Markov decision processes with unbounded rewards

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 $S := \{1,2,3,\dots\}$. The system can be controlled at discrete points in time $t = 0,1,2,\dots$ 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) $$p^a(i,j) \geq 0, \ \sum_j p^a(i,j) \leq 1, \qquad \text{for all } i,j \in S \text{ and all } a \in A.$$

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 $\mathcal{RM}$.

(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 $\mathcal{F}$ 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 $\mathcal{M}$.

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 $\{(X_t, Z_t)\}_{t=0}^{\infty}$ 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.

<u>ASSUMPTION 2.2.</u> 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

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

<u>ASSUMPTION 2.3.</u>

(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 $\bar{r}(i) := \sup_{a \in A} r(i,a)$.

<u>REMARK 2.3.</u> 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 $\pi \in \mathcal{M}$

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

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 $\sup_{\pi \in \mathcal{M}} v_i(\pi) < \infty$. In [9] VAN HEE shows that under assumption 2.3.i $v_i(\pi)$ is properly defined for all $\pi \in \mathcal{RM}$ since

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

Moreover, he proves that

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

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

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

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

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

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

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 $v \le v_0$.

<u>LEMMA 3.1.</u>

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

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

<u>PROOF.</u>

$$\begin{aligned} P(f_2)P(f_1)\bar{r} &\le P(f_2)(\rho\bar{r} + M_1\mu) \\ &\le \rho^2\bar{r} + \rho M_1\mu + \rho_* M_1\mu \\ &\le \rho^2\bar{r} + 2\rho_0 M_1\mu \end{aligned}$$

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

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

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

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* $\|P(f)\| \le \rho_* < 1$. *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.

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

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

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

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) $\qquad v(\pi) \le v^*$ *for all* $\pi \in \mathcal{D}$

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

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

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

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

then

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

hence

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

iterating this inequality gives

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

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* $Uv' - \delta\mu \le v$ *then*

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

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

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

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

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

or

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

Similarly

$$\begin{aligned} U^2v &\le U(v+\varepsilon\mu) = U(v'+v-v'+\varepsilon\mu) \\ &\le Uv' + \rho_* \|v - v'\|\mu + \rho_*\varepsilon\mu \\ &\le v + \delta\mu + \rho_*\|v - v'\|\mu + \rho_*\varepsilon\mu = v + \varepsilon(1+\rho_*)\mu. \end{aligned}$$

Iterating in the same way gives

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

This implies

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

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

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

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.

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

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.

<u>REMARK 5.1.</u>

(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 $\beta \le 1$) 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 $\alpha \ge 0$ 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

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

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

similarly

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

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

Hence, we have

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

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

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 $\tilde{v}_i(\pi)$.

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

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

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

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

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

where the third equality is allowed since

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

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.

<u>CONDITIONS OF LIPPMAN</u>. 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 $i \in S$, $a \in A$

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

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

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) $\|r(f)\| \le M.$

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.

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

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* $\mu(i) \ge \delta > 0$ *if and only if the lifetimes of the process are exponentially bounded, i.e.*

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

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 $P(f)\mu \le \rho_* \mu$.
"only if" Note that for $\pi := (f_0, f_1 \ldots)$

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

with $e := \{1,1,\ldots\}$. □

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

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

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*

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

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* $\pi \in M$

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

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

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

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

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

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.