On stationary strategies

van der Wal, J.

Published: 01/01/1981

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 author's 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

Citation for published version (APA):
On stationary strategies

by

Jan van der Wal
ON STATIONARY STRATEGIES

by

Jan van der Wal

Abstract

This paper deals with total reward Markov decision processes with countable state space and extends various results on (nearly-)optimal stationary strategies. Strauch proved that if the rewards are nonpositive and the action space is finite then an optimal stationary strategy exists. For the case of nonnegative rewards Ornstein proved the existence of a stationary strategy \( f \) which is uniformly nearly-optimal in the multiplicative sense: \( v(f) \geq (1 - \varepsilon)v^* \).

Van der Wal showed that if the action space is finite then for each initial state a stationary nearly-optimal strategy exists. These partial results are connected and extended in the following theorem. If in each state where the value is nonpositive a conserving action exists then there exists a stationary strategy \( f \) which is uniformly nearly optimal in the following sense:

\[
v(f) \geq v^* - cu^*,
\]

where \( u^* \) is the value of the problem if only the positive rewards are counted.

Further the following result is established: if an optimal strategy exists then also an optimal stationary strategy exists. This generalizes results of Strauch and Ornstein for the negative and positive dynamic programming cases respectively.
1. Introduction

Consider a Markov decision process (MDP) with countable state space \( S \) and arbitrary action space \( A \), with a \( \sigma \)-field \( \mathcal{A} \) containing all one-point sets. If in state \( i \in S \) action \( a \in A \) is taken two things happen: a (possibly negative) immediate reward \( r(i,a) \) is earned and the system moves to a new state \( j, j \in S \), with probability \( p(i,a,j) \), where \( \sum_j p(i,a,j) = 1 \). The functions \( r(i,a) \) and \( p(i,a,j) \) are assumed to be a measurable in \( a \).

One may distinguish three sets of strategies, namely the set \( \Pi \) of all (possibly randomized and history dependent) strategies satisfying the usual measurability conditions, the set \( M \) of all nonrandomized Markov strategies and the set \( F \) of all nonrandomized stationary strategies. So \( F \subset M \subset \Pi \). The elements of \( F \) will be called policies and are treated as functions on \( S \).

For each strategy \( \pi \in \Pi \) and each initial state \( i \in S \) we define in the usual way a probability measure \( \mathbb{P}_{i,\pi} \) on \((S \times A)^\omega\) and a stochastic process \([(X_n,A_n), n = 0,1,2,...]\) where \( X_n \) denotes the state of the system at time \( n \) and \( A_n \) the action chosen at time \( n \). Expectations with respect to \( \mathbb{P}_{i,\pi} \) will be denoted by \( E_{i,\pi} \).

Now we can define the total expected reward when the process starts in \( i \in S \) and strategy \( \pi \in \Pi \) is used

\[
(1.1) \quad v(i,\pi) := E_{i,\pi} \sum_{n=0}^{\infty} r(X_n,A_n),
\]

whenever the expectation at the right hand side is well-defined. To guarantee this the following assumption will be made.

**Condition 1:** For all \( i \in S \) and all \( \pi \in \Pi \)

\[
(1.2) \quad u(i,\pi) := E_{i,\pi} \sum_{n=0}^{\infty} r^+(X_n,A_n) < \infty,
\]

where

\[
r^+(i,a) := \max\{0, r(i,a)\}, \quad i \in S, \quad a \in A.
\]
This condition allows us to interchange expectation and summation in (1.1) and implies

\[ \lim_{n \to \infty} v_n(i, \pi) = v(i, \pi), \]

where

\[ v_n(i, \pi) = \mathbb{E}_{i, \pi} \sum_{k=0}^{n-1} r(X_k, A_k). \]

The value of the total reward MDP is defined by

\[ v^*(i) := \sup_{\pi \in \Pi} v(i, \pi). \]

Further we denote the value of the MDP where the negative rewards are neglected by \( u^* \), so

\[ u^*(i) := \sup_{\pi \in \Pi} u(i, \pi). \]

If the argument \( i \) corresponding to the state is deleted the function on \( S \) is meant. So for example \( v(\pi) \) and \( v^* \) are the functions with \( i \)-th coordinate \( v(i, \pi) \) and \( v^*(i) \) respectively. Often these functions will be treated as column vectors.

