Strongly convergent dynamic programming : some results

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 76-26

Strongly convergent dynamic programming: some results

by

K.M. van Hee and J. van der Wal

Eindhoven, December 1976 (Revised February 1977)

The Netherlands

Strongly convergent dynamic programming: some results

by

K.M. van Hee and J. van der Wal

1. Introduction

In this paper we consider Markov decision processes with respect to the total expected reward criterion. We work under a convergence condition which guarantees that the total expected reward from time n onwards, tends to zero uniformly in the strategy. This condition is weaker than the contraction conditons considered by Wessels (1974) and Van Nunen (1976) which are extensions of the discounted model studied by Blackwell (1965). A nice feature of our condition is the fact that convergence of the method of successive approximations can be shown by elementary calculus, see Van Hee, Hordijk and Van der Wal (1977). Here we concentrate the attention on Howard's policy iteration method and the existence of nearly optimal stationary strategies. Although our results are partially known they seem to be unpublished. Before we formulate our condition in detail, we first sketch the framework of dynamic programming, using notations of Hordijk (1974). Consider a countable set S, the *state space* and an arbitrary set A, endowed with a σ-field containing all one-point sets, the *action space*. There is a transition probability Q from $S \times A$ to S, and a *reward function* r from $S \times A$ to $\mathbb{R}$ such that $r(i,\cdot)$ is measurable for all $i \in S$ and if $Q(\cdot|i,a_1) = Q(\cdot|i,a_2)$ then $r(i,a_1) = r(i,a_2)$, $a_1,a_2 \in A$, $i \in S$. With Q one can compose the set $\mathcal{P}$ of all transition probabilities P from S to S such that, for any $i \in S$ $P(\cdot|i) = Q(\cdot|i,a)$ for some $a \in A$. A (Markov) *strategy* R may now be defined as a sequence $P_0,P_1,P_2,..$ with $P_n \in \mathcal{P}, n = 0,1,2,\dots$. Each $i \in S$ and R determine a probability $\mathbb{P}_{i,R}$ on $(S \times A)^\infty$ and a stochastic proces $\{(X_n,A_n)\ ,\ n = 0,1,2,\dots\}$ where X_n is the state and A_n is the action at time n. (The expectation with respect to $\mathbb{P}_{i,R}$ is denoted by $\mathbb{E}_{i,R}$ and if we omit the subscript i in $\mathbb{E}_{i,R}$ we mean the function on S.)
Throughout this paper we assume

$$\sup_R \mathbb{E}_{i,R}[\sum_{n=0}^{\infty} r^+(X_n,A_n)] < \infty \text{ for all } i \in S$$

(note that $x^+ := \max(0,x)$)

As shown in Van Hee (1975) this assumption guarantees that the restriction to pure Markov strategies gives no loss of generality.
On S we define the following functions:

i) $v := \sup_R \mathbb{E}_R[\sum_{n=0}^{\infty} r(X_n,A_n)]$, the criterion function.

for a function $s : S \to \mathbb{R}$ with $\sup_R \mathbb{E}_R [s^+(X_N)] < \infty$:

ii) $v_N^s := \sup_R \mathbb{E}_R [\sum_{n=0}^{N-1} r(X_n,A_n) + s(X_N)]$

for a sequence $a := (a_0,a_1,a_2,\ldots)$ of functions $a_n : S \to [1,\infty), n = 0,1,2,\ldots$:

iii) $w_a(i) := \sup_R \sum_{n=0}^{\infty} a_n(i) \mid \mathbb{E}_{i,R} r(X_n,A_n) \mid , \quad i \in S$

iv) $z_a(i) := \sup_R \sum_{n=0}^{\infty} a_n(i) \mathbb{E}_{i,R} |r(X_n,A_n)|, \; i \in S$

