Successive approximations for the average Markov reward game : the communicating case

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

PROBABILITY THEORY, STATISTICS AND OPERATIONS RESEARCH GROUP

Memorandum COSOR 81-03

Successive approximations for the average Markov game; the communicating case.

by

J. van der Wal

Eindhoven, January 1981
The Netherlands

SUCCESSIVE APPROXIMATIONS FOR THE AVERAGE REWARD MARKOV GAME; THE COMMUNICATING CASE

J. van der Wal
Department of Mathematics
Eindhoven University of Technology
Eindhoven, The Netherlands

This paper considers the two-person zero-sum Markov game with finite state and action spaces at the criterion of average reward per unit time. For two types of Markov games, the communicating game and the simply connected game, it is shown that the method of successive approximations provides good bounds on the value of the game and nearly-optimal stationary strategies for the two players.

INTRODUCTION

This paper deals with two-person zero-sum average reward Markov games with finite state and action spaces. Recently Monash [6] and Mertens and Neyman [5] have shown that these games always have a value, though not necessarily within the class of stationary strategies nor within the class of Markov strategies (cf. Gillette [4] and Blackwell and Ferguson [2]).
Federgruen [3] showed that, if the underlying Markov games corresponding to a pair of (pure) stationary strategies all have the same number of irreducible subchains, then the game has a value within the class of stationary strategies.
In Van der Wal [9,11] it is shown that in the unichain case one can obtain good bounds on the value of the game and nearly-optimal stationary strategies for the two players by the method of standard successive approximations. Here we want to extend the results in [9,11] to the communicating case (cf. Bather [1]) and the simply connected case (cf. Platzman [7]).

So we consider a dynamical system with finite state space $S := \{1,2,\dots,N\}$ and finite action spaces A and B for players $1(P_1)$ and $2(P_2)$ respectively. The system is observed at equidistant points in time, $t = 0,1,\dots$ say, and controlled by the two players. At each time t, having seen the state of the system, they choose an action. As a joint result of the actions a by P_1 and b by P_2 in state i, P_1 receives a (possibly negative) reward $r(i,a,b)$ from P_2 and the system moves to state j with probability $p(i,a,b,j)$, $\sum_j p(i,a,b,j) = 1$.

In general a strategy π for P_1 is any sequence $(f_0,f_1,\dots)$ of mappings $f_n : S \times A \to [0,1]$ with $\sum_a f_n(i,a) = 1$ for all $i \in S$. The functions f_n are called policies and $f_n(i,a)$ denotes the probability that action a is taken if the system is observed in state i at time n. A strategy is called stationary if $f_n = f_0$ for all $n \geq 1$. Notation $f_0^{(\infty)}$ or simply f_0. Similarly we define strategies $\gamma = (h_0, h_1, \ldots)$ for P_2.

It is sufficient to consider only Markov strategies since in the cases treated here the game has a value within the class of Markov strategies and as one may show a (nearly-) optimal Markov strategy in the game with Markov strategies only is also (nearly-) optimal in the game with arbitrary strategies.

For any two policies f and h define the vector r(f,h) and the matrix P(f,h) by

$$r(f,h)(i) := \sum_a \sum_b f(i,a)h(i,b)r(i,a,b) \quad , \quad i \in S$$

$$P(f,h)(i,j) := \sum_a \sum_b f(i,a)h(i,b)p(i,a,b,j), \ i,j \in S.$$

Further define the operators $L(f,h)$ and U on $\mathbb{R}^N$ by

$$L(f,h)v = r(f,h) + P(f,h)v$$

$$Uv \quad = \max_f \min_h L(f,h)\ v.$$

Then the average reward per unit time vector corresponding to a pair of strategies $\pi = (f_0, f_1, \ldots)$, $\gamma = (h_0, h_1, \ldots)$ is defined by

$$g(\pi,\gamma) = \liminf_{n \to \infty} \ n^{-1} L(f_0,h_0) \ldots\ldots L(f_{n-1},h_{n-1})0 .$$

The game is said to have the value g^* if

$$\sup_\pi \inf_\gamma \ g(\pi,\gamma) = \inf_\gamma \sup_\pi \ g(\pi,\gamma) = g^*.$$