It is useful to have the following notations. Let \( f \) be any policy then the immediate reward function \( r(f) \) and the transition probability function \( P(f) \), which will be treated as a column vector and a matrix respectively, are defined by

\[ r(f)(i) := r(i, f(i)), \quad i \in S \]
\[ P(f)(i, j) := p(i, f(i), j), \quad i, j \in S. \]
On suitable subsets of functions on $S$ we define the operators $L(f)$ and $U$ by

\begin{align}
L(f)v &= r(f) + P(f)v \\
Uv &= \sup_{f \in \mathcal{F}} L(f)v 
\end{align}

Van Hae [1978] has shown that Condition 1 implies $u^* < \infty$ (i.e. $u^*(i) < \infty$ for all $i \in S$), whence also $v^* < \infty$, and further that

\begin{equation}
(1.11) \quad v^*(i) = \sup_{\pi \in \Pi} v(i, \pi).
\end{equation}

Ornstein [1969] proved that if all rewards are nonnegative then for each $\varepsilon > 0$ a stationary strategy $f$ exists satisfying

\begin{equation}
(1.12) \quad v(f) \geq (1 - \varepsilon)v^*.
\end{equation}

Strauch [1966] showed that if $r(i, a) \leq 0$ for all $i \in S$, $a \in A$ and if $A$ is finite then an optimal stationary strategy exists, i.e. an $f \in \mathcal{F}$ with

\begin{equation}
(1.13) \quad v(f) = v^*.
\end{equation}

This result has been generalized by Hordijk [1974, Theorem 6.3.c]: if $v^* < 0$ and if $f$ is a conserving policy, i.e.

\begin{equation}
(1.14) \quad r(f) + P(f)v^* = v^*,
\end{equation}

then

\begin{equation}
(1.15) \quad v(f) = v^*.
\end{equation}
Van der Wal [1981, Theorem 2.22] proved that if \( A \) is finite then

\[
(1.15) \quad v^{*}(i) = \sup_{f \in F} v(i, f),
\]

where the condition \( A \) is finite can be weakened to any other condition guaranteeing the existence of an optimal stationary strategy for the MDP with rewards \( r^{-}(i, a) \) with \( r^{-}(i, a) := \min\{0, r(i, a)\} \).

One of the two main results in this paper is the following theorem which connects the partial results mentioned above.

**Theorem 1.1.** If in each state \( i \in S \) for which \( v^{*}(i) \leq 0 \) a conserving action exists (cf. (1.14)) then for each \( \varepsilon > 0 \) a stationary strategy \( f \) exists satisfying

\[
(1.16) \quad v(f) \geq v^{*} - \varepsilon u^{*}.
\]

So in the positive and the negative dynamic programming cases this theorem yields the results of Ornstein, Strauch and Hordijk. And in the case of both positive and negative rewards it generalizes the result of Van der Wal to uniform nearly optimality.

Our second result is the following.

**Theorem 1.2.** If an optimal strategy exists then also an optimal stationary strategy exists.

This theorem extends results of Strauch [1966, Theorem 8.3] and Ornstein [1969, Proposition B] to the case of both positive and negative immediate rewards.
The organisation of the paper is as follows. The next section gives a brief outline of the proof of Theorem 1.1. Then in the Sections 3 - 6 various parts of the proof will be worked out. Finally in Section 7 the proof of Theorem 1.2 is given.

2. Outline of the proof

In this section the proof of Theorem 1.1 will be sketched. Therefore we first split up the state space $S$ into the subsets $S^- := \{i \in S \mid v^*(i) \leq 0\}$ and $S^+ := \{i \in S \mid v^*(i) > 0\}$. In the first part of the proof (Section 3) in each state $i \in S^-$ an arbitrary conserving action is fixed and it is shown that the value of this restricted MDP is the same as the value of the original MDP. As a result of this one can embed, having fixed conserving actions on $S^-$, the MDP on $S^+$ thus obtaining an MDP with strictly positive value function.

