

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
<b>2</b>	<b>The stochastic game model</b>	<b>7</b>
2.1	The game and the rules . . . . .	7
2.2	Strategies . . . . .	8
2.2.1	Strategy classes . . . . .	8
2.2.2	Induced probability measures on the set of histories . . . . .	10
2.2.3	Induced stochastic processes on the set of states . . . . .	11
2.3	Rewards . . . . .	12
2.3.1	The average reward . . . . .	12
2.3.2	The discounted reward . . . . .	13
2.3.3	The rewards for stationary strategies . . . . .	13
2.4	Playing against a fixed strategy . . . . .	17
2.5	Zero-sum stochastic games and optimality . . . . .	18
2.6	General-sum stochastic games and equilibria . . . . .	22
2.7	Special classes of stochastic games . . . . .	24
<b>I</b>	<b>Zero-sum stochastic games</b>	<b>27</b>
<b>3</b>	<b>Simplifying optimal strategies</b>	<b>29</b>
3.1	Introduction . . . . .	29
3.2	Preliminaries . . . . .	30
3.3	The construction . . . . .	32
3.4	The proof . . . . .	34
3.5	Concluding remarks . . . . .	41
<b>4</b>	<b>Improving and non-improving strategies</b>	<b>45</b>
4.1	Introduction . . . . .	45
4.2	Preliminaries . . . . .	47
4.3	The construction . . . . .	49
4.4	The proof . . . . .	50
<b>5</b>	<b>Markov strategies are better</b>	<b>53</b>
5.1	Introduction . . . . .	53
5.2	An example where $\mathcal{A} < \mathcal{B}$ for some initial states . . . . .	54

5.3	Sufficient conditions for $\mathcal{A} = \mathcal{B}$ . . . . .	66
5.3.1	Repeated games with absorbing states . . . . .	66
5.3.2	Games with constant $\mathcal{A}$ or $\mathcal{B}$ . . . . .	68
5.3.3	Games with optimal strategies or with best-Markov strategies . . . . .	70
5.4	Concluding remarks . . . . .	71
5.5	Appendix . . . . .	71
<b>6</b>	<b>Almost stationary <math>\varepsilon</math>-equilibria</b>	<b>75</b>
6.1	Introduction . . . . .	75
6.2	Preliminaries . . . . .	76
6.3	The construction . . . . .	78
6.4	Examples . . . . .	83
<b>II</b>	<b>General-sum stochastic games</b>	<b>85</b>
<b>7</b>	<b>Recursive repeated games</b>	<b>87</b>
7.1	Introduction . . . . .	87
7.2	Preliminaries . . . . .	87
7.3	The construction . . . . .	91
7.4	The proof . . . . .	93
7.5	Concluding remarks . . . . .	95
<b>8</b>	<b>Average-discounted equilibria</b>	<b>97</b>
8.1	Introduction . . . . .	97
8.2	Stationary $\varepsilon$ -equilibria . . . . .	98
8.3	Ultimately stationary $\varepsilon$ -equilibria . . . . .	100
8.4	A game without average-discounted 0-equilibria . . . . .	101
8.5	Special classes of stochastic games . . . . .	103
8.6	Concluding remarks . . . . .	104
<b>9</b>	<b>More than two players</b>	<b>105</b>
9.1	Introduction . . . . .	105
9.2	A cyclic three-person game . . . . .	106
9.3	Concluding remarks . . . . .	116
<b>10</b>	<b>Appendix: uniform optimality and equilibria</b>	<b>117</b>
<b>11</b>	<b>References</b>	<b>123</b>

# Chapter 1

## Introduction

This monograph is devoted to the study of stochastic games, which can be seen as decision processes with a certain number of decision makers (players). In this monograph we will always assume that there are at least two players; stochastic games with only one player are better known as Markov decision processes and the theory on such games has developed in another direction. We will now describe stochastic games with two players for the sake of simplicity; the description of stochastic games with more players is analogous. A stochastic game with two players can be given by a state space  $S$ , and related to each state  $s \in S$ , a bimatrix  $A_s$  in which each entry contains two real numbers (payoffs to the respective players) and a probability vector (transition vector) over the state space  $S$ . The rows and the columns of bimatrix  $A_s$  represent the available decisions (actions) in state  $s$  for player 1 and player 2 respectively. The play of the game evolves at decision moments (stages) in  $\mathbb{N}$  as follows. The play starts at stage 1 in an initial state  $s \in S$ , where, simultaneously and independently, both players are to choose an action: player 1 has to choose a row of bimatrix  $A_s$  while player 2 has to choose a column of  $A_s$ . Then, each player receives his payoff corresponding to the entry determined by these choices, and next the play moves to a new state  $t \in S$ , according to the transition vector. In the new state  $t$  at stage 2, the players have to choose actions again, and just like before, depending on their choices, they receive the corresponding payoffs and the play moves to a new state again, and so on.

Note that the game is non-cooperative, meaning that the players are not allowed to make binding agreements. It is furthermore assumed that the players have complete information (they know the bimatrices) and have perfect recall about the past history of the play. Consequently, when the players have to choose actions in the current state, they may take the entire past history into account.

A plan which tells a player how to make his decisions during the play is called a strategy. Instead of choosing an action with probability 1, a strategy may as well prescribe to apply a probability distribution on the set of available actions (mixed action) for the selection. The most complex strategies are the history dependent strategies, which prescribe mixed actions in the current state depending on the past history of the play. If the prescribed mixed actions in the current state only depend on the current stage then the strategy is called a Markov strategy; while if the prescribed mixed actions are fixed for each state then the strategy is called stationary.

Thus, as a result of the play, each player obtains an infinite sequence of payoffs. These sequences need to be evaluated in some manner. We will mainly consider the average reward (simply referred to as reward) which uses the long term average payoffs as an evaluation. The goal of each player in the game is simply to maximize his own reward by means of applying an effective strategy.

Zero-sum stochastic games are special stochastic games with two players in which the two players have completely opposite interests, namely one player pays the other player and the gain of one player is the loss of the other player. We assume that player 1 is paid by player 2, hence player 1 is trying to maximize his own reward while player 2 aims to minimise player 1's reward (player 2's maximization of his own reward now coincides with the minimization of player 1's reward). It is fortunate to know that there is always a certain reward which satisfies the following properties: for any  $\varepsilon > 0$ , player 1 has a strategy that guarantees him at least this reward (up to this  $\varepsilon$ ) against any strategy of player 2, while there are available strategies for player 2 which ensure him of not needing to pay more than this reward (up to this  $\varepsilon$ ) regardless of player 1's strategy. This unique reward is called the value of the game and the above strategies are called  $\varepsilon$ -optimal strategies. Clearly, the value as a reward is an acceptable outcome of the game for both players as neither of them is able to force a better reward in his favour. It is an interesting fact that 0-optimal strategies do not necessarily exist and achieving  $\varepsilon$ -optimality can often only be possible by employing history dependent strategies.

Stochastic games (not necessarily with only two players) without the requirement that the players have completely opposite interests are called general-sum stochastic games. As, in these games, some players may as well have matching interests up to some extent, the concepts 'value' and 'optimality' are no longer applicable. Here the usual solution concept is that of  $\varepsilon$ -equilibria,  $\varepsilon > 0$ , which is a collection of strategies from the players with the property that no player can improve his own reward by more than  $\varepsilon$  if he unilaterally deviates to another strategy. Hence, for small  $\varepsilon$ , the rewards corresponding to an  $\varepsilon$ -equilibrium are an appealing solution of the stochastic game. It is known that 0-equilibria do not always exist and history dependent strategies are often indispensable for obtaining  $\varepsilon$ -equilibria. The existence of  $\varepsilon$ -equilibria in stochastic games, however, is not yet known and is the most challenging open problem in stochastic game theory these days, even though the existence problem has been answered in the affirmative for several special classes.

This monograph is structured as follows. Chapter 2 describes the stochastic game model in detail and provides a summary of the most important results. This is followed by several chapters on zero-sum stochastic games. Chapter 3 deals with possible simplifications of 0-optimal strategies: we show that the existence of 0-optimal strategies implies the existence of stationary  $\varepsilon$ -optimal strategies and Markov 0-optimal strategies. This means that it is not necessary to play complex history dependent 0-optimal strategies, whenever they exist, as stationary and Markov strategies are equally effective. In chapter 4, we extend the results to possible simplifications of so-called nonimproving strategies. Next, a thorough analysis on the comparison of the effectiveness of stationary strategies and Markov strategies follows in chapter 5. We present an interesting game which demonstrates the advantage of applying Markov strategies, however, we also provide several conditions under which the two classes of

strategies perform equally well. Chapter 6 deals with the structure of  $\varepsilon$ -equilibria in zero-sum stochastic games (the concept of equilibria is also applicable for zero-sum stochastic games, as they are special general-sum stochastic games). In the second part of this monograph we turn our attention to general-sum stochastic games. We only consider games with only two players; games with more players are treated in chapter 9. In chapter 7, we show the existence of stationary  $\varepsilon$ -equilibria under conditions on the payoff and the transition structure. Next, chapter 8 provides some results on the existence of equilibria when player 1 is still interested in the average reward, but player 2 uses the so-called discounted reward. Chapter 9 is devoted to the extension of the results in stochastic games with more than two players. Finally, the appendix deals with some other important evaluations similar to the average reward. Furthermore, some important issues are discussed regarding the stochastic game model.



## Chapter 2

# The stochastic game model

In this chapter we restrict the description of the model and the discussion of the most important issues to two-person stochastic games. Extensions of the model and the results to  $K$ -person stochastic games are treated in chapter 9.

### 2.1 The game and the rules

**Definition 1** *A two-person stochastic game  $\Gamma$  is a tuple*

$$\langle S, (I_s)_{s \in S}, (J_s)_{s \in S}, (r_s^1)_{s \in S}, (r_s^2)_{s \in S}, (p_s)_{s \in S} \rangle,$$

where

- $S$  is a nonempty and finite set, called the state space;
- $I_s$  is a nonempty and finite set, called the action space for player 1 in state  $s \in S$ ;
- $J_s$  is a nonempty and finite set, called the action space for player 2 in state  $s \in S$ ;
- $r_s^k$  is a payoff function for player  $k \in \{1, 2\}$  in state  $s \in S$  assigning a real number  $r_s^k(i_s, j_s)$ , called payoff, to each action pair  $(i_s, j_s) \in I_s \times J_s$ ;
- $p_s$  is the transition map in state  $s \in S$  assigning a probability distribution on the state space  $p_s(i_s, j_s) = (p_s(t|i_s, j_s))_{t \in S}$ , called transition vector, to each action pair  $(i_s, j_s) \in I_s \times J_s$ .

In the sequel, whenever we talk about stochastic games we will have two-person stochastic games in mind, unless mentioned otherwise.

In view of the above definition, a two-person stochastic game can be represented as a collection of bimatrices  $\{\text{Bimatrix}(s) : s \in S\}$ , where entry  $(i_s, j_s)$  of  $\text{Bimatrix}(s)$  consists of the two corresponding payoffs  $r_s^k(i_s, j_s)$ ,  $k = 1, 2$ , and the corresponding transition vector  $p_s(i_s, j_s)$ , and is given as

$$\begin{array}{c}
r_s^1(i_s, j_s), r_s^2(i_s, j_s) \\
\\
\\
\\
\\
p_s(i_s, j_s)
\end{array}$$

When the transition vector  $p_s(i_s, j_s)$  has a component  $p_s(t|i_s, j_s)$  equal to 1, for some  $t \in S$ , then the transition vector  $p_s(i_s, j_s)$  shall be abbreviated by  $t$ . Further abbreviations are explained later with the help of examples 11 and 17.

In any state  $s \in S$ , in this bimatrix representation, the actions of player 1 are simply the rows and the actions of player 2 are simply the columns of Bimatrix( $s$ ).

The play of the game evolves at stages in  $\mathbb{N}$  as follows. 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  $r_{s^1}^1(i_{s^1}^1, j_{s^1}^1)$  to player 1 and an immediate payoff  $r_{s^1}^2(i_{s^1}^1, j_{s^1}^1)$  to player 2. Next, the play moves to a new state according to the transition vector  $p_{s^1}(i_{s^1}^1, j_{s^1}^1)$ , namely a transition occurs to state  $s^2 \in S$  with probability  $p_{s^1}(s^2|i_{s^1}^1, j_{s^1}^1)$ . At stage 2 in the new state  $s^2$ , 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. Afterwards the players receive the corresponding payoffs  $r_{s^2}^1(i_{s^2}^2, j_{s^2}^2)$  and  $r_{s^2}^2(i_{s^2}^2, j_{s^2}^2)$ , and the play moves to some state  $s^3$  according to the transition vector  $p_{s^2}(i_{s^2}^2, j_{s^2}^2)$  again, and so on.

It is assumed that the players know the structure of the game, namely the tuple in definition 1, and are aware of the present state and the past history  $(s^m, i_{s^m}^m, j_{s^m}^m)_{m=1}^{n-1}$  at any stage  $n$  of the play.

It should be noticed that, depending on the initial state, the stochastic game situation is different. However, it often appears useful to treat these games simultaneously.

## 2.2 Strategies

### 2.2.1 Strategy classes

#### Definition 2

- (a) A sequence  $h^n = (s^m, i_{s^m}^m, j_{s^m}^m)_{m=1}^n$ , with  $n \in \mathbb{N}$ , where  $s^m \in S$ ,  $i_{s^m}^m \in I_{s^m}$ ,  $j_{s^m}^m \in J_{s^m}$  for all  $m = 1, \dots, n$ , is called a history up to stage  $n$ . Here state  $s^1$  is called the initial state of history  $h^n$ . The initial history up to stage 0 is the empty sequence  $h^0 := ()$ . Let  $H^n$  denote the set of histories up to stage  $n$ , and let  $H^0 := \{h^0\}$ . Let  $H := \cup_{n=0}^{\infty} H^n$  denote the set of finite histories.
- (b) An infinite history is a sequence  $h^\infty = (s^n, i_{s^n}^n, j_{s^n}^n)_{n \in \mathbb{N}}$  where  $s^n \in S$ ,  $i_{s^n}^n \in I_{s^n}$ ,  $j_{s^n}^n \in J_{s^n}$  for all  $n \in \mathbb{N}$ . The set of infinite histories is denoted by  $H^\infty$ .
- (c) For  $s \in S$ , let  $H_s$  denote the set of finite histories with initial state  $s$ , and let  $H_s^\infty$  denote the set of infinite histories with initial state  $s$ .

When playing a stochastic game, the players may randomize over their actions, namely instead of choosing an action with probability 1, they may use a probability distribution on the action spaces. Such probability distributions are called mixed actions.

**Definition 3** *A mixed action  $x_s$  for player 1 in state  $s \in S$  is a probability distribution on  $I_s$ . A mixed action  $y_s$  for player 2 in state  $s \in S$  is a probability distribution on  $J_s$ . The respective sets of mixed actions in state  $s$  are denoted by  $X_s$  and  $Y_s$ .*

Note that the sets  $X_s$  and  $Y_s$ ,  $s \in S$ , are nonempty polytopes. Moreover, any action  $i_s \in I_s$ , in any state  $s \in S$ , can be naturally identified with the mixed action in  $X_s$  which puts probability 1 on  $i_s$ . Actions in  $J_s$  can be similarly identified with mixed actions in  $Y_s$ , in any state  $s \in S$ . By these identifications,  $I_s$  and  $J_s$  become the respective sets of extreme points of  $X_s$  and  $Y_s$ .

Next we define the most important classes of strategies.

**Definition 4**

- (a) *A (history dependent) strategy for player 1 is a map  $\pi$  assigning a mixed action  $\pi_s(h) \in X_s$  in any present state  $s \in S$  for any past history  $h \in H$ . A (history dependent) strategy for player 2 is a map  $\sigma$  assigning a mixed action  $\sigma_s(h) \in Y_s$  in any present state  $s \in S$  for any past history  $h \in H$ . For the sake of simplicity let  $\pi_s := \pi_s(h^0)$  and  $\sigma_s := \sigma_s(h^0)$  for all  $s \in S$ . The respective sets of history dependent strategies are denoted by  $\Pi$  and  $\Sigma$ .*
- (b) *A Markov strategy for player 1 is a map  $f$  assigning a mixed action  $f_s^n \in X_s$  in any present state  $s \in S$  and stage  $n \in \mathbb{N}$ . A Markov strategy for player 2 is a map  $g$  assigning a mixed action  $g_s^n \in Y_s$  in any present state  $s \in S$  and stage  $n \in \mathbb{N}$ . The respective sets of Markov strategies are denoted by  $F$  and  $G$ .*
- (c) *A stationary strategy for player 1 is a collection of mixed actions  $x = (x_s)_{s \in S}$  belonging to  $X := \times_{s \in S} X_s$ . A stationary strategy for player 2 is a collection of mixed actions  $y = (y_s)_{s \in S}$  belonging to  $Y := \times_{s \in S} Y_s$ . Here  $X$  and  $Y$  are called the spaces of stationary strategies, respectively.*

History dependent strategies are the most general strategies. When a player uses a history dependent strategy, he takes the entire past history into account when choosing his action in the present state. Markov strategies, however, have a substantially easier structure than history dependent strategies, since Markov strategies only consider the present state and present stage when prescribing a mixed action. A fundamental role in the analysis of stochastic games will be played by stationary strategies, where the prescribed mixed actions only depend on the present state. We wish to distinguish strategies which do not use randomization. These strategies always prescribe one action to be used with probability 1.

**Definition 5** *A strategy  $\pi \in \Pi$  for player 1 is called pure if  $\pi_s(h) \in I_s$  in any present state  $s \in S$  and for any past history  $h \in H$ . For player 2, pure strategies are defined analogously. Let  $\Pi^p$  and  $\Sigma^p$  denote the respective spaces of pure (history dependent)*

strategies,  $F^p$  and  $G^p$  the respective spaces of pure Markov strategies, and  $I$  and  $J$  the respective spaces of pure stationary strategies. Pure stationary strategies are denoted by  $i$  for player 1 and by  $j$  for player 2.

In fact,  $I = \times_{s \in S} I_s$  and  $J = \times_{s \in S} J_s$  and they are the respective sets of extreme points of the polytopes  $X$  and  $Y$ .

Finally, we define what we mean by a strategy conditional on a past history.

**Definition 6** *Let  $\pi \in \Pi$  and  $h \in H$ . The strategy  $\pi$  conditional on the history  $h$ , denoted by  $\pi[h]$ , is the strategy which prescribes a mixed action  $\pi_s[h](\bar{h})$  in any present state  $s \in S$  for any history  $\bar{h} \in H$  as if  $h$  had happened before  $\bar{h}$ , namely  $\pi_s[h](\bar{h}) = \pi_s(h \oplus \bar{h})$ , where  $h \oplus \bar{h}$  is the history consisting of  $h$  concatenated with  $\bar{h}$ . The definition is analogous for strategies of player 2.*

Notice that any strategy conditional on the initial history  $h^0$  is simply itself.

## 2.2.2 Induced probability measures on the set of histories

Take an initial state  $s \in S$  and a pair of (history dependent) strategies  $(\pi, \sigma)$ . In this section, we will consider probability measures that are induced by the triple  $(s, \pi, \sigma)$  on sets of histories; the measure theoretic concepts that we will use can be found in elementary books on measure theory. For  $n \in \mathbb{N}$ , consider the finite set  $H_s^n$  of histories up to stage  $n$ . Let  $\mathcal{M}_s^n$  denote the set of all subsets of  $H_s^n$ . The pair  $(H_s^n, \mathcal{M}_s^n)$  is a measurable space, on which the triple  $(s, \pi, \sigma)$  induces a probability measure  $\mathcal{P}_{s\pi\sigma}^n$  in a natural way. So, if  $U \in \mathcal{M}_s^n$  then  $\mathcal{P}_{s\pi\sigma}^n(U)$  gives the probability that the history up to stage  $n$  will belong to  $U$ . We have thus obtained a probability measure space  $(H_s^n, \mathcal{M}_s^n, \mathcal{P}_{s\pi\sigma}^n)$  which describes the play up to stage  $n$  in a probabilistic manner. For the sake of completeness, we may also consider the probability measure space  $(H_s^0, \mathcal{M}_s^0, \mathcal{P}_{s\pi\sigma}^0)$  where  $H_s^0 = \{h^0\}$ ,  $\mathcal{M}_s^0 = \{\emptyset, H_s^0\}$ ,  $\mathcal{P}_{s\pi\sigma}^0(h^0) = 1$ .

We will now define a probability measure space in order to be able to describe the infinite play in a probabilistic way with respect to  $(s, \pi, \sigma)$ . This probability measure space is generated by the family of probability measure spaces  $(H_s^n, \mathcal{M}_s^n, \mathcal{P}_{s\pi\sigma}^n)_{n \in \mathbb{N} \cup \{0\}}$  as follows. We may naturally identify each history  $h^n \in H_s^n$ , where  $n \in \mathbb{N} \cup \{0\}$ , with the set  $H_s^\infty[h^n]$  of infinite histories which coincide with  $h^n$  up to stage  $n$ . Similarly, we identify each  $U \subset H_s^n$  with the set

$$H_s^\infty[U] := \cup_{h^n \in U} H_s^\infty[h^n];$$

if  $U = \emptyset$  then  $H_s^\infty[U] := \emptyset$ . On basis of these identifications, we may define the following set: let

$$\mathcal{M}_s^\infty := \cup_{n \in \mathbb{N} \cup \{0\}} \mathcal{M}_s^n.$$

In fact, one can check that  $\mathcal{M}_s^\infty$  is an algebra of subsets of  $H_s^\infty$ . Let  $\mathcal{S}(\mathcal{M}_s^\infty)$  denote the sigma-algebra generated by the algebra  $\mathcal{M}_s^\infty$ . By Kolmogorov's extension theorem (cf. Kolmogorov [1933]), there is a unique probability measure  $\mathcal{P}_{s\pi\sigma}$  on the measurable space  $(H_s^\infty, \mathcal{S}(\mathcal{M}_s^\infty))$  with the consistency property that

$$\mathcal{P}_{s\pi\sigma}(U) = \mathcal{P}_{s\pi\sigma}^n(U) \quad \forall U \in \mathcal{M}_s^n, \forall n \in \mathbb{N} \cup \{0\}.$$

So we have obtained a probability measure space  $(H_s^\infty, \mathcal{S}(\mathcal{M}_s^\infty), \mathcal{P}_{s\pi\sigma})$  which is consistent with the family of probability measure spaces  $(H_s^n, \mathcal{M}_s^n, \mathcal{P}_{s\pi\sigma}^n)_{n \in \mathbb{N} \cup \{0\}}$  in the sense that the probability measure  $\mathcal{P}_{s\pi\sigma}$  coincides with  $\mathcal{P}_{s\pi\sigma}^n$  on the set  $\mathcal{M}_s^n$ .

From this point on, we will only deal with sets of histories that belong to  $\mathcal{S}(\mathcal{M}_s^\infty)$ , and whenever we talk about random variables we will have random variables with respect to the space  $(H_s^\infty, \mathcal{S}(\mathcal{M}_s^\infty), \mathcal{P}_{s\pi\sigma})$  in mind. We will use the notation  $\mathcal{E}_{s\pi\sigma}$  for the expectation taken with respect to  $\mathcal{P}_{s\pi\sigma}$ .

### 2.2.3 Induced stochastic processes on the set of states

A pair of history dependent strategies together with an initial state induce a stochastic process on the state space  $S$ , where the transitions are history dependent. In the case of Markov strategies, this stochastic process reduces to a nonhomogeneous Markov chain, while this Markov chain is even homogeneous when stationary strategies are used by the players.