A strategy $\tilde{\pi}$ is called ε-optimal for P_1 if $\inf_\gamma g(\tilde{\pi},\gamma) \geq g^* - \varepsilon e$, $e^T = (1,1,\ldots,1)$. Similarly $\tilde{\gamma}$ is ε-optimal for P_2 if

$$\sup_\pi g(\pi,\tilde{\gamma}) \leq g^* + \varepsilon e.$$

In order to approximate g^* and to find nearly-optimal stationary strategies for the two players we use the method of standard successive approximations (sa)

$$(1) \qquad \begin{aligned} & v_0 := 0 \\ & v_{n+1} := Uv_n, \quad n = 0,1,\ldots \ . \end{aligned}$$

As will be shown for communicating and for simply connected games the value g^* is independent of the initial state. Therefore the following lemma is of considerable interest.

Lemma 1

Let $v \in \mathbb{R}^N$ be arbitrary and let $\hat{f}$ and $\hat{h}$ satisfy $L(f,\hat{h})v \leq Uv \leq L(\hat{f},h)v$ for all f and h, then

(i) $\quad g(\hat{f},\gamma) \geq \min\limits_{i \in S} (Uv - v)(i).e \quad$ for all γ

(ii) $\quad g(\pi,\hat{h}) \leq \max\limits_{i \in S} (Uv - v)(i).e \quad$ for all π

(iii) $\quad \min\limits_{i \in S} (Uv - v)(i).e \leq g^* \leq \max\limits_{i \in S} (Uv - v)(i).e$.

Proof

The proof is rather straightforward, see e.g [9].

From this lemma we see that if $v_{n+1} - v_n$ converges to a constant vector then g^* is state independent and the method of sa yields good bounds on g^* and nearly-optimal stationary strategies for the two players.

In order to avoid period behaviour of $v_{n+1} - v_n$ the following assumption is made.

Strong aperiodicity assumption (SAA)

For some constant $\alpha > 0$ and for all $i \in S$, $a \in A$, $b \in B$

$$p(i,a,b,i) \geq \alpha .$$

This is no serious restriction. Any Markov game can be transformed into an equivalent game satisfying SAA by means of a data transformation due to Schweitzer [8] (cf. [9]). Now let us define the communicating and the simply connected game.

Definition 1 (cf. Bather [1])

A Markov game is called *communicating* if for any pair $i,j \in S$ each of the two players can force the system to reach state j from state i with positive probability in a finite number of steps whatever actions his opponent takes.

As a consequence of the SAA we have that if the Markov game is communicating then there exists some constant $\eta > 0$ and Markov strategies $\tilde{\pi} = (\tilde{f}_0,\ldots,\tilde{f}_{N-2})$ and $\tilde{\gamma} = (\tilde{h}_0,\ldots,\tilde{h}_{N-1})$ such that for all $i,j \in S$ and all $\pi = (f_0,\ldots,f_{N-2})$ and $\gamma = (h_0,\ldots,h_{N-2})$

(2) $P(\tilde{f}_0,h_0)\dots P(\tilde{f}_{N-2},h_{N-2})\,(i,j) \ge \eta$ and

(3) $P(f_0,\tilde{h}_0)\dots P(f_{N-2},\tilde{h}_{N-2})\,(i,j) \ge \eta$.

(Recall that N is the number of states in S). This can be shown along similar lines as lemma 13.3 in [11].

A somewhat weaker condition is the condition of simply connectivity.

Definition 2 (cf. Platzman [7])

A Markov game is called *simply connected* if the state space S can be divided into two disjoint subjects $\hat{S}$ and $\bar{S}$ such that

(i) $p(i,a,b,j) = 0$ for all $i \in S$, $j \in S$, $a \in A$, $b \in B$

(ii) the game is communicating on $\hat{S}$

(iii) the game is transient on $\bar{S}$, i.e. there is some constant $\theta > 0$ such that for all $i \in \bar{S}$ and for all π,γ

(4) $\sum_{j \in \hat{S}} P(f_0,h_0)\dots P(f_{N-2},h_{N-2})\,(i,j) \ge \theta$.

In the sequel it will be shown that the conditions communicatingness and simply connectivity each guarantee that g* is constant and (together with the SAA) imply that $v_{n+1} - v_n \to g^*$ $(n \to \infty)$.

