Successive approximation methods for Markov games

van der Wal, J.; Wessels, J.

Published: 01/01/1976

Citation for published version (APA):
Successive approximation methods for Markov games

by

J. van der Wal and J. Wessels

Eindhoven, November 1976

The Netherlands
Successive approximation methods for Markov-games

by

J. van der Wal and J. Wessels

Summary

In this paper an overview will be presented of the applicability of successive approximation methods for Markov games. The larger part of the paper will be devoted to two-person zero-sum Markov games with the total expected reward criterion. The analysis will include policy-iteration algorithms. Finally there are sections on the average-reward case and on the nonzero-sum case.

1. Introduction

The main purpose of this paper is to investigate the following question: can the theory of successive approximations for Markov decision processes be extended to Markov games?

A preliminary answer to this question can be very short, since Shapley [14] introduced already in 1953 successive approximations for Markov games, which were only introduced in 1957 for Markov decision processes [Bellman, 1]. However, for Markov decision processes, under relatively weak conditions, several types of successive approximations methods have been derived, together with sophisticated extrapolation procedures, see e.g. [8] and [9] in this volume. So the present paper will be mainly concerned with the question of generalizing this theory to Markov games. For an elementary treatment of dynamic programming in Markov games we refer to [20]. For other aspects of the theory of Markov games we refer to the recent bibliography and survey by Parthasarathy and Stern [10].

In section 2 the model will be introduced, also the finite stage case will be treated. We will allow unbounded rewards, but as in [8] and [9] contraction will be assumed. In section 3 the infinite stage case will be treated. In that section we will show that the standard successive approximations technique (extrapolations included) for the expected total reward criterion may be extended to contracting Markov games. In section 4 it will be shown that this is less easy for the policy iteration and
other value oriented methods. However, a suitable extension will be presented. In section 5 positive Markov games with stopping actions for the second player will be considered. These games are not necessarily contracting. Section 6 is devoted to Markov games with the average reward criterion, and section 7 to the nonzero-sum case.

2. The model

As in [8J and [9J, we consider a system, which is observed at discrete points in time $t = 0, 1, 2, \ldots$. The system can be in one of a countable number of states: $\mathcal{S} = \{1, 2, \ldots\}$. In each state $i$ and at each time $t$ the proceedings of the system may be influenced. This may be done by two players $P_1$ and $P_2$. Except in section 7 these players are supposed to have completely opposite aims. In each state $i$ there are two finite (nonempty) sets $\mathcal{K}_i$ and $\mathcal{L}_i$ of allowed actions for $P_1$ and $P_2$ respectively. If at some time $t$ the system is in state $i$ and the players choose actions $k$ and $\ell$ from $\mathcal{K}_i$ and $\mathcal{L}_i$ respectively, then this results in an immediate reward $r(i, k, \ell)$ for $P_1$ (to be paid by $P_2$) and it further results in a transition of the system to state $j$ with probability $p(j | i, k, \ell)$. We suppose $\sum_j p(j | i, k, \ell) \leq 1$.

A strategy $\pi$ for $P_1$ specifies for all times $t$ and all possible histories $h_t$ the probability $\pi_t(k|h_t)$ of choosing action $k$. Here the history $h_t$ equals the sequence of states and actions in the past:

$$h_t = (s_0, k_0, \ell_0, \ldots, s_{t-1}, k_{t-1}, \ell_{t-1}, s_t)$$

where $s_t$ is the state of the system at time $t$ and $k_t$, $\ell_t$ are the actions chosen at time $t$ by $P_1$ and $P_2$. If these probabilities only depend on $s_t$ instead of $h_t$, then $\pi$ is called a Markov strategy. If, moreover, $\pi_t$ does not depend on $t$ explicitly then $\pi$ is called a stationary strategy.

Stationary strategies correspond to policies, where a policy $f$ for $P_1$ is any function on $\mathcal{S}$ such that $f(i)$ is a probability distribution on $\mathcal{K}_i$, we will use the notation $f$ both for policies and stationary strategies,
by \( f(i,k) \) we denote the probability of \( k \in \mathcal{K}_i \). The set of policies for \( P_1 \) is denoted by \( F \), the set of strategies by \( \pi \). Similarly one defines strategies \( \gamma \in \Gamma \) and policies \( g \in G \) for the player \( P_2 \).

