1,\ldots)$ of functions a_n: E → $\hat{\mathbb{R}}$, $\hat{\mathbb{R}} := \{x \in \mathbb{R} \mid x \geq 1\}$ we define the functions w_a and z_a on E:

$$(1.5)\qquad w_a(i) := \sup_{R\in M} \sum_{n=0}^{\infty} a_n(i)\ \big|\ \mathbb{E}_{i,R}[r(X_n,Y_n)]\big|, \qquad i \in E$$

$$(1.6)\qquad z_a(i) := \sup_{R\in M} \sum_{n=0}^{\infty} a_n(i)\,\mathbb{E}_{i,R}\big|r(X_n,Y_n)\big|, \qquad i \in E$$

we write w for w_a and z for z_a if $a_n \equiv 1$ for $n = 0,1,\ldots$

$$(1.7)\qquad y_1 := z, y_n := \sup_{R\in M} \sum_{k=0}^{\infty} \mathbb{E}_R[y_{n-1}(X_k)] \quad n = 2,3,\ldots .$$

A dynamic programming model is said to be *stable* with respect to *scrapfunction* s if

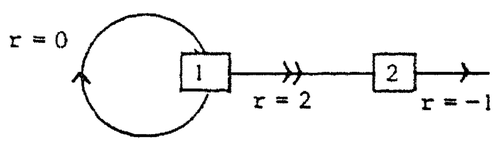
$$\lim_{n\to\infty} v_n^s(i) = v(i), \qquad \text{for all } i \in E.$$

It is well-known that positive, negative and discounted dynamic programming models with finite E and A are stable. But this is not true in *convergent dynamic programming*, the case that z is finite (see [13], [14]), as is shown by the following example.

COUNTEREXAMPLE: $E = \{1,2\}$, $A = \{1,2\}$, $p(1|1,1) = p(2|1,2) = 1$,

$r(1,1) = 0$, $r(1,2) = 2$, $p(\cdot|2,1) = p(\cdot|2,2) = 0$,

$r(2,1) = r(2,2) = -1$.



Then $v_n^0(1) = 2$ and $v(1) = 1$.

It is well-known that stability (with respect to scrapfunction 0) is guaranteed, if the expected total return from time n onwards, tends to zero as n tends to infinity uniformly in the strategy. In 1.8 this *uniform tail property* is defined:

$$(1.8)\qquad \lim_{n\to\infty} \sup_{R\in M} \sum_{k=n}^{\infty} \big|\mathbb{E}_{i,R}[r(X_k,Y_k)]\big| = 0.$$

In this paper two types of assumptions are considered to guarantee this uniform tail convergence. In section 2 the *strong convergence conditions* are introduced. A model is called strongly convergent if w_a or z_a is finite for a sequence of functions $a = (a_0,a_1,\ldots)$ with $\lim_{n\to\infty} a_n(i) = \infty$ for all $i \in E$. It turns out that property 1.8 is equivalent to a strong convergence

condition. In section 3 *Liapunov functions* are introduced and the existence of finite Liapunov functions is related to strong convergence. In section 4 Liapunov functions turn out to be important tools in successive approximations because they provide bounds for $|v-v_n^s|$ and procedures for excluding suboptimal actions.

In section 5 the connection with *contracting dynamic programming* is made and in section 6 a waiting line model with controllable input is presented, which satisfies the strong convergence condition but which is not contracting. Finally in section 7 some results on (nearly) optimal strategies are collected.

We conclude this section with some remarks and notations. Models with for each $i \in E$ a different action space A_i can easily be transformed into our frame work. In [13] and [14] convergent dynamic programming ($z<\infty$) was studied extensively. In this paper we are almost always working within this framework, since besides the overall assumption 1.1 we work with additional assumptions which are at least as strong as: w is finite. Hence with $w < \infty$ and

$$\mathbb{E}_{i,R}|r(X_n,Y_n)| \leq 2\mathbb{E}_{i,R}r^+(X_n,Y_n) + |\mathbb{E}_{i,R}r(X_n,Y_n)|$$

we have

$$z(i) \leq 2 \sup_{R\in M} \mathbb{E}_{i,R} \sum_{n=0}^{\infty} r^+(X_n,Y_n) + w(i) < \infty.$$

For two extended real valued functions a and b on E we write $a \leq b$ iff $a(i) \leq b(i)$ for all $i \in E$, and let x be an extended real number then $a \leq x$ iff $a(i) \leq x$ for all $i \in E$ (the same holds if $\leq$ is replaced by $<$ or $=$). With the convergence of a sequence of functions on E we mean pointwise convergence and the supremum of a collection of functions is the pointwise supremum. With convergence of a sequence of elements of P we mean elementwise convergence. For an extended real valued function a and a positive function b on E we write $\frac{a}{b}$ for the function $c(i) := \frac{a(i)}{b(i)}$. For a nonnegative function μ on E we introduce the set

$$V(\mu) := \{v \in \mathbb{R}^{\infty} \mid |v| \leq k\mu \quad \text{for some } k \in \mathbb{R}\}.$$

On $V(\mu)$ we define the norm μ by

$$\|f\|_{\mu} = \sup\{\mu^{-1}(i)|f(i)|\,\big|\, i \in E,\ \mu(i) > 0\}.$$

The function μ is called a *bounding function* (cf. section 5). For functions f on E with

$$\sup_{P\in\mathcal{P}} Pf^{+} < \infty$$

we define two well-known operators

$$Uf := \sup_{P\in\mathcal{P}} \{r_P + Pf\} \tag{1.10}$$

$$\tilde{U}f := \sup_{P\in\mathcal{P}} Pf.$$

Finally we formulate Bellman's optimality equations:

$$v_n^s = U^n s \tag{1.11}$$

$$v = Uv. \tag{1.12}$$

The Liapunov-approach was presented by Hordijk at the Advanced Seminar on Markov decision Theory, Amsterdam 1976. So he inspired van Hee and van der Wal to investigate the problem of successive approximations under very general conditions, which resulted in the strong convergence approach. Then the three of us joined the investigations which led to this paper.

2. STRONG CONVERGENCE

One of the main results in this section is the equivalence of the strong convergence condition with the uniform tail property expressed in 1.8. We first give some simple, but useful inequalities. Throughout this section let $a = (a_0, a_1, \ldots)$ be a nondecreasing sequence of functions, $a_n : E \to \bar{\mathbb{R}}$.

THEOREM 2.1.

$$\text{(i)} \qquad \sup_{R\in M} \sum_{k=n}^{\infty} \left| \mathbb{E}_R\, r(X_k, Y_k) \right| \le \frac{w_a}{a_n}$$

$$\text{(ii)} \qquad \sup_{R\in M} \sum_{k=n}^{\infty} \mathbb{E}_R \left| r(X_k, Y_k) \right| \le \frac{z_a}{a_n} .$$

PROOF. Since $a_k(i)$ is nondecreasing in k and $a_1(i) > 0$, we have, for all $i \in E$:

$$\sup_{R\in M} \sum_{k=n}^{\infty} \left| \mathbb{E}_{i,R}\, r(X_k, Y_k) \right| \le \sup_{R\in M} \sum_{k=n}^{\infty} \frac{a_k(i)}{a_n(i)} \left| \mathbb{E}_{i,R}\, r(X_k, A_k) \right| \le \frac{w_a(i)}{a_n(i)} .$$

The proof of (ii) is similar. □

LEMMA 2.2.

$$\text{(2.1)} \qquad \sup_{R\in M} \mathbb{E}_R\, z(X_n) = \sup_{R\in M} \mathbb{E}_R \left[\sum_{k=n}^{\infty} \left| r(X_k, Y_k) \right| \right] = \tilde{U}^n z .$$

PROOF.

$$\sup \mathbb{E}_R z(X_n) = \sup_{P_0 \cdots P_{n-1}} P_0 \cdots P_{n-1} z = \tilde{U}^n z .$$

And further:

$$\sup_{P_0 \cdots P_{n-1}} P_0 \cdots P_{n-1} z =$$

$$\sup_{P_0 \cdots P_{n-1}} P_0 \cdots P_{n-1} \sup_{P_n, \ldots} \sum_{k=0}^{\infty} P_n \cdots P_{n+k-1} \left| r_{P_{n+k}} \right| =$$

$$= \sup_{P_0, P_1, \ldots} \sum_{k=0}^{\infty} P_0 \cdots P_{n+k-1} \left| r_{P_{n+k}} \right| = \sup_{R\in M} \mathbb{E}_R \left[\sum_{k=n}^{\infty} \left| r(X_k, Y_k) \right| \right] . \quad \square$$

A direct consequence of theorem 2.1 and lemma 2.2 is

$$\text{(2.2)} \qquad \sup_{R\in M} \mathbb{E}_R \left| v(X_n) \right| \le \sup_{R\in M} \mathbb{E}_R\, z(X_n) \le \frac{z_a}{a_n} .$$

And in a similar way one may prove

$$\text{(2.3)} \qquad \sup_{R\in M} \left| \mathbb{E}_R\, v(X_n) \right| \le \frac{w_a}{a_n} .$$

One of the consequences of the above inequalities is that if $z_a < \infty$ for some sequence a with $\lim_{n\to\infty} a_n = \infty$, then $\lim_{n\to\infty} \mathbb{E}_R \left| v(X_n) \right| = 0$ for any strategy.

Hence any strategy is *equalizing* (see chapter 4 of [13]). See also theorem 7.8.

Theorem 2.3. states that $w_a < \infty$ and $\lim_{n\to\infty} a_n = \infty$ guarantee stability. Note that $w_a \leq z_a$.