v) $w := w_a, \; z := z_a$ if $a_n \equiv 1$ for $n = 0,1,2,\ldots$

The conditions we are working with in this paper, state the existence of a sequence $a = (a_0,a_1,a_2,\ldots)$ of functions $a_n : S \to [1,\infty)$ with $a_n \to \infty$ (pointwise) while still $w_a < \infty$ or even $z_a < \infty$ holds.
We suggest to use the term 'strongly convergent' for models satisfying the weaker ($w_a < \infty$) condition.
We conclude this section with some notational conventions. It is easy to see that for $P \in \mathcal{P}$ there is a function $f_p : S \to A$ such that $P(j|i) = Q(j|i,f_p(i))$, $i, j \in S$ and we sometimes write $r_p(i) := r(i,f_p(i))$. For $R = (P_0,P_1,P_2,\ldots)$ we easily obtain $\mathbb{E}_{i,R}[r(X_n,A_n)] = P_0 \ldots P_{n-1} r_{P_n}(i)$. An empty product of elements of $\mathcal{P}$ is defined as the identity operator.
For two (extended) real functions a and b on S we write $\frac{a}{b}$ for the (extended) real function c defined by $c(i) := \frac{a(i)}{b(i)}$ if $b(i) \neq 0$. With convergence of a sequence of functions on S we mean pointwise convergence; the supremum of a sequence of functions is the pointwise supremum.

2. Standard successive approximations

In this section we present some inequalities which imply, for strongly convergent models, the convergence of the method of successive approximations. Further we give a sufficient condition for a Markov decision process to be strongly convergent. For proofs, not given here, we refer to Van Hee, Hordijk and Van der Wal (1977). In this section we assume that $a = (a_0, a_1, a_2, \dots)$ is a nondecreasing sequence of functions $a_n : S \to [1,\infty)$

Theorem 1.

The following holds:

$$\sup_R \sum_{k=n}^{\infty} | \mathbb{E}_R r(X_k, A_k) | \le \frac{w_a}{a_n} \quad \text{and} \quad \sup_R \sum_{k=n}^{\infty} \mathbb{E}_R | r(X_k, A_k) | \le \frac{z_a}{a_n} .$$

Proof:

$$\sup_R \sum_{k=n}^{\infty} | \mathbb{E}_{iR} r(X_k, A_k) | \le \frac{1}{a_n(i)} \sup_R \sum_{k=n}^{\infty} a_k(i) | \mathbb{E}_{i,R} r(X_k, A_k) | \le$$

$$\le \frac{w_a(i)}{a_n(i)} .$$

The proof of the second inequality is identical. □

Corollary 1.

$$\sup_R |\mathbb{E}_R v(X_n)| \le \frac{w_a}{a_n}, \text{ since}$$

$$\sup_R |\mathbb{E}_R v(X_n)| \le \sup_R \sum_{k=n}^{\infty} | \mathbb{E}_R r(X_k, A_k)| .$$

Another direct consequence of theorem 1 is the following.

Theorem 2.

Let $s : S \to \mathbb{R}$ be such that $\mathbb{E}_R s^+(X_n) < \infty$ for all R then:

$$|v_n^s - v| \le \frac{w_a}{a_n} + \sup_R | \mathbb{E}_R s(X_n) | .$$

Hence if $a_n \to \infty$ and $w_a < \infty$ the method of successive approximations converges to the value function v for any scrapfunction s satisfying $\sup_R \mathbb{E}_R s(X_n) \to 0$. The bound given in theorem 2 is rather rough, which becomes clear if we set s equal to v and note that $v_n^v = v$ for $n = 0,1,2,\dots$

In corollary 2 we give sufficient conditions for scrapfunctions to guarantee convergence:

Corollary 2.

Let $a_n \to \infty$ and $z_a < \infty$. If the real valued function s satisfies $|s| \le k\, z$ for some $k \in \mathbb{R}$ we have

$$|v_n^s - v| \le \frac{w_a + kz_a}{a_n} .$$

It follows from theorem 1 that the existence of a sequence a with $a_n \to \infty$ and $w_a < \infty$ implies that

$$\lim_{n\to\infty} \sup_R \sum_{k=n}^{\infty} |\mathbb{E}_R\, r(X_k,A_k)| = 0 .$$

The following theorem states that this limit property almost implies the existence of such a sequence $a = (a_0,a_1,a_2\ldots)$.

