Successive approximations for Markov decision processes and Markov games with unbounded rewards

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 78-06

Successive approximations for Markov decision processes and Markov games with unbounded rewards

by

Jo van Nunen and Jaap Wessels

Eindhoven, February 1978

The Netherlands

Successive approximations for Markov decision processes and Markov games with unbounded rewards

by

Jo van Nunen* and Jaap Wessels**

Summary: The aim of this paper is to give an overview of recent developments in the area of successive approximations for Markov decision processes and Markov games. We will emphasize two aspects, viz. the conditions under which successive approximations converge in some strong sense and variations of these methods which diminish the amount of computational work to be executed. With respect to the first aspect it will be shown how much unboundedness of the rewards may be allowed without violation of the convergence.
With respect to the second aspect we will present four ideas, that can be applied in conjunction, which may diminish the amount of work to be done. These ideas are: 1. the use of the actual convergence of the iterates for the construction of upper and lower bounds (McQueen bounds), 2. the use of alternative policy improvement procedures (based on stopping times), 3. a better evaluation of the values of actual policies in each iteration step by a value oriented approach, 4. the elimination of suboptimal actions not only permanently, but also temporarily.
The general presentation is given for Markov decision processes with a final section devoted to the possibilities of extension to Markov games.

* Graduate School of Management, Delft, The Netherlands.
** Eindhoven University of Technology (dept. of Math.), Eindhoven, The Netherlands.

1. Introduction

In recent years quite a lot of research effort has been dedicated to successive approximations methods in Markov decision processes and Markov games for the total expected reward as well as for the average reward criterion. In this paper we only consider the total expected reward criterion. The reasons for this attention are both theoretical and numerical. It appeared that from the theoretical point of view the successive approximations approach gives a basic understanding of the processes involved. From the numerical point of view, it turned out that the more sophisticated methods like policy iteration and linear programming are not suitable for very large problems. Furthermore, it appeared that the policy iteration method and therefore the strongly related (see e.g. [21],[46]) linear programming approach, is essentially an extreme example of a successive approximations method (see section 5).

In this paper we will give a review of recent developments with regard to successive approximations for Markov decision processes and Markov games with the total expected reward criterion. First we will be concerned with the conditions under which successive approximations converge, in some uniform sense, to the value of the decision problem (section 2). In section 3 we investigate these conditions further. For the sake of simplicity we initially treat the whole theory for Markov decision problems only and consider the extensions to Markov games later on (section 7).
Essential for our conditions is that unbounded rewards are to a certain extent allowed. Furthermore, we do not require strict discounting. Actually-as will appear in section 2 and will be worked out further in section 3 - our conditions give a joint restriction on the allowed unboundedness of the rewards and the drift/fading of the system (or equivalently: the uniformity of discounting). With our conditions we combine the shift approach of Harrison [5] with the weighted supremum norm approach of Wessels [43]. For countable state space and arbitrary action space this combination has been presented first by van Nunen in his monograph [22]. A slight generalization has been given by the present authors in [25]. A generalization to more general state spaces has been given by Couwenbergh in [2] and [3]. In order to avoid measure theoretic and topological complexities, we will only treat the countable state space situation in this paper.

Besides the conditions for convergence of successive approximations, we will also be concerned with ways to diminish the amount of work required for the computation of good strategies and good estimates of the value function. We will present essentially four work saving ideas. The first one already appears in section 2 and consists of using the subsequent approximations for the construction of upper and lower bounds. Compared with the conventional estimates (using only the geometric convergence rate), these generalizations of the Mac Queen-bounds [19] accelerate the convergence considerably, without requiring more work per iteration step. The second device which can be used in order to accelerate convergence is an alternative policy improvement step which can be defined with a stopping time. In section 4 it is shown how different stopping times generate different successive approximation methods. This stopping time approach has been presented first in [44] and has been generalized and refined in [22]; [26] gives a short overview together with some new points of view. In section 5 it is demonstrated how all these successive approximation methods can be refined by the introduction of a better value estimation for the current policy in each iteration step. Here all types of policy iteration methods appear to be extreme cases of successive approximations procedures. This value oriented approach has been introduced in [23], generalized in [27] , and further generalized in [22], see also [26]. There is still one other idea to exploit (section 6), viz. the elimination of suboptimal actions. In [20] MacQueen used his upper and lower estimates for the value function for the detection of actions which cannot be optimal. Namely, in each iteration step some actions may be eliminated resulting in less work during the remaining iteration steps. This idea can be adopted to our more general conditions and to our alternative procedures (see [22]).
However, it is also possible to eliminate actions only temporarily. Based on an idea of Hastings (see [7]), this has been established in [8] for finite actions, finite state discounted Markov decision problems. This will be demonstrated for our situation in section 6 (an example will be included).

Finally, in section 7, all these features are reconsidered for Markov games. It appears that all the ideas allow some sort of generalization to the zero-sum game situation. A striking point is that the standard successive approximations approach for Markov games is older than the analogous (but actually more specific) approach for Markov decision processes (see Shapley [38]). For Markov games the MacQueen bounds have been introduced in [40] by van der Wal. In the same paper the stopping time approach for introducing alternative procedures has been given. The more general convergence

conditions have been given in [45]; Further generalizations of the conditions- viz. to noncountable state spaces- may be found in the papers [2], [3] by Couwenbergh. The value oriented approach has been given in [41] again by van der Wal. The elimination of suboptimal actions has been treated in [34]. For a partial survey of these results (together with some other topics) we refer to [42].

2. Markov decision processes with unbounded rewards

We will first introduce our Markov decision process.

A system is observed at discrete points of time ($t = 0,1,2,\ldots$). The state of the system at any time t is an element of the countable state space $S:=\{1,2,\ldots\}$. If at time t the state of the system is $i \in S$, an action a may be chosen from a given arbitrary set A, which incurs a reward $r(i,a)$. The current state i at time t and the action a determine the probability $p^a(i,j)$ of observing the system in state j at time $t + 1$ (regardless of the earlier history of the process). We suppose

$$\sum_{j\in S} p^a(i,j) \leq 1 \quad \text{for all } i \in S,\ a \in A.$$

Hence a positive probability for fading of the system is allowed.
A policy f is a map from S into A. A strategy π is a sequence of policies: $\pi=(f_0,f_1,\ldots)$. If strategy π is used, we choose action $f_t(i)$ when the system is in state i at time t. The set of all policies is denoted by F, the set of all strategies by M. A stationary strategy consists of equal policies $\pi = (f,f,\ldots)$, so we actually may use the terms policy and stationary strategy deliberately.
As optimality criterion we choose total expected rewards, which is defined (if the sum converges absolutely) for a strategy $\pi = (f_0,f_1,\ldots)$ by

$$v(\pi) := \sum_{t=0}^{\infty} \{ \prod_{n=0}^{t-1} P(f_n)\}\, r(f_t) ,$$

where $r(f)$ is the column vector with i-th component $r(i,f(i))$, $P(f)$ is the matrix with (i,j)-component $p^{f(i)}(i,j)$ and empty products of matrices are equal to the identity matrix I. Matrix products, matrix-vector products and sums of vectors are defined in the usual way. Hence $v(\pi)$ is a column vector with i-th component: the total expected rewards under stragey π, if the process starts in i.

