stichting
mathematisch
centrum

AFDELING MATHEMATISCHE BESLISKUNDE BW 68/76 NOVEMBER
(DEPARTMENT OF OPERATIONS RESEARCH)

A. HORDIJK, O.J. VRIEZE & G.L. WANROOIJ

SEMI-MARKOV STRATEGIES IN STOCHASTIC GAMES

2e boerhaavestraat 49 amsterdam

Printed at the Mathematical Centre, 49, 2e Boerhaavestraat, Amsterdam.

The Mathematical Centre, founded the 11-th of February 1946, is a non-profit institution aiming at the promotion of pure mathematics and its applications. It is sponsored by the Netherlands Government through the Netherlands Organization for the Advancement of Pure Research (Z.W.O), by the Municipality of Amsterdam, by the University of Amsterdam, by the Free University at Amsterdam, and by industries.

AMS(MOS) subject classification scheme (1970): 93E05

Semi-Markov strategies in stochastic games

by

A. Hordijk*), O.J. Vrieze and G.L. Wanrooij

ABSTRACT

For a stochastic game with countable state and action spaces we proof that solutions in the game where all players are restricted to semi-Markov strategies are solutions for the unrestricted game. An example shows that while the unrestricted game is solvable we cannot always find solutions in the restricted game.

KEY WORDS & PHRASES: *Stochastic game; discounted model; average return model; N-person game; semi-Markov strategies; equilibrium point.*

*) University of Leiden

1. INTRODUCTION

The concept of a stochastic game was introduced by SHAPLEY [6]; his model belongs to the two person zero sum games. A two person non zero sum version was treated by ROGERS [5]; SOBEL [7] introduced the N-person stochastic game. Due to different specifications for state- and action spaces there are many models refered to as a stochastic game.

In this paper a stochastic game will be a discrete time dynamic system with a countable state space: $\{1,2,\ldots\}$. At times $0,1,2,\ldots$ players $\{1,2,\ldots,N\}$ choose simultaniously an action out of a countable action space: $\{1,2,\ldots\}$. If the system is in state s at time t and the players choose actions $a_1,\ldots,a_N$ there will be a payment $r_i(s,a,\ldots,a_N)$ to player i and the system has probability $q(s'|s,a_1,\ldots,a_N)$ to be in state s' at time t+1.

Games with finite state space or finite action spaces for some players in some states can be viewed as a special case of this model, since we can enlarge the state or action spaces with a sequence of states or actions that are essentially the same as already existing states or actions.

A strategy for player i is a mechanism for choosing actions in all circumstances that can occur during the play. At every time t the state s^t at time t and the history before time t (the sequence of states and actions choosen at times $1,\ldots,t-1$) is known to the players. So the game is of perfect recall and by a result of AUMANN [1] for each strategy for a player we can find an equivalent behavior strategy. Let s^t be the state at time t and a_i^t the action choosen by player i at time t then a behavior strategy for player i π_i specifies for each t and each history $h^t = (s^0,a_1^0,\ldots,a_N^0,s^1,\ldots,a_N^{t-1},s^t)$ a probability distribution $\pi_i^t(h^t)$ on the action space. $\pi_i^t(a|h^t)$ is the probability with which player i chooses action a at time t if history h_2^t occured. More formally π_i is a sequence $\pi_i^1,\pi_i^2,\ldots$ where π_i^t is a mapping from the product set of tN+N+1 times the positive integers to het set of probability distributions on the positive integers.

A semi-Markov strategy for player i is a behavior strategy for which

$\pi_i^t(h^t)$ depends only on h^t through the s^0 and s^t; so $\pi_i^t(h^t) = \pi_i^t(s^0,s^t)$.

A Markov strategy for player i is a semi-Markov strategy for which $\pi_i^t(s^0,s^t)$ does not depend on s^0; so $\pi_i^t(s^0,s^t) = \pi_i^t(s^t)$.