The next step in the proof (Section 4) is to consider an MDP with $v^* > 0$ but positive as well as negative immediate rewards. For such an MDP we construct a modified MDP with immediate rewards $\tilde{r}(i,a)$ and transition probabilities $\tilde{p}(i,a,j)$ as follows:

\[
(2.1) \begin{cases}
\text{(i)} & \text{if } r(i,a) \geq 0 \text{ then } \tilde{r}(i,a) := r(i,a) \text{ and } \tilde{p}(i,a,j) := p(i,a,j) \\
\text{(ii)} & \text{if } r(i,a) < 0 \text{ and } r(i,a) + \sum_j p(i,a,j)v^*(j) \geq 0 \text{ then } \\
& \tilde{r}(i,a) := 0 \text{ and } \\
& \tilde{p}(i,a,j) := \frac{r(i,a) + \sum_k p(i,a,k)v^*(k)}{\sum_k p(i,a,k)v^*(k)} p(i,a,j) \\
\text{(iii)} & \text{if } r(i,a) < 0 \text{ and } r(i,a) + \sum_j p(i,a,j)v^*(j) < 0 \text{ then } \\
& \tilde{r}(i,a) := r(i,a) + \sum_j p(i,a,j)v^*(j) \text{ and } \tilde{p}(i,a,j) := 0.
\end{cases}
\]
So in this modified MDP one "borrows from the future" if the immediate reward is negative but never more then to pay the immediate loss and also not more then possible, and the transition probabilities are adapted accordingly.

It will be shown that this modified MDP has the same value as the original one (with \( \tilde{v}^* > 0 \)) and that in each state \( i \) all actions for which \( \tilde{r}(i,a) < 0 \) (case (iii)) can be eliminated without affecting the value of the MDP. This way we obtain a positive dynamic programming problem. Using Ornsteins result we have for each \( \varepsilon > 0 \) the existence of a stationary strategy \( f \) satisfying

\[
(2.2) \quad \tilde{v}(f) \geq (1 - \varepsilon)v^*
\]

(tildes are used for the modified MDP).

Then, in Section 5, it will be shown that a stationary strategy \( f \) satisfying \( (2.2) \) also satisfies

\[
(2.3) \quad v(f) \geq v^* - \varepsilon(1 - \varepsilon)^{-1}u^*.
\]

This proves Theorem 1.1 for positive valued MDP's.

In Section 6 the proof is extended to the case that \( S^- \) is nonempty.

3. On \( S^- \) conserving actions are enough

In this section we prove the following result.

Theorem 3.1. Let \( f \) be any conserving policy on \( S^- \), i.e.

\[
\tilde{r}(f) + P(f)v^* = v^* \text{ on } S^-.
\]
Let further \( \Pi_f \) denote the set of all strategies in \( \Pi \) which on \( S^- \) act according to \( f \). Then

\[
\sup_{\pi \in \Pi_f} v(i, \pi) = v^*(i) \quad \text{for all } i \in S
\]

(So restricting to \( f \) on \( S^- \) does not affect the value of the MDP).

In order that this theorem holds one certainly needs the following lemma.

Let the stopping time \( \tau \) be the time of the first switch from \( S^- \) to \( S^+ \) or vice versa:

\[
\tau(i_0, i_1, \ldots) = \begin{cases} 
\inf \{ n \mid i_n \in S^+ \} & \text{if } i_0 \in S^- \\
\inf \{ n \mid i_n \in S^- \} & \text{if } i_0 \in S^+
\end{cases}
\]

for all \( i_0, i_1, \ldots \in S \), and where \( \inf \emptyset := \infty \).

(At this time we only use that \( \tau \) is the first exit time on \( S^- \), the ampler definition is for later use).

**Lemma 3.2.** Let \( f \) be any conserving policy on \( S^- \). Then

\[
E_{i,f} \left[ \sum_{n=0}^{\tau-1} r(X_n, A_n) + v^*(X_{\tau}) \right] = v^*(i) \quad \text{for all } i \in S^-,
\]

where

\[
v^*(X_\tau) := 0 \quad \text{if } \tau = \infty.
\]

(Note that if \( v^*(i) > 0 \) an equality like (3.2) need not hold as in the positive dynamic programming case conserving actions need not be optimal.)
Proof. The expression at the left hand side of (3.2) is clearly equal to the total expected reward for the Markov process on \( S^- \) with rewards \( \tilde{r}(i,a) \) and transition probabilities \( \tilde{p}(i,a,j) \) defined by

\[
\begin{align*}
\tilde{r}(i,a) &:= r(i,a) + \sum_{j \in S^+} p(i,a,j) v^*(j), \quad i \in S^- \\
\tilde{p}(i,a,j) &:= p(i,a,j), \quad i,j \in S^-.
\end{align*}
\]

(In order to have the transition probabilities add up to 1 we could add an extra absorbing state where no more returns are obtained; we will not do this explicitly.)