THEOREM 2.3. *Let* $w(a) < \infty$ *and* $\lim_{n\to\infty} a_n = \infty$. *Then the problem is stable with respect to any scrapfunction* s *satisfying* $\sup_R \mathbb{E}_R s^+(X_n) < \infty$, $n = 0,1,\ldots$ *and* $\sup_R |\mathbb{E}_R s(X_n)| \to 0$ $(n\to\infty)$.

PROOF.

$$v - v_n^s = \sup_{R\in M} \mathbb{E}_R\left[\sum_{k=0}^{\infty} r(X_k,Y_k)\right] - \sup_{R\in M} \mathbb{E}_R\left[\sum_{k=0}^{n-1} r(X_k,Y_k) + s(X_n)\right]$$

$$\leq \sup_{R\in M} \left|\mathbb{E}_R \sum_{k=n}^{\infty} r(X_k,Y_k)\right| + \sup_{R\in M} \left|\mathbb{E}_R s(X_n)\right|$$

$$\leq \frac{w_a}{a_n} + \sup_{R\in M} \left|\mathbb{E}_R s(X_n)\right|.$$

Similarly one shows

$$v_n^s - v \leq \frac{w_a}{a_n} + \sup_{R\in M} \left|\mathbb{E}_R s(X_n)\right|.$$

Hence $\lim_{n\to\infty} |v_n^s - v| = 0$. □

So theorem 2.3 gives a new criterion for stability. If $z_a < \infty$ and $\lim_{n\to\infty} a_n = \infty$ we may use scrapfunctions s satisfying, for some $K \in \mathbb{R}$, $|s| \leq Kz$ since by theorem 2.1 and lemma 2.2 $\lim_{n\to\infty} \sup_{R\in M} \mathbb{E}_R z(X_n) = 0$.

Consider a dynamic programming model with bounded rewards, say $|r(i,a)| \leq b$ for all $i \in E$, $a \in A$ and let E_0 be an absorbing subset of E with $r(i,a) = 0$ for all $i \in E_0$, $a \in A$. Let T be the entrance time in E_0. If $\sup_{R\in M} \mathbb{E}_R T^2 < \infty$ then this model satisfies the strong convergence condition in a natural way since $z_a \leq b \sup_{R\in M} \mathbb{E}_R T^2$ for $a_n \equiv n + 1$, $n = 0,1,\ldots$.

In fact $|v_n - v| \leq b/(n+1) \sup_{R\in M} \mathbb{E}_R T^2$. Similar expressions can be derived with higher moments of the entrance time. In general one may say if $w(a) < \infty$ then $|v_n(i) - v(i)|$ tends to zero at a rate at least as fast as $[a_n(i)]^{-1}$. From the foregoing results the question arises under which conditions there exists a sequence of functions a with $a_n \to \infty$ and $w_a < \infty$.

The following theorem gives the already announced characterization.

THEOREM 2.4. *There exists a nondecreasing sequence of functions* $a = (a_0,a_1,\ldots)$ *on* E *with* $\lim_{n\to\infty} a_n = \infty$ *and* $w_a < \infty$ *if and only if*

$$w < \infty \quad \textit{and} \quad \lim_{n\to\infty} \sup_{R\in M} \sum_{k=n}^{\infty} \left| \mathbb{E}_{i,R}[r(X_k,Y_k)] \right| = 0.$$

PROOF. First the if part. Define

$$b_n(i) := \sup_{R\in M} \sum_{k=n}^{\infty} \left| \mathbb{E}_{i,R}[r(X_k,Y_k)] \right|, \qquad i \in E.$$

Obviously, $b_n \geq b_{n+1}$. Now let $a_n(i) = \ell + 1$ if $N_\ell(i) \leq n < N_{\ell+1}(i)$ with $N_0(i) := 0$ and $N_\ell(i) := \min\{n \mid b_n(i) \leq 2^{-\ell}\}$, $\ell = 1,2,\ldots$. Then

$$\sup_{R\in M} \sum_{n=N_\ell(i)}^{N_{\ell+1}(i)} a_n(i) \left| \mathbb{E}_{i,R}[r(X_n,Y_n)] \right| \leq (\ell+1)2^{-\ell}, \qquad \ell = 1,2,\ldots$$

and consequently

$$w_a(i) \leq \sup_{R\in M} \sum_{n=0}^{N_1(i)} \left| \mathbb{E}_{i,R}[r(X_n,Y_n)] \right| + \sum_{\ell=1}^{\infty} (\ell+1)2^{-\ell}$$

The only if part is immediate from $w \leq w(a) < \infty$ and theorem 2.1.(i). □

In theorem 2.5 we collect two sufficient conditions for stability which are weaker than the strong convergence condition. It is well known that positive dynamic programming models are stable, but the strong convergence condition need not be fulfilled there. The following theorem covers also the positive case.

THEOREM 2.5. *Each of the following conditions guarantees stability for scrap-function* 0.

(i) $\liminf_{n\to\infty} \inf_{R\in M} \mathbb{E}_R v(X_n) \geq 0$

(ii) *there exists a nondecreasing sequence* $a = (a_0,a_1,\ldots)$ *of functions* $a_n : E \to \hat{\mathbb{R}}$ *with* $\lim_{n\to\infty} a_n = \infty$ *and*

$$d_a(i) := \sup_{R\in M} \mathbb{E}_{i,R}\left[\sum_{n=0}^{\infty} a_n(i) r^-(X_n,Y_n) \right] < \infty \quad (x^- = \max(0,-x)). \tag{2.4}$$

PROOF. For all $R \in M$

$$v_n \geq \mathbb{E}_R\left[\sum_{k=0}^{n-1} r(X_k,Y_k)\right].$$

Hence $\liminf_{n\to\infty} v_n \geq v(\cdot,R)$ for all $R \in M$ and consequently

$$\liminf_{n\to\infty} v_n \geq \sup_{R\in M} v(\cdot,R) = v.$$

Hence to prove stability we have to show $\limsup_{n\to\infty} v_n \leq v$.
Part (i). By the optimality equation we have $r_P + Pv \leq v$, $P \in \mathcal{P}$.
Hence by iteration

$$\sum_{k=0}^{n-1} P_0 \cdots P_{k-1} r_{P_k} + P_0 \cdots P_{n-1} v \leq v$$

or

$$\mathbb{E}_R\left[\sum_{k=0}^{n-1} r(X_k,Y_k)\right] + \mathbb{E}_R[v(X_n)] \leq v.$$

Consequently, $v_n + \inf_R \mathbb{E}_R[v(X_n)] \leq v$. So with

$$\liminf_{n\to\infty} \inf_{R\in M} \mathbb{E}_R[v(X_n)] \geq 0$$

we find $\limsup_{n\to\infty} v_n \leq v$.
Part (ii). For $R \in M$

$$v(i,R) = \mathbb{E}_R\left[\sum_{k=0}^{n-1} r(X_k,Y_k)\right] + \mathbb{E}_R\left[\sum_{k=n}^{\infty} r(X_k,Y_k)\right] \geq$$

$$\geq \mathbb{E}_R\left[\sum_{k=0}^{n-1} r(X_k,Y_k)\right] - \sup_{R\in M} \mathbb{E}_R\left[\sum_{k=n}^{\infty} r^-(X_k,Y_k)\right].$$

Hence, by taking the supremum over $R \in M$

$$v \geq v_n - \sup_{R\in M} \mathbb{E}_R\left[\sum_{k=n}^{\infty} r^-(X_k,Y_k)\right].$$

Using 2.4 one proves in a way similar as in the proof of theorem 2.1

$$\sup_{R\in M} \mathbb{E}_R\left[\sum_{k=n}^{\infty} r^-(X_k,Y_k)\right] \leq \frac{d_a}{a_n}.$$

Hence

$$v \geq \limsup_{n\to\infty} v_n - \lim_{n\to\infty} \frac{d_a}{a_n} = \limsup_{n\to\infty} v_n. \quad \square$$

If for some sequence $P_0, P_1, \ldots$ we have

$$v_n = r_{P_n} + P_n v_{n-1}$$

then $\liminf_{n\to\infty} P_n \ldots P_0 v \geq 0$ is sufficient for stability, since iteration of the inequality $v \geq r_P + Pv$ yields

$$v \geq r_{P_n} + \sum_{k=1}^{n} P_n P_{n-1} \ldots P_{n-k+1} r_{P_{n-k}} + P_n \ldots P_0 v = v_n + P_n \ldots P_0 v$$

and, by the proof of theorem 2.5, $\limsup_{n\to\infty} v_n \leq v$ is sufficient for stability.

3. LIAPUNOV FUNCTIONS AND STRONG CONVERGENCE

We first introduce Liapunov functions. Consider a sequence of non-negative extended real functions $\ell_1, \ell_2, \ldots$ on E satisfying for all $P \in \mathcal{P}$ the inequalities

$$\ell_1 \geq |r_P| + P\ell_1$$

(3.1)

$$\ell_k \geq \ell_{k-1} + P\ell_k, \qquad k = 2,3,\ldots$$

Finite solutions of 3.1 are called Liapunov functions. If ℓ_k is finite ℓ_k is called a *Liapunov function* of *order* k. Note that $\ell_k < \infty$ implies $\ell_{k-1} < \infty$. Liapunov functions are powerfull tools in dynamic programming. They were first studied in a context of dynamic programming in [13] chapter 4 for the convergent dynamic programming model and in chapter 5 of [13] and in [15] Liapunov functions are studied in connection with the average return criterion for models in which some state is recurrent under each strategy and in [14] they are used to obtain (partial) Laurent expansions for the expected total discounted return. In section 4 the existence of a Liapunov function of order 2 is assumed to obtain bounds for $|v_n^s - v|$.