Remarks:

a. Actually, we only introduced the so-called nonrandomized Markov strategies. It would be very well possible to work with more general types of strategies allowing e.g. actions based on the complete history of the process and mixing of actions. However, under the convergence conditions we need anyhow for our theory it is not necessary to consider these more complicated strategies. This can be proved easily using a basic theorem of van Hee (see [9]), as has been demonstrated in [25] for a somewhat more general situation than we have in this paper.

b. This set-up contains (semi)-Markov decision processes with discounting, since the resulting discount factors may be incorporated in the transition probabilities. This approach leads to the fact that probabilities not necessarily sum to one, as mentioned in the beginning of this section (see e.g. [22] section 9.1).

Supposing for the moment that $v(\pi)$ is properly defined for all strategies $\pi \in M$, we may state the aim of the decision maker:

find $\pi^* \in M$, such that

$$v(\pi^*) = \sup_{\pi \in M} v(\pi) =: v,$$

or, if the sup is not attained, a π° is asked for such that $v(\pi^\circ)$ approximates v in some sense.

The key to the solution of this problem is the so-called optimality equation (which holds under suitable conditions to be specified in the sequel):

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

where the sup is taken componentwise.

Analogously to linear equation systems and linear integral equations, a standard approach for solving such an equation is to use contraction properties of a suitably chosen operator. For our kind of equations, this approach has been initiated by Blackwell in [1] and extended by Denardo in [4]. In the following it will be shown how this approach can be generalized further and how it can be used to find relatively good extrapolations for v and $v(\pi^\circ)$.

We will first introduce the assumptions on the transition probabilities and the rewards. Therefore we assume a positive function μ on S to be given (μ and μ^{-1}, the function with values $\mu^{-1}(i)$, will also be interpreted as column vectors). 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 B (real valued functions on $S \times S$) we introduce the corresponding operator norm:

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

Using this, our main assumptions become:

(i) a. there is a number $m > 0$, such that for all policies $f \in F$:

$$\|r(f) - \bar{r}\| \leq m,$$

where $\bar{r}$ is the vector with $\bar{r}(i) := \sup_{a\in A} r(i,a)$ or $\bar{r} = \sup_{f\in F} r(f)$.

b. $\sup_{\pi\in M} \sum_{t=0}^{\infty} \{ \prod_{n=0}^{t-1} P(f_n) \} \cdot |\bar{r}| < \infty$ (componentwise)

where $|\bar{r}|$ is the vector with components $|\bar{r}(i)|$.

(ii) $\rho_+ := \sup_{f\in F} \|P(f)\| < 1$, hence $P(f)\mu \leq \rho_+ \mu$ for all $f \in F$.

(iii) $m_1 := \sup_{f\in F} \|P(f)\bar{r} - \rho\bar{r}\| < \infty$ for some ρ with $0 < \rho < 1$.

In section 3 we will consider the question in what situations a function μ exists such that (i), (ii) and (iii) hold. For the moment we will take these assumptions for granted. Note that ρ and ρ_+ in the assumptions are not necessarily equal.

From these assumptions it follows not only that P(f) is a properly defined operator on W, but also that P(f) is monotone and contracting on W with contraction radius $\|P(f)\| \leq \rho_+ < 1$.

Since $r(f) + P(f)w$ is not necessarily in W, we define a translated space U of W on which $r(f) + P(f)u$ is a properly defined operator: U is the set of vectors u (real valued functions on S) which satisfy $u - (1-\rho)^{-1}\bar{r} \in W$. As W is a Banach space, the space U is completely metric with the distance $\|u - u'\|$ between u and $u' \in U$.

Now one easily verifies the following properties (compare [25]):

Lemma 2.1: (i) For any $f \in F$, the mapping $L(f)$ on U, defined by

$$L(f)u := r(f) + P(f)u,$$

maps U into U.

(ii) $L(f)$ is monotone and contracting on U with contraction radius $\|P(f)\| \leq \rho_+ < 1$.

(iii) Being contracting, $L(f)$ has a unique fixed point in U. This fixed point is $v(f)$, the total expected reward vector of the stationary strategy $(f,f,f,\ldots)$.

(vi) If $u \in U$, then $L^n(f)u$ converges to $v(f)$ if $n \to \infty$.

However, for application of the successive approximations idea to the optimality equation, it is not the operator $L(f)$ that counts, but the operator T on U defined by

$$Tu := \sup_{f \in F} L(f)u.$$

Using the properties of $L(f)$ one verifies for T (see [25]):

Lemma 2.2: (i) T maps U into U.

(ii) T is monotone and contracting on U with contraction radius $\gamma \leq \rho_+ < 1$.

(iii) Being contracting, T has a unique fixed point in U. This shows that the optimality equation $u = \sup_{f \in F} \{r(f) + P(f)u\}$ has a unique solution in U.

(iv) The unique fixed point of T in U, and hence the unique solution of the optimality equation in U, is $v = \sup_{\pi \in M} v(\pi)$.

(v) If $u \in U$, then $T^n u$ converges to v if $n \to \infty$.

The proofs of point (i), (ii) and (v) being straightforward, we only sketch the proof of (iv):

Suppose the fixed point of T is $u^* \in U$, then we have for any policy f_0 : $u^* \geq r(f_0) + P(f_0)u^* = L(f_0)u^*$

This implies for any strategy $\pi = (f_0, f_1, \ldots)$ by iteration

$$u^* \geq L(f_0)L(f_1)\ldots L(f_t)u^*.$$

By taking $t \to \infty$ the left hand side converges to $v(\pi)$, so $v(\pi) \leq u^*$ for any π, Hence $v \leq u^*$.

That $u^* \leq v$ can be shown by taking a policy f with $u^* \leq r(f) + P(f)u^* + \varepsilon\mu$

for an arbitrary $\varepsilon > 0$. By iterating this inequality one shows

$$u^* \le v(f) + \varepsilon(1 - \rho_+)^{-1}\mu.$$

The proof of lemma 2.2 (iv) exhibited above actually shows more, viz. the existence of an ε-optimal stationary strategy ($\|v(f) - v\| \le \varepsilon$) for any positive ε. If the sup in the optimality equation is attained by some f^* for $u = v$, then ε can be taken zero and the stationary strategy f^* is optimal.

Since T is contracting with contraction radius γ, $T^n u$ converges to v geometrically. This can be used to give upper and lowerbounds for v based on $T^n u$. However, using the actual convergence of the iterates $T^n u$ much better bounds can be given without much extra work. These bounds are based on the following properties:

Lemma 2.3: Let $\delta > 0$, $u, u' \in U$, $f \in F$:

(i) if $Tu' - \delta\mu \le u$, then $v \le u + \dfrac{\delta + \rho_+\|u - u'\|}{1 - \rho_+}\mu$.

