Note on the optimal strategies for the finite-stage Markov game

EINDHOVEN UNIVERSITY OF TECHNOLOGY

Department of Mathematics

STATISTICS AND OPERATIONS RESEARCH GROUP

memorandum COSOR 75-06

Note on the optimal strategies for the finite-stage Markov game

by

J. van der Wal

Eindhoven, June 1975

Note on the optimal strategies for the finite-stage Markov game

by

J. van der Wal

Abstract. In this note we consider the finite-stage Markov game with finitely many states and actions as described by Zachrisson [5]. Zachrisson proves that this game has a value and shows that value and optimal strategies may be determined with a dynamic programming approach. However, he silently assumed that both players would use only Markov strategies. Here we will give a simple proof which shows this restriction to be irrelevant.

1. Introduction and notations

The finite-stage Markov game considered here is a game between two players which proceeds as follows. At each of a finite number of time instants both players select an action out of a finite set of allowed actions. As a result of these two actions the state of the game is changed and one of the players receives some amount, specified by the rules of the game, from the other. This we formalize as follows.

We will consider a dynamic system with finite state space $S := \{1,\dots,N\}$, the behavior of which is influenced by two players, P_1 and P_2, having opposite aims. For each state $x \in S$ two finite non-empty sets of actions exist, one for each player, denoted by K_x for P_1 and L_x for P_2. At T equidistant time instants, numbered in reversed order $n = T, T-1, \dots, 1$, both players select an action out of the set available to them. As a joint result of the two selected actions, k for P_1 and ℓ for P_2, the system moves to a new state y with probability $p(y|x,k,\ell)$, with $\sum_{y \in S} p(y|x,k,\ell) = 1$, and P_1 will receive some (possibly negative) amount from P_2, denoted by $r(x,k,\ell)$. Moreover we will assume, that if - as a result of the actions at $n = 1$ - the system moves to state y at the end of the game, P_1 will receive a final payoff $q(y)$ from P_2.

We will call this game the T-stage Markov game with final payoff q.

In this note we will prove that this game has a value and we will derive some properties of the strategies which maximize the total expected income for a player over the duration of the game. Moreover we will give a way to determine value and optimal strategies. First we give some definitions and notations.

A strategy π for P_1 for the game is any function that specifies for each time instant $n = T, T-1, \ldots, 1$, and for each state $x \in S$, the probability $\pi(k|x,n,h_n)$ that action $k \in K_x$ will be taken as a function of x,n and the history h_n. By h_n we mean the history of the game upto time-instant n, the sequence $h_n = (x_T, k_T, \ell_T, \ldots, x_{n+1}, k_{n+1}, \ell_{n+1})$ of prior states and actions (h_T is the empty sequence). We will call π a Markov strategy if all $\pi(k|x,n,h_n)$ are independent of h_n.

A policy f for P_1 will be defined as any function such that f(x) is a probability distribution on K_x for all $x \in S$. Thus a Markov strategy π consists of T policies and we will denote it by $\pi = (f_T, \ldots, f_1)$ (f_n is the policy to be used at time instant n). Similarly we defined strategies ρ and policies g for P_2.

Let $V(\pi,\rho)$ denote the N-column vector with x-th component equal to the total expected reward for P_1 when the game starts in state x, P_1 plays strategy π and P_2 plays strategy ρ. Strategies π^* and ρ^* satisfying $V(\pi,\rho^*) \leq V(\pi^*,\rho^*) \leq V(\pi^*,\rho)$ for all π and ρ will be called optimal and $V(\pi^*,\rho^*)$ is called the value of the game.

The finite-stage Markov game has already been considered by Zachrisson [5]. However, he (silently) assumed that both players would use only Markov strategies. Under this assumption Zachrisson proves that the game has a value and that the value and optimal strategies for both players can be determined by a dynamic programming approach. In the early days of Markov decision processes the same restriction was made. Derman [1] proved that the "intuitively obvious" restriction to Markov strategies was correct. Here we will do the same for finite-stage Markov games.

So we will show that there exist Markov strategies π^* and ρ^* satisfying for all strategies π and ρ $V(\pi,\rho^*) \leq V(\pi^*,\rho^*) \leq V(\pi^*,\rho)$.

