A principle for generating optimization procedures for discounted Markov decision processes

Citation for published version (APA):
Wessels, J., & van Nunen, J. A. E. E. (1974). *A principle for generating optimization procedures for discounted Markov decision processes*. (Memorandum COSOR; Vol. 7411). Technische Hogeschool Eindhoven.

Document status and date:
Published: 01/01/1974

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 23. Aug. 2022

TECHNOLOGICAL UNIVERSITY EINDHOVEN

Department of Mathematics

STATISTICS AND OPERATIONS RESEARCH GROUP

memorandum COSOR 74-11

A principle for generating optimization procedures for discounted Markov decision processes

by

J. Wessels
and
J.A.E.E. van Nunen

Eindhoven, October 1974

A principle for generating optimization procedures for discounted Markov decision processes

by

J. Wessels
and
J.A.E.E. van Nunen

§ 0. Introduction

In this paper we will show how all existing optimization procedures (and a number of new ones) for discounted Markov decision processes may be derived from one point of view.
So we consider a finite-state discrete time Markov system which is controlled by a decision maker. After each transition the system may be identifies as being in one of N possible states. Let $S := \{1,2,\dots,N\}$ be the set of states. Transitions occure at discrete points in time $n = 0,1,2,\dots$. After observing state i at time n the decisionmaker selects an action k from a nonempty finite set K(i). Now $p_{ij}^{k} (\geq 0)$ is the probability of a transition to state $j \in S$ if the system's actual state is $i \in S$ and decision $k \in K(i)$ has been selected. An expected reward $r^{k}(i)$ is earned immediately while future income is discounted by a constant factor β, $0 < \beta < 1$.

The problem is to choose a policy which maximizes the total expected discounted reward over an infinite time horizon.

In the literature a great number of optimization procedures for solving this kind of problems has been presented. Each procedure requires its own proof of convergence and possesses its own properties. We divide the proposed procedures into two classes:

policy improvement procedures;
policy improvement-value determination procedures.

In procedures of the second class in each iteration step some extra work is done in order to estimate or compute the values for the current policy ([3], [4], [5], [6], [7], [8]). Procedures of the first class have been presented in [1], [2], [3], [7] and [11].

In § 1 we will use (as in [12]) the concept of stopping times for the generation of policy improvement procedures.
In § 2 we will show that any policy improvement procedure may be used to generate a whole set of policy improvement-value determination procedures (including a Howard like one).
In § 3 we will present upper and lower bounds for the values corresponding to the policies which appear during the iteration process.
This has been done already for specific procedures [1], [3], [7], [9].
We will present a general approach.
Finally some extensions to more general problems will be indicated.

§ 1. Policy improvement procedures

For the Markov decision process as described in the introduction the set of allowed paths until time n is S^{n+1}. So $S^\infty := S \times S \times S \ldots$ is the set of all allowed paths.

Definition 1.1.

a) The function τ on S^∞ with nonnegative integer values is called a *stopping time*, if and only if its inverse satisfies $\overset{\leftarrow}{\tau}(n) = B \times S^\infty$ with $B \subset S^{n+1}$;

b) a nonempty subset A of $\bigcup_{k=0}^{\infty} S^k$ is called a *go ahead set*, if and only if

$(\alpha,\beta) \in A \Rightarrow \alpha \in A$ for all $(\alpha,\beta) \in \bigcup_{k=0}^{\infty} S^k$.

(S^0 only consists of the null-tuple which concatenates to α with any α: our definition implies that any go ahead set contains this null-tuple.)

Notations. $\equiv$ $A_n := \bigcup_{k=0}^{n} S^k \quad (0 \le n \le \infty)$;

$\equiv$ the i-th component of $\alpha \in S^n$, $(n \ge 1)$ is denoted by $[\alpha]_{i-1}$;

$\equiv$ if $\alpha \in S^n$ $(n \ge 0)$ k_α is defined to be n;

$\equiv$ hence $\alpha \in S^n$ $(n \ge 1)$ may be written as $([\alpha]_0,[\alpha]_1,\ldots,[\alpha]_{k_\alpha -1})$;

$\equiv$ hence $k_\gamma = k_\alpha + k_\beta$ if $\gamma = (\alpha,\beta)$;

$\equiv$ $A(i) := \{\alpha \in A \mid [\alpha]_0 = i \text{ if } k_\alpha \ge 1\}$.