(ii) if $L(f)u' = u$, then $u + \dfrac{\rho_f^-\|u - u'\|_-}{1 - \rho_f^-}\mu \le v(f) \le u + \dfrac{\rho_+\|u - u'\|}{1 - \rho_+}\mu$

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

Proof: The proof of (ii) is similar, but simpler, as the proof of (i) which will be sketched:

$Tu = T(u' + u - u') \le Tu' + \rho_+\|u - u'\|\mu \le u + \delta\mu + \rho_+\|u - u'\|\mu$.

Hence $Tu \le u + \varepsilon\mu$, with $\varepsilon = \delta + \rho_+\|u - u'\|$.

Repeating this argument we obtain

$$T^n u \le u + \varepsilon(1 + \rho_+ + \ldots + \rho_+^{n-1})\mu.$$

Taking the limit for $n \to \infty$ we obtain the required inequality.

These properties make it possible to construct an algorithm which generates vectors $u^{(n)}$ and policies f_n such that $u^{(n)}$ converges monotonically to v and producing the following bounds:

$$u^{(n)} + \frac{\rho_{f_n}^-\|u^{(n)} - u^{(n-1)}\|_-}{1 - \rho_{f_n}^-}\mu \le v(f_n) \le v \le u^{(n)} + \frac{\delta + \rho_+\|u^{(n)} - u^{(n-1)}\|}{1 - \rho_+}\mu.$$

This can be obtained by choosing $\delta > 0$, $u^{(0)} \in U$ with $u^{(0)} \leq Tu^{(0)}$ and determining f_n $(n = 1,2,\ldots)$ such that

$$u^{(n)} := L(f_n)\, u^{(n-1)} \geq \max \{u^{(n-1)}, Tu^{(n-1)} - \delta\mu\}.$$

If δ satisfies $\delta < \alpha(1 - \rho_+)$ for some chosen $\alpha > 0$, then the upper and lower bound will differ at most $\alpha\mu$ for some finite n. For details see [43], [22],[25].

3. Analysis of the assumptions

We will first make some miscellaneous remarks on the assumptions.

a. $\bar{r}$ may be replaced by any vector b with $b - r(f) \in W$ for some f, so it is not necessary to compute $\bar{r}$ exactly (see [22]).

b. the translation function $(1 - \rho)^{-1}\, \bar{r}$ - as introduced in a slightly different way by Harrison in [5] - can be related to an approach by Porteus in [31]. Porteus introduced his so-called return transformation which transforms the original problem into an equivalent problem satisfying $\|r(f)\| \leq m_0$, $\|P(f)\| \leq \rho_+$, which can be treated in W itself (for details see [25]).

c. In [18], Lippman presents conditions for problems with unbounded rewards to allow convergent successive approximations. As we have proved in [28], the models for which Lippman's approach works are covered by our approach (in fact the translation is not necessary, so even the conditions of [43] are satisfied). As demonstrated in [28], our conditions are easier to verify.

d. As remarked in the introduction the conditions of section 2 have been weakened somewhat in [25] by exploiting the fact that very bad policies cannot influence convergence essentially.

e. For convergence of successive approximations it is not necessary to have contraction. However, the contraction gives elegant and efficient estimates. For treatment of the convergence problem without contraction we refer to Schäl [36], Couwenbergh [2] and the papers[10], [11] by van Hee, Hordijk, van der Wal.

f. Another possible extension is to weaken the countability of the state space. This has been executed by Schäl in [36] and by Couwenbergh in [2], [3]. The main new difficulties in that case are measurability problems. These problems are solved by using selection theorems.

g. A similar theory as given in section 2 can be given for finite-stage problems. Then strict contraction (i.e. $\rho_+ < 1$) is not necessary. For an overview and further references see [13] by Hinderer and Hübner.

h. In the procedure at the end of section 2, the requirement $u^{(0)} < Tu^{(0)}$ is not essential. It only makes a definition of the sequence $u^{(n)}$ possible such that this sequence is nondecreasing. Here and in later variants we only require this monotonicity for reasons of elegance and - in some cases - simplicity of the proofs.

In the sequel of this section we will discuss the assumptions of section 2, so in this part we will not presuppose them.
Our assumption (ii) requires some sort of transient behaviour of the process involved. Assumption (ii) may be written as

$$P(f)\mu \leq \rho_+ \mu \quad \text{for all } f \in F, \text{ with } \rho_+ < 1.$$

In this form the function μ is clearly required to satisfy some sort of strong excessivity property (for excessive functions in Markov decision theory see Hordijk's monograph [15]). If such a strongly excessive and positive function μ exists, we will call the Markov decision problem strongly excessive.
In the following lemmas we demonstrate the relation between strong excessiveness and the transient behaviour of the process. In order to do so, we denote by $\mathbb{P}_i^\pi(\cdot)$ the probability of some event given that the strategy π is used and the process starts in i. X_t denotes the random variable indicating the state of the system at time t. So $X_t \in S$ denotes that the system is still "alive" (has not faded) at time t.

Lemma 3.1: A Markov decision process is strongly excessive if and only if there exist some partition $\{S_k | k \text{ integer}\}$ of S and numbers $\alpha > 1$, $\beta \geq 1$, such that for all strategies $\pi \in M$

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

In fact this lemma gives a bound for the expected number of visits to S_k. This number is bounded in k and ℓ but also decreases exponentially with increasing k for $k \geq \ell$. For a proof we refer to [12] by van Hee and Wessels.

Lemma 3.2: A Markov decision process is strongly excessive with $\mu(i) \geq \delta > 0$ if and only if the lifetimes of the process are exponentially bounded, i.e.

$$\mathbb{P}_i^\pi (X_t \in S) \leq a(i)\lambda^t \quad \text{for all } t = 0,1,2,\ldots, \text{ all } \pi \in M,$$

$$i \in S \text{ and}$$

positive function a on S, $\lambda < 1$.
Similarly, strong excessivity with $\Delta \geq \mu(i) \geq \delta > 0$ corresponds to uniform, exponential boundedness of lifetimes, i.e. the function a is constant on S.
Again we refer for proofs to [12].

Lemma 3.3: A Markov decision process is strongly excessive with $\Delta \geq \mu(i) \geq \delta > 0$ if and only if the supremal expected lifetime is bounded as a function of the starting state, i.e.

$$\sup_{i \in S} \sup_{\pi \in M} \sum_{t=0}^{\infty} \mathbb{P}_i^{\pi}(X_t \in S) < \infty.$$

This has been remarked for finite state and action Markov decision processes by Veinott in [39] and by Denardo in [4] . For our situation a proof may be found in [12].

There is a close relation between strong excessivity and so-called N-stage contraction. The following lemma (see [12]) even shows more, viz. if the spectral radius of the Markov decision process with respect to some norm μ is less than one, then the process is strongly excessive:

Lemma 3.4: Let μ be a positive function on S, such that $P(f)\mu \leq m_2\mu$ for some number m_2 and all policies f.
If $\lim_{n\to\infty} \|P^n(f)\|^{1/n} \leq \rho_* < 1$ for all f and some ρ_*, then the Markov decision process is strongly excessive with respect to some positive function μ' on S.