For the MDP with state space \( S^- \) defined by (3.3) we have, since \( f \) is conserving,

\[
\tilde{r}(f) + \tilde{p}(f)v^* = v^*.
\]

Further

\[
\tilde{r}(f) + \tilde{p}(f)\nu(f) = \nu(f),
\]

where

\[
\nu(i,f) := E_{i,f} \left[ \sum_{n=0}^{N-1} r(X_n,A_n) + v^*(X_N) \right], \quad i \in S^-.
\]

Clearly \( v^* < 0 \) (we only consider \( S^- \)) and \( \nu(f) \leq v^* \). But also \( \nu(f) \) is the largest nonpositive solution of \( \tilde{r}(f) + \tilde{p}(f)v = v \) (see e.g. Van der Wal [1981, Theorem 2.18]), hence \( \nu(f) \geq v^* \). So \( \nu(f) = v^* \), which proves the lemma. \( \square \)
The next step in the proof of Theorem 3.1 is the construction of a strategy 
\( \pi \in \Pi_{\mathcal{F}} \) (i.e. using \( f \) on \( S^- \)) which is nearly optimal.

Let \( \pi_n^{(n)}, n = 1, 2, \ldots \) be a strategy satisfying

\[ \nu(i, \pi_k^{(k)}) \geq \nu^*(i) - \epsilon 2^{-k} \text{ for all } i \in S^+. \]

(Clearly such a strategy exists within \( \Pi \) but not necessarily within \( M \).) Then also

\[ \mathbb{E}_{i, \pi} (k) \left[ \sum_{n=0}^{\tau-1} r(X_n, A_n) + \nu^*(X_{\tau}) \right] \geq \nu^*(i) - \epsilon 2^{-k}, \]

for all \( k = 1, 2, \ldots \) and all \( i \in S^+ \).

Now let \( \pi^* \) be the strategy which on \( S^- \) acts according to \( f \) and on \( S^+ \) uses.
strategy \( \pi_k^{(k)} \) during the \( k \)-th stay in \( S^+ \), pretending the process to restart.
at the time of the \( k \)-th entry to \( S^+ \). Then \( \pi^* \) satisfies the following lemma.

**Lemma 3.3.** \( \nu(\pi^*) \geq \nu^* - \epsilon \alpha \).

**Proof.** To prove this we introduce the following notations. Define the func-
tions \( b_{\tau}(\pi^{(k)}) \) on \( S^+ \) and \( c_{\tau}(f) \) on \( S^- \) by

\[ b_{\tau}(i, \pi^{(k)}) := \mathbb{E}_{i, \pi} (k) \sum_{n=0}^{\tau-1} r(X_n, A_n) , \quad i \in S^+, \]

\[ c_{\tau}(i, f) := \mathbb{E}_{i, f} \sum_{n=0}^{\tau-1} r(X_n, A_n) , \quad i \in S^- . \]
Further define the transition probability matrices $Q_t(\pi(k))$ from $S^+$ to $S^-$ and $R_t(f)$ from $S^-$ to $S^+$ by

$$Q_t(\pi(k))(i,j) = \mathbb{P}_{i,\pi(k)}(\tau < \omega, X_\tau = j), \quad i \in S^+, j \in S^-$$

$$R_t(f)(i,j) = \mathbb{P}_{i,f}(\tau < \omega, X_\tau = j), \quad i \in S^-, j \in S^+.$$  

Then we have

$$(3.4) \quad b_t(\pi(k)) + Q_t(\pi(k))v^* \geq v^* - \varepsilon 2^{-k}e \quad \text{on } S^+,$$

where $e$ is the unit function on $S^+$, and by Lemma 3.2

$$(3.5) \quad c_t(f) + R_t(f)v^* = v^* \quad \text{on } S^-.$$

Now define on $S^+$ (suppressing the dependence on $\pi^*$)

$$\begin{align*}
    v_0 &:= 0 \\
    v_1 &:= b_t(\pi(1)) + Q_t(\pi(1))c_t(f) \\
    &\vdots \\
    v_n &:= v_{n-1} + P_{\tau,n-1} [b_t(\pi(n)) + Q_t(\pi(n))c_t(f)],
\end{align*}$$