There is a one to one correspondence between stopping times and go ahead sets:

$$A = \bigcup_{n=0}^{\infty} \{\alpha \in S^n \mid \forall_{\beta \in S^\infty} \tau(\alpha,\beta) \geq n\} ,$$

$$\alpha \in A,\ \ell \in S,\ (\alpha,\ell) \notin A \Leftrightarrow \tau(\alpha,\ell,\beta) = k_\alpha \text{ for all } \beta \in S^\infty .$$

Definition 1.2. A stopping time τ (or its go ahead set A) is said to be *nonzero* if and only if $\tau(\alpha) \geq 1$ for all $\alpha \in S^\infty$ (or equivalently $S \subset A$).

The only nonzero stopping time which is an entry time (memoryless) is $\tau \equiv \infty$ ($A = A_\infty$).

Examples of nonzero stopping times

1.1. A_n : $(1 \leq n \leq \infty)$, $(\tau \equiv n)$;

1.2. A_H : defined by $A_H(i) := S^0 \cup \{(i) \cup (i,\alpha) \mid \alpha \in \bigcup_{j=1}^{i-1} A_H(j)\}$

1.3. A_R : defined by $A_R(i) := \bigcup_{n=0}^{\infty} \{\alpha \in S^n \mid [\alpha]_j = i,\ j = 0,1,2,\ldots,n-1, \text{ if } n \geq 1\}$

1.4. A_E : with E a subset of S defined by:

$$A_E := \bigcup_{n=2}^{\infty} \{\alpha \in S^n \mid [\alpha]_j \in E,\ j = 1,2,\ldots,n-1\} \cup S \cup S^0$$

$$(E = S \Rightarrow A_E = A_\infty;\ E = \emptyset \Rightarrow A_E = A_1) .$$

Definition 1.3.

≡ A *decision rule* D is a function ascribing to each $\alpha \in \bigcup_{k=1}^{\infty} S^k$ an element $D(\alpha)$ of $K([\alpha]_{k_\alpha - 1})$;

≡ the decision rule D is said to be *memoryless* (stationiary Markov) if

$$D(\alpha) = D([\alpha]_{k_\alpha - 1}) \text{ of each } \alpha \in \bigcup_{k=1}^{\infty} S^k;$$

≡ the set of decision rules is denoted by $\mathcal{D}$; the set of memoryless decision rules by M.

Let a decision rule $D \in \mathcal{D}$ be given. This decision rule determines a stochastic process $\{x_n \mid n = 0,1,\dots\}$ on S.

As in [12] we now introduce the operator L_τ^D where τ and D are given.

Definition 1.4. $D \in \mathcal{D}$, τ is a stopping time, A its corresponding go ahead set. The operator L_τ^D (or L_A^D) on $\mathbb{R}^N$ is defined by:

$$(L_\tau^D v)(i) := \mathbb{E}_D\Big(\sum_{k=0}^{\tau-1} \beta^k r^{D(x_0,\dots,x_k)}(x_k) + \beta^\tau v(x_\tau) \mid x_0 = i\Big)$$

(where $\mathbb{E}_D$ denotes the expectation given that decision rule D is used), or equivalently:

$$(L_A^D v)(i) = \sum_{\alpha\in A(i)} \mathbb{P}_D(\alpha|i)\beta^{k_\alpha - 1} r^{D(\alpha)}([\alpha]_{k_\alpha - 1}) + \sum_{\substack{\alpha\in A(i)\\ \ell\in S\\ (\alpha,\ell)\notin A(i)}} \mathbb{P}_D(\alpha,\ell|i)\beta^{k_\alpha} v(\ell) .$$

$\mathbb{P}_D(\alpha|i)$ is the probability of path α given that $x_0 = i$ and decision rule D is used.

Lemma 1.1. Let τ be an arbitrary stopping time. For any $v \in \mathbb{R}^N$, there exists a decision rule D_0 such that

$$L_\tau^{D_0} v \geq L_\tau^D v$$

componentwise for all $D \in \mathcal{D}$. For a proof see [12].

Notation. The vector $L_\tau^{D_0} v$ will be denoted by:

$$\max_D L_\tau^D v,\ U_\tau v,\ \max_D L_A^D v,\ U_A v .$$

The operators U_τ serve for some specific choices of τ to construct optimization procedures, which aim actually at finding $U_{A_\infty} 0$ (sometimes denoted by $U_\infty 0$, 0 denotes the null-vector in $\mathbb{R}^N$). The i-th component of $U_\infty 0$ gives the total expected discounted reward over an infinite time horizon when the initial state is i and an optimal decision rule is used.