A state is called absorbing, if the probability of leaving the state is zero for any available pair of actions; otherwise the state is called non-absorbing. So absorbing states cannot be left with respect to any stochastic process on the state space, induced by any strategy pair. When the stochastic process enters an absorbing state we speak of absorption.

Take now a pair of stationary strategies  $(x, y) \in X \times Y$ . As mentioned above, we obtain a homogeneous Markov chain with respect to  $(x, y)$ , without specifying the initial state now. The transition matrix for this Markov chain is denoted by  $P(x, y)$ . The matrix  $P(x, y)$  is a stochastic matrix and entry  $(s, t)$  of  $P(x, y)$ , where  $s, t \in S$ , is given as

$$p_s(t|x_s, y_s) := \sum_{i_s \in I_s} \sum_{j_s \in J_s} x_s(i_s) y_s(j_s) \cdot p_s(t|i_s, j_s),$$

which equals the probability of a transition from state  $s$  to state  $t$ , if the players use the mixed actions  $x_s$  and  $y_s$  in state  $s$ . With respect to the Markov chain corresponding to  $(x, y)$ , or equivalently with respect to  $P(x, y)$ , we can classify the states in the usual way (cf. Kemeny & Snell [1960], section 2.4). A state  $s$  is called transient if it has the property that, if the process starts in  $s$ , then visiting state  $s$  infinitely often has probability zero; otherwise the state is called recurrent. Note that a recurrent state  $s$  has the property that, if the process starts in state  $s$ , then visiting state  $s$  infinitely often has probability 1. A set of states  $A \subset S$  is called closed if it has the property that, if the process starts in any state  $s \in A$ , then ever leaving  $A$  has probability zero. (The above probabilities are obviously taken with regard to the probability measure  $\mathcal{P}_{sxy}$  on the set of infinite histories). A minimal closed set of states is called an ergodic set. It is known that ergodic sets form a disjoint partition of the set of recurrent states.

With respect to the Markov chain induced by  $(x, y)$ , the  $n$ -stage transition probabilities are clearly given by the matrix  $P^n(x, y)$ . For completeness, we let  $P^0(x, y) := I$ , where  $I$  is the identity matrix of size  $|S| \times |S|$ . We use the notation  $P^n(x, y)(s, t)$  for entry  $(s, t)$  of  $P^n(x, y)$ .

We define a stochastic matrix

$$Q(x, y) := \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N P^{n-1}(x, y);$$

here the limit is known to exist (cf. Doob [1953], theorem 2.1, page 175). Entry  $(s, t)$  of  $Q(x, y)$  is denoted by  $q_s(t|x, y)$  and expresses the expected average number of visits to state  $t$  if the stationary strategy pair  $(x, y)$  is used and the initial state is state  $s$ . Obviously, if  $t \in S$  is a transient state then  $q_s(t|x, y) = 0$  for all  $s \in S$ ; while if  $s, t \in S$  are recurrent then  $q_s(t|x, y) > 0$  if and only if  $s$  and  $t$  belong to the same ergodic set. If  $E \subset S$  is an ergodic set then the probability distribution  $(q_s(t|x, y))_{t \in E}$  is the same for all  $s \in E$ , so the  $s$ -th row and the  $t$ -th row of  $Q(x, y)$  are equal if  $s$  and  $t$  belong to the same ergodic set  $E$ . In fact,  $(q_s(t|x, y))_{t \in E}$ , for any  $s \in E$ , is the unique stationary distribution of the Markov chain corresponding to the ergodic set  $E$  if the players use  $(x, y)$ .

The matrix  $Q(x, y)$  has the property

$$P(x, y) \cdot Q(x, y) = Q(x, y) \cdot P(x, y) = Q(x, y), \quad (2.1)$$

which follows from the definitions. Using equations (2.1) inductively, we obtain for all  $n \in \mathbb{N}$  that

$$P^n(x, y) \cdot Q(x, y) = Q(x, y) \cdot P^n(x, y) = Q(x, y).$$

Hence by the definition of  $Q(x, y)$  we also have

$$Q(x, y) \cdot Q(x, y) = Q(x, y). \quad (2.2)$$

## 2.3 Rewards

As we have already discussed, during the play the players receive infinite sequences of payoffs. These sequences must be evaluated in some manner. We mainly deal with the so-called average reward for an evaluation of these sequences.

### 2.3.1 The average reward

The average reward was introduced by Gillette [1957] and is defined as follows.

**Definition 7** *The average reward with respect to a strategy pair  $(\pi, \sigma) \in \Pi \times \Sigma$  and initial state  $s \in S$  is defined for player  $k \in \{1, 2\}$  as*

$$\gamma_s^k(\pi, \sigma) := \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \mathcal{E}_{s\pi\sigma} \left( R_n^k \right) = \liminf_{N \rightarrow \infty} \mathcal{E}_{s\pi\sigma} \left( \frac{1}{N} \sum_{n=1}^N R_n^k \right),$$

where  $R_n^k$  denotes the random variable for the payoff for player  $k$  at stage  $n$ . We also use the vector notations

$$\gamma^k(\pi, \sigma) := \left( \gamma_s^k(\pi, \sigma) \right)_{s \in S}, \quad \gamma_s(\pi, \sigma) := \left( \gamma_s^k(\pi, \sigma) \right)_{k=1,2},$$

$$\gamma(\pi, \sigma) := \left( \gamma_s^k(\pi, \sigma) \right)_{s \in S; k=1,2}.$$

The average reward uses the long term expected average payoff for the evaluation of the infinite sequences of payoffs. As the limit does not necessarily exist, we need to take a limit point of the sequences. In the above definition we chose the limit inferior, even though all the further results remain valid with respect to the limit superior as well. The appendix provides more details on these issues.

This monograph mainly deals with the average reward. So whenever we talk about rewards in the sequel we will have the average reward in mind, unless mentioned otherwise.

### 2.3.2 The discounted reward

In the literature of stochastic game theory, the discounted reward was the first mentioned reward in Shapley [1953], and since then it has been one of the most widely used ones.

**Definition 8** *Let  $\beta \in (0, 1)$ . The  $\beta$ -discounted reward with respect to the strategy pair  $(\pi, \sigma)$  and initial state  $s \in S$  is defined for player  $k \in \{1, 2\}$  as*

$$\gamma_{\beta s}^k(\pi, \sigma) := (1 - \beta) \cdot \sum_{n=1}^{\infty} \beta^{n-1} \cdot \mathcal{E}_{s\pi\sigma} \left( R_n^k \right),$$

where  $R_n^k$  denotes the random variable for the payoff for player  $k$  at stage  $n$ . We also use the vector notations

$$\gamma_{\beta}^k(\pi, \sigma) := \left( \gamma_{\beta s}^k(\pi, \sigma) \right)_{s \in S}, \quad \gamma_{\beta s}(\pi, \sigma) := \left( \gamma_{\beta s}^k(\pi, \sigma) \right)_{k=1,2},$$

$$\gamma_{\beta}(\pi, \sigma) := \left( \gamma_{\beta s}^k(\pi, \sigma) \right)_{s \in S; k=1,2}.$$

The idea of the  $\beta$ -discounted reward is that the payoff at stage  $n$  has to be discounted  $n - 1$  times, as it is received  $n - 1$  stages later than the first payoff at stage 1. In economic applications, this discount factor  $\beta$  reflects an interest rate  $(1 - \beta)/\beta$ . The factor  $(1 - \beta)$  is just a normalizing factor so that the discounted reward becomes  $c$  if all the payoffs equal the same constant  $c$ .

The discounted reward itself is not only used in economic applications, but also provides one of the most frequently used tools for the analysis of the average reward.

### 2.3.3 The rewards for stationary strategies

A pair of stationary strategies induces a homogeneous Markov chain as discussed in section 2.2.3. Due to the simple structure of such stochastic processes, the average reward and the discounted reward can be easier calculated for stationary strategies. When using stationary strategies, the prescribed mixed actions only depend on the present state, therefore whenever a state  $s \in S$  is visited, the expected payoff for player  $k$  is

$$r_s^k(x_s, y_s) := \sum_{i_s \in I_s} \sum_{j_s \in J_s} x_s(i_s) y_s(j_s) \cdot r_s^k(i_s, j_s)$$

and the expected transition vector is the  $s$ -th row of the stochastic matrix  $P(x, y)$ . We also use the vector notations

$$r^k(x, y) := (r_s^k(x_s, y_s))_{s \in S}, \quad r_s(x_s, y_s) := \left( r_s^k(x_s, y_s) \right)_{k=1,2},$$

$$r(x, y) := \left( r_s^k(x_s, y_s) \right)_{s \in S; k=1,2}.$$

**Lemma 9** *Take a stationary strategy pair  $(x, y) \in X \times Y$ . Then*

- (a)  $\gamma(x, y) = Q(x, y) \cdot r(x, y)$ ;
- (b)  $\gamma(x, y) = P(x, y) \cdot \gamma(x, y)$ ;
- (c)  $\gamma(x, y) = Q(x, y) \cdot \gamma(x, y)$ ;
- (d)  $\gamma_s(x, y) = \gamma_t(x, y)$  if  $s$  and  $t$  belong to the same ergodic set for  $(x, y)$ ;

**Proof.**

(a) Let  $R_n$  denote the random variable for the payoff vector at stage  $n$ . Then for all  $s \in S$  we have

$$\mathcal{E}_{sxy}(R_n) = \sum_{t \in S} P^{n-1}(x, y)(s, t) \cdot r_t(x_t, y_t).$$

Using definition 7, for all  $s \in S$

$$\begin{aligned} \gamma_s(x, y) &= \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \mathcal{E}_{sxy}(R_n) \\ &= \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \sum_{t \in S} P^{n-1}(x, y)(s, t) \cdot r_t(x_t, y_t). \end{aligned}$$

Therefore

$$\gamma(x, y) = \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N P^{n-1}(x, y) \cdot r(x, y) = Q(x, y) \cdot r(x, y),$$

which completes the proof of (a).

(b) It immediately follows from (a) by using (2.1).

(c) Using (b) inductively, we have for all  $n \in \mathbb{N}$  that

$$\gamma(x, y) = P^n(x, y) \cdot \gamma(x, y).$$

Now by the definition of  $Q(x, y)$  we obtain (c).

(d) It is a consequence of (c) using the fact that if  $s$  and  $t$  belong to the same ergodic set then the  $s$ -th row and the  $t$ -th row of  $Q(x, y)$  are equal.  $\square$

Next we discuss an important continuity property of the average reward on the spaces of stationary strategies.

**Lemma 10** For any sequence  $(x^n, y^n)$ ,  $n \in \mathbb{N}$ , in  $X \times Y$  converging to some  $(x, y)$  in  $X \times Y$ , if for large  $n \in \mathbb{N}$ , the ergodic sets with respect to  $(x^n, y^n)$  coincide with the ergodic sets with respect to  $(x, y)$  then  $\gamma(x^n, y^n)$  has a limit as  $n$  tends to infinity and

$$\gamma(x, y) = \lim_{n \rightarrow \infty} \gamma(x^n, y^n).$$

**Proof.** Using that  $Q(x, y) = \lim_{n \rightarrow \infty} Q_{x^n y^n}$  holds for such a sequence  $(x^n, y^n)$ ,  $n \in \mathbb{N}$  (cf. Schweitzer [1968], theorem 5), lemma 9-(a) implies the statement.  $\square$

Note that, in general, the average reward is not continuous on the spaces of stationary strategies, as illustrated by the next example.

**Example 11**

$T$	0,0		1,0
	1		2
$B$	1,0		1,0
	2		2
	1		2

Notice that state 2 is absorbing and both players have only one action in state 2, hence state 2 is not an interesting initial state and strategies only need to be defined for state 1. So assume the initial state to be state 1. By suppressing the absorbing state and the transition for action  $T$  which keeps the play in the same state with probability 1, and by denoting the absorption for action  $B$  by a  $*$  we may represent the game as follows:

$T$	0,0
	*
$B$	1,0
	*
	1

Such a shorter representation shall often be used later. Let  $y$  denote player 2's only strategy. In state 1 player 1 has two actions:  $T$  standing for top and  $B$  standing for bottom. Now we have  $\gamma_1^1(x, y) = 1$  for all  $x \in \{(u, 1 - u) \mid u \in [0, 1]\} \subset X$ , however  $\gamma_1^1((1, 0), y) = 0$ , which demonstrates that the average reward is not continuous on  $X \times Y$ . In fact, state 1 is transient with respect to  $(x, y)$  for all  $x \in \{(u, 1 - u) \mid u \in [0, 1]\}$ , while it becomes recurrent for  $((1, 0), y)$ . This clarifies the necessity of the condition on the ergodic structures in lemma 10.  $\triangleleft$

**Lemma 12** Let  $(x, y) \in X \times Y$  and  $\beta \in (0, 1)$ . Let  $I$  denote the identity matrix of size  $|S| \times |S|$ . Then the inverse of the matrix  $(I - \beta \cdot P(x, y))$  exists, and

$$\gamma_\beta(x, y) = (1 - \beta) \cdot (I - \beta \cdot P(x, y))^{-1} \cdot r(x, y).$$

**Proof.** Since

$$(I - \beta \cdot P(x, y)) \cdot \left( \sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x, y) \right) = I$$

$$\left( \sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x, y) \right) \cdot (I - \beta \cdot P(x, y)) = I,$$

we may conclude that the matrix  $(I - \beta \cdot P(x, y))$  has an inverse and

$$(I - \beta \cdot P(x, y))^{-1} = \sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x, y).$$

Let  $R_n$  denote the random variable for the payoff vector at stage  $n$ . As for any  $s \in S$

$$\mathcal{E}_{sxy}(R_n) = \sum_{t \in S} P^{n-1}(x, y)(s, t) \cdot r_t(x_t, y_t),$$

it follows from definition 8 that for all  $s \in S$

$$\gamma_{\beta s}(x, y) = (1 - \beta) \cdot \sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x, y)(s, t) \cdot r_t(x_t, y_t).$$

Therefore

$$\begin{aligned} \gamma_{\beta}(x, y) &= (1 - \beta) \cdot \sum_{n=1}^{\infty} \beta^{n-1} \cdot P^{n-1}(x, y) \cdot r(x, y) \\ &= (1 - \beta) \cdot (I - \beta \cdot P(x, y))^{-1} \cdot r(x, y), \end{aligned}$$

which completes the proof.  $\square$

The following continuity property of the discounted reward makes the analysis of discounted games substantially easier.

**Lemma 13** *The function  $\gamma_{\beta}(\cdot, \cdot)$  is continuous on  $X \times Y$  for any  $\beta \in (0, 1)$ .*

**Proof.** It follows from lemma 12, since each factor in continuous on  $X \times Y$ .  $\square$

There is a strong relation between the average reward and the discounted rewards for stationary strategies. This is stated in the next lemma.

**Lemma 14** *Let  $(x, y) \in X \times Y$ . Then*

$$\gamma(x, y) = \lim_{\beta \uparrow 1} \gamma_{\beta}(x, y).$$

**Proof.** The result follows from lemma 9-(a) and lemma 12, using the equality

$$Q(x, y) = \lim_{\beta \uparrow 1} (1 - \beta) \cdot (I - \beta \cdot P(x, y))^{-1};$$

which is shown in Blackwell [1962].  $\square$

## 2.4 Playing against a fixed strategy

In this section we examine what happens when a player fixes a strategy in a stochastic game. Each player aims to maximize his own individual reward, so it is of interest to see how a player has to play against a fixed strategy.

**Definition 15** *Let  $\sigma \in \Sigma$  be a fixed strategy for player 2. Then a strategy  $\pi \in \Pi$  for player 1 is an  $\varepsilon$ -best reply against  $\sigma$  for initial state  $s \in S$ , where  $\varepsilon \geq 0$ , if*

$$\gamma_s^1(\bar{\pi}, \sigma) \leq \gamma_s^1(\pi, \sigma) + \varepsilon \quad \forall \bar{\pi} \in \Pi.$$

*The strategy  $\pi$  is called an  $\varepsilon$ -best reply against  $\sigma$ , if it is an  $\varepsilon$ -best reply against  $\sigma$  for all initial states  $s \in S$ . 0-best replies are simply called best replies. Similar definitions hold for the best replies of player 2, and for the discounted reward as well.*

First we treat existence of ( $\varepsilon$ -)best replies with regard to the average reward.

### Theorem 16

- (a) *Against a fixed strategy  $\sigma \in \Sigma$ , for any  $\varepsilon > 0$ , player 1 has a pure  $\varepsilon$ -best reply  $\pi \in \Pi^p$ . A similar statement holds for player 2 as well.*
- (b) *Against any fixed stationary strategy  $y \in Y$ , player 1 has a pure stationary best reply  $i \in I$ . A similar statement holds for player 2 as well.*

The proof can be found in Monash [1980] (theorem 1, page 6) and Hordijk et al. [1983]. It is still an open problem whether, against a fixed Markov strategy, pure Markov  $\varepsilon$ -best replies exist for all  $\varepsilon > 0$ .

The next example demonstrates that a player does not necessarily have best replies against a fixed strategy.

### Example 17

	<i>L</i>	<i>R</i>	
<i>T</i>	0,0	0,0	
<i>B</i>	1,0	0,0	
		*	*
	1		

In order to explain the notation once more we give the game in its full form as well.

	<i>L</i>	<i>R</i>			
<i>T</i>	0,0	0,0	1	1	
<i>B</i>	1,0	0,0	2	3	
	1		2	3	

Consider the Markov strategy  $g$  for player 2 which prescribes action  $L$  with probability  $1 - 1/n$  and action  $R$  with probability  $1/n$  at stage  $n$ . It is clear that, against the strategy  $g$ , player 1 is not able to get reward 1 as action  $L$  will never be chosen with probability 1. However, for any  $\varepsilon > 0$ , player 1 can get at least  $1 - \varepsilon$  by playing a Markov strategy which prescribes to play action  $T$  until  $1 - 1/n \geq 1 - \varepsilon$  for the current stage  $n$  and then to play action  $B$  afterwards. This clearly implies that player 1 does not have best replies against the Markov strategy  $g$ .  $\triangleleft$

On discounted best replies we will need the following useful lemma, which follows from Hordijk et al. [1983].

**Theorem 18** *Against any fixed stationary strategy  $y \in Y$ , for any  $\beta \in (0, 1)$ , player 1 has a pure stationary  $\beta$ -discounted best reply  $i \in I$ . A similar statement holds for player 2 as well.*

## 2.5 Zero-sum stochastic games and optimality

The theory of zero-sum stochastic games was started by the seminal work of Shapley [1953]. Zero-sum stochastic games are special stochastic games in which the two players have completely opposite interests. These opposite interests are expressed by two assumptions. The first assumption is that

$$r_s^1(i_s, j_s) = -r_s^2(i_s, j_s) \quad \forall i_s \in I_s, \forall j_s \in J_s, \forall s \in S,$$

so in fact the payoffs can be assumed to be payed to player 1 by player 2. Obviously, (??) means that the sum of the payoffs of the players is always equal to zero. The second assumption is that

$$\gamma_s^1(\pi, \sigma) = -\gamma_s^2(\pi, \sigma) \quad \forall \pi \in \Pi, \forall \sigma \in \Sigma, \forall s \in S,$$

so the sum of the rewards of the players is also equal to zero. When the discounted reward is used then the second assumption is of course that

$$\gamma_s^1(\pi, \sigma) = -\gamma_s^2(\pi, \sigma) \quad \forall \pi \in \Pi, \forall \sigma \in \Sigma, \forall s \in S.$$

Note that the second assumption follows from the first one for the discounted reward, but not in the case of the average reward, because player 2 has to use the limit superior instead of the limit inferior in the definition of the average reward (cf. definition 7) in order to make the sum of the rewards zero. (Recall that all the previously discussed issues remain valid with respect to the limit superior.)

As the players have completely opposite interests, the questions in these games are the following: (i) what are the rewards that the players can guarantee against any strategy of their opponents, (ii) which strategies of the players can guarantee these rewards. We will discuss both these issues for the average reward and for the discounted reward as well.

Technically, instead of considering two payoff functions and two rewards, it is easier to consider only player 1's payoffs and player 1's reward and to assume that player 1

tries to maximize his own reward while player 2 tries to minimize player 1's reward. So instead of  $r^1$ ,  $\gamma^1$ , and  $\gamma_\beta^1$  we simply use  $r$ ,  $\gamma$ , and  $\gamma_\beta$ .

As the players have completely opposite interests, it is natural to evaluate a strategy for a player by the reward that it guarantees against any strategy of the opponent. On basis of this evaluation we now define the solution concepts of zero-sum stochastic games.

**Definition 19** *In a zero-sum stochastic game, for strategies  $\pi \in \Pi$  and  $\sigma \in \Sigma$  let*

$$\begin{aligned} \underline{v}_s(\pi) &:= \inf_{\sigma' \in \Sigma} \gamma_s(\pi, \sigma') \quad \forall s \in S, & \underline{v}(\pi) &:= (\underline{v}_s(\pi))_{s \in S} \\ \bar{v}_s(\sigma) &:= \sup_{\pi' \in \Pi} \gamma_s(\pi', \sigma) \quad \forall s \in S, & \bar{v}(\sigma) &:= (\bar{v}_s(\sigma))_{s \in S}. \end{aligned}$$

We say that strategy  $\pi$  guarantees reward  $c_s \in \mathbb{R}$  for initial state  $s \in S$ , if  $\underline{v}_s(\pi) \geq c_s$ , and guarantees  $c \in \mathbb{R}^{|S|}$ , if  $\underline{v}(\pi) \geq c$ . Similarly, a strategy  $\sigma$  is said to guarantee reward  $c_s \in \mathbb{R}$  for initial state  $s \in S$ , if  $\bar{v}_s(\sigma) \leq c_s$ , and to guarantee  $c \in \mathbb{R}^{|S|}$ , if  $\bar{v}(\sigma) \leq c$ .

If there exists a real valued vector  $v = (v_s)_{s \in S}$  such that

$$v_s = \sup_{\pi \in \Pi} \underline{v}_s(\pi) = \inf_{\sigma \in \Sigma} \bar{v}_s(\sigma) \quad \forall s \in S,$$

then  $v$  is called the value of the zero-sum stochastic game.

Assume that the value  $v$  exists. Then, for initial state  $s \in S$ , a strategy  $\pi \in \Pi$  is called  $\varepsilon$ -optimal for player 1, where  $\varepsilon \geq 0$ , if

$$\underline{v}_s(\pi) \geq v_s - \varepsilon.$$

The strategy  $\pi$  is called  $\varepsilon$ -optimal, if it is  $\varepsilon$ -optimal for all initial states  $s \in S$ . 0-optimal strategies are simply called optimal. Similar definitions hold for player 2.

For  $\beta \in (0, 1)$ , the  $\beta$ -discounted value  $v_\beta$  and  $\beta$ -discounted optimality are analogously defined.

Notice that for all  $\pi \in \Pi$  and  $\sigma \in \Sigma$  we have  $\underline{v}(\pi) \leq \gamma(\pi, \sigma) \leq \bar{v}(\sigma)$ . This implies

$$\sup_{\pi \in \Pi} \underline{v}(\pi) \leq \inf_{\sigma \in \Sigma} \bar{v}(\sigma).$$

Note furthermore that, by definition 19, if the value exists then both players must have  $\varepsilon$ -optimal strategies for all  $\varepsilon > 0$ .

Shapley [1953] showed the following result for discounted zero-sum stochastic games. The continuity of the discounted reward on the stationary strategy spaces plays a crucial role here (cf. lemma 13).

**Theorem 20** *In any zero-sum stochastic game, for any  $\beta \in (0, 1)$ , the  $\beta$ -discounted value  $v_\beta$  exists and both players have stationary  $\beta$ -discounted optimal strategies. Moreover, a stationary strategy  $x \in X$  is  $\beta$ -discounted optimal if and only if*

$$v_\beta \leq (1 - \beta) \cdot r(x, y) + \beta \cdot P(x, y) \cdot v_\beta \quad \forall y \in Y.$$

A similar statement holds for player 2 as well.

