A policy improvement-value approximation algorithm for the ergodic average reward Markov decision process

van der Wal, J.

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 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):

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

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 07. Dec. 2018
Memorandum COSOR 78-27

A policy improvement-value approximation algorithm for the ergodic average reward Markov decision process

by

J. van der Wal

Eindhoven, december 1978

The Netherlands
A policy improvement-value approximation algorithm for the ergodic average
reward Markov decision process

by

J. van der Wal

Abstract. This paper presents a policy improvement-value approximation
algorithm for the average reward Markov decision process when all
transition matrices are unichained. In contrast with Howard's algorithm
we do not solve for the exact gain and relative value vector but only
approximate them. It is shown that the value approximation algorithm
produces a nearly optimal strategy. This paper extends the results of
a previous paper in which transient states were not allowed. Also the
algorithm is slightly different.

1. Introduction and notations

In this paper we consider the average reward Markov decision process
(MDP) with finite state and action spaces. In [4] we introduced a value
approximation algorithm for the case that all chains are completely
ergodic (i.e. only one recurrent subchain and no transient states).
This value approximation algorithm is a variant of Howard's policy
iteration algorithm [2]. The difference is that we do not solve for the
exact gain and relative value vector of the actual policy but only
approximate them. For our convergence proof, however, these approximations
have to be sufficiently accurate. Here we extend these results to the
case that all possible transition matrices are ergodic, i.e. have only
one recurrent subchain but there may be transient states. Also the
convergence proof we give here is more transparent.

So we are concerned with a MDP with finite state space $S := \{1, 2, \ldots, N\}$
and finite action space $A$. If in state $i \in S$ action $a \in A$ is taken, then
the immediate reward is $r(i, a)$ and the process moves to state $j$ with
probability $P_{ij}^a$.

A policy or stationary strategy is a map $f : S \rightarrow A$. With a policy $f$
we associate the immediate reward vector $r_f$ and the transition matrix $P_f$
deefined by
\( r_f(i) := r(i, f(i)) , \quad i \in S . \)

\( P_f(i, j) := \frac{f(i)}{i_j} , \quad i, j \in S . \)

We will assume that all matrices \( P_f \) are aperiodic (this gives no loss of generality, as one may use Schweitzer's data transformation \( \tilde{r}_f := r_f , \) \( \tilde{P}_f := \alpha I + (1 - \alpha)P_f , \) \( 0 < \alpha < 1 , \) to transform any MDP into an equivalent aperiodic one, see [3]).

Let \( g_f \) be the gain and \( v_f \) be the relative value vector of a policy \( f \).

Then, at least in the unichain case we consider here, \( g_f , v_f \) uniquely solve

\[
\begin{align*}
(1.1) & \quad \sum r_f + P_f v = v + g_f e , \quad (e = (1, \ldots , 1)^T) \\
(1.2) & \quad P_f^* v = 0 ,
\end{align*}
\]

where \( P_f^* \) is defined by

\[
P_f^* := \lim_{T \to \infty} T^{-1} \sum_{n=0}^{T-1} P_f^n .
\]

From the aperiodic assumption we also have

\[
(1.3) \quad P_f^* = \lim_{n \to \infty} P_f^n .
\]

Further we have

\[
P_f^* r_f = g_f e .
\]

Let \( g^* \) be the gain of the MDP, \( g^* := \max_f g_f \).

We are interested in finding an \( \varepsilon \)-optimal policy, i.e. a policy \( f \) satisfying

\[
g_f \geq g^* - \varepsilon .
\]
In order to determine such an $\epsilon$-optimal policy we propose the following approximated version of Howards policy iteration algorithm (which we already presented in [4] in a slightly different form).

**Policy improvement-value approximation algorithm**

The main iteration step reads as follows. (The constants $\alpha$ and $\epsilon$ occurring in this iteration step are chosen beforehand)

**Value approximation.** Let $f$ be the current policy, then determine

$$v_t = r_f + P_f v_{t-1}, \quad t = 1, 2, \ldots \text{ until}$$

$$\text{sp}(v_t - v_{t-1}) \leq \epsilon \tag{1.4}$$

where $v_0$ is obtained in the previous iteration step and $\text{sp}(v)$ denotes the span of a vector $v$:

$$\text{sp}(v) := \max_i v(i) - \min_i v(i).$$

**Policy improvement.** Let $n$ be the first index for which (1.4) holds. Then define $\gamma$ by

$$\gamma := \max_h \left( r_h + P_h v_n \right) - r_f - P_f v_n \tag{1.5}$$

Clearly $\gamma \geq 0$. Now we distinguish two cases

(i) $\gamma \leq \alpha$.