**Notations**

\( r(f,g) \) will denote for \( f \in F, g \in G \) a real-valued function on \( \$ \) with

\[
 r(f,g)(i) := \sum_{K \in \mathcal{K}_i} \sum_{L \in \mathcal{K}_L} f(i,k) g(i,L) r(i,k,L)
\]

\( P(f,g) \) will denote a nonnegative function on \( \$ \times \$ \) with

\[
 P(f,g)(i,j) := \sum_{K \in \mathcal{K}_i} \sum_{L \in \mathcal{K}_L} f(i,k) g(i,L) p(j|i,k,L)
\]

Functions on \( \$ \) and \( \$ \times \$ \) respectively will be treated as column-vectors and matrices, with matrix products and matrix-vector products defined in the obvious way.

We will work under the following assumptions in this section and in section 3-4.

**Assumptions**

It is supposed that there is a positive function \( \mu \) on \( \$ \), which defines (as in [8] and [9]) a Banach space \( \mathcal{W} \) of vectors \( w \) with the norm

\[
\| w \|_\mu = \sup_i |w(i)| \mu^{-1}(i), \text{ such that}
\]

a. \( \| r(f,g) \|_\mu \leq M \) for some \( M \) and all \( f \in F, g \in G \).

b. \( \| P(f,g) \|_\mu \leq \rho < 1 \) for some \( \rho \) and all \( f \in F, g \in G \).

In order to simplify our notations we will write \( W \) and \( \| \| \) instead of \( \mathcal{W} \) and \( \| \|_\mu \) whenever this is possible.

**Remark**

These assumptions are somewhat more restrictive than those in [8] and even than those in [9]. However, as shown in [21], assumption a. may be weakened to
$\bar{r} := \sup_{f \in F} \inf_{g \in G} r(f,g) \in W$.

Furthermore the use of the Harrison translation does not present essential difficulties. To avoid technical details we will stick in this and the following two sections to the assumptions stated before.

As in [8] and [9] the transition probabilities may be defective, i.e. $\sum_j p(j|i,k,\ell) \leq 1$. This may be repaired by the introduction of an absorbing state. We will not do this explicitly.

Also the condition $K_i^j, L_i^j$ finite is not very essential. It may be replaced by $K_i^j, L_i^j$ compact, $p(j|i,k,\ell), r(i,k,\ell)$ continuous in $k, \ell$.

A starting state $i$ and strategies $\pi \in \Pi, \gamma \in \Gamma$ determine a stochastic process $\{(S_t, K_t, L_t)\}_{t=0}^\infty$ in an obvious way, where $S_t$ is the state at time $t$ and $K_t, L_t$ the actions chosen at time $t$ by $P_1$ and $P_2$ respectively. Probabilities referring to this process are denoted by $\mathbb{E}_i^{\pi,\gamma}$, expectations by $E_i^{\pi,\gamma}$. If the index $i$ is deleted, a column vector of probabilities or expectations is meant.

The assumptions guarantee (compare [8])

$$\mathbb{E}_i^{\pi,\gamma} \sum_{t=0}^\infty |r(S_t, K_t, L_t)| < \infty$$

and even

$$\|\mathbb{E}_i^{\pi,\gamma} \sum_{t=N}^\infty |r(S_t, K_t, L_t)| \| \leq M(1 - \rho)^{-1} \rho^N.$$ 

Therefore the total expected rewards (for $P_1$) are properly defined for any pair of strategies:

$$V(\pi, \gamma) := \mathbb{E}_i^{\pi,\gamma} \sum_{t=0}^\infty r(S_t, K_t, L_t).$$

Strategies $\pi^*, \gamma^*$ are said to be optimal if

$$V(\pi, \gamma^*) \leq V(\pi^*, \gamma^*) =: V^* \leq V(\pi^*, \gamma)$$

for all $\pi \in \Pi, \gamma \in \Gamma$.

$V^*$ will be called the value of the game.
Analogous to [8] we introduce the following operators in W.

\[ L(f,g)w := r(f,g) + P(f,g)w \]

\[ Uw := \max_{f \in F} \min_{g \in G} L(f,g)w \]

with max-min taken componentwise.