2. The existence of optimal Markov strategies

In order to simplify the notations we introduce two operators.
Let f and g be arbitrary policies then the operators $L(f,g)$ and U on $\mathbb{R}^N$ are defined by

$$(L(f,g)v)(x) := \sum_{k\in K_x} f^k(x) \sum_{\ell\in L_x} g^\ell(x)[r(x,k,\ell) + \sum_{y\in S} p(y|x,k,\ell)], \; x\in S$$

with $f^k(x)$ ($g^\ell(x)$) denoting the probability that in state x action $k(\ell)$ will be taken when policy $f(g)$ is used.

$$Uv := \max_f \min_g L(f,g)v$$

(where maxmin is taken componentwise).

Now the sequence v_n, $n = 0,1,\dots,T$, $v_n \in \mathbb{R}^N$ is defined by

$$\begin{cases} v_0(x) := q(x), \; x \in S \\ v_n \quad := Uv_{n-1}, \; n = 1,\dots,T . \end{cases}$$

We expect v_T to be the value of the game. Before we prove this we first give two lemmas.

Lemma 1. The 1-stage Markov game with final payoff v has value Uv and there exist policies f^* and g^* satisfying $L(f,g^*)v \le L(f^*,g^*)v \le L(f^*,g)v$ for all f and g.

Proof. For any $x \in S$ the game with initial state x is a matrix game with value $(Uv)(x)$. For this game (randomized) optimal actions $f^*(x)$ and $g^*(x)$ exist. Thus the game has value Uv and the policies f^* and g^* are optimal. □

Let f_n^* and g_n^* be optimal policies in the 1-stage Markov game with final payoff v_{n-1}, $n = 1,\dots,T$. That is f_n^* and g_n^* satisfy

$L(f,g_n^*)v_{n-1} \le L(f_n^*,g_n^*)v_{n-1} = v_n \le L(f_n^*,g)$ for all policies f and g.

Define the strategies π^* and ρ^* by $\pi^* = (f_T^*,\dots,f_1^*)$, $\rho^* = (g_T^*,\dots,g_1^*)$.

Let $v_n(\pi,\rho^*,h_n,x)$, $n = 1,\dots,T$ denote the conditional expected reward for P_1 from the n-th epoch onwards if the system is in state x at epoch n, strategies π and ρ^* are used and history h_n has been observed.
And define $v_0(\pi,\rho^*,h_0,x) := q(x)$ for all π, h_0 and $x \in S$.

Lemma 2. Strategy π^* satisfies $V(\pi^*,\rho^*) \geq V(\pi,\rho^*)$ for all π.

Proof. We will prove the assertion by induction. By definition we have for all π and h_0

$$v_0(\pi,\rho^*,h_0,x) \leq v_0(\pi^*,\rho^*,h_0,x) = v_0(x), \quad x \in S .$$

Now assume $v_t(\pi,\rho^*,h_t,x) \leq v_t(\pi^*,\rho^*,h_t,x) = v_t(x)$, $t = 0,\dots,n$ for all π, h_t and x. So for all π, h_{n+1} and x we have

$$\begin{aligned} v_{n+1}(\pi,\rho^*,h_{n+1},x) &= \sum_{k\in K_x} \pi(k|x,n,h_{n+1}) \sum_{\ell\in L_x} g^{*\ell}_{n+1}(x)[r(x,k,\ell) + \\ &+ \sum_{y\in S} p(y|x,k,\ell)v_n(\pi,\rho^*,h_{n+1} \circ (x,k,\ell),y)] \leq \\ &\leq \sum_{k\in K_x} \pi(k|x,n,h_{n+1}) \sum_{\ell\in L_x} g^{*\ell}_{n+1}(x)[r(x,k,\ell) + \\ &+ \sum_{y\in S} p(y|x,k,\ell)v_n(y)] \leq \\ &\leq v_{n+1}(x) = v_{n+1}(\pi^*,\rho^*,h_{n+1},x) , \end{aligned}$$

where $h_{n+1} \circ (x,k,\ell)$ denotes the concatenation of h_{n+1} and (x,k,ℓ) with result h_n. The first inequality follows from the induction assumption and the latter one from the definition of v_{n+1} and g^*_n. The latter equality follows from $v_{n+1} = L(f^*_{n+1},g^*_{n+1})v_n$ and the induction assumption. Hence for all $x \in S$

$$v_T(\pi,\rho^*,h_T,x) \leq v_T(\pi^*,\rho^*,h_T,x) \quad \text{or} \quad V(\pi,\rho^*) \leq V(\pi^*,\rho^*). \qquad \square$$

The proof of the above Lemma is a shortcut of the proof given by Derman [1] for the existence of memoryless optimal strategies in finite stage Markov decision processes.

