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

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

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 $f : S \to A$. With a policy f we associate the immediate reward vector r_f and the transition matrix P_f defined by

$$r_f(i) := r(i,f(i)) , \qquad i \in S .$$

$$P_f(i,j) := p_{ij}^{f(i)} , \qquad i,j \in S .$$

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

$$(1.1) \qquad r_f + P_f v = v + g_f e , \qquad (e = (1,\ldots,1)^T)$$

$$(1.2) \qquad P_f^* v = 0 ,$$

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

$$(1.3) \qquad P_f^* = \lim_{n\to\infty} P_f^n .$$

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 ε-optimal policy, i.e. a policy f satisfying

$$g_f \geq g^* - \varepsilon .$$

In order to determine such an ε-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 α and ε occurring in this iteration step are chosen beforehand)

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

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

$$(1.4) \qquad sp(v_t - v_{t-1}) \leq \varepsilon$$

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

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

Policy improvement. Let n be the first index for which (1.4) holds. Then define γ by

$$(1.5) \qquad \gamma := \max_h \{r_h + P_h v_n\} - r_f - P_f v_n .$$

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

(i) $\gamma \leq \alpha e$. Then f is nearly optimal.

$$(1.6) \qquad g_f \geq g^* - \alpha - \varepsilon$$ (as we will prove later on).

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

$h(i) = f(i)$ if $\gamma(i) \leq \alpha$, and

$h(i) =$ a maximizer of $r(i,a) + \sum_j p_{ij}^a v_n(j)$ 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

(1.7) $$r_h + P_h v_n = r_f + P_f v_n + \zeta$$

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 ε is sufficiently small compared to α 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 α/ε is sufficiently large, $g_h > g_f$, 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 $g_h = g_f$, $v_h \geq v_f$ and if $\zeta(i) > \alpha$ then $v_h(i) > v_f(i)$, provided that α/ε sufficiently large. This we prove in section 5.

In section 6 we collect some concluding remarks.

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_f^n to P_f^* formulated in (1.3):

Lemma 2.1. There exist constants b and $0 \leq \rho < 1$ such that

$$|P_f^*(i,j) - P_f^n(i,j)| \leq b\rho^n$$

for all $i,j \in S$, f and $n \in \mathbb{N}$.

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

algorithm.

Lemma 2.2. For any $v \in \mathbb{R}^N$

(i) $\quad P_f^*(L_f v - v) = g_f e$

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

(iii) $\quad L_f^{n+1} v - L_f^n v = P_f^n (L_f v - v)$

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

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

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

(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. $L_f^n v - ng_f e - v_f \to P_f^* v$ as $n \to \infty$,

so the difference between $L_f^n v$ and v_f becomes constant.

Proof. Clearly $L_f^n v - L_f^n w = P_f^n (v - w)$ and as v_f satisfies (1.1)

$$L_f^n v_f = v_f + ng_f e \ .$$

So

$$L_f^n v - ng_f e - v_f = L_f^n v - L_f^n v_f = P_f^n (v - v_f) \to P_f^* (v - v_f) = P_f^* v$$

as $n \to \infty$, where we used that v_f satisfies (1.2). □

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

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

$$g_h e \leq (g_f + \alpha + \varepsilon) e \ .$$ □

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 $sp(v_n - v_{n-1}) \leq \varepsilon$ implies $sp(L_f v_n - L_f v_{n-1}) \leq \varepsilon$. □

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 ε is sufficiently small compared to α.
First define θ by

$$\theta := \min_{i,j,f} \{P_f^*(i,j) \mid P_f^*(i,j) > 0\} . \tag{4.1}$$

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 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 ϵ/α is sufficiently small then a 'case B type' improvement gives a policy h with a higher relative value vector $v_h \succ v_f$. Where we write $v \succ w$ 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

$$(5.1) \qquad L_f^n v = v + n g_f e + \sum_{t=0}^{n-1} (P_f^t - P_f^*)\delta \quad .$$

Proof. We can rewrite $L_f^n v$ in the form

$$L_f^n v = v + \sum_{t=1}^{n} (L_f^t v - L_f^{t-1} v) \quad .$$

Or with lemma 2.2(iii)

$$L_f^n v = v + \sum_{t=1}^{n} P_f^{t-1}(L_f v - v) = v + \sum_{t=1}^{n} P_f^{t-1}(g_f e + \delta) = v + n g_f e + \sum_{t=1}^{n} P_f^{t-1}\delta \quad .$$