It should be remarked here that μ' is equivalent to μ in the sense that r(f)-vectors which are bounded in μ-norm are also bounded in μ'-norm (see the construction of μ' in [12.).
The proof of lemma 3.4 is relatively complicated. If we do not require the new norm μ' to be of the weighted supremum norm type, but content ourselves with contractingness with respect to an arbitrary norm, then a similar statement is much easier to prove (see [28]).

4. Stopping times and successive approximation methods

As mentioned in the introduction we will present alternative ways for generating sequences like $\{u^{(n)}\}$ at the end of section 2. These alternatives for the so called policy improvement step amount to replacement of the operators $L(f)$ - and hence T - by alternative ones. The alternative operators will be defined by stopping times.

Actually, several variants of the policy improvement step are well known. E.g. a Gauss-Seidel procedure (see Hastings[6] or Kushner and Kleinmann [17]), an overrelaxation procedure (see Reetz [33] or Schellhaas [37]) and several other variants (see van Nunen [24]). All variants require their own convergence proofs and their own construction of upper and lower bounds, although clearly the techniques are similar. Furthermore, the question arises whether this set of variants exhausts all possibilities. In [24] there is already a first attempt to a unified approach. In [44] it is demonstrated how stopping times can be used to generate alternative policy improvement operators and how a general proof and general bounds can be given for all stopping time based operators at the same time. This approach has been generalized to countable state spaces and randomized stopping times in [22]. In order to keep the presentation simple, we will only consider nonrandomized stopping times here.

Let us go back to the set-up of section 2. The set of allowed paths until time t is S^{t+1}. So, S^{∞} is the set of paths.
A stopping time τ is a function on S^{∞} with nonnegative integer values and satisfying $\tau^{\leftarrow}(t) = B \times S^{\infty}$ for some $B \subset S^{t+1}$. This means that τ prescribes stopping of the process at time t depending on the path of the process until time t. With this definition τ is a stopping time in the ordinary sense (see e.g. Ross [35])with respect to the random process $X_0, X_1, \ldots$.
An alternative way of introducing a stopping time is by way of its so-called "go ahead set". Each stopping time corresponds to a go ahead set (and reversely) in the following way:

$$\text{Let } G_\tau := \bigcup_{t=1}^{\infty} \{\alpha \in S^t \mid \tau(\alpha,\beta) \geq t \text{ for all } \beta \in S^{\infty}\} ,$$

then for all t we have:

$$\alpha \in G_\tau \cap S^t,\ \ell \in S,\ (\alpha,\ell) \notin G_\tau \Leftrightarrow \tau(\alpha,\ell,\beta) = t \text{ for all } \beta \in S^{\infty}$$

So G_τ is the set of paths until some time, for which the stopping time τ prescribes: go ahead. The set G_τ determines the stopping time τ completely (see [44], [26], [22]).
For our purpose (as will appear shortly) the only interesting stopping times are those with $\tau(\alpha) \geq 1$ for all $\alpha \in S^\infty$ (or equivalently $S \subset G_\tau$).
So from now on we restrict ourselves to such nonzero stopping times. Before going on, we give some simple examples:

a. $\tau \equiv n$ for some fixed n $(1 \leq n \leq \infty)$;

b. τ defined by the go ahead set G^H containing all paths $(i_0,i_1,\ldots,i_t)$ until time t (for any $t \geq 0$) with $i_0 > i_1 > \ldots > i_t$.

c. τ defined by the go ahead set G^R containing all paths $(i,i,\ldots,i)$ for any $i \in S$ until any time t.

Now we will explain how a stopping time τ generates a policy improvement step. Consider the stopping time $\tau \equiv 1$. Using this stopping time we may redefine the operators $L(f)$ on U as follows:

$$L(f)u = r(f) + P(f)u.$$

Then clearly $L(f)u = E^f[r(X_0,f(X_0) + u(X_1)]$,

$$L(f)u = E^f[\sum_{t=0}^{\tau-1} r(X_t,f(X_t) + u(X_\tau)],$$

where E_i^f denotes the expectation with respect to the stationary strategy f if the process starts in i; so E^f is the corresponding vector of expectations.
Now we can introduce similar operators $L_\tau(\pi)$
For arbitrary stopping times τ and arbitrary strategies $\pi = (f_0,f_1,\ldots)$:

$$L_\tau(\pi)u := E^\pi[\sum_{t=0}^{\tau-1} r(X_t,f_t(X_t) + u(X_\tau)],$$

with $u(X_\tau) = 0$ if $\tau = \infty$.
As for the case $\tau \equiv 1$, it is easy to show that $L_\tau(\pi)$ is a proper operator on U, which is monotone and strictly contracting with a contraction radius not larger that ρ_+ (see [22] or [44]). Therefore. $L_\tau(\pi)$ possesses a unique fixed point u_τ^π in U.
For the examples a,b,c with stationary strategy f the i-th component of $L_\tau(f)u$ becomes:

a. $(n = 1)\quad r(i,f(i)) + \sum_j p^{f(i)}(i,j)u(j)$.

b. $r(i,f(i)) + \sum_{j<i} p^{f(i)}(i,j)\,(L_\tau(f)u)(j) + \sum_{j\geq i} p^{f(i)}(i,j)u(j)$.

c. $(1 - p^{f(i)}(i,i))^{-1}\{r(i,f(i) + \sum_{j\neq i} p^{f(i)}(i,j)u(j)\}$.

Furthermore, it is easily verified that $L_\tau(f)v(f) = v(f)$, so for any τ we obtain $u_\tau^f = v(f)$.
Now a policy improvement operator T_τ on U may be defined in an obvious way

$$T_\tau u := \sup_{\pi\in M} L_\tau(\pi)u.$$

For computational purposes it is desirable if the supremum in the definition of T_τ can be restricted to stationary strategies.
Furthermore, it would be nice if $T_\tau u$ is also the supremum over all thinkable (history-based, randomized) strategies. All this is true if τ is a so-called transition memoryless stopping time, i.e. whether the process is stopped after a path $(i_0,i_1,\ldots,i_t)$ or not only depends on the pair (i_{t-1},i_t) for $t \geq 1$ and neither on the earlier history nor on t. So for a transition memoryless stopping time the stopping or going ahead is based on the most recent transition (for a detailed treatment see [22], [27], [44]).
If τ is not transition memoryless, then there are $\{p^a(i,j),\ r(i,a)\}$ such that the supremum of $L_\tau(\pi)u$ over all (very general) strategies π may not be replaced by the supremum over the stationary strategies only (see [22]).
From now on we will restrict ourselves to transition memoryless stopping times. For the operator T_τ on U we then have the following properties (for proofs see [22], [44]):