Note that \((Uw)(i)\) is the value of the matrix game with entries

\[ r(i,k,l) + \sum_j p(j|i,k,l)w(j) \]

Now define \(w_n := Uw_{n-1}\) \((n = 1, \ldots, T)\) for some \(w_0 \in W\) and find policies which satisfy for \(n = 1, \ldots, T\)

\[ L(f,g)w_{n-1} \leq L(f,g)w_n \leq L(f,g)w_{n-1} \quad \text{for all } f, g. \]

Then we get the following result for the T-stage Markov game with terminal reward \(w\) (actually for this result the assumption \(p < 1\) may be replaced by \(p < \infty\):)

**Theorem 1**

The T-stage Markov game with terminal reward \(w_0 \in W\), i.e. the game with criterion function

\[ V_T(\pi, \gamma) := \mathbb{E}^{\pi, \gamma} \left[ \sum_{t=0}^{T-1} r(S_t, K_t, L_t) + w_0(S_T) \right] \]

has the value \(w_T\) and the strategies \(\pi_T\) and \(\gamma_T\), which might be denoted by \((f_T, \ldots, f_1), (g_T, \ldots, g_1)\), are optimal.

The proof proceeds by induction. For details see [20].

This shows that in the finite-stage case optimal strategies may be found by dynamic programming or successive approximation. In the following section we will extend this result to the infinite-stage case. For that case our methods of proof bear more heavily on the assumptions. Especially (compare e.g. [21], [17]) it is very essential that the assumptions imply
Lemma 1

$L(f, g)$ and $U$ are contracting with contraction radii $\|P(f, g)\|$ and $\nu$ respectively, with

$$\nu \leq \max_{f, g} \|P(f, g)\| \leq \rho < 1$$

As a consequence of this lemma the operators $L(f, g)$ and $U$ possess unique fixed points. For $L(f, g)$ this fixed point is exactly $V(f, g)$, the criterion value for the stationary strategies $f, g$ in the infinite-stage Markov game. For $U$ this fixed point will be shown to be equal to the value $V^*$ of the game.

3. The $\infty$-stage Markov game

Let $w^* \in W$ be the unique fixed point of $U$ in $W$, so $Uw^* = w^*$.

Let $f^*$, $g^*$ satisfy

$$L(f, g^*)w^* \leq L(f^*, g^*)w^* \leq L(f^*, g)w^* \text{ for all } f \in F, g \in G.$$  

We will prove the following result, which has been proved already in 1953 by Shapley [14] for the finite state case with

$$\sum_{j} p(j|i, k, \ell) \leq \beta < 1 :$$

Theorem 2

The stationary strategies $f^*$ and $g^*$ are optimal in the $\infty$-stage Markov game and $w^*$ is the value of the game, i.e. $V^* = w^*$.

Proof

Obviously (theorem 1) the $T$-stage game with terminal reward $w^*$ has value $w^*$ and $f^*$, $g^*$ are optimal stationary strategies for that game.

Suppose $P_1$ plays $f^*$ and $P_2$ an arbitrary strategy $\gamma$. Then for all $T$,

$$V(f^*, \gamma) \geq w^* - p(T)w^* - \frac{MD_T}{1-\mu},$$

where $p(t)$ is the matrix with $(i,j)$ entry $P_{i, j}^f (S_T = j)$. 
One may show \( P(T)^* \leq P(T)^{\infty} \|w^*\| \leq \rho P(T)^* \|w^*\| \).
Hence \( V(f^*, \gamma) \geq w^* = V(f^*, g^*) \).
Similarly one shows
\[
V(\pi, g^*) \leq w^* \quad \text{for all } \pi \in \Pi.
\]
Hence \( w^* = V^* \).

So the \( \infty \)-stage game possesses a value and optimal stationary strategies.
It will now be investigated whether successive approximations produce
\( \varepsilon \)-optimal stationary strategies and bounds for \( V^* \) which are arbitrarily
close.

**Definition**

\( \pi_{\varepsilon} \in \Pi \) is called \( \varepsilon \)-optimal if
\( V(\pi_{\varepsilon}, \gamma) \geq V^* - \varepsilon \mu \) for all \( \gamma \in \Gamma \).

\( \gamma_{\varepsilon} \in \Gamma \) is called \( \varepsilon \)-optimal if
\( V(\pi, \gamma_{\varepsilon}) \leq V^* + \varepsilon \mu \) for all \( \pi \in \Pi \).