For each initial state s^0 and each set of strategies $\pi_1,\ldots,\pi_N$ for the players the game yields a stochastic process with rewards for the N players. Because for each player there will be realized a sequence of payments we have to specify a criterion. In the discounted game the criterion for player i will be

$$V_i(s^0,\pi_1,\ldots,\pi_N) = \limsup_{t'\to\infty} \sum_{t=0}^{t'} \beta^t V_i^t(s^0,\pi_1,\ldots,\pi_N)$$

or

$$\liminf_{t'\to\infty} \sum_{t=0}^{t'} \beta^t V_i^t(s^0,\pi_1,\ldots,\pi_N)$$

or any convex linear combination of lim sup and lin inf; where $V_i^t(s^0,\pi_1,\ldots,\pi_N)$ is the expected payment to player i at time t and $\beta \in [0,1)$ the discount factor. In the game with average return criterion:

$$V_i(s^0,\pi_1,\ldots,\pi_N) = \limsup_{t'\to\infty} \frac{1}{t'} \sum_{t=0}^{t'} V_i^t(s^0,\pi_1,\ldots,\pi_N)$$

or lim inf or any convex linear combination of lim sup and lim inf.

For $\varepsilon \geq 0$ an ε-equilibrium point of strategies given the criterion is a set of strategies for the players: $\pi_1^*,\ldots,\pi_N^*$ such that:

$$V_i(s^0,\pi_1^*,\ldots,\pi_{i-1}^*,\pi_i,\pi_{i+1}^*,\ldots,\pi_N^*) \leq$$

$$V_i(s^0,\pi_1^*,\ldots,\pi_N^*) + \varepsilon \qquad \text{for all strategies } \pi_i$$

for player i, for all players i and for all inital states s^0.

An 0-equilibrium point is called an equilibrium point.

Using the approach of DERMAN and STRAUCH [3] in the Markov decision process (one person stochastic game), we investigate whether the players can restrict themselves to semi-Markov strategies.

2. TWO PERSON ZERO SUM STOCHASTIC GAMES

We will call the game a two person zero sum game if $N = 2$ and $V_1(s^0,\pi_1,\pi_2) = - V_2(s^0,\pi_1,\pi_2)$ for all s^0,π_1 and π_2. If the limit in the definition of V_i always exists, $r_1(s,a_1,a_2) = - r_2(s,a_1,a_2)$ for all s,a_1 and a_2 is sufficient for the game to be zero sum. In general this is not true.

EXAMPLE 1.
State space: $\{1,2,\ldots\}$; in each state both players have only 1 action; if the state at time t is s then the state at time t+1 is s+1 with probability 1; $r_1(s,1,1) = - r_2(s,1,1) = (-2)^s$.
The game is discounted with $\beta = \frac{1}{2}$, we take the lim sup for both players.

$$V_1(1,\pi_1,\pi_2) = \limsup_{t'\to\infty} \sum_{t=0}^{t'} (\tfrac{1}{2})^t(-2)^{t+1} = 0$$

$$V_2(1,\pi_1,\pi_2) = \limsup_{t'\to\infty} \sum_{t=0}^{t'} (\tfrac{1}{2})^t(-2)^{t+1} = 2$$

EXAMPLE 2.
The game has one state where both players have 2 actions; whatever the actions chosen the game returns to the state with probability 1. in the next period; $r,(1,1,1) = - r_2(1,1,1) = 1$, $r_1(1,2,2) = - r_2(1,2,2) = -1$ all other rewards being zero. In symbolic notation:

$$\Gamma : \begin{bmatrix} 1 + \Gamma & \Gamma \\ \Gamma & -1 + \Gamma \end{bmatrix}$$

We consider the average return criterion with lim sup for both players. By cooperation both players can get an average reward 1; for example by playing n^n times action 1 followed by $(n+1)^{n+1}$ times action 2 etc.

LEMMA. *If for the two person zero sum game there exists an* ε*-equilibrium point* $\pi_1^\varepsilon,\pi_2^\varepsilon$ *for each* $\varepsilon > 0$ *then the game is strictly determined and the value of the game is* $\lim_{\varepsilon\downarrow 0} V_1(s^0,\pi_1^\varepsilon,\pi_2^\varepsilon)$ *for any criterion.*

PROOF. Since $\pi_1^\varepsilon,\pi_2^\varepsilon$ is an ε-equilibrium point. We have:

$$V_1(s^0,\pi_1,\pi_2^\varepsilon) - \varepsilon \le V_1(s0,\pi_1^\varepsilon,\pi_2^\varepsilon) \le V_1(s^0,\pi_1^\varepsilon,\pi_2) + \varepsilon.$$