Then $f$ is nearly optimal.

(ii) $\gamma(i) > \alpha$ for some $i \in S$. Then replace $f$ by a policy $h$ with

$$h(i) = f(i) \text{ if } \gamma(i) \leq \alpha,$$

$$h(i) = \text{ a maximizer of } r(i,a) + \sum_j p_{ij} v_n(j) \text{ if } \gamma(i) > \alpha.$$ 

Return to the value approximation step with policy $f$ replaced by $h$ and $v_0 := v_n$. 


Note that for the policy $h$ determined according to (ii) we now have
\begin{equation}
(1.7) \quad r_h + P_h v_n = r_f + P_f v_n + \zeta
\end{equation}
with $\zeta(i) = 0$ if $\gamma(i) \leq \alpha$ and $\zeta(i) = \gamma(i)$ if $\gamma(i) > \alpha$.

In the remainder we prove that this algorithm converges.

In section 2 we derive some preliminary results.

In section 3 we prove the correctness of formula (1.6). In the sections 4 and 5 we prove that if $c$ is sufficiently small compared to $\alpha$ then a replacement of a policy $f$ by a policy $h$ is indeed an improvement. In order to prove this we will distinguish two cases.

A : $\zeta(i) > \alpha$ for some state $i$ which is recurrent under $P_h$. Then, if $\alpha/c$ is sufficiently large, $q_h > q_f$, which will be proved in section 4.

B : $\zeta(i) = 0$ (hence $h(i) = f(i)$) in all states which are recurrent under $P_h$. Then $q_h = q_f$, $v_h \geq v_f$ and if $\zeta(i) > \alpha$ then $v_h(i) > v_f(i)$, provided that $\alpha/c$ sufficiently large. This we prove in section 5.

In section 6 we collect some concluding remarks.

\section*{2. Some preliminary results}

In this section we collect some preliminary results needed for the proofs in the following sections.

First we give a stronger result than the mere convergence of $P^n_f$ to $P^*_f$ formulated in (1.3):

\begin{lemma}
There exist constants $b$ and $0 < \rho < 1$ such that
\[ |P^*_f(i,j) - P^n_f(i,j)| \leq b \rho^n \]
for all $i, j \in S$, $f$ and $n \in \mathbb{N}$.
\end{lemma}

A proof of this basic result can be found for example in Doob [1].

For notational convenience we use the operator $L_f$ on $\mathbb{R}^N$, defined by
\[ L_f v = r_f + P_f v \]

In the following lemma we give some results concerning the behaviour of the approximation $v_t = L_f v_{t-1}$ in the value approximation step of our
Lemma 2.2. For any $v \in \mathbb{R}^n$

(i) $p_f^*(L_f v - v) = g_f e$

(ii) $\min_i (L_f v - v)(i) \leq g_f \leq \max_i (L_f v - v)(i)$

(iii) $L_f^{n+1} v - L_f^n v = p_f^n (L_f v - v)$

(iv) $L_f^{n+1} v - L_f^n v - g_f e = (p_f^n - p_f^*) (L_f v - v)$.

Proof.

(i) $p_f^* p_f = p_f^*$, so

$$p_f^* (L_f v - v) = p_f^* (r_f + p_f v - v) = p_f^* r_f = g_f e.$$ 

(ii) Immediate from (i) as $L_f v - v$ must have components at least equal to $g_f$ as well as components at most equal to $g_f$.

(iii) $L_f^{n+1} v - L_f^n v = L_f L_f^n v - L_f L_f^{n-1} v = p_f^n (L_f v - L_f^{n-1} v)$

$$= \ldots = p_f^n (L_f v - v).$$

(iv) From (ii) and (iii).

From these two lemmas, it is clear, that $L_f^{n+1} v - L_f^n v$ converges to $g_f e$ exponentially fast. So we see that the differences $v_t - v_{t-1}$ in the value approximation step converge to $g_f e$ exponentially fast. Hence this step (provided that $\varepsilon > 0$) terminates.

In the policy improvement step the most important role however is played by the term $v_n = L_f^n v_0$.