It is fairly appealing in discounted games that optimal strategies of the players can be found in terms of stationary strategies.

Bewley & Kohlberg [1976-I] showed that the discounted values have a unique limit point as the discount factor tends to 1.

**Theorem 21** *In any zero-sum stochastic game,  $\lim_{\beta \uparrow 1} v_\beta$  exists.*

Based on deep results of Bewley & Kohlberg [1976-I,II] on discounted values and stationary discounted optimal strategies, Mertens & Neyman [1981] achieved the following fundamental result.

**Theorem 22** *In any zero-sum stochastic game, the value  $v$  exists and*

$$v = \lim_{\beta \uparrow 1} v_\beta.$$

*Moreover, both players have  $\varepsilon$ -optimal strategies for all  $\varepsilon > 0$ .*

The fact that the average value is the limit of the discounted values, as the discount factor  $\beta$  tends to 1, is often used in the analysis of stochastic games.

Next we provide an illustration for theorem 22 by examining the famous stochastic game, called the Big Match, which was introduced by Gillette [1957]. For a long time it was unclear whether the game had a value and whether player 1 had  $\varepsilon$ -optimal strategies for all  $\varepsilon > 0$ . The game was only solved 11 years later by Blackwell & Ferguson [1968].

**Example 23** *The Big Match*

	$L$	$R$
$T$	0	1
$B$	1 *	0 *
	1	

The beauty of the Big Match is that the structure of the game is so simple. In state 1 each player has two actions. Player 1's actions are  $T$  and  $B$  standing for top and bottom, while player 2's actions are  $L$  and  $R$  standing for left and right. 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 place of absorption fully depends on the action chosen by player 2. We discuss several important issues regarding the Big Match.

**Lemma 24** *The Big Match has the following properties.*

- (a) The value for state 1 equals  $v_1 = 1/2$ .
- (b) Player 2 has a stationary optimal strategy  $y = (1/2, 1/2) \in Y$ .

- (c) For any  $\varepsilon > 0$ , player 1 has an  $\varepsilon$ -optimal strategy for initial state 1 which, for present state 1 and any past history  $h \in H$ , prescribes action  $T$  with probability  $1 - \left(\frac{1}{k(h)+N}\right)^2$  and action  $B$  with probability  $\left(\frac{1}{k(h)+N}\right)^2$ , 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  $h$ , and where  $N$  is sufficiently large.
- (d) Player 1 has no optimal strategy for initial state 1.
- (e) Player 1 has neither stationary nor Markov  $\varepsilon$ -optimal strategy for initial state 1, if  $\varepsilon > 0$  is small. In fact, player 1 can only guarantee reward 0 by stationary strategies and by Markov strategies, namely  $\sup_{x \in X} v_1(x) = \sup_{f \in F} v_1(f) = 0$ .

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 strategy 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 the strategy that player 1 uses against  $(1/2, 1/2)$  the reward always equals  $1/2$ , since the expected payoff equals  $1/2$  for each stage.

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

Now we show (d). Take an arbitrary strategy  $\pi$ . Consider the strategy for player 2 which prescribes to play action  $L$  as long as  $\pi$  chooses action  $T$  with probability 1, to play action  $R$  at the first stage when  $\pi$  puts a positive probability on action  $B$ , and to play the optimal strategy  $(1/2, 1/2)$  afterwards. Then if  $\pi$  always chooses  $T$  with probability 1 then the reward is 0. On the other hand, if  $\pi$  ever puts a positive probability on  $B$ , say at stage  $n$  for the first time, then with a positive probability absorption occurs with payoff zero at stage  $n$ , while with the rest of the probability all further expected payoffs equal  $1/2$ , as player 2 uses  $(1/2, 1/2)$ . This means that the reward is strictly less than  $1/2$ , so  $\pi$  cannot be optimal for initial state 1. Hence we have shown (d).

Now we prove (e). Clearly, it suffices to show the statement for Markov strategies, as all stationary strategies are Markov strategies as well. Take an arbitrary Markov strategy  $f$  for player 1. Consider the stationary strategy  $y = (1, 0)$ , which prescribes action  $L$  with probability 1. Let  $\rho^n(f)$  denote the overall probability that absorption occurs at any of the stages  $n + 1, n + 2, n + 3, \dots$  with respect to  $f$  when the initial state is state 1 (clearly, this probability is independent of the strategy used by player 1, 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  $\rho^0(f) = \lim_{n \rightarrow \infty} (\rho^0(f) - \rho^n(f))$ , hence  $\lim_{n \rightarrow \infty} \rho^n(f) = 0$ . 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  $L$  afterwards. Then, with probability at least  $1 - \varepsilon$ , either absorption occurs with payoff zero during the first  $N$  stages or entry  $(T, L)$  is played at each

stage after stage  $N$ . Therefore  $\gamma_1(f, g) \leq \varepsilon$ . As  $\varepsilon > 0$  was arbitrary, we have shown (e).  $\triangleleft$

## 2.6 General-sum stochastic games and equilibria

The study of general-sum stochastic games has been started by Fink [1964] and Takahashi [1964]. In general-sum games, in contrast with the previously discussed zero-sum games, the players do not necessarily have strictly opposite interests. The solution concepts value and ( $\varepsilon$ -)optimal strategies therefore lose their meanings in the context of general-sum games. The most widely applied solution concept here is the concept of (Nash-)equilibria. The idea is to find pairs of strategies which reflect strategically stable situations in the sense that neither player can individually improve his reward by choosing another strategy.

### Definition 25

- (a) A pair of strategies  $(\pi, \sigma) \in \Pi \times \Sigma$  is called a (Nash)  $\varepsilon$ -equilibrium for initial state  $s \in S$ , where  $\varepsilon \geq 0$ , if

$$\gamma_s^1(\bar{\pi}, \sigma) \leq \gamma_s^1(\pi, \sigma) + \varepsilon \quad \forall \bar{\pi} \in \Pi$$

$$\gamma_s^2(\pi, \bar{\sigma}) \leq \gamma_s^2(\pi, \sigma) + \varepsilon \quad \forall \bar{\sigma} \in \Sigma.$$

The strategy pair  $(\pi, \sigma)$  is an  $\varepsilon$ -equilibrium, if it is an  $\varepsilon$ -equilibrium for all initial states  $s \in S$ . 0-equilibria are simply called equilibria.

- (b) Let  $\beta_1, \beta_2 \in (0, 1)$ . A pair of strategies  $(\pi, \sigma) \in \Pi \times \Sigma$  is called a (Nash)  $(\beta_1, \beta_2)$ -discounted equilibrium for initial state  $s \in S$ , if

$$\gamma_{\beta_1 s}^1(\bar{\pi}, \sigma) \leq \gamma_{\beta_1 s}^1(\pi, \sigma) \quad \forall \bar{\pi} \in \Pi$$

$$\gamma_{\beta_2 s}^2(\pi, \bar{\sigma}) \leq \gamma_{\beta_2 s}^2(\pi, \sigma) \quad \forall \bar{\sigma} \in \Sigma.$$

The strategy pair  $(\pi, \sigma)$  is a  $(\beta_1, \beta_2)$ -discounted equilibrium, if it is a  $(\beta_1, \beta_2)$ -discounted equilibrium for all initial states  $s \in S$ .

When we speak of stationary ( $\varepsilon$ -)equilibria or Markov ( $\varepsilon$ -)equilibria we obviously mean ( $\varepsilon$ -)equilibria which consists of stationary or Markov strategies, respectively. Note that  $\varepsilon$ -equilibria can be simply seen as pairs of  $\varepsilon$ -best replies against each other with respect to the rewards used by the players.

Fink [1964] and Takahashi [1964] have proven the following result.

**Theorem 26** *In any stochastic game, there exists a stationary  $(\beta_1, \beta_2)$ -discounted equilibrium for any  $\beta_1, \beta_2 \in (0, 1)$ .*

Regarding the average reward, the existence of  $\varepsilon$ -equilibria has been the most challenging open problem in stochastic game theory for a long time. Although this question had been previously answered in the affirmative for several classes of stochastic games, it was only recently that Vieille [1997,I,II] derived the existence of  $\varepsilon$ -equilibria for all two-person stochastic games, for all  $\varepsilon > 0$ . The existence problem, however, is still open in stochastic games with more than two players.

The concept of ( $\varepsilon$ -)equilibria naturally extends to stochastic games where the players use different rewards. In this sense we can speak of ( $\varepsilon$ -)equilibria in zero-sum stochastic games as well (recall that, in the case of the average reward, player 2 uses the limit superior instead of the limit inferior). It is easy to see that, in zero-sum stochastic games, for any  $\varepsilon > 0$ , any pair of  $\varepsilon$ -optimal strategies forms a  $2\varepsilon$ -equilibrium, and any  $\varepsilon$ -equilibrium must be formed by  $2\varepsilon$ -optimal strategies.

**Definition 27** *The I-zero-sum game of a general-sum stochastic game is the zero-sum game where player 1 maximizes his own reward while player 2 minimizes player 1's reward. The II-zero-sum game of a general-sum stochastic game is the zero-sum game where player 2 maximizes his own reward while player 1 minimizes player 2's reward. The value of the  $k$ -zero-sum game,  $k \in \{1, 2\}$ , is denoted by  $v^k$ .*

Clearly,  $v^k$  is the reward that player  $k$  can guarantee on his own, therefore in an  $\varepsilon$ -equilibrium situation player  $k$  must receive a reward at least  $v^k - \varepsilon$ .

For zero-sum stochastic games, in view of theorem 22, we have  $v = \lim_{\beta \uparrow 1} v^\beta$ . For general-sum stochastic games the relation between the discounted and the average game is much weaker, which is fully clarified by the game in Sorin [1986].

**Example 28**

	$L$	$R$	
$T$	0,1	1,0	
$B$	1,0	0,2	
	*	*	
	1		

The I-zero-sum game is exactly the Big Match (cf. example 23), while the II-zero-sum stochastic game is also the Big Match except for payoff 2 in entry  $(B, R)$ . Sorin [1986] showed that the set of rewards which, for all  $\varepsilon > 0$ , correspond to  $\varepsilon$ -equilibria is

$$L = \text{conv} \{(1/2, 1), (2/3, 2/3)\},$$

where  $\text{conv}$  stands for the convex hull of a set. On the other hand, the set of rewards corresponding to  $(\beta_1, \beta_2)$ -discounted equilibria is the same singleton for all  $\beta_1, \beta_2 \in (0, 1)$ :

$$L_{\text{disc}} = \{(1/2, 2/3)\}.$$

Therefore there is a gap between the average solutions and the discounted solutions of this game.  $\triangleleft$

We would like to stress that, despite the above example, the discounted solutions are frequently used in the analysis of equilibria for the average reward.

## 2.7 Special classes of stochastic games

In this section we give a brief overview of the most important classes of stochastic games with the most important issues.

**Definition 29** *A stochastic game is called*

- (a) *a unichain stochastic game, if with respect to any pair of strategies there is only one ergodic set of states.*
- (b) *a perfect information stochastic game, if  $S$  can be partitioned into  $S^1$  and  $S^2$  such that  $|J_s| = 1$  for all  $s \in S^1$  and  $|I_s| = 1$  for all  $s \in S^2$ .*
- (c) *a switching control stochastic game, if  $S$  can be partitioned into  $S^1$  and  $S^2$  such that  $p_s(i_s, j_s)$  is independent of  $j_s$  for all  $i_s \in I_s$ ,  $s \in S^1$ , and  $p_s(i_s, j_s)$  is independent of  $i_s$  for all  $j_s \in J_s$ ,  $s \in S^2$ .*
- (d) *a stochastic game with additive reward and additive transition (ARAT) structure, if  $r_s^k(i_s, j_s)$  and  $p_s(i_s, j_s)$  can be decomposed as  $r_s^k(i_s, j_s) = r_{s1}^k(i_s) + r_{s2}^k(j_s)$  and  $p_s(i_s, j_s) = p_{s1}(i_s) + p_{s2}(j_s)$  for all  $i_s \in I_s$ ,  $j_s \in J_s$ ,  $s \in S$ .*
- (e) *a stochastic game with state independent transitions (SIT), if the cardinality of the action spaces is independent of the state and, assuming that  $I_s = I_t =: \bar{I}$  and  $J_s = J_t =: \bar{J}$  for all  $s, t \in S$ , it holds that  $p_s = p_t$  for all  $s, t \in S$ .*
- (f) *a stochastic game with separable rewards and state independent transitions (SER-SIT), if it is a SIT stochastic game and, assuming that  $I_s = I_t =: \bar{I}$  and  $J_s = J_t =: \bar{J}$  for all  $s, t \in S$ , the function  $r_s^k(i, j)$  can be decomposed as  $r_s^k(i, j) = c_s^k + d^k(i, j)$  for all  $s \in S$ ,  $i \in \bar{I}$ ,  $j \in \bar{J}$ .*
- (g) *a repeated game with absorbing states, if all the states but one are absorbing.*
- (h) *a recursive stochastic game, if the payoffs are equal to zero in all non-absorbing states.*

For the zero-sum case the following theorem summarizes the most important results.

### Theorem 30

- (a) *In zero-sum unichain stochastic games both players have stationary optimal strategies, and the value is independent of the initial state (cf. Hoffman & Karp [1966], Thuijsman [1992]).*

- (b) *In zero-sum perfect information stochastic games both players have pure stationary optimal strategies (cf. Liggett & Lippman [1969]).*
- (c) *In zero-sum switching control stochastic games both players have stationary optimal strategies (cf. Filar [1981]).*
- (d) *In zero-sum ARAT stochastic games both players have pure stationary optimal strategies (cf. Raghavan et al. [1985]).*
- (e) *In zero-sum SIT stochastic games both players have stationary optimal strategies, and the value is independent of the initial state (cf. Thuijsman [1992]).*
- (f) *In zero-sum SER-SIT stochastic games both players have stationary optimal strategies such that the prescribed mixed actions are state independent, and the value is independent of the initial state (cf. Parthasarathy et al. [1984]).*
- (g) *In zero-sum repeated games with absorbing states both players have  $\varepsilon$ -optimal strategies for all  $\varepsilon > 0$  (cf. Kohlberg [1974]).*
- (h) *In zero-sum recursive stochastic games both players have stationary  $\varepsilon$ -optimal strategies for all  $\varepsilon > 0$  (cf. Everett [1957], Thuijsman & Vrieze [1992]).*

In the general-sum case the following results are known.

**Theorem 31**

- (a) *In unichain stochastic games stationary equilibria exist (cf. Rogers [1969], Sobel [1971], Federgruen [1978], Thuijsman [1992]).*
- (b) *In perfect information stochastic games equilibria exist (cf. Liggett & Lippman [1969], Thuijsman & Raghavan [1997]).*
- (c) *In switching control stochastic games  $\varepsilon$ -equilibria exist for all  $\varepsilon > 0$  (cf. Thuijsman & Raghavan [1997]).*
- (d) *In ARAT stochastic games equilibria exist (cf. Thuijsman & Raghavan [1997].)*
- (e) *In SIT stochastic games  $\varepsilon$ -equilibria exist for all  $\varepsilon > 0$  (cf. Thuijsman [1992]).*
- (f) *In SER-SIT stochastic games stationary equilibria exist such that the prescribed mixed actions are state independent (cf. Parthasarathy et al. [1984]).*
- (g) *In repeated games with absorbing states  $\varepsilon$ -equilibria exist for all  $\varepsilon > 0$  (cf. Vrieze & Thuijsman [1989]).*

—Recently, Vieille derived that, for all  $\varepsilon > 0$ ,  $\varepsilon$ -equilibria exist in all two-person stochastic games. In order to achieve this statement, first he proved that a specific class of two-person recursive games possesses  $\varepsilon$ -equilibria for all  $\varepsilon > 0$  (cf. Vieille [1997,II]) and afterwards, he showed that the general existence problem in two-person stochastic games reduces to the existence problem in the previously mentioned class of two-person recursive games (cf. Vieille [1994] and [1997,I]).



## Part I

# Zero-sum stochastic games



## Chapter 3

# Simplifying optimal strategies

### 3.1 Introduction

In zero-sum stochastic games, the value exists and the players have  $\varepsilon$ -optimal strategies for all  $\varepsilon > 0$ , as stated in theorem 22. Moreover, the Big Match (cf. example 23 and lemma 24-(d)) illustrated that optimal strategies need not always exist. In this chapter, which is based on Flesch et al. [1998,I], we examine how optimal strategies, when they exist, can be simplified.

The main result, which will follow from theorem 36, is the following one.

**Main Theorem 3** *In a zero-sum stochastic game, if a player has an optimal strategy then he has stationary  $\varepsilon$ -optimal strategies, for all  $\varepsilon > 0$ , and he has Markov optimal strategies as well.*

In other words, we show that optimal strategies, when they exist, can be simplified by stationary  $\varepsilon$ -optimal strategies, for all  $\varepsilon > 0$ , and by Markov optimal strategies as well. So instead of playing a complex history dependent optimal strategy, the player can also play a stationary  $\varepsilon$ -optimal strategy with an arbitrary small  $\varepsilon > 0$ , or he can even achieve optimality in the class of Markov strategies. We present such a construction for which we do not even need to know any optimal strategy, which makes the result even stronger.

We provide two examples showing the sharpness of the result. Example 37 will demonstrate that the existence of stationary optimal strategies is not implied by the existence of optimal strategies, while example 43 will clarify that the existence of stationary  $\varepsilon$ -optimal strategies, for all  $\varepsilon > 0$ , is not sufficient for the existence of optimal strategies.

In several stochastic games, it is easy to show that stationary  $\varepsilon$ -optimal strategies do not exist for all  $\varepsilon > 0$ . In such games, by the above theorem, we may exclude the existence of optimal strategies as well. Note that, without knowing this result, it would be a much harder problem to check the existence of optimal strategies, since optimal strategies could only exist in terms of complex history dependent strategies. Moreover, the above theorem provides a sufficient condition for the existence of stationary  $\varepsilon$ -optimal strategies and Markov optimal strategies. For many classes of

stochastic games, where on the payoff and transition structures special conditions are imposed, stationary  $\varepsilon$ -optimal strategies exist for all  $\varepsilon > 0$  (cf. theorem 30), while about sufficient conditions for the existence of Markov optimal strategies comparatively little is known. Here, instead of providing such structural conditions, the existence of optimal strategies will be proven to be sufficient.

At the end of this chapter we make several remarks regarding the construction and the proofs. There we also treat possible simplifications of strategies that are only optimal for particular initial states.

### 3.2 Preliminaries

The following lemma was shown by von Neumann [1928].

**Lemma 32** *Let  $s \in S$  and let  $c_s : X_s \times Y_s \mapsto \mathbb{R}$  be linear in both components. Then there exist  $x_s \in X_s$ ,  $y_s \in Y_s$ , and a unique  $C_s \in \mathbb{R}$  such that*

$$c_s(x_s, y'_s) \geq C_s \geq c_s(x'_s, y_s) \quad \forall x'_s \in X_s, \forall y'_s \in Y_s.$$

**Lemma 33** *Let  $s \in S$ . Let  $c_s$  and  $C_s$  be as in lemma 32. Then the sets*

$$O_s^1 := \{x_s \in X_s \mid c_s(x_s, y'_s) \geq C_s \quad \forall y'_s \in Y_s\}$$

$$O_s^2 := \{y_s \in Y_s \mid c_s(x'_s, y_s) \leq C_s \quad \forall x'_s \in X_s\}$$

$$\bar{O}_s^1 := \{x_s \in X_s \mid c_s(x_s, y'_s) = C_s \quad \forall y'_s \in O_s^2\}$$

$$\bar{O}_s^2 := \{y_s \in Y_s \mid c_s(x'_s, y_s) = C_s \quad \forall x'_s \in O_s^1\}$$

are nonempty polytopes. Furthermore, if  $\bar{I}_s$  and  $\bar{J}_s$  denote the extreme points of the sets  $\bar{O}_s^1$  and  $\bar{O}_s^2$ , respectively, then

$$\bar{I}_s = \{i_s \in I_s \mid \exists x_s \in O_s^1 : x_s(i_s) > 0\}$$

$$\bar{J}_s = \{j_s \in J_s \mid \exists y_s \in O_s^2 : y_s(j_s) > 0\}.$$

**Proof.** The above sets  $O_s^1$ ,  $O_s^2$ ,  $\bar{O}_s^1$ ,  $\bar{O}_s^2$  are nonempty by lemma 32. One can also show that these sets are polytopes by using the linearity of  $c_s$  in both components, which can be found in most elementary books on the theory of matrix games. The last part of the statement is shown in Gale & Sherman [1950] and Bohnenblust et al. [1950].  $\square$

**Definition 34** *For  $s \in S$ ,  $x_s \in X_s$ ,  $y_s \in Y_s$  let*

$$V_s(x_s, y_s) := \sum_{t \in S} p_s(t \mid x_s, y_s) \cdot v_t.$$

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

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

Here  $V_s(x_s, y_s)$  is the expected value after transition from state  $s$  with regard to the pair of mixed actions  $(x_s, y_s)$ .

The following lemma intuitively says that player 1 can guarantee that the value does not decrease in expectation after transition, while player 2 can make sure that the value does not increase in expectation after transition.

**Lemma 35** *For any  $s \in S$ , there exist  $x_s \in X_s$  and  $y_s \in Y_s$  such that*

$$V_s(x_s, y'_s) \geq v_s \geq V_s(x'_s, y_s) \quad \forall x'_s \in X_s, \forall y'_s \in Y_s.$$

**Proof.** Let  $s \in S$ . In view of lemma 32, there exist  $x_s \in X_s$ ,  $y_s \in Y_s$ , and a unique  $C_s \in \mathbb{R}$  such that

$$c_s(x_s, y'_s) \geq C_s \geq c_s(x'_s, y_s) \quad \forall x'_s \in X_s, \forall y'_s \in Y_s.$$

So we have to show that  $v_s = C_s$ . Assume by way of contradiction that  $v_s > C_s$ ; the proof is similar when  $v_s < C_s$  is assumed.

