A stopping time-based policy iteration algorithm for Markov decision processes with discountfactor tending to 1

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 78-24

A stopping time-based policy iteration algorithm for Markov decision processes with discountfactor tending to 1

by

J. van der Wal

Eindhoven, November 1978

The Netherlands

A stopping time-based policy iteration algorithm for Markov decision processes with discountfactor tending to 1

by

J. van der Wal

Abstract. This paper considers the Markov decision process with finite state and action spaces, when the discountfactor tends to 1. Miller and Veinott have shown the existence of n-discount optimal policies and Veinott has given an algorithm to determine one. In this paper we use the stopping times as introduced by Wessels to generate a set of modified policy iteration algorithms for the determination of an n-discount optimal strategy.

Introduction and notations. In this paper we consider the discounted Markov decision process (MDP) with finite state and action spaces when the discountfactor β tends to 1. We are interested in finding n-discount optimal policies. The notion of $n^{(+)}$-discount optimality stems from Miller and Veinott [3]. As we know (-1)-discount optimality corresponds to average (or gain) optimality and 0-discount optimality to bias optimality. In [3] the existence of n-discount optimal policies has been shown and Veinott [4] has shown how to determine n-discount optimal policies with an extended (and adapted) version of Howard's Policy Iteration Algorithm (PIA) [2].
In a previous paper [6] we gave a variant of Howard's PIA based on a finite transition memoryless stopping time to determine an average optimal policy. Here we extend this stopping time based approach to determine n-discount optimal policies. An example of such a stopping time based algorithm is the Gauss-Seidel version of Howard's PIA.
So, we are looking at a discrete-time MDP with finite state space $S = \{1,2,..,N\}$ and finite action space A. If in state i action a is taken then the immediate reward is r(i,a) and the system moves to state j with probability p_{ij}^a. A policy or stationary strategy is a map from S into A. Each $i \in S$ and policy f determine a probability measure $\mathbb{P}_{i,f}$ on $(S \times A)^\infty$ and a stochastic process $\{(X_n,A_n),\ n = 0,1,...\}$ where X_n is the state and A_n the action taken at time n. The expectation with respect to $\mathbb{P}_{i,f}$ will be denoted by $\mathbb{E}_{i,f}$.
In Wessels [7] stopping times are used to generate successive approximation algorithms. Following the same approach we define a nonzero, finite and transition memoryless stopping time τ as a map from S^∞ into $\bar{\mathbb{N}} = \{1,2,...,\infty\}$ such that for all i and f $\mathbb{P}_{i,f}(\tau < \infty) = 1$ and that τ can be completely characterized by a set $T \subset S^2$ such that (cf. [7,6])

$$\tau(x_0,x_1,\ldots) = n \Leftrightarrow (x_0,x_1) \notin T,\ldots,(x_{n-2},x_{n-1}) \notin T,\ (x_{n-1},x_n) \in T.$$

Here we consider only this type of stopping times. As a consequence of this transition memorylessness we can restrict ourselves to policies (cf. lemma 3.1 and 3.2 in [6]). In the remainder of this paper τ and T are fixed.
We want to introduce a few more notations. Let f be a policy then define the vectors r_f and $r_{\beta,\tau,f}$ and the matrices P_f, P_f^* and $P_{\beta,\tau,f}$ by

$$r_f(i) = r(i,f(i))$$

$$r_{\beta,\tau,f}(i) = \mathbb{E}_{i,f} \sum_{n=0}^{\tau-1} \beta^n r(X_n,A_n) \qquad \text{(cf. [7])}$$

$$P_f(i,j) = p_{ij}^{f(i)}$$

$$P_f^* = \lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N-1} P_f^n$$

$$P_{\beta,\tau,f}(i,j) = \sum_{n=1}^{\infty} \beta^n \mathbb{P}_{i,f}(X_\tau = j,\ \tau = n) \ .$$

Further we define the matrices $\bar{P}_f$ and $\tilde{P}_f$ (suppressing the dependence on τ) by

$$\bar{P}_f(i,j) = \begin{cases} p_{ij}^{f(i)} & \text{if } (i,j) \notin T \\ 0 & \text{if } (i,j) \in T \end{cases}$$

$$\tilde{P}_f(i,j) = \begin{cases} 0 & \text{if } (i,j) \notin T \\ p_{ij}^{(i)} & \text{if } (i,j) \in T \ . \end{cases}$$

Then we have

Lemma 1.1.

i) $P_f = \bar{P}_f + \tilde{P}_f$

ii) $P_{\beta,\tau,f} = \beta\tilde{P}_f + \beta^2\bar{P}_f\tilde{P}_f + \beta^3\bar{P}_f^2\tilde{P}_f + \ldots = (I - \beta\bar{P}_f)^{-1}\beta\tilde{P}_f$

iii) $r_{\beta,\tau,f} = r_f + \beta\bar{P}_f r_f + \beta^2\bar{P}_f^2 r_f + \ldots = (I - \beta\bar{P}_f)^{-1} r_f \ .$

From the finiteness of τ it follows that $I - \bar{P}_f$ is nonsingular so that ii) and iii) also hold for $\beta = 1$. We will write $r_{\tau,f}$ and $P_{\tau,f}$ instead of $r_{1,\tau,f}$ and $P_{1,\tau,f}$.

The total expected discounted reward under policy f, denoted by $v_{\beta,f}$, satisfies

$$v_{\beta,f} = \sum_{n=0}^{\infty} (\beta P_f)^n r_f .$$

A policy f is n-discount optimal (n = -1,0,...) if

$$(1.1) \qquad \limsup_{\beta \uparrow 1} (1-\beta)^n (v_{\beta,f} - v_{\beta,g}) \geq 0 \quad \text{for all } g .$$

And policy f is called ∞-discount optimal if f is n-discount optimal for all n = -1,0,1,... .

For $v_{\beta,f}$ we also have the Laurent series expansion in $(1-\beta)$ for $\beta \uparrow 1$

$$v_{\beta,f} = \sum_{n=-1}^{\infty} (1-\beta)^n c_{n,f} .$$

(Miller and Veinott [3] used the expansion in ρ, with $\beta = (1+\rho)^{-1}$, but in our case the expansion in $(1-\beta)$ gives the simpler expressions).