THE COMMUNICATING CASE

In this section it will be shown that the communicating condition together with the SAA guarantees that $v_{n+1} - v_n$ converges to the value function g* which is independent of the initial state.

Consider the sa scheme (1) and define for n = 0,1,...

$$g_n := v_{n+1} - v_n$$

$$\ell_n := \min_i g_n(i)$$

$$u_n := \max_i g_n(i) .$$

From

$$\min_i\, (v-w)(i).e \le Uv-Uw \le \max_i\, (v-w)(i).e \quad \text{for all } v,w \in \mathbb{R}^N$$

we immediately have the following lemma.

Lemma 2

$$\ell_n.e \leq \ell_{n+1}.e \leq g^* \leq u_{n+1}.e \leq u_n.e \quad \text{for all } n = 0,1,\dots\ .$$

So the sequences $\{\ell_n\}$ and $\{u_n\}$ are monotonically nondecreasing and nonincreasing respectively and bounded by g^*. Thus we can define

$$\ell^* := \lim_{n\to\infty} \ell_n \quad , \quad u^* := \lim_{n\to\infty} u_n\ .$$

Now our aim is to prove that $\ell^* = u^*$. To prove this we follow the line of reasoning in Van der Wal [10,11]. First it will be shown that $sp(v_n)$ is bounded,where the span of a vector v is defined by

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

Next this is used together with the SAA to prove $\ell^* = u^*$.
Let K be defined by

$$K := \max_{i,a,b} |r(i,a,b)|\ .$$

and let $i \in S$ and n be arbitrary. Then it follows from (2) that

$$v_{n+N-1}(i) = (U^{N-1}v_n)(i)$$

$$\text{(5)} \qquad \geq -(N-1)K + \max_{f_0,\dots,f_{N-2}} \min_{h_0,\dots,h_{N-2}} P(f_0,h_0)\dots P(f_{N-2},h_{N-2})v_n(i)$$

$$\geq -(N-1)K + \eta \max_j v_n(j) + (1-\eta) \min_j v_n(j).$$

And similarly if follows from (3) that for all $k \in S$

$$\text{(6)} \qquad v_{n+N-1}(K) \leq (N-1)K + \eta \min_j v_n(j) + (1-\eta) \max_j v_n(j).$$

So from (5) and (6)

$$sp(v_{n+N-1}) \leq 2(N-1)K + (1-2\eta)\ sp(v_n).$$

Thus for all $k = 0,1,\dots,N-3$ and all $\ell = 0,1,\dots.$

$$sp(v_{k+\ell(N-1)}) \leq \frac{1-(1-2\eta)^\ell}{1-(1-2\eta)}\ 2(N-1)K + (1-2\eta)^\ell\ sp(v_k)$$

$$\leq \eta^{-1}(N-1)K + sp(v_k)$$

Hence for all n

$$sp(v_n) \leq n^{-1}(N-1)K + \max_{k=0,..,N-3} sp(v_k),$$

so $sp(v_n)$ is bounded.

In order to prove $\ell^* = u^*$ we first have to derive some inequalities. Let $f_1,f_2...$ and $h_1,h_2,...$ be policies satisfying

$$L(f,h_{n+1})v_n \leq v_{n+1} \leq L(f_{n+1},h)v_n \quad \text{for all f and h.}$$

Then

$$v_{n+2} - v_{n+1} = L(f_{n+2},h_{n+2})v_{n+1} - L(f_{n+1},h_{n+1})v_n$$

$$\geq L(f_{n+1},h_{n+2})v_{n+1} - L(f_{n+1},h_{n+2})v_n = P(f_{n+1},h_{n+2})(v_{n+1}-v_n).$$

So for all $s,t = 0,1,...$

(7) $$v_{s+t+1} - v_{s+t} \geq P(f_{s+t},h_{s+t+1})...P(f_{s+1},h_{s+2})(v_{s+1} - v_s).$$