The functions $y_1, y_2, \ldots$ defined in 1.7 satisfy Bellman's optimality

equation, hence

$$y_1 = \sup_{P\in\mathcal{P}}\{|r_P| + Py_1\}, \qquad \text{if } y_1 < \infty$$

and

$$y_k = \sup_{P\in\mathcal{P}}\{y_{k-1} + Py_k\}, \qquad \text{if } y_k < \infty, \qquad k = 2,3,\ldots .$$

Hence, if y_k is finite, $y_1,\ldots,y_k$ are Liapunov functions and moreover it is easy to verify that $\ell_k < \infty$ implies $\ell_n \geq y_n$ $n = 1,2,\ldots,k$. Although we can work with y_k in stead of ℓ_k for theoretical purposes, it may happen in applications that one can find, in a relative simple way, Liapunov functions $\ell_1,\ell_2,\ldots,\ell_k$, while the functions $y_1,y_2,\ldots,y_k$ are hard to obtain. Since there is a large class of Liapunov functions there still is some freedom to choose an appropriate one. Specially this might improve the bounds in the approximation procedure (see also section 4). In this section we concentrate on the relations between Liapunov functions and strong convergence.

We recall that the existence of a Liapunov function of order k is equivalent to the finiteness of $y_1,\ldots,y_k$.

THEOREM 3.1.

$$y_n \geq \sup_{R\in M} \mathbb{E}_R \sum_{k=0}^{\infty} \binom{k+n-1}{k} |r(X_k,Y_k)| .$$

REMARK. Hence $y_n < \infty$ implies $z_a < \infty$ for a sequence functions $a_k \equiv \binom{k+n-1}{k}$, and consequently the strong convergence condition holds.

PROOF. By induction. For n = 1 the statement holds by definition 1.7. Suppose it holds for n - 1 (n≥2) then:

$$y_n = \sup_{P_0,\ldots} \sum_{k=0}^{\infty} P_0\ldots P_{k-1}y_{n-1} \geq$$

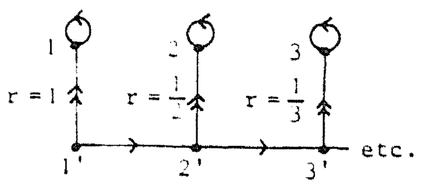
$$\geq \sup_{P_0,\ldots} \sum_{k=0}^{\infty} P_0\ldots P_{k-1} \sup_{P_k,\ldots} \sum_{\ell=0}^{\infty} \binom{\ell+n-2}{\ell} P_k\ldots P_{k+\ell-1}|r_{P_{k+\ell}}|$$

$$= \sup_{P_0,\ldots} \sum_{m=0}^{\infty} \sum_{k=0}^{m} \binom{k+n-2}{k} P_0\ldots P_{m-1}|r_{P_m}|$$

$$= \sup_{P_0,\ldots} \sum_{m=0}^{\infty} \binom{m+n-1}{m} P_0\ldots P_{m-1}|r_{P_m}| . \qquad \square$$

So $y_n < \infty$ implies $z_a < \infty$ for $a_k(i) = 0(k^{n-1})$, $k \to \infty$. The converse is not true, as shown by the following example.

COUNTEREXAMPLE 3.2. The states 1,2,... are absorbing with reward 0. In the states n', n = 1,2,..., there are two actions. Action 1 yields reward 0 and a transition to state (n+1)', action 2 yields reward n^{-1} and a transition to state n. Obviously we have for all $R \in M$



$$\mathbb{E}_R \sum_{n=0}^{\infty} (n+1)\,|r(X_n,Y_n)| \le 1$$

but since $y_1(n') = n^{-1}$ we have for the strategy R^* yielding transitions from n' to (n+1)' etc. that

$$\mathbb{E}_{1,R^*} \sum_{n=0}^{\infty} y_1(X_n) = \infty.$$

But if we make a slightly stronger assumption then

$$\sup_{R\in M} \sum_{n=0}^{\infty} n^{N-1}\, \mathbb{E}_R\,|r(X_n,Y_n)| < \infty$$

the finiteness of the functions $y_1,\dots,y_N$ defined in 1.7 can be shown.

THEOREM 3.3. *If for a nondecreasing sequence of numbers* $a_0,a_1,\dots,$ with $a_n \in \bar{\mathbb{R}}$ *and* $b = \sum_{n=0}^{\infty} a_n^{-1} < \infty$ *it holds that*

$$u := \sup_{R\in M} \mathbb{E}_R \sum_{n=0}^{\infty} a_n^{N-1}\,|r(X_n,Y_n)| < \infty$$

then the functions $y_1,\dots,y_N$ *defined in* 1.7 *are finite and satisfy the inequalities* $y_k \le ub^{k-1}a_0^{k-N}$, $k = 1,\dots,N$.

PROOF. We will prove by induction

$$\sup_{R\in M} \mathbb{E}_R\, y_k(X_n) \le ub^{k-1}a_n^{k-N} \qquad \text{for } k = 1,2,\dots,N-1,\ n = 0,1,2,\dots .$$

Set k = 1. Using $y_1 = z$ (by definition) and

$$\sup_{R\in M} \mathbb{E}_R\, z(X_n) \le ua_n^{1-N}$$

(from lemma 2.2 and theorem 2.1(ii) we get

$$\sup_{R\in M} \mathbb{E}_R\, y_1(X_n) \le ua_n^{1-N}, \qquad n = 0,1,\dots .$$

Now let us assume

$$\sup_{R\in M} \mathbb{E}_R y_k(X_n) \le ub^{k-1}a_n^{k-N} \qquad \text{for } k = 1,\dots,m \le N-2 \text{ and } n = 0,1,\dots$$

and prove that the inequalities hold for $k = m+1$.

$$\begin{aligned}\sup_{R\in M} \mathbb{E}_R y_{m+1}(X_n) &= \sup_{R\in M} P_0\cdots P_{n-1} \sup_{\tilde{R}\in M} \sum_{\ell=0}^{\infty} \tilde{P}_0\cdots\tilde{P}_{\ell-1}y_m \\ &= \sup_{P_0\cdots P_{n-1}} \sup_{\tilde{P}_0\cdots} \sum_{\ell=0}^{\infty} P_0\cdots P_{n-1}\tilde{P}_0\cdots\tilde{P}_{\ell-1}y_m \\ &= \sup_{R\in M} \sum_{\ell=0}^{\infty} \mathbb{E}_R y_m(X_{n+\ell}) \le ub^{m-1} \sum_{\ell=0}^{\infty} a_{n+\ell}^{m+N} \\ &\le ub^{m-1} \sum_{\ell=0}^{\infty} a_n^{m+1-N} a_{n+\ell}^{-1} \le ub^m a_n^{m+1-N} .\end{aligned}$$

Thus we proved

$$\sup_{R\in M} \mathbb{E}_R y_k(X_n) \le ub^{k-1}a_n^{k-N}, \qquad k = 1,2,\dots,N-1,\ n = 0,1,\dots .$$

Setting $n = 0$ we get $y_k \le ub^{k-1}a_0^{k-N}$, $k = 1,\dots,N-1$ and with

$$y_N = \sup_{R\in M} \mathbb{E}_R \sum_{n=0}^{\infty} y_{N-1}(X_n)$$

we get $y_N \le ub^{N-1}$. (And obviously $y_1,\dots,y_N$ are finite). □

COROLLARY 3.4. *If* $a_n \equiv n^{k+\varepsilon}$ *for* $n = 0,1,$ *and some* $\varepsilon > 0$, *then* $z_a < \infty$ *implies the existence of (finite) Liapunov functions* $\ell_1,\dots,\ell_{k+1}$ *satisfying* 3.1. *This is immediate from theorem 3.3 with*

$$\sum_{n=0}^{\infty} n^{1+\varepsilon/k} < \infty.$$

4. LIAPUNOV FUNCTIONS AND SUCCESSIVE APPROXIMATIONS

In this section we first formulate sufficient conditions for stability in terms of Liapunov functions ℓ_1 and ℓ_2 (of order 1 and order 2 respectively).

LEMMA 4.1. *If some Liapunov function* ℓ_1 *(of order 1) exists and if in addition*

$$\lim_{n\to\infty} \tilde{U}^n \ell_1 = 0$$

then the problem is stable with respect to scrapfunctions $s \in V(\ell_1)$.

PROOF. Since $z \leq \ell_1$ we have $\lim_{n\to\infty} \tilde{U}^n z = 0$. By lemma 2.2, theorems 2.4 and 2.3 we have the desired result. □

LEMMA 4.2. *If Liapunov functions* ℓ_1 *and* ℓ_2 *exist, then*

$$\lim_{n\to\infty} \tilde{U}^n \ell_1 = 0.$$

PROOF. Consider a new reward structure: $\tilde{r}_P := \ell_1 - P\ell_1$, $P \in \mathcal{P}$. For all $R \in M$ we have

$$\ell_1 = \sum_{n=0}^{\infty} P_0 \ldots P_{n-1} \tilde{r}_{P_n} + \lim_{n\to\infty} P_0 \ldots P_n \ell_1.$$

Since

$$\ell_2 \geq \sum_{n=0}^{\infty} P_0 \ldots P_{n-1} \ell_1 \qquad \text{for all } R \in M,$$

we have

$$\lim_{n\to\infty} P_0 \ldots P_n \ell_1 = 0.$$

Hence ℓ_1 is the function y_1, defined in 1.7, for this new model. Therefore, by theorem 3.1, lemma 2.2 and theorem 2.1 we have the desired result. □

As a direct consequence of lemma's 4.1 and 4.2 we have

THEOREM 4.3. *If Liapunov functions* ℓ_1 *and* ℓ_2 *exist, then the problem is stable with respect to scrapfunctions* $s \in V(\ell_1)$.

We note that sometimes Liapunov functions ℓ_1 and ℓ_2 can be found in a rather simple way, while y_1 and y_2 are difficult to obtain.

