Improving strategies in stochastic games

J. Flesch, F. Thuijsman, O.J. Vrieze
Department of Mathematics, Maastricht University
P.O. Box 616, 6200 MD Maastricht, The Netherlands
frank@math.unimaas.nl

Abstract

In a zero-sum limiting average stochastic game, we evaluate a strategy π for the maximizing player, player 1, by the reward $\phi_s(\pi)$ that π guarantees to him when starting in state s. A strategy π is called non-improving if $\phi_s(\pi) \geq \phi_s(\pi[h])$ for any state s and for any finite history h, where $\pi[h]$ is the strategy π conditional on the history h; otherwise the strategy is called improving. We investigate the use of improving and non-improving strategies, and explore the relation between (non-)improvingness and (ε-)optimality. Improving strategies appear to play a very important role for obtaining ε-optimality, while 0-optimal strategies are always non-improving. Several examples will clarify all these issues.

1 Introduction

We deal with zero-sum stochastic games with finite state and action spaces. These games model conflict situations in which two players are involved with completely opposite interests. Such a game Γ can be described by a state space $S := \{1, \ldots, z\}$, and a corresponding collection $\{M_1, \ldots, M_z\}$ of matrices, where matrix M_s has size $m_s^1 \times m_s^2$ and, for $i_s \in I_s := \{1, \ldots, m_s^1\}$ and $j_s \in J_s := \{1, \ldots, m_s^2\}$, entry (i_s, j_s) of M_s consists of a payoff $r_s(i_s, j_s) \in \mathbb{R}$ and a probability vector $p_s(i_s, j_s) = (p_s(t|i_s, j_s))_{t \in S}$. The elements of S are called states and for each state $s \in S$ the elements of I_s and J_s are called 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 an $i_{s^1}^1 \in I_{s^1}$, while player 2 chooses a $j_{s^1}^1 \in J_{s^1}$. These choices induce an immediate payoff $r_{s^1}(i_{s^1}^1, j_{s^1}^1)$ from player 2 to player 1. Next, the play moves to a new state according to the probability vector $p_{s^1}(i_{s^1}^1, j_{s^1}^1)$, say to state s^2. At stage 2 new actions $i_{s^2}^2 \in I_{s^2}$ and $j_{s^2}^2 \in J_{s^2}$ are to be chosen by the players in state s^2. Then player 1 receives payoff $r_{s^2}(i_{s^2}^2, j_{s^2}^2)$ from player 2 and the play moves to some state s^3 according to the probability vector $p_{s^2}(i_{s^2}^2, j_{s^2}^2)$, etc.

The sequence $(s^1, i_{s^1}^1, j_{s^1}^1; \ldots; s^n, i_{s^n}^n, j_{s^n}^n)$ is called the history up to 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. A strategy is a decision rule that prescribes a mixed action for any history of play. Such general strategies, so-called history dependent strategies, will be denoted by π for player 1 and by σ for player 2, and $\pi_s(h)$ and $\sigma_s(h)$ will denote the mixed actions for present state s and history h. 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. We will use the respective notations x and y for stationary strategies and f and g for Markov strategies for players 1 and 2.

For a strategy π and a history h, we can also define the strategy $\pi[h]$ which prescribes a mixed action $\pi_s[h](\bar{h})$ for each history $\bar{h}$ and present state s, as if h had happened before $\bar{h}$, i.e., $\pi_s[h](\bar{h}) = \pi_s(h\bar{h})$, where $h\bar{h}$ is the history consisting of h concatenated with $\bar{h}$.

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

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

where r_n is the random variable for the payoff at stage $n \in N$, and R_N for the average payoff up to stage N. In [4] it is shown that

$$\sup_\pi \inf_\sigma \gamma_s(\pi,\sigma) = \inf_\sigma \sup_\pi \gamma_s(\pi,\sigma) =: v_s \qquad \forall s \in S,$$

where $v := (v_s)_{s \in S}$ is called the limiting average value. A strategy π of player 1 is called optimal for initial state $s \in S$ if $\gamma_s(\pi, \sigma) \geq v_s$ for all σ, and is called ε-optimal for initial state $s \in S$, $\varepsilon > 0$, if $\gamma_s(\pi, \sigma) \geq v_s - \varepsilon$ for all σ. If a strategy of player 1 is optimal or ε-optimal for all initial states in S then the strategy is called optimal or ε-optimal respectively. Optimality for

0-7803-4394-8/98 $10.00 (c) 1998 IEEE

strategies of player 2 is analogously defined. Although for all $\varepsilon > 0$, by the definition of the value, there exist ε-optimal strategies for both players, the Big Match, a famous example introduced in [3] and analyzed in [1], demonstrates that in general the players need not have optimal strategies and for achieving ε-optimality history dependent strategies are indispensable.