where the transition (sub)probability matrix $P_{\tau,n}$ on $S^+$ is defined by

$$\begin{align*}
    P_{\tau,0} &:= I \\
    P_{\tau,n} &:= P_{\tau,n-1} Q_t(\pi(n)) R_t(f), \quad n = 1, 2, \ldots.
\end{align*}$$
Then
\[ v_n + P_{\tau,n} v^* = \]
\[ = v_{n-1} + P_{\tau,n-1} \left[ b_\tau(\pi(n)) + Q_\tau(\pi(n)) \left[ c_\tau(\epsilon) + R_\tau(\epsilon)v^* \right] \right] \]
\[ = v_{n-1} + P_{\tau,n-1} \left[ b_\tau(\pi(n)) + Q_\tau(\pi(n))v^* \right] \]
\[ = v_{n-1} + P_{\tau,n-1} v^* - \epsilon 2^{-n} \epsilon \]
\[ = \ldots \geq v^* - \epsilon [2^{-1} + \ldots + 2^{-n}] \epsilon > v^* - \epsilon \epsilon . \]

Clearly \( v_n \) converges to \( v(\pi^*) \) if \( n \) tends to \( \infty \), so if \( \limsup_{n \to \infty} P_{\tau,n} v^* \leq 0 \) then \( v(\pi^*) \geq v^* - \epsilon \epsilon \) on \( S^+ \). To prove this observe that on \( S^+ \) we have

\[ \limsup_{n \to \infty} P_{\tau,n} b_\tau(\pi(n+1)) \leq 0 \]

since \( \tau \geq 1 \) and for a fixed strategy the sum of the positive rewards from time \( n \) onwards tends to 0 if \( n \) tends to \( \infty \). Also

\[ b_\tau(\pi(n+1)) \geq v^* - \epsilon 2^{-(n+1)} \epsilon - Q_\tau(\pi(n+1))v^* \]
\[ \geq v^* - \epsilon 2^{-(n+1)} \epsilon \quad \text{(since } Q_\tau(\pi(n+1))v^* \leq 0 \text{)} . \]

So from (3.6) and (3.7) we have on \( S^+ \)

\[ \limsup_{n \to \infty} P_{\tau,n} v^* \leq \limsup_{n \to \infty} P_{\tau,n} [b_\tau(\pi(n+1)) + \epsilon 2^{-(n+1)} \epsilon] \leq 0 . \]
Remains to prove the lemma on $S^-$. 

Let $i \in S^-$ then

$$v(i, \pi) = c_T(i, \pi) + \sum_{j \in S^+} R_T(\pi)(i, j)v(j, \pi^*) \geq c_T(i, \pi) + \sum_{j \in S^+} R_T(\pi)(i, j)v^*(j) - \varepsilon$$

$$= v^*(i) - \varepsilon .$$

This completes the proof of the lemma. \qed

Since in Lemma 3.3 $\varepsilon > 0$ can be chosen arbitrarily Lemma 3.3 also proves Theorem 3.1.

Note that as a consequence of Theorem 3.1 we can fix some arbitrary conserving policy on $S^-$ and then embed the MDP on $S^+$. The embedded MDP on $S^+$ then has a strictly positive value function.

4. Positive valued MDP's

In this section we consider an MDP with strictly positive value: $v^*(i) > 0$ for all $i \in S$. It will be shown that transformation (2.1) yields a new MDP with the same value function and essentially only actions with nonnegative immediate rewards.

All objects in the transformed MDP will be labeled by a ~.
Lemma 4.1. \( \bar{v}^* = v^* \).

Proof. First it will be shown that \( \bar{v}^* \leq v^* \).

As one easily verifies:

\[
L(f)v^* = \bar{L}(f)v^*.
\]

Hence, as \( Uv^* = v^* \), also

\[
\bar{U}v^* = v^*.
\]

So

\[
\bar{v}^* \leq \limsup_{n \to \infty} U^n 0 \leq \limsup_{n \to \infty} U^n v^* = v^*,
\]

where the first inequality can be found in Schäl [1975, Formula (2.5)] and the second one is immediate from \( v^* > 0 \). Remains to show that \( v^* \leq \bar{v}^* \).

It follows from (1.11) that it suffices to prove that for all \( \pi \in M \) we have \( v(\pi) \leq \bar{v}(\pi) \).