The terms $c_{n,f}$ can be obtained as follows

$$\begin{aligned} v_{\beta,f} &= [I + \beta P_f + \beta^2 P_f^2 + \ldots] r_f \\ &= [I + \beta(P_f - P_f^*) + \beta^2 (P_f - P_f^*) + \ldots] r_f + (1-\beta)^{-1} P_f^* r_f - P_f^* r_f . \end{aligned}$$

With $P_f^n - P_f^* = (P_f - P_f^*)^n$, n = 1,2,... (from $P_f P_f^* = P_f^* P_f = P_f^*$, cf. [1]) we get

$$(1.2) \qquad v_{\beta,f} = (1-\beta)^{-1} P_f^* r_f - P_f^* r_f + [I - \beta(P_f - P_f^*)]^{-1} r_f .$$

If I - S is nonsingular and β is sufficiently close to 1 then we have the expansion

$$\begin{aligned} (1.3) \qquad (I - \beta S)^{-1} &= (I - S + (1-\beta)S)^{-1} = [I + (1-\beta) S (I-S)^{-1}]^{-1} (I-S)^{-1} \\ &= \sum_{k=0}^{\infty} (-1)^k (1-\beta)^k [S(I-S)^{-1}]^k (I-S)^{-1} . \end{aligned}$$

Since $I - P_f + P_f^*$ is nonsingular (lemma 1d in [1]) we may substitute (1.3) in (1.2) to obtain

$$(1.4) \qquad \begin{cases} c_{-1,f} = P_f^* r_f \\ c_{0,f} = [(I-S)^{-1} - P_f^*] r_f \\ c_{k,f} = (-1)^k [S(I-S)^{-1}]^k (I-S)^{-1} r_f \end{cases}$$

with $S = P_f - P_f^*$.

For any two policies f and g we define

$$\Delta c_{n,f,g} := c_{n,f} - c_{n,g}, \quad n = -1,0,\ldots . \tag{1.5}$$

And we define

$$f \overset{n}{\geq} g$$

if for all $i \in S$ the first nonzero element, if any, in the row $\Delta c_{-1,g,f}(i)$, $\Delta c_{0,g,f}(i),\ldots,\Delta c_{n,f,g}(i)$ is positive (cf. Miller and Veinott [3]). Further we write $f \geq g$ if $f \overset{n}{\geq} g$ for all $n = -1,0,\ldots$. So $\overset{n}{\geq}$ and $\geq$ are partial orderings on the set of policies.

We see that a policy f is n-discount optimal [∞-discount optimal] if and only if $f \overset{n}{\geq} g$ [$f \geq g$] for all g.

It is straightforward that our notion of n-discount optimality is identical to the n^+ discount optimality in Veinott [5] as $\lim_{\beta\uparrow 1} (1-\beta)/\rho = 1$ $(\beta = (1+\rho)^{-1})$.

In section 2 we will derive a Laurent series expansion for $r_{\beta,\tau,g} + P_{\beta,\tau,g}v_{\beta,f}$, from which we obtain the PIA formulated in section 4. In section 5 we show that the policy improvement step of this algorithm indeed improves the policy. And in section 6 we show that our modified PIA produces an n-discount optimal policy.

2. The Laurent series expansion for $r_{\beta,\tau,g} + P_{\beta,\tau,g}v_{\beta,f}$. Performing a stopping time based successive approximation step on $v_{\beta,f}$ means maximize over g

$$r_{\beta,\tau,g} + P_{\beta,\tau,g}v_{\beta,f} \quad \text{(cf. Wessels [7])} . \tag{2.1}$$

For (2.1) we can derive a Laurent series expansion as follows: Substitute in (2.1) lemma 1.1(ii) and (iii) and use expansion (1.3) with $S = \bar{P}_g$ to obtain

$$\begin{aligned} r_{\beta,\tau,g} + P_{\beta,\tau,g}v_{\beta,f} &= (I - \beta\bar{P}_g)^{-1}[r_g + (\tilde{P}_g - (1-\beta)\tilde{P}_g)v_{\beta,f}] \\ &= \sum_{n=0}^{\infty} (-1)^n(1-\beta)^n[\bar{P}_g(I-\bar{P}_g)^{-1}]^n(I-\bar{P}_g)^{-1}\{r_g + \\ &\qquad [1-(1-\beta)]\tilde{P}_g \sum_{k=-1}^{\infty} (1-\beta)^k c_{k,f}\} . \end{aligned} \tag{2.2}$$

And we find for the coefficient $d_{k,g,f}$ of $(1-\beta)^k$ in (2.2)

$$\begin{aligned} d_{-1,g,f} &= (I - \bar{P}_g)^{-1}\tilde{P}_g c_{-1,f} \\ d_{n,g,f} &= (-1)^n[\bar{P}_g(I-\bar{P}_g)^{-1}]^n(I-\bar{P}_g)^{-1}r_g + \end{aligned} \tag{2.3}$$

$$\sum_{\ell=0}^{n+1} (-1)^{\ell}[\bar{P}_g(I-\bar{P}_g)^{-1}]^{\ell}(I-\bar{P}_g)^{-1}\tilde{P}_g c_{n-\ell,f} +$$
$$\sum_{\ell=0}^{n} (-1)^{\ell}[\bar{P}_g(I-\bar{P}_g)^{-1}]^{\ell}(I-\bar{P}_g)^{-1}\tilde{P}_g c_{n-\ell-1,f} .$$

With the notations $r_{\tau,g}$, $P_{\tau,g}$ and

$$Q_{\tau,g} = (I - \bar{P}_g)^{-1}$$
$$R_{\tau,g} = \bar{P}_g(I - \bar{P}_g)^{-1}$$

(2.3) simplifies to

(2.4) $$d_{-1,g,f} = P_{\tau,g}c_{-1,f}$$

(2.5) $$d_{0,g,f} = r_{\tau,g} + P_{\tau,g}c_{0,f} - Q_{\tau,g}P_{\tau,g}c_{-1,f}$$

$$d_{n,g,f} = (-1)^n R^n_{\tau,g} r_{\tau,g} + \sum_{\ell=0}^{n+1} R^{\ell}_{\tau,g}P_{\tau,g}c_{n-\ell,f} + \sum_{\ell=0}^{n} (-1)^{\ell+1} R^{\ell}_{\tau,g}P_{\tau,g}c_{n-\ell-1,f}.$$

The expression for $d_{n,g,f}$ can be simplified further to the recursion

(2.6) $$d_{n,g,f} = (-R_{\tau,g})d_{n-1,g,f} + P_{\tau,g}(c_{n,f} - c_{n-1,f}), \quad n \geq 1 .$$

If we maximize (2.1) for β sufficiently close to 1 then we maximize "lexicografically" the first terms of the expansion (2.2), i.e.