Let $\varepsilon_1,\varepsilon_2,\ldots$ be a sequence of non-negative numbers such that $\lim_{i\to\infty} \varepsilon_i = 0$ then:

$$V_1(s^0,\pi_1^{\varepsilon_j},\pi_2^{\varepsilon_j})-\varepsilon_i-\varepsilon_j \le V_1(s^0,\pi_1^{\varepsilon_j},\pi_2^{\varepsilon_i})-\varepsilon_i \le V_1(s^0,\pi_1^{\varepsilon_i},\pi_2^{\varepsilon_i}) \le$$

$$V_1(s^0,\pi_1^{\varepsilon_i},\pi_2^{\varepsilon_j})+\varepsilon_i \le V_1(s^0,\pi_1^{\varepsilon_j},\pi_2^{\varepsilon_j}) + \varepsilon_i + \varepsilon_j,$$

$$\Rightarrow |V_1(s^0,\pi_1^{\varepsilon_i},\pi_2^{\varepsilon_i}) - V_1(s^0,\pi_1^{\varepsilon_j},\pi_2^{\varepsilon_j})| \le \varepsilon_i + \varepsilon_j$$

so the sequence $V_1(s^0,\pi_1^{\varepsilon_i},\pi_2^{\varepsilon_i})$ converges and $V(s^0) = \lim_{\varepsilon\downarrow 0} V_1(s^0,\pi_1^\varepsilon,\pi_2^\varepsilon)$ exists.

For each $\varepsilon > 0$ there exists a $\delta \in (0,\frac{1}{2}\varepsilon)$ such that

$$|V_1(s^0,\pi_1^\delta,\pi_2^\delta) - V(s^0)| \le \tfrac{1}{2}\varepsilon \Rightarrow$$

$$V_1(s^0,\pi_1^\delta,\pi_2) \ge V_1(s^0,\pi_1^\delta,\pi_2^\delta) - \tfrac{1}{2}\varepsilon \ge V(s^0) - \varepsilon$$

and

$$V_1(s^0,\pi_1,\pi_2^\delta) \le V_1(s^0,\pi_1^\delta,\pi_2^\delta) + \tfrac{1}{2}\varepsilon \le V(s^0) + \varepsilon.$$

So π_1^δ and π_2^δ are ε-optimal strategies for player 1 and player 2 respectively and $V(s^0)$ is the value of the game. □

3. EQUILIBRIUM POINTS OF SEMI-MARKOV STRATEGIES

THEOREM 1. *Let* $\pi_1,\ldots,\pi_N$ *be a set of behavior strategies for the players* 1,...,N. *If* π_j *is a semi-Markov strategy for all* j ≠ i *then there exists a semi-Markov strategy* π_i^{SM} *for player* i *such that:*

$$V_k^t(s^0,\pi_1,\ldots,\pi_{i-1},\pi_i^{SM},\pi_{i+1},\ldots,\pi_N) = V_k^t(s^0,\pi_1,\ldots,\pi_N)$$

for all times t, *initial states* s^0 *and players* k.

PROOF. Given initial state s^0 and behavior strategies $\pi_1,\ldots,\pi_N$ let $\underline{s}^t$ be the random variable whose value is the state at time t and $\underline{a}_i^t$ the random variable whose value is the action chosen by player i at time t.

For each set of strategies for the players and each initial state we have a corresponding probability measure on the space of sequences of states and actions that can be realized. As σ-field structure for this space we take the σ-field generated by finite sequences of states and actions.

Let P_{s^0} denote the probability measure corresponding to $\pi_1,\ldots,\pi_N$ as strategies and s^0 as initial state.

$$P_{s^0}(\underline{a}_j^t=a_j^t\ \forall j;\underline{s}^t=s^t) =$$

$$P_{s^0}(\underline{a}_i^t=a_i^t|\underline{a}_j^t=a_j^t\ \forall j\neq i;\underline{s}^t=s^t)\cdot P_{s^0}(\underline{a}_j^t=a_j^t\ \forall j\neq i;\underline{s}^t=s^t).$$