In Howard's algorithm we there have $v_f$. But this difference, is not so large, as we have
Lemma 2.3. \( L^n_f v - n g_f e - v_f \rightarrow P^n_f v \) as \( n \to \infty \),
so the difference between \( L^n_f v \) and \( v_f \) becomes constant.

Proof. Clearly \( L^n_f v - L^n_w = P^n_f (v - w) \) and as \( v_f \) satisfies (1.1)
\[ L^n_f v_f = v_f + n g_f e. \]
So
\[ L^n_f v - n g_f e - v_f = L^n_f v - L^n_f v_f = P^n_f (v - v_f) + P^*_f (v - v_f) = P^*_f v. \]
as \( n \to \infty \), where we used that \( v_f \) satisfies (1.2). \( \square \)

And as we see adding a constant to the vector \( v_n \) in (1.5) does not change \( \gamma \).

3. The correctness of formula (1.6)

In this section we prove that if the algorithm terminates then the policy we find is \((\alpha + \varepsilon)\)-optimal.

Lemma 3.1. If \( sp(L_f v - v) \leq \varepsilon \) and \( L_h v \leq L_f v + \alpha e \) then \( g_h \leq g_f + \alpha + \varepsilon \).

Proof. From \( sp(L_f v - v) \leq \varepsilon \) and lemma 2.2(ii) we have
\[ L_f v \leq g_f e + v + \alpha e. \]
So
\[ L_h v - v \leq (g_f + \alpha + \varepsilon) e. \]
Premultiplying this with \( P^*_h \) and using lemma 2.2(i) we get
\[ g_h e \leq (g_f + \alpha + \varepsilon) e. \]
\( \square \)

From this we get

Corollary 3.1. If \( f \) is the current policy in the algorithm,
\( sp(v_n - v_{n-1}) \leq \varepsilon \) and \( \max_{h} L_h v_n - L_f v_n \leq \alpha e \) then
\[ g_f \geq g^* - \alpha - \varepsilon. \]
Proof. Almost immediately from lemma 3.1. We only have to observe that
\[ \text{sp}(v_n - v_{n-1}) \leq \varepsilon \] implies
\[ \text{sp}(L_f v_n - L_f v_{n-1}) \leq \varepsilon. \]

4. Case A: improvement in a recurrent state

In this section we show that, if \( h \) is an improvement of \( f \) according to
case A, then \( h \) has a higher gain than \( f \), provided that \( \varepsilon \) is sufficiently
small compared to \( \alpha \).

First define \( \theta \) by
\[
\theta := \min_{i,j} \{ P_f^*(i,j) | P_f^*(i,j) > 0 \}. 
\]

Clearly \( \theta > 0 \). And we get

Theorem 4.1. If \( f \) is replaced by \( h \) and \( \zeta(i) > \alpha \) for some \( i \) which is
recurrent under \( h \) and if \( \varepsilon/\alpha \leq \theta \), then

\[ g_h > g_f. \]

Proof. \( L_h v_n - v_n = L_f v_n - v_n + \zeta \geq (g_f - \varepsilon)e + \zeta. \)

Multiplying both sides with \( P_h^* \) we get
\[
g_h e = P_h^*(L_h v_n - v_n) \geq P_h^*(g_f - \varepsilon)e + P_h^* \zeta = (g_f - \varepsilon)e + P_h^* \zeta.
\]

\( \zeta(i) > \alpha \) for some recurrent state under \( h \), so \( P_h^* \zeta > \alpha \theta e. \)

Hence, if \( \alpha \theta \geq \varepsilon \) we have \( g_h > g_f \).

5. Case B: improvement in transient states only

Case B is the more difficult one. If \( f \) and \( h \) are equal on the set \( \text{Rec}(h) \)
of recurrent states under \( P_h \), then we must have \( g_f = g_h \). So, as in the
standard policy iteration algorithm, we have to investigate the relative
value vectors.

We will show in this section that if \( \varepsilon/a \) is sufficiently small then a 'case B type' improvement gives a policy \( h \) with a higher relative value vector \( v_h > v_f \). Where we write \( v > w \) if \( v \geq w \) and \( v \neq w \).

First we will derive some lemmas.