Lemma 4.1: 1. T_τ is a monotone operator on U.
2. T_τ is strictly contracting with contraction radius ν_τ not larger that ρ_+. Actually, ν_τ is the supremum of the radii of $L_\tau(f)$ over all policies f.
3. the unique fixed point of T_τ in U is the value function v.
4. for all $\varepsilon > 0$, $u \in U$ there exists a policy f such that

$$L_\tau(f)u \geq T_\tau u - \varepsilon\mu\ .$$

Now we may define for any (nonzero, transition memoryless) stopping time τ a successive approximation procedure (completely analogous to the procedure at the end of section 2):
Choose $\delta > 0$, $u^{(0)} \in U$ with $u^{(0)} < T_\tau u^{(0)}$ and determine f_n (for $n = 1,2,\ldots$) such that

$$u^{(n)} := L_\tau(f_n)\, u^{(n-1)} \geq \max \{u^{(n-1)}, T_\tau u^{(n-1)} - \delta\mu\} .$$

Then we have the following upper bound for v:

$$v \leq u^{(n)} + \frac{\delta + \nu_\tau \| u^{(n)} - u^{(n-1)} \|}{1 - \nu_\tau} \mu .$$

and a similar lower bound for v and $v(f_n)$. Furthermore, by definition $u^{(n)} \geq u^{(n-1)}$.

The examples a,b,c induce numerically well-executable procedures:

a. only $n = 1$ gives a transition memoryless stopping time and the procedure is the standard Gauss-Jordan procedure.

b. this stopping time yields the Gauss-Seidel procedure (see [6], [17].

c. this stopping time yields the Jacobi procedure (see [31]).

Also combinations of b. and c. are interesting as well as some other variants (compare [22]).

Porteus' preinverse transformation [31] for the data of the Markov decision process is strongly related to our stopping time approach (see [26]). Actually, this stopping time approach might give a possibility of weakening conditions for convergence of successive approximations. Namely, if for some τ the operator T_τ is N-stage contracting, then there will be convergence in norm of $T_\tau^n u$ to v. Whether this idea actually gives a real weakening of the conditions or not, has not been investigated so far.

5. Value oriented methods

In this section we will introduce another acceleration technique. As has been said already in the introduction, it will consist of better approximation of $v(f_n)$ in the n-th step of the procedure.
Suppose for simplicity that the supremum in $T_\tau u$ is actually attained for some policy. Then we find in the n-th step of the algorithm some policy f_n with

$$T_\tau u^{(n-1)} = L_\tau(f_n)\, u^{(n-1)} .$$

$T_\tau u^{(n-1)}$ is, of course, an approximation of $v(f_n)$. However, a better approximation would be

$$L_\tau^\lambda(f_n)\, u^{(n-1)}$$

for some natural - preferably large - number λ.
So, let us fix λ (λ may depend on n) and define

$$u^{(n)} := L_\tau^\lambda(f_n)\, u^{(n-1)} .$$

In this way one obtains a procedure which seems to converge better. However, this becomes less certain as soon as one realizes that the operator which maps an arbitrary $u^{(0)}$ on $u^{(1)}$ is neither necessarily monotone, nor contracting as one can see from some simple examples (see [23], [22] or [27]). Nevertheless the convergence proof is simple (at least if we enforce monotonicity of the iterates) as will be demonstrated.
Suppose that $u^{(0)}$ satisfies

$$T_\tau u^{(0)} \geq u^{(0)} \quad \text{and} \quad L_\tau(f_1)\, u^{(0)} = T_\tau u^{(0)} .$$

Hence $u^{(0)} \leq L_\tau(f_1)\, u^{(0)} \leq \ldots \leq L_\tau^\lambda(f_1)\, u^{(0)} = u^{(1)} \leq \lim_{n\to\infty} L^n(f_1) u^{(0)} = v(f_1)$

Furthermore, $T_\tau u^{(1)} = L_\tau(f_2)\, u^{(1)} \geq L_\tau(f_1)\, u^{(1)}$. Thus one obtains $u^{(1)} \leq u^{(2)} \leq v(f_2)$. By induction this gives $u^{(n-1)} \leq u^{(n)} \leq v(f_n) \leq v$.
On the other hand one has $u^{(n)} \geq T_\tau^n\, u^{(0)}$, where the right hand side tends to v for $n \to \infty$. Therefore $u^{(n)} \to v$ when $n \to \infty$. Moreover

$$\| u^{(n)} - v \| \leq \| T_\tau^n\, u^{(0)} - T_\tau^n\, v \| \leq \nu_\tau^n\, \| u^{(0)} - v \| .$$

In the same way as in section 2 and 4, one may obtain efficient bounds. As an example we give the upper bound for v:

$$v \leq u^{(n)} + (1 - \nu_\tau)^{-1} \| T_\tau u^{(n)} - u^{(n)} \| \cdot \mu$$

where the only remarkable point is, that the extrapolation is not based on $u^{(n+1)} - u^{(n)}$, but on $T_\tau\, u^{(n)} - u^{(n)}$.

It should be remarked that the restriction to finite values of λ is not essential. If we interpret $L_\tau^\infty(f)u$ as $\lim_{n\to\infty} L_\tau^n(f)u$, then all properties of our value oriented methods also hold for $\lambda = \infty$.
However, $L_\tau^\infty(f)u = v(f)$, which implies that for $\lambda = \infty$ we obtain the following procedure (without the stopping criterion):

Choose $u^{(0)}$, such that $u^{(0)} \leq T_\tau u^{(0)}$.
Determine f_n for $n = 1,2,\ldots$ with

$$T_\tau u^{(n-1)} = L_\tau(f_n)\, u^{(n-1)}, \quad u^{(n)} := v(f_n).$$

This means, that for $\lambda = \infty$ we have a policy iteration procedure. Namely, for $\tau \equiv 1$ this results in the standard policy iteration procedure of Howard [16](extended to more general situations), for the stopping time from example b (section 4) this results in Hasting's policy iteration procedure [6] based on Gauss-Seidel iteration. In fact, we developed in this way a whole set of policy iteration procedures (for some examples, see [24]).

The idea of value oriented methods has been initiated in [23], where also examples are given showing the numerical advantages. The combination with non-standard policy improvement steps has been given in [24] and generalized to stopping time based policy improvements in [27]. The most general theory (countable state space, randomized stopping times) has been presented in [22].

