Math Meth Oper Res (2002) 56:315–322

Mathematical Methods
of Operations Research

Optimality in different strategy classes in zero-sum stochastic games

J. Flesch, F. Thuijsman, O. J. Vrieze

Department of Mathematics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, the Netherlands (e-mail: frank@math.unimaas.nl)

Manuscript received: December 2001/Final version received: March 2002

Abstract. We present a *complete* picture of the relationship between the existence of 0-optimal strategies and ε-optimal strategies, $\varepsilon > 0$, in the classes of stationary, Markov and history dependent strategies.

Key words: zero-sum stochastic games, optimality, stationary strategies, Markov strategies, average rewards

1 Introduction

A zero-sum stochastic game Γ can be described by a state space $S := \{1, \dots, z\}$, and a corresponding collection $\{M_1, \dots, M_z\}$ of matrices, where matrix M_s has size $m_s^1 \times m_s^2$ and, for $i \in I_s := \{1, \dots, m_s^1\}$ and $j \in J_s := \{1, \dots, m_s^2\}$, entry (i, j) of M_s consists of a payoff $r_s(i, j) \in \mathbb{R}$ and a probability vector $p_s(i, j) = (p_s(t|i, j))_{t \in S}$. The elements of S are called states and for each state $s \in S$ the elements of I_s and J_s are called (pure) actions of player 1 and player 2 in state s. The game is to be played at stages in $\mathbb{N}$ in the following way. The play starts at stage 1 in an initial state, say in state $s^1 \in S$, where, simultaneously and independently, both players are to choose an action: player 1 chooses a row $i^1 \in I_{s^1}$, while player 2 chooses a column $j^1 \in J_{s^1}$. These choices induce an immediate payoff $r_{s^1}(i^1, j^1)$ from player 2 to player 1. Next, the play moves to a new state according to the probability vector $p_{s^1}(i^1, j^1)$, say to state s^2. At stage 2 new actions $i^2 \in I_{s^2}$ and $j^2 \in J_{s^2}$ are to be chosen by the players in state s^2. Then player 1 receives payoff $r_{s^2}(i^2, j^2)$ from player 2 and the play moves to some state s^3 according to the probability vector $p_{s^2}(i^2, j^2)$, and so on.

The sequence $h^n = (s^1, i^1, j^1; \dots; s^n, i^n, j^n)$ is called the history upto stage n. The players are assumed to have complete information and perfect recall.

A mixed action for a player in state s is a probability distribution on the set of his actions in state s. Mixed actions in state s will be denoted by x_s for player 1 and by y_s for player 2, and the sets of mixed actions in state s by X_s and Y_s respectively. A (history dependent) strategy π for player 1 is a decision rule that prescribes a mixed action $\pi_s(h)$ in the present state s for any past history h of the play. For player 2, (history dependent) strategies σ are defined similarly. We use the notations Π and Σ for the respective (history dependent) strategy spaces of the players. If the mixed actions prescribed by a strategy only depend on the current stage and state then the strategy is called Markov, while if they only depend on the current state then the strategy is called stationary. Thus the stationary strategy spaces are $X := \times_{s \in S} X_s$ for player 1 and $Y := \times_{s \in S} Y_s$ for player 2; while the Markov strategy spaces are $F := \times_{n \in \mathbb{N}} X$ for player 1 and $G := \times_{n \in \mathbb{N}} Y$ for player 2. We will use the respective notations x and y for stationary strategies and f and g for Markov strategies for players 1 and 2.

A pair of strategies (π, σ) together with an initial state $s \in S$ determine a stochastic process on the payoffs. The sequences of payoffs are evaluated by the average reward, given by

$$\gamma_s(\pi, \sigma) := \liminf_{N \to \infty} \mathbb{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N} r_n\right),$$

where r_n denotes the random variable for the payoff at stage n.