Let

$$d := v_s - C_s > 0.$$

We derive a contradiction by showing that player 1 does not have  $\varepsilon$ -optimal strategies for initial state  $s$ , for any  $\varepsilon \in (0, d)$ . Let  $\varepsilon \in (0, d)$  and take an arbitrary strategy  $\pi \in \Pi$ . Recall that the mixed action prescribed by  $\pi$  for stage 1 in the initial state  $s$  is denoted by  $\pi_s$ . Consider a strategy  $\sigma^\varepsilon$  for player 2 which prescribes to play  $y_s$  at stage 1 in state  $s$  and to play a  $((d - \varepsilon)/2)$ -optimal strategy afterwards. Then

$$V_s(\pi_s, y_s) \leq C_s = v_s - d.$$

From stage 2 on player 2 plays a  $((d - \varepsilon)/2)$ -optimal strategy, so if  $s^2$  denotes the random variable for the state at stage 2, then

$$\begin{aligned} \gamma_s(\pi, \sigma^\varepsilon) &\leq \mathcal{E}_{s\pi\sigma^\varepsilon} \left( v_{s^2} + \frac{d - \varepsilon}{2} \right) \\ &= \sum_{t \in S} p_s(t | \pi_s, y_s) \cdot \left( v_t + \frac{d - \varepsilon}{2} \right) \\ &= V_s(\pi_s, y_s) + \frac{d - \varepsilon}{2} \\ &\leq (v_s - d) + \frac{d - \varepsilon}{2} \\ &< v_s - \varepsilon. \end{aligned}$$

Thus player 1 cannot have an  $\varepsilon$ -optimal strategy for initial state  $s$ , which is a contradiction.  $\square$

We will deal with restricted games derived from the original game  $\Gamma$ . Assume that  $S' \subset S$  is a non-empty set of states and  $X'_s \subset X_s$ ,  $Y'_s \subset Y_s$  are nonempty polytopes for all  $s \in S'$ . If all pairs of mixed actions in  $X'_s \times Y'_s$ , for any  $s \in S'$ , only induce transitions to states in  $S'$ , then we may define a restricted game  $\Gamma'$ , derived from

the original game  $\Gamma$ , 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  $s \in S'$ . Let  $\Pi'$  and  $\Sigma'$  denote the sets of these strategies. The stationary strategy spaces in  $\Gamma'$  are  $X' := \times_{s \in S'} X'_s$  and  $Y' := \times_{s \in S'} Y'_s$ . For the restricted game  $\Gamma'$ , for any  $\beta \in (0, 1)$ , using the continuity of the  $\beta$ -discounted reward on the polytope  $X' \times Y'$ , it can be shown similarly to theorem 20 that the  $\beta$ -discounted value  $v'_\beta$  exists and both players have stationary  $\beta$ -discounted optimal strategies. Moreover,  $x \in X'$  is  $\beta$ -discounted optimal in  $\Gamma'$  if and only if

$$v'_\beta \leq (1 - \beta) \cdot r(x, y) + \beta \cdot P(x, y) \cdot v'_\beta \quad \forall y \in Y'. \quad (3.1)$$

Theorem 21 applies for  $\Gamma'$  as well, so  $\lim_{\beta \uparrow 1} v'_\beta$  exist. Let

$$v' := \lim_{\beta \uparrow 1} v'_\beta. \quad (3.2)$$

Note that we do not claim that  $v'$  is the average value of  $\Gamma'$  as in theorem 22 for the original game, because even though the players only observe pure actions, these do not correspond one-to-one to extreme points of the restricted spaces of mixed actions. However one can show, by using an appropriate sequence of discount factors as in Mertens & Neyman [1981], that, against any fixed strategy in  $\Pi'$ , for any  $\varepsilon > 0$  player 2 can make sure that player 1's reward is at most  $v' + \varepsilon$ , namely

$$\sup_{\pi \in \Pi'} \inf_{\sigma \in \Sigma'} \gamma_s(\pi, \sigma) \leq v'_s \quad \forall s \in S'. \quad (3.3)$$

### 3.3 The construction

Let

$$\begin{aligned} X_s^* &:= \{x_s \in X_s \mid V_s(x_s, y_s) \geq v_s \quad \forall y_s \in Y_s\} & \forall s \in S, & \quad X^* := \times_{s \in S} X_s^*, \\ Y_s^* &:= \{y_s \in Y_s \mid V_s(x_s, y_s) = v_s \quad \forall x_s \in X_s^*\} & \forall s \in S, & \quad Y^* := \times_{s \in S} Y_s^*. \end{aligned}$$

Note the asimilarity in the definitions of  $X_s^*$  and  $Y_s^*$ ,  $s \in S$ . In view of lemmas 35 and 33, for all  $s \in S$ , the sets  $X_s^*$ ,  $Y_s^*$  are nonempty polytopes and there exists a  $J_s^* \subset J_s$  such that  $Y_s^* = \text{conv}(J_s^*)$ , where  $\text{conv}$  stands for the convex hull of a set. Let

$$J^* := \times_{s \in S} J_s^*.$$

As in section 3.2, we may define a restricted game  $\Gamma^*$ , derived from the original game  $\Gamma$ , 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  $s \in S$ . The sets of these strategies are denoted by  $\Pi^*$  and  $\Sigma^*$ . For the restricted game  $\Gamma^*$ , let  $v_\beta^*$  denote the  $\beta$ -discounted value and let  $v^* := \lim_{\beta \uparrow 1} v_\beta^*$ .

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  $\beta$ -discounted optimal strategies  $x^\beta \in X^*$  in the restricted game  $\Gamma^*$  such that the sets  $\{i_s \in I_s \mid x_{\beta s}(i_s) > 0\}$ ,  $s \in S$ , 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.

If  $Z$  is a polytope then let  $\text{Relint}(Z)$  denote 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.

**Theorem 36** *Assume that player 1 has an optimal strategy in  $\Gamma$ .*

- (a) *For any  $\beta \in \mathcal{B}$ , let  $x_\beta \in X^*$  be a  $\beta$ -discounted optimal strategy in the restricted game  $\Gamma^*$  and let  $x \in \text{Relint}(X^*)$ . Then, for any  $\varepsilon > 0$ , if  $\beta \in \mathcal{B}$ ,  $\tau \in (0, 1)$  are sufficiently large then the stationary strategy*

$$x_\beta^\tau := \tau \cdot x_\beta + (1 - \tau) \cdot x \in X^*$$

*is  $\varepsilon$ -optimal in  $\Gamma$ .*

- (b) *Let  $\varepsilon_n$ ,  $n \in \mathbb{N}$ , be an arbitrary monotonously decreasing sequence converging to 0. Let  $x_n \in X^*$  be  $\varepsilon_n$ -optimal in  $\Gamma$  for all  $n \in \mathbb{N}$ . Then there exist 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, and so on, is optimal in  $\Gamma$ .*

*A similar statement holds for player 2 as well.*

With the help of the following example we provide an illustration of the above constructions. Furthermore, this example also demonstrates that the existence of optimal strategies does not yield the existence of stationary optimal strategies.

**Example 37**

	$L$	$R$
$T$	0	2 *
$B$	1	0
	1	

The value for 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).

Following the construction for stationary  $\varepsilon$ -optimal strategies, we have  $X^* = X$ ,  $Y^* = \{(1, 0)\}$ . Now the unique  $\beta$ -discounted optimal strategy of player 1 in  $\Gamma^*$  is  $x_\beta = (0, 1)$  for all  $\beta \in (0, 1)$ . The role of  $x_\beta$  is to play well as long as player 2 plays in the restricted game  $\Gamma^*$ , 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

$x \in \text{Relint}(X^*)$ , 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  $x_\beta^\tau$  is  $\varepsilon$ -optimal for large  $\tau$  and  $\beta$  indeed, as the stationary strategies  $(p, 1 - p)$  are  $\varepsilon$ -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 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  $v = 1$ .

A Markov optimal strategy can be constructed as in theorem 36. The idea is to increase  $\beta$  and  $\tau$  simultaneously during the play so that player 1 plays better and better in the restricted game. However,  $\tau$  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  $\varepsilon_n$ -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 36. 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. (A rigorous proof for the optimality of  $f$  can be done by using techniques as in chapter 5).  $\triangleleft$

### 3.4 The proof

The following lemma clarifies why the sets  $X^*$  and  $Y^*$  play an important role when player 1 has an optimal strategy in the original game  $\Gamma$ . This lemma states that if  $\pi$  is an optimal strategy for player 1 in  $\Gamma$  then, for any present state  $s \in S$  and past history with a positive occurrence probability with respect to  $(\pi, \sigma)$  for some  $\sigma \in \Sigma^*$ , the strategy  $\pi$  prescribes a mixed action belonging to  $X_s^*$ . In other words, if player 2 uses a strategy  $\sigma \in \Sigma^*$  then the optimal strategy  $\pi$  will behave as a strategy in  $\Pi^*$ .

**Lemma 38** *Let  $\pi \in \Pi$  be an optimal strategy for player 1 in the game  $\Gamma$ . For  $s \in S$ , let*

$$U_s := \{(h, t) \in H_s \times S \mid \mathcal{P}_{s\pi\sigma}(h) > 0 \text{ and } \mathcal{P}_{s\pi\sigma}(t|h) > 0 \text{ for some } \sigma \in \Sigma^*\},$$

where  $\mathcal{P}_{s\pi\sigma}(t|h)$  is the probability that, with respect to  $(\pi, \sigma)$ , the new state becomes state  $t$  after history  $h$ . Then  $\pi_t(h) \in X_t^*$  for all  $(h, t) \in U_s$  and for all  $s \in S$ .

**Proof.** Suppose the opposite. Then, for some  $s \in S$ , 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  $\sigma \in \Sigma^*$  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

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

By lemma 35, there also exists a  $y \in Y$  such that  $V_z(x_z, y_z) \leq v_z$  for all  $x_z \in X_z$  and  $z \in S$ .

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

$$\delta \in (0, \mathcal{P}_{s^1 \pi \sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1 \pi \sigma}(t|\bar{h}^n) \cdot \tau).$$

Let  $\sigma^\delta \in \Sigma$  be the strategy that prescribes to play as follows: play  $\sigma$  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  $y_{s^{n+1}}$ ; and finally, play a  $\delta$ -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 that above specified property, by the definitions of  $X^*$  and  $Y^*$  we have

$$\mathcal{E}_{s^1 \pi \sigma^\delta}(v_{s^{n+1}}) = v_{s^1},$$

and by the used mixed actions at stage  $n + 1$

$$\mathcal{E}_{s^1 \pi \sigma^\delta}(v_{s^{n+2}}) \leq \mathcal{E}_{s^1 \pi \sigma^\delta}(v_{s^{n+1}}) - \mathcal{P}_{s^1 \pi \sigma^\delta}(\bar{h}^n) \cdot \mathcal{P}_{s^1 \pi \sigma}(t|\bar{h}^n) \cdot \tau.$$

From stage  $n + 2$  player 2 plays a  $\delta$ -best reply, so the choice of  $\delta$  yields

$$\begin{aligned} \gamma_{s^1}(\pi, \sigma^\delta) &\leq \mathcal{E}_{s^1 \pi \sigma^\delta}(v_{s^{n+2}}) + \delta \\ &\leq \mathcal{E}_{s^1 \pi \sigma^\delta}(v_{s^{n+1}}) - \mathcal{P}_{s^1 \pi \sigma^\delta}(\bar{h}^n) \cdot \mathcal{P}_{s^1 \pi \sigma}(t|\bar{h}^n) \cdot \tau + \delta \\ &= v_{s^1} - \mathcal{P}_{s^1 \pi \sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1 \pi \sigma}(t|\bar{h}^n) \cdot \tau + \delta \\ &< v_{s^1}, \end{aligned}$$

which contradicts the optimality of  $\pi$ .  $\square$

The condition that player 1 has an optimal strategy in  $\Gamma$  is only needed for the next lemma. We show that, if player 1 has an optimal strategy in  $\Gamma$ , then he can guarantee a reward at least  $v$  in the restricted game  $\Gamma^*$ . This is based on the facts that optimal strategies of player 1 guarantee a reward at least the value  $v$  in the original game  $\Gamma$  and, in view of the previous lemma, they can only prescribe mixed actions in  $X_s^*$ , if the play is in any state  $s$ , against any strategy of player 2 in  $\Sigma^*$ . The second part of the statement says that player 1 cannot guarantee more than the limit of the  $\beta$ -discounted values in  $\Gamma^*$ .

**Lemma 39** *Suppose that player 1 has an optimal strategy  $\pi \in \Pi$ . Then*

$$v_s \leq \sup_{\pi^* \in \Pi^*} \inf_{\sigma^* \in \Sigma^*} \gamma_s(\pi^*, \sigma^*) \leq v_s^* \quad \forall s \in S.$$

**Proof.** The second inequality follows from (3.3), so we only have to show the first one.

Let  $s \in S$  be arbitrary and let  $U_s$  be as in lemma 38. Take an arbitrary  $x \in X^*$ . By using lemma 38, we may define a strategy  $\pi^* \in \Pi^*$  as follows: for  $h \in H$  and  $t \in S$  let

$$\pi_t^*(h) := \begin{cases} \pi_t(h) & \text{if } (h, t) \in U_s \\ x_t & \text{otherwise.} \end{cases}$$

Then by the optimality of  $\pi$  and by the definition of  $\pi^*$ , we have

$$v_s \leq \gamma_s(\pi, \sigma^*) = \gamma_s(\pi^*, \sigma^*) \quad \forall \sigma^* \in \Sigma^*,$$

which implies the first inequality.  $\square$

The next result shows the effectiveness of the  $\beta$ -discounted optimal strategies in the restricted game  $\Gamma^*$ .

**Lemma 40** *Let  $\varepsilon > 0$ . For  $\beta \in \mathcal{B}$ , let  $x_\beta \in X^*$  be a  $\beta$ -discounted optimal strategy of player 1 in  $\Gamma^*$ , and let  $y \in Y^*$ . Suppose that  $E \subset S$  is a closed set of states with respect to  $(x_\beta, y)$  for all  $\beta \in \mathcal{B}$ . Then for large  $\beta \in \mathcal{B}$*

$$\gamma_s(x_\beta, y) \geq \min_{t \in E} v_t^* - \varepsilon \quad \forall s \in E.$$

**Proof.** Using inequality (3.1) for  $\Gamma^*$  we have

$$(1 - \beta) \cdot r(x_\beta, y) + \beta \cdot P(x_\beta, y) \cdot v_\beta^* \geq v_\beta^* \quad \forall \beta \in \mathcal{B}.$$

By (2.1), multiplying this inequality with  $Q(x_\beta, y)$  yields

$$Q(x_\beta, y) \cdot r(x_\beta, y) \geq Q(x_\beta, y) \cdot v_\beta^* \quad \forall \beta \in \mathcal{B}.$$

The closedness of  $E$  implies that, for any  $s \in E$ , if  $q_s(t|x_\beta, y) > 0$  then  $t \in E$ . Hence for all  $s \in E$  and for large  $\beta \in \mathcal{B}$ , using theorem 9-(a) and the definition of  $v^*$ , we have

$$\begin{aligned} \gamma_s(x_\beta, y) &= \sum_{t \in E} q_s(t|x_\beta, y) r_t(x_\beta, y) \\ &\geq \sum_{t \in E} q_s(t|x_\beta, y) v_{\beta t}^* \\ &\geq \sum_{t \in E} q_s(t|x_\beta, y) (v_t^* - \varepsilon) \\ &\geq \min_{t \in E} v_t^* - \varepsilon, \end{aligned}$$

so the proof is complete.  $\square$

Next we discuss some properties of stationary strategies belonging to  $X^*$  or to  $\text{Relint}(X^*)$ .

**Lemma 41** *Let  $x \in X^*$  and  $y \in Y$ . Suppose  $E$  is an ergodic set with respect to  $(x, y)$ . Then  $v_s = v_t$  for all  $s, t \in E$ . Furthermore, if  $x \in \text{Relint}(X^*)$  then necessarily  $y_s \in Y_s^*$  for all  $s \in E$ .*

**Proof.** By  $x \in X^*$  and by the closedness of  $E$  for  $(x, y)$  we obtain

$$v_s \leq V_s(x_s, y_s) = \sum_{t \in E} p_s(t|x_s, y_s) v_t \quad \forall s \in E.$$

Let  $\bar{E} := \{s \in E | v_s = \max_{t \in E} v_t\}$ . The above inequalities imply that  $\bar{E}$  is a closed set of states for  $(x, y)$ , so since  $E$  is an ergodic set for  $(x, y)$ , we have  $\bar{E} = E$ . Therefore  $v_s = v_t =: v_E$  for all  $s, t \in E$ .

Now suppose that  $x \in \text{Relint}(X^*)$ . Then  $(\bar{x}_s, y_s)$  only induces transitions to states in  $E$  for any  $\bar{x}_s \in X_s^*$ ,  $s \in E$ , hence

$$V_s(\bar{x}_s, y_s) = \sum_{t \in E} p_s(t|\bar{x}_s, y_s) v_E = v_E = v_s \quad \forall \bar{x}_s \in X_s^*, \forall s \in E,$$

which implies that  $y_s \in Y_s^*$  for all  $s \in E$ .  $\square$

An important property of convex combinations of stationary strategies is stated in the next lemma.

**Lemma 42** For  $\tau \in (0, 1)$ ,  $x^1, x^2 \in X$  let  $x^\tau := \tau x^1 + (1 - \tau)x^2$ . Suppose that  $E$  is an ergodic set with respect to  $(x^\tau, y)$  for some  $y \in Y$ . Let  $\varepsilon > 0$  and  $d \in \mathbb{R}$ . If

$$\gamma_s(x^1, y) \geq d \quad \forall s \in E$$

then for sufficiently large  $\tau$

$$\gamma_s(x^\tau, y) \geq d - \varepsilon \quad \forall s \in E.$$

**Proof.** Let  $\delta \in (0, 1)$ . Since

$$\gamma_s(x^1, y) \geq d \quad \forall s \in E,$$

there exists a  $K^\delta$  satisfying

$$\frac{1}{N} \sum_{n=1}^N \mathcal{E}_{sx^1y}(R_n) \geq d - \delta \quad \forall N \geq K^\delta, \forall s \in E,$$

where  $R_n$  denotes the random variable for the payoff at stage  $n$ . Choose a sufficiently large  $\tau \in (0, 1)$  such that

$$\tau^{K^\delta} \geq 1 - \delta.$$

The strategy  $x^\tau$  can be interpreted as playing  $x^1$  with probability  $\tau$  and  $x^2$  with probability  $1 - \tau$  at each stage, so the last inequality means that  $x^1$  will be played at each  $K^\delta$  consecutive stages with probability at least  $1 - \delta$ . Hence with probability at least  $1 - \delta$ , the average of the expected payoffs will be at least  $d - \delta$  for any  $K^\delta$  consecutive stages. Therefore, if  $r$  denotes the smallest payoff in the game, then for small  $\delta > 0$  we have

$$\gamma_s(x^\tau, y) \geq (1 - \delta)(d - \delta) + \delta r \geq d - \varepsilon \quad \forall s \in E,$$

so the proof is complete.  $\square$

Now we are ready to prove theorem 36.

**Proof of theorem 36-(a).**

We only show the statement for player 1. First notice that by the convexity of  $X^*$  and by  $x \in \text{Relint}(X^*)$  we have  $x_\beta^\tau \in \text{Relint}(X^*)$  for all  $\beta \in \mathcal{B}$  and  $\tau \in (0, 1)$ .

Since there are only finitely many pure stationary strategies, by theorem 16-(b), it suffices to show that, for all  $j \in J$ , if  $\tau \in (0, 1)$  and  $\beta \in \mathcal{B}$  are large then

$$\gamma(x_\beta^\tau, j) \geq v - \varepsilon 1_{|S|},$$

where  $1_{|S|} = (1, \dots, 1) \in \mathbb{R}^{|S|}$ . Take a  $j \in J$  and let  $E \subset S$  be an arbitrary ergodic set with respect to  $(x_\beta^\tau, j)$ . We start with showing that for large  $\tau \in (0, 1)$ ,  $\beta \in \mathcal{B}$  we have

$$\gamma_s(x_\beta^\tau, j) \geq v_s - \varepsilon \quad \forall s \in E. \quad (3.4)$$

Since  $x_\beta^\tau \in \text{Relint}(X^*)$ , by lemma 41 we obtain  $v_s = v_t := v_E$  for all  $s, t \in E$  and  $j_s \in J_s^*$  for all  $s \in E$ . Let  $j_s^* := j_s$  for all  $s \in E$  and let  $j_s^* \in J_s^*$  for all  $s \notin E$ ; so  $j^* \in J^*$ . By the definition of  $x_\beta^\tau$  and by the properties of  $\mathcal{B}$ , the set of states  $E$  is closed with respect to  $(x_\beta, j)$  for all  $\beta \in \mathcal{B}$ , so with respect to  $(x_\beta, j^*)$  for all  $\beta \in \mathcal{B}$  as well, thus applying lemma 40 for  $\Gamma^*$  and using lemma 39 yield that for large  $\beta \in \mathcal{B}$

$$\gamma_s(x_\beta, j) = \gamma_s(x_\beta, j^*) \geq \min_{t \in E} v_t^* - \frac{1}{2} \varepsilon \geq \min_{t \in E} v_t - \frac{1}{2} \varepsilon = v_E - \frac{1}{2} \varepsilon \quad \forall s \in E.$$

Now lemma 42 yields that for large  $\tau \in (0, 1)$  and for large  $\beta \in \mathcal{B}$

$$\gamma_s(x_\beta^\tau, j) \geq v_E - \varepsilon = v_s - \varepsilon \quad \forall s \in E,$$

which proves (3.4).

Using that  $x_\beta^\tau \in X^*$  we have

$$P(x_\beta^\tau, j) v \geq v,$$

therefore inductively we obtain for all  $n \in \mathbb{N}$

$$P^n(x_\beta^\tau, j) v \geq v,$$

which by the definition of  $Q$  yields

$$Q(x_\beta^\tau, j) v \geq v.$$

For any  $s \in S$ ,  $q_s(t|x_\beta^\tau, j) > 0$  implies that  $t \in E$  for some ergodic set  $E$  with respect to  $(x_\beta^\tau, j)$ , hence by theorem 9-(c) and (3.4) for large  $\tau \in (0, 1)$  and  $\beta \in \mathcal{B}$  we obtain

$$\begin{aligned} \gamma(x_\beta^\tau, j) &= Q(x_\beta^\tau, j) \gamma(x_\beta^\tau, j) \\ &\geq Q(x_\beta^\tau, j) (v - \varepsilon 1_{|S|}) \\ &= Q(x_\beta^\tau, j) v - \varepsilon 1_{|S|} \\ &\geq v - \varepsilon 1_{|S|}, \end{aligned}$$

which completes the proof.  $\square$

**Proof of theorem 36-(b).**

We only show the statement for player 1. For  $n \in \mathbb{N}$ , let  $\varepsilon_n$  and  $x_n$  be as in theorem 36-(b). We will choose an appropriate sequence  $K_n$  in  $\mathbb{N}$  so that the Markov strategy described in theorem 36-(b) is optimal. Using the results of Bewley & Kohlberg [1978] (theorem 5.2), for any  $n \in \mathbb{N}$ , there exists a stage  $\bar{K}_n$  such that

$$\frac{1}{N} \sum_{m=1}^N \mathcal{E}_{sx_n\sigma}(R_m) \geq v_s - 2\varepsilon_n \quad \forall N \geq \bar{K}_n, \forall s \in S, \forall \sigma \in \Sigma, \quad (3.5)$$

where  $R_m$  denotes the random variable for the payoff at stage  $m$ . Let  $r$  denote the smallest payoff in the game minus  $\varepsilon_1$ :

$$r := \min_{i_s \in I_s, j_s \in J_s, s \in S} r_s(i_s, j_s) - \varepsilon_1.$$

Given  $\bar{K}_n$ ,  $n \in \mathbb{N}$ , choose an arbitrary  $K_1 \geq \bar{K}_1$  and choose  $K_n \geq \bar{K}_n$ ,  $n \geq 2$ , inductively so that

$$\frac{\sum_{l=1}^n K_l \cdot (v_s - 2\varepsilon_l) + \bar{K}_{n+1} \cdot r}{\sum_{l=1}^n K_l + \bar{K}_{n+1}} \geq v_s - 2\varepsilon_{n-1} \quad \forall s \in S, \forall n \geq 2. \quad (3.6)$$

By the definition of  $r$ , inequality (3.6) implies

$$\frac{\sum_{l=1}^n K_l \cdot (v_s - 2\varepsilon_l)}{\sum_{l=1}^n K_l} \geq v_s - 2\varepsilon_{n-1} \quad \forall s \in S, \forall n \geq 2. \quad (3.7)$$

Let  $s^1$  be an arbitrary initial state and let  $s^m$ ,  $m \geq 2$ , denote the random variable for the state at stage  $m$ . Let the Markov strategy  $f$  be as in theorem 36-(b). Using that  $f$  only prescribes mixed actions in  $X_s^*$ ,  $s \in S$ , we have for all  $\sigma \in S$  that

$$\mathcal{E}_{s^1 f \sigma}(v_{s^m}) \geq v_{s^1} \quad \forall m \in \mathbb{N}, \forall \sigma \in \Sigma. \quad (3.8)$$