Let \( \pi \in M \) be an arbitrary Markov strategy, then \( \pi \) can be characterized by the policies to be followed at each time: \( \pi = (f_0,f_1,\ldots) \). Let \( d(f) \) be the function defined by

\[
d(f) = v^* - L(f)v^*,
\]

so \( d(f) \geq 0 \). Then we have for the \( n \)-period reward \( v_n(\pi) \) (see (1.4))
Similarly

\begin{align}
(4.1) \quad v_n(\pi) &= L(f_0) L(f_1) \ldots L(f_{n-1}) 0 = \\
&= L(f_0) \ldots L(f_{n-1}) v^* - P(f_0) \ldots P(f_{n-1}) v^* \\
&= L(f_0) \ldots L(f_{n-2}) (v^* - d(f_{n-1})) - P(f_0) \ldots P(f_{n-1}) v^* \\
&= \ldots = v^* - \sum_{k=0}^{n-1} \prod_{j=0}^{k-1} P(f_j) d(f_k) - \prod_{j=0}^{n-1} P(f_j) v^* .
\end{align}

Since \( P(f) \geq \tilde{P}(f) \) for all \( f \) we have from (4.1) and (4.2)

\begin{align}
(4.2) \quad \tilde{v}_n(\pi) &= v^* - \sum_{k=0}^{n-1} \prod_{j=0}^{k-1} \tilde{P}(f_j) d(f_k) - \prod_{j=0}^{n-1} \tilde{P}(f_j) v^* .
\end{align}

Next we want to show that this modified MDP can be reduced to an MDP with non-negative immediate rewards only.

\begin{align}
\tilde{v}_n(\pi) &\leq \tilde{v}_n(\pi) \quad \text{for all } n, \\
\text{whence also} \\
v(\pi) &\leq \tilde{v}(\pi) .
\end{align}

As remarked before this implies \( v^* \leq \tilde{v}^* \) which completes the proof of the lemma.
Lemma 4.2. Let $\Pi_+$ be the set of all strategies which use only actions for which $\tilde{r}(i,a) \geq 0$. Then

$$\sup_{\pi \in \Pi_+} \tilde{v}(\pi) = v^*.$$ 

L.e., we can eliminate in each state $i \in S$ all those actions for which $\tilde{r}(i,a) < 0$ without affecting the value.

Proof. Let $\pi \in \Pi$ be an arbitrary Markov strategy. Let further $f$ be some policy with $\tilde{r}(f) \geq 0$. (Since $v^* > 0$ such a policy exists.) Now consider the strategy $\tilde{\pi}$ which is the following combination of $\pi$ and $f$. In words: play $\pi$ until you first reach a state where $\pi$ prescribes an action for which the immediate "-reward" is negative; instead of taking this action you switch to $f$ and you play $f$ for ever after. Clearly $\tilde{v}(f) \geq 0$, so this switch yields you a better total expected reward, i.e.

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

Hence

$$\sup_{\Pi_+} v(\pi) \geq \sup_{\Pi_+} \tilde{v}(\pi) = v^*.$$ 

Clearly $v(\pi) \leq v^*$ for all $\pi$, which completes the proof of the lemma. $\square$

Now in each state all actions yielding negative immediate payoffs can be eliminated which gives us a positive dynamic programming problem. So by Ornsteins result we obtain the following corollary.
Corollary 4.3. For each $\epsilon > 0$ a policy $f$ exists such that

$$\gamma(v(f)) \geq (1 - \epsilon)v^*.$$  

5. Nearly-optimality for the positive valued MDP

The main result of this section is the following theorem.

Theorem 5.1. Suppose we have an MDP with $v^* > 0$ and let $f$ satisfy $\gamma(f) \geq 0$ and $\gamma(v(f)) \geq (1 - \epsilon)v^*$. Then

$$\gamma(v(f)) \geq v^* - \epsilon(1 - \epsilon)^{-1}u^*.$$  

This theorem extends Ornstein's result and already establishes Theorem 1.1 for a special case.

In order to prove this theorem we need two lemmas. The main tool in our approach is the concept of the so-called stationary randomized and action dependent go ahead function as used in Van Nunen en Stidham [1981] and Van der Wal [1981, Chapter 3].

This gives us a different view upon the datatransformation (2.1).

Let $\delta$ be a go-ahead function on $S \times A$ with

$$\delta(i,f(i)) = 1 \text{ if } r(i,f(i)) \geq 0$$

$$\delta(i,f(i)) = \frac{r(i,f(i)) + \sum_j p(i,f(i),j)v^*(j)}{\sum_j p(i,f(i),j)v^*(j)} \text{ if } r(i,f(i)) < 0.$$  

So $\gamma(p(i,f(i),j)) = \delta(i,f(i))p(i,f(i),j).$
Since we are only interested in the Markov process where policy \( f \) is used it suffices here to define \( \delta \) for the policy \( f \). The idea of this go-ahead function is that if in a state an action is taken with negative immediate reward, then with probability \( 1 - \delta \) the process stops after the next transition and with probability \( \delta \) the process continues. Let \( \tau_\delta \) be the time upon which the process is stopped. (A more formal introduction of go-ahead functions and stopping times can be found in Van Nunen and Stidham [1981] and Van der Wal [1981].)