Since π_j for all $j \neq i$ are semi-Markov strategies the random variables $\underline{a}_i^t$ and $\underline{a}_j^t$, given s^0 and s^t with $j \neq i$ are independent, so

$$P_{s^0}(\underline{a}_i^t=a_i^t|\underline{a}_j^t=a_j^t\ \forall j\neq i;\underline{s}^t=s^t) = P_{s^0}(\underline{a}_i^t=a_i^t|\underline{s}^t=s^t).$$

$$\Rightarrow P_{s^0}(\underline{a}_j^t=a_j^t\ \forall j;\underline{s}^t=s^t) = P_{s^0}(\underline{a}_i^t=a_i^t|\underline{s}^t=s^t)\circ P_{s^0}(\underline{a}_j^t=a_j^t\ \forall j\neq i;\underline{s}^t=s^t)\quad (*)$$

Define π_i^{SM} as follows: if initial state is s^0 and the state at time t is s^t then choose action a_i^t with probability $P_{s^0}(\underline{a}_i^t=a_i^t|\underline{s}^t=s^t)$.

Let $P^*_{s^0}$ denote the probability measure on the sequences of states and actions if player i switches his strategy to π_i^{SM}.

We will show by induction with respect to t that

$$P^*_{s^0}(\underline{a}_j^t=a_j^t\ \forall j;\underline{s}^t=s^t) = P_{s^0}(\underline{a}_j^t=a_j^t\ \forall j;\underline{s}^t=s^t).$$

This equality is easily checked for t = 0; suppose it holds for t = T then

$$P_{s^0}(\underline{s}^{T+1}=s^{T+1}) =$$

$$\sum_{s^T,a_1^T,\ldots a_N^T} P_{s^0}(\underline{a}_j^T=a_j^T\ \forall j;\underline{s}^T=s^T)\ q(s^{T+1}|s^T,a_1^T,\ldots,a_N^T) =$$

$$\sum_{s^T,a_1^T,\ldots,a_N^T} P_{s^0}(\underline{a}_j^T=a_j^T\ \forall j;\underline{s}^T=s^T)\ q(s^{T+1}|s^T,a_1^T,\ldots,a_N^T) =$$

$$P_{s^0}(\underline{s}^{T+1}=s^{T+1}).$$

Since the players $j \neq i$ play semi-Markov strategies we have

$$P^*_{s^0}(\underline{a}_j^{T+1}=a_j^{T+1}\ \forall j\neq i;\underline{s}^{T+1}=s^{T+1}) = P_{s^0}(\underline{a}_j^{T+1}=a_j^T\ \forall j\neq 1;\underline{s}^{T+1}=s^{T+1}).$$

The equality for $t = T + 1$ then follows from the definition of π_i^{SM} and equality (*).
Since

$$V_k^t(s^0,\pi_1,\ldots,\pi_N) =$$

$$\sum_{a_1^t,\ldots,a_N^t,s^t} r_k(s^t,a_1^t,\ldots,a_N^t)\cdot P_{s^0}(\underline{a}_j^t=a_j^t\ \forall j;\underline{s}^t=s^t)$$

this proves the theorem. □

THEOREM 2. *If for any criterion $\pi_1^*,\ldots,\pi_N^*$ is an ε-equilibrium-point in the game where all players are restricted to play semi-Markov strategies then $\pi_1^*,\ldots,\pi_N^*$ is also an ε-equilibrium point for that criterion.*

PROOF. $V_i(s^0,\pi_1,\ldots,\pi_N)$ is some function of the $V_i^t(s^0,\pi_1,\ldots,\pi_N)$, $t = 1,2,\ldots$
By theorem 1 for each behavior strategy π_i there exists a semi-Markov strategy π_i^{SM} such that:

$$V_i(s^0,\pi_1^*,\ldots,\pi_{i-1}^*,\pi_i,\pi_{i+1}^*,\ldots,\pi_N^*) =$$

$$V_i(s^0,\pi_1^*,\ldots,\pi_{i-1}^*,\pi_i^{SM},\pi_{i+1}^*,\ldots,\pi_N^* \qquad \text{for all } s^0.$$

while

$$V_i(s^0,\pi_1^*,\ldots,\pi_{i-1}^*,\pi_i^{SM},\pi_i^*,\ldots,\pi_N^*) \leq$$

$$V_i(s^0,\pi_1^*,\ldots,\pi_N^*) + \varepsilon$$

for all s^0. therefore $\pi_1^*,\ldots,\pi_N^*$ is an ε-equilibrium point. □

However the existence of an ε-equilibrium point does not imply the existence of an ε-equilibrium point in the restricted game. The following example is a two person zero sum game that is strictly determined and whose restricted game is not.