And for all $i \in S$

(8) $$g_{s+t}(i) \geq \alpha^t g_s(i) + (1-\alpha^t)\ell_s.$$

Now let us fix for the time being n and m and let state i satisfy $g_{n+m}(i) = \ell_{n+m}$. Then from (8), with $s = n+k$, $t = m-k$,

$$\ell_{n+m} = g_{n+m}(i) \geq \alpha^{m-k} g_{n+k}(i) + (1-\alpha^{m-k})\ell_{n+k}$$

$$\geq \alpha^{m-k} g_{n+k}(i) + (1-\alpha^{m-k})\ell_n.$$

Hence for all $k = 0,1,...,m$

$$g_{n+k}(i) \leq \alpha^{k-m}(\ell_{n+m} - \ell_n) + \ell_n \leq \alpha^{-m}(\ell^* - \ell_n) + \ell_n.$$

So

$$v_{n+m}(i) = v_n(i) + \sum_{k=0}^{m-1} g_{n+k}(i)$$

$$\leq v_n(i) + m\ell_n + m\alpha^{-m}(\ell^* - \ell_n).$$

On the other hand there must exist a state j such that $g_{n+k}(j) = u_{n+k} \geq u^*$ for at least $\frac{m}{N}$ of the indices $k = 0,1,\ldots, m-1$ and $g_{n+k}(j) \geq \ell_{n+k} \geq \ell_n$ for the other indices. Then for this state j

$$(10) \qquad v_{n+m}(j) = v_n(j) + \sum_{k=0}^{m-1} g_{n+k}(j) \geq v_n(j) + \frac{m}{N} u^* + (m - \frac{m}{N})\ell^*.$$

Hence by (9) and (10)

$$sp(v_{n+m}) \geq -sp(v_n) + \frac{m}{N}(u^* - \ell^*) - m\alpha^{-m}(\ell^* - \ell_n).$$

Now suppose $\ell^* < u^*$, then we can choose m as to make $\frac{m}{N}(u^* - \ell^*)$ arbitrary large and next choose m so that $m\alpha^{-m}(\ell^* - \ell_n)$ is small. Further $sp(v_n)$ is bounded, so if $\ell^* < u^*$ then we can choose n and m so that $sp(v_{n+m})$ becomes arbitrarily large. This however violates the boundedness of $\{sp(v_k)\}$. Hence $\ell^* = u^*$.

With lemma 1 (iii) this implies that the communicating Markov game has a value independent of the initial state and that the sa scheme (1) yields good bounds on this value and nearly-optimal stationary strategies for the two players.

THE SIMPLY CONNECTED CASE

For the simply connected game it is clear that on the set $\hat{S}$ the sequence $\{v_{n+1} - v_n\}$ converges to a constant vector.
From (4) we see that the system reaches $\hat{S}$ from $\bar{S}$ exponentially fast thus it follows from inequalities like (7) that $\{v_{n+1} - v_n\}$ becomes constant on the whole state space. So also the simply connected game has a constant value and the method of sa yields good bounds on g^* and nearly-optimal stationary strategies for the two players.

References

[1] Bather, J.A., Optimal decision procedures for finite Markov chains, Part II, Adv. Appl. Prob. 5 (1973), 521-540.

[2] Blackwell, D. and Ferguson, T.S., The big match, Ann. Math. Statist. 39 (1968), 159-163.

[3] Federgruen, A., Successive approximation methods in undiscounted stochastic games, Oper. Res. 28 (1980). 794-809.

[4] Gillette,D., Stochastic games with zero stop probabilities, in Contributions to the theory of games, Vol.III, eds. M. Dresher, A. Tucker and P. Wolfe, Princeton Univ. Press, Princeton, New Jersey, 179-187, 1957.

[5] Mertens, J.F. and Neyman, A., Stochastic games, Univ. Catholique de Louvain, Dept. of Math., 1980.

[6] Monash, C.A., Stochastic games: the minimax theorem, Harvard Univ., Cambridge, Massachusetts, 1979.

[7] Platzman, L., Improved conditions for convergence in undiscounted Markov renewal programming, Oper. Res. 25 (1977), 529-533.

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

[9] Van der Wal, J., Successive Approximations for average reward Markov games, Int. J. Game Theory 9 (1980), 13-24.

[10] Van der Wal, J., The method of value oriented successive approximations for the average reward Markov decision process, OR Spektrum 1 (1980), 233-242.

[11] Van der Wal, J., Stochastic dynamic programming, to appear in the series of Mathematical Centre Tracts, Mathematisch Centrum, Amsterdam.