Theorem 3.

Let $w < \infty$ and $\lim_{n\to\infty} \sup_R \sum_{k=n}^{\infty} |\mathbb{E}_R\, r(X_k,A_k)| = 0$, then there is a nondecreasing sequence of functions $a_n : S \to [1,\infty)$ such that: $a_n \to \infty$ and $w_a < \infty$.

Finally we remark that our restriction to a countable state space is not essential; it seems that these results carry over to the general case without any difficulty.

3. The policy iteration method

In this section we assume the existence of a nondecreasing sequence of functions $a_n : S \to [1,\infty)$ such that $w_a < \infty$. In section 2 we have seen that in this situation the method of successive approximations converges and now we show that the same holds for the policy iteration method given by Howard (1960). In fact the convergence of both methods is wellknown for the contracting dynamic programming model. The proofs given here are quite simple and use the same ideas as in the contracting case.

We first introduce Howard's iteration method.

Let for $P \in \mathcal{P}$ $R_P := (P,P,P,\ldots)$.

3.1. i) choose $P_0 \in \mathcal{P}$ and define $v_0 := \mathbb{E}_{R_{P_0}} [\sum_{n=0}^{\infty} r(X_n,A_n)]$, choose a sequence $\varepsilon_1,\varepsilon_2,\ldots$ such that $\varepsilon_n \downarrow 0$.

ii) Determine $P_n \in \mathcal{P}$ such that

$$r_{P_n} + P_n v_{n-1} \geq \max\{\sup_P [r_P + Pv_{n-1} - \varepsilon_n e] \, , \, v_{n-1}\}$$

and define

$$v_n := \mathbb{E}_{R_{P_n}} [\sum_{n=0}^{\infty} r(X_n, A_n)]$$

(e is the unit function on S).

In the remainder of this section we show that v_n converges monotonically to the criterion function v. First we prove two lemma's.

<u>Lemma 1</u>.

$$v_n \geq v_{n-1} \quad , \quad n = 1,2,3,\ldots$$

<u>Proof</u>:

From 3.1. ii) we have $r_{P_n} + P_n v_{n-1} \geq v_{n-1}$. Iterating this equation k times yields

$$\sum_{\ell=0}^{k} P_n^{\ell} r_{P_n} + P_n^{k+1} v_{n-1} \geq v_{n-1} \; .$$

Since $\sum_{\ell=0}^{k} P_n^{\ell} r_{P_n}$ converges to v_n and $|P_n^{k+1} v_{n-1}| \leq \frac{w_a}{a_{k+1}}$ we get $v_n \geq v_{n-1}$. □

Obviously $v_n \leq v$. Defining $\hat{v} := \lim_{n\to\infty} v_n$ we get $\hat{v} \leq v$.

<u>Lemma 2</u>.

$$\sup_P \{r_P + P\hat{v}\} \leq \hat{v} \; .$$

<u>Proof</u>:

$r_{P_n} + P_n v_n = v_n$ and so, by lemma 1, we have $v_n \geq r_{P_n} + P_n v_{n-1}$. Hence

$v_n \geq r_P + Pv_{n-1} - \varepsilon_n e$ for all $P \in \mathcal{P}$. Using the monotone convergence theorem we derive $\hat{v} \geq r_P + P\hat{v}$ for all $P \in \mathcal{P}$. □

Now we are ready to prove $\hat{v} = v$.