REMARK 4.4. If we assume besides the existence of a first order Liapunov function ℓ_1, the compactness of $\mathcal{P}$ and the continuity of $P\ell_1$, as function of

P, then a sufficient condition for $\lim_{n\to\infty} \tilde{U}^n \ell_1 = 0$ is $\lim_{n\to\infty} P^n \ell_1 = 0$ for all $P \in \mathcal{P}$. The proof of this statement proceeds in a way similar to the proof of lemma 5.7 in [13].

THEOREM 4.5. *Let ℓ_1 and ℓ_2 be Liapunov functions(of order 1 and 2 respectively) and define for a function $s \in V(\ell_1)$*

$$b_1 := \inf\{\ell_1^{-1}(i)(Us-s)(i) \mid i \in E,\ \ell_1(i) > 0\}$$

$$b_2 := \sup\{\ell_1^{-1}(i)(Us-s)(i) \mid i \in E,\ \ell_1(i) > 0\}$$

then

$$s - b_1^- \ell_2 \le v \le s + b_2^+ \ell_2 \ . \tag{4.1}$$

PROOF. First observe that $s \in V(\ell_1)$ then also $Us \in V(\ell_1)$ so the set $\{i \mid \ell_1(i) = 0\}$ gives no trouble. Since $Us \le s + b_2\ell_1$ and $\ell_1 \le \ell_2$ we have

$$U^2 s \le \sup_P\{r_P + P(s+b_2\ell_1)\} \le Us + b_2^+ \tilde{U}\ell_1 \le Us + b_2^+ \tilde{U}\ell_2$$

$$\le s + b_2^+ \sup_P\{\ell_1 + P\ell_2\} \le s + b_2^+ \ell_2.$$

Similarly, from $U^k s \le s + b_2^+ \ell_2$ it follows that $U^{k+1} s \le s + b_2^+ \ell_2$, hence $U^n s \le s + b_2^+ \ell_2$ for $n = 1,2,\ldots$. Since the problem is stable (theorem 4.3) we have $v = \lim_{n\to\infty} U^n s \le s + b_2^+ \ell_2$. The proof of the left inequality is similar. □

The following, somewhat weaker, but more elegant inequality is now immediate.

$$\|v - s\|_{\ell_2} \le \|Us - s\|_{\ell_1}. \tag{4.2}$$

REMARK. If we have functions ℓ_1 and ℓ_2 satisfying the inequalities 3.1 but $\ell_2(i) = \infty$ for some i then we may separate the state space into $E_1 := \{i \in E \mid \ell_2(i) < \infty\}$ and $E_2 := E \backslash E_1$. Since $\ell_2(i) < \infty$ implies $\ell_2(j) < \infty$ for all $j \in E$ which can be reached under some strategy from state i, we have that ℓ_1 and ℓ_2 are Liapunov functions on the smaller model with state space E_1. Hence all results can be generalized to that situation.

If for some P, $r_P + Ps = Us$, $\|Us - s\|_{\ell_1}$ is small and $\ell_2 < \infty$ one may

use the stationary strategy R := (P,P,...). In section 7 th.7.2 we give bounds for the value of this strategy.

It is well-known that the β-discounted dynamic programming model

$$\Big(\sum_{j\in E} p(j|i,a) \le \beta < 1 \text{ for all } i \in E \text{ and } |r_p| \le Me \text{ for some } M \in \mathbb{R} \text{ and all } P \in \mathcal{P} \Big)$$

can be brought into our framework by defining an extra absorbing state -1 with $r(-1,a) = 0$ for all $a \in A$ and

$$p(-1 \mid i,a) = 1 - \sum_{j\in E} p(j \mid i,a), \qquad i \in E.$$

In this new model we can take as Liapunov functions the functions defined by $\ell_k(i) = M(1-\beta)^{-k}$, $i \in E$, $\ell_k(-1) = 0$, $k = 1,2$ and then 4.1 becomes slightly weaker than the MacQueen bounds [19] since we work with b_1^- and b_2^+ instead of b_1 and b_2.

In the following theorem s is an approximation for v with known bounds b_1 and b_2. At the price of extra calculation of $\tilde{U}^n(b_1)$ and $\tilde{U}^n(b_2)$ we obtain bounds for v_n^s. (Note that b_1 and b_2 are arbitrary bounds here.)

THEOREM 4.6. *If* $s - b_1 \le v \le s + b_2$ *then* $v_n^s - \tilde{U}^n b_1 \le v \le v_n^s + \tilde{U}^n b_2$.

PROOF. For n = 0 the statement is trivial. Suppose it holds for n = k. Then

$$v - v_{k+1}^s \le \sup_P\{r_p + Pv\} - \sup_P\{r_P + Pv_k^s\} \le \sup_P P\tilde{U}^k b_2 = \tilde{U}^{k+1} b_2$$

and

$$v_{k+1}^s - v \le \sup_P\{r_P + Pv_k^s\} - \sup_P\{r_P + Pv\} \le \tilde{U}^{k+1} b_1. \qquad \square$$

If there is a sequence $P_1, P_2, \ldots$ such that

$$\sup_P\{r_P + Pv_n^s\} = r_{P_n} + P_{n+1} v_n^s \qquad \text{for } n = 1,2,\ldots$$

then we can use $P_n P_{n-1} \ldots P_1 b_1$ instead of $\tilde{U}^n b_1$. Note that we may choose $b_2 \equiv 0$ if $s \ge v$.

Finally we can use these bounds to eliminate suboptimal actions. (We use the notation with explicitly written actions). Action a is called

suboptimal or *nonconserving* if

$$r(i,a) + \sum_j p(j \mid i,a)v(j) < \sup_{a\in A}\{r(i,a) + \sum_j P(j \mid i,a)v(j)\}$$

Hence if b_1 and b_2 are bounds on v, $b_1 \le v \le b_2$ it holds that action a is suboptimal if

$$r(i,a) + \sum_{j\in E} p(j \mid i,a)b_2 < \sup_{a\in A}\{r(i,a) + \sum_{j\in E} P(j \mid i,a)b_1\}.$$

Below we prove that elimination of some suboptimal actions gives a new model with the same value function. We only assume the model satisfies some strong convergence condition. In [14] a similar property is proved without this condition.

THEOREM 4.7. *Suppose that some strong convergence condition holds. Consider a new model with* $\tilde{P} \subset P$ *such that for all* $\varepsilon > 0$ *there is a* $P \in \tilde{P}$ *with* $r_P + Pv \ge v - \varepsilon$. *Then the new model has the same value function.*

PROOF. Fix $\varepsilon > 0$, let $\varepsilon_n := \varepsilon.2^{-(n+1)}$ e and choose $P_n \in \tilde{P}$ such that

$$r_{P_n} + \varepsilon_n + P_n v \ge v.$$

Iteration of this inequality yields

$$\sum_{n=0}^{N} P_0 \ldots P_{n-1}(r_{P_n} + \varepsilon_n) + P_0 \ldots P_N v \ge v.$$

Hence

$$\sum_{n=0}^{\infty} P_0 \ldots P_{n-1} r_{P_n} + \sum_{n=0}^{\infty} \varepsilon_n \ge v$$

since by the strong convergence condition $\lim_{n\to\infty} P_0 \ldots P_n v = 0$. Therefore

$$\sum_{n=0}^{\infty} P_0 \ldots P_{n-1} r_{P_n} \ge v - \varepsilon,$$

Consequently the supremum in this model equals v. □

As in [7] and [8] we can also exclude actions for a finite number of iterations instead of all future iterations. Fix some scrapfunction s. For notational convenience we omit the dependence on s in the following

definitions:

$$v_n(i,a) := r(i,a) + \sum_{j\in E} p(j \mid i,a) v^s_{n-1}(j)$$

$$d_n(i,a) := v^s_n(i) - v_n(i,a)$$

$$b_{1,n} := \inf_{i\in E}\{v^s_n(i) - v^s_{n-1}(i)\},\ b_{2,n} := \sup_{i\in E}\{v^s_n(i) - v^s_{n-1}(i)\}$$

$$\Phi_n := b_{2,n} - b_{1,n}.$$

THEOREM 4.8.

(i) $d_{n+k+1}(i,a) \geq d_n(i,a) - \sum_{\ell=0}^{k} \Phi_{n+\ell}$

(ii) *If* $d_n(i,a) - \sum_{\ell=0}^{k} \Phi_{n+\ell} > 0$ *then action* a *is suboptimal at stage* n+k+1.