6. Elimination of non-optimal actions

The final work saving idea which will be presented in this paper, is the elimination of actions which are clearly not optimal. If such an elimination can be executed without much extra work, then it will save work, since the amount of work in the policy improvement step heavily depends on the number of available actions.
The basic idea for using upper and lower bounds for elimination of actions stems from MacQueen [20]. A considerable improvement has been suggested by Hastings in [7] for average reward problems by eliminating actions only temporarily. For total expected reward problems this idea has been extended by Hastings and van Nunen in [8].
We will present the action-elimination algorithm for the standard successive approximations procedure (i.e. $\tau \equiv 1$, $\lambda = 1$), however the more general case may be treated in the same way.
For simplicity of notations we will also suppose that the supremum in the operator T is attained, so δ can be taken zero. Now consider the algorithm of section 2 (with $\delta = 0$), which produces a monotone nondecreasing sequence $\{u^{(n)}\}_{n=0}^{\infty}$. Suppose for some n we have $u^{(n)} = u$, $u^{(n-1)} = w$, then we have the following lower bound for $u^{(n+1)} = Tu^{(n)} = Tu$:

$$u + \rho_- \|u - w\| \cdot \mu \leq Tu,$$

where $\rho_- = \inf_f \rho_f^-$. This can be proved with the same trick as in the first step of the proof of lemma 2.3 (i).
For an arbitrary f we find the following bound for L(f)u with the analogous trick:

$$L(f)u \le r(f) + P(f)w + \rho_f^+ \|u - w\| \mu,$$

where $\rho_f^+ := \|P(f)\|$.

Combining these bounds one may conclude that f can only be a candidate for the maximizer in Tu if

$$u + \rho_- \|u - w\|_- \mu \le r(f) + P(f)w + \rho_f^+ \|u - w\| \mu .$$

This property can be used to eliminate some actions for state i as candidates for the maximizing action $f_{n+1}(i)$. Namely, by replacing ρ_f^+ by the more conservative value ρ_+ we see that $f_{n+1}(i) \neq a$ if

$$u^{(n)}(i) + \rho_- \|u - w\|_- \mu(i) > r(i,a) + \sum_j p^a(i,j)\, w(j) + \rho_+ \|u - w\| \mu(i),$$

or

$$r(i,a) + \sum_j p^a(i,j)\, u^{(n-1)}(j) <$$

$$< u^{(n)}(i) + \{\rho_- \|u^{(n)} - u^{(n-1)}\|_- - \rho_+ \|u^{(n)} - u^{(n-1)}\|\}\mu(i)$$

Using this idea one can easily test action a in state i as a candidate for step n + 1 with the data computed in step n.
Analogously one can test whether action a for state i can be eliminated for more than one step or not. Namely, again using the trick from the proof of lemma 2.3, we can give as lower bound for $T^m u$:

$$u + (\rho_- + \rho_-^2 + \ldots + \rho_-^m) \|u - w\|_- \mu \le T^m u.$$

As upper bound for $L(f)\, T^{m-1}u$ we find

$$L(f)\, T^{m-1}u \le r(f) + P(f)w + (\rho_+ + \ldots + \rho_+^m) \|u - w\| \mu .$$

Hence a test for action a in state i to be no candidate for $f_{n+m}(i)$ becomes

$$r(i,a) + \sum_j p^a(i,j)\, u^{(n-1)}(j) <$$

$$< u^{(n)}(i) + \{(\rho_- + \ldots + \rho_-^m)\| u - w \|_- - (\rho_+ + \ldots \rho_+^m)\| u - w \|\}\mu$$

Since the right hand side of this testinequality is nonincreasing in m, we can use this inequality with varying m to decide for how many steps action a in state i may be disregarded. This test does not cost much extra work but can save much computational work. In [8] Hastings and van Nunen illustrate this by Howard's autoreplacement problem. Here we will present the results for an inventory problem. In this example of a discounted Markov decision problem where β is the discountfactor, we have $\mu \equiv 1$ and hence $\rho_- = \rho_+ = \beta$.

Inventory planning example:

$\beta = 0.99$.

$S = \{0, 1, \ldots, 60\}$, where $i \in S$ denotes an inventory level.

$A = \{0, 1, \ldots, 60\}$, where $a \in A$ denotes the inventory level after ordering; in state i only actions a with $a \geq i$ are available. Hence 1891 combinations (i,a) are available. With some realistic cost data and a demand distribution ranging from zero to forty we obtained the elimination results as stated in the table. For comparison, we also give the elimination results for MacQueen's test [20], which only removes action a in state i if it satisfies the test for $m = \infty$, however, in that case the removal is permanent.

no of iteration step	Hastings/v. Nunen no of (i,a)-combinations which have not been eliminated temporarily	MacQueen no of (i,a)-combinations which have not been eliminated permanently
2	1891	1891
3	1890	1891
4	1859	1891
5	1585	1891
6	1131	1891
7	1038	1891
8	648	1891
9	434	1891
10	237	1891
11	186	1891
12	150	1889
13	103	1848
14	90	1809
15	98	1200
16	65	883
17	65	258
18	62	177
19	65	102
20	62	100
21	63	78
22	61	68
23	61	62
24	61	62
25	61	61

Table of performance of two action elimination procedures. Note that the Hastings/v. Nunen procedure does not produce a monotone column.

7. Markov games

Most of the ideas explained in the preceding sections can be extended to so called Markov games. Actually, as stated in the introduction, the basic dynamic programming approach for Markov games is older than the similar approach for Markov decision processes (see Shapley [38]). However, the refinements as presented in this paper have been developed first for Markov decision processes. Most refinements can be generalized in a simple way to Markov games. At least, after the generalization has been found, it appears to be simple.

In this section we will give a short introduction to Markov games and demonstrate how the ideas of the preceding section can be applied.

7.1. Markov games with unbounded rewards

We now consider a system as in section 2, however, two players may choose at any time $t = 0,1,2,\ldots$ actions a and b from sets A and B after having observed the state i of the system. These (independently made) choices determine the immediate reward $r(i,a,b)$ for the first player (choosing from A); this reward has to be paid by the second player (choosing from B). Another result of these choices is a state transition which will result in state j with probability $p^{a,b}(i,j)$.

The conditions have to be stronger than in section 2 with respect to the action spaces. For simplicity we suppose here that A and B are finite. For more general cases see Couwenbergh [2], [3] and the survey paper of Parthasarathy and Stern [29]. In all generalizations there are compactness requirements for A and B. Again we only introduce Markov strategies, since it can be proved that more general strategies can be discarded (see e.g. [2], [3], [40],[42],[45]). However, we need randomized Markov strategies. So we call f a policy for the first player if $f(i)$ is a probability distribution on A: $f^a(i) \geq 0$, $\sum_{a\in A} f^a(i) = 1$. Similarly a policy g for the second player is defined as a probability distribution on B. Strategies for the players are sequences of policies: $\pi = (f_0,f_1,\ldots)$ for the first player and $\rho = (g_0,g_1,\ldots)$ for the second player.

Now the total expected reward $v(\pi,\rho)$ for the first player (= costs for the second player if they play the strategies π,ρ is defined as:

$$v(\pi,\rho) := \sum_{t=0}^{\infty} \{ \prod_{n=0}^{t-1} P(f_n,g_n)\} \, r(f_t,g_t) \ ,$$

where $P(f,g)$, $r(f,g)$ are defined analogous to $P(f)$, $r(f)$ in section 2.

The goal is to find strategies π^*, ρ^* such that

$$v(\pi,\rho^*) \leq v(\pi^*,\rho^*) \leq v(\pi^*,\rho) \quad \text{for all strategies } \pi,\rho.$$

Furthermore, one is interested in finding the value of $v(\pi^*,\rho^*)$, which will be denoted by v and called: the value of the game, π^* and ρ^* will be called optimal strategies.