The strategy  $x_1$  is to be played at stages  $1, 2, \dots, K_1$ , hence using (3.5)

$$\sum_{m=1}^{K_1} \mathcal{E}_{s^1 f \sigma}(R_m) \geq K_1 \cdot (v_{s^1} - 2\varepsilon_1) \quad \forall \sigma \in \Sigma; \quad (3.9)$$

while the strategy  $x_n$ ,  $n \geq 2$ , is to be played at stages  $w_n, \dots, w_n + K_n - 1$ , where  $w_n := \sum_{l=1}^{n-1} K_l + 1$ , hence using (3.5) and (3.8) we have

$$\begin{aligned} \sum_{m=w_n}^{w_n+K_n-1} \mathcal{E}_{s^1 f \sigma}(R_m) &\geq L \cdot (\mathcal{E}_{s^1 f \sigma}(v_{s^{w_n}}) - 2\varepsilon_n) \\ &\geq L \cdot (v_{s^1} - 2\varepsilon_n) \quad \forall K_n \geq L \geq \bar{K}_n, \forall \sigma \in \Sigma. \end{aligned} \quad (3.10)$$

As a special case we have for all  $n \geq 2$  that

$$\sum_{m=w_n}^{w_n+K_n-1} \mathcal{E}_{s^1 f \sigma}(R_m) \geq K_n \cdot (v_{s^1} - 2\varepsilon_n) \quad \forall \sigma \in \Sigma. \quad (3.11)$$

Assume first that stage  $N$  has the property that for some  $n(N) \geq 2$  we have

$$\sum_{l=1}^{n(N)} K_l < N \leq \sum_{l=1}^{n(N)} K_l + \bar{K}_{n(N)+1};$$

which means that at stage  $N$  the strategy  $x_{n(N)+1}$  is used and it has not yet been used at more than  $\bar{K}_{n(N)+1}$  stages. Using (3.9), (3.11), by the definition of  $r$  and by (3.6) we have for all  $\sigma \in \Sigma$

$$\begin{aligned} \frac{1}{N} \sum_{m=1}^N \mathcal{E}_{s^1 f \sigma}(R_m) &= \frac{\sum_{l=1}^{n(N)} \sum_{m=w_l}^{w_l+K_l-1} \mathcal{E}_{s^1 f \sigma}(R_m) + \sum_{m=w_{n(N)+1}}^N \mathcal{E}_{s^1 f \sigma}(R_m)}{N} \\ &\geq \frac{\sum_{l=1}^{n(N)} K_l \cdot (v_{s^1} - 2\varepsilon_l) + \left(N - \sum_{l=1}^{n(N)} K_l\right) \cdot r}{N} \\ &\geq \frac{\sum_{l=1}^{n(N)} K_l \cdot (v_{s^1} - 2\varepsilon_l) + \bar{K}_{n(N)+1} \cdot r}{\sum_{l=1}^{n(N)} K_l + \bar{K}_{n(N)+1}} \\ &\geq v_{s^1} - 2\varepsilon_{n(N)-1}. \end{aligned} \quad (3.12)$$

Assume now that stage  $N$  has the property that for some  $n(N) \geq 2$  we have

$$\sum_{l=1}^{n(N)} K_l + \bar{K}_{n(N)+1} < N \leq \sum_{l=1}^{n(N)} K_l + K_{n(N)+1};$$

which means that at stage  $N$  the strategy  $x_{n(N)+1}$  is used and it has already been used more than  $\bar{K}_{n(N)+1}$  stages. Using (3.9), (3.11), (3.10), by (3.7) we have for all  $\sigma \in \Sigma$

$$\begin{aligned} \frac{1}{N} \sum_{m=1}^N \mathcal{E}_{s^1 f \sigma}(R_m) &= \frac{\sum_{l=1}^{n(N)} \sum_{m=w_l}^{w_l+K_l-1} \mathcal{E}_{s^1 f \sigma}(R_m) + \sum_{m=w_{n(N)+1}}^N \mathcal{E}_{s^1 f \sigma}(R_m)}{N} \\ &\geq \frac{\sum_{l=1}^{n(N)} K_l \cdot (v_{s^1} - 2\varepsilon_l) + \left(N - \sum_{l=1}^{n(N)} K_l\right) \cdot (v_{s^1} - 2\varepsilon_{n(N)+1})}{N} \\ &\geq \frac{\sum_{l=1}^{n(N)} K_l \cdot (v_{s^1} - 2\varepsilon_l)}{\sum_{l=1}^{n(N)} K_l} \\ &\geq v_{s^1} - 2\varepsilon_{n(N)-1}. \end{aligned} \quad (3.13)$$

Inequalities (3.12) and (3.13) together imply that

$$\gamma_{s^1}(f, \sigma) = \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{m=1}^N \mathcal{E}_{sf\sigma}(R_m) \geq v_{s^1} \quad \forall \sigma \in \Sigma.$$

Since the initial state  $s^1$  was arbitrary, we have shown that  $f$  is optimal in  $\Gamma$ .  $\square$

The next example shows that the existence of stationary  $\varepsilon$ -optimal strategies, for all  $\varepsilon > 0$ , does not imply the existence of optimal strategies.

**Example 43**

	$L$	$R$	
$T$	1 *	0	
$B$	0 *	1 *	
			1

Here the value for initial state 1 is  $v_1 = 1$ . The stationary strategy for player 1 which, in state 1, prescribes action  $T$  with probability  $1 - \varepsilon$  and action  $B$  with probability  $\varepsilon$  is  $\varepsilon$ -optimal for all  $\varepsilon > 0$ . However, we show that player 1 does not have optimal strategies for initial state 1. Take an arbitrary strategy  $\pi$ . We show that player 2 can make sure that the reward is strictly less than 1. Indeed, player 2 has to choose action  $R$  as long as  $\pi$  prescribes action  $T$  with probability 1 and to play action  $L$  at the first stage when  $\pi$  prescribes action  $B$  with a positive probability. Then either entry  $(T, R)$  is played forever or absorption occurs with payoff zero with a positive probability, thus the reward is strictly less than 1 indeed.  $\triangleleft$

### 3.5 Concluding remarks

*Remarks on the restricted game  $\Gamma^*$ .* In lemma 39 we showed that  $v_s^* \geq v_s$  for all  $s \in S$ . In fact, this is the only statement for which we needed the condition that player 1 has an optimal strategy. Therefore if in a zero-sum game  $v_s^* \geq v_s$  holds for all  $s \in S$ , then stationary  $\varepsilon$ -optimal strategies,  $\varepsilon > 0$ , and Markov optimal strategies can be constructed exactly as before. It also means that  $v_s^* \geq v_s$  for all  $s \in S$  holds if and only if player 1 has an optimal strategy.

We also remark that one can find examples in which, although player 1 has an optimal strategy,  $v_s^* > v_s$  holds for some state  $s \in S$ . Such an example is presented next.

**Example 44**

	$L$	$R$	
$T$	1	0 *	
$B$	0 *	1 *	
			1

The value for initial state 1 is  $v_1 = 0$ . Indeed, one can easily verify that against the stationary strategy  $y = (1 - \varepsilon, \varepsilon) \in Y$ , for any  $\varepsilon \in (0, 1)$ , player 1 cannot get more than  $\varepsilon$ . Notice that any strategy of player 1 is optimal. We have  $X^* = X$ ,  $Y^* = \{(1, 0)\}$ , hence  $v_1^* = 1 > v_1$ . (Obviously, for any absorbing state  $s$  in the game  $v_s = v_s^*$ .)  $\triangleleft$

However, if  $E$  is an ergodic set with respect to some  $(x, y) \in \text{Relint}(X^*) \times \text{Relint}(Y^*)$  then there exists a state  $s \in E$  such that  $v_s^* = v_E$  (recall that the value  $v$  is a constant on  $E$  by lemma 41). To see this one can argue as follows. Suppose to the contrary that  $v_s^* \geq v_E + \mu$  for all  $s \in E$ , where  $\mu > 0$ . Let  $x_\beta^\tau \in \text{Relint}(X^*)$  be defined as in theorem 36-(a). Then lemmas 40 and 42 imply that for large  $\tau$  and  $\beta$  we have

$$\gamma_s(x_\beta^\tau, j) \geq \min_{t \in E} v_t^* - \frac{\mu}{2} \geq v_E + \frac{\mu}{2} \quad \forall s \in E, \forall j \in J^*. \quad (3.14)$$

Here we used that there are only finitely many pure stationary strategies. Let player 1 play the strategy  $\pi^\delta$ ,  $\delta > 0$ , which prescribes to play as follows: play  $x_\beta^\tau$  as long as player 2 chooses actions in  $J_s^*$ ,  $s \in E$ , and start playing a  $\delta$ -optimal strategy as soon as player 2 chooses an action in  $J_s \setminus J_s^*$  in some state  $s \in E$ . Note that if player 2 always chooses actions in  $J_s^*$ ,  $s \in E$ , then (3.14) assures that the reward is at least  $v_E + \frac{\mu}{2}$  (recall that against a stationary strategy there always exists a pure stationary best reply). On the other hand, if player 2 chooses an action in  $J_s \setminus J_s^*$  in some state  $s \in E$ , then one can show that  $x_{\beta_s}^\tau \in \text{Relint}(X_s^*)$  yields that the original value  $v$  increases in expectation by at least some  $\nu > 0$ , so if  $\delta \in (0, \frac{\nu}{2})$ , by the definition of  $\pi^\delta$ , the reward is at least  $v_E + \frac{\nu}{2}$  in this case. Therefore  $\pi^\delta$ , with  $\delta \in (0, \frac{\nu}{2})$ , guarantees a reward at least  $v_E + \frac{1}{2} \min(\mu, \nu) > v_E$ , which contradicts the definition of the value. So we have shown that  $v_s^* = v_E$  holds for some state  $s \in E$ .

*Optimal strategies for particular initial states.* We briefly discuss a generalization of the results, which concerns strategies that are only optimal for particular initial states. Let  $\tilde{S}$  denote the set of states for which player 1 has an optimal strategy. First note that, in each stochastic game, there exists at least one initial state for which player 1 has optimal strategies (cf. Thuijsman & Vrieze [1993]), so the set  $\tilde{S}$  is always nonempty. Using similar techniques as before, one can show that, for any  $\varepsilon > 0$ , player 1 has a strategy  $\xi^\varepsilon$  which for all initial states  $\alpha \in \tilde{S}$  satisfies: (a)  $\xi^\varepsilon$  is  $\varepsilon$ -optimal, (b)  $\xi^\varepsilon$  is stationary until leaving  $\tilde{S}$ , (c) there exist stationary best replies of player 2 against  $\xi^\varepsilon$ , (d) the probability of ever leaving  $\tilde{S}$  is zero with respect to  $(\xi^\varepsilon, \sigma)$  and initial state  $\alpha$ , if  $\sigma$  is a best reply. The weakness of this result is mainly due to the fact that stationary strategies are not effective in states outside  $\tilde{S}$ , so player 1 may have to start playing a history dependent  $\delta$ -optimal strategy if the play leaves  $\tilde{S}$ , for some  $\delta > 0$ . Furthermore, one can also show that player 1 has a strategy  $\chi$  which for all initial states  $\alpha \in \tilde{S}$  satisfies: (e)  $\chi$  is optimal, (f)  $\chi$  is Markov until leaving  $\tilde{S}$ , (g) there exist Markov best replies of player 2 against  $\chi$ , (h) the probability of ever leaving  $\tilde{S}$  is zero with respect to  $(\chi, \sigma)$  and initial state  $\alpha$ , if  $\sigma$  is a best reply.

We remark here that Markov best replies do not necessarily exist against a Markov strategy, but the Markov strategy  $\chi$  can be constructed so that (f) holds. At this point it is important that if  $E$  is an ergodic set for some  $(x, y) \in \text{Relint}(X^*) \times \text{Relint}(Y^*)$

then, although  $v_s^* \geq v_E$  for all  $s \in E$ , there exists a state  $s \in E$  such that  $v_s^* = v_E$ , as discussed above. This guarantees that player 2 can get satisfied in the restricted game when playing against  $\chi$ .

**Example 45**

	$L_1$	$R_1$	
$T_1$	$\frac{1}{4}$	$\frac{1}{4}$	
	1	2	
$B_1$	$\frac{1}{4}$	$\frac{1}{4}$	
	$(2, 4)$	$(2, 4)$	
	1		

	$L_2$	$R_2$	
$T_2$	0	1	
	2	2	
$B_2$	1	0	
	3	4	
	2		

1	3	3
---	---	---

0	4	4
---	---	---

This example clarifies the existence of such “almost stationary”  $\varepsilon$ -optimal strategies and “almost Markov” optimal strategies for initial states in  $\tilde{S}$ . The two “mixed” transition vectors in entries  $(B_1, L_1)$  and  $(B_1, R_1)$  lead to state 2 with probability  $\frac{1}{2}$  and to state 4 with probability  $\frac{1}{2}$ . For the sake of simplicity, we only focus on the possible simplifications by “almost stationary”  $\varepsilon$ -optimal strategies. Notice that if the initial state is state 2 then this game reduces to the Big Match (cf. example 23). So here the value is  $v = (\frac{1}{4}, \frac{1}{2}, 1, 0)$ . By lemma 24-(d), for initial state 2 player 1 has no optimal strategy, so  $\tilde{S} = \{1, 3, 4\}$ . Since initial states 3, 4  $\in \tilde{S}$  are trivial, we assume the initial state to be  $1 \in \tilde{S}$ . Consider the strategy  $\xi$  for player 1 which prescribes to play action  $T_1$  as long as the play is in state 1 and as soon as the play visits state 2 then prescribes to start playing a history dependent  $\frac{1}{8}$ -optimal strategy. This strategy  $\xi$  is optimal and clearly satisfies properties (a), (b), (c), and (d). Note that switching to a history dependent strategy when entering state 2 is crucial, because by stationary strategies player 1 could only guarantee 0 for initial state 2 (cf. lemma 24-(e)). Note also that even though  $(0, 1) \in X_1^*$ , player 1 should not choose action  $B_1$ , because it would violate property (d).

*Subgame optimality.* Note that the Markov strategy  $f$ , constructed in theorem 36-(b), is “subgame optimal”; namely the strategy  $f[h]$  is optimal for any finite history  $h \in H$ . (The stationary  $\varepsilon$ -optimal strategies,  $\varepsilon > 0$ , in theorem 36-(a) are obviously “subgame  $\varepsilon$ -optimal”, as  $x = x[h]$  for any stationary strategy  $x \in X$  and for any finite history  $h \in H$ ).



## Chapter 4

# Improving and non-improving strategies

### 4.1 Introduction

In zero-sum stochastic games the players have completely opposite interests, so it is natural to evaluate a strategy of a player by the reward it guarantees against any strategy of the opponent. So as in definition 22, for strategies  $\pi \in \Pi$  and  $\sigma \in \Sigma$  let

$$\underline{v}_s(\pi) := \inf_{\sigma' \in \Sigma} \gamma_s(\pi, \sigma') \quad \forall s \in S, \quad \underline{v}(\pi) := (\underline{v}_s(\pi))_{s \in S},$$
$$\bar{v}_s(\sigma) := \sup_{\pi' \in \Pi} \gamma_s(\pi', \sigma) \quad \forall s \in S, \quad \bar{v}(\sigma) := (\bar{v}_s(\sigma))_{s \in S}.$$

These evaluations enable us to compare strategies.

#### Definition 46

(a) A strategy  $\pi^1$  is called  $\varepsilon$ -better than  $\pi^2$ , where  $\varepsilon \geq 0$ , if  $\underline{v}(\pi^1) \geq \underline{v}(\pi^2) - \varepsilon$  holds. 0-better strategies are simply called better. A similar definition

of  $\varepsilon$ -betterness holds for strategies of player 2.

(b) A strategy  $\pi$  is called non-improving if for any history  $h \in H$  we have  $\underline{v}(\pi) \geq \underline{v}(\pi[h])$ ; otherwise  $\pi$  is called improving. Non-improving strategies for player 2 are similarly defined.

Intuitively, a non-improving strategy 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  $h \in H$ . In the following simple example we show an instance of an improving strategy.

#### Example 47

$T$	1
$B$	0
	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  $\underline{v}_1(f) = 1/2$ . 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  $\underline{v}_1(f[h]) = 1$ . Thus  $\underline{v}_1(f) < \underline{v}_1(f[h])$ , which means that  $f$  is improving.  $\triangleleft$

The main results of this chapter, which is mainly based on Flesch et. al. [1998,IV], are summarized by the following theorem, which will follow from theorem 53.

**Main Theorem 4** *In any zero-sum stochastic game, for any non-improving strategy, there exists an  $\varepsilon$ -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 Main Theorem 3 in chapter 3.

Just as in section 3.5, the above theorem and further results can also be generalized to strategies that are only non-improving for particular initial states.

In fact, the above theorem and Main Theorem 3 together have the following corollary, which shows the insufficiency of the class of non-improving strategies as well as the indispensability of improving strategies for achieving  $\varepsilon$ -optimality, for small  $\varepsilon > 0$ .

**Corollary 48** *In a zero-sum stochastic game, if a player has no stationary  $\varepsilon$ -optimal strategies for small  $\varepsilon > 0$ , then he has no optimal strategies either and all his  $\varepsilon$ -optimal strategies, with small  $\varepsilon > 0$ , are improving.*

The next example provides a illustration for the above corollary.

#### Example 49

	$L$	$R$
$T$	0	1
$B$	1 *	0 *
	1	

This example is the Big Match (cf. example 23 and lemma 24). The value for initial state 1 is  $v_1 = 1/2$  and player 1 has no stationary  $\varepsilon$ -optimal strategies for small  $\varepsilon > 0$ . Moreover, player 1 has no optimal strategy either. Now consider the history dependent  $\varepsilon$ -optimal strategy  $\pi^N$  as in lemma 24-(c). One can show now that the strategy  $\pi^N$  is improving, since for the history  $h = (1, T, R)$  we have  $\pi^N[h] = \pi^{N+1}$ . Furthermore, notice that there are no stationary and Markov strategies that are  $\varepsilon$ -better than  $\pi^N$ , for large  $N$  and small  $\varepsilon > 0$ , since player 1 cannot guarantee more than 0 by stationary strategies and Markov strategies, as stated in lemma 24-(e).  $\triangleleft$

The next example demonstrates that the class of non-improving strategies may admit  $\varepsilon$ -optimal strategies as well, for all  $\varepsilon > 0$ , even when player 1 has no optimal strategy. So non-improving strategies are not only important for achieving optimality.

**Example 50**

	$L$	$R$
$T$	0	1 *
$B$	1 *	0 *
	1	

It is easy to check that player 1 can obtain  $\varepsilon$ -optimality by using non-improving strategies. Indeed, the stationary strategy  $x^\varepsilon = (1 - \varepsilon, \varepsilon)$  is  $\varepsilon$ -optimal for player 1 for any  $\varepsilon > 0$ . On the other hand, player 1 has no optimal strategy for initial state 1. One can argue as follows. Take an arbitrary strategy  $\pi$ . Let  $\sigma$  be the strategy for player 2 that prescribes action  $L$  whenever  $\pi$  prescribes action  $T$  with probability 1 and prescribes action  $R$  otherwise. Then whenever player 2 chooses  $L$  the cell  $(T, L)$  is played with payoff 0, and whenever player 2 chooses  $R$  absorption through the cell  $(B, R)$  has a positive probability. Hence  $\pi$  cannot guarantee reward 1, so  $\pi$  is not optimal.  $\triangleleft$

## 4.2 Preliminaries

The definitions will be very similar to those in chapter 3. Let  $\pi$  denote a fixed non-improving strategy and let

$$a := \underline{v}(\pi).$$

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

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

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

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

Let

$$\tilde{X}_s := \{x_s \in X_s \mid A_s(x_s, y_s) \geq a_s \quad \forall y_s \in Y_s\} \quad \forall s \in S, \quad \tilde{X} := \times_{s \in S} \tilde{X}_s,$$

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 51** *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  $\pi_s \in \tilde{X}_s$ . (Recall that  $\pi_s$  denotes the mixed action prescribed by  $\pi$  for stage 1 if the initial state is state  $s$ ). By the definition of  $\underline{v}_s(\pi)$

$$\underline{v}_s(\pi) = \min_{y_s \in Y_s} \sum_{t \in S} \sum_{i_s \in I_s, j_s \in J_s} [\pi_s(i_s) \cdot y_s(j_s) \cdot p_s(t|i_s, j_s)] \cdot \underline{v}_t(\pi[s, i_s, j_s]),$$

hence using the definition of  $a$  and the non-improvingness of  $\pi$  we have

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

so the proof is complete.  $\square$

As in chapter 3, if  $Z$  is a polytope then  $\text{Relint}(Z)$  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  $x \in \text{Relint}(\tilde{X})$ , 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 52** *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  $x \in \text{Relint}(\tilde{X})$*

(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  $y_s \in \text{Relint}(Y'_s)$  for all  $s \in S'$ , all states  $s \in S'$  are recurrent with respect to  $(x, y)$ .

**Proof.** Take an arbitrary  $x \in \text{Relint}(\tilde{X})$ . 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

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

$$Y'_s := \text{conv} \{J'_s\}, \quad Y' := \times_{s \in S'} Y'_s,$$

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

### 4.3 The construction

Recall that we have fixed a non-improving strategy  $\pi$  for player 1. Let  $\tilde{X}$  be as above, let  $S'$  and  $Y'$  be as in lemma 52, and let  $X' := \times_{s \in S'} \tilde{X}_s$ . In view of lemma 52, we may define a restricted stochastic game  $\Gamma'$ , as in section 3.2, in the following way. Let  $\Gamma'$  be the game, derived from  $\Gamma$ , 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  $\Gamma'$  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  $\beta$ -discounted optimal strategies  $x^\beta \in X'$  in the restricted game  $\Gamma'$  such that the sets  $\{i_s \in I_s \mid x_{\beta s}(i_s) > 0\}$ ,  $s \in S$ , 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.

The following result is very similar to theorem 36. In fact, the next theorem is a generalization of theorem 36, since optimal strategies are always non-improving, as we have already mentioned.

**Theorem 53** *Let  $\pi \in \Pi$  be a non-improving strategy in a zero-sum stochastic game. By using the strategy  $\pi$ , define  $S', X', Y'$ , and the restricted game  $\Gamma'$  as above.*

(a) *For any  $\beta \in \mathcal{B}$ , let  $x_\beta \in X'$  be a  $\beta$ -discounted optimal strategy in the restricted game  $\Gamma'$  and let  $x \in \text{Relint}(\tilde{X})$ . 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  $\varepsilon$ -better than  $\pi$  in  $\Gamma$ .

- (b) Let  $\varepsilon_n$ ,  $n \in \mathbb{N}$ , be an arbitrary monotonously decreasing sequence converging to 0. Let the stationary strategy  $x_n \in X'$  be  $\varepsilon_n$ -better than  $\pi$  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, and so on, is better than  $\pi$ .

A similar statement holds for player 2 as well.

One may read example 37 for an illustration; just instead of an optimal strategy one has to think of a non-improving strategy which guarantees the value, so  $a$  would be equal to the value  $v$ . We will now explain with the help of the following example why the proof of theorem 36 does not apply directly.

#### Example 54

$T$	0
$B$	1
	1

Consider the stationary strategy  $x$  which prescribes action  $T$  in state 1 with probability 1. Clearly, this strategy is non-improving and we have  $a_1 = 0$ . Notice that  $x$  is not effective in state 1, since by choosing action  $B$  player 1 would be better off. Formally, it means that there exists a mixed action  $\bar{x}_1 \in X_1$  in state 1 so that we have

$$A_1(\bar{x}_1, y_1) > a_1 \quad \forall y_1 \in Y_1.$$

Generally, such states cannot be treated in the same way as in the proof of theorem 36, since a version of lemma 35 would not hold here.