PROOF. (ii) is a direct consequence of (i). Since

$$v_{n+1}(i,a) - v_n(i,a) = \sum_{j\in E} p(j) \mid i,a)\{v^s_n(j) - v^s_{n-1}(j)\} \leq b_{2,n}$$

and

$$v^s_{n+1}(i) - v^s_n(i) \geq \inf_{a\in A} \sum_{j\in E} p(j \mid i,a)\{v^s_n(j) - v^s_{n-1}(j)\} \geq b_{1,n}$$

we have by subtraction of these inequalities:

$$d_{n+1}(i,a) = v^s_{n+1}(i) - v_{n+1}(i,a) \geq d_n(i,a) - \Phi_n.$$

Iteration of this inequality yields the desired result. □

Hence, if we determine at stage n: $d_n(i,a)$ and at each following stage: Φ_{n+k}, we need not compute $v_{n+k+1}(i,a)$ as long as

$$d_n(i,a) - \sum_{\ell=0}^{k} \Phi_{n+\ell} > 0.$$

5. CONTRACTING DYNAMIC PROGRAMMING, STRONG CONVERGENCE AND LIAPUNOV FUNCTIONS

In this section we show how the contracting dynamic programming model introduced by van NUNEN [20] fits into the framework of strong convergence

and Liapunov functions. The model assumptions are as follows:
There exist a finite function b and a bounding function μ and there are constants $k,k' > 0$ and ρ,ρ' with $0 \leq \rho,\rho' < 1$, such that

$$\text{(i)} \quad \sup_R \sum_{n=0}^{\infty} \mathbb{E}_R[|b(X_n)|] < \infty$$

$$\text{(ii)} \quad \|r_P - b\|_\mu \leq k, \quad P \in \mathcal{P}$$

(5.1)

$$\text{(iii)} \quad P\mu \leq \rho\mu, \quad P \in \mathcal{P}$$

$$\text{(iv)} \quad \|Pb - \rho' b\|_\mu \leq k', \quad P \in \mathcal{P}.$$

In the papers of SHAPLEY [22], BLACKWELL [1] and DENARDO [3] it is assumed that the rewards are bounded and that the operator U (def.1.10) is a contraction with respect to the supremum norm. VEINOTT [23] showed that transient models can be transformed into discounted models using a similarity transformation which is equivalent to working with a bounding function (see below). HARRISON [6] noticed that in many practical models with a countable state space the reward function is unbounded and he suggested a modification: he introduced the translation function b. But he worked with $\mu \equiv 1$. LIPPMAN [17, 18] remarked that Harrison's model is too restrictive to include for example the M/M/1 queueing system with quadratic cost. He introduced a special bounding function: a polynomial. WIJNGAARD [25] considered exponential bounding functions to study inventory models with the average cost criterion. WESSELS [24] gave the first systematic treatment of general bounding functions for total return models with a countable state space. Van HEE and WESSELS [11] studied necessary and sufficient conditions for the existence of a bounding function μ such that for all $P \in \mathcal{P}$: $P\mu \leq \rho\mu$, $0 \leq \rho < 1$. HINDERER [12] used bounding functions for finite stage dynamic programming models with a general state space. We shall consider the contracting dynamic programming model in more detail. Let us denote

$$w_p := (1-\rho)^{-1}(b-Pb)$$

then by iteration, we find:

$$\sum_{n=0}^{N} P_0 \ldots P_{n-1} w_{P_n} + P_0 \ldots P_N (1-\rho)^{-1} b = (1-\rho)^{-1} b.$$

Since by 5.1(i), $\lim_{N\to\infty} P_0 \ldots P_N b = 0$, we have

$$\sum_{n=0}^{\infty} P_0 \ldots P_{n-1} w_{P_n} = (1-\rho)^{-1} b.$$

Hence the dynamic programming model with reward function $\tilde{r}_P = r_P - w_P$ for $P \in \mathcal{P}$ is equivalent to the original problem. However $\|\tilde{r}_P\|_\mu < \infty$. Indeed with 5.1 ii) and iv) we find

$$\|r_P\|_\mu = \|r_P - w_P\|_\mu = \|r_P - (1-\rho)^{-1} b + (1-\rho)^{-1} Pb\|_\mu =$$

$$= (1-\rho)^{-1} \|(1-\rho) r_P - b + Pb\|_\mu = (1-\rho)^{-1} \|(1-\rho)(r_P - b) - \rho b + Pb\|_\mu$$

$$\leq (1-\rho)^{-1} \{(1-\rho)\|r_P - b\|_\mu + \|Pb - \rho b\|_\mu\} < \infty.$$

Hence the contracting dynamic programming model is equivalent to a model satisfying for $P \in \mathcal{P}$ and some $k > 0$:

(5.2)
(i) $P\mu \leq \rho\mu$

(ii) $\|r_P\|_\mu \leq k.$

Note that this model can be reduced in a similar way to a discounted dynamic programming model by the transformations:

$$Q(i,j) := P(i,j)\mu(j)\mu(i)^{-1}, \quad \bar{r}_Q(i) := r_P(i)\mu(i)^{-1}.$$

This is in fact the similarity transformation studied by VEINOTT [23].
From 5.2(i) and ii) we have immediately

$$\sup_R \mathbb{E}_R[|r(X_n, Y_n)|] \leq \sup_R P_0 \ldots P_{n-1} |r_{P_n}| \leq k\rho^n \mu$$

and therefore, we have for $1 < \lambda < \rho^{-1}$

$$\sup_R \mathbb{E}_R\left[\sum_{n=0}^{\infty} \lambda^n |r(X_n, Y_n)|\right] \leq k(1-\lambda\rho)^{-1}\mu < \infty.$$

Thus the contracting dynamic programming model satisfies the strong convergence condition for the sequence $a_n \equiv \lambda^n$. And since $n^k = O(\lambda^n)$ $(n\to\infty)$ for all $k \geq 1$ we have by corollary 3.4 that there exist Liapunov functions ℓ_k

satisfying 3.1 for $k \geq 1$. Apart from this one immediately sees that

$$\mu + (1-\rho)^{-1}P\mu \leq (1-\rho)^{-1}\mu$$

thus

$$|r_p| + k(1-\rho)^{-1}P\mu \leq k(1-\rho)^{-1}\mu.$$

Hence $k(1-\rho)^{-1}\mu$ suffices as Liapunov function ℓ_1, and it is easily checked that $k(1-\rho)^{-n}$, $n \geq 1$ is a system of Liapunov functions satisfying 3.1.

6. WAITING LINE MODEL WITH CONTROLLABLE INPUT; AN EXAMPLE WHICH IS STRONGLY CONVERGENT BUT NOT NECESSARILY CONTRACTING

In this section we consider as an example the waiting line model with controllable input which was studied in chapter 5 in [13] and in [15]. In this queuing model the arrival process is Poisson with expected number of arrivals per unit time λ_a where a denotes the service cost. We assume that we can control the arrival process by choosing a from the interval $[a_1,a_2]$. And we make the reasonable assumption that λ_a decreases as a increases. The service time distribution F is general. At each time a customer completes service, the service cost may be changed. We will be looking at the embedded Markov chain.

The state space becomes $E = \{0,1,\ldots\}$ and the transition probabilities satisfy

$$p(j \mid i,a) = \begin{cases} 0 & \text{if } j < i-1 \\ k_{j-i+1}(a) & \text{if } j \geq i-1 \end{cases}$$

with

$$k_r(a) = \int_0^\infty e^{-\lambda_a s}(\lambda_a s)^r (r!)^{-1} dF(s).$$

Furthermore we assume

$$\lambda_{a_1} \int_0^\infty s\,dF(s) < 1$$

and $r(i,a) \geq \delta > 0$ for $i = 1,2,\ldots$ and all $a \in A := [a_1,a_2]$. If one is

looking for an average optimal strategy for this problem then one is interested in the behaviour of the system upto the first time the system empties again. In order to study the behaviour until this time we modify the transition probabilities and rewards in state 0 as follows

$$p(j \mid 0,a) := \delta_{0j} \quad \text{and} \quad r(0,a) := 0$$

If this model is contracting then there exists a bounding function μ satisfying

(i) $|r_P| \leq k\mu$ for some $k \in \mathbb{R}$ over all $P \in \mathcal{P}$

(ii) $P\mu \leq \rho\mu$, for some $0 \leq \rho < 1$ and all $P \in \mathcal{P}$.

Now (i) implies $|r(i,a)| \leq k\mu(i)$ and with $r(i,a) \geq \delta > 0$ follows $\mu(i) \geq \delta k^{-1}$, $i \geq 1$. Now we may use theorem 2 in [11] which states that there exists a function μ satisfying (ii) and $\inf_{i\geq 1} \mu(i) > 0$ if and only if the lifetime N of the process (here the number of transitions until state 0 is reached) is exponentially bounded. So in order that this model is contracting at least all moments of the life time must be finite and with the inequality

$$\mathbb{E}\, N(N-1)\ldots(N-k+1) \geq \sum_{\ell=k}^{\infty} \int_0^{\infty} \frac{(\lambda s)^{\ell} e^{-\lambda s}}{\ell!}\, \ell(\ell-1)\ldots(\ell-k+1)\, dF(s) = \lambda^k \int_0^{\infty} s^k dF(s)$$

(cf. [15]) we see that all moments of the service time must be finite as well. Hence the model is certainly not contracting if not all moments of the service time are finite. On the other hand it is shown in [15] that if the k-th moment of the service time is finite and if

$$\sup_a |r(i,a)| \leq Ai^{\ell}$$

for some $A \in \mathbb{R}$ and all $i \in S$ then there exist Liapunov functions $y_1, \ldots, y_{k-\ell}$, $\ell < k$. We will prove this here using a completely different approach. First one may show that if the k-th moment of the service time is finite then also the k-th moment of the lifetime of the embedded process is finite. This may be seen as follows. It is clear that the lifetime is maximized if we use the strategy $\bar{R}$ which corresponds to the minimal service cost in each state. For that strategy we have an M|G|1 queue. And the lifetime of the embedded process is now equal to the number of customers N in the busy period of the M|G|1 queue. And the lifetime of the embedded process

is now equal to the number of customers N in the busy period of the M|G|1 queue. Let F^* be the Laplace transform of the service time and N^* the transform of the distribution of the number of customers in a busy period. Then we have the following relation between F^* and N^*

$$N^*(t) = e^{-t}F^*(\lambda-\lambda N^*(t)), \qquad t > 0$$

where λ is the Poisson parameter (cf. COHEN [2] p.250). Differentiating this equation once with respect to t gives

$$N^{*'}(t) = \frac{-e^{-t}F^*(\lambda-\lambda N^*(t))}{1+\lambda F^{*'}(\lambda-\lambda N^*(t))} \quad . \tag{6.1}$$