An alternative, well known, evaluation criterion is that of β-discounted rewards, with $\beta \in [0, 1)$, defined for a pair of strategies (π, σ) with initial state $s \in S$ by:

$$\gamma_{\beta s}(\pi, \sigma) = \mathbb{E}_{s\pi\sigma}\left((1-\beta)\sum_{n=1}^{\infty}\beta^{n-1} r_n\right),$$

where r_n is as above. The β-discounted value and β-discounted optimality are defined as for limiting average rewards, and in [5] the β-discounted value v_β and stationary β-discounted optimal strategies are shown to exist.

2 Preliminaries

In zero-sum games the players have completely opposite interests, so it is natural to evaluate a strategy π of player 1 by the reward $\phi(\pi)$ it guarantees against any strategy of the opponent. For a strategy π let

$$\phi_s(\pi) := \inf_\sigma \gamma_s(\pi, \sigma) \quad \forall s \in S, \qquad \phi(\pi) := (\phi_s(\pi))_{s \in S}.$$

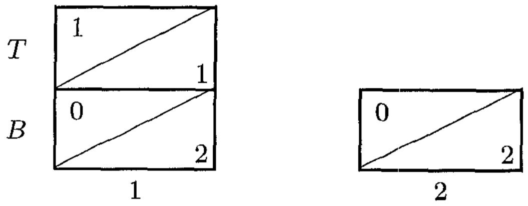
Using this evaluation ϕ we may naturally define the relation "ε-better" between strategies of player 1. A strategy π^1 is called ε-better than π^2, where $\varepsilon \geq 0$, if $\phi_s(\pi^1) \geq \phi_s(\pi^2) - \varepsilon$ holds for all $s \in S$. 0-better strategies will be simply called better. We will call a strategy π *non-improving* if for any state $s \in S$ and for any history h with final state s we have

$$\phi_s(\pi) \geq \phi_s(\pi[h]);$$

otherwise π is called *improving*. Intuitively, a non-improving strategy, for any state, cannot guarantee a larger reward conditional on any past history than initially. On the other hand, improving strategies may become better during the play than initially.

For example, all stationary strategies are clearly non-improving strategies, because $x = x[h]$ for any history h. In the following simple example we show an instance of an improving strategy.

Example 1:



Here player 1 chooses rows and player 2 chooses columns. Notice that player 2 has no influence on the play at all, as he has only one action in both states. In each entry, the corresponding payoff is placed in the up-left corner, while the transition is placed in the bottom-right corner. In this game each transition is represented by the number of the state to which transition should occur with probability 1. Notice that state 2 is absorbing, i.e., if the play visits state 2 then it stays there forever. Since player 1 has only one action in state 2, strategies for player 1 only need to be defined in state 1. Consider the Markov strategy f for player 1 which prescribes to play action T with probability 1/2 and action B with probability 1/2 at stage 1, and if the play does not absorb then to play action T at all further stages. Clearly, f yields reward 1/2, hence we obtain $\phi_1(f) = 1/2$. However, if h^* denotes the history up to stage 1 when player 1 chooses action T at stage 1, then the strategy $f[h^*]$ prescribes action T for each stage, hence $\phi_1(f[h^*]) = 1$. Thus $\phi_1(f) < \phi_1(f[h^*])$, which means that f is improving.

3 Results

In this paper the main result is given in theorem 5, which, verbally and less detailed, can be presented as:

Main Theorem *In any zero-sum stochastic game, for any non-improving strategy, there exists an ε-better stationary strategy, for any $\varepsilon > 0$, and there exists a better Markov strategy as well.*

The above theorem says, that, surprisingly, non-improving strategies are not more effective than stationary strategies or Markov strategies. This also means that, instead of using a complex history dependent non-improving strategy, the player could also use a simple stationary strategy which guarantees at least the same reward up to some arbitrarily small $\varepsilon > 0$, or he could even achieve the same reward by employing a Markov strategy.

