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

Citation for published version (APA):

Document status and date:
Published: 01/01/1978

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.
Memorandum COSOR 78-24
A stopping time-based policy iteration algorithm for Markov decision processes with discount factor 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 discount factor tending to 1

by

J. van der Wal

Abstract. This paper considers the Markov decision process with finite state
and action spaces, when the discount factor tends to 1. Miller and Veinott
have shown the existence of n-discount optimal policies and Veinott has gi­
gen 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 al­
gorithms 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 discount­
factor \( \beta \) 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) optima­
lity 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 op­
timal 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 \( S = \{1,2,\ldots,N\} \)
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 \( p^a_{i,j} \). A po­
licy 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
\( \{(X_n,A_n), n = 0,1,\ldots\} \) 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 tran­
sition memoryless stopping time \( \tau \) as a map from \( S^\infty \) into \( \mathbb{N} = \{1,2,\ldots,\infty\} \) such
that for all \( i \) and \( f \), \( \mathbb{P}_{i,f}(\tau < \infty) = 1 \) and that \( \tau \) can be completely characte­
rized by a set \( T \subset S^2 \) such that (cf. [7,6])
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 \( \tau \) and \( T \) are fixed.

We want to introduce a few more notations. Let \( f \) be a policy then define the vectors \( r_f \) and \( \beta_\tau r_f \), and the matrices \( P_f \), \( P^*_f \) and \( P^\beta,\tau,f \) by

\[
r_f(i) = r(i, f(i))
\]

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

\[
P_f(i, j) = P^f_{ij}
\]

\[
P^*_f = \lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} P^n_f
\]

\[
P^\beta,\tau,f(i, j) = \sum_{n=1}^{\infty} \beta^n P_{i,f}(X = j, \tau = n).
\]

Further we define the matrices \( \tilde{P}_f \) and \( \tilde{P}^\beta_f \) (suppressing the dependence on \( \tau \)) by

\[
\tilde{P}_f(i, j) = \begin{cases} f(i) & \text{if } (i, j) \not\in T \\ P_{ij} & \text{if } (i, j) \in T \\ 0 & \text{if } (i, j) \not\in T \\ \tilde{P}_f(i, j) = \begin{cases} f(i) & \text{if } (i, j) \in T \\ P_{ij} & \text{if } (i, j) \in T 
\end{cases}
\]

Then we have

Lemma 1.1.

i) \( P_f = \tilde{P}_f + \tilde{P}^\beta_f \)

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

iii) \( r^\beta,\tau,f = r_f + \beta \tilde{P}_f r_f + \beta^2 \tilde{P}^\beta_f r_f + \ldots = (I - \beta \tilde{P}_f)^{-1} r_f \).

From the finiteness of \( \tau \) it follows that \( I - \tilde{P}_f \) 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 (\( n = -1, 0, \ldots \)) if
\[
\limsup_{\beta \uparrow 1} (1 - \beta)^n (v_{\beta, f} - v_{\beta, g}) \geq 0 \quad \text{for all } g.
\]
And policy \( f \) is called \( \infty \)-discount optimal if \( f \) is \( n \)-discount optimal for all \( n = -1, 0, 1, \ldots \).

For \( v_{\beta, f} \), we also have the Laurent series expansion in \( (1 - \beta) \) for \( \beta + 1 \)
\[
v_{\beta, f} = \sum_{n=-1}^{\infty} (1 - \beta)^n c_{n,f}.
\]
(Miller and Veinott [3] used the expansion in \( \rho \), 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
\[
v_{\beta, f} = [1 + \beta P_f + \beta^2 P_f^2 + \ldots] r_f
\]
\[
= [1 + \beta (P^*_f - P_f) + \beta^2 (P^*_f - P_f) + \ldots] r_f + (1 - \beta)^{-1} P^*_f r_f - P_f r_f.
\]