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

$$L_f^n v = v + n g_f e + \sum_{t=0}^{n-1} (P_f^t - P_f^*)\delta \quad . \qquad \square$$

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 $i \in$ Rec (h) then

$$(5.2) \qquad L_h^n v - L_f^n v = \sum_{t=0}^{n-1} (P_h^t - P_h^*)\delta - \sum_{t=0}^{n-1} (P_f^t - P_f^*)\delta + \sum_{t=0}^{n-1} P_h^t \zeta \quad .$$

Proof. We have

$$L_h v = v + g_f e + \delta + \zeta ,$$

and since f = h on Rec(h) also $P_f^* = P_h^*$ and hence $P_h^* \delta = 0$.
The approach of the proof of lemma 5.1 now yields

$$L_h^n v = v + n g_f e + \sum_{t=0}^{n-1} P_h^t (\delta + \zeta)$$

$$= v + n g_f e + \sum_{t=0}^{n-1} (P_h^t - P_h^*)\delta + \sum_{t=0}^{n-1} P_h^t \zeta .$$

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 $L_h^n v - L_f^n v$ becomes clear from the following lemma.

Lemma 5.3. Let f and h be two policies with f = h on Rec(h) then $g_f = g_h$ and we have for any $v \in \mathbb{R}^N$

$$\lim_{n\to\infty} (L_f^n v - L_h^n v) = v_f - v_h .$$

Proof. From f = h on Rec(h) we have $g_f = g_h$ and $P_f^* = P_h^*$. So we get with lemma 2.3.

$$L_f^n v - L_h^n v = n g_f e + v_f + P_f^* v - (n g_h e + v_h + P_h^* v) + o(1)$$

$$= v_f - v_h + o(1) \qquad (n \to \infty) .$$ □

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

Corollary 5.1. Suppose $sp(L_f v - v) \le \varepsilon$ and h is a policy with h = f on Rec(h) and $L_h v = L_f v + \zeta$ for some $\zeta \ge 0$.
If for component i we have $\zeta(i) > 2Nb\varepsilon/(1 - \rho)$, then

$$v_h(i) > v_f(i) .$$

Proof. From $sp(L_f v - v) \leq \varepsilon$ we have the existence of a vector δ, with $|\delta| \leq \varepsilon e$, such that $L_f v = v + g_f e + \delta$.
Now consider (5.2). By lemma 2.1 we have

$$\left| \sum_{t=0}^{n-1} (P_h^t - P_h^*)\delta \right| \leq \sum_{t=0}^{n-1} |P_h^t - P_h^*| \varepsilon e \leq Nb\varepsilon(1 - \rho^n)/(1 - \rho)$$

$$\leq Nb\varepsilon/(1 - \rho) \ .$$

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

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

$$-2Nb\varepsilon/(1 - \rho)e + \zeta \ .$$

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

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

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 $g_f = g_h$ and, as f and h are equal on Rec(h), also $v_f = v_h$ on Rec(h). From corollary 5.1 we know that if $h(i) \neq f(i)$ (then also $\zeta(i) > \alpha$) and if α/ε 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 $sp(L_f v - v) \leq \varepsilon$ and h is a policy with $h = f$ on 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 $\alpha > 2Nb\varepsilon/(1 - \rho)$, then $v_h \succ v_f$.

Proof. We already have $v_h = v_f$ on 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 $r_f(i) = r_h(i)$, $p_{ij}^{f(i)} = p_{ij}^{h(i)}$ and $g_f = g_h$)

$$\sum p_{ij}^{h(i)} (v_h - v_f)(j) = (v_h - v_f)(i) \tag{5.3}$$

Suppose that $\min_{i \in I} (v_h - v_f)(i) < 0$ and define

$$J := \{ j \in I \mid (v_h - v_f)(j) = \min_{i \in I} (v_h - v_f)(i) \} \ .$$

On $S \backslash I$ 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 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 $v_h \succ v_f$. □

6. Concluding remarks

As we have seen in the preceding sections the iterationstep 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 ε is sufficiently small compared to α. 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, ρ and θ 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 $\varepsilon_n \downarrow 0$. And use in the n-th iterationstep ε_n. Then if n if sufficiently large α/ε_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 α. 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

[1] Doob, J.L., Stochastic Processes, Wiley, New York, 1953.

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

[3] Schweitzer, P.J., Iterative solution of the functional equations of undiscounted Markov renewal programming. J. Math. Anal. Appl. 34 (1971), 495-501.

[4] Van der Wal, J., A successive approximation algorithm for an undiscounted Markov decision process, Computing 17 (1976), 157-162.