Notice that optimal strategies are always non-improving, since they guarantee the value and no higher reward can be guaranteed by the definition of the value. Using this observation the above result can be seen as a generalization of the following theorem, which is proved in [2]:

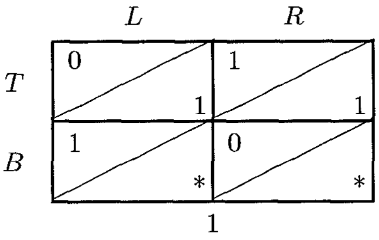
Theorem 1 *In any zero-sum stochastic game, if player 1 has an optimal strategy, then he also has a stationary ε-optimal strategy, for any $\varepsilon > 0$, and a Markov optimal strategy as well.*

The above theorem and the Main Theorem together have the following corollary. This shows the insufficiency of the class of non-improving strategies as well as the indispensability of improving strategies for achieving ε-optimality, for small $\varepsilon > 0$.

Corollary 2 *In a zero-sum stochastic game, if a player has no stationary ε-optimal strategies for small $\varepsilon > 0$, then he neither has optimal strategies and all his ε-optimal strategies, with small $\varepsilon > 0$, are improving.*

The next example, the Big Match (cf. [1]), provides an illustration for the above corollary.

Example 2:



The notation is the same as in example 1. In fact, this example is a 3-state game in which states 2 and 3 are absorbing, i.e. once play reaches such a state, play remains there forever. State 2 has a payoff 1 to player 1 and is reached (with transition probability 1) from state 1 by playing (B, L); state 3 has payoff 0 for player 1 and is reached from state 1 by playing (B, R). For initial state 1, the limiting average value is $v_1 = \frac{1}{2}$ and player 1 is known to have neither optimal strategies nor stationary ε-optimal strategies for small $\varepsilon > 0$. But for any $N \in \mathbb{N}$ player 1 can guarantee $\frac{1}{2} - \frac{1}{2(N+1)}$ by playing the following strategy π^N: for any history h without absorption, if $k(h)$ denotes the number of stages where player 2 has chosen action R minus the number of stages where player 2 has chosen action L, player 1 has to play the mixed action:

$$\pi^N(h) := \left(1 - \frac{1}{(k(h)+N+1)^2}, \frac{1}{(k(h)+N+1)^2}\right).$$

These results can all be found in [1]. Clearly, the latter strategy π^N is improving, since for the history $h = (1, T, R)$ we have $\pi^N[h] = \pi^{N+1}$.

4 Proof

First we introduce some more notations. Let π denote a fixed non-improving strategy and let

$$a := \phi(\pi).$$

For $x_s \in X_s$, $y_s \in Y_s$ for some $s \in S$ let

$$A_s(x_s, y_s) := \sum_{t \in S} p_s(t|x_s, y_s)\, a_t,$$

where

$$p_s(t|x_s, y_s) = \sum_{i_s, j_s} x_s(i_s)\, y_s(j_s)\, p_s(t|i_s, j_s).$$

For $x \in X$ and $y \in Y$ let

$$A(x, y) := (A_s(x_s, y_s))_{s \in S}.$$

For all $s \in S$ let

$$\begin{aligned} \tilde{X}_s \ &: \ = \{x_s \in X_s \,|\, A_s(x_s, y_s) \geq a_s \ \ \forall y_s \in Y_s\} \\ \tilde{X} \ &: \ = \times_{s \in S} \tilde{X}_s, \end{aligned}$$

so $\tilde{X}_s$ is the set of mixed actions of player 1 in state s which assure that after transition a will not decrease in expectation.