With \( P^n_f - P^*_f = (P_f - P^*_f)^n \), \( n = 1, 2, \ldots \) (from \( P_f P_f^* = P^* f = P_f^* \), cf. [1]) we get
\[
(1.2) \quad v_{\beta, f} = (1 - \beta)^{-1} P^*_f r_f - P_f r_f + [1 - \beta (P^*_f - P_f)]^{-1} r_f.
\]
If \( I - S \) is nonsingular and \( \beta \) is sufficiently close to 1 then we have the expansion
\[
(1.3) \quad (I - \beta S)^{-1} = (I - S + (1 - \beta) S)^{-1} = [I + (1 - \beta) S (I - S)^{-1}]^{-1} (I - S)^{-1}
\]
\[
= \sum_{k=0}^{\infty} (-1)^k (1 - \beta)^k [S (I - S)^{-1}]^k (I - S)^{-1}.
\]

Since \( I - P_f + P^*_f \) is nonsingular (lemma 1d in [1]) we may substitute (1.3) in (1.2) to obtain
\[
(1.4) \quad \begin{cases}
    c_{-1,f} = P^*_f r_f \\
    c_{0,f} = [(I - S)^{-1} - P^*_f] r_f \\
    c_{k,f} = (-1)^k [S (I - S)^{-1}]^k (I - S)^{-1} r_f
\end{cases}
\]
with \( S = P_f - P^*_f \).
For any two policies \( f \) and \( g \) we define

\[
\Delta c_{n,f,g} := c_{n,f} - c_{n,g}, \quad n = -1,0,\ldots.
\]

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

We see that a policy \( f \) is \( n \)-discount optimal \([\omega\text{-discount optimal}]\) if and only if \( f \succeq g \) \([f \preceq g]\) for all \( g \).

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

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 \( g \)

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

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 \( S = P_g \) to obtain

\[
(2.2) \quad r_{\beta,\tau,g} + P_{\beta,\tau,g} v_{\beta,f} = (I - P_g)^{-1} r_g + (P_g - (1 - \beta) P_g) v_{\beta,f}.
\]

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

\[
(2.3) \quad d_{-1,g,f} = (I - P_g)^{-1} P_{-1,f} c_{-1,f},
\]

\[
(2.3) \quad d_{n,g,f} = (1 - \beta)^n r_g + (P_g - (1 - \beta) P_g) (I - P_g)^{-1} r_g \quad \text{for} \quad n = 0,1,\ldots.
\]
With the notations $r_{\tau,g}$, $P_{\tau,g}$ and $T_{\tau,g}$, simplifies to

\[
Q_{\tau,g} = (I - \tilde{P}_{g})^{-1}
\]

\[
R_{\tau,g} = \tilde{P}_{g} (I - \tilde{P}_{g})^{-1}
\]

(2.3) simplifies to

\[
d_{-1,g,f} = P_{\tau,g} c_{-1,f}
\]

\[
d_{0,g,f} = r_{\tau,g} + P_{\tau,g} c_{0,f} - Q_{\tau,g} P_{\tau,g} c_{-1,f}
\]

\[
d_{n,g,f} = (-1)^{n} R_{\tau,g}^n d_{n-1,g,f} + \sum_{k=0}^{n+1} R_{\tau,g}^k P_{\tau,g} c_{n-k,f} + \sum_{k=0}^{n} (-1)^{k} R_{\tau,g}^k P_{\tau,g} c_{n-k-1,f}
\]

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

\[
d_{n,g,f} = (-R_{\tau,g})^{n} d_{n-1,g,f} + P_{\tau,g} (c_{n,f} - c_{n-1,f}), n \geq 1.
\]

If we maximize (2.1) for $\beta$ sufficiently close to 1 then we maximize "lexicographically" 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 $d_{-1,g,f}, ..., d_{n,g,f}$, 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

\[
\psi_{k,g,f} := d_{k,g,f} - c_{k,f}, k = -1, 0, ...
\]