From a computational point of view it is desirable to maximize only over the memoryless decision rules when $U_\tau v$ is computed. This is allowed when the stopping time is transition memoryless (see [12]):

Definition 1.5. A stopping time τ (and its corresponding go ahead set A) is said to be *transition memoryless*, if and only if there exists a subset T_1 of S^2 and a subset S_0 of S such that

$$\tau(\alpha) = 0 \Leftrightarrow [\alpha]_0 \in S_0$$

$$\tau(\alpha) = n \ (n > 0) \Leftrightarrow [\alpha]_0 \notin S_0,\ ([\alpha]_k,[\alpha]_{k+1}) \in T_1 \text{ for } k = 0,1,\ldots,n-2$$
$$([\alpha]_{n-1},[\alpha]_n) \in T_1 \ .$$

Lemma 1.2. If τ is transition memoryless, then for all $v \in \mathbb{R}^N$

$$U_\tau v = \max_{D \in M} L_\tau^D v \ .$$

For a proof see [12].

Theorem 1.1.

a) The operators L_τ^D and U_τ are monotone, i.e.:
if $v \geq w$ (componentwise) then:

$$L_\tau^D v \geq L_\tau^D w \text{ and } U_\tau v \geq U_\tau w \ .$$

b) The operators L_τ^D and U_τ are strictly contracting (with respect to the supnorm in $\mathbb{R}^H$: $\|v\|_\infty = \max_i |v(i)|$) if and only if τ is nonzero, the corresponding contraction radii ρ_τ^D and ν_τ are equal to:

$$\rho_\tau^D := \max_{i \in S} \mathbb{E}_D(\beta^\tau | x_0 = i), \ \nu_\tau := \max_D \rho_\tau^D \ .$$

c) If D is memoryless then for any nonzero τ the fixed point of L_τ^D equals $L_{A_\infty}^D 0$.

d) For all nonzero τ the operators U_τ possess the fixed point $U_{A_\infty} 0$ (= $U_\infty 0$).

The stopping times used in the examples of this section are all nonzero and transition memoryless (hence: $S_0 = \emptyset$).

Lemma 1.3. Let τ be transition memoryless; suppose $r^k(i) \geq 0$ for all $i \in S$ and all $k \in K(i)$ then the sequence

$$v_0^\tau := 0$$

$$v_n^\tau := U_\tau v_{n-1} = (U_\tau)^n 0 \qquad (n = 1,2,\ldots) ,$$

is nondecreasing and converges to $U_\infty 0$, i.e.

$$v_{n-1}^\tau \leq v_n^\tau \leq L_{A_\infty}^{D_n} 0 \leq U_\infty 0$$

$$\lim_{n\to\infty} v_n^\tau = U_\infty 0 \; .$$

Here D_n is the memoryless decision rule found by applying U_τ on v_{n-1}^τ. The proof follows in a direct way from Theorem 1.1 and lemma 1.2.

Remark. The restriction $r^k(i) \geq 0$ which is permitted without loss of generality, is made in order to enable us to start each algorithm with the same starting vector $v_0^\tau = 0$. Without this restriction it is sufficient for the preservation of the monotonicity of the sequence v_n^τ, if v_0^τ satisfies:

$$U_\tau v_0^\tau \geq v_0^\tau \quad .$$

Examples. 1.1. $\nu_{A_n} = \beta^n \qquad (1 \leq n \leq \infty)$

1.2. $\nu_{A_H} = \beta$

1.3. $\nu_{A_R} = \beta \, \frac{1-p}{1-\beta p}$, with $p := \min\limits_{i,D(i)} p_{ii}^{D(i)}$.

§ 2. Policy improvement-value determination procedures

Now, for each stopping time τ which is nonzero and transition memoryless, we introduce a class of value oriented extensions of the operator U_τ.

Definition 2.1. For τ transition memoryless, $\lambda \in \mathbb{N}$, $v \in \mathbb{R}^N$ we define the operator

$$U_\tau^{(\lambda)} v := (L_\tau^{D_v})^\lambda v$$

where D_v is the memoryless strategy which is found by applying U_τ on v. Now $U_\tau^{(\lambda)}$ is neither necessarily strictly contracting nor necessarily monotone.