For any initial state $s \in S$, it is in the spirit of the game to evaluate a strategy π of player 1 or a strategy σ of player 2 by the rewards $\phi_s(\pi)$ and $\psi_s(\sigma)$, respectively, that π and σ guarantee when starting in s; so let

$$\phi_s(\pi) := \inf_{\sigma \in \Sigma} \gamma_s(\pi, \sigma), \quad \psi_s(\sigma) := \sup_{\pi \in \Pi} \gamma_s(\pi, \sigma).$$

Mertens and Neyman (1981) showed that

$$\sup_{\pi \in \Pi} \phi_s(\pi) = \inf_{\sigma \in \Sigma} \psi_s(\sigma) =: v_s \quad \forall s \in S.$$

Here $v := (v_s)_{s \in S}$ is called the average value of the game. A strategy π of player 1 is called ε-optimal $\varepsilon \geq 0$, if

$$\phi_s(\pi) \geq v_s - \varepsilon \quad \forall s \in S.$$

Similarly, a strategy σ of player 2 is called ε-optimal $\varepsilon \geq 0$, if

$$\psi_s(\sigma) \leq v_s + \varepsilon \quad \forall s \in S.$$

Because of the definition of the value v, both players have ε-optimal strategies for all $\varepsilon > 0$, but, generally, 0-optimal strategies need not exist. This is completely different from the situation in finite state Markov decision problems, where stationary 0-optimal strategies allways exist. In stochastic games even stationary ε-optimal strategies may fail to exist and history dependent strategies, where players have to respond to the behavior of the opponent, are generally indispensable for achieving ε-optimality.

2 The relationship

The goal of this paper is to present a *complete* picture of the relationship between the existence of 0-optimal strategies and ε-optimal strategies, $\varepsilon > 0$, in the classes of stationary, Markov and history dependent strategies. More precisely, we present the following relationship, which shall be proven in sections 2.1 and 2.2.

Theorem 1. *Between the existence of 0-optimal and ε-optimal strategies, $\varepsilon > 0$, in the classes of stationary (S), Markov (M) and history dependent (H) strategies, the following relations apply:*

$$\boxed{S0 \Rightarrow M0 \Leftrightarrow H0 \Rightarrow S\varepsilon \Rightarrow M\varepsilon \Rightarrow H\varepsilon}$$

and none of the one-sided implications can be reversed.

So for example, $H0 \Rightarrow S\varepsilon$ means that the existence of history dependent 0-optimal strategies implies the existence of stationary ε-optimal strategies for all $\varepsilon > 0$, but it also means that the existence of stationary ε-optimal strategies does not imply the existence of history dependent 0-optimal strategies in general.

Thus, in the above picture, implications always hold from left to right, but never from right to left except for $H0 \Rightarrow M0$. For example, $S0 \Rightarrow H\varepsilon$ or $H0 \Rightarrow M\varepsilon$ are both true (by transitivity), but $S\varepsilon \Rightarrow M0$ or $M\varepsilon \Rightarrow S0$ are both generally false.

Despite the simple structure of the picture above, some implications are far from straightforward (for instance $H0 \Rightarrow S\varepsilon$) and some counterexamples are fairly subtle (especially $H\varepsilon \not\Rightarrow M\varepsilon$, but also $M\varepsilon \not\Rightarrow S\varepsilon$).

2.1 The implications

The implications

$$S0 \Rightarrow M0, \quad M0 \Rightarrow H0, \quad S\varepsilon \Rightarrow M\varepsilon, \quad M\varepsilon \Rightarrow H\varepsilon$$

follow from the inclusions of the strategy classes. The other implications

$$H0 \Rightarrow M0, \quad H0 \Rightarrow S\varepsilon$$

are far from straightforward and have been shown in Flesch et al. (1998).

2.2 The counterexamples

In the examples below, we will only indicate the non-absorbing states (states which the play can leave for at least one pair of actions). Moreover, we will assume that, in each absorbing state, each player has only one action. Note that if the play moves to an absorbing state s (absorption occurs in state s)

then the play is strategically over and the average reward will equal the payoff in state s. Since the absorbing states are trivial, we will always assume that the initial state is one of the non-absorbing ones.

- **Example 1:** $M0 \not\Rightarrow S0$

	L	R
T	0	1 *
B	1	0

1

This is a game with one non-absorbing state, state 1. The actions of player 1 are the rows (T for Top and B for Bottom) and the actions of player 2 are the columns (L for Left and R for Right). The payoffs are placed in the upper-left corners of the entries. If one of the entries (T, L), (B, L) or (B, R) is chosen then the play remains in state 1, while entry (T, R) yields absorption (indicated by $*$) with absorbing payoff 1.

One can check that the value for initial state 1 is $v_1 = 1$. Indeed, the stationary strategy $x_\varepsilon = (1 - \varepsilon, \varepsilon)$ of player 1 guarantees $1 - \varepsilon$ for initial state 1, namely $\phi_1(x_\varepsilon) = 1 - \varepsilon$. Since the highest payoff of the game is 1, we must have $v_1 = 1$.