Nevertheless, in states in  $S'$  the construction remains almost the same. Here  $S'$  only consists of the absorbing state, which is a trivial state.  $\triangleleft$

## 4.4 The proof

In this section we provide a proof of theorem 53. which will go along similar lines as the proof of theorem 36 in section 3.4. Recall that we have fixed a non-improving strategy  $\pi$ . In the restricted game  $\Gamma'$ , let  $H'$  denote the set of finite histories,  $\Pi'$  and  $\Sigma'$  the sets of history dependent strategies,  $\gamma'$  the average reward,  $v'_\beta$  the  $\beta$ -discounted value for all  $\beta \in (0, 1)$ . Let  $v' := \lim_{\beta \uparrow 1} v'_\beta$  (as discussed in section 3.2, the limit must exist). Moreover, let

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

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

so  $\bar{\Pi}$  and  $\bar{\Sigma}$  are the set of strategies in the original game  $\Gamma$  with the property that, as long as the play is in the restricted game  $\Gamma'$ , they behave as strategies in  $\Pi'$  and  $\Sigma'$ . By using the definition of  $\tilde{X}$ , the following lemma follows analogously to the first part of lemma 41.

**Lemma 55** *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  $Y'_s$ ,  $s \in S'$ .

**Lemma 56** *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  $\bar{y}_t \in \text{Relint}(Y'_t)$ . In view of lemma 52-(b), state  $s$  belongs to an ergodic set  $E$  with regard to  $(\bar{x}, \bar{y})$ , hence by lemma 55, 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.  $\square$

The next lemma, which is similar to lemma 38, says that, as long as player 2 plays in the restricted game  $\Gamma'$ , so does the non-improving strategy  $\pi$ .

**Lemma 57** *Let  $s \in S'$  be an arbitrary initial state. Let*

$$U_s := \{(h, t) \in H_s \times S \mid \mathcal{P}_{s\pi\sigma}(h) > 0 \text{ and } \mathcal{P}_{s\pi\sigma}(t|h) > 0 \text{ for some } \sigma \in \bar{\Sigma}\},$$

where  $\mathcal{P}_{s\pi\sigma}(t|h)$  is the probability that, with respect to  $(\pi, \sigma)$ , 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  $\phi_z(h) \in Y_z$  as follows: if  $\pi_z(h) \in X'_z$  then let  $\phi_z(h) \in Y'_z$ ; while if  $\pi_z(h) \in X_z \setminus X'_z$  then let  $\phi_z(h) \in Y_z$  such that  $A_z(\pi_z(h), \phi_z(h)) \leq a_z$ . By lemma 56, we have in both cases that

$$A_z(\pi_z(h), \phi_z(h)) \leq a_z. \tag{4.1}$$

Let

$$\delta \in (0, \mathcal{P}_{s^1\pi\sigma}(\bar{h}^n) \cdot \mathcal{P}_{s^1\pi\sigma}(t|\bar{h}^n) \cdot \tau).$$

Let  $s^1 := s$ , and let  $s^m$ ,  $m \geq 2$ , denote the random variable for the state at stage  $m$ , and let  $\theta^m$  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  $\sigma$  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  $\delta$ -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  $\Gamma'$ . By lemma 52-(b), we must have  $s^n \in S'$ , and by the definitions of  $X'$  and  $Y'$ , we obtain

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

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

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

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

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

which contradicts the definition of  $a$ .  $\square$

Using the previous lemma, the next result follows similarly to lemma 39.

**Lemma 58** *We have*

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

**Proof of theorem 53:**

By using the above lemmas, the proof is almost the same as the proof of theorem 36. Note that lemma 52-(a) is needed for achieving the following crucial property: if a pure stationary strategy  $j \in J$  is a best reply to some  $x_\beta^\tau$ , then, in any ergodic set with respect to  $(x_\beta^\tau, j)$ , the play is in fact going on in the restricted game  $\Gamma'$ .  $\square$

## Chapter 5

# Markov strategies are better than stationary strategies

### 5.1 Introduction

As before, we evaluate a strategy of a player by the highest reward it guarantees against any strategy of the opponent. For the sake of simplicity, we only focus on strategies of player 1 here. So as in definition 22, for a strategy  $\pi \in \Pi$  let

$$\underline{v}_s(\pi) := \inf_{\sigma \in \Sigma} \gamma_s(\pi, \sigma) \quad \forall s \in S, \quad \underline{v}(\pi) := (\underline{v}_s(\pi))_{s \in S}.$$

For an initial state  $s \in S$ , we will call the highest reward that can be guaranteed by stationary strategies the stationary utility, denoted by  $\mathcal{A}_s$ , and the highest reward that can be guaranteed by Markov strategies the Markov utility, denoted by  $\mathcal{B}_s$ . Formally,

$$\mathcal{A}_s := \sup_{x \in X} \underline{v}_s(x), \quad \mathcal{B}_s := \sup_{f \in F} \underline{v}_s(f), \quad \mathcal{A} := (\mathcal{A}_s)_{s \in S}, \quad \mathcal{B} := (\mathcal{B}_s)_{s \in S}.$$

The fact that all stationary strategies are Markov strategies as well, and the definition of the value yield

$$\mathcal{A} \leq \mathcal{B} \leq v. \tag{5.1}$$

Although the class of Markov strategies is much richer than the class of stationary strategies, so far no substantial difference has been found in the use of stationary and Markov strategies in zero-sum stochastic games (with finite state and action spaces). Most classes of stochastic games, examined so far, have the property that both players have stationary  $\varepsilon$ -optimal strategies for all  $\varepsilon > 0$  (cf. theorem 30), thus, in view of (5.1), in those classes Markov strategies do not yield higher rewards than stationary strategies. The only class in which stationary  $\varepsilon$ -optimal strategies are not available is the class of repeated games with absorbing states. Later we provide a direct proof that the equality  $\mathcal{A} = \mathcal{B}$  holds for this class as well (the very same result also follows from a more general result in Coulomb [1992]). Thus, the goal of this chapter, which is based on Flesch et al. [1997,II], is to explore the way how Markov strategies can be

more effective than stationary strategies, as well as to find sufficient conditions under which these two classes of strategies are equally good.

The main results can be summarized as follows.

**Main Theorem 5**

(a) *In the following game we have  $\mathcal{A}_t < \mathcal{B}_t$  for initial states  $t = 1, 2$ :*

		$L_1$	$R_1$	
$T_1$	1		0	
$B_1$	1		1	
		2	*	
		1		

		$L_2$	$R_2$	
$T_2$	1		0	
$B_2$	0	1		
		*	*	
		2		

(b) *In every zero-sum stochastic game there exists an initial state  $s \in S$  for which  $\mathcal{A}_s = \mathcal{B}_s$ .*

(c) *We have  $\mathcal{A} = \mathcal{B}$  in*

- *repeated games with absorbing states;*
- *games where for all  $s, t \in S$  either  $\mathcal{A}_s = \mathcal{A}_t$  or  $\mathcal{B}_s = \mathcal{B}_t$  holds; particularly in games with constant  $\mathcal{A}$  or  $\mathcal{B}$ ;*
- *games in which player 1 has an optimal strategy;*
- *games in which player 1 has a best Markov strategy, namely a Markov strategy  $f \in F$  with  $\underline{v}(f) \geq \underline{v}(f')$  for all Markov strategies  $f' \in F$ .*

The above theorem will follow from theorems 60, 70, 71, 75, and 76.

## 5.2 An example where $\mathcal{A} < \mathcal{B}$ for some initial states

This section is devoted to the following example demonstrating that  $\mathcal{B}$  may be strictly larger than  $\mathcal{A}$  for some initial states. Consider the following game  $\Gamma$ .

**Example 59**

		$L_1$	$R_1$	
$T_1$	1		0	
$B_1$	1		1	
		2	*	
		1		

		$L_2$	$R_2$	
$T_2$	1		0	
$B_2$	0		1	
		*	*	
			2	

The main result of this section is the next theorem, which will follow from lemmas 61 and 69 below.

**Theorem 60** *In the game  $\Gamma$  we have  $0 = \mathcal{A}_t < \mathcal{B}_t = 1 = v_t$  for initial states  $t = 1, 2$ .*

This theorem states that, for initial states 1 and 2 in the game  $\Gamma$ , player 1 can get at most 0 by using stationary strategies, although using Markov strategies he can get as close to 1 as he likes.

Since there are two actions for each player in states 1 and 2, we may represent each mixed action in state 1 and in state 2 by the probability assigned to the first action, which makes the stationary and Markov strategy spaces

$$X = Y = [0, 1] \times [0, 1], \quad F = G = \times_{n \in \mathbb{N}} ([0, 1] \times [0, 1]).$$

First we intuitively discuss the main steps of the proof. We will start with an easy proof that  $\mathcal{A}_t = 0$  for initial states  $t = 1, 2$  (cf. lemma 61). Since the largest payoff in the game is 1, in view of (5.1), it remains to show that  $\mathcal{B}_t = 1$  for initial states  $t = 1, 2$ . However, for this step we need to analyze the game in detail. We define a Markov strategy  $f^K$  for player 1 where  $K \in \mathbb{N}$ : let

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

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. Note that the sequence  $u^K(n)$  converges to 1 as  $n$  tends to infinity, so  $f^K$  assigns less and less probabilities to actions  $B_1$  and  $B_2$ .

We will show that, for all  $\varepsilon > 0$ , for initial states 1 and 2, player 1 can guarantee a reward at least  $1 - \varepsilon$  by playing the Markov strategy  $f^K$  with a large  $K \in \mathbb{N}$ . Now the question 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, but apparently player 2 can only achieve absorption in entry  $(B_2, L_2)$  with probability at most  $\varepsilon$  (cf. lemma 66). 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 in fact, then whenever the play is in state 2 a transition occurs to state 1 with a large probability, and from state 1 it takes a long time, and for large stages even a longer

and longer time, until the play comes back to state 2 again, so using that the strategy  $f^K$  assigns less and less probabilities to  $B_2$ , the probability of absorption in entry  $(B_2, L_2)$  turns out to be at most  $\varepsilon$  indeed. On the other hand, using that 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  $R_2$  “often enough” and hope that the play will not absorb, but in that case it will appear that the play will eventually absorb with probability 1 (cf. lemma 67), and then the zero payoffs in entries  $(T_1, R_1)$  and  $(T_2, R_2)$  will have no influence on the reward. First we show that, by playing stationary strategies, player 1 can get at most 0 for initial states 1 and 2.

**Lemma 61**  $\mathcal{A}_t = 0$  for initial states  $t = 1, 2$  in the game  $\Gamma$ .

**Proof.** For each strategy  $x = (x_1, x_2)$  we define a strategy  $y^x = (y_1^x, y_2^x)$  for player 2. Let

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

Notice that, for  $t = 1, 2$ , we have  $\gamma_t(x, y^x) = 0$  for all  $x \in X$ , so

$$\mathcal{A}_t = \sup_{x \in X} \underline{v}_t(x) = \sup_{x \in X} \inf_{\sigma \in \Sigma} \gamma_t(x, \sigma) \leq \sup_{x \in X} \gamma_t(x, y^x) = 0 \quad \forall t = 1, 2.$$

Since the smallest payoff in the game is 0, the proof is complete.  $\square$

For the analysis of  $f^K$ , defined above, we need two important properties of the speed of convergence when  $u^K(n)$  tends to 1 as  $n$  goes to infinity. The first property says that the convergence is fast in the sense that, intuitively, for any  $\varepsilon > 0$ , if  $K \in \mathbb{N}$  is sufficiently large then the probability of ever playing action  $B_1$  or action  $B_2$  at stages  $2^{n-1}$ ,  $n \in \mathbb{N}$ , is at most  $\varepsilon/2$ . However, on the other hand the second property tells us that, in a “dense” set of stages, one of the bottom actions  $B_1$  and  $B_2$  will eventually be chosen, so the convergence of  $u^K(n)$  is not too fast.

**Lemma 62** The sequences  $(u^K(n))_{n \in \mathbb{N}}$ , where  $K \in \mathbb{N}$ , have the following properties:

(a) For any  $\varepsilon > 0$ , if  $K \in \mathbb{N}$  is sufficiently large then

$$\prod_{n=1}^{\infty} u^K(2^{n-1}) \geq 1 - \frac{\varepsilon}{2}.$$

(b) If  $A \subset \mathbb{N}$  satisfies

$$\omega(A) := \limsup_{N \rightarrow \infty} \frac{1}{N} \cdot |A \cap \{1, \dots, N\}| > 0$$

then for any  $K \in \mathbb{N}$

$$\prod_{n \in A} u^K(n) = 0.$$

**Proof.**

(a) Let  $\varepsilon > 0$ . For any sequence  $(w^n)_{n \in \mathbb{N}}$  in  $[0, 1]$  we have

$$\prod_{n=1}^{\infty} w^n = 1 - [(1 - w^1) + w^1(1 - w^2) + w^1 w^2(1 - w^3) + \dots],$$

thus

$$\begin{aligned} \prod_{n=1}^{\infty} u^K(2^{n-1}) &= \prod_{n=1}^{\infty} \sqrt[K]{\frac{2^{n-1}}{2^{n-1} + 1}} \\ &= \sqrt[K]{\prod_{n=1}^{\infty} \frac{2^{n-1}}{2^{n-1} + 1}} \\ &= \sqrt[K]{1 - \left( \frac{1}{2} + \frac{11}{23} + \frac{121}{235} + \frac{1241}{2359} + \dots \right)}. \end{aligned}$$

Let

$$d := 1 - \left( \frac{1}{2} + \frac{11}{23} + \frac{121}{235} + \frac{1241}{2359} + \dots \right).$$

Notice that

$$d > 1 - \left( \frac{1}{2} + \frac{11}{22} + \frac{11}{24} + \frac{11}{28} + \dots \right) = 0.$$

Since  $d$  is positive, there exists a  $\bar{K} \in \mathbb{N}$  such that for all  $K \geq \bar{K}$

$$\prod_{n=1}^{\infty} u^K(2^{n-1}) = \sqrt[K]{d} \geq 1 - \frac{\varepsilon}{2},$$

so the proof of the first part is complete.

(b) By the definition of  $\omega(A)$ , there exists an increasing sequence  $(n_k)_{k \in \mathbb{N}}$  in  $A$  such that

$$\frac{1}{n_k} \cdot |A \cap \{1, \dots, n_k\}| \geq \frac{1}{2} \omega(A) \quad \forall k \in \mathbb{N}. \quad (5.2)$$

As  $\omega(A) > 0$ , by taking a subsequence we may assume without loss of generality that

$$\frac{1}{8} n_{k+1} \cdot \omega(A) \geq n_k \quad \forall k \in \mathbb{N}. \quad (5.3)$$

Then (5.2) and (5.3) imply

$$\begin{aligned} |A \cap \{n_k + 1, \dots, n_{k+1}\}| &\geq |A \cap \{1, \dots, n_{k+1}\}| - n_k \\ &\geq \frac{1}{2} n_{k+1} \cdot \omega(A) - n_k \\ &\geq \frac{1}{4} n_{k+1} \cdot \omega(A). \end{aligned}$$

Since the left hand side is a natural number or zero we obtain

$$|A \cap \{n_k + 1, \dots, n_{k+1}\}| \geq \left\lceil \frac{1}{4} n_{k+1} \cdot \omega(A) \right\rceil,$$

where  $\lceil r \rceil$  denotes  $\min\{n \in \mathbb{N} | r \leq n\}$ . Using that  $u^K(n)$  is increasing in  $n$  and applying (5.3) yields

$$\begin{aligned} \prod_{n \in A \cap \{n_k + 1, \dots, n_{k+1}\}} u^K(n) &\leq \prod_{n = n_{k+1} - \lceil \frac{1}{4} n_{k+1} \cdot \omega(A) \rceil + 1}^{n_{k+1}} u^K(n) \\ &= \sqrt[K]{\frac{n_{k+1} - \lceil \frac{1}{4} n_{k+1} \cdot \omega(A) \rceil + 1}{n_{k+1} + 1}} \\ &\leq \sqrt[K]{\frac{n_{k+1} - \frac{1}{4} n_{k+1} \cdot \omega(A) + 1}{n_{k+1}}} \\ &= \sqrt[K]{1 - \frac{1}{4} \omega(A) + \frac{1}{n_{k+1}}} \\ &\leq \sqrt[K]{1 - \frac{1}{8} \omega(A)}. \end{aligned}$$

Therefore

$$\begin{aligned} \prod_{n \in A} u^K(n) &= \prod_{n \in A \cap \{1, \dots, n_1\}} u^K(n) \cdot \prod_{k \in \mathbb{N}} \left[ \prod_{n \in (A \cap \{n_k + 1, \dots, n_{k+1}\})} u^K(n) \right] \\ &\leq \prod_{k \in \mathbb{N}} \sqrt[K]{1 - \frac{1}{8} \omega(A)} \\ &= 0, \end{aligned}$$

so the proof is complete.  $\square$

The next lemma says that, for initial states 1 and 2, if player 2 chooses actions  $L_1$  and  $L_2$  whenever the play is in state 1 or in state 2, then the strategy  $f^K$ , with a large  $K \in \mathbb{N}$ , guarantees that the frequency of visits to state 2 rapidly decreases during the play. At the first sight the reason might seem to be absorption in entry  $(B_2, L_2)$ , but as it will turn out in lemma 66, absorption in entry  $(B_2, L_2)$  does not play an important role here. The very reason is in fact that the lengths of periods when staying in state 1 increase during the play, which is due to the gradually decreasing probabilities for playing  $B_1$  in state 1.

**Lemma 63** *Let  $\varepsilon > 0$ ,  $t \in \{1, 2\}$ , and let  $y = (1, 1) \in Y$ . For a history  $h^\infty \in H_t^\infty$ , let  $m(h^\infty)$  be the number of stages at which the play is in state 2 during  $h^\infty$ . Let  $M(h^\infty) := \{n \in \mathbb{N} | n \leq m(h^\infty)\}$ . Let  $(a^n(h^\infty))_{n \in M(h^\infty)}$  denote the sequence of stages at which state 2 is visited during  $h^\infty$ . Then for large  $K \in \mathbb{N}$*

$$\mathcal{P}_{tfK_y}(a^n(\theta^\infty) \geq 2^{n-1} \quad \forall n \in M(\theta^\infty)) \geq 1 - \frac{\varepsilon}{2},$$

where  $\theta^\infty$  denotes the random variable for the infinite history.

**Proof.** We only show the statement for initial state 2; for initial state 1 a similar proof can be given. So suppose that the initial state is state 2. Then notice that  $a^1(h^\infty) = 1$ ,  $m(h^\infty) \geq 1$ , and  $M(h^\infty) \neq \emptyset$  for all  $h^\infty \in H_2^\infty$  (for initial state 1 we would have that if  $M(h^\infty) \neq \emptyset$  then  $a^1(h^\infty) \geq 2$ , which would only slightly modify the rest of the proof). For all  $h^\infty \in H_2^\infty$ , whenever  $m(h^\infty) < \infty$ , we define inductively

$$a^n(h^\infty) := \max \{2^{n-1}, 8a^{n-1}(h^\infty)\} \quad \forall n = m(h^\infty) + 1, m(h^\infty) + 2, \dots \quad (5.4)$$

In view of (5.4), we have to show that for large  $K \in \mathbb{N}$

$$\mathcal{P}_{2fKy} (a^n(\theta^\infty) \geq 2^{n-1} \quad \forall n \in \mathbb{N}) \geq 1 - \frac{\varepsilon}{2}. \quad (5.5)$$

Observe that if the play is in state 2 at stage  $w$ , then the probability, with respect to  $(f^K, y)$ , that the play does not return to state 2 before stage  $8w$  is at least the probability that the play moves to state 1 and it stays there till stage  $8w - 1$ ; so at least

$$u^K(w) \cdot \prod_{n=w+1}^{8w-2} u^K(n) = \prod_{n=w}^{8w-2} u^K(n).$$

Hence, for any  $w, k \in \mathbb{N}$ , if  $\mathcal{P}_{2fKy}(a^k(\theta^\infty) = w, k \in M(\theta^\infty)) > 0$  then

$$\mathcal{P}_{2fKy} (a^{k+1}(\theta^\infty) \geq 8a^k(\theta^\infty) | a^k(\theta^\infty) = w, k \in M(\theta^\infty)) \geq \prod_{n=w}^{8w-2} u^K(n). \quad (5.6)$$

On the other hand, if  $\mathcal{P}_{2fKy}(a^k(\theta^\infty) = w, k \notin M(\theta^\infty)) > 0$  then by (5.4) we have

$$\mathcal{P}_{2fKy} (a^{k+1}(\theta^\infty) \geq 8a^k(\theta^\infty) | a^k(\theta^\infty) = w, k \notin M(\theta^\infty)) = 1. \quad (5.7)$$

Therefore, for all  $w \in \mathbb{N}$  and for all  $k \in \mathbb{N}$  satisfying  $\mathcal{P}_{2fKy}(a^k(\theta^\infty) = w) > 0$ , by (5.6) and (5.7) we have

$$\begin{aligned} \mathcal{P}_{2fKy} (a^{k+1}(\theta^\infty) \geq 8a^k(\theta^\infty) | a^k(\theta^\infty) = w) &\geq \prod_{n=w}^{8w-2} u^K(n) & (5.8) \\ &= \sqrt[K]{\frac{w}{(8w-2)+1}} \\ &= \sqrt[K]{\frac{w}{8w-1}} \\ &\geq \sqrt[K]{\frac{1}{8}}. \end{aligned}$$

For  $h^\infty \in H_2^\infty$ , let  $\eta^0(h^\infty) := 0$  and for  $n \in \mathbb{N}$  let

$$\eta^n(h^\infty) := \begin{cases} \eta^{n-1}(h^\infty) + 1 & \text{if } a^{n+1}(h^\infty) \geq 8a^n(h^\infty) \\ \eta^{n-1}(h^\infty) - 1 & \text{otherwise.} \end{cases}$$

We now show that for large  $K \in \mathbb{N}$

$$\mathcal{P}_{2fKy} (\eta^n(\theta^\infty) \geq 1 \quad \forall n \in \mathbb{N}) \geq 1 - \frac{\varepsilon}{2}. \quad (5.9)$$

For simplicity, let  $\xi^K := \sqrt[K]{\frac{1}{8}}$ . On the set of integers, for any  $K \in \mathbb{N}$ , we define a birth and death process  $\bar{\eta}_K^n$ ,  $n = 0, 1, 2, \dots$ , as follows. Let  $\bar{\eta}_K^0 := 0$  and for  $n \in \mathbb{N}$  let

$$\bar{\eta}_K^n := \begin{cases} \bar{\eta}_K^{n-1} + 1 & \text{with probability } \xi^K \\ \bar{\eta}_K^{n-1} - 1 & \text{with probability } 1 - \xi^K. \end{cases}$$

Since  $\xi^K$  converges to 1 as  $K$  tends to infinity, for the birth and death process  $\bar{\eta}_K^n$ ,  $n = 0, 1, 2, \dots$ , we clearly have that for large  $K \in \mathbb{N}$

$$\mathcal{P}(\bar{\eta}_K^n \geq 1 \quad \forall n \in \mathbb{N}) \geq 1 - \frac{\varepsilon}{2}.$$

Hence by the definitions of  $\eta^n$  and  $\bar{\eta}_K^n$  for  $n = 0, 1, 2, \dots$ , and by (5.8) we have for large  $K \in \mathbb{N}$  that

$$\mathcal{P}_{2fKy}(\eta^n(\theta^\infty) \geq 1 \quad \forall n \in \mathbb{N}) \geq \mathcal{P}(\bar{\eta}_K^n \geq 1 \quad \forall n \in \mathbb{N}) \geq 1 - \frac{\varepsilon}{2},$$

which completes the proof of (5.9).