From the definitions of $r_{\tau,g}$, $P_{\tau,g}$, $Q_{\tau,g}$ and $R_{\tau,g}$ we have
If we substitute (3.1) and (3.2) in (2.4)-(2.6) we get

\[ \tilde{\rho}_g c_{-1,f} + \tilde{\rho}_g (c_{-1,f} + \psi_{-1,g,f}) = c_{-1,f} + \psi_{-1,g,f} \]

\[ r_g + \tilde{\rho}_g c_{0,f} - (c_{-1,f} + \psi_{-1,g,f}) = c_{0,f} + \psi_{0,g,f} \]

\[ -\tilde{\rho}_g (c_{k-1,f} + \psi_{k-1,g,f}) + \tilde{\rho}_g (c_{k,f} - c_{k-1,f}) + \tilde{\rho}_g (c_{k,f} + \psi_{k,g,f}) = c_{k,f} + \psi_{k,g,f}. \]

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

\[ T_1 := \{ j \in S \mid (i,j) \in T \}. \]

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

\[ (\tilde{\rho}_g v)(i) = \sum_{j \in T_1} p_{ij} g(i) v(j) \quad \text{and} \quad (\tilde{\rho}_g v)(i) = \sum_{j \notin T_1} p_{ij} g(i) v(j). \]

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

\[ \sum_{j \in T_1} p_{ij} g(i) c_{-1,f}(j) + \sum_{j \notin T_1} p_{ij} g(i) (c_{-1,f} + \psi_{-1,g,f})(j) = (c_{-1,f} + \psi_{-1,g,f})(i) \]

\[ r(i,g(i)) + \sum_{j \in T_1} p_{ij} g(i) c_{0,f}(j) - (c_{-1,f} + \psi_{-1,g,f})(i) + \]

\[ + \sum_{j \notin T_1} p_{ij} g(i) (c_{0,f} + \psi_{0,g,f})(j) = (c_{0,f} + \psi_{0,g,f})(i) \]

\[ - \sum_{j \notin T_1} p_{ij} g(i) (c_{k-1,f} + \psi_{k-1,g,f})(j) + \sum_{j \in T_1} p_{ij} g(i) (c_{k,f} - c_{k-1,f})(j) + \]

\[ + \sum_{j \notin T_1} p_{ij} g(i) (c_{k,f} + \psi_{k,g,f})(j) = (c_{k,f} + \psi_{k,g,f})(i). \]

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 \( \tau \).

This we see as follows. Clearly (3.8)-(3.10) and (3.3)-(3.5) are equivalent. And as \( \tau \) is finite \( 1 - \tilde{\rho}_g \) is nonsingular. Multiplying (3.3)-(3.5) by \((1 - \tilde{\rho}_g)^{-1}\) gives us (2.4)-(2.6).
In the second part of this section we derive some relations between the
\( \Delta k, g, f \) and the \( \psi_k, g, f \). Clearly we have from \( r_\beta, \tau, f + P_\beta, \tau, f \nu_\beta, f = \nu_\beta, f \) (cf.
lemma 1.1 in Wessels [7]) that \( d_k, f, f = c_k, f \) so

\[
(3.11;f) \quad P_{\tau, f} c_{-1, f} = c_{-1, f}
\]

\[
(3.12;f) \quad r_{\tau, f} + P_{\tau, f} c_0, f - Q_{\tau, f} P_{\tau, f} c_{-1, f} = c_0, f
\]

\[
(3.13;f) \quad (-R_{\tau, f}) c_{k-1, f} + P_{\tau, f} (c_k, f - c_{k-1, f}) = c_k, f.
\]

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

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

\[
(3.15) \quad P_{\tau, g} \Delta c_0, g, f - Q_{\tau, g} \Delta c_{-1, g, f} = \Delta c_0, g, f - \psi_0, g, f
\]

\[
(3.16) \quad (-R_{\tau, g}) \Delta c_{k-1, g, f} - \psi_{k-1, g, f} + P_{\tau, g} (\Delta c_k, g, f - c_{k-1, g, f}) =
\]

\[= \Delta c_{k, g, f} - \psi_{k, g, f}, \quad k \geq 1.
\]