Similarly as in section 2, the key to the solution is the fact that v satisfies (under suitable conditions) a kind of optimality equation

$$v = \sup_f \inf_g \{r(f,g) + P(f,g)v\}.$$

As in section 2 this may be proved by dynamic programming. In this case it implies the introduction of operators $L(f,g)$ and Q with the similar tasks as the operators $L(f)$ and T:

$$L(f,g)u = r(f,g) + P(f,g)u,$$

$$Qu = \sup_f \inf_g \{r(f,g) + P(f,g)u\} = \sup_g \inf_f L(f,g)u.$$

In order to guarantee that $L(f,g)$ and Q are well-defined operators with nice properties, we need similar assumptions as in (i) - (iii) of section 2:

(i) a. there is a number $m > 0$, such that for all policies f,g:

$$\| r(f,g) - \bar{r} \| \leq m,$$

where $\bar{r}$ is the vector with $\bar{r}(i) := \sup_{a\in A} \inf_{b\in B} r(i,a,b)$ or

$$\bar{r} = \sup_f \inf_g r(f,g)$$

b. $$\sup_{\pi,\rho} \sum_{t=0}^{\infty} \{ \prod_{n=0}^{t-1} P(f_n,g_n) \} |\bar{r}| < \infty$$

(ii) $$\rho_+ := \sup_{f,g} \| P(f,g) \| < 1$$

(iii) $$m_1 := \sup_{f,g} \| P(f,g)\bar{r} - \rho\bar{r} \| < \infty \text{ for some } \rho \text{ with } 0 < \rho < 1.$$

(Actually, for some of these assumptions only degenerated policies are necessary; furthermore, sup and inf may often be replaced by max and min).

With these assumptions $L(f,g)$ and Q are proper operators in U which is defined - exactly as in section 2 - as the translation over $(1-\rho)^{-1}r$ of the space W of functions on the state space with finite norm. $(Q,u)(i)$ now represents the value of the matrix game on A,B with rewards $(L(f,g)u)(i)$ if policies f,g are chosen. That $(Qu)(i)$ is really the value of this matrix game follows from the finiteness of A and B. Interpreting $L(f,g)u$, we see that $(Qu)(i)$ represents the value of the one-step Markov game starting in i and with terminal reward u. So - exactly as in section 2 - we obtain that $(Q^n u)(i)$ is the value of the n-step Markov game with terminal value u. Using the fact that - again similar to section 2 - Q is contracting on U, we obtain that $Q^n u$ tends to the unique fixed point of Q if n tends to infinity. Then it is simple to prove that this fixed point must be the value v of the ∞-step Markov game. As in section 2 this can be used to construct a successive approximation method for finding v and relatively good strategies: choose $u^{(0)} \in U$, compute successively $u^{(n)} = Qu^{(n-1)}$ and f_n, g_n such that

$$L(f,g_n)\, u^{(n-1)} \leq u^{(n)} = Qu^{(n-1)} = L(f_n,g_n)\, u^{(n-1)} \leq L(f_n,g)\, u^{(n-1)}$$

Again $u^{(n)}$ and $u^{(n-1)}$ may be used for the computation of simple but efficient upper and lower bounds for v and $v(f_n,g_n)$. Because of the similarity with section 2 we will skip this (see e.g. [42]), moreover, we have here the simpler situation that the sup and inf in Q are always attained.

The results of section 3 apply completely to the Markov game situation, since the lemmas 3.1 - 3.4 in fact only require a set of transition matrices with the property that any combination of rows from some of the allowed matrices forms again an allowed matrix (see [12]).
For a slight extension of the assumptions in this section see [45].

7.2. Stopping times generating approximation methods

As in section 4 we may introduce a stopping time τ and define $L_\tau(\pi,\rho)$ by

$$L_\tau(\pi,\rho)u := E^{\pi,\rho}\Big[\sum_{t=0}^{\tau-1} r(X_t,f_t(X_t),g_t(X_t)) + u(X_t)\Big],$$

with $u(X_\tau) = 0$ if $\tau = \infty$ and $E^{\pi,\rho}$ is defined as in sections 2 and 4.

This leads to the definition of Q_τ by

$$Q_\tau u = \sup_\pi \inf_\rho L_\tau(\pi,\rho)\, u.$$

Again, we obtain for nonzero transition memoryless stopping times τ, that $u^{(n)} = Q_\tau\, u^{(n-1)}$ produces a proper successive approximations method. For details see van der Wal [40].

7.3. Value oriented methods for Markov games

The value oriented methods for Markov decision processes as introduced in section 5 contain policy iteration methods as extreme cases ($\lambda = \infty$). For Markov games the standard ($\tau \equiv 1$) extension of the policy iteration method has been suggested by Pollatschek and Avi-Itzhak [30]. They give a convergence proof under fairly strong conditions and only conjecture the convergence under milder conditions. A more general proof of Rao, Chandrasekaran and Nair [32] appeared to be incorrect as has been demonstrated by van der Wal in a forthcoming paper [41]. Actually, a simple example (see [41] or [42]) shows that the algorithm may start cycling.

Hence the straightforward generalization of the policy iteration method to Markov games is not feasible in general. However, an idea of Hoffman and Karp [14] for average reward games can be applied here as has also been suggested by Pollatschek and Avi-Itzhak [30]. Van der Wal [41] has used this set up for the construction of value oriented methods for Markov games. The case $\tau \equiv 1$, $\lambda = 1$ represents the standard successive approximations method; arbitrary τ, $\lambda = 1$ represents the methods of subsection 7.2; $\tau \equiv 1$, $\lambda = \infty$ represents the policy iteration method as introduced by Pollatschek and Avi-Itzhak according to the idea of Hoffman and Karp. Here, we describe the method for fixed τ and λ:

choose $u^{(0)} \in \mathbf{U}$ with $Q_\tau\, u^{(0)} \leq u^{(0)}$

Policy improvement step: Determine $Q_\tau u^{(n)}$ for successive values of n and find a policy g_{n+1} satisfying $L_\tau(f,g_{n+1})\, u^{(n)} \leq$ $\leq Q_\tau\, u^{(n)}$ for all f.

Value approximation step: Determine $u^{(n+1)} := Q_\tau^\lambda(g_{n+1})\, u^{(n)}$, where the operator $Q_\tau(g)$ is defined by $Q_\tau(g)u := \max_f L_\tau(f,g)u$.

Having the formulation of this procedure, the convergence proof is very similar to the proof for Markov decision processes (see [41]). Also the stopping criteria and the bounds for v and $v(f_n,g_n)$ are completely similar (see [41], [42],[45]).

7.4. Elimination of nonoptimal actions

As in section 6 we restrict attention to the standard successive approximation method ($\tau \equiv 1$, $\lambda = 1$). Then an action a' is nonoptimal at stage n in state i, if any policy $f_n^{(i)}$ being optimal for the matrix game $r(i,a,b) + \sum_j p^{a,b}(i,j)u^{(n)}(j)$ satisfies $f_n^{a'}(i) = 0$.