> first maximize $d_{-1,g,f}$, then maximize $d_{0,g,f}$ over the set of maximizers of $d_{-1,g,f}$ etc.

In [6] we showed that a policy improvement step which subsequently maximizes $d_{-1,g,f}$ and $d_{0,g,f}$ gives a convergent algorithm and produces an average optimal strategy. Here we extend this result and we show that an algorithm with as improvement step the maximization of $d_{-1,g,f},\ldots,d_{n,g,f}$, produces an $(n-1)$-discount optimal strategy.

3. Some equations. In this section we collect a number of equations we need in the sequel.

In the first part of this section we derive from equations (2.4)-(2.6) a set of equivalent equations.

Let f be the current policy and g an arbitrary policy. Define

(3.1) $$\psi_{k,g,f} := d_{k,g,f} - c_{k,f}, \quad k = -1,0,\ldots .$$

From the definitions of $r_{\tau,g}$, $P_{\tau,g}$, $Q_{\tau,g}$ and $R_{\tau,g}$ we have

$$
\begin{aligned}
r_{\tau,g} &= r_g + \bar{P}_g r_{\tau,g} \\
P_{\tau,g} &= \tilde{P}_g + \bar{P}_g P_{\tau,g} \\
Q_{\tau,g} &= I + \bar{P}_g Q_{\tau,g} \\
R_{\tau,g} &= \bar{P}_g + \bar{P}_g R_{\tau,g} \ .
\end{aligned} \tag{3.2}
$$

If we substitute (3.1) and (3.2) in (2.4)-(2.6) we get

$$\tilde{P}_g c_{-1,f} + \bar{P}_g (c_{-1,f} + \psi_{-1,g,f}) = c_{-1,f} + \psi_{-1,g,f} \tag{3.3}$$

$$r_g + \tilde{P}_g c_{0,f} - (c_{-1,f} + \psi_{-1,g,f}) + \bar{P}_g (c_{0,f} + \psi_{0,g,f}) = c_{0,f} + \psi_{0,g,f} \tag{3.4}$$

$$-\bar{P}_g (c_{k-1,f} + \psi_{k-1,g,f}) + \tilde{P}_g (c_{k,f} - c_{k-1,f}) + \bar{P}_g (c_{k,f} + \psi_{k,g,f}) = c_{k,f} + \psi_{k,g,f} . \tag{3.5}$$

In order to rewrite (3.3)-(3.5) componentwise, define

$$T_i := \{ j \in S \mid (i,j) \in T \} \ . \tag{3.6}$$

Then we have for all $v \in \mathbb{R}^N$

$$(\tilde{P}_g v)(i) = \sum_{j \in T_i} p_{ij}^{g(i)} v(j) \quad \text{and} \quad (\bar{P}_g v)(i) = \sum_{j \notin T_i} p_{ij}^{g(i)} v(j) \ . \tag{3.7}$$

If we substitute this into (3.3)-(3.5) we get the componentwise formulation of (3.3)-(3.5).

$$\sum_{j \in T_i} p_{ij}^{g(i)} c_{-1,f}(j) + \sum_{j \notin T_i} p_{ij}^{g(i)} (c_{-1,f} + \psi_{-1,g,f})(j) = (c_{-1,f} + \psi_{-1,g,f})(i) \tag{3.8}$$

$$\begin{aligned}
r(i,g(i)) + \sum_{j \in T_i} p_{ij}^{g(i)} c_{0,f}(j) - (c_{-1,f} + \psi_{-1,g,f})(i) + \\
+ \sum_{j \notin T_i} p_{ij}^{g(i)} (c_{0,f} + \psi_{0,g,f})(j) = (c_{0,f} + \psi_{0,g,f})(i)
\end{aligned} \tag{3.9}$$

$$\begin{aligned}
-\sum_{j \notin T_i} p_{ij}^{g(i)} (c_{k-1,f} + \psi_{k-1,g,f})(j) + \sum_{j \in T_i} p_{ij}^{g(i)} (c_{k,f} - c_{k-1,f})(j) + \\
+ \sum_{j \notin T_i} p_{ij}^{g(i)} (c_{k,f} + \psi_{k,g,f})(j) = (c_{k,f} + \psi_{k,g,f})(i) \ .
\end{aligned} \tag{3.10}$$

So (3.8)-(3.10) follow from (2.4)-(2.6). That (3.8)-(3.10) is even equivalent to (2.4)-(2.6) is immediate from the finiteness of the stopping time τ. This we see as follows. Clearly (3.8)-(3.10) and (3.3)-(3.5) are equivalent. And as τ is finite $I - \bar{P}_g$ is nonsingular. Multiplying (3.3)-(3.5) by $(I - \bar{P}_g)^{-1}$ gives us (2.4)-(2.6).

In the second part of this section we derive some relations between the $\Delta c_{k,g,f}$ and the $\psi_{k,g,f}$. Clearly we have from $r_{\beta,\tau,f} + P_{\beta,\tau,f} v_{\beta,f} = v_{\beta,f}$ (cf. lemma 1.1 in Wessels [7]) that $d_{k,f,f} = c_{k,f}$ so

(3.11;f) $P_{\tau,f} c_{-1,f} = c_{-1,f}$

(3.12;f) $r_{\tau,f} + P_{\tau,f} c_{0,f} - Q_{\tau,f} P_{\tau,f} c_{-1,f} = c_{0,f}$

(3.13;f) $(-R_{\tau,f}) c_{k-1,f} + P_{\tau,f}(c_{k,f} - c_{k-1,f}) = c_{k,f}$.