Lemma 3 *The sets $\tilde{X}_s$, $s \in S$, are nonempty polytopes.*

Proof: Let $s \in S$. One can verify that the linearity of A_s in both components implies that the set $\tilde{X}_s$ is a polytope. Now we prove that $\tilde{X}_s$ is nonempty by showing that $\pi_s \in \tilde{X}_s$, where π_s denotes the mixed action prescribed by π for stage 1 if the initial state is state s. By the definition of $\phi_s(\pi)$

$$\begin{aligned} \phi_s(\pi) \ = \ & \min_{y_s \in Y_s} \sum_{t \in S} \sum_{\substack{i_s \in I_s \\ j_s \in J_s}} \pi_s(i_s)\, y_s(j_s)\, p_s(t|i_s, j_s) \cdot \\ & \cdot \phi_t(\pi[s, i_s, j_s]), \end{aligned}$$

hence using the definition of a and the non-improvingness of π we have

$$\begin{aligned} a_s \ &= \ \phi_s(\pi) \\ &= \ \min_{y_s \in Y_s} \sum_{t \in S} \sum_{\substack{i_s \in I_s \\ j_s \in J_s}} \pi_s(i_s)\, y_s(j_s)\, p_s(t|i_s, j_s) \cdot \\ & \quad \cdot \phi_t(\pi[s, i_s, j_s]) \\ &\leq \ \min_{y_s \in Y_s} \sum_{t \in S} \sum_{\substack{i_s \in I_s \\ j_s \in J_s}} \pi_s(i_s)\, y_s(j_s)\, p_s(t|i_s, j_s)\, \phi_t(\pi) \\ &= \ \min_{y_s \in Y_s} \sum_{t \in S} p_s(t|\pi_s, y_s) \cdot a_t = \min_{y_s \in Y_s} A_s(\pi_s, y_s), \end{aligned}$$

so the proof is complete. □

If Z is a polytope then $\mathrm{Relint}(Z)$ denotes the relative interior of the polytope Z, which is defined as the set of points in Z which can be written as a convex combination of all the extreme points of Z with only strictly positive coefficients.

The following technical lemma is needed later for the construction of a restricted game. Here, on condition that player 1 uses a strategy $x \in \mathrm{Relint}(\tilde{X})$, we are looking for the largest set S' of states which can be made recurrent and the largest sets Y_s', $s \in S'$, of mixed actions which keep all the states in S' recurrent.

Lemma 4 *There exist a nonempty $S' \subset S$ and a nonempty $Y' = \times_{s \in S'} Y_s'$, where $Y_s' \subset Y_s$ are polytopes for all $s \in S'$, such that for any $x \in \mathrm{Relint}(\tilde{X})$*

(a) *for any $y \in Y$, if $s \in S$ is recurrent with respect to (x,y) then $s \in S'$ and $y_s \in Y_s'$;*

(b) *for any $y \in Y$ with $y_s \in \mathrm{Relint}(Y_s')$ for all $s \in S'$, all states $s \in S'$ are recurrent with respect to (x,y).*

Proof: Take an arbitrary $x \in \mathrm{Relint}(\tilde{X})$. For $j \in J$, let $R(j)$ denote the set of recurrent states with respect to (x,j). Now let

$$S' := \cup_{j \in J} R(j).$$

For $s \in S'$ let

$$J_s' := \{j_s \in J_s | \exists \bar{j} \in J : \bar{j}_s = j_s,\ s \in R(\bar{j})\},$$

$$Y_s' := \mathrm{conv}\ \{J_s'\}, \quad Y' := \times_{s \in S'} Y_s',$$

where conv stands for the convex hull of a set. Note that these sets are independent of the choice of $x \in \mathrm{Relint}(\tilde{X})$, because all $x \in \mathrm{Relint}(\tilde{X})$ put positive probabilities on the same actions in any state. It is not hard to check that S' and Y' satisfy the required properties. □

Recall that we have fixed a non-improving strategy π for player 1. Let $\tilde{X}$ be as above, let S' and Y' be as in lemma 4, and let $X' := \times_{s \in S'} \tilde{X}_s$. In view of lemma 4, we may define a restricted stochastic game Γ' in the following way. Let Γ' be the game, derived from Γ, where the state space is S' and the players are restricted to use strategies that only prescribe mixed actions in X_s' and Y_s' if the play is in any state $s \in S'$. Clearly, X' and Y' are respective stationary strategy spaces in Γ' for the players.

By the finiteness of the state and action spaces, there exists a countable subset of discount factors $\mathcal{B} \subset (0,1)$ such that 1 is a limit point of $\mathcal{B}$ and there are stationary β-discounted optimal strategies $x_\beta \in X'$ in the restricted game Γ' such that the sets $\{i_s \in I_s | x_{\beta s}(i_s) > 0\}$, $s \in S$, are independent of $\beta \in \mathcal{B}$. In the sequel each time that we are dealing with discount factors, discounted optimal strategies, or with limits when the discount factors converge to 1, we will have such a subset of discount factors $\mathcal{B}$ in mind.