An obvious way of approximating \( V^* \) is suggested by the fixed point property
of \( U \):

**Theorem 3**

Choose \( w_0 \in W \). Then \( w_n := Uw_{n-1} (n = 1, 2, \ldots) \) converges (in norm) to \( V^* \) and
one actually gets the following bounds
\[
w_n - \nu(1 - \nu)^{-1} \|w_n - w_{n-1}\| \leq V^* \leq w_n + \nu(1 - \nu)^{-1} \|w_n - w_{n-1}\| \mu.
\]
However, somewhat better estimates can be given and one may simultaneously
give bounds for the policies \( f_n \) and \( g_n \) (see section 2) found in the \( n \)-th
iteration.

We first introduce some notations:
\[\lambda_n := \inf_i (w_n(i) - w_{n-1}(i)) \mu^{-1}(i),\]

\[\nu_n := \sup_i (w_n(i) - w_{n-1}(i)) \mu^{-1}(i),\]

\[a_n := \begin{cases} 
\sup_{i,g} P(f_n,g)(i,j) \mu(j) & \text{if } \lambda_n < 0, \\
\inf_{i,f} P(f_n,g)(i,j) \mu(j) & \text{if } \lambda_n \geq 0,
\end{cases}\]

\[b_n := \begin{cases} 
\inf_{i,f} P(f_n,g)(i,j) \mu(j) & \text{if } \nu_n < 0, \\
\sup_{i,g} P(f_n,g)(i,j) \mu(j) & \text{if } \nu_n \geq 0.
\end{cases}\]

Theorem 4

Choose \(w_0 \in W\), define \(w_n := Uw_{n-1}\) (\(n = 1, 2, \ldots\)). Let \(f \in F, g_n \in G\) satisfy

\[L(f,g_n)w_{n-1} \leq L(f_n,g_n)w_{n-1} = w_n \leq L(f_n,g)w_{n-1}.\]

Then we have the following bounds for \(V, V(f_n,g_n), V(f_n,\gamma),\) and \(V(\pi,g_n)\):

a. \(w_n + a_n \lambda_n (1 - a_n)^{-1} \mu \leq V \leq w_n + b_n \nu_n (1 - b_n)^{-1} \mu,\)

b. \(V(f_n,\gamma) \geq w_n + a_n \lambda_n (1 - a_n)^{-1} \mu\) for all \(\gamma \in \Gamma,\)

c. \(V(\pi,g_n) \leq w_n + b_n \nu_n (1 - b_n)^{-1} \mu\) for all \(\pi \in \Pi,\)

d. \(w_n + a_n \lambda_n (1 - a_n)^{-1} \mu \leq V(f_n,g_n) \leq w_n + b_n \nu_n (1 - b_n)^{-1} \mu.\)

Proof

a, d are direct consequences of b and c.

The proof of c will be sketched (for more detailed proofs in somewhat different situations, see [17], [21]).
It suffices to prove \( c \) for stationary strategies \( \pi \) (compare [8]). Consider a policy (or stationary strategy) \( f \).

\[
L(f, g_n) w_{n-1} \leq w_n \leq w_{n-1} + v_n \mu \quad \text{(by definition)}.
\]

Hence

\[
L^2(f, g_n) w_{n-1} \leq L(f, g_n) [w_{n-1} + v_n \mu] = L(f, g_n) w_{n-1} + v_n P(f, g_n) \mu \leq w_n + v_n b_n \mu.
\]

In this way one obtains

\[
L^n(f, g_n) w_{n-1} \leq w_n + v_n (b_n + \ldots + b_{n-1}) \mu.
\]

Hence

\[
V(f, g_n) = \lim_{N \to \infty} L^n(f, g_n) w_{n-1} \leq w_n + b_n v_n (1 - b_n)^{-1} \mu.
\]

In this way the standard successive approximations technique may be extended to Markov games. On the upper- and lowerbounds of theorem 4 one may base tests for suboptimality (see [16] and for a more detailed treatment [12]).

In [9] it has been shown that an extensive class of successive approximation techniques may be generated by using stopping times. This also holds for Markov games. This will not be worked out in this paper, since the concepts and proofs are rather straightforward (for finite state discounted Markov games this has been worked out in [16] and [17]).