We are now ready to show:

<u>Theorem</u>. The T-stage Markov game with final payoff q has the value v_T and the Markov strategies π^* and ρ^* are optimal, that is
$V(\pi,\rho^*) \le V(\pi^*,\rho^*) = v_T \le V(\pi^*,\rho)$ for all strategies π and ρ.

<u>Proof</u>. From Lemma 2 we have $V(\pi,\rho^*) \le V(\pi^*,\rho^*)$. By interchanging the roles of π and ρ we may show in the same way $V(\pi^*,\rho^*) \le V(\pi^*,\rho)$. This proves the assertion. □

Summarizing we see that we have shown that the following algorithm provides the value v_T of the game and optimal strategies π^* and ρ^*.

(i) Set $v_0(x) = q(x)$, $x = 1,\dots,N$.

(ii) Determine for $n = 1,\dots,T$ policies f_n^*, and g_n^* satisfying for all f and g

$$L(f,g_n^*)v_{n-1} \le L(f_n,g_n^*)v_{n-1} \le L(f_n^*,g_n^*)v_{n-1}$$

and define $v_n := L(f_n^*,g_n^*)v_{n-1}$.

(iii) v_T is the value of the game and $\pi^* = (f_T^*,\dots,f_1^*)$ and $\rho^* = (g_T^*,\dots,g_1^*)$ are optimal strategies for P_1 and P_2 respectively.

3. Extensions and remarks

We considered the case that neither the state space nor the action spaces depend on the time t. And we demanded $\sum_{y\in S} p(y|x,k,\ell) = 1$ for all x, k and ℓ and the times at which the system is influenced to be equidistant.
None of these restrictions however, is essential. It is easily seen that we may allow the state space and the action spaces to depend on t. And only trivial changes in the proofs are needed if we allow $\sum_{y\in S} p(y|x,k,\ell) < 1$ for some or all x, k and ℓ. If the time between two epochs is a random variable with probability distribution $F(.|y,x,k,\ell)$ if in state x actions k and ℓ are taken and the system moves to y we must be careful. In order to avoid difficulties we demand these random variables to have finite expectations. For these finite-stage semi-Markov games only minor changes in the proofs are needed to obtain the same results. E.g. we would have to extend the history of the system with the time elapsed before the next state is reached.

Instead of considering the criterion of total expected rewards it is also possible to use the criterion of total expected discounted rewards. For the game with equidistant time instants we may use any discount factor $\beta \in [0,\infty)$. For the semi-Markov game we may use $\beta \in [0,1]$ but if we want to use $\beta > 1$ we must demand $\int_0^\infty \beta^t dF(t|y,x,k,\ell) < \infty$ for all y,x,k and ℓ.
Here we only considered finite-stage Markov games. However, our results may easily be extended to some infinite-horizon Markov games. For example consider the infinite-horizon Markov game as described by Shapley [2] with the criterion of total expected reward (Shapley considers the case $\sum_{y \in S} p(y|x,k,\ell) < s < 1$ for all x,k and ℓ) or the β-discounted ($\beta \in [0,1)$) infinite horizon Markov game. In order to prove that these games have a value and to find (near) optimal strategies for both players one usually approximates the game by a finite-stage Markov game. If we let v_n denote the value of the n-stage Markov game we may easily show that v_n tends to the value v^* of the infinite horizon Markov game if n tends to infinity. Moreover, one may prove that if f(g) is an optimal policy for the 1-stage (discounted) Markov game with final payoff v^* the strategy $f^{(\infty)} = (f,f,\dots)$ $(g^{(\infty)})$ will be optimal in the infinite horizon Markov game. This is shown in Van der Wal [4]. Two other types of infinite horizon Markov games with the criterion of total expected rewards may be found in Van der Wal [3].

References.

[1] Derman, C., Finite state Markovian decision processes. Academic Press, New York and London, 1970.

[2] Shapley, L.S., Stochastic games. Proc. Nat. Acad. Sci. USA 39 (1953), 1095-1100.

[3] Van der Wal, J., The solution of Markov games by successive approximation. Master's thesis, Department of Mathematics, Technological University Eindhoven, 1975.

[4] Van der Wal, J., The method of successive approximations of the discounted Markov game. Memorandum COSOR 75-02. Technological University Eindhoven, March 1975 (Department of Mathematics).

[5] Zachrisson, L.E., Markov games. Annals of Mathematics Studies No. 52, Princeton, New Yersey, 1964, 211-253.