Theorem 5 *Let $\pi \in \Pi$ be a non-improving strategy in a zero-sum stochastic game. By using the strategy π, define S', X', Y', and the restricted game Γ' as above.*

(a) *For any $\beta \in \mathcal{B}$, let $x_\beta \in X'$ be a β-discounted optimal strategy in the restricted game Γ' and let $x \in \mathrm{Relint}(\tilde{X})$. Then, for any $\varepsilon > 0$, if $\beta \in \mathcal{B}$, $\tau \in (0,1)$ are sufficiently large then the stationary strategy $x_\beta^\tau \in \tilde{X}$, given for state $s \in S$ by*

$$x_{\beta s}^\tau := \begin{cases} \tau \cdot x_{\beta s} + (1-\tau) \cdot x_s & \text{if } s \in S' \\ x_s & \text{if } s \in S \setminus S' \end{cases},$$

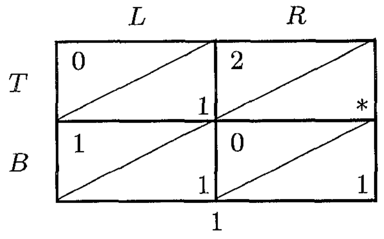
is ε-better than π in Γ.

(b) *Let ε_n, $n \in \mathbb{N}$, be an arbitrary monotonously decreasing sequence converging to 0. Let the stationary strategy $x_n \in X'$ be ε_n-better than π for all $n \in \mathbb{N}$. Then there exists a sequence K_n in $\mathbb{N}$ such that the Markov strategy f which prescribes to play x_1 for the first K_1 stages, then to play x_2 for the next K_2 stages, etc., is better than π.*

A similar statement holds for player 2 as well.

To illustrate this theorem we present the following example in which we focus on optimal stratgies as non-improving strategies.

Example 3:



The value for the only non-trivial initial state 1 is $v_1 = 1$. It is not hard to show that there are optimal strategies for player 1 (later we will construct optimal

Markov strategies). Therefore we have $S' = S$ for this game.

Following the construction for stationary ε-optimal strategies, we have $X' = X$, $Y' = \{(1,0)\}$. Now the unique β-discounted optimal strategy of player 1 in Γ' is $x_\beta = (0,1)$ for all $\beta \in (0,1)$. The role of x_β is to play well as long as player 2 plays in the restricted game Γ', namely to guarantee the value v as long as player 2 chooses action L in state 1. However, an enforcement is needed to make sure that player 2 is not better off by playing outside Y', namely by choosing action R. Therefore we take a strategy $x \in \text{Relint}(\tilde{X})$, for example $x = (1/2, 1/2)$, which will force player 2 not to choose action R, since then R leads to absorption with payoff 2. Now for $\tau, \beta \in (0,1)$ we have

$$x_\beta^\tau = \tau \cdot x_\beta + (1-\tau) \cdot x = (1/2 - \tau/2, 1/2 + \tau/2).$$

The strategy x_β^τ is ε-optimal for large τ and β indeed, as the stationary strategies $(p, 1-p)$ are ε-optimal for all $p \in (0, \varepsilon]$.

Note that player 1 has no stationary optimal strategy 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 then 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 $v = 1$.

A Markov optimal strategy can be constructed as above. The idea is to increase β and τ simultaneously during the play so that player 1 plays better and better in the restricted game. However, τ must be increased sufficiently slowly so that player 2 cannot choose action R "too often" without absorption. Formally, let $\varepsilon_n = 1/n$ and take the stationary ε_n-optimal strategy $x_n = (\varepsilon_n, 1 - \varepsilon_n) \in X'$ for all $n \in \mathbb{N}$. Let $K_n = 1$ for all $n \in \mathbb{N}$. Let f be the Markov strategy as in theorem 5. So at stage n, the strategy f 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.