4. Value oriented methods

In this volume Van Nunen and Wessels [9] consider a set of value oriented methods for MDP which can be viewed upon as a special type of successive approximations method. One of these methods being Howard's policy iteration method. A straightforward generalisation of Howard's method to Markov games has been proposed by Pollatschek and Avi-Itzhak [11]. This generalisation may be formulated as follows
Algorithm

step 1 \( v_0(i) := 0 \) for all \( i \in S \).

step 2 (Policy iteration). Determine policies \( f_n \) and \( g_n \), such that

\[
L(f_n, g_n) v_{n-1} \leq L(f_{n-1}', g_n) v_{n-1} \leq L(f_n, g) v_{n-1}.
\]

step 3 (Value determination) \( v_n := V(f_n, g_n) \).

Pollatschek and Avi-Itzhak proved in the finite state case that the algorithm converges under the following condition

\[
\max \left[ \sum_{j} \left( \max_{k\neq k'} p(j|i,k) - \min_{k}\ p(j|i,k) \right) \right] < 1 - \max i,j,k,l \sum_{j} p(j|i,k,l).
\]

In [18] essentially the following example has been given which proves that this algorithm does not converge in general for finite state discounted Markov games.

Example

There is but one state. In this state both players have two actions. If \( P_1 \) picks action 2 and \( P_2 \) action 1 then \( P_2 \) pays \( P_1 \) 2 units and the system stays in state 1 with probability 1/4, etc.

A policy \( f \) is completely determined by the probability \( f(1,1) \). If we apply the algorithm we find \( f_1(1,1) = g_1(1,1) = 1, v_1 = 12, f_2(1,1) = g_2(1,1) = 0, v_2 = 4, f_3(1,1) = g_3(1,1) = 1, v_3 = 12 \) etc. So the algorithm cycles without ever finding an optimal pair of strategies.

A somewhat more refined extension of Howard's method is the following. This extension has been inspired by Hoffman and Karp's method [5] for the average reward Markov game.
Algorithm (H,K)

step 1. Choose $v_0$ such that $Uv_0 \leq v_0$.

step 2. Determine $Uv_n$ and a policy $g_{n+1}$ with $L(f,g_{n+1})v_n \leq Uv_n$ for all $f$.

step 3. Determine $v_{n+1} := \max_{f} V(f,g_{n+1})$.

As in the case of MDP one may consider this algorithm as an extreme element of the following set of value oriented methods:

Algorithm ($\lambda$)

step 1. Choose $v_0$ such that $Uv_0 \leq v_0$.

step 2. Determine $Uv_n$ and a policy $g_{n+1}$ with $L(f,g_{n+1})v_n \leq Uv_n$ for all $f$.

step 3. Determine $v_{n+1} := \max_{g} U^\lambda v_n$, where that operator $U^\lambda$ is defined by

$$U^\lambda v := \max_{f} L(f,g)v.$$

For $\lambda = 1$ we have again the standard successive approximations method treated in section 3. For $\lambda = \infty$ we have Hoffman and Karp's algorithm.

One may prove, using the monotonicity of the operators and $Uv_0 \leq v_0$, that $v_n$ converges monotonically to $V^*$.

For the finite state case the proof is given in [18]. The extension of this proof to the case we deal with here is straightforward. One just has to prove by induction $V^* \leq v_n \leq Uv_{n-1} \leq v_{n-1}$, and $v_n \leq U^nv_0$. Since $\|U^nv_0 - V^*\| \leq \|v_n - V^*\|$ we also have $\|v_n - V^*\| \leq \|U^n v_0 - V^*\|$.

A possible extension is again the introduction of the stoppingtime-based $L_0$ and $U_0$ operators as in [17]. Another extension is that instead of using a fixed $\lambda$ one may use a different $\lambda_n$ in each iteration step. Note also, that if the first player has only one action in each state we get the value oriented methods presented by Van Nunen and Wessels [9] for MDP.
5. **Strictly positive Markov games with stopping actions**

In this section we will consider a type of Markov game for which successive approximations still converge but where the U and L(f,g) operators are no longer strictly contracting. We release the assumptions of section 2 and replace them by: $\mathcal{L}_1$, $\mathcal{X}_1$, all finite

$$\sum_{j \in \mathcal{S}} p(j|i,k,\ell) \leq 1, \quad r(i,k,\ell) > 0 \quad \text{for all } i, k \text{ and } \ell$$

and moreover

$$\mathcal{L}^{\text{STOP}}_i := \{ \ell \in \mathcal{L}_1 | \sum_{j \in \mathcal{S}} p(j|i,k,\ell) = 0 \quad \text{for all } k \in \mathcal{X}_1 \} \neq \emptyset \quad \text{for all } i \in \mathcal{S}.$$ 

By $\|v\|$ we mean standard maximum norm, $\|v\| = \max_{i \in \mathcal{S}} |v(i)|$.