Theorem 2.1. Suppose $r^k(i) \geq 0$ for all $i \in S$ and $k \in K(i)$, let τ be transition memoryless and $\lambda \in \mathbb{N}$ then the sequence

$$v_0^{\lambda\tau} := 0;\ v_n^{\lambda\tau} := U_\tau^{(\lambda)} v_{n-1}^{\lambda\tau} \text{ ,}$$

is nondecreasing and converges to $U_{A_\infty} 0$. Furthermore

$$v_{n-1}^\tau \leq v_{n-1}^{\lambda\tau} \leq v_n^{\lambda\tau} \leq L_{A_\infty}^{D_n} 0 \leq U_{A_\infty} 0 \text{ ,}$$

where D_n is the memoryless strategy found by applying U_τ on $v_{n-1}^{\lambda\tau}$.

Proof. Since $r^k(i) \geq 0$ (see the remark at the end of section 1) we have

$$U_\tau v_0^{\lambda\tau} = U_\tau v_0^\tau = U_\tau 0 = L_\tau^{D_1} 0 \geq 0$$

so because of the monotony of $L_\tau^{D_1}$

$$v_1^{\lambda\tau} = (L^{D_1})^\lambda 0 \geq (L_\tau^{D_1})^{\lambda-1} \geq \ldots \geq L_\tau^{D_1} 0 = U_\tau 0 \geq v_1^\tau \text{ .}$$

The proof proceeds further in an inductive way using the fact that U_τ and L_τ^D are monotone contractions and the fact that v_n^τ converges monotonously from below to $U_{A_\infty} 0$.

Assertion. Actually $L_\tau^{(\lambda)} v$ is a better estimate for $L_{A_\infty}^{D_v} 0$ than $U_\tau v$, where D_v is the strategy that is found by applying U_τ on v.

For $\tau \equiv 1$ this assertion is illustrated in [7]. In general the statement follows from the following considerations:
Let τ be transition memoryless, let v and w be given such that $w \geq v$ and $w := U_\tau v = L_\tau^{D_v} v$. Now from the previous section we know that

$$L_{A_\infty}^{D_v} 0 = \lim_{n\to\infty} (L_\tau^{D_v})^n v = \lim_{n\to\infty} \left\{ w + \sum_{k=1}^{n-1} \left[(L_\tau^{D_v})^k w - (L_\tau^{D_v})^k v \right] \right\} .$$

$w \geq v$ and the contraction property of $L_\tau^{D_v}$ imply

$$0 \leq (L_\tau^{D_v})^k w - (L_\tau^{D_v})^k v \leq (\rho_\tau^{D_v})^k \, \| w-v \|_\infty .$$

Since

$$U_\tau^{(\lambda)} v = w + \sum_{k=1}^{\lambda} \left[(L_\tau^{D_v})^k w - (L_\tau^{D_v})^k v \right]$$

the statement will be clear.

Remark. If τ is nonzero and $\lambda \equiv \infty$, then the algorithm of theorem 2.1 is clearly of the policy iteration type: in each step the values of the current policy are computed exactly. The choice of τ only influences the way of looking for possible improvement: If $\tau = 1$, the methode equals Howard's policy iteration algorithm [4], [11]. If τ is replaced by the stopping time induced by the go ahead set A_H, we get Hasting's modified policy iteration algorithm [8]. A great number of other choices is possible, e.g. τ as induced by A_R.
Now, regardless of the restriction $r^k(i) \geq 0$, each iteration step brings a strict improvement in the values $v_n^{\infty\tau}$, until the optimum is reached, which occurs after a finite number of steps (since only finitely many memoryless strategies are available).

§ 3. Upper and Lower bounds

If the theory developed in the previous sections is used for generating successive approximation algorithms it will be necessary to construct upper- and lower bounds for the optimal return $U_\infty 0$ and for the return of $L_\infty^{D_n} 0$ of the strategy D_n occurring in the n-th iteration step.

Furthermore upper and lower bounds enable us to incorporate a test for the suboptimality of policies see for instance [13], [14], [15]. Such a test may be based on the following idea:

Lemma 3.1. Let the upper bound $\bar{x}$ and the lower bound $\underline{x}$ for the optimal return $U_\infty 0$ be given i.e. $\underline{x} \le U_\infty 0 \le \bar{x}$ then decision rule D_0 is not optimal if $L_\tau^{D_0} \bar{x} < U_\tau \underline{x}$ (where $v < w$ means $v(i) \le w(i)$ and for at least one component: $v(i) < w(i)$).