We now provide a proof for theorem 5. Recall that we have fixed a non-improving strategy π. In the restricted game Γ', let H' denote the set of finite histories, Π' and Σ' the sets of history dependent strategies, γ' the limiting avarage reward, v'_β the β-discounted value for all $\beta \in (0,1)$. Let $v' := \lim_{\beta\uparrow 1} v'_\beta$ (here "limit" is understood to be taken for a sequence of β's). Also, let

$$\bar{\Pi} := \{\pi \in \Pi |\, \pi_s(h) \in X'_s \quad \text{for all } s \in S' \text{ and } h \in H'\}$$

$$\bar{\Sigma} := \{\sigma \in \Sigma |\, \sigma_s(h) \in Y'_s \quad \text{for all } s \in S' \text{ and } h \in H'\};$$

so $\bar{\Pi}$ and $\bar{\Sigma}$ are the set of strategies in the original game Γ with the property that, as long as the play is in the restricted game Γ', they behave as strategies in Π' and Σ'.

By using the definition of $\tilde{X}$, the following lemma is straightforward.

Lemma 6 *Let $x \in \tilde{X}$ and $y \in Y$. Suppose E is an ergodic set with respect to (x,y). Then $a_s = a_t$ for all $s,t \in E$.*

Next, we show an important property of the sets X'_s, Y'_s, $s \in S'$.

Lemma 7 *For any $s \in S'$, we have that $A_s(x_s, y_s) = a_s$ for all $x_s \in X'_s$ and $y_s \in Y'_s$.*

Proof: Take arbitrary $s \in S'$, $x_s \in X'_s$, and $y_s \in Y'_s$. Let $\bar{x} \in \text{Relint}(\tilde{X})$ and $\bar{y} \in Y$ with $\bar{y}_t \in \text{Relint}(Y'_t)$ for all $t \in S'$. In view of lemma 4-(b), state s belongs to an ergodic set E with regard to $(\bar{x}, \bar{y})$, hence by lemma 6, we obtain $a_t = a_w$ for all $t, w \in E$. As $p_s(t|\bar{x}_s, \bar{y}_s) > 0$ implies $t \in E$, we must have $p_s(t|x_s, y_s) > 0$ also implies $t \in E$, which completes the proof. □

Lemma 8 *Let $s \in S'$ be an arbitrary initial state and let H_s be the set of histories starting in s. Also, let*

$$U_s := \{(h,t) \in H_s \times S |\, \mathcal{P}_{s\pi\sigma}(h) > 0 \quad \textit{and}$$

$$\mathcal{P}_{s\pi\sigma}(t|h) > 0 \quad \textit{for some } \sigma \in \bar{\Sigma}\},$$

where $\mathcal{P}_{s\pi\sigma}(t|h)$ is the probability that, with respect to (π, σ), the new state becomes state t after history h. Then $\pi_t(h) \in X'_t$ for all $(h,t) \in U_s$.

Proof: Suppose the opposite. Then there exists a shortest history $\bar{h}^n \in H_s$, say up to stage n, and a state t such that $\mathcal{P}_{s\pi\sigma}(\bar{h}^n) > 0$ and $\mathcal{P}_{s\pi\sigma}(t|\bar{h}^n) > 0$ for some $\sigma \in \bar{\Sigma}$ and $\pi_t(\bar{h}^n) \notin X'_t$. Since $\pi_t(\bar{h}^n) \notin X'_t$ there exists a $\bar{y}_t \in Y_t$ such that

$$\tau := a_t - A_t(\pi_t(\bar{h}^n), \bar{y}_t) > 0.$$

For any present state $z \in S'$ and past history $h \in H'$, we define a mixed action $\xi_z(h) \in Y_z$ as follows: if $\pi_z(h) \in X'_z$ then let $\xi_z(h) \in Y'_z$; while if $\pi_z(h) \in X_z \setminus$

X'_z then let $\xi_z(h) \in Y_z$ such that $A_z(\pi_z(h), \xi_z(h)) \leq a_z$. By lemma 7, we have in both cases that

$$A_z(\pi_z(h), \xi_z(h)) \leq a_z.$$

Let

$$\delta \in \left(0, \mathcal{P}_{s\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s\pi\sigma}(t|\bar{h}^n) \cdot \tau\right).$$

Let $s^1 := s$, and let s^m, $m \geq 2$, denote the random variable for the state at stage m, and let θ^m denote random variable for the history up to stage $m \in \mathbb{N}$.