This gives the possibility to eliminate some actions for both players in some states for one iteration step. This can be executed completely similar to the procedure presented for Markov decision processes (section 6) using the upper and lower bounds which have not been stated explicity in sub section 7.1. Such a procedure has been worked out in detail by Reetz and van der Wal in [34]. It will be self-evident that the same idea may be used to eliminate actions for m steps (see [34]).

References:

[1] Blackwell, D., Discounted dynamic programming.
Ann. Math. Statist. 36 (1965) 226-235.

[2] Couwenbergh, H.A.M., Stochastic games with general state space.
Master's thesis. Dept. of Mathematics, Eindhoven University of Technology, February 1978.

[3] Couwenbergh, H.A.M., Stochastic games with metric state space.
Memorandum COSOR-78-05, February 1978, Dept. of Math., Eindhoven University of Technology.

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

[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. Q. 19 (1968) 453-464.

[7] Hastings, N.A.J., A test for nonoptimal actions in undiscounted finite Markov decision chains.
Management Sci: 23 (1976) 87-91.

[8] Hastings, N.A.J. & J.A.E.E. van Nunen, The action elimination algorithm for Markov decision processes.
p. 161 - 170 in the same volume as [25]

[9] Hee van, K.M., Markov strategies in dynamic programming.
Mathematics of Operations Research (to appear).

[10] Hee van, K.M., A Hordijk & J. van der Wal, Successive approximations for convergent dynamic programming.
p. 183-211 in the same volume as [25].

[11] Hee van, K.M. & J. van der Wal, Strongly convergent dynamic programming: some results.
p. 165 - 172 in M. Schäl (ed.), Dynamische Optimierung, Bonn, Bonner Mathematische Schriften nr. 98, 1977.

[12] Hee van,K.M. & J. Wessels, Markov decision processes and strongly excessive functions.
Memorandum COSOR 77-11, May 1977, Dept. of Math., Eindhoven University of Technology.

[13] Hinderer, K.& G.Hübner, On approximate and exact solutions for finite stage dynamic programs.
p. 57-76 in the same volume as [25].

[14] Hoffman, A.J. & R.M. Karp, On nonterminating stochastic games.
Management Science 12 (1966) 359-370.

[15] Hordijk, A., Dynamic programming and Markov potential theory.
Amsterdam, Mathematical Centre (Mathematical Centre Tract no. 51) 1974.

[16] Howard, R.A., Dynamic programming and Markov decision processes, Cambridge (Mass.), M.I.T.-press, 1960.

[17] Kushner, H.J. & A.J. Kleinmann, Accelerated procedures for the solution of discrete Markov control problems.
IEEE-Trans. Autom. Contr. 16 (1971) 147-152.

[18] Lippman, S.A., On dynamic programming with unbounded rewards.
Management Science 21 (1975) 1225-1233.

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

[20] MacQueen, J., A test for suboptimal actions in Markovian decision problems.
Oper. Res. 15 (1967) 559-561.

[21] Mine, H. & S. Osaki, Markovian decision processes. New York etc., Elsevier 1965.

[22] Nunen van, J.A.E.E., Contracting Markov decision processes. Amsterdam, Mathematical Centre (Mathematical Center Tract no. 71) 1976.

[23] Nunen van, J.A.E.E., A set of successive approximation methods for discounted Markovian decision problems.
Zeitschrift für Oper. Res. 20 (1976) 203-208.

[24] Nunen van, J.A.E.E., Improved successive approximation methods for discounted Markov decision processes.
p. 667-682 in A. Prékopa (ed.), Progress in Operation Research, Amsterdam, North-Holland Publ. Comp. 1976.

[25] Nunen van,J.A.E.E. & J. Wessels, Markov decision process with unbounded rewards.
p. 1-24 in H.C. Tijms & J. Wessels (eds.), Markov decision theory, Amsterdam, Mathematical Centre (Mathematical Centre Tract no. 93) 1977.

[26] Nunen van, J.A.E.E. & J. Wessels, The generation of successive approximations for Markov decision processes by using stopping times.
p. 25-37 in the same volume as [25].

[27] Nunen van, J.A.E.E. & J. Wessels, A principle for generating optimization procedures for discounted Markov decision process. p. 683-695 in the same volume as [24].

[28] Nunen van, J.A.E.E. & J. Wessels, A note on dynamic programming with unbounded rewards.
Management Science 24 (1978) 576-580.

[29] Parthasarathy, T.& M. Stern, Markov games - a survey.University of Illinois at Chicago Circle, Chicago (1976).

[30] Pollatschek, M.A. & B. Avi-Itzhak, Algorithms for stochastic games.
Management Science 15 (1969) 399-415.

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

[32] Rao, S.S., R. Chandrasekaran & K.P.K. Nair, Algorithms for discounted stochastic games.
J. Opt. Th. Appl. 11 (1973) 627-637.

[33] Reetz, D., Solution of a Markovian decision problem by overrelaxation.
Z. Oper. Res. 17 (1973) 29-32.

[34] Reetz, D. & J. van der Wal, On suboptimality in two-person zero-sum Markov games.
Memorandum COSOR 76-19, October 1976 (revised October 1977), Dept. of Math., Eindhoven University of Technology.

[35] Ross, S.M., Applied probability models with optimization applications.
San Francisco, Holden-Day 1970.

[36] Schäl, M., Conditions for optimality in dynamic programming and for the limit of N-stage optimal policies to be optimal.
Z.f. Wahrscheinlichkeitstheorie 32 (1975) 179-196.

[37] Schellhaas, H., Zur Extrapolation in Markoffschen Entscheidungsmodellen mit Diskontierung.
Z. Oper. Res. 18 (1974) 91-104.

[38] Shapley, L.S., Stochastic games.
Proceed. Nat. Acad. Sci. 39 (1953) p. 1095-1100.

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

[40] Wal van der, J., Discounted Markov games; successive approximations and stopping times.
Intern J. Game Th. 6 (1977) 11-22.

[41] Wal van der, J., Discounted Markov games; the generalized policy iteration method.
J. Optim. Th. Appl. 25 (1978) no. 1

[42] Wal van der, J., & J. Wessels, Successive approximation methods for Markov games.
p. 39-55 in the same volume as [25].

[43] Wessels, J., Markov programming by successive approximations with respect to weighted supremum norms.
J. Math. Anal. Appl. 58 (1977) 326-335.

[44] Wessels, J., Stopping times and Markov programming.
p. 575-585 in Transactions of the 7 th Prague Conference on Information theory, Statistical Decision Functions, Random Processes, Prague, Academia 1977.

[45] Wessels, J., Markov games with unbounded rewards.
p. 133-147 in the same volume as [11].

[46] Wessels, J. & J.A.E.E. van Nunen, Discounted semi-Markov decision processes: linear programming and policy iteration.
Statistica Neerlandica 29 (1975) 1-7.