So all rewards are strictly positive and -since $\mathcal{L}_1$, $\mathcal{X}_1$, are finite- also bounded away from zero. The assumptions allow $V(\pi,\gamma)(i) = \infty$ for some $\pi, \gamma$ and $i$. But since $\mathcal{L}^{\text{STOP}}_i$ is nonempty $P_i$ can stop playing immediately in each state and thus restrict his loss to some finite amount.

As in section 2 we have the following lemma.

**Lemma**

The n-stage game with terminal reward $w_0$ has the value $U^n w_0$ with terminal reward $w_0$ has the value $U^n w_0$ with optimal strategies $(f_1, \ldots, f_n)$ and $(g_1, \ldots, g_n)$ satisfying $L(f,g_k) w_{k-1} \leq L(f_{k-1},g_k) w_{k-1} =: w_k \leq L(f,g) w_{k-1}$ for all $f$ and $g$.

The problem remains to investigate how $w$ behaves as $n$ tends to infinity.

Let $r^{\text{STOP}}(i)$ be defined as the value of the matrix game with entries $r(i,k,\ell), k \in \mathcal{X}_1, \ell \in \mathcal{L}_1$. Then for any $w_0$ we obviously have $w_n \leq r^{\text{STOP}}, n \in \mathbb{N}$, since in state $i$ the second player may restrict his loss to $r^{\text{STOP}}(i)$ by choosing a good randomized action in $\mathcal{L}^{\text{STOP}}_i$.

We also have $0 \leq w^{n-1} \leq U^n w_0, \leq n = 2, 3, \ldots$, hence $\lim_{n \to \infty} U^n w_0$ exists. Call it $w^*$. Hence $w^*$ is a fixed point of $U : w^* = Uw^*$. 
Theorem 5

\( w^* \) is the unique fixed point of \( U \) and \( U^nv + w^* \ (n \to \infty) \) for any \( v \in \mathbb{R}^n \).

Proof

First we prove the uniqueness. Let \( u \) and \( v \) be fixed points of \( U \) which have \((f_u, g_u)\) and \((f_v, g_v)\) as optimal strategies in the one-stage game with terminal payoff \( u \) and \( v \), respectively. Then

\[
U^nu = L^n(f_u, g_u)u \geq L^n(f_v, g_u)u \geq L^n(f_v, g_v)v - \mathbb{P}^v, g_v(S_n \in \$) ||u - v||
\]

\[
\geq v - \mathbb{P}^v, g_v(S_n \in \$) ||u - v||.
\]

Obviously for all \( i \in \$ \mathbb{P}^i(i) \to 0 \ (n \to \infty) \) since otherwise \( V(f_v, g_u)(i) = \infty \) contradicting \( V(f_v, g_u) \leq u \). Hence \( u \geq v \).

Similarly \( u \leq v \) and thus \( u = v \).

So it remains to show \( U^nv + w^* \) for any \( v \). This follows from

\[
U^nv \geq L(f_w, g_n) \ldots L(f_w, g_1)v \geq L(f_w, g_n) \ldots L(f_w, g_1)w^* - \mathbb{P}^w, (g_n, \ldots, g_1) (S_n \in \$) ||v - w^*||
\]

\[
\geq w^* - \mathbb{P}^w, (g_n, \ldots, g_1) (S_n \in \$) ||v - w^*||.
\]

Again it is obvious that \( \mathbb{P}^i (S_n \in \$) \to 0 \ (n \to \infty) \).

Therefore \( \limsup_{n \to \infty} U^n v \geq w^* \). Similarly one may show \( \liminf_{n \to \infty} U^n v \leq w^* \).