If we subtract (2.4)-(2.6) from (3.11;g)-(3.13;g) and substitute (3.1) and (1.5) we get

(3.14) $P_{\tau,g} \Delta c_{-1,g,f} = \Delta c_{-1,g,f} - \psi_{-1,g,f}$

(3.15) $P_{\tau,g} \Delta c_{0,g,f} - Q_{\tau,g} \Delta c_{-1,g,f} = \Delta c_{0,g,f} - \psi_{0,g,f}$

(3.16) $(-R_{\tau,g})(\Delta c_{k-1,g,f} - \psi_{k-1,g,f}) + P_{\tau,g}(\Delta c_{k,g,f} - c_{k-1,g,f}) =$
$= \Delta c_{k,g,f} - \psi_{k,g,f}, \quad k \geq 1$.

4. <u>The modified policy improvement step</u>. In section 2 we have seen that if $\beta \uparrow 1$ the stopping time-based successive approximation step first maximizes $d_{-1,g,f}$ then $d_{0,g,f}$ etc. In [6] where we only considered $d_{-1,g,f}$ and $d_{0,g,f}$ we gave the following approach.

Define $\psi_{-1,f}$ by

(4.1) $\psi_{-1,f} := \max_g \psi_{-1,g,f} = \max_g P_{\tau,g} c_{-1,f} - c_{-1,f}$.

Then we have for all a

(4.2) $\sum_{j \in T_i} p_{ij}^a c_{-1,f}(j) + \sum_{j \notin T_i} p_{ij}^a (c_{-1,f} + \psi_{-1,f})(j) \leq (c_{-1,f} + \psi_{-1,f})(i)$.

Since, suppose the lhs in (4.2) is greater than the rhs for some a. And let g be a maximizer in (4.1) then we see from (3.8) that (4.2) holds with equality for g(i). Now consider the policy h with h(i) = a and h(j) = g(j), $j \neq i$. Then from (4.2)

(4.3) $\tilde{P}_h c_{-1,f} + \bar{P}_h (c_{-1,f} + \psi_{-1,f}) \geq (c_{-1,f} + \psi_{-1,f})$,

so

(4.4) $(I - \bar{P}_h)^{-1} \tilde{P}_h c_{-1,f} = P_{\tau,h} c_{-1,f} \geq c_{-1,f} + \psi_{-1,f}$,

with strict inequality in the i-th component. But this contradicts (4.1).

Define

(4.5) $A_{-1}(i,f)$:= the set of actions for which (4.2) holds with equality.

And

(4.6) $G_{-1}(f) := \{g \mid g(i) \in A_{-1}(i,f) \text{ for all } i \in S\}$.

For any policy $g \in G_{-1}(f)$ (4.3) and (4.4) will hold with equality, so $G_{-1}(f)$ is the set of maximizers of (4.1). Continuing in this way we define

(4.7) $$\psi_{0,f} := \max_{g \in G_{-1}(f)} \psi_{0,g,f} .$$

Then for all $a \in A_{-1}(i,f)$

(4.8) $$r(i,a) + \sum_{j \in T_i} p_{ij}^a c_{0,f}(j) - (c_{-1,f} + \psi_{-1,f})(i) + \sum_{j \notin T_i} p_{ij}^a (c_{0,f} + \psi_{0,f})(j) \leq$$

$$\leq (c_{0,f} + \psi_{0,f})(i) .$$

If we define further

(4.9) $A_0(i,f)$:= the set of $a \in A_{-1}(i,f)$ for which (4.8) holds with equality

(4.10) $G_0(f) := \{g \mid g(i) \in A_0(i,f) \text{ for all } i \in S\}$.

Then again $G_0(f)$ is precisely the set of maximizers of (4.7). In [6] we proved that a policy iteration algorithm with as improvement step the determination of a policy g in $G_0(f)$ with g equal to f whenever possible ($g(i) = f(i)$ if $f(i) \in A_0(i,f)$), converges and produces an average optimal policy. I.e. a policy h with $h \overset{-1}{\geq} g$ for all g.

Here we extend the policy improvement step in the following way. Define

(4.11) $$\psi_{k,f} := \max_{g \in G_{k-1}(f)} \psi_{k,g,f}, \quad k = 1,2,...$$

(4.12) $A_k(i,f)$:= the set of $a \in A_{k-1}(i,f)$ for which (4.13) below holds with equality, $k = 1,2,...$

(4.13) $$-\sum_{j \notin T_i} p_{ij}^a (c_{k-1,f} + \psi_{k-1,f})(j) + \sum_{j \in T_i} p_{ij}^a (c_{k,f} - c_{k-1,f})(j) +$$

$$+ \sum_{j \notin T_i} p_{ij}^a (c_{k,f} + \psi_{k,f})(j) \leq (c_{k,f} + \psi_{k,f})(i)$$

(4.14) $G_k(f) := \{g \mid g(i) \in A_k(i,f) \text{ for all } i \in S\}$, $k = 1,2,...$.

In the same way as before one may show that (4.13) holds for all $a \in A_{k-1}(i,f)$ and that g maximizes (4.11) within $G_{k-1}(f)$ if and only if $g \in G_k(f)$.
Now we can propose the following modified policy iteration algorithm.

Value determination step

Let f be the current policy. Determine $c_{-1,f}, \ldots, c_{n,f}$.

(4.15) *Policy improvement step*

Determine a policy $g \in G_n(f)$ with $g(i) = f(i)$ whenever $f(i) \in A_n(i,f)$.

In the next sections we will show that this modified PIA converges and terminates with an (n - 1)-discount optimal strategy.

5. The policy improvement step. In this section we prove that the policy improvement step (4.15) produces a policy g which is at least as good as f with respect to the first n + 2 terms of the Laurent series expansions for $v_{\beta,g}$ and $v_{\beta,f}$. And that these terms can only be two by two equal if the newly produced policy is identical to the old one:

Theorem 5.1. Let f be an arbitrary policy and $g \in G_n(f)$ with $g(j) = f(j)$ whenever $f(j) \in A_n(j,f)$, $j \in S$ then

i) $g \overset{n}{\geq} f$.

ii) $g \overset{n}{=} f$ only if $g = f$ ($g \overset{n}{=} f \Leftrightarrow g \overset{n}{\geq} f$ and $f \overset{n}{\geq} g$).