The denominator is bounded from below by $1 - \lambda \int_0^\infty sdf(s) > 0$. It is well-known (see for example FELLER [5] p.412) that $N^{*(k)}(t)$ has a finite limit for $t \to 0$ if and only if

$$\sum_{n=0}^{\infty} n^k P(N=n) < \infty.$$

Then

$$\sum_{n=0}^{\infty} n^k P(N=n) = (-1)^k N^{*(k)}(0).$$

Differentiating (6.1) one may show by induction that if $F^{*(\ell)}(t)$ has a finite limit for $t \to 0$ for $\ell = 1,\ldots,k$ then $N^{*(k)}$ has a finite limit for $t \to 0$ as well. So we conclude that if the k-th moment of the service time is finite then also the k-th moment of the lifetime of the embedded process is finite. Now suppose

$$\mathbb{E}_{\tilde{R}} N^k < \infty \text{ and } \sup_P |r_P(i)| \le Ai^{k-m-1} \text{ for some } A \in \mathbb{R} \text{ and all } i \in E.$$

Then we have for all R

$$\mathbb{E}_R \sum_{\ell=0}^{\infty} \ell^m |r(X_\ell, Y_\ell)| = \sum_{t=0}^{\infty} \mathbb{P}_R(N=t) \sum_{\ell=0}^{t} \mathbb{E}_R \ell^m [|r(X_\ell, Y_\ell)| \mid N = t]$$

$$\le \sum_{t=0}^{\infty} \mathbb{P}_R(N=t) \sum_{\ell=1}^{t} t^m A t^{k-m-1}$$

$$= A \sum_{t=0}^{\infty} t^k \mathbb{P}_R(N=t) \le A \sum_{t=0}^{\infty} t^k \mathbb{P}_{\tilde{R}}(N=t) < \infty$$

Where the inequality

$$\sum_{\ell=0}^{t} \mathbb{E}_R[|r(X_\ell,Y_\ell| \ |N=t] \leq At^{k-m-1}$$

follows immediately from the fact that in the embedded process only one customer is served per unit of time.
So we see that

$$\int_0^\infty s^k F(s) < \infty \text{ and } \sup_P |r_P(i)| \leq Ai^{k-m-1}$$

for some $A \in \mathbb{R}$ and all $i \in E$ imply, using corollary 3.4, the finiteness of the functions $y_1,\ldots,y_m$. Reasoning in a similar way one may show that for $m = 0$ the model is strongly convergent (and thus $y_1 < \infty$).

7. NEARLY OPTIMAL STRATEGIES

In this section we collect some results with respect to nearly optimal strategies for the strongly convergent case. But before we do so we first give an example which shows that there need not exist for all $\varepsilon > 0$ a stationary strategy $P^{(\infty)}$ satisfying

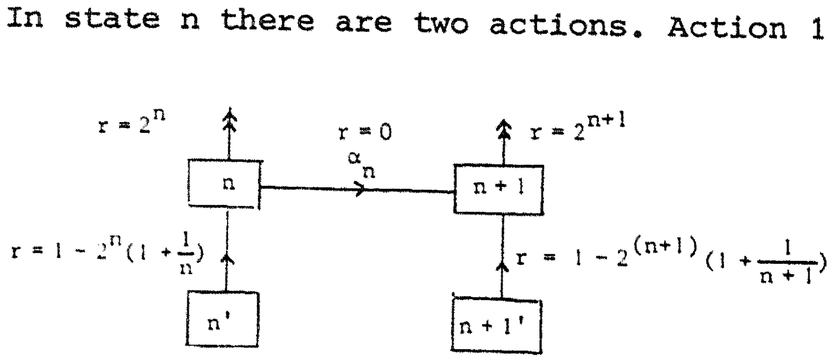
$$v(\cdot,P^{(\infty)}) \geq v - \varepsilon(1+|v|)e \tag{7.1}$$

if we only assume

$$\sup_R \mathbb{E}_R \sum_{n=0}^{\infty} |r(X_n,Y_n)| < \infty$$

but not 1.8, the uniform tail property or positivity of all $r(i,a)$. For the positive case ORNSTEIN [21] proved the existence of a $P^{(\infty)}$ satisfying 7.1.

EXAMPLE 7.1. $E\{1,1',2,2',\ldots\}$. In the states n' there is only one available action yielding an immediate reward $1 - 2^n(1+\frac{1}{n})$ and a transition to state n. In state n there are two actions. Action 1 gives reward 0 and a transition to state $n+1$ with probability $\alpha_n = b_n/2b_{n+1}$ where $b_n = 1 + n^{-1}$ and with probability $1-\alpha_n$ the system leaves E. Action 2 gives a reward 2^n and the system leaves E with probability 1. v may be found as follows.



$$v(n) = \sup(2^n, \alpha_n 2^{n+1}, \alpha_n\alpha_{n+1}2^{n+2}, \ldots) = 2^n \sup(1, \frac{b_n}{b_{n+1}}, \frac{b_n}{b_{n+2}}, \ldots)$$

$$= 2^n b_n = 2^n(1+\frac{1}{n}), \quad \text{since } b_n \downarrow 1 \text{ as } n \to \infty.$$

Thus $v(n') = 1 - 2^n(1+n^{-1}) + 2^n(1+n^{-1}) = 1$. We will show that there does not exist a stationary strategy $P^{(\infty)}$ for which $v(n',P^{(\infty)}) \geq 0$ for all $n \geq 1$. Any stationary strategy may be characterized by the probabilities γ_n by which action 2 is taken in state n, $n = 1,2,\ldots$. (We consider randomized strategies since when we were looking for an example we have seen that it may occur that though there is no pure ε-optimal strategy there does exist a randomized one). We see that for this strategy

$$v(n',R) \leq 1 - 2^n(1+\frac{1}{n}) + \gamma_n 2^n + (1-\gamma_n)2^n(1+\frac{1}{n}).$$

So strategy R gives for state n' an immediate loss of $\gamma_n 2^n/n$ compared to what could be gained. In order that this loss is smaller than 1 we must have $\gamma_n \leq n2^{-n}$. Now let us consider an arbitrary strategy R with $\gamma_n \leq n2^{-n}$ for all n and see what its total expected reward for state n is. Using the inequalities $\alpha_n \leq 2/3$, $1 - \gamma_n \leq 1$ and $\gamma_n \leq n2^{-n}$, $n = 1,2,\ldots$ we get

$$v(n,R) = 2^n\gamma_n + \alpha_n(1-\gamma_n)\gamma_{n+1}2^{n+1} + \alpha_n\alpha_{n+1}(1-\gamma_n)(1-\gamma_{n+1})\gamma_{n+2}2^{n+2}+\ldots$$

$$\leq 2^n n2^{-n} + \frac{2}{3}(n+1)2^{-(n+1)}2^{n+1} + (\frac{2}{3})^2(n+2)2^{-(n+2)}2^{n+2} + \ldots$$

$$= n + \frac{2}{3}(n+1) + \frac{4}{9}(n+2) + \ldots = 3n + 6.$$

So for $n \geq 4$ $v(n',R) \leq -1$. Hence no stationary strategy $P^{(\infty)}$ exists with $v(i,R) \geq v(i) - \varepsilon(1+|v(i)|)$ e for all $\varepsilon < \frac{1}{2}$. This concludes our counterexample.

Now we continue with some positive results.

If the model is strongly convergent then Howards' policy iteration algorithm converges. And as a result we conclude that in the strong convergence case it holds that for all $i \in E$ and all $\varepsilon > 0$ there exists a stationary strategy $P^{(\infty)}$ such that

(7.2) $\qquad v(i,P^{(\infty)}) \geq v(i) - \varepsilon.$ $\qquad$ (cf. [10]).

The following results deal with *uniform* ε-*optimality* on E, in some sense. If for a sequence $a = (a_0,a_1,\ldots)$ with $a_n \to \infty$ uniformly on E it holds that $z_a < \infty$, then for all $\varepsilon > 0$ there exists a stationary strategy $P^{(\infty)}$ such that

$$v(\cdot,P^{(\infty)}) \geq v - \varepsilon z_a \quad \text{(cf.[10])}.$$

And if $w_a/a_n \to 0$ uniformly on E then there exists for all $\varepsilon > 0$ a stationary strategy $P^{(\infty)}$ satisfying $v(\cdot,P^{(\infty)}) \geq v - \varepsilon e$ (cf.[10]).

THEOREM 7.2. *Let* ℓ_1 *and* ℓ_2 *be Liapunov functions of order* 1 *and* 2 *and let either* $s \in V(\ell_1)$ *or* $\limsup_{T\to\infty} P^T s \leq 0$. *If furthermore* $r_P + Ps \geq Us - \varepsilon\ell_1$ *then*

$$v(\cdot,P^{(\infty)}) \geq s - \varepsilon\ell_2.$$

PROOF. Iterating $r_P + Ps \geq s - \varepsilon\ell_1$ gives us

$$\sum_{n=0}^{T-1} P^n r_P + P^T s \geq s - \varepsilon \sum_{n=0}^{T-1} P^n \ell_1 \geq s - \varepsilon\ell_2.$$

Letting $T \to \infty$ yields the desired result. □

The next theorem presents a result under a partly weaker and partly stronger assumption than 1.8.

THEOREM 7.3. *If* $z < \infty$ *and*

$$\sup_{P\in\mathcal{P}} \sum_{n=1}^{\infty} nP^{n-1}r_P^- < \infty$$

then there exists for any state $i \in S$ *and for all* $\varepsilon > 0$ *a stationary strategy* $P^{(\infty)}$ *with*

$$v(i,P^{(\infty)}) \geq v(i) - \varepsilon.$$

PROOF. The proof proceeds analogous to the proof of theorem 13.6 in [13].

Fix $i \in E$ and $\varepsilon > 0$. Let strategy R be such that $v(i,R) \geq v(i) - \frac{\varepsilon}{4}$.