Now define the reward functions \( r^+_\delta(f) \) and \( r^-_\delta(f) \) by

\[
\begin{align*}
    r^+_{\delta}(f) &:= \mathbb{E}_f \sum_{n=0}^{\tau_\delta - 1} r^+(X_n, A_n) \\
    r^-_{\delta}(f) &:= \mathbb{E}_f \sum_{n=0}^{\tau_\delta - 1} r^-(X_n, A_n)
\end{align*}
\]

the matrix \( P^\delta(f) \) by

\[
P^\delta(f)(i, j) = \mathbb{P}_{i, f} (\tau_\delta < \infty, X_{\tau_\delta} = j),
\]

and the operator \( L^\delta(f) \) by

\[
L^\delta(f)v = r^\delta(f) + P^\delta(f)v,
\]

where

\[
r^\delta(f) = r^+_\delta(f) + r^-_\delta(f).
\]

Then we have

\[
(5.3) \quad \tilde{v}(f) = L^\delta(f)v^* = r^+_\delta(f) + r^-_\delta(f) + P^\delta(f)v^* = r^+_\delta(f).
\]
Since stopping occurs to cover up the negative immediate rewards one has

\[(\tau^-(f) + P_\delta(f)v^* = 0).\]

From (5.3) and \(f\) being a policy satisfying \(\widetilde{v}(f) \geq (1 - \varepsilon)v^*\) we also have

\[(5.4)\quad L_\delta(f)v^* \geq (1 - \varepsilon)v^*.

Further, since \(\tau_\delta \geq 1\), (cf. the proof of Lemma 4.24 in Van der Wal [1981])

\[(5.5)\quad v(f) = \lim_{n \to \infty} L^n_\delta(f)0.

By (5.4) we have

\[
L^n_\delta(f)0 = L^n_\delta(f)v^* - P^n_\delta(f)v^*
\]

\[
= L^{n-1}_\delta(f)(1 - \varepsilon)v^* - P^n_\delta(f)v^*
\]

\[(5.6)\quad \geq \ldots \geq v^* - \varepsilon \sum_{k=0}^{n-1} P^k_\delta(f)v^* - P^n_\delta(f)v^*.

From (5.5) and (5.6) one sees that in order to prove (5.1) it suffices to prove that

\[(5.7)\quad \sum_{n=0}^{\infty} P^n_\delta(f)v^* \leq (1 - \varepsilon)^{-1}u^*,

as then clearly \(P^n_\delta(f)v^* \to 0\) \((n \to \infty)\).

To prove this we need the following lemma.
Lemma 5.2.

\[ u(\delta) = \sum_{n=0}^{\infty} P^R_\delta(\delta)r^\nu_\delta(\delta) \cdot \]

**Proof.** Immediate from \( v_\delta \geq 1 \) (cf. also (5.5)). \( \square \)

From this we obtain almost immediately

Lemma 5.3.

\[ \sum_{n=0}^{\infty} P^R_\delta(\delta)v^* \leq (1 - \epsilon)^{-1}u^* \cdot \]

**Proof.** \( u(\delta) \leq u^* \) and by (5.3) and (5.4) also \( r_\delta^\nu(\delta) \geq (1 - \epsilon)v^* \).

So

\[ \sum_{n=0}^{\infty} P^R_\delta(\delta)v^* \leq (1 - \epsilon)^{-1} \sum_{n=0}^{\infty} P^R_\delta(\delta)r^\nu_\delta(\delta) = (1 - \epsilon)^{-1}u(\delta) \leq (1 - \epsilon)^{-1}u^* \cdot \square \]

As remarked before this proves Theorem 5.1.

6. Proof of Theorem 1.1

Finally we show how the proof of Theorem 5.1 can be extended to Theorem 1.1.