In order to prove this we need the following lemma.

Lemma 5.2. Let f be an arbitrary policy then

i) $\psi_{-1,f} \geq 0$

and if $\psi_{-1,f}(i) = \ldots = \psi_{k,f}(i) = 0$ then

ii) $\psi_{-1,f}(j) = \ldots = \psi_{k,f}(j) = 0$ for all $j \in V(i,f) := \{\ell \notin T_i \mid p_{i\ell}^{f(i)} > 0\}$.

iii) $f(i) \in A_k(i,f)$.

iv) $\psi_{k+1,f}(i) \geq 0$.

Proof. i) From (3.11;f) we have

$$c_{-1,f} + \psi_{-1,f} = \max_g P_{\tau,g} c_{-1,f} \geq P_{\tau,f} c_{-1,f} = c_{-1,f} ,$$

hence $\psi_{-1,f} \geq 0$.

ii)-iv) we prove by induction.

k = -1. Assume $\psi_{-1,f}(i) = 0$. Then from (4.2)

$$(5.1)\qquad \sum_{j\in T_i} p_{ij}^{f(i)} c_{-1,f}(i) + \sum_{j\notin T_i} p_{ij}^{f(i)} (c_{-1,f} + \psi_{-1,f})(j) \le$$

$$\le (c_{-1,f} + \psi_{-1,f})(i) = c_{-1,f}(i) \ .$$

Also from (3.11;f) and (3.2) and (3.7)

$$(5.2)\qquad \sum_{j\in T_i} p_{ij}^{f(i)} c_{-1,f}(j) + \sum_{j\notin T_i} p_{ij}^{f(i)} c_{-1,f}(j) = c_{-1,f}(i) \ .$$

Subtracting (5.2) from (5.1) we get

$$\sum_{j\notin T_i} p_{ij}^{f(i)} \psi_{-1,f}(j) \le 0$$

which together with $\psi_{-1,f} \ge 0$ yields

$$\psi_{-1,f}(j) = 0 \text{ for all } j \in V(i,f)$$

and (5.1) [(4.2) with $a = f(i)$] holds with equality so

$$f(i) \in A_{-1}(i,f) \ .$$

Next, let $W_{-1}(f) := \{j \mid \psi_{-1,f}(j) = 0\}$, then $W_{-1}(f)$ is closed under $\bar{P}_f$. Further $f(i) \in A_{-1}(i,f)$ for all $i \in W_{-1}(f)$. Now let $\bar{f}$ be any policy with $\bar{f}(i) = f(i)$ on $W_{-1}(f)$ and $\bar{f}(i) \in A_{-1}(i,f)$ else, then $\bar{f} \in G_{-1}(f)$. So if the system starts in $i \in W_{-1}(f)$ and we use policy $\bar{f}$ then the system will not leave $W_{-1}(f)$ before τ, therefore it uses only actions from f. So

$$\psi_0(i) = \max_{g\in G_{-1}(f)} \psi_{0,g,f}(i) \ge \psi_{0,\bar{f},f}(i) = \psi_{0,f,f}(i) = 0$$

which completes the proof for $n = -1$.

Let $W_\ell(f) := \{j \mid \psi_{-1,f}(j) = \ldots = \psi_{\ell,f}(j) = 0\}$ then we have from the induction assumption $f(i) \in A_{k-1}(i,f)$ for $i \in W_{k-1}(f)$ and $\psi_{k,f} \ge 0$ on $W_{k-1}(f)$. Assume $\psi_{k,f}(i) = 0$, $f(i) \in A_{k-1}(i,f)$ so (4.13) holds for $a = f(i)$ $(k \ge 1)$:

$$(5.3)\qquad -\sum_{j\notin T_i} p_{ij}^{f(i)} (c_{k-1,f} + \psi_{k-1,f})(j) + \sum_{j\in T_i} p_{ij}^{f(i)} (c_{k,f} - c_{k-1,f})(j) +$$

$$+ \sum_{j\notin T_i} p_{ij}^{f(i)} (c_{k,f} + \psi_{k,f})(j) \le (c_{k,f} + \psi_{k,f})(i) = c_{k,f}(i) \ .$$

And from (3.13;f), (3,2) and (3.7) we have

$$(5.4)\qquad -\sum_{j\notin T_i} p_{ij}^{f(i)} c_{k-1,f}(j) + \sum_{j\in T_i} p_{ij}^{f(i)} (c_{k,f} - c_{k-1,f})(j) +$$

$$+ \sum_{j\notin T_i} p_{ij}^{f(i)} c_{k,f}(j) = c_{k,f}(i) \ .$$

If we subtract (5.4) from (5.3) we get

$$(5.5)\qquad -\sum_{j\notin T_i} p_{ij}^{f(i)} \psi_{k-1}(j) + \sum_{j\notin T_i} p_{ij}^{f(i)} \psi_k(j) \le 0 \ .$$

(For $k = 0$ (5.3) and (5.4) will look different but after the subtraction we get again (5.5).)
In the induction assumption the first term on the rhs of (5.5) disappears, so

$$\sum_{j\notin T_i} p_{ij}^{f(i)} \psi_{k,f}(j) \le 0 \ .$$

But $\psi_{k,f} \ge 0$ on W_{k-1} so also for all $j \in V(i,f)$. Hence $\psi_k(j) = 0$ for all $j \in V(i,f)$. As a result (5.3) holds with equality so $f(i) \in A_k(i,f)$. Finally the same reasoning as before gives us $\psi_{k+1,f}(i) \ge 0$. □

Now we return to the proof of the theorem.