4. The modified policy improvement step. In section 2 we have seen that if \( \beta + 1 \)
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

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

Then we have for all a

\[
(4.2) \quad \sum_{j \in T_i} a_{ij} c_{-1, f}(j) + \sum_{j \notin T_i} a_{ij} (c_{-1, f} + \psi_{-1, f})(j) \leq (c_{-1, f} + \psi_{-1, f})(i).
\]

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 equal­
ity for \( g(i) \). Now consider the policy \( h \) with \( h(i) = a \) and \( h(j) = g(j), j \neq i. \)
Then from (4.2)

\[
(4.3) \quad \bar{P}_{h} c_{-1, f} + \bar{P}_{h} (c_{-1, f} + \psi_{-1, f}) \geq (c_{-1, f} + \psi_{-1, f}),
\]

so

\[
(4.4) \quad (I - \bar{P}_{h})^{-1} h_{-1, f} = P_{\tau, h} c_{-1, f} \geq c_{-1, f} + \psi_{-1, f},
\]

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

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

And

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

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

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

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

\[(4.8) \quad r(i,a) + \sum_{j \in T_{1}} p_{ij}^{a} c_{0,f}(j) - (c_{-1,f} + \psi_{-1},f)(i) + \sum_{j \notin T_{1}} p_{ij}^{a} (c_{0,f} + \psi_{0},f)(j) \leq (c_{0,f} + \psi_{0},f)(i). \]

If we define further

\[(4.9) \quad A_{0}(i,f) := \text{the set of } a \in A_{-1}(i,f) \text{ for which } (4.8) \text{ holds with equality} \]

\[(4.10) \quad G_{0}(f) := \{g \mid g(i) \in A_{0}(i,f) \text{ for all } i \in S\}. \]

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 \( (g(i) = f(i) \text{ if } f(i) \in A_{0}(i,f)) \), converges and produces an average optimal policy. I.e. a policy \( h \) with \( h \perp g \) for all \( g \).

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

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

\[(4.12) \quad A_{k}(i,f) := \text{the set of } a \in A_{k-1}(i,f) \text{ for which } (4.13) \text{ below holds with equality, } k = 1,2,\ldots \]

\[(4.13) \quad \sum_{j \notin T_{1}} p_{ij}^{a} (c_{k-1,f} + \psi_{k-1},f)(j) + \sum_{j \in T_{1}} p_{ij}^{a} (c_{k,f} - c_{k-1,f})(j) + \sum_{j \notin T_{1}} p_{ij}^{a} (c_{k,f} + \psi_{k,f})(j) \leq (c_{k,f} + \psi_{k,f})(i) \]

\[(4.14) \quad G_{k}(f) := \{g \mid g(i) \in A_{k}(i,f) \text{ for all } i \in S\}, \quad k = 1,2,\ldots \]
In the same way as before one may show that (4.13) holds for all \( a \in \mathcal{A}_{k-1}(i,f) \) and that \( g \) maximizes (4.11) within \( \mathcal{C}_{k-1}(f) \) if and only if \( g \in \mathcal{C}_k(f) \).

Now we can propose the following modified policy iteration algorithm.

\[
\begin{align*}
\text{Value determination step} \\
\text{Let } f \text{ be the current policy. Determine } c_{-1,f}^{n}, \ldots, c_{n,f}^{n}.
\end{align*}
\]

\[
\begin{align*}
\text{Policy improvement step} \\
\text{Determine a policy } g \in \mathcal{G}_n(f) \text{ with } g(i) = f(i) \text{ whenever } f(i) \in \mathcal{A}_n(i,f).
\end{align*}
\]

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 \( g \in \mathcal{G}_n(f) \) with \( g(j) = f(j) \) whenever \( f(j) \in \mathcal{A}_n(j,f) \), \( j \in S \) then

i) \( g \succeq f \).

ii) \( g \succeq f \) only if \( g = f \) (\( g \succeq f \Leftrightarrow g \succeq f \) and \( f \succeq g \)).