For  $h^\infty \in H_2^\infty$ , let  $\nu^0(h^\infty) := 0$  and let  $\nu^n(h^\infty)$  denote the number of jumps with +1 in the sequence  $\eta^0(h^\infty), \eta^1(h^\infty), \dots, \eta^n(h^\infty)$ . Since for all  $n \in \mathbb{N}$

$$\eta^n(h^\infty) = (+1) \cdot \nu^n(h^\infty) + (-1) \cdot (n - \nu^n(h^\infty)) = 2\nu^n(h^\infty) - n,$$

for large  $K \in \mathbb{N}$ , inequality (5.9) implies

$$\mathcal{P}_{2fKy} \left( \nu^n(\theta^\infty) \geq \frac{n+1}{2} \quad \forall n \in \mathbb{N} \right) \geq 1 - \frac{\varepsilon}{2}. \quad (5.10)$$

Recall that  $a^1(h^\infty) = 1$  for all  $h^\infty \in H_2^\infty$  and notice that if  $\nu^n(h^\infty) \geq \frac{n+1}{2}$  for some  $n \in \mathbb{N}$ ,  $h^\infty \in H_2^\infty$ , then

$$a^n(h^\infty) \geq 8^{\nu^n(h^\infty)-1} \geq 8^{\frac{n-1}{2}} = 2^{\frac{3}{2}(n-1)} \geq 2^{n-1},$$

hence (5.10) implies (5.5), which completes the proof.  $\square$

In the remainder of this chapter, if  $h^\infty$  is an infinite history, then  $h^n$  denotes the history up to stage  $n$  which coincides with  $h^\infty$  up to stage  $n$ .

For  $t \in S$ ,  $\pi \in \Pi$ ,  $\sigma \in \Sigma$ , let

$$H_t^\infty(\pi, \sigma) := \{h^\infty \in H_t^\infty \mid \mathcal{P}_{t\pi\sigma}(h^n) > 0 \text{ for all } n \in \mathbb{N}\}.$$

Clearly,

$$\mathcal{P}_{t\pi\sigma}(\theta^\infty \in H_t^\infty(\pi, \sigma)) = 1,$$

where  $\theta^\infty$  denotes the random variable for the infinite history.

Furthermore, if  $U_t^\infty \subset H_t^\infty$  for some  $t \in S$  then let

$$U_t^n := \{h^n \in H_t^n \mid h^n = \bar{h}^n \text{ for some } \bar{h}^\infty \in U_t^\infty\}.$$

The next lemma, which is not specific for this game  $\Gamma$  at all, provides useful lower and upper-bounds for the probability that the infinite history belongs to a set  $V_t^\infty$  of infinite histories, on condition that it belongs to some other set  $U_t^\infty$ . The proof involves some technical difficulties, therefore it can be found in the appendix of this chapter.

**Lemma 64** *Let  $t \in S$ ,  $\pi \in \Pi$ ,  $\sigma \in \Sigma$ . Let  $V_t^\infty, U_t^\infty \subset H_t^\infty(\pi, \sigma)$  such that  $\emptyset \neq V_t^\infty \subset U_t^\infty$ . Let  $\theta^\infty$  denote the random variable for the infinite history, and for all  $h^\infty \in V_t^\infty$  let*

$$Z_{t\pi\sigma, V_t^\infty | U_t^\infty}(h^\infty) := \prod_{k=0}^{\infty} \mathcal{P}_{t\pi\sigma} \left( \theta^{k+1} \in V_t^{k+1} \mid \theta^k = h^k, \theta^\infty \in U_t^\infty \right).$$

Then, if

$$\mathcal{P}_{t\pi\sigma}(\theta^\infty \in U_t^\infty) > 0,$$

then

$$\inf_{h \in V_t^\infty} Z_{t\pi\sigma, V_t^\infty | U_t^\infty}(h^\infty) \leq \mathcal{P}_{t\pi\sigma}(\theta^\infty \in V_t^\infty \mid \theta^\infty \in U_t^\infty) \leq \sup_{h \in V_t^\infty} Z_{t\pi\sigma, V_t^\infty | U_t^\infty}(h^\infty).$$

Note that, in the game  $\Gamma$ , the history up to any stage  $n \in \mathbb{N}$  already determines the state at stage  $n+1$ , because for each action pair in  $\Gamma$ , the transition occurs to a certain state with probability 1. Therefore, in the game  $\Gamma$ , the mixed actions prescribed by the strategies at stage  $n+1$  are already determined by the history up to stage  $n$ . The following result, which will follow from the previous lemma, intuitively states that the set of infinite histories in which absorption should occur with probability 1 but no absorption occurs has probability zero.

**Lemma 65** *Let  $t \in \{1, 2\}$ ,  $K \in \mathbb{N}$ ,  $\sigma \in \Sigma^p$ . Let*

$$\tilde{H}_t^\infty := \{h^\infty \in H_t^\infty(f^K, \sigma) \mid \text{no absorption occurs in } h^\infty\},$$

$$\bar{H}_t^\infty := \{h^\infty \in \tilde{H}_t^\infty \mid \prod_{n \in C(h^\infty)} u^K(n) = 0\},$$

where  $C(h^\infty)$  is the set of stages  $n$  when, according to the pure strategy  $\sigma$ , player 2 plays actions  $R_1$ ,  $R_2$ , or  $L_2$  after history  $h^{n-1}$ . Let  $\theta^\infty$  denote the random variable for the infinite history. Then

$$\mathcal{P}_{tf^K\sigma}(\theta^\infty \in \bar{H}_t^\infty) = 0.$$

**Proof.** Let  $Z_{tf^K\sigma, \bar{H}_t^\infty | H_t^\infty(f^K, \sigma)}$  be defined as in lemma 64. By the definition of  $\bar{H}_t^\infty$  we have for all  $h^\infty \in \bar{H}_t^\infty$

$$\begin{aligned} Z_{tf^K\sigma, \bar{H}_t^\infty | H_t^\infty(f^K, \sigma)}(h^\infty) &= \prod_{k=0}^{\infty} \mathcal{P}_{tf^K\sigma} \left( \theta^{k+1} \in \bar{H}_t^{k+1} \mid \theta^k = h^k \right) \\ &\leq \prod_{k=0}^{\infty} \mathcal{P}_{tf^K\sigma} \left( \theta^{k+1} \in \tilde{H}_t^{k+1} \mid \theta^k = h^k \right) \\ &= \prod_{n \in C(h^\infty)} u^K(n) \\ &= 0, \end{aligned}$$

hence lemma 64 yields

$$\begin{aligned} \mathcal{P}_{tf^K\sigma}(\theta^\infty \in \bar{H}_t^\infty) &= \mathcal{P}_{tf^K\sigma}(\theta^\infty \in \bar{H}_t^\infty | \theta^\infty \in H_t^\infty(f^K, \sigma)) \\ &\leq \sup_{h \in \bar{H}_t^\infty} Z_{tf^K\sigma, \bar{H}_t^\infty | H_t^\infty(f^K, \sigma)}(h^\infty) \\ &\leq 0, \end{aligned}$$

which completes the proof.  $\square$

It turns out that the strategy  $f^K$ , with a large  $K \in \mathbb{N}$ , keeps the probability of absorption in entry  $(B_2, L_2)$  small. In fact, the absorption probability in entry  $(B_2, L_2)$  is maximal when player 2 always chooses actions  $L_1$  and  $L_2$  whenever the play is in state 1 or in state 2, but even then, in view of lemma 63, the play does not visit state 2 “frequently enough”, so using that  $f^K$  assigns less and less probabilities to action  $B_2$ , the probability of absorption in entry  $(B_2, L_2)$  turns out to be small indeed.

**Lemma 66** *Let  $\varepsilon > 0$ . If  $K \in \mathbb{N}$  is sufficiently large then, for initial states  $t = 1, 2$  in the game  $\Gamma$ , the probability of absorption in entry  $(B_2, L_2)$  is at most  $\varepsilon$  with respect to  $(f^K, \sigma)$ , for any  $\sigma \in \Sigma$ .*

**Proof.** It is easy to see that the stationary strategy  $y$  which chooses actions  $L_1$  and  $L_2$  with probability 1 maximizes the probability of absorption in entry  $(B_2, L_2)$  against  $f^K$  with any  $K \in \mathbb{N}$ . Therefore it is sufficient to show the statement for  $y$ .

We only show the statement for initial state 2. Then for initial state 1 the statement becomes immediate, since from the stage on when the play moves to state 2, the strategy  $f^K$  assigns even less probabilities to actions  $B_1$  and  $B_2$  than when starting from stage 1. So assume that the initial state is state 2.

Let

$$\begin{aligned} \tilde{H}_2^\infty &: = \{h^\infty \in H_2^\infty(f^K, y) \mid \text{no absorption occurs in } h^\infty\} \\ \hat{H}_2^\infty &: = \{h^\infty \in H_2^\infty(f^K, y) \mid a^n(h^\infty) \geq 2^{n-1} \quad \forall n \in M(h^\infty)\}, \end{aligned}$$

where  $a^n(h^\infty)$  and  $M(h^\infty)$  are defined as in lemma 63. Observe that, for large  $K \in \mathbb{N}$ , by lemma 63 we have

$$\mathcal{P}_{2f^K y}(\theta^\infty \in \hat{H}_2^\infty) \geq 1 - \frac{\varepsilon}{2}. \quad (5.11)$$

Now for  $h^\infty \in \tilde{H}_2^\infty \cap \hat{H}_2^\infty$  let  $Z_{2f^K y, \tilde{H}_2^\infty \cap \hat{H}_2^\infty | \hat{H}_2^\infty}(h^\infty)$  be defined as in lemma 64. Then, using lemma 62-(a), if  $K \in \mathbb{N}$  is sufficiently large then for any  $h^\infty \in \tilde{H}_2^\infty \cap \hat{H}_2^\infty$

$$\begin{aligned}
Z_{2f^K y, \tilde{H}_2^\infty \cap \hat{H}_2^\infty | \hat{H}_2^\infty}(h^\infty) &= \prod_{k=0}^{\infty} \mathcal{P}_{2f^K y} \left( \theta^{k+1} \in \tilde{H}_2^{k+1} \cap \hat{H}_2^{k+1} \mid \theta^k = h^k, \theta^\infty \in \hat{H}_2^\infty \right) \\
&= \prod_{k=0}^{\infty} \mathcal{P}_{2f^K y} \left( \theta^{k+1} \in \tilde{H}_2^{k+1} \mid \theta^k = h^k, \theta^\infty \in \hat{H}_2^\infty \right) \\
&= \prod_{n \in M(h^\infty)} u^K(a^n(h^\infty)) \\
&\geq \prod_{n \in M(h^\infty)} u^K(2^{n-1}) \\
&\geq \prod_{n=1}^{\infty} u^K(2^{n-1}) \\
&\geq 1 - \frac{\varepsilon}{2}.
\end{aligned}$$

Hence, by applying (5.11) and lemma 64, for large  $K \in \mathbb{N}$  we get

$$\begin{aligned}
\mathcal{P}_{2f^K y} \left( \theta^\infty \in \tilde{H}_2^\infty \right) &\geq \mathcal{P}_{2f^K y} \left( \theta^\infty \in \tilde{H}_2^\infty \cap \hat{H}_2^\infty \right) \\
&= \mathcal{P}_{2f^K y} \left( \theta^\infty \in \tilde{H}_2^\infty \cap \hat{H}_2^\infty \mid \theta^\infty \in \hat{H}_2^\infty \right) \cdot \mathcal{P}_{2f^K y} \left( \theta^\infty \in \hat{H}_2^\infty \right) \\
&\geq \mathcal{P}_{2f^K y} \left( \theta^\infty \in \tilde{H}_2 \cap \hat{H}_2^\infty \mid \theta^\infty \in \hat{H}_2^\infty \right) \cdot \left( 1 - \frac{\varepsilon}{2} \right) \\
&\geq \inf_{h^\infty \in \tilde{H}_2^\infty \cap \hat{H}_2^\infty} Z_{2f^K y, \tilde{H}_2^\infty \cap \hat{H}_2^\infty | \hat{H}_2^\infty}(h^\infty) \cdot \left( 1 - \frac{\varepsilon}{2} \right) \\
&\geq \left( 1 - \frac{\varepsilon}{2} \right) \cdot \left( 1 - \frac{\varepsilon}{2} \right) \\
&\geq 1 - \varepsilon,
\end{aligned}$$

which means that if  $K \in \mathbb{N}$  is large then, with respect to  $(f^K, y)$  and initial state 2, the probability of absorption in entry  $(B_2, L_2)$  is at most  $\varepsilon$ .  $\square$

Now we show that, when player 1 uses  $f^K$  with any  $K \in \mathbb{N}$  for initial states 1 or 2, if player 2 chooses actions  $R_1$  and  $R_2$  “too frequently” then absorption occurs with probability 1. (Recall again that, in the game  $\Gamma$ , the history up to any stage  $n \in \mathbb{N}$  already determines the state for stage  $n + 1$ .)

**Lemma 67** *Let  $t \in \{1, 2\}$ ,  $K \in \mathbb{N}$ ,  $\sigma \in \Sigma^p$ . Let*

$$\tilde{H}_t^\infty := \{h^\infty \in H_t^\infty(f^K, \sigma) \mid \text{no absorption occurs in } h^\infty\}.$$

For  $A \subset \mathbb{N}$  let

$$\omega(A) := \limsup_{N \rightarrow \infty} \frac{1}{N} \cdot |A \cap \{1, \dots, N\}|.$$

For a history  $h^\infty \in H_t^\infty$ , let  $A(h^\infty)$  denote the set of stages  $n$  when, according to the pure strategy  $\sigma$ , player 2 chooses actions  $R_1$  or  $R_2$  after history  $h^{n-1}$ . Then, if

$$\mathcal{P}_{tfK\sigma}(\theta^\infty \in \tilde{H}_t^\infty) > 0$$

then

$$\mathcal{P}_{tfK\sigma}(\omega(A(\theta^\infty)) = 0 | \theta^\infty \in \tilde{H}_t^\infty) = 1,$$

where  $\theta^\infty$  denotes the random variable for the infinite history.

**Proof.** Suppose that  $\omega(A(h^\infty)) > 0$  for some history  $h^\infty \in H_t^\infty$ . Then, clearly, no absorption occurs in  $h^\infty$ , thus  $h^\infty \in \tilde{H}_t^\infty$ . Let

$$\bar{H}_t^\infty := \{h^\infty \in \tilde{H}_t^\infty | \prod_{n \in A(h^\infty) \cup B(h^\infty)} u^K(n) = 0\},$$

where  $B(h^\infty)$  is the set of stages  $n$  when, according to the pure strategy  $\sigma$ , player 2 plays action  $L_2$  after history  $h^{n-1}$ . By lemma 62-(b) we have

$$\prod_{n \in A(h^\infty)} u^K(n) = 0,$$

therefore

$$\{h^\infty \in H_t^\infty(f^K, \sigma) | \omega(A(h^\infty)) > 0\} \subset \bar{H}_t^\infty.$$

Now lemma 65 yields

$$\mathcal{P}_{tfK\sigma}(\omega(A(\theta^\infty)) > 0) \leq \mathcal{P}_{tfK\sigma}(\theta^\infty \in \bar{H}_t^\infty) = 0,$$

which implies the statement.  $\square$

The next result tells us that, when player 1 uses  $f^K$  with any  $K \in \mathbb{N}$  for initial states 1 and 2, then, given that no absorption occurs (and this has a positive probability), the average of the payoffs along the infinite history equals 1 almost surely.

**Lemma 68** *Let  $t \in \{1, 2\}$ ,  $K \in \mathbb{N}$ ,  $\sigma \in \Sigma^p$ . Let*

$$\tilde{H}_t^\infty := \{h^\infty \in H_t^\infty(f^K, \sigma) | \text{no absorption occurs in } h^\infty\}.$$

Then, if

$$\mathcal{P}_{tfK\sigma}(\theta^\infty \in \tilde{H}_t^\infty) > 0$$

then

$$\mathcal{P}_{tfK\sigma}\left(\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n = 1 | \theta^\infty \in \tilde{H}_t^\infty\right) = 1,$$

where  $\theta^\infty$  denotes the random variable for the infinite history and  $R_n$  the random variable for the payoff at stage  $n$ .

**Proof.** Let  $\omega(A)$  for  $A \subset \mathbb{N}$  and  $A(h^\infty)$  be defined as in lemma 67. Let  $R_n(h^\infty)$  be the payoff at stage  $n$  according to the history  $h^\infty$ . Clearly, we have  $R_n = R_n(\theta^\infty)$ . Now for any  $h^\infty \in \tilde{H}_t^\infty$

$$\begin{aligned}
\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n(h^\infty) &= \lim_{m \rightarrow \infty} \inf_{N \geq m} \frac{\sum_{n=1}^N R_n(h^\infty)}{N} \\
&= \lim_{m \rightarrow \infty} \inf_{N \geq m} \frac{|\{n \in \{1, \dots, N\} | R_n(h^\infty) = 1\}|}{N} \\
&= \lim_{m \rightarrow \infty} \inf_{N \geq m} \frac{N - |\{n \in \{1, \dots, N\} | R_n(h^\infty) = 0\}|}{N} \\
&= \lim_{m \rightarrow \infty} \inf_{N \geq m} \frac{N - |A(h^\infty) \cap \{1, \dots, N\}|}{N} \\
&= 1 + \lim_{m \rightarrow \infty} \inf_{N \geq m} \frac{-|A(h^\infty) \cap \{1, \dots, N\}|}{N} \\
&= 1 - \lim_{m \rightarrow \infty} \sup_{N \geq m} \frac{|A(h^\infty) \cap \{1, \dots, N\}|}{N} \\
&= 1 - \omega(A(h^\infty)),
\end{aligned}$$

hence lemma 67 implies the result.  $\square$

Now we are ready to prove that  $\mathcal{B}_t = 1$  for initial states  $t = 1, 2$  and also that the Markov strategy  $f^K$  is  $\varepsilon$ -optimal for large  $K \in \mathbb{N}$ . More specifically,  $K$  can be any number that satisfies lemma 66.

**Lemma 69** *For all  $t = 1, 2$ , in the game  $\Gamma$  we have that  $\mathcal{B}_t = v_t = 1$ , and also that, for any  $\varepsilon > 0$ , if  $K \in \mathbb{N}$  is sufficiently large then  $\underline{v}_t(f^K) \geq 1 - \varepsilon$ .*

**Proof.** Let  $t \in \{1, 2\}$  and let  $\varepsilon > 0$ . We only need to show that  $\underline{v}_t(f^K) \geq 1 - \varepsilon$  for large  $K \in \mathbb{N}$ , because then  $\mathcal{B}_t = v_t = 1$  follows from (5.1) and from the fact that the largest payoff in the game is 1. Let  $\theta^\infty$  denote the random variable for the infinite history and  $R_n$  the random variable for the payoff at stage  $n$ . By lemma 68 we have with respect to  $(f^K, \sigma)$ , for any  $K \in \mathbb{N}$  and for any  $\sigma \in \Sigma^p$ , and initial state  $t$  with probability 1 that

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n = \begin{cases} 0 & \text{if absorption occurs in entry } (B_2, L_2) \text{ in } \theta^\infty \\ 1 & \text{otherwise.} \end{cases}$$

Take  $K \in \mathbb{N}$  as in lemma 66. Then the probability of absorption in entry  $(B_2, L_2)$  is at most  $\varepsilon$  with respect to  $(f^K, \sigma)$  and initial state  $t$ , for any  $\sigma \in \Sigma^p$ , hence

$$\mathcal{E}_{t, f^K, \sigma} \left( \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n \right) \geq 1 - \varepsilon \quad \forall \sigma \in \Sigma^p. \quad (5.12)$$

By applying Fatou's lemma (cf. Fatou [1906]) we obtain for all  $\sigma \in \Sigma^P$  that

$$\begin{aligned} \gamma_t(f^K, \sigma) &= \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \mathcal{E}_{tf^K\sigma}(R_n) \\ &\geq \mathcal{E}_{tf^K\sigma} \left( \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n \right) \\ &\geq 1 - \varepsilon. \end{aligned}$$

In view of theorem 16-(a), it suffices to consider pure replies from player 2, thus

$$v_t(f^K) \geq 1 - \varepsilon,$$

so the proof is complete.  $\square$

### 5.3 Sufficient conditions for $\mathcal{A} = \mathcal{B}$

Example 59 in the previous section demonstrated that  $\mathcal{B}$  may be strictly larger than  $\mathcal{A}$  for some initial states. However, this cannot hold for all initial states, as stated in the next theorem.

**Theorem 70** *In every zero-sum stochastic game  $\mathcal{A}_s = \mathcal{B}_s (= v_s)$  for all states  $s \in S^{\min} := \{t \in S \mid v_t = \min_{w \in S} v_w\}$ .*

**Proof.** Let  $s \in S^{\min}$ . Then by the results of Thuijsman & Vrieze [1993], for any  $\varepsilon > 0$ , player 1 has a stationary  $\varepsilon$ -optimal strategy  $x^\varepsilon$  for initial state  $s$ . Hence  $\mathcal{A}_s = v_s$ , thus (5.1) yields  $\mathcal{A}_s = \mathcal{B}_s (= v_s)$ .  $\square$

We know by now that  $\mathcal{A}$  may be strictly smaller than  $\mathcal{B}$  and also that  $\mathcal{A}$  equals  $\mathcal{B}$  for at least one initial state in every zero-sum game. Now the question is what conditions would guarantee that  $\mathcal{A}$  equals  $\mathcal{B}$  for all initial states. We will present several sufficient conditions, however, first we would like to recall some of the most important classes of zero-sum games in which  $\mathcal{A} = \mathcal{B}$  is already known.

Clearly, we have  $\mathcal{A} = \mathcal{B}(= v)$  for any class of games where player 1 has stationary  $\varepsilon$ -optimal strategies, for all  $\varepsilon > 0$ . The existence of stationary ( $\varepsilon$ -)optimal strategies is known in several special classes of stochastic games (cf. theorem 30). Moreover, the condition that the value is constant ( $v_s = v_t$  for all  $s, t \in S$ ) is also sufficient for the existence of stationary  $\varepsilon$ -optimal strategies (cf. Thuijsman & Vrieze [1993], theorem 3.1). (In such games actually both players have Markov optimal strategies as well.)

#### 5.3.1 Repeated games with absorbing states

Repeated games with absorbing states are stochastic games where there is only one non-absorbing state. Kohlberg [1974] showed that these games have a value (cf. theorem 30-(g)). However, to achieve this value history dependent strategies are

indispensable. We will show for these games that  $\mathcal{A} = \mathcal{B}$ . For the specific case of the Big Match we have already shown this equality (cf. lemma 24-(e)). In fact, we generalize this proof to all repeated games with absorbing states. We will not discuss all the technical details in the proof, we only give a brief sketch. We wish to mention that the same result also follows from Coulomb [1992], who used quite similar techniques in his proof as well.

**Theorem 71** *In every zero-sum repeated game with absorbing states  $\mathcal{A} = \mathcal{B}$ .*

**Proof.** Take a zero-sum repeated game with absorbing states. We may suppose without loss of generality that, in each absorbing state, both players have only one action (otherwise we may replace the state by another absorbing state containing only the value of the original state as payoff). Suppose that the initial state is state 1, the only non-absorbing state. Any action of player 1 or player 2 in state 1 will also denote the stationary strategy which prescribes this action for each stage.