EXAMPLE 3. This example is due to GILETTE [4] and BLACKWELL and FERGUSON [2] showed that starting in state 1 the game is strictly determined with value ½. Blackwell and Ferguson called this game "the big match"; we write it in symbolic notation:

$$\Gamma_1 : \begin{bmatrix} 1 + \Gamma_1 & \Gamma_1 \\ \Gamma_2 & 1 + \Gamma_3 \end{bmatrix}$$

$$\Gamma_2 : \quad [\Gamma_2]$$

$$\Gamma_3 : \quad [1 + \Gamma_3]$$

The stochastic game has state space: {1,2,3}; in state 1 both players have action space: {1,2}; in state 2 and 3 both players have action space: {1}. If in state 1 both players choose action 1 then one unit is payed by player 2 to player 1 and the next state is state 1 with probability 1. etc. If the game is in state 2 or 3 both players have only one action available and the game stays forever in that same state. We consider the average return criterion with lim sup for player 1 and lim inf for player 2.

In this example the set of semi-Markov strategies is the same as the set of Markov strategies. Blackwell and Ferguson used non-Markov strategies for player 1, dependent on the actions taken by player 2 in the past, to show that the game starting in state 1 is strictly determined. However if the players stick to (semi-)Markov strategies the game is not strictly determined. Stochastic games where the players are restricted to semi-Markov strategies can be considered as repeated games with incomplete information. ZAMIR [8] gives an equivalent example. We show that player 1 has no ε-optimal strategies for $\varepsilon < \frac{1}{2}$.

<u>PROOF</u>. Let $\pi = (\pi^1, \pi^2, \ldots)$ be a Markov strategy for player 1 that is ε-optimal (π^t is the probability of choosing action 1 at time t); p^t the probability that player 1 chooses action 2 for the first time at time t and $p = \sum_{t=1}^{\infty} p^t$ the probability that player 1 not always chooses action 1.
For each $\delta > 0$ there exists a t^0 such that : $\sum_{t=1}^{t^0} p^t \geq p-\delta$. We construct a strategy ρ for player 2 as follows: choose action 1 at time $1,\ldots,t^0$ and action 2 thereafter. If player 1 plays π and player 2 plays ρ the game reduces to a stochastic process that realizes exactly one of the following events:

1. player 1 uses action 2 before time t^0+1
2. player 1 uses action 2 for the first time at t^0+1 or thereafter
3. player 1 never uses action 2.

The probability that the first event occurs is at least $p-\delta$ and the average return in this case is 0. The second event has probability at most δ and average return 1. The third event has probability $1-p$ and average return 0. So the overall average return is at most δ.

The value of the restricted game, if it exists, is the same as the value of "the big match" by theorem 2 and the lemma. If $\varepsilon < \frac{1}{2}$ then choose $\delta < \frac{1}{2}-\varepsilon$; this contradicts the fact that π is an ε-optimal strategy for player 1. □

REFERENCES

[1] AUMANN, R.J., *Mixed and behavior strategies in infinite extensive games*, <u>in</u> Advances in game theory, pp 627-650, M. Dresher, L.S. Shapley and A.W. Tucker eds., Princeton university press (1964).

2 BLACKWELL, D. & T.S. FERGUSON, *The big match*, Annals of Math. Stat. 39, pp 159-163 (1968).

3 DERMAN, C. & R.E. STRAUCH, *A note on memoryless rules for controlling sequential control processes*, Annals of Math. Stat. 37, pp 276-278 (1966).

4 GILETTE, D., *Stochastic games with zero stop probabilities*, in Contributions to the theory of games 3, pp 179-187, M. Dresher, A.W. Tucker and P. Wolfe eds., Princeton university press (1957).

5 ROGERS, Ph.D., *Non zerosum stochastic games*, Berckeley ORC 69-8, april 1969.

6 SHAPLEY, L.S., *Stochastic games*, Proc. Nat. Acad. Sci. U.S.A. 39, pp 327-332 (1958).

7 SOBEL, M.J., *Noncooperative stochastic game*, Annals of Math. Stat. 42, pp 1930-1935 (1971).

8 ZAMIR, S., *On the notion of value for games with infinitely many stages*, The Annals of Statistics 1, pp 791-796 (1973).