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 \( j \in V(i,f) := \{ k \notin T_i | p_{ik}^{f(i)} > 0 \} \).

iii) \( f(i) \in \mathcal{A}_k(i,f) \).

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

**Proof.**

i) From (3.11;f) we have

\[
c_{-1,f} + \psi_{-1,f} = \max_{g} p_{f,g} c_{-1,f} \geq p_{f,c_{-1,f}} c_{-1,f} = c_{-1,f},
\]

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) \[ \sum_{j \in T_i} p_{ij} f(i) c_{-1,f}(i) + \sum_{j \not\in T_i} p_{ij} f(i) (c_{-1,f} + \psi_{-1,f}(j)) \leq (c_{-1,f} + \psi_{-1,f})(i) = c_{-1,f}(i). \]

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

(5.2) \[ \sum_{j \in T_i} p_{ij} f(i) c_{-1,f}(j) + \sum_{j \not\in T_i} p_{ij} f(i) c_{-1,f}(j) = c_{-1,f}(i). \]

Subtracting (5.2) from (5.1) we get

\[ \sum_{j \not\in T_i} p_{ij} f(i) \psi_{-1,f}(j) \leq 0 \]

which together with \( \psi_{-1,f} \geq 0 \) yields

\[ \psi_{-1,f}(j) = 0 \] 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 \( \tau, \) therefore it uses only actions from \( f \). So

\[ \psi_0(i) = \max_{g \in G_{-1}(f)} \psi_0,g,f(i) \geq \psi_0,f,f(i) = \psi_0,f,f(i) = 0 \]

which completes the proof for \( n = -1 \).

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

(5.3) \[ - \sum_{j \not\in T_i} p_{ij} f(i) (c_{k-1,f} + \psi_{k-1,f}(j)) + \sum_{j \not\in T_i} p_{ij} f(i) (c_{k,f} - c_{k-1,f})(j) + \sum_{j \not\in T_i} p_{ij} f(i) c_{k,f}(j) \leq (c_{k,f} + \psi_{k,f})(i) = c_{k,f}(i). \]

And from (3.13; f), (3.2) and (3.7) we have
If we subtract (5.4) from (5.3) we get

\[ (5.5) \quad - \sum_{j \notin T} p_{ij} f(i) \psi_{k-1}(j) + \sum_{j \notin T} p_{ij} f(i) \psi_k(j) \leq 0. \]

(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

\[ \sum_{j \notin T} p_{ij} f(i) \psi_k(j) \leq 0. \]

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

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

\[ \Delta c_{-1,g,f} \geq 0 \quad \text{and if } i \in Z_{k-1} \text{ then } \Delta c_{k,g,f}(i) \geq 0, \quad k = 0,1,\ldots,n \]

\( \psi_{k,f} = 0 \) on \( Z_k \) and \( Z_k \) is closed under \( P_{g,k} \), \( k = -1,0,\ldots,n \).

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

\[ (5.6) \quad P_{\tau,g} \Delta c_{-1,g,f} = \Delta c_{-1,g,f} - \psi_{-1,f} \]

\[ (5.7) \quad P_{\tau,g} \Delta c_{0,g,f} = \Delta c_{0,g,f} - \psi_{0,f} \]

\[ (5.8) \quad -P_{\tau,g} (\Delta c_{-1,g,f} - \psi_{-1,f}) + P_{\tau,g} (\Delta c_{k,g,f} - \psi_{k,f}) = \Delta c_{k,g,f} - \psi_{k,f} \]