Note that player 1 has no stationary optimal strategy for initial state 1 in this game. One can argue as follows. If a stationary strategy x prescribes action T with a positive probability then x only gives a reward strictly less than 1 if player 2 always chooses action L. On the other hand, if x chooses action B with probability 1, then if player 2 always takes action R then the reward is 0. Thus no stationary strategy can guarantee $v_1 = 1$.

Nevertheless, a Markov optimal strategy can be constructed as follows: let f be the Markov strategy that, at stage n, chooses action T with probability $1/n$ and action B with probability $1 - 1/n$. One can verify that f is optimal. We only give an intuitive argument. If player 2 chooses action R with a "positive frequency" then absorption occurs with probability 1 due to the slowly decreasing probabilities on action T; while almost always choosing action L yields reward 1 since the probabilities on action B converge to 1.

- **Example 2:** $S\varepsilon \not\Rightarrow H0$

	L	R
T	1 *	0
B	0 *	1 *

1

The notation is similar to that of example 1.

Here the value for initial state 1 is $v_1 = 1$, and the stationary strategy $x = (1 - \varepsilon, \varepsilon)$ for player 1, which prescribes action T with probability $1 - \varepsilon$ and action B with probability ε, is ε-optimal for all $\varepsilon > 0$.

However, we show that player 1 does not have optimal strategies for initial state 1. Take an arbitrary strategy π. We show that player 2 can make sure that player 1's reward is strictly less than 1. Indeed, player 2 has to choose action R as long as π prescribes action T with probability 1 and to play action L at the first stage when π prescribes action B with a positive probability. Then either entry (T, R) is played forever or absorption occurs in entry (B, L) with payoff zero with a positive probability, thus player 1's reward is strictly less than 1 indeed.

- **Example 3:** $M\varepsilon \nRightarrow S\varepsilon$

$$
\begin{array}{c|c|c|}
 & L_1 & R_1 \\ \hline
T_1 & \begin{array}{ll} 1 & \\ & \end{array} & \begin{array}{ll} 0 & \\ & \end{array} \\ \hline
B_1 & \begin{array}{ll} 1 & \\ & 2 \end{array} & \begin{array}{ll} 1 & \\ & * \end{array} \\ \hline
\multicolumn{3}{c}{1}
\end{array}
\qquad
\begin{array}{c|c|c|}
 & L_2 & R_2 \\ \hline
T_2 & \begin{array}{ll} 1 & \\ & 1 \end{array} & \begin{array}{ll} 0 & \\ & \end{array} \\ \hline
B_2 & \begin{array}{ll} 0 & \\ & * \end{array} & \begin{array}{ll} 1 & \\ & * \end{array} \\ \hline
\multicolumn{3}{c}{2}
\end{array}
$$

The notation is similar to that of example 1. Here entries (B_1, L_1) and (T_2, L_2) lead to the other non-absorbing state.

This game, which has been analyzed in Flesch et al. (1997), has the following properties for initial states 1 and 2:

(a) The value is $v_1 = v_2 = 1$.

(b) Player 1 has Markov ε-optimal strategies for initial states 1 and 2, for all $\varepsilon > 0$. Indeed, define a Markov strategy f^K for player 1, where $K \in \mathbb{N}$, as follows:

$$u^K(n) := \sqrt[K]{\frac{n}{n+1}} \quad \text{for all } n \in \mathbb{N},$$

$$f^K := [(u^K(n), 1 - u^K(n)), (u^K(n), 1 - u^K(n))]_{n \in \mathbb{N}}.$$

Observe that the Markov strategy f^K is symmetric in the sense that the prescribed mixed actions in state 1 and state 2 are the same for any stage n. Note that the sequence $u^K(n)$ "slowly" converges to 1 as n tends to infinity, so f^K assigns less and less probabilities to actions B_1 and B_2.

For initial states 1 and 2, for all $\varepsilon > 0$, if $K \in \mathbb{N}$ is large then player 1 can guarantee a reward at least $1 - \varepsilon$ by playing the Markov strategy f^K, namely

$$\phi_1(f^K) \geq 1 - \varepsilon, \quad \phi_2(f^K) \geq 1 - \varepsilon.$$

(c) Player 1 has no stationary ε-optimal strategy for initial states 1 and 2, if $\varepsilon \in [0, 1)$. In fact, player 1 can get at most 0 for initial states 1 and 2 by playing stationary strategies, namely

$$\phi_1(x) = \phi_2(x) = 0 \quad \forall x.$$

Note that (b) implies (a), because the highest payoff in the game is 1.