So we start with some arbitrary countable state MDP for which in all states with \( v^*(i) \leq 0 \) a conserving action exists. Now fix a conserving action in each state in \( S^- \). As shown in Section 3 this does not affect the value of the MDP. Next we embed the MDP on \( S^+ \) thus obtaining an MDP with strictly positive value function. For this MDP apply Theorem 5.1. Further it is clear that the \( u^* \) function of the original MDP (on \( S^+ \)) is larger than or equal to the \( u^* \) of the embedded MDP. So a policy which is conserving on \( S^- \) and satisfies (5.1) for the embedded MDP also satisfies (5.1) for the original MDP. This completes the argument for \( S^+ \).
Now consider $S^-$. Let $f$ be some policy which is conserving on $S^-$ and satisfies (1.16) on $S^+$. Then for any $i \in S^-
abla$

$$v(i, f) = E_{i, f} \left[ \sum_{n=0}^{\tau-1} r(X_n, A_n) + v(X_{\tau}, f) \right]$$

$$\geq E_{i, f} \left[ \sum_{n=0}^{\tau-1} r(X_n, A_n) + v^*(X_{\tau}) \right] - \epsilon E_{i, f} u^*(X_{\tau})$$

$$= v^*(i) - \epsilon E_{i, f} u^*(X_{\tau}) \geq v^*(i) - \epsilon u^*(i) \nabla$$

since

$$E_{i, f} u^*(X_{\tau}) \leq E_{i, f} \left[ \sum_{n=0}^{\tau-1} r(X_n, A_n) + u^*(X_{\tau}) \right] \leq u^*(i) \nabla$$

And this completes the proof of Theorem 1.1.

7. Proof of Theorem 1.2

This section shows that the line of proof of Theorem 1.1 yields almost immediately also a proof for Theorem 1.2.

First the proof is given for the case that all actions are conserving. At the end of this section it will be shown that if an optimal strategy exists the MDP is equivalent with an MDP with only conserving actions.

As all actions are conserving we can fix again actions on $S^-$ without affecting the value, embed the MDP on $S^+$, and transform this positive valued MDP to an MDP with nonnegative immediate rewards. In this MDP all actions are still conserving. Then Proposition B in Ornstein [1969] states that even an optimal stationary strategy exists. (One may easily show that in the positive dynamic programming case a stationary strategy $f$ satisfying $v(f) \geq \delta v^*$
with \( \delta > 0 \) is also optimal.\) So by Theorem 5.1 an optimal stationary strategy exists in any positive valued MDP. Finally it is immediate from Section 6 that this strategy combined with the fixed actions on \( S^- \) is optimal for the original MDP. This completes the proof of Theorem 1.2 for the "conserving case". 

Note that we actually proved that if in an MDP all actions are conserving, then an optimal stationary strategy exists. For this the condition that an optimal strategy exists was not needed.

To complete the proof we will show that if an optimal strategy exists then the action set in each state can be restricted to the set of conserving actions without affecting the value of the MDP (or the existence of an optimal strategy). Let \( \pi \in \Pi \) be optimal, i.e. \( v(\pi) = v^* \). And let \( \pi(i) \) be a randomized Markov strategy having the same marginal distributions for initial state \( i \) as the optimal strategy \( \pi \). Further let \( \pi_c(j) \) be the randomized action prescribed by \( \pi(i) \) if at time \( t \) the system is in state \( j \).

Define \( A_c(i) \) to be the set of conserving actions in state \( i \), so

\[
A_c(i) := \{ a \in A \mid r(i, a) + \sum_j p(i, a, j) v^*(j) = v^*(i) \} .
\]

Then \( \mathbb{P}_{i,\pi}(X_t = j) > 0 \) implies \( \mathbb{P}_{i,\pi}(A_c \in A_c(j) \mid X_t = j) = 1 \) so \( \pi(t)(j) \) yields with probability 1 a conserving action.

Since \( \pi(t)(i) \) appoints with probability 1 a conserving action \( A_c(i) \) is nonempty for all \( i \). Further if \( \mathbb{P}_{i,\pi}(X_t = j) > 0 \) then \( \pi(t)(j) \) induces a randomized action on \( A_c(j) \). So \( \pi(t)(i) \) induces a strategy \( \tilde{\pi}(i) \) which uses conserving actions only. (If \( \mathbb{P}_{i,\pi}(X_t = j) = 0 \) then \( \tilde{\pi}(i)(j) \) can be chosen to be any conserving action from \( A_c(j) \).) And this strategy \( \tilde{\pi}(i) \) has with probability 1 the same marginal distributions, so \( v(i, \tilde{\pi}(i)) = v^*(i) \). So restricting the action sets to the conserving actions does not affect the value which completes the proof of Theorem 1.2.
References