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

\[ P_{\tau,g} \Delta c_{-1,g,f}(i) = \sum_{j \notin T} p_{ij} \Delta c_{-1,g,f}(i) + \sum_{j \notin T} p_{ij} (\Delta c_{-1,g,f}(i) + \psi_{-1,f}(j)) = 0. \]
From \( \psi_1, f(i) = 0 \) and lemma 5.2(ii) also \( \sum_{j \in T_1} P_{ij} g(i) \psi_1, f(j) = 0 \) which gives us
\[
\sum_{j \in S} g(i) \Delta c_{-1, g, f}(j) = 0.
\]
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 \( Z_{k-1} \) is closed under \( P_g \).
-1 < k < n: From the induction assumption and (5.7) and (5.8) we get on \( Z_k \) the following two equations
\[
\begin{align*}
(P) & P_{\tau, g} \Delta c_{k, g, f} = \Delta c_{k, g, f} - \psi_{k, f} \\
\text{(Q)} & -R_{\tau, g} (\Delta c_{k, g, f} - \psi_{k, f}) + P_{\tau, g} (\Delta c_{k+1, g, f} - \Delta c_{k, g, f}) = \Delta c_{k+1, g, f} + \psi_{k+1, f}.
\end{align*}
\]
With (5.9) and \( I + R = 0 \), we may rewrite (5.10) as
\[
(P) P_{\tau, g} \Delta c_{k+1, g, f} = \Delta c_{k+1, g, f} - \psi_{k, f}.
\]
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.
\[
P_{\tau, g} \Delta c_{k, g, f} = \Delta c_{k, g, f} - \psi_{k, f}.
\]
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
\[
(P) P_{\tau, g} = \lim_{N \to \infty} \frac{1}{N+1} \sum_{n=0}^{N} P_{\tau, g} \text{ on } Z_{k-1} \text{ (}Z_{k-1} \text{ is also closed under } P^*_{\tau, g})
\]
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 \( \psi_{k, f}(j) = 0 \) for all \( j \in V(i, f) \). But also
\[ \psi_{k, f}(j) = 0 \text{ if } j \in T_i \text{ and } P_{i j} g(i) \text{ since then } j \in B_{\tau, k-1}. \]
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 $\Delta c_{-1,g,f} = \cdots = \Delta c_{n,g,f} = 0$ then (by induction) $\psi_{-1,f} = \cdots = \psi_{n,f} = 0$, 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 $g \not\equiv f$. And $g \equiv f$ 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. $f \not\equiv g$ 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 $\psi_{-1,g,f}(i) = \cdots = \psi_{k,g,f}(i) = 0$ then

ii) $\psi_{-1,g,f}(j) = \cdots = \psi_{k,g,f}(j) = 0$ 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

$$\sum_{j \in T_1} p_{ij} \psi_{-1,g,f}(j) \geq \psi_{-1,g,f}(i).$$

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 $\bar{g}$ be an arbitrary policy in $G_{-1}(f)$ with $\bar{g}(i) = g(i)$

if $g(i) \in A_{-1}(i,f)$ then for $i$ with $\psi_{-1,g,f}(i) = 0$

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

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 \( \Delta_{c_{-1,g,f}}(i) = 0 \), then also 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 \( f(i) \in \Lambda_n(i,f) \), therefore we can only prove \((n-1)\) - and not \( n \)-discount - optimality.

7. \( \infty \)-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 \( \infty \)-discount optimal.

In order to do this we first copy the result of Miller & Veinott for the case \( \tau = 1 \), 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

\[
(7.1) \quad r_g + \beta P_g V_{\beta f} - V_{\beta f} = r_g + \frac{P g V_{\beta f} - (1-\beta) P g V_{\beta f} - V_{\beta f}}{n=1} \gamma_{n,g,f} (1-\beta)^n
\]

with

\[
\gamma_{-1,g,f} = P g c_{-1,f} - c_{-1,f}
\]

\[
\gamma_{0,g,f} = r_g + P g (c_{0,f} - c_{-1,f}) - c_{0,f}
\]

\[
\gamma_{n,g,f} = P g (c_{n,f} - c_{n-1,f}) - c_{n,f}, \quad n = 1, 2, \ldots .
\]

Of course \( c_{n,f,f} = 0 \) so

\[
(7.3) \quad c_{n,f,f} = P f (c_{n,f} - c_{n-1,f}), \quad n = 1, 2, \ldots .
\]