Choose $0 < \alpha < 1$ such that

$$\mathbb{E}_{i,R} \sum_{n=0}^{\infty} \alpha^n r(X_n,Y_n) \geq v(i) - \frac{\varepsilon}{4}$$

and

$$(1-\alpha)\sup_P \sum_{n=0}^{\infty} nP^n r_P^-(i) \le \frac{\varepsilon}{2} .$$

The α-discounted problem is strongly convergent, hence by 7.2, there exists a Q such that

$$\sum_{n=0}^{\infty} \alpha^n Q^n r_Q(i) \ge \sup_R \mathbb{E}_{i,R} \sum_{n=0}^{\infty} \alpha^n r(X_n,Y_n) - \frac{\varepsilon}{4} \ge v(i) - \frac{\varepsilon}{2} .$$

Since $1 - \alpha^n \le (1-\alpha)n$ for $0 < \alpha < 1$ and $n = 0,1,\ldots$ we have

$$\sum_{n=0}^{\infty} Q^n r_Q(i) \ge \sum_{n=0}^{\infty} \alpha^n Q^n r_Q(i) - \sum_{n=0}^{\infty} (Q^n - \alpha^n Q^n) r_Q^-(i)$$

$$\ge v(i) - \frac{\varepsilon}{2} - \sum_{n=0}^{\infty} (1-\alpha) n Q^n r_Q^-(i) \ge v(i) - \varepsilon.$$

Hence $v(i,Q) \ge v(i) - \varepsilon$. □

Finally a result on optimal strategies.

THEOREM 7.4. *If the model is strongly convergent then any conserving* P, i.e. $r_P + Pv = v$, *constitutes a stationary optimal strategy.*

PROOF. Iterating $r_P + Pv = v$ we get

$$\sum_{n=0}^{N-1} P^n r_P + P^N v = v.$$

Since

$$\sum_{n=0}^{N-1} P^n r_P \to v(\cdot,P^{(\infty)}) \quad (N\to\infty)$$

and $P^N v \to 0$ $(N\to\infty)$ (2.3) we have $v(\cdot,P^{(\infty)}) = v$. □

Hence if the model is strongly convergent, $\mathcal{P}$ compact, $w < \infty$, r_P and Pw continuous on $\mathcal{P}$ then there exists a stationary optimal strategy. Since with the compactness and continuity assumptions one may show the existence of a conserving P. See also chapter 4 in [13].

REFERENCES

[1] BLACKWELL, D., *Discounted dynamic programming,* Ann. Math. Statist. 36, 226-235, 1965.

[2] COHEN, J.W., *The single server queue,* North-Holland publishing Company, 1969.

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

[4] DERMAN, C. & R. STRAUCH, *A note on memoryless rules for controlling sequential control processes,* Ann. Math. Statist. 37, 276-278, 1966.

[5] FELLER, W., *An introduction to probability theory and its applications,* Vol. II, Wiley, New York, 1966.

[6] HARRISON, J., *Discrete dynamic programming with unbounded rewards,* Ann. Math. Statist. 43, 636-644, 1972.

[7] HASTINGS, N.A.J., *A test for nonoptimal actions in undiscounted finite Markov decision chains,* Management Science 23, 87-91, 1976.

[8] HASTINGS, N.A.J. & J.A.E.E. van NUNEN, *The action elimination algorithm for Markov decision processes,* this volume, 1977.

[9] HEE, K.M. van, *Markov strategies in dynamic programming,* Memorandum COSOR 75-20. University of Technology, Eindhoven, 1975.

[10] HEE, K.M. van, & J. van der WAL, *Strongly convergent dynamic programming,* Memorandum COSOR 76-26, University of Technology, Eindhoven, 1976.

[11] HEE, K.M. van, & J. WESSELS, *Markov decision processes and strongly excessive functions,* Memorandum COSOR 75-22, University of Technology, Eindhoven, 1975.

[12] HINDERER, K., *Bounds for stationary finite-stage dynamic programs with unbounded reward functions*, to be published.

[13] HORDIJK, A., *Dynamic programming and Markov potential theory,* Mathematical Centre Tracts No. 51, Amsterdam, 1974.

[14] HORDIJK, A., *Convergent dynamic programming,* Technical Report, Department of Operations Research, Stanford University, 1974.

[15] HORDIJK, A., *Regenerative Markov decision models*, to appear in Stochastic systems: Modeling, Identification and Optimization II, Mathematical Programming Studies 6, North-Holland, Amsterdam, 1975.

[16] HORDIJK, A. & K. SLADKY, *Sensitive optimality criteria in countable state dynamic programming*, to appear in Mathematics of Operations Research, 1975.

[17] LIPPMAN, S.A., *Semi-Markov decision processes with unbounded rewards*, Management Science 19, 717-731, 1973.

[18] LIPPMAN, S.A., *On dynamic programming with unbounded rewards*, Management Science 21, 1225-1233, 1975.

[19] MACQUEEN, J., *A modified dynamic programming method for Markovian decision problems*, J. Math. Anal. Appl. 14, 38-43, 1966.

[20] NUNEN, J.A.E.E. van, *Contracting Markov decision processes*, Mathematical Centre Tracts No. 71, Amsterdam, 1976.

[21] ORNSTEIN, D., *On the existence of stationary optimal strategies*, Proc. Amer. Math. Soc. 20, 563-569, 1969.

[22] SHAPLEY, L.S., *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A. 39, 1095-1100, 1953.

[23] VEINOTT, A.F., *Discrete dynamic programming with sensitive discount optimality criteria*, Ann. Math. Statist. 40 , 1635-1660, 1969.

[24] WESSELS, J., *Markov programming by successive approximations with respect to weighted supremum norms*, J. Math. Anal. Appl., 58, 326-335, 1977.

[25] WIJNGAARD, J., *Stationary Markovian Decision Problems*, Dissertation Eindhoven University of Technology 1975.

A SIMPLE BANDIT PROBLEM

J.A.Bather
University of Sussex, Brighton, England

1. INTRODUCTION

These notes provide a brief outline, without proofs, of some recent work. Consider two sequences of Bernoulli trials with probabilities p_1 and p_2 of success. The two-armed bandit problem is usually specified by assuming that p_1 and p_2 are both unknown and by fixing the total number of trials allowed with either process. The aim is to maximise the expected number of successes obtained.

2. RESULTS

The results here are mainly concerned with a special case. Let α be a known constant $\frac{1}{2} < \alpha < 1$ and write $\theta = \alpha/(1-\alpha) > 1$. Assume that $p_2 = \frac{1}{2}$. There are two simple hypotheses about p_1;

$$H^+ : p_1 = \alpha, \qquad H^- : p_1 = 1 - \alpha.$$

Thus, under H^+, it is preferable to carry out the trials on process 1, rather than process 2, but the opposite applies under H^-. A policy must determine, at each stage, which process is used for the next trial. For convenience, we shall use a criterion based on the number of "mistakes", i.e. trials on the process with the smaller probability of success.

Suppose that after n trials with process 1, the number of successes obtained on these trials is r and let $j = 2r - n$. Obviously, any trials on process 2 make no contribution to the information on H^+ versus H^-. In fact, this information can be represented by j alone. The likelihood ratio is θ^j and if we start with equal prior probabilites for the two hypotheses, the posterior probability of H^+ is

$$\Pi_j = \frac{\theta^j}{1+\theta^j} .$$

Using a Bayesian approach, let $R_j(t)$ be the minimum expected number of mistakes, given state j and t further trials with either process. Let $\phi_j(t)$ be the conditional probability of choosing process 1 at the next trial in a (possibly randomised) optimal policy. We have $R_j(0) = 0$ for $j = 0, \pm 1, \ldots$ and

$$(1) \qquad R_j(t) = \min \begin{cases} \phi[1 - \Pi_j + (\Pi_j\alpha+(1-\Pi_j)(1-\alpha))\ R_{j+1}(t-1) \\ \qquad\qquad + (\Pi_j(1-\alpha) + (1-\Pi_j)\alpha)R_{j-1}(t-1)], \\ (1-\phi)\ [\Pi_j+R_j(t-1)]. \end{cases}$$

We must set $\phi_j(t) = 1$ if the first of the terms in square brackets is smaller and $\phi_j(t) = 0$ otherwise. For simplicity, let $\phi_j(t) = 0$ in case of equality.

Results:

1.1. There is a sequence $k(1) = 0 \geq k(2) \geq \ldots \geq k(t) \to -\infty$, such that $\phi_j(t) = 1$ if $j > k(t)$, $\phi_j(t) = 0$ if $j \leq k(t)$.
It is enough to consider pure strategies: $\phi = 0$ or 1 in every case. The optimal policy always uses process 1 first, if at all, and then switches to process 2 at most once.

1.2. The limiting policy is trivial: $\phi_j(\infty) = 1$ for all j. In fact the limiting form of (1) as $t \to \infty$ has no solution. The dynamic programming equation is inappropriate for investigating policies over an infinite period.

BELLMAN [1] studied a more general problem in which p_1 is arbitrary, $0 < p_1 < 1$. The above results are special cases of his and he obtained similar results for the form of the optimal policy over an infinite number of trials, with a discount factor.