Proof. $U_\infty 0 = U_\tau(U_\infty 0) \ge U_\tau \underline{x} > L_\tau^{D_0} \bar{x} \ge L_\tau^{D_0}(U_\infty 0)$ where the monotony of U_τ and L_τ is used.

Let us now return to the upper and lower bounds.

Lemma 3.2. For τ transition memoryless. The sequence

$$\bar{v}_n^\tau := v_n^\tau + \frac{\nu_\tau}{1-\nu_\tau} \max_{i \in S} \left(v_n^\tau(i) - v_{n-1}^\tau(i)\right) \cdot e$$

yields a sequence of nonincreasing upper bounds for $U_\infty 0$; and $\lim_{n\to\infty} \bar{v}_n^\tau = U_\infty 0$. Here $e \in \mathbb{R}^N$ and $e(i) = 1$, $i \in \{1,2,\ldots,N\}$ and ν_τ is the contraction radius of U_τ.

Proof. $U_\infty 0 = \lim_{k\to\infty} \left(L_\tau^{D^*}\right)^k v_{n-1}^\tau$ where D is an optimal decision rule.

However

$$\left(L_\tau^{D^*}\right)^\ell v_{n-1}^\tau = v_{n-1}^\tau + \left(L_\tau^{D^*} v_{n-1}^\tau - v_{n-1}^\tau\right) + \ldots + \left(\left(L_\tau^{D^*}\right)^\ell v_{n-1}^\tau - \left(L_\tau^{D^*}\right)^{\ell-1} v_{n-1}^\tau\right)$$

$$\le v_{n-1}^\tau + \sum_{k=0}^{\ell-1} \left(\rho_\tau^{D^*}\right)^k \max_{i\in S} \left(L_\tau^{D^*} v_{n-1}^\tau(i) - v_{n-1}^\tau(i)\right) \cdot e$$

$$\le v_{n-1}^\tau + \sum_{k=0}^{\ell-1} (\nu_\tau)^k \max_{i\in S} \left(v_n^\tau(i) - v_{n-1}^\tau(i)\right) \cdot e$$

taking the limit for ℓ to infinity gives the assertion.

Lemma 3.2. For τ transition memoryless, the sequence $\{\underline{v}_n^\tau\}$ defined as follows:

$$\underline{v}_n^\tau = \max\{v_n^\tau + \frac{\eta_\tau^{D_n}}{1 - \eta_\tau^{D_n}} \cdot \min_{i \in S} (v_n^\tau(i) - v_{n-1}^\tau(i)) \cdot e,\ \underline{v}_{n-1}^\tau\}$$

where $\eta_\tau^{D_n} := \min_{i \in S} \{\mathbb{E}_{D_n} (\beta^\tau | x_0 = i)\}$, yields a nondecreasing sequence of lower bounds for $L_\infty^{D_n} 0$ and thus $U_\infty 0$. Furthermore

$$\lim_{n\to\infty} \underline{v}_n^\tau = U_\infty 0 .$$

Lemma 3.3. For τ transition memoryless, $\lambda \in \mathbb{N}$, the sequence $\{\bar{v}_n^{\lambda\tau}\}$ defined as follows:

$$\bar{v}_1^{\lambda\tau} = v_o^{\lambda\tau} + \frac{1}{1 - \nu_\tau} \max_{i \in S} (U_\tau v_0^{\lambda\tau}(i) - v_0^{\lambda\tau}(i)) \cdot e$$

$$\bar{v}_n^{\lambda\tau} = \min\{\bar{v}_{n-1}^{\lambda\tau},\ v_{n-1}^{\lambda\tau} + \frac{1}{1 - \nu_\tau} \max_{i \in S} (U_\tau v_{n-1}^{\lambda\tau}(i) - v_{n-1}^{\lambda\tau}(i)) \cdot e\},\ n > 1$$

yields a nonincreasing sequence of upper bounds for $U_\infty 0$, with

$$\lim_{n\to\infty} \bar{v}_n^{\lambda\tau} = U_\infty 0 .$$

Lemma 3.4. For τ transition memoryless, $\lambda \in \mathbb{N}$, the sequence $\{\underline{v}_n^{\lambda\tau}\}$ defined as follows:

$$\underline{v}_1^{\lambda\tau} := v_0^{\lambda\tau} + \frac{1}{1 - \eta_\tau^{D_1}} \min_{i \in S} (U_\tau v_0^{\lambda\tau}(i) - v_0^{\lambda\tau}(i)) \cdot e$$

$$\underline{v}_n^{\lambda\tau} := \max\{\underline{v}_{n-1}^{\lambda\tau},\ v_{n-1}^{\lambda\tau} + \frac{1}{1 - \eta_\tau^{D_n}} \min_{i \in S} (U_\tau v_{n-1}^{\lambda\tau}(i) - v_{n-1}^{\lambda\tau}(i)) \cdot e\}$$

yields a nondecreasing sequence of lower bounds for $L_\infty^{D_n} 0$ and thus for $U_\infty 0$, again we have

$$\lim_{n\to\infty} \underline{v}_n^{\lambda\tau} = U_\infty 0 .$$

The proofs of the last three lemma's proceed in a similar way as the proof of lemma 3.1. For special stopping times see also [3].

Examples. 3.1. For $\tau \equiv k$, $\nu_\tau = \beta^k$; $\eta_\tau^{D_n} = \beta^k$ independent of D_n.

3.2. If τ corresponds with A_H $\nu_\tau = \beta$ and $\eta_\tau^{D_n} = \beta^N$, again independent of D_n.

3.3. If τ corresponds with A_R $\nu_\tau = \max\limits_{i,k} \beta \dfrac{1 - p_{ii}^k}{1 - \beta p_{ii}^k}$

$$\eta_\tau^{D_n} := \min_{i \in S} \beta \frac{1 - p_{ii}^{D_n(i)}}{1 - \beta p_{ii}^{D_n(i)}} .$$

See also [3].

§ 4. Extensions and remarks.

The ideas which have been presented in the previous sections may also be used in the case of a semi-Markov decision process (e.g. [5], [6]).

In this paper we only considered pure stopping times. We avoided the use of mixed stopping times in order to maintain a better sight of the basic ideas. However, the introduction of mixing for stopping times produces many more algorithms and even two already published ones: viz. the policy improvement algorithm of Reetz [2] and a linear programming algorithm (e.g. [5], [6]) with a random choice of the new basic variable from the relevant ones.

In section 2 we introduced policy improvement-value determination procedures characterized by a stopping time τ and a natural number λ. For the proofs it is not essential that λ is fixed for all random steps. The value of λ may depend on the number of the iteration and even on specific aspects of the actual iteration process, see also [3].

For numerical experience with a number of the methods treated in this paper we refer to [7].

References

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

[2] D. Reetz, Solution of a Markovian decision problem by successive over-relaxation, Z.f. Oper. Res. 17 (1973) 29-32.

[3] J.A.E.E. van Nunen, Improved successive approximation methods for discounted Markov decision processes, in these Proceedings.

[4] R.A. Howard, Dynamic programming and Markov processes, MIT-Press, Cambridge, 1960.

[5] G.T. de Ghellinck, G.D. Eppen, Linear programming solutions for separable Markovian decision problems, Man. Sci. 13 (1967) 371-394.

[6] J. Wessels, J.A.E.E. van Nunen, Discounted semi-Markov decision processes: linear programming and policy iteration, to appear in Statistica Neerlandica 29 (1975) nr.1.

[7] J.A.E.E. van Nunen, A set of successive approximation methods for discounted Markovian decision problems, submitted to Z.f. Oper. Res.

[8] N. Hastings, Some notes on dynamic programming and replacement, Oper. Res. Q. 19 (1968) 453-464.

[9] H. Schellhaas, Zur Extrapolation in Markoffschen Entscheidungsmodellen mit Diskontierung, Preprint nr. 84 (1973), Technische Hochschule Darmstadt.

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

[11] H. Mine, S. Osaki, Markovian decision processes, New York 1970, submitted to Proceedings of 1974 EMS-meeting and 7th Prague Conference on Information theory, Statistical decision functions, and random processe

[12] J. Wessels, Stopping times and Markov programming, submitted to Proceedings of 1974 EMS-meeting and 7th Prague conference on Information Theory, Statistical decision functions, and random processes.

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

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

[15] R.C. Grinold, Elimination of suboptimal actions in Markov decision problems. Oper. Res. 21 (1973) 848-851.