Proof of theorem 5.1. Define $Z_k := \{i \in S \mid \Delta c_{-1,g,f}(i) = \ldots = \Delta c_{k,g,f}(i) = 0\}$, $k = -1,0,\ldots$. We will prove by induction

$\Delta c_{-1,g,f} \ge 0$ and if $i \in Z_{k-1}$ then $\Delta c_{k,g,f}(i) \ge 0$, $k = 0,1,\ldots,n$

$\psi_{k,f} = 0$ on Z_k and Z_k is closed under P_g, $k = -1,0,\ldots,n$.

From $g \in G_n(f)$ we have $\psi_{k,g,f} = \psi_{k,f}$, $k = -1,\ldots,n$. So from (3.14)-(3.16)

$$(5.6)\qquad P_{\tau,g}\Delta c_{-1,g,f} = \Delta c_{-1,g,f} - \psi_{-1,f}$$

$$(5.7)\qquad P_{\tau,g}\Delta c_{0,g,f} - Q_{\tau,g}P_{\tau,g}\Delta c_{-1,g,f} = \Delta c_{0,g,f} - \psi_{0,f}$$

$$(5.8)\qquad -R_{\tau,g}(\Delta c_{k-1,g,f} - \psi_{k-1,f}) + P_{\tau,g}(\Delta c_{k,g,f} - \Delta c_{k-1,g,f}) = \Delta c_{k,g,f} - \psi_{k,f} \ .$$

$k = -1$: In [6] we used (5.6) and (5.7) to prove $\Delta c_{-1,g,f} \ge 0$. Assume $\Delta c_{-1,g,f}(i) = 0$ then we have from (5.6), $\psi_{-1,f} \ge 0$ and $P_{\tau,g}\Delta c_{-1,g,f} \ge 0$ that $\psi_{-1,f}(i) = 0$ and

$$P_{\tau,g}\Delta c_{-1,g,f}(i) = \sum_{j\in T_i} p_{ij}^{g(i)} \Delta c_{-1,g,f}(i) + \sum_{j\notin T_i} p_{ij}^{g(i)} (\Delta c_{-1,g,f} + \psi_{-1,f})(j) = 0.$$

From $\psi_{-1,f}(i) = 0$ and lemma 5.2(ii) also $\sum_{j \notin T_i} p_{ij}^{g(i)} \psi_{-1,f}(j) = 0$ which gives us

$$\sum_{j \in S} p_{ij}^{g(i)} \Delta c_{-1,g,f}(j) = 0 \ .$$

So with $\Delta c_{-1,g,f} \geq 0$ we get

$$\Delta c_{-1,g,f}(j) = 0 \text{ for all } j \in W(i,g) := \{\ell \mid p_{i\ell}^{g(i)} > 0\} \ .$$

And the set Z_{-1} is closed under P_g.

$-1 < k < n$: From the induction assumption and (5.7) and (5.8) we get on Z_{k-1} the following two equations

$$(5.9) \qquad P_{\tau,g} \Delta c_{k,g,f} = \Delta c_{k,g,f} - \psi_{k,f}$$

$$(5.10) \qquad -R_{\tau,g}(\Delta c_{k,g,f} - \psi_{k,f}) + P_{\tau,g}(\Delta c_{k+1,g,f} - \Delta c_{k,g,f}) = \Delta c_{k+1,g,f} + \psi_{k+1,f} \ .$$

With (5.9) and $I + R_{\tau,g} = Q_{\tau,g}$ we may rewrite (5.10) as

$$(5.11) \qquad P_{\tau,g} \Delta c_{k+1,g,f} - Q_{\tau,g} P_{\tau,g} \Delta c_{k,g,f} = \Delta c_{k+1,g,f} - \psi_{k+1,f} \ .$$

Now (5.9) and (5.11) have the same form as (5.6)-(5.7) so in exactly the same way as there (cf. [6]) we obtain that $\Delta c_{k,g,f} \geq 0$ on Z_{k-1}, $\psi_{k,f} = 0$ on Z_k and Z_k closed under P_g.

$k = n$. In this case we only have one equation on Z_{k-1}, viz.

$$(5.12) \qquad P_{\tau,g} \Delta c_{k,g,f} = \Delta c_{k,g,f} - \psi_{k,f} \ .$$

But we also have $g(i) = f(i)$ if $f(i) \in G_k(i,f)$.

The situation is identical to the case $n = k = 0$ in [6]. If we multiply (5.12) with $P^*_{\tau,g} = \lim_{N \to \infty} \frac{1}{N+1} \sum_{n=0}^{N} P^n_{\tau,g}$ on Z_{k-1} (Z_{k-1} is also closed under $P^*_{\tau,g}$) then we get $P^*_{\tau,g} \psi_{k,f} = 0$, hence $\psi_{k,f} = 0$ on the set $B_{\tau,k-1}$ of recurrent states in Z_{k-1} under $P_{\tau,g}$ and therefore $f(i) \in A_k(i,f)$, $i \in B_{\tau,k-1}$ (lemma 5.2). We need however $\psi_{k,f} = 0$ on B_{k-1}, the set of recurrent states of Z_{k-1} under P_g. Let $i \in B_{k-1}$ with $\psi_{k,f}(i) = 0$ then $\psi_{k,f}(j)$ for all $j \in V(i,f)$. But also $\psi_{k,f}(j) = 0$ if $j \in T_i$ and $p_{ij}^{f(i)}$ since then $j \in B_{\tau,k-1}$. So if $\psi_{k,f}(i) = 0$ then $\psi_{k,f}(j) = 0$ and $f(j) \in A_k(j,f)$ for all $j \in W(i,g)$. Therefore the set of states with $\psi_{k,f}(i) = 0$ is closed under P_g. It contains $B_{\tau,k-1}$ hence also B_{k-1}.

As a result $f = g$ on B_{k-1}, and since B_{k-1} is closed under P_g also $c_{k,g} = c_{k,g}$ or $\Delta c_{k,g,f} = 0$ on B_{k-1}.