Let $\sigma^\delta \in \Sigma$ be the strategy that prescribes to play as follows: play σ during the first n stages; at stage $n+1$, if $\theta^n = \bar{h}^n$ and $s^{n+1} = t$ then play $\bar{y}_t$ while if $\theta^n \neq \bar{h}^n$ or $s^{n+1} \neq t$ then play the mixed action $\phi_{s^{n+1}}(\theta^n)$; and finally, play a δ-best reply against $\pi[\theta^{n+1}]$ from stage $n+2$ on. Note that

$$\mathcal{P}_{s^1\pi\sigma^\delta}(\bar{h}^n) = \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) > 0.$$

Since we have chosen a shortest history $\bar{h}^n$ with the above property, the play up to stage n has been going in the restricted game Γ'. By lemma 4-(b), we must have $s^n \in S'$, and by the definitions of X' and Y', we obtain in expectation

$$\mathcal{E}_{s^1\pi\sigma^\delta}\left(a_{s^{n+1}}\right) = a_{s^1}.$$

The choices of the used mixed actions at stage $n+1$ imply

$$\begin{aligned}\mathcal{E}_{s^1\pi\sigma^\delta}\left(a_{s^{n+2}}\right) \leq\ & \mathcal{E}_{s^1\pi\sigma^\delta}\left(a_{s^{n+1}}\right) - \\ & \mathcal{P}_{s^1\pi\sigma^\delta}(\bar{h}^n) \cdot \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) \cdot \tau.\end{aligned}$$

Since from stage $n+2$ player 2 plays a δ-best reply and π is non-improving, the choice of δ yields

$$\begin{aligned}\gamma_{s^1}(\pi, \sigma^\delta) \leq\ & \sum_{\substack{h^{n+1} \in H^{n+1} \\ z \in S}} \mathcal{P}_{s^1\pi\sigma^\delta}(h^{n+1})\, \mathcal{P}_{s^1\pi\sigma}(z|h^{n+1}) \\ & \gamma_z(\pi[h^{n+1}], \sigma^\delta[h^{n+1}]) \\ \leq\ & \sum_{\substack{h^{n+1} \in H^{n+1} \\ z \in S}} \mathcal{P}_{s^1\pi\sigma^\delta}(h^{n+1})\, \mathcal{P}_{s^1\pi\sigma}(z|h^{n+1}) \\ & (a_z + \delta) \\ =\ & \mathcal{E}_{s^1\pi\sigma^\delta}\left(a_{s^{n+2}}\right) + \delta \\ \leq\ & \mathcal{E}_{s^1\pi\sigma^\delta}\left(a_{s^{n+1}}\right) \\ & -\mathcal{P}_{s^1\pi\sigma^\delta}(\bar{h}^n)\, \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n)\, \tau + \delta \\ =\ & a_{s^1} - \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n)\, \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n)\, \tau + \delta \\ <\ & a_{s^1},\end{aligned}$$

which contradicts the definition of a. □

The next result follows similarly to lemma 2.3 in [2].

Lemma 9 *We have*

$$v_s \leq \sup_{\pi' \in \Pi'} \inf_{\sigma' \in \Sigma'} \gamma_s(\pi', \sigma') \leq v'_s \qquad \forall s \in S.$$

Proof of theorem 5: By using the above lemmas, the proof is almost the same as the proof of theorem 1 in [2]. Note that lemma 4-(a) is needed for achieving the following crucial property: if a pure stationary strategy $j \in J$ is a best reply to some x^T_β, then, in any ergodic set with respect to (x^T_β, j), the play is in fact taking place in the restricted game Γ'. □

5 References

[1] D. Blackwell & T.S. Ferguson [1968]: "The big match", *Annals of Mathematical Statistics* 33, 159-163.

[2] J. Flesch, F. Thuijsman, O.J. Vrieze [1998]: "Simplifying optimal strategies in stochastic games", *SIAM Journal on Control and Optimization* 36, 1331-1347.

[3] D. Gillette [1957]: "Stochastic games with zero stop probabilities", In: M. Dresher, A.W. Tucker & P. Wolfe (eds.), Contributions to the Theory of Games III, *Annals of Mathematical Studies* 39, Princeton University Press, 179-187.

[4] J.F. Mertens & A. Neyman [1981]: "Stochastic games", *International Journal of Game Theory* 10, 53-66.

[5] L.S. Shapley [1953]: "Stochastic games", *Proceedings of the National Academy of Sciences USA* 39, 1095-1100.