Even for the simple bandit problem, the optimal policy is quite complicated and it is sensible to consider various suboptimal policies. From now on, let $\phi = \phi_j(t)$ represent a prescribed policy with $0 \le \phi_j(t) \le 1$ always. The corresponding Bayes risk is $R_0(t) = \frac{1}{2}(U_0(t) + V_0(t))$, where $U_j(t)$ and $V_j(t)$ are the components of risk under H^+ and H^- respectively. In general, $U_j(0) = 0$ and

$$(2) \qquad U_j(t) = \phi_j(t)\,[\alpha U_{j+1}(t-1) + (1-\alpha)\,U_{j-1}(t-1)] + (1-\phi_j(t))\,[1+U_j(t-1)].$$

$V_j(0) = 0$ for all j and

$$(3) \qquad V_j(t) = \phi_j(t)\,[1+(1-\alpha)\,V_{j+1}(t-1) + \alpha V_{j-1}(t-1)] + (1-\phi_j(t))\,V_j(t-1).$$

We can also evaluate the error probabilities $\xi_j(t)$ and $\eta_j(t)$ under H^+ and H^-. These are associated with a given policy by demanding a final decision in favour of H^+ or H^- after t trials. For a suitable decision rule, we have

$$(4) \qquad \xi_j(0) = 0 \quad \text{if} \quad j > 0, \quad \xi_j(0) = 1 \quad \text{if} \quad j \le 0,$$

$$\xi_j(t) = \phi_j(t)\,[\alpha\xi_{j+1}(t-1) + (1-\alpha)\,\xi_{j-1}(t-1)] + (1-\phi_j(t))\,\xi_j(t-1)$$

$$(5) \qquad \eta_j(0) = 1 \quad \text{if} \quad j > 0, \quad \eta_j(0) = 0 \quad \text{if} \quad j \le 0,$$

$$\eta_j(t) = \phi_j(t)\,[(1-\alpha)\eta_{j+1}(t-1) + \alpha\eta_{j-1}(t-1)] + (1-\phi_j(t))\,\eta_j(t-1).$$

Different policies can be compared by examining the appropriate values of $\varepsilon_0(t) = \frac{1}{2}(\xi_0(t)+\eta_0(t))$. Note that the trivial policy, with $\phi_j(t) = 1$ always, has the property that $\varepsilon_0(t)$ is a minimum for every t.

ROBBINS [9] introduced a useful definition. A policy is asymptotically optimal (a.o.) if the proportion of mistakes tends to zero with probability 1 under either hypothesis. This means that the corresponding quantities

$$\frac{U_0(t)}{t}, \quad \frac{V_0(t)}{t} \quad \text{and} \quad \frac{R_0(t)}{t} \quad \text{all} \to 0 \text{ as } t \to \infty.$$

Stationary policies. Let $\phi = \phi_j$, not depending on t, $0 \leq \phi_j \leq 1$. The policy $\{\phi_j\}$ is pure if $\phi_j = 0$ or 1 for every j. It is worth considering whether any stationary policies have good behaviour in the long run.

Results:

2.1. No pure statonary policy can be a.o.

2.2. Let $\{\phi_j\}$ be a stationary policy such that $\phi_j > 0$ always, $\phi_j \to 1$ as $j \to \infty$ and $\phi_j \to 0$ as $j \to -\infty$. Then it is a.o.

Thus, randomisation is important for good long-term properties of a stationary policy. However, it is not easy to choose a particular policy from those indicated by 2.2. Intuitively, we should demand that

(i) $\phi_{j+1} \geq \phi_j$ for all j,

(ii) $\phi_j = 1$ for $j > 0$.

2.3. Let $\phi_j = 1$ for $j > 0$ and let $U_j(t)$, $V_j(t)$ be the corresponding components of risk. Let $\{\phi_j^*\}$ be any other stationary policy with risk functions such that

$$U_0^*(t) \leq U_0(t) \quad \text{and} \quad V_0^*(t) \leq V_0(t) \quad \text{for all} \quad t.$$

Then $\phi_j^* = \phi_j$ for every j.

In other words (ii) ensures that none of the quantities $U_0(t)$, $V_0(t)$, $t = 1,2,\ldots$, can be reduced without increasing another of them.

A further result is concerned with the possibility of minimising the Bayes risk $R_0(t)$ within the class of stationary policies.

2.4. Let $\{\phi_j\}$ be a stationary policy with $\phi_j > 0$ for each j. Then $R_0(t)$ does not attain a minimum value for any particular $t \geq 2$.

This applies to a.o. policies and it can be deduced that there is no stationary policy which minimises $R_0(t)$ for all sufficiently large t.

3. COMPUTATIONS

If we are interested in large values of t, or if t is an unknown parameter as must often be the case, there is a large class of stationary policies which may be useful. The table below examines the behaviour of just two a.o. policies over t = 50 and t = 100 trials. The computations are based on relations (1),...,(5). We recall that $\theta = \alpha/(1-\alpha)$, so that $\alpha = \theta/(1+\theta)$.

The tabulated values for the Bayes risk $R_0(t)$ and the average error probability $\varepsilon_0(t)$ give some idea of the relative merits of five different policies. Note that $R_0(t)$ is minimised by policy (a), whereas $\varepsilon_0(t)$ is minimised by (e). The table illustrates the incompatibility between these two different criteria. The optimal policy (a) depends on the value of the parameter θ but computations not included here indicate that the effect of using incorrect values of θ in the determination of the policy is relatively small. The effect of misjudging the total number of trials is illustrated by (d). This is more substantial. Policy (c) is considerably better than (b). It compares reasonably well with the optimal policy, especially when one takes into account the gain in simplicity.

Policy	θ	t = 50				t = 100			
		U_0	V_0	R_0	ε_0	U_0	V_0	R_0	ε_0
(a)	2	3.59	10.59	6.99	0.0500	4.25	13.21	8.73	0.0252
(b)	2	5.02	12.43	8.73	0.0448	7.41	15.52	11.46	0.0218
(c)	2	2.14	13.93	8.04	0.0279	3.79	15.80	9.80	0.0160
(d)	2	23.51	2.99	13.25	0.2503	8.49	10.42	9.45	0.0492
(e)	2	0.00	50.00	25.00	0.0079	0.00	100.00	50.00	0.0003
(a)	1.5	7.59	16.26	11.92	0.1391	10.57	23.53	17.05	0.0780
(b)	1.5	10.17	16.95	13.56	0.1468	18.48	22.83	20.65	0.0956
(c)	1.5	5.83	19.76	12.80	0.1177	12.40	24.46	18.43	0.0779
(d)	1.5	30.22	4.67	17.45	0.3353	20.30	17.06	18.68	0.1334
(e)	1.5	0.00	50.00	25.00	0.0776	0.00	100.00	50.00	0.0220
(a)	1.1	16.74	26.19	21.46	0.3945	30.04	50.07	40.05	0.3518
(b)	1.1	18.97	25.00	21.98	0.3996	42.23	41.05	41.64	0.3713
(c)	1.1	13.91	29.54	21.73	0.3872	35.32	46.13	40.73	0.3593
(d)	1.1	37.66	8.58	23.12	0.4579	47.87	34.19	41.03	0.3896
(e)	1.1	0.00	50.00	25.00	0.3688	0.00	100.00	50.00	0.3173

Policies:
(a) Optimal policy $\phi = \phi_j(t)$ determined from the solution of (1).
(b) Stationary policy $\phi = \phi_j$, $\phi_j = 1$ for $j \geq 0$, $\phi_j = 2^j$ for $j < 0$.
(c) Stationary policy $\phi = \phi_j$, $\phi_j = 1$ for $j \geq -2$, $\phi_j = 4^{j+2}$ for $j < -2$.
(d) Lagged application of optimal policy $\phi = \phi_j(t-50)$ for $t > 50$, $\phi = \phi_j(1)$ for $t \leq 50$.
(e) Trivial policy $\phi = 1$ always.

REMARKS

The general two-armed bandit problem mentioned earlier has received a good deal of attention in the literature. Optimal policies can be determined, in principle, by assuming prior distributions for the unknown p_1 and p_2. Such policies are much more complicated than those discussed here and the danger of critical dependence on the total number of trials allowed and the prior distributions seems greater: see Fabius and Van Zwet [3]. On the other hand, reasonably simple stationary policies can be devised which are asymptotically optimal. Of course, much more extensive computations will be needed to discover really useful ones.

ACKNOWLEDGEMENT

I am grateful to a colleague, Hilary Stevens, for carrying out most of the computations.

REFERENCES

[1] BELLMAN, R. (1956), *A problem in the sequential design of experiments*, Sankhya 16, 221-229.

[2] BRADT, R.N., S.M. JOHNSON & S. KARLIN (1956), *On sequential designs for maximising the sum of* n *observations*, Ann. Math. Statist. 27, 1060-1074.

[3] FABIUS, J. & W.R. VAN ZWET (1970), *Some remarks on the two-armed bandit*, Ann. Math. Statist. 41, 1906-1916.

[4] FELDMAN, D. (1962), *Contributions to the two-armed bandit problem*, Ann. Math. Statist. 33, 847-856.

[5] GITTINS, J.C. & D.M. JONES (1973), *A dynamic allocation index for the sequential design of experiments*. Proceedings of the European Meeting of Statisticians, Budapest 1972. Hungarian Academy of Sciences.

[6] HOEL, D.G., M. SOBEL & G.H. WEISS (1972), *A two-stage procedure for choosing the better of two binomial populations*, Biometrika 59, 317-322.

[7] ISBELL, J.R. (1959), *On a problem of Robbins,* Ann. Math. Statist. 30, 606-610.

[8] NASH, P. (1973), *Optimal allocation of resources between research projects,* Ph.D. thesis. University of Cambridge.

[9] ROBBINS, H. (1952), *Some aspects of the sequential design of experiments,* Bull. Amer. Math. Soc. 58, 527-535.

[10] ROBBINS, H. (1956), *A sequential decision problem with finite memory,* Proc. Nat. Acad. Sci. USA 42, 920-923.

[11] ROBBINS, H. & D.O. SIEGMUND (1974), *Sequential tests involving two populations,* J.A.S.A. 69, 132-139.

[12] TAHERI, H. & D.H. YOUNG (1974), *A comparison of sequential sampling procedures for selecting the better of two populations,* Biometrika 61, 585-592.

[13] ZELEN, M. (1969), *Play the winner rule and the controlled clinical trial,* J. Amer. STatist. Assoc. 64, 131-146.