**Lemma 5.1.** If \( L_f v = v + g_f e + \delta \) then

\[
L^n_f v = v + n g_f e + \sum_{t=0}^{n-1} (P_f^t - P^*_f) \delta .
\]

**Proof.** We can rewrite \( L^n_f v \) in the form

\[
L^n_f v = v + \sum_{t=1}^{n} (L^t_f v - L^{t-1}_f v) .
\]

Or with lemma 2.2(iii)

\[
L^n_f v = v + \sum_{t=1}^{n} P^{t-1}_f (L^t_f v - v) = v + \sum_{t=1}^{n} P^{t-1}_f (g_f e + \delta) = v + n g_f e + \sum_{t=1}^{n} P^{t-1}_f \delta .
\]

Also by lemma 2.2(i) we have \( P_f^* \delta = 0 \). Therefore

\[
L^n_f v = v + n g_f e + \sum_{t=0}^{n-1} (P_f^t - P_f^*) \delta .
\]

And from this lemma we get

**Lemma 5.2.** If \( L_f v = v + g_f e + \delta \) and \( L_h v = L_f v + \zeta \) while \( h(i) = f(i) \) for all \( i \in \text{Rec}(h) \) then

\[
L^n_h v - L^n_f v = \sum_{t=0}^{n-1} (P_h^t - P^*_h) \delta - \sum_{t=0}^{n-1} (P_f^t - P^*_f) \delta + \sum_{t=0}^{n-1} P_h^t \zeta .
\]
Proof. We have

\[ L^*_h v = v + g^*_f e + \delta + \zeta, \]

and since \( f = h \) on \( \text{Rec}(h) \) also \( P^*_f = P^*_h \) and hence \( P^*_h \delta = 0 \).

The approach of the proof of lemma 5.1 now yields

\[
L^n_h v = v + ng_f e + \sum_{t=0}^{n-1} P^*_h (\delta + \zeta)
= v + ng_f e + \sum_{t=0}^{n-1} (P^*_h - P^*_h \delta) + \sum_{t=0}^{n-1} P^*_h \zeta.
\]

If we subtract equation (5.1) from this result we get (5.2), which completes the proof.

The reason why we are interested in the difference \( L^n_h v - L^n_f v \) becomes clear from the following lemma.

Lemma 5.3. Let \( f \) and \( h \) be two policies with \( f = h \) on \( \text{Rec}(h) \) then \( g^*_f = g^*_h \) and we have for any \( v \in \mathbb{R}^N \)

\[
\lim_{n \to \infty} (L^n_f v - L^n_h v) = v_f - v_h.
\]

Proof. From \( f = h \) on \( \text{Rec}(h) \) we have \( g^*_f = g^*_h \) and \( P^*_f = P^*_h \). So we get with lemma 2.3.

\[
L^n_f v - L^n_h v = ng_f e + v_f + P^*_f v - (ng_h e + v_h + P^*_h v) + o(1)
= v_f - v_h + o(1) \quad (n \to \infty).
\]

From the lemmas 5.2 and 5.3 we get the following important corollary.

Corollary 5.1. Suppose \( sp(L^n_f v - v) \leq \varepsilon \) and \( h \) is a policy with \( h = f \) on \( \text{Rec}(h) \) and \( L^*_h v = L_f v + \zeta \) for some \( \zeta \geq 0 \).

If for component \( i \) we have \( \zeta(i) > 2N\varepsilon/(1 - \rho) \), then

\[
v_h(i) > v_f(i).
\]
Proof. From $\text{sp}(L_f^v - v) \leq \epsilon$ we have the existence of a vector $\delta$, with

$$|\delta| \leq \epsilon e,$$

such that $L_f^v = v + g_f^e + \delta$.

Now consider (5.2). By lemma 2.1 we have

$$|\sum_{t=0}^{n-1} (P_h^t - P_h^*)\delta| \leq \sum_{t=0}^{n-1} |P_h^t - P_h^*| \epsilon e \leq N\epsilon(1 - \rho^n)/(1 - \rho) \leq N\epsilon/(1 - \rho).$$

Similarly with $h$ replaced by $f$. So we get for all $n$

$$L_f^n v - L_h^n v \geq -2N\epsilon/(1 - \rho)e + \sum_{t=0}^{n-1} P_h^t \zeta - 2N\epsilon/(1 - \rho) e + \zeta.$$

So, if for component $i$ we have $\zeta(i) > 2N\epsilon/(1 - \rho)$ then

$$\lim_{n \to \infty} (L_f^n v - L_h^n v)(i) > 0.$$

Hence by lemma 5.3 $v_h(i) > v_f(i)$.