Theorem 5.

$$\hat{v} = v .$$

Proof:

Since $\hat{v} \le v$ it suffices to show $\hat{v} \ge v$. Let $R = (P_0, P_1, P_2, \ldots)$ be an arbitrary strategy. Then, by lemma 2, we get

$$\hat{v} \ge r_{P_0} + P_0 \hat{v} \ge \ldots \ge \sum_{k=0}^{n} P_0 \ldots P_{k-1} r_{P_k} + P_0 \ldots P_n \hat{v} .$$

Since, $\hat{v} \ge v_0$ and $|P_0 \ldots P_n v_0| \le \frac{w_a}{a_n}$ (by theorem 1)

we have

$$\hat{v} \ge \mathbb{E}_R[\sum_{n=0}^{\infty} r(X_n, A_n)] .$$

Since this holds for all R the theorem is proved. □

4. Nearly optimal stationary strategies

In this section we again assume the existence of a nondecreasing sequence of functions $a = (a_0, a_1, a_2, \ldots)$, $a_n : S \to [1,\infty)$ such that $a_n \to \infty$ and $w_a < \infty$. It follows from theorem 5 that there is for each finite subset $S_0 \subset S$ and for all $\varepsilon > 0$ a stationary strategy $R = (P,P,P,\ldots)$ such that

$$v_R(i) := \mathbb{E}_{i,R} [\sum_{n=0}^{\infty} r(X_n, A_n)] \ge v(i) - \varepsilon \text{ for } i \in S_0 .$$

We show in this section under some additional assumptions the existence of everywhere nearly optimal stationary strategies.

Theorem 6.

If $\frac{w_a}{a_n} \to 0$ uniformly on S, then there exists for any $\varepsilon > 0$ a stationary strategy R such that

$$v_R \ge v - \varepsilon e .$$

Proof:

Choose $\varepsilon > 0$, N such that $\frac{w_a}{a_N} \le \frac{\varepsilon}{3} e$ and P such that $v \le r_P + Pv + \frac{\varepsilon}{3N} e$.
Then iterating this inequality N times and using theorem 1 one easily shows $v \le v_R + \varepsilon e$. □

Under a weaker additional assumption we have a weaker sense of ε-optimality.

Theorem 7.

Let $a_n \to \infty$ uniformly and $z_a < \infty$ then there exists for any $\varepsilon > 0$ a stationary strategy R such that $v_R \ge v - \varepsilon z_a$.

Proof:

Choose $\varepsilon > 0$, N such that $a_N \ge \frac{3}{\varepsilon}$ and P such that

$$r_P + Pv \ge v - \varepsilon \{ \sum_{n=0}^{N-1} a_n^{-1} \}^{-1} z \ .$$

Iterating this inequality N times yields:

$$\sum_{n=0}^{N-1} P^n r_P + P^N v + \varepsilon \{ \sum_{n=0}^{N-1} a_n^{-1} \}^{-1} \sum_{n=0}^{N-1} P^n z \ge v \ .$$

Since $P^n z \le \frac{z_a}{a_n}$, $| \sum_{n=N}^{\infty} P^n r_P | \le \frac{\varepsilon}{3} z_a$ and $P^n v \le \frac{\varepsilon}{3} z_a$ we get for $R = (P,P,P,\dots)$

$$v - v_R \le \varepsilon z_a \ .$$ □

References

Blackwell, D., (1965) Discounted Dynamic Programming, Ann. Math. Statist. 36, 226-235.

Van Hee, K.M., (1975) Markov strategies in dynamic programming, Eindhoven, University of Technology (Dept. of Math.) Memorandum COSOR 75-20. Submitted for publication.

Van Hee, K.M., A. Hordijk, J. van der Wal. (1977) To appear in the proceedings of the advanced seminar on Markov decision theory in Amsterdam, in the series of Mathematical Centre Tracts.

Hordijk, A. (1974) Dynamic programming and Markov potential theory. Amsterdam, Mathematical Centre Tracts, no. 51.

Howard, R.A. (1960) Dynamic programming and Markov processes. Cambridge (Mass.) M.I.T. Press.

Van Nunen, J.A.E.E. (1976) Contracting Markov decision processes, Amsterdam, Mathematical Centre Tracts, no. 71.

Wessels, J. (1974) Markov programming by successive approximations with respect to weighted suppremum norms to appear in: Journ. of Math. Anal. and Appl.