Now we briefly explain (b). The question here is how player 2 can reply to the strategy f^K. Intuitively, player 2 has two hopes to decrease player 1's reward. The first one is achieving absorption in entry (B_2, L_2) with payoff 0. Player 2's best candidate would be playing actions L_1 and L_2 whenever the play is in state 1 or in state 2. But then whenever the play is in state 2 a transition occurs to state 1 with a large probability, and it takes a long time until the play comes back to state 2 again. Because the strategy f^K assigns decreasing probabilities to action B_1, the lengths of stay in state 1 will increase fast during the play and the frequency of visits to state 2 will tend to zero. As a consequence, the frequency of stages when absorption could occur is zero (in the limit) and the probabilities on action B_2 at those stages will decrease "rapidly". Therefore, the overall probability of absorption in entry (B_2, L_2) will be small. In conclusion, playing L_1 and L_2 gives player 2 little hope.

On the other hand, since the payoffs in entries (T_1, R_1) and (T_2, R_2) equal 0, player 2 could try to play actions R_1 and R_2 with a "positive" frequency and hope that the play will not absorb. But in that case, the frequency of stages when absorption could occur is positive and the probabilities on B_1 and B_2 at those stages decrease "slowly". Hence, it will appear that the play must eventually absorb with probability 1, and then the zero payoffs in entries (T_1, R_1) and (T_2, R_2) will have no influence on the reward.

Finally, we discuss (c). For each stationary strategy $x = (x_1, x_2)$ of player 1, we define a strategy $y^x = (y_1^x, y_2^x)$ for player 2: let

$$y_1^x := \begin{cases} (0,1) & \text{if } x_1 = (1,0) \\ (1,0) & \text{otherwise} \end{cases}, \quad y_2^x := \begin{cases} (0,1) & \text{if } x_2 = (1,0) \\ (1,0) & \text{otherwise.} \end{cases}$$

Notice that, for $s = 1, 2$, we have $\gamma_s(x, y^x) = 0$ for all x, so the proof of (c) is complete.

- **Example 4:** $H\varepsilon \not\Rightarrow M\varepsilon$

	L	R
T	0	1
B	1 *	0 *

1

The notation is similar to that of example 1.

This game is the famous Big Match introduced in Gillette (1957), which was only solved 11 years later by Blackwell & Ferguson (1968). The beauty of the Big Match is that the structure of the game is so simple. Action T keeps the play in state 1 with probability 1, while action B leads to an absorbing state, so it ends the game in a strategic sense. Player 1's trouble is that if he uses action B then the entry of absorption fully depends on the action chosen by player 2.

The Big Match has the following properties:

(a) The value for initial state 1 equals $v_1 = 1/2$.
(b) Player 2 has a stationary optimal strategy $y = (1/2, 1/2)$.
(c) For $N \in \mathbb{N}$, let π^N be the strategy for player 1 which, for present state 1 and any past history h, prescribes action T with probability $1 - (k(h) + N)^{-2}$ and action B with probability $(k(h) + N)^{-2}$, where $k(h)$ denotes the number of stages where player 2 has chosen action R minus the number of stages where player 2 has chosen action L with respect to the history h.

    Then for any $\varepsilon > 0$, if $N \in \mathbb{N}$ is at least $\frac{1}{2\varepsilon}$, then the strategy π^N is an ε-optimal strategy for player 1.
(d) Player 1 has no Markov ε-optimal strategy for initial state 1, if $\varepsilon \in \left[0, \frac{1}{2}\right)$. In fact, player 1 can only guarantee reward 0 by Markov strategies, namely

$$\phi_1(f) = 0 \quad \forall f.$$

The proofs of (a), (b), and (c) can be done by showing that, for initial state 1, player 2 can guarantee 1/2 by playing $y = (1/2, 1/2)$, and player 1 can guarantee $1/2 - \varepsilon$ by the strategies in (c) for any $\varepsilon > 0$.

It is easy to verify that, for initial state 1, the strategy $y = (1/2, 1/2)$ guarantees 1/2 for player 2. In fact, regardless of the strategy that player 1 uses against y, the reward always equals 1/2, since the expected payoff equals 1/2 for each stage.