We now have that if $h$ is a case B type improvement of $f$ then $g_f^e = g_h$ and, as $f$ and $h$ are equal on $\text{Rec}(h)$, also $v_f = v_h$ on $\text{Rec}(h)$. From corollary 5.1 we know that if $h(i) \neq f(i)$ (then also $\zeta(i) > \alpha$) and if $\alpha/e$ sufficiently large then $v_h(i) > v_f(i)$. What remains to be shown (in order to establish $v_h \succ v_f$) is that also $v_h(i) \geq v_f(i)$ in all transient states under $P_h$ in which $h(i) = f(i)$.

**Theorem 5.1.** If $\text{sp}(L_f^v - v) \leq \epsilon$ and $h$ is a policy with $h = f$ on $\text{Rec}(h)$ and $L_h^v = L_f^v + \zeta$ with $\zeta \geq 0$, such that, for at least one $i$ we have $\zeta(i) > \alpha$ and for all $i \in S$, either $h(i) = f(i)$ or $\zeta(i) > \alpha$, with $\alpha > 2N\epsilon/(1 - \rho)$, then $v_h \succ v_f$.

**Proof.** We already have $v_h = v_f$ on $\text{Rec}(h)$ and $v_h(i) > v_f(i)$ if $h(i) \neq f(i)$. Let $I$ be the set of states where $h(i) = f(i)$.

Now consider (1.1) for $f$ and $h$. 

\[ r_f + P_f v_f = v_f + g_f e \]
\[ r_h + P_h v_h = v_h + g_h e . \]

Subtracting the first of these two equations from the second we get for a state \( i \in I \) (as \( r_f(i) = r_h(i) \), \( P_{ij}^f = P_{ij}^h \) and \( g_f = g_h \))

\[ \begin{equation}
\sum_{j} P_{ij}^h (v_h - v_f)(j) = (v_h - v_f)(i)
\end{equation} \]

Suppose that \( \min_{i \in I} (v_h - v_f)(i) < 0 \) and define \( J := \{ j \in I | (v_h - v_f)(j) = \min_{i \in I} (v_h - v_f)(i) \} \).

On \( S \setminus I \) we have \( v_h \geq v_f \) hence we see from (5.3) that the set \( J \) must be closed under \( P_h \). But this contradicts the fact that \( v_h = v_f \) on \( \text{Rec}(h) \). Hence \( v_h \geq v_f \) on \( J \), thus on \( S \). And as there is at least one \( i \in S \) for which \( v_h(i) > v_f(i) \) we get \( v_h > v_f \).

6. Concluding remarks

As we have seen in the preceding sections the iteration step formulated in section 1 yields an improved policy: either a higher gain (section 4) or with equal gains a higher relative value (section 5). Provided, of course, that \( \varepsilon \) is sufficiently small compared to \( \alpha \). There are two ways in which we can try to construct a convergent algorithm from this iteration step.

(i) Select \( \varepsilon > 0 \) sufficiently small. In order to do this we will have to know the constants \( b, \rho \) and \( \theta \) appearing in lemma 2.1 and formula (4.1), which makes this approach practically infeasible (we would need \( \varepsilon \leq \min(\alpha \theta, \alpha(1 - \rho)/2Nb)) \).

(ii) Select a sequence \( \varepsilon_1, \varepsilon_2, \ldots \) with \( \varepsilon_n > 0 \) for all \( n \) and \( \varepsilon_n \to 0 \).
And use in the \( n \)-th iteration step \( \varepsilon_n \). Then if \( n \) is sufficiently large \( \alpha/\varepsilon_n \) is large enough to assure that each policy replacement is indeed an improvement. So this algorithm will converge.
Another remark we have to make concerns the following. In the policy improvement step we demand that actions are replaced only if the effect of this improvement is more than \( \alpha \). In the case that there are no transient states we do not need this restriction. Then, we only have improvements of type A for which the proof of section 4 goes on. This is the case we treated in [4]. In the case of transient states however we see that small improvements in recurrent states may, because of the inaccuracy of the value approximation step, yield a policy with the same gain. But in that case we may no longer have \( P_f^* = P_h^* \) so the essential lemma 5.3 may fail to hold.

7. References