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

\[
(7.4) \quad \gamma_{n,g,f} = (P_g - P_f) (c_{n,f} - c_{n-1,f}).
\]

Further we have from (1.3a)

\[
(7.5) \quad c_{n,f,f} - c_{n-1,f} = (-1)^{n-1}[S(I - S)^{-1}]^{n-1} [S(I - S)^{-1} - I][I - S]^{-1} r_f = -(-1)^{n-1}[S(I - S)^{-1}]^{n-1} (I - S)^{-2} r_f
\]

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 \( x = (I - P_x + P_x)^{-1}x_f \), \( M = -(P_f - P_f^*)(I - P_x + P_x)^{-1} \) and \( L \) the null space of \( P - P_f \).

Note that the result of this lemma can be obtained directly from theorem 4 in Miller and Veinott using \( \lim \frac{1}{\beta+1} \).

For an arbitrary stopping time \( \tau \) we have with lemma 1.1

\[
\begin{align*}
\gamma_{\tau},g + \gamma_{\tau}^\beta,f = & \frac{\pi\beta}{\beta + 1} - \gamma_{\tau},g, f = (I - \beta P_g)^{-1}(\gamma_{\tau} + \beta P_g^\beta,f) - \gamma_{\tau},g, f \quad \text{(7.6)} \\
= & (I - \beta P_g)^{-1}(\gamma + \beta P_g^\beta v, f - \gamma_{\beta},f) = (I - \beta P_g)^{-1} \sum_{n=-1}^{\infty} \gamma_n,g,f (1 - \beta)^n \\
= & \sum_{k=1}^{\infty} (1 - \beta)^k \sum_{l=0}^{k} (-1)^{k-l} P_g (I - \beta P_g)^{-1} \gamma_{k-l} (I - \beta P_g)^{-1} \gamma_{k},g,f
\end{align*}
\]

From which we obtain

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

**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 \( f \parallel g \) for all \( n \) and all \( g \).

**Proof.** Suppose we have a policy \( g \) with \( g \parallel f \) for some \( n > S-1 \), then clearly \( g^S = f \) and \( g \parallel f \). From lemma 6.2 we have \( \Delta_k,g,f = 0 \), \( k = -1,0,\ldots,S-1 \), \( \psi_k,g,f = 0 \), \( k = -1,0,\ldots,S-1 \) and \( \psi_S,g,f \leq 0 \), otherwise \( g \) would be an improvement of \( f \). Further (3.16) with \( k = S \) reduces to

\[
(7.7) \quad P_{\tau},g \Delta S,g,f = \Delta S,g,f - \psi_S,g,f
\]
And if we multiply this with $P_{\tau,g}^*$ then we get

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

where $\text{Rec}(\tau,g)$ 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 $i \in \text{Rec}(\tau,g)$. So even

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

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

So on $\text{Rec}(g)$ $\psi_{-1,g,f} = \cdots = \psi_{S,g,f} = 0$. Hence by lemma 7.2 also

$$\gamma_{-1,g,f} = \cdots = \gamma_{S,g,f} = 0$$

and by lemma 7.1 $\gamma_{n,g,f} = 0$ for all $n$. Thus with

$$\gamma_{g} + \beta p \gamma_{g} = \gamma_{f}$$

we have $\gamma_{g} = \gamma_{f}$ and especially $\Delta c_{S,g,f} = 0$ on $\text{Rec}(g)$. From (7.7) we have with

$$\psi_{S,g,f} \leq 0$$

$$\text{Rec}(g) = \{i \mid \Delta c_{S,g,f}(i) = \max_j \Delta c_{S,g,f}(j)\}$$

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 $g \geq f$ for some $n > S - 1$, hence with $\Delta c_{-1,g,f} = \cdots = \Delta c_{S-1,g,f} = 0$ 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 $g \leq f$ for all $n$. Summarizing, we have shown that if for some $n \geq S$ we have $g \geq f$ then $g \leq f$. Hence $f \leq g$ for all $n$ and all $g$.

8. References