For any $\varepsilon > 0$, the strategies in (c) have been found and have been shown to guarantee $1/2 - \varepsilon$ for initial state 1, by Blackwell & Ferguson (1968). Notice that this strategy is rather complex and player 1 has to make use of the whole past history of the play when choosing his actions.

Finally, we explain (d). Take an arbitrary Markov strategy f for player 1. Let $\rho^n(f)$ denote the overall probability that absorption occurs at any of the stages $n+1, n+2, n+3, \ldots$ with respect to f when the initial state is state 1 (clearly, this probability is indepedent of the strategy used by player 2, due to the fact that f is a Markov strategy and the transition structure of the game). Since the probability that absorption occurs up to stage n converges to $\rho^0(f)$ as n tends to infinity, we have $\rho^0(f) = \lim_{n\to\infty}(\rho^0(f) - \rho^n(f))$, hence $\lim_{n\to\infty} \rho^n(f) = 0$. Let $\varepsilon > 0$ be arbitrary. Then there exists a stage N such that $\rho^N(f) \leq \varepsilon$. Now consider the Markov strategy g for player 2 which prescribes action R up to stage N and action L for all further stages. Then, with respect to the (f, g), the following 3 events can occur:

(i) Absorption takes place in entry (B, R) at some stage in $1, \ldots, N$;
(ii) Absorption takes place in entry (B, L) at some stage in $N+1, N+2, \ldots$;
(iii) No absorption occurs at all, and entry (T, L) is played at all stages in $N+1, N+2, \ldots$

By the choice of N, event (ii) has probability at most ε, therefore $\gamma_1(f, g) \leq \varepsilon$. As $\varepsilon > 0$ was arbitrary, we have shown (d).

3 Concluding Remarks

Instead of examining optimality for the infinite horizon game, one could alternatively study optimality for games with a (possibly unknown) finite, but sufficiently long, horizon. This yields uniform (ε-)optimality, i.e., for player 1 a strategy π is uniform optimal for state $s \in S$ if

$$\forall \delta > 0 \; \exists N^{\delta} : \mathbb{E}_{s\pi\sigma}\left(\frac{1}{N}\sum_{n=1}^{N} r_n\right) \geq v_s - \delta \quad \forall N \geq N^{\delta}, \; \forall \sigma \in \Sigma.$$

The definition of uniform ε-optimality is similar. The result of Mertens & Neyman (1981) also applies for the existence of uniform ε-optimal strategies. Also, any stationary (ε-)optimal strategy is necessarilly uniform (ε-)optimal as is shown by Bewley & Kohlberg (1978). We would like to emphasize that Theorem 1 is also valid for uniform (ε-)optimality; all proofs and counterexamples still apply.

Bewley & Kohlberg (1978) give necessary and sufficient conditions for the existence of stationary optimal strategies by using Puiseux expansions for the λ-discounted value. The λ-discounted reward ($\lambda \in (0, 1)$) is defined by

$$\gamma_s^{\lambda}(\pi, \sigma) := \mathbb{E}_{s\pi\sigma}\left(\sum_{n=1}^{\infty} \lambda(1-\lambda)^{n-1} r_n\right).$$

For λ close to 0, the λ-discounted value can be expanded as a Puiseux series, i.e. a Laurent series with fractional powers of λ.

A characterization for the existence of stationary (ε-)optimal strategies in terms of mathematical programming is provided by Filar et al. (1991).

For (ε-)optimality in terms of Markov strategies we refer to Flesch et al. (1997).

4 References

Bewley T, Kohlberg E (1978) On stochastic games with stationary optimal strategies. Mathematics of Operations Research 3:104–125

Blackwell D, Ferguson TS (1968) The big match. Annals of Mathematical Statistics 39:159–163

Filar JA, Schultz TA, Thuijsman F, Vrieze OJ (1991) Nonlinear programming and stationary equilibria in stochastic games. Mathematical Programming 50:227–237

Flesch J, Thuijsman F, Vrieze OJ (1997) Markov strategies are better than stationary strategies. International Game Theory Review 1:9–31

Flesch J, Thuijsman F, Vrieze OJ (1998) Simplifying optimal strategies in stochastic games. SIAM Journal of Control and Optimization 36(4):1331–1347

Gillette D (1957) Stochastic games with zero stop probabilities. In: Dresher M, Tucker AW, Wolfe P (eds.), Contributions to the Theory of Games III, Annals of Mathematical Studies 39, Princeton University Press, pp. 179–187

Mertens JF, Neyman A (1981) Stochastic games. International Journal of Game Theory 10:53–66