Hence \( \lim_{n \to \infty} U^n v = w^* \). \( \square \)

Here it is again possible to determine bounds for \( w^* \) using that \( \mathbb{P}^i (S_n \in \$) \) converges to zero geometrically. This has been worked out on [19].
It is not necessary to assume that $P_2$ can quit playing in any state. It is sufficient to assume that $P_2$ can restrict his loss to some finite amount. This, more general case, has been treated by Kushner and Chamberlain [6].

6. Average reward Markov game

In this section the state space will be assumed to be finite.

In the previous sections we have seen that it is possible to extend many of the results with respect to successive approximations in MDP to Markov games. In the average reward case however, we encounter substantial difficulties. This is illustrated by the following example called the big match. It is due to Gillette [4] and studied by Blackwell and Ferguson [3].

Example

$S_1 = \{1, 2, 3\}, \quad \mathcal{K}_1 = \mathcal{L}_1 = \{1, 2\}, \quad \mathcal{K}_2 = \mathcal{L}_2 = \mathcal{K}_3 = \mathcal{L}_3 = \{1\}$.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

If in state 1 $P_1$ picks action 2 and $P_2$ action 1 the payoff will be zero and the system moves to state 2, etc.

So states 2 and 3 are absorbing.

One easily argues that, if $P_1$ takes in state 1 action 1 with probability 1/2, the average reward for $P_1$ will be 1/2, whatever, strategy he uses. But it is not very clear how $P_1$ can guarantee himself an average payoff of 1/2. Any Markov strategy guarantees only 0. This is seen as follows: Let $p(n)$ denote the probability that $P_1$ has picked action 2 before or on time $n$, and define $p := \lim_{n \to \infty} p(n)$. Now let $\varepsilon > 0$ be given arbitrarily and let $N_\varepsilon$ be such that

$$p - p(N_\varepsilon) \leq \varepsilon.$$

Then $P_2$'s strategy "play action 1 until time $N_\varepsilon$ and action 2 thereafter" gives an average pay off of at most $\varepsilon$. 
Blackwell and Ferguson show that $P_1$ can guarantee himself the average payoff $N/2(N + 1)$ by playing strategy $\pi_N$ defined as follows: Let $P_2$'s first $n$ choices be $\ell_1, \ldots, \ell_n$, $\ell_k \in \{1, 2\}$, and let $c_n$ be the excess of 1's over 2's among $\ell_1, \ldots, \ell_n$. Then take action 2 with probability $(N + c_n + 1)^{-2}$.

The difficulties here arise from the fact that there are strategies with more than one recurrent subchain.

Under the assumption that all pairs of stationary strategies induce an irreducible Markov chain (one recurrent subchain and no transient states) Hoffman and Karp [5] show that the game has a value and that their algorithm $(H, K)$ from section 4 yields $\varepsilon$-optimal stationary strategies. Rios and Yanez [13] consider the game with for all $i, j, k$ and $P(j|i, k, \ell) \geq \rho > 0$. (Then obviously Hoffman and Karp's irreducibility assumption is satisfied.) They show that in this case the standard successive approximations method converges. Recently Tanaka and Wakuta [15], dealing with compact state and action spaces under appropriate continuity assumptions, consider the following condition: $\mathbb{P}_{i}^{\pi, \gamma}(S_n = s_0) \geq \alpha > 0$ for some $s_0 \in \mathcal{S}$ and all $i, \pi$ and $\gamma$. And show that in this case the game has a value and that successive approximations converge.

7. Nonzero-sum two-person Markov games

This section shows that finite-stage two-person-nonzero-sum Markov games do have at least one Nash equilibrium point [7] which may be determined by successive approximations.

The main difference with the zero-sum games of the previous sections is that now we have two reward functions $r_1(x, k, \ell)$ and $r_2(x, k, \ell)$, where $r_i$ denotes the reward for $P_i$, $i = 1, 2$. Furthermore we have two terminal reward functions $w_1$ and $w_2$. As a result we have to define two total expected reward functions $V_1(\pi, \gamma)$ and $V_2(\pi, \gamma)$ for $P_1$ and $P_2$ respectively. Now we are looking for a Nash equilibrium pair (cf. [7]) for this game; that is a pair of strategies $\pi^*, \gamma^*$ satisfying $V_1(\pi, \gamma^*) \leq V_1(\pi^*, \gamma^*)$ and $V_2(\pi^*, \gamma^*) \leq V_2(\pi^*, \gamma)$ for all $\pi$ and $\gamma$. In bimatrix games (1-stage games) there can in general be more than one equilibrium pair.
The assumptions in this section are the following:

(i) $\mathcal{S}$ is countable, $\mathbb{M}_1, \mathbb{M}_2$ finite

(ii) There exist two positive vectors, $\mu_1$ and $\mu_2$ such that

$$\|r_1(f,g)\|_{\mu_1} \leq M_1 \quad \text{and} \quad \|P(f,g)\|_{\mu_1} \leq \rho_1 \quad \text{for all } f \text{ and } g$$

$$\|r_2(f,g)\|_{\mu_2} \leq M_2 \quad \text{and} \quad \|P(f,g)\|_{\mu_2} \leq \rho_2 \quad \text{for all } f \text{ and } g$$

where $M_1, M_2, \rho_1, \rho_2$ are real numbers.

Analogous to section 2 we define the operators $L_1$ and $L_2$ on $\mu_1$ and $\mu_2$ respectively by

$$L_i(f,g)w(x) := \sum_{k \in \mathcal{K}} \sum_{x \in \mathcal{X}} f(x,k)g(x,\ell)[r_i(x,k,\ell) + \sum_{j \in \mathcal{S}} p(j|x,k,\ell)w(j)] \quad i = 1,2.$$ 

Now for all $x \in \mathcal{S}$, $w_1 \in \mu_1$, $w_2 \in \mu_2$, $L_1(f,g)w_1$ and $L_2(f,g)w_2$ determine a bimatrix game. Note that assumption (ii) guarantees that $L_1(f,g)w_1$ lies again in $\mu_1$.

Let us consider the $n$-stage game with terminal payoffs $w_1$ and $w_2$ for $P_1$ and $P_2$ respectively, with $w_1 \in \mu_1$.

Now define $w_1^0 := w_1$, $w_2^0 := w_2$. Let $f_n$ and $g_n$ be a pair of policies satisfying

$$L_1(f_n,g_n)w_1^{n-1} \leq L_1(f_n,g_n)w_1^n \quad \text{and} \quad L_2(f_n,g_n)w_2^{n-1} \leq L_2(f_n,g_n)w_2^n \quad \text{for all } f_n \text{ and } g_n$$

and define $w_1^n := L_1(f_n,g_n)w_1^{n-1}$, $w_2^n := L_2(f_n,g_n)w_2^{n-1}$, $i = 1,2$. Then we have the following result: The pair of strategies $\pi_n := (f_n, \ldots, f_1)$, $\gamma_n := (g_n, \ldots, g_1)$ is a Nash equilibrium pair of strategies for the $n$-stage game under consideration.

The proof of this statement goes along the same lines as the proof in [20] for zero-sum games, essentially using the monotonicity of the $L$ operators.

For infinite stage games there are a number of theorems about the existence of a pair of equilibrium strategies. See for example the survey paper by Parthasarathy and Stern [10]. Beniest [2] considers a game with $\mathcal{S}$ finite.
and \( \sum_{j \in S} p(j|i,k,\ell) < 1 \) for all \( i, k \) and \( \ell \)

under two different cooperation schemes and shows that in both cases there exists a unique pair of value vectors \( v^*_1, v^*_2 \) which may be determined by successive approximations.

For the case of noncooperation the following example shows one of the problems we encounter when considering infinite stage games.

**Example.**

\[
\begin{array}{ccc}
5,5,3/4 & 1,7,3/4 \\
7,1,3/4 & 2,2,3/4 \\
\end{array}
\]

There is only one state. If \( P_1 \) picks action 1 and \( P_2 \) action 2 then \( P_1 \) receives 1, \( P_2 \) 7 and the system vanishes with probability 1/4, etc.

For each finite horizon game there is only one equilibrium pair of strategies, namely pick always action 2. In the infinite horizon games however there is still another equilibrium pair consisting of non-Markov strategies. Namely pick action 1 until your opponent has picked action 2, then continue to play action 2. One easily argues that if both players use this strategy this is indeed an equilibrium pair.

**Acknowledgement**

With respect to this section, the authors gratefully acknowledge the contribution of their student Mr. Pulskens.

**References**