The equivalent of lemma 5.4 in [6] gives

$$\min_{i \in Z_{k-1}} \Delta c_{k,g,f}(i) = 0 \ ,$$

hence $\Delta c_{k,g,f}(i) \geq 0$ on Z_{k-1}.
Finally if $\Delta c_{-1,g,f} = \ldots = \Delta c_{n,g,f} = 0$ then (by induction) $\psi_{-1,f} = \ldots = \psi_{n,f} = 0$, and by lemma 5.2(iii) $f(i) \in A_n(i,f)$ for all $i \in S$, hence $g = f$. □

6. The convergence to an $(n-1)$-discount optimal policy. In the previous section we showed that the policy improvement step gives a policy $g \in G_n(f)$ with $g \overset{n}{\geq} f$. And $g \overset{n}{=} f$ only if $g = f$. As there are only finitely many policies, the policy iteration algorithm terminates with a policy, f say, with $f \in G_n(f)$. Now we prove

Theorem 6.1. If $f \in G_n(f)$, then f is $(n-1)$-discount optimal, i.e. $f \overset{n-1}{\geq} g$ for all g.

The way we prove this is similar to the approach of section 5. First we need the following equivalent of lemma 5.2.

Lemma 6.2. Let g be an arbitrary policy and $f \in G_n(f)$ then

i) $\psi_{-1,g,f} \leq 0$
and if $\psi_{-1,g,f}(i) = \ldots = \psi_{k,g,f}(i) = 0$ then
ii) $\psi_{-1,g,f}(j) = \ldots = \psi_{k,g,f}(j) = 0$ for all $j \in V(i,g)$, $k \leq n$
iii) $g(i) \in A_k(i,f)$, $k \leq n$
iv) $\psi_{k+1,g,f}(i) \leq 0$, $k \leq n-1$.

Proof. The proof is similar to the proof of lemma 5.2. i) $\psi_{-1,g,f} \leq \psi_{-1,f} = 0$.
ii)-iv) we show again by induction.
$k = -1$: If we subtract (4.2) with $a = g(i)$ from (3.8) and substitute $\psi_{-1,f} = 0$ we get

$$\sum_{j \notin T_i} p_{ij}^{g(i)} \psi_{-1,g,f}(j) \geq \psi_{-1,g,f}(i) .$$

So if $\psi_{-1,g,f}(i) = 0$ then from $\psi_{-1,g,f} \leq 0$ also $\psi_{-1,g,f}(j) = 0$ for $j \in V(i,g)$.
As a result $g(i)$ satisfies (4.2) (with $\psi_{-1,f} = 0$) with equality, so $g(i) \in A_{-1}(i,f)$. And let $\hat{g}$ be an arbitrary policy in $G_{-1}(f)$ with $\hat{g}(i) = g(i)$ if $g(i) \in A_{-1}(i,f)$ then for i with $\psi_{-1,g,f}(i) = 0$

$$\psi_{0,g,f}(i) = \psi_{0,\hat{g},f}(i) \leq \max_{h \in G_{-1}(f)} \psi_{0,h,f}(i) = \psi_{0,f}(i) = 0$$

which completes the proof for $k = -1$.
The case $k \geq 0$ is completely analogous to the case $k \geq 0$ in lemma 5.2. We omit it here. □

Proof of theorem 6.1. The reasoning is almost identical to the one in theorem 5.1. We only give a brief outline. First we have from (3.14), (3.15) and $\psi_{-1,g,f} \leq 0$ that $\Delta c_{-1,g,f} \leq 0$. If $\Delta c_{-1,g,f}(i) = 0$, then - from (3.14) - also $\psi_{-1,g,f}(i) = 0$, and - by lemma 6.2 - $\psi_{-1,g,f}(j) = 0$ for all $j \in V(i,g)$. And again the set $\{i \in S \mid \Delta c_{-1,g,f}(i) = 0\}$ is closed under P_g. For $0 \leq k \leq n-1$ the reasoning is similar to the reasoning in theorem 5.1. We miss however the condition $g(i) = f(i)$ if $f(i) \in A_n(i,f)$, therefore we can only prove $(n-1)$ - and not n-discount - optimality. □

7. ∞-discount optimality. In this section we prove the result which corresponds to theorem 4 in Miller and Veinott [3]. I.e., we show that a policy f obtained from the modified policy iteration algorithm with $n = N$, the number of states in S, $[f \in G_N(f)]$, is not only $(N-1)$-discount optimal but even ∞-discount optimal.

In order to do this we first copy the result of Miller & Veinott for the case $\tau \equiv 1$, i.e. the standard successive approximation step in (2.1). (We have to do this because we expand in $(1-\beta)$ and Miller and Veinott [3] used the expansion in $\rho(\beta = (1+\rho)^{-1})$.) Then we have

$$(7.1) \qquad r_g + \beta P_g v_{\beta,f} - v_{\beta,f} = r_g + P_g v_{\beta,f} - (1-\beta) P_g v_{\beta,f} - v_{\beta,f}$$

$$= \sum_{n=-1}^{\infty} \gamma_{n,g,f} (1-\beta)^n$$

with

$$\gamma_{-1,g,f} = P_g c_{-1,f} - c_{-1,f}$$

$$\gamma_{0,g,f} = r_g + P_g (c_{0,f} - c_{-1,f}) - c_{0,f}$$

$$(7.2) \qquad \gamma_{n,g,f} = P_g (c_{n,f} - c_{n-1,f}) - c_{n,f}, \quad n = 1,2,\ldots .$$

Of course $\gamma_{n,f,f} = 0$ so

$$(7.3) \qquad c_{n,f} = P_f (c_{n,f} - c_{n-1,f}), \quad n = 1,2,\ldots .$$

From (7.2) and (7.3) we get for $n \geq 1$

$$(7.4) \qquad \gamma_{n,g,f} = (P_g - P_f)(c_{n,f} - c_{n-1,f}) .$$

Further we have from (1.3a)

$$(7.5) \qquad c_{n,f} - c_{n-1,f} = (-1)^{n-1} [S(I-S)^{-1}]^{n-1} [-S(I-S)^{-1} - I][I-S]^{-1} r_f$$

$$= -(-1)^{n-1} [S(I-S)^{-1}]^{n-1} (I-S)^{-2} r_f$$

with $S = P_f - P_f^*$.

And we get the following variant of theorem 4 in Miller and Veinott [3].

Lemma 7.1. If $\gamma_{n,g,f} = 0$, $n = -1,\ldots,S$ then $\gamma_{n,g,f} = 0$ for all n.

Proof. Substitute (7.5) in (7.4) and use lemma 4 in Miller and Veinott [3] with $x = (I - P_f + P_f^*)^{-2} r_f$, $M = -(P_f - P_f^*)(I - P_f + P_f^*)^{-1}$ and L the null space of $P_g - P_f$. □

Note that the result of this lemma can be obtained directly from theorem 4 in Miller and Veinott using $\lim_{\beta \uparrow 1} (1 - \beta)/\rho = 1$.

For an arbitrary stopping time τ we have with lemma 1.1

$$
\begin{aligned}
& r_{\beta,\tau,g} + P_{\beta,\tau,g} v_{\beta,f} - v_{\beta,f} = (I - \beta \bar{P}_g)^{-1} (r_g + \beta \tilde{P}_g v_{\beta,f}) - v_{\beta,f} \\
& = (I - \beta \bar{P}_g)^{-1} (r_g + \beta \tilde{P}_g v_{\beta,f} + \beta \bar{P}_g v_{\beta,f} - v_{\beta,f}) \\
& = (I - \beta \bar{P}_g)^{-1} (r_g + \beta P_g v_{\beta,f} - v_{\beta,f}) = (I - \beta \bar{P}_g)^{-1} \sum_{n=-1}^{\infty} \gamma_{n,g,f} (1 - \beta)^n \\
& = \sum_{k=-1}^{\infty} (1 - \beta)^k \sum_{\ell=-1}^{k} (-1)^{k-\ell} \{\bar{P}_g (I - \bar{P}_g)^{-1}\}^{k-\ell} (I - \bar{P}_g)^{-1} \gamma_{\ell,g,f} .
\end{aligned}
\tag{7.6}
$$

From which we obtain

Lemma 7.2. $\gamma_{-1,g,f} = \ldots = \gamma_{k,g,f} = 0$ if and only if $\psi_{-1,g,f} = \ldots = \psi_{k,g,f} = 0$, $k = -1,0,\ldots$.

Proof. The if part follows by induction. The only if part is immediate from (7.6). □

Now we are able to prove

Theorem 7. If $f \in G_S(f)$ then $f \overset{n}{\geq} g$ for all n and all g.

Proof. Suppose we have a policy g with $g \overset{n}{\geq} f$ for some $n > S - 1$, then clearly $g \overset{S-1}{=} f$ and $g \overset{S}{\geq} f$. From lemma 6.2 we have $\Delta c_{k,g,f} = 0$, $k = -1,0,\ldots,S-1$, $\psi_{k,g,f} = 0$, $k = -1,0,\ldots,S-1$ and $\psi_{S,g,f} \leq 0$, otherwise g would be an improvement of f. Further (3.16) with $k = S$ reduces to

$$
P_{\tau,g} \Delta c_{S,g,f} = \Delta c_{S,g,f} - \psi_{S,g,f} . \tag{7.7}
$$

And if we multiply this with $P^*_{\tau,g}$ then we get

$$\psi_{S,g,f} = 0 \text{ on } Rec(\tau,g)$$

where $Rec(\tau,g)$ is the set of recurrent states under $P_{\tau,g}$. And with lemma 6.2 ii) also $\psi_{S,g,f}(j)$ if $j \in V(i,g)$ for some $i \in Rec(\tau,g)$. So even

$$\psi_{S,g,f} = 0 \text{ on } Rec(g)$$

with $Rec(g)$ the set of recurrent states under P_g (cf. the proof of theorem 5.1).

So on $Rec(g)$ $\psi_{-1,g,f} = \ldots = \psi_{S,g,f} = 0$. Hence by lemma 7.2 also $\gamma_{-1,g,f} = \ldots = \gamma_{S,g,f} = 0$ and by lemma 7.1 $\gamma_{n,g,f} = 0$ for all n. Thus with (7.1)

$$r_g + \beta P_g v_{\beta,f} = v_{\beta,f} .$$

So $v_{\beta,g} = v_{\beta,f}$ and especially $\Delta c_{S,g,f} = 0$ on $Rec(g)$. From (7.7) we have with $\psi_{S,g,f} \leq 0$

$$P_{\tau,g} \Delta c_{S,g,f} \geq \Delta c_{S,g,f} .$$

So $V := \{i \mid \Delta c_{S,g,f}(i) = \max_j \Delta c_{S,g,f}(j)\}$ is closed under $P_{\tau,g}$. Hence on V we have $\Delta c_{S,g,f} = 0$ and therefore $\Delta c_{S,g,f} \leq 0$ on S. But we assumed $g \overset{n}{\geq} f$ for some $n > S-1$, hence with $\Delta c_{-1,g,f} = \ldots = \Delta c_{S-1,g,f} = 0$ we must have $\Delta c_{S,g,f} = 0$ on S. But then $\psi_{S,g,f} = 0$ on S and by lemma 7.2 also $\gamma_{S,g,f} = 0$. So by lemma 7.1 we have $\gamma_{n,g,f} = 0$ for all n and also $\Delta c_{n,g,f} = 0$ for all n, or $g \overset{n}{=} f$ for all n. Summarizing, we have shown that if for some $n \geq S$ we have $g \overset{n}{\geq} f$ then $g \overset{n}{=} f$. Hence $f \overset{n}{\geq} g$ for all n and all g. □

8. References

[1] Blackwell, D., Discrete dynamic programming, Ann. Math. Statist. 33 (1962), 719-726.

[2] Howard, R.A., Dynamic Programming and Markov Processes, MIT Press, Cambridge (Mass), 1960.

[3] Miller, B.L. and Veinott, A.F., Discrete dynamic programming with a small interest rate, Ann. Math. Statist. 40 (1969), 366-370.

[4] Veinott, A.F., On finding optimal policies in discrete dynamic programming with no discounting, Ann. Math. Statist. 37 (1966), 1284-1294.

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

[6] Wal, J. van der, A stopping time-based policy iteration algorithm for average reward Markov decision processes, Memorandum COSOR 78-11, University of Technology Eindhoven, 1978.

[7] Wessels, J., Stopping times and Markov programming, Transactions of the 7th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Prague, 1977, vol. A, 575-585.