In view of theorem 16-(b), against any stationary strategy  $x \in X$  there exists a best reply  $j^x \in J$ , hence we have  $\gamma(x, j^x) \leq \mathcal{A}$ . This means that, for initial state 1, either  $(x, j^x)$  is absorbing (namely it eventually leads to absorption with probability 1, or equivalently,  $p_1(1|x, j^x) < 1$ ) and then the expected absorption payoff is at most  $\mathcal{A}_1$ , or  $(x, j^x)$  is non-absorbing and then the expected one-shot payoff  $r_1(x, j^x)$  is at most  $\mathcal{A}_1$ .

Take an arbitrary Markov strategy  $f = (x^n)_{n \in \mathbb{N}} \in F$ . Let  $\varepsilon > 0$ . It suffices to show that there exists a Markov strategy  $g \in G$  such that  $\gamma_1(f, g) \leq \mathcal{A}_1 + \varepsilon$ .

*Step 1.* Let  $f_1 := f$  and  $x_1^n := x^n$  for all  $n \in \mathbb{N}$ . Let  $g_1 = (j^{x_1^n})_{n \in \mathbb{N}}$ . Let  $\xi$  denote the random variable for the stage when absorption occurs, if no absorption occurs at all then let  $\xi = 0$ . For  $N \in \mathbb{N} \cup \{0\}$  let

$$p_1^N := \mathcal{P}_{f_1 g_1}(\xi > N),$$

so  $p_1^N$  is the probability of absorption after stage  $N$  with respect to  $(f_1, g_1)$ . Let  $p_1 := p_1^0$ . Clearly,  $p_1$  is the probability of absorption with respect to  $(f_1, g_1)$ . We have

$$\begin{aligned} p_1 &= \mathcal{P}_{f_1 g_1}(1 \leq \xi) \\ &= \lim_{N \rightarrow \infty} \mathcal{P}_{f_1 g_1}(1 \leq \xi \leq N) \\ &= \lim_{N \rightarrow \infty} [\mathcal{P}_{f_1 g_1}(1 \leq \xi) - \mathcal{P}_{f_1 g_1}(\xi > N)] \\ &= p_1 - \lim_{N \rightarrow \infty} p_1^N, \end{aligned}$$

hence  $p_1^N$  converges to 0. Therefore, for any small  $\delta \in (0, 1)$ , there exists a stage  $N_1$  such that  $p_1^{N_1} \leq p^* \cdot \delta$ , where  $p^*$  is the smallest positive absorption probability in state 1:

$$p^* := \min \{p_{ij}^* \mid i \in I, j \in J, p_{ij}^* := 1 - p_1(1|i, j) \text{ and } p_{ij}^* > 0\};$$

(we may assume that there exist an  $i \in I$  and  $j \in J$  such that  $p_1(1|i, j) < 1$ , otherwise the game is trivial).

If  $p_1^{N_1} = 0$  then we have  $\gamma_1(f, g_1) = \gamma_1(f_1, g_1) \leq \mathcal{A}_1$ , because, with respect to  $(f_1, g_1)$ , the expected absorption payoff is at most  $\mathcal{A}_1$  at each stage  $n \leq N_1$ ; the probability

of absorption after stage  $N_1$  is zero; and the expected payoff in state 1 is at most  $\mathcal{A}_1$  at each stage  $n > N_1$ .

Assume now that  $p_1^{N_1} > 0$ . By the definition of  $N_1$ , the probability of absorption after stage  $N_1$  for  $(f_1, g_1)$  is at most  $p^* \cdot \delta$ . Now let  $I_1^n := \{i \in I \mid (i, j^{x_1^n}) \text{ is non-absorbing}\}$ . Thus the probability that, with respect to  $(f_1, g_1)$ , player 1 will ever choose an action outside  $I_1^n$  at stages  $n > N_1$  is at most  $\delta$ .

*Step 2.* Let  $x_2^n := x_1^n$  for  $n \leq N_1$  and let  $x_2^n$  be the normalization of  $x_1^n$  on  $I_1^n$  for  $n > N_1$ :

$$x_2^n(i) := \frac{x_1^n(i)}{\sum_{i \in I_1^n} x_1^n(i)} \quad \text{for all } i \in I_1^n, \quad x_2^n(i) := 0 \quad \text{for all } i \in I \setminus I_1^n.$$

Let  $f_2 := (x_2^n)_{n \in \mathbb{N}}$ . Intuitively,  $f_2$  coincides with  $f_1$  up to stage  $N_1$ , and, after stage  $N_1$ , the strategy  $f_2$  equals the strategy  $f_1$  on condition that no action outside  $I_1^n$  will ever be chosen at stages  $n > N_1$ . Let  $g_2 := (j^{x_2^n})_{n \in \mathbb{N}}$ , so by the definitions,  $g_1$  and  $g_2$  are the same for the first  $N_1$  stages. One can show, using the properties of the construction, that, with respect to  $(f, g_2)$ , the probability of absorption outside  $I_1^n$  at stages  $n > N_1$  is at most  $\delta$ . Similarly to step 1, choose an  $N_2 > N_1$  such that

$$p_2^{N_2} := \mathcal{P}_{f_2 g_2}(\xi > N_2) \leq \delta \cdot p^*.$$

Assume first that  $p_2^{N_2} = 0$ . Then we have  $\gamma_1(f, g_2) \leq \mathcal{A}_1 + \varepsilon$  for small  $\delta$ , because, with respect to  $(f, g_2)$ , the expected absorption payoff at each stage in  $n \leq N_1$  is at most  $\mathcal{A}_1$ ; the probability of absorption outside  $I_1^n$  at stages  $n = N_1 + 1, \dots, N_2$  is at most  $\delta$ ; the expected absorption payoff in  $I_1^n$  at each stage in  $n = N_1 + 1, \dots, N_2$  is at most  $\mathcal{A}_1$ ; the probability of absorption after stage  $N_2$  is zero; and the expected payoff in state 1 at each stage in  $n > N_2$  is at most  $\mathcal{A}_1$ .

Assume now that  $p_2^{N_2} > 0$ . Let  $I_2^n := \{i \in I \mid (i, j^{x_2^n}) \text{ is non-absorbing}\}$ , and repeat the above steps, in such a way that  $N_{k+1} > N_k$  for all  $k$ , until at some step  $K$  we have  $p_K^{N_K} = 0$ . This results in a strategy  $g_K$  for player 2. Note that for  $p_K^{N_K} = 0$  it is sufficient that  $I_K^n = I_{K-1}^n$  holds for all  $n > N_K$ , hence we only need at most  $K \leq |I|$  steps because, for any stage  $n > N_k$ , either  $I_{k+1}^n$  becomes smaller than  $I_k^n$ , or  $I_k^n = I_{k+1}^n$  and then nothing changes at further steps for stage  $n$ . Using similar arguments as before, one can show now that  $p_K^{N_K} = 0$  implies that  $\gamma_1(f, g_K) \leq \mathcal{A}_1 + \varepsilon$  if  $\delta > 0$  is small enough.  $\square$

### 5.3.2 Games with constant $\mathcal{A}$ or $\mathcal{B}$

In this section we show that  $\mathcal{A} = \mathcal{B}$  in games where  $\mathcal{A}$  or  $\mathcal{B}$  is constant. We need the following lemma.

**Lemma 72** *Let  $\varepsilon > 0$ . For  $\beta \in (0, 1)$ , let  $x_\beta \in X^*$  be a  $\beta$ -discounted optimal strategy. Then for large  $\beta \in (0, 1)$*

$$\underline{v}_s(x_\beta) \geq \min_{t \in S} v_t - \varepsilon \quad \forall s \in S.$$

**Proof.** By theorem 16-(b) and by the finiteness of the state space  $S$  and the space  $J$  of pure stationary strategies for player 2, it suffices to show that for any  $s \in S$  and  $j \in J$ , if  $\beta \in (0, 1)$  is large, then

$$\gamma_s(x_\beta, j) \geq \min_{t \in S} v_t - \varepsilon.$$

Let  $s \in S$  and  $j \in J$ . Using theorem 20 we have

$$(1 - \beta) \cdot r(x_\beta, j) + \beta \cdot P(x_\beta, j) \cdot v_\beta \geq v_\beta \quad \forall \beta \in (0, 1).$$

By (2.1), multiplying this inequality with  $Q(x_\beta, j)$  yields

$$Q(x_\beta, j) \cdot r(x_\beta, j) \geq Q(x_\beta, j) \cdot v_\beta \quad \forall \beta \in (0, 1).$$

Using theorem 9-(a) and theorem 22, we have for large  $\beta \in (0, 1)$  that

$$\begin{aligned} \gamma_s(x_\beta, j) &= \sum_{t \in S} q_s(t|x_\beta, j) r_t(x_{\beta t}, j_t) \\ &\geq \sum_{t \in S} q_s(t|x_\beta, j) v_{\beta t} \\ &\geq \sum_{t \in S} q_s(t|x_\beta, j) (v_t - \varepsilon) \\ &\geq \min_{t \in S} v_t - \varepsilon, \end{aligned}$$

so the proof is complete.  $\square$

With the help of the above lemma we show the following result.

**Theorem 73** *In every zero-sum stochastic game*

$$\min_{s \in S} \mathcal{A}_s = \min_{s \in S} \mathcal{B}_s = \min_{s \in S} v_s, \quad \max_{s \in S} \mathcal{A}_s = \max_{s \in S} \mathcal{B}_s = \max_{s \in S} v_s.$$

**Proof.** By lemma 72, for any  $\varepsilon > 0$  player 1 has a stationary strategy  $x^\varepsilon$  satisfying

$$\underline{v}_t(x^\varepsilon) \geq \min_{s \in S} v_s - \varepsilon \quad \forall t \in S,$$

hence

$$\min_{s \in S} \mathcal{A}_s \geq \min_{s \in S} v_s,$$

which in view of (5.1) implies the first part of the statement.

By the results of Thuijsman & Vrieze [1993], there is always a state  $t$  in  $S^{\max} := \{s \in S \mid v_s = \max_{w \in S} v_w\}$  for which player 1 has a stationary optimal strategy  $x$ . Hence

$$\max_{s \in S} \mathcal{A}_s \geq \mathcal{A}_t \geq \underline{v}_t(x) = v_t = \max_{s \in S} v_s,$$

thus (5.1) yields the second part of the statement.  $\square$

The above theorem yields the following corollary.

**Corollary 74** *In every zero-sum stochastic game where any of  $\mathcal{A}$ ,  $\mathcal{B}$ , or  $v$  is constant,  $\mathcal{A} = \mathcal{B}(= v)$  is constant.*

The following theorem provides a more relaxed view on constant values.

**Theorem 75** *In every zero-sum stochastic game where for all  $s, t \in S$  either  $\mathcal{A}_s = \mathcal{A}_t$  or  $\mathcal{B}_s = \mathcal{B}_t$  we have that  $\mathcal{A} = \mathcal{B}(= v)$  is constant.*

**Proof.** Using the inequality  $\mathcal{A} \leq \mathcal{B}$  and theorem 73, it is clear that if state  $s$  has the property that  $\mathcal{B}_s = \min_{w \in S} \mathcal{B}_w$  then  $\mathcal{A}_s = \min_{w \in S} \mathcal{A}_w$ . Similarly, if state  $t$  has the property that  $\mathcal{A}_t = \max_{w \in S} \mathcal{A}_w$  then  $\mathcal{B}_t = \max_{w \in S} \mathcal{B}_w$ . Now by the condition we have either  $\mathcal{A}_s = \mathcal{A}_t$  or  $\mathcal{B}_s = \mathcal{B}_t$ . Therefore by theorem 73 either  $\mathcal{A}$  or  $\mathcal{B}$  is constant, thus corollary 74 completes the proof.  $\square$

An interesting equivalent formulation of theorem 75 is the following: if  $\mathcal{A} \neq \mathcal{B}$  then there must exist two states  $s$  and  $t$  such that  $\mathcal{A}_s \neq \mathcal{A}_t$  and  $\mathcal{B}_s \neq \mathcal{B}_t$ .

### 5.3.3 Games with optimal strategies or with best-Markov strategies

By a best-Markov strategy we mean a Markov strategy  $f$  with the property that  $\underline{v}(f) \geq \underline{v}(\bar{f})$  for all  $\bar{f} \in F$ , or equivalently  $\underline{v}(f) = \mathcal{B}$ . Optimal strategies and best-Markov strategies do not necessarily exist, but if they do then their existence surprisingly implies  $\mathcal{A} = \mathcal{B}$ , as stated in the next theorem.

**Theorem 76** *In every zero-sum stochastic game, if player 1 has an optimal strategy or a best-Markov strategy then  $\mathcal{A} = \mathcal{B}$ .*

**Proof.** Suppose first that player 1 has an optimal strategy. Then by Main Theorem 3, player 1 has stationary  $\varepsilon$ -optimal strategies for all  $\varepsilon > 0$  and Markov optimal strategies as well, hence (5.1) yields the result.

Assume now that player 1 has a best-Markov strategy  $f$ , so  $\underline{v}(f) \geq \underline{v}(\bar{f})$  for all  $\bar{f} \in F$ . Since, for any history  $h \in H$ , the strategy  $f[h]$  is also a Markov strategy, we have  $\underline{v}(f) \geq \underline{v}(f[h])$  for all  $h \in H$ , hence  $f$  must be a non-improving strategy (cf. definition 46-(b)). Then by Main Theorem 4, for all  $\varepsilon > 0$ , player 1 has stationary strategies guaranteeing  $\underline{v}_s(f) - \varepsilon = \mathcal{B}_s - \varepsilon$  for all initial states  $s$ , hence  $\mathcal{A}_s \geq \mathcal{B}_s$ . Now (5.1) completes the proof.  $\square$

Note that in example 59 player 1 has neither optimal strategies nor best-Markov strategies for initial states 1 and 2. We only show it for initial state 2. One can argue as follows. Since  $\mathcal{B}_2 = v_2 = 1$  in that game, it suffices to show that player 1 has no strategy guaranteeing 1 for initial state 2. Assume by way of contradiction that a strategy  $\pi$  guarantees 1 for initial state 2. As the largest payoff in the game is 1,  $\pi$  has to prescribe action  $T_2$  with probability 1 whenever the play is in state 2 (otherwise the probability of absorption in entry  $(B_2, L_2)$  with payoff 0 would be positive if player 2 chooses action  $L_2$ ). Thus if player 2 always plays action  $R_2$  in state 2, then the reward is 0, which is a contradiction. Therefore, player 1 has neither optimal nor best-Markov strategies for initial state 2 indeed.

## 5.4 Concluding remarks

By the definition of  $\mathcal{A}$ , for each  $s \in S$  and for any  $\delta > 0$ , player 1 has a stationary strategy  $x^{s\delta} \in X$  such that  $\underline{v}_s(x^{s\delta}) \geq \mathcal{A}_s - \delta$ . In this finite state model it can be shown however that for any  $\delta > 0$  we can take  $x^{s\delta}$  independent of the initial state, so for all  $\delta > 0$  there exists a  $x^\delta \in X$  such that  $\underline{v}_s(x^\delta) \geq \mathcal{A}_s - \delta$  for all  $s \in S$ . This means that the following equality for stationary strategies makes sense:

$$\mathcal{A} = \sup_{x \in X} \underline{v}(x).$$

So we could have used this state independent equality as the definition of  $\mathcal{A}$  as well. Note that for games with countable state space this equivalence of definitions is not valid. Nowak & Raghavan [1991] presented a game with countable state space, where even though, for each initial state, player 1 has stationary  $\varepsilon$ -optimal strategies for all  $\varepsilon > 0$ , he has no stationary strategies that are  $\varepsilon$ -optimal for all initial states if  $\varepsilon$  is small.

Finally, we wish to remark that it is not known whether or not  $\mathcal{B}$  can be defined state independently.

## 5.5 Appendix

Here we provide a proof for lemma 64.

**Proof of lemma 64.** We prove the statement in steps. First, however, for all  $k \in \mathbb{N} \cup \{0\}$  and  $h^k \in V_t^k$  we introduce two events

$$A^k := \left( \theta^k = h^k, \theta^\infty \in U_t^\infty \right)$$

$$B^k := \left( \theta^{k+1} \in V^{k+1}, \theta^k = h^k, \theta^\infty \in U_t^\infty \right).$$

For  $n \in \mathbb{N}$  let

$$Z_{t\pi\sigma, V_t|U_t}^n(h^n) : = \prod_{k=0}^{n-1} \mathcal{P}_{t\pi\sigma} \left( \theta^{k+1} \in V_t^{k+1} | A^k \right),$$

$$T_{t\pi\sigma, V_t|U_t}^n(h^n) : = \prod_{k=0}^{n-1} \mathcal{P}_{t\pi\sigma} \left( \theta^{k+1} = h^{k+1} | B^k \right).$$

Notice that for all  $h \in V_t$

$$Z_{t\pi\sigma, V_t|U_t}(h) = \lim_{n \rightarrow \infty} Z_{t\pi\sigma, V_t|U_t}^n(h^n).$$

**Step 1.** First we show that for all  $n \in \mathbb{N}$

$$\sum_{h^n \in V_t^n} T_{t\pi\sigma, V_t|U_t}^n(h^n) = 1. \tag{5.13a}$$

We use induction on  $n$ . For  $n = 1$  we have

$$\begin{aligned} \sum_{h^1 \in V_t^1} T_{t\pi\sigma, V_i|U_t}^1(h^1) &= \sum_{h^1 \in V_t^1} \mathcal{P}_{t\pi\sigma}(\theta^1 = h^1 | B^0) \\ &= \mathcal{P}_{t\pi\sigma}(\theta^1 \in V_t^1 | B^0) \\ &= 1, \end{aligned}$$

thus (5.13a) holds for  $n = 1$ .

Now suppose that (5.13a) holds for  $n = m$ , where  $m \in \mathbb{N}$ . We need to show (5.13a) for  $n = m + 1$ . We have

$$\begin{aligned} \sum_{h^{m+1} \in V_t^{m+1}} T_{t\pi\sigma, V_i|U_t}^{m+1}(h^{m+1}) &= \sum_{h^m \in V_t^m} \sum_{\substack{\bar{h}^{m+1} \in V_t^{m+1} \\ \bar{h}^m = h^m}} T_{t\pi\sigma, V_i|U_t}^{m+1}(\bar{h}^{m+1}) \\ &= \sum_{h^m \in V_t^m} T_{t\pi\sigma, V_i|U_t}^m(h^m) \cdot \left[ \sum_{\substack{\bar{h}^{m+1} \in V_t^{m+1} \\ \bar{h}^m = h^m}} \mathcal{P}_{t\pi\sigma}(\theta^{m+1} = \bar{h}^{m+1} | B^m) \right] \\ &= \sum_{h^m \in V_t^m} T_{t\pi\sigma, V_i|U_t}^m(h^m) \\ &= 1, \end{aligned}$$

where the last equality holds by the assumption that (5.13a) is valid for  $n = m$ . Hence (5.13a) holds for all  $n \in \mathbb{N}$ .

**Step 2.** Now we show that

$$\mathcal{P}_{t\pi\sigma}(\theta \in V_t | \theta \in U_t) = \lim_{n \rightarrow \infty} \sum_{h^n \in V_t^n} Z_{t\pi\sigma, V_i|U_t}^n(h^n) \cdot T_{t\pi\sigma, V_i|U_t}^n(h^n). \quad (5.14)$$

For any  $h^{k+1} \in V_t^{k+1}$  we have

$$\mathcal{P}_{t\pi\sigma}(\theta^{k+1} = h^{k+1} | A^k) = \mathcal{P}_{t\pi\sigma}(\theta^{k+1} \in V_t^{k+1} | A^k) \cdot \mathcal{P}_{t\pi\sigma}(\theta^{k+1} = h^{k+1} | B^k),$$

hence for any  $h^n \in V_t^n$  we obtain

$$\begin{aligned} \mathcal{P}_{t\pi\sigma}(\theta^n = h^n | \theta \in U_t) &= \mathcal{P}_{t\pi\sigma}(\theta^1 = h^1 | A^0) \cdots \mathcal{P}_{t\pi\sigma}(\theta^n = h^n | A^{n-1}) \\ &= Z_{t\pi\sigma, V_i|U_t}^n(h^n) \cdot T_{t\pi\sigma, V_i|U_t}^n(h^n). \end{aligned}$$

Therefore

$$\begin{aligned} \mathcal{P}_{t\pi\sigma}(\theta \in V_t | \theta \in U_t) &= \lim_{n \rightarrow \infty} \mathcal{P}_{t\pi\sigma}(\theta^n \in V_t^n | \theta \in U_t) \\ &= \lim_{n \rightarrow \infty} \sum_{h^n \in V_t^n} \mathcal{P}_{t\pi\sigma}(\theta^n = h^n | \theta \in U_t) \\ &= \lim_{n \rightarrow \infty} \sum_{h^n \in V_t^n} Z_{t\pi\sigma, V_i|U_t}^n(h^n) \cdot T_{t\pi\sigma, V_i|U_t}^n(h^n), \end{aligned}$$

so the proof of (5.14) is complete.

**Step 3.** We show the validity of the lower-bound in the lemma. Using

$$Z_{t\pi\sigma, V_t|U_t}^n(h^n) \geq Z_{t\pi\sigma, V_t|U_t}(h) \quad \forall n \in \mathbb{N}, \forall h \in V_t, \quad (5.15)$$

equality (5.14) yields

$$\begin{aligned} \mathcal{P}_{t\pi\sigma}(\theta \in V_t | \theta \in U_t) &= \lim_{n \rightarrow \infty} \sum_{h^n \in V_t^n} Z_{t\pi\sigma, V_t|U_t}^n(h^n) \cdot T_{t\pi\sigma, V_t|U_t}^n(h^n) \\ &\geq \liminf_{n \rightarrow \infty} \left[ \left( \inf_{h^n \in V_t^n} Z_{t\pi\sigma, V_t|U_t}^n(h^n) \right) \cdot \sum_{h^n \in V_t^n} T_{t\pi\sigma, V_t|U_t}^n(h^n) \right] \\ &\geq \liminf_{n \rightarrow \infty} \left[ \left( \inf_{h \in V_t} Z_{t\pi\sigma, V_t|U_t}(h) \right) \cdot \sum_{h^n \in V_t^n} T_{t\pi\sigma, V_t|U_t}^n(h^n) \right] \\ &= \inf_{h \in V_t} Z_{t\pi\sigma, V_t|U_t}(h), \end{aligned}$$

where the last equality follows from (5.13a).

**Step 4.** We show that

$$\lim_{n \rightarrow \infty} \sup_{h^n \in V_t^n} Z_{t\pi\sigma, V_t|U_t}^n(h^n) = \sup_{h \in V_t} Z_{t\pi\sigma, V_t|U_t}(h). \quad (5.16)$$

Let

$$\rho := \sup_{h \in V_t} Z_{t\pi\sigma, V_t|U_t}(h).$$

Then in view of (5.15) it suffices to show that

$$\lim_{n \rightarrow \infty} \sup_{h^n \in V_t^n} Z_{t\pi\sigma, V_t|U_t}^n(h^n) \leq \rho. \quad (5.17)$$

Suppose by way of contradiction that there exists a  $d > 0$  such that

$$\lim_{n \rightarrow \infty} \sup_{h^n \in V_t^n} Z_{t\pi\sigma, V_t|U_t}^n(h^n) \geq \rho + d. \quad (5.18)$$

Let

$$\begin{aligned} W_t^n &:= \{h^n \in V_t^n | Z_{t\pi\sigma, V_t|U_t}^n(h^n) \geq \rho + d\} && \forall n \in \mathbb{N} \\ W_t^n[h^m] &:= \{\bar{h}^n \in W_t^n | \bar{h}^m = h^m\} && \forall h^m \in V_t^m, \forall m \leq n. \end{aligned}$$

It is clear that for all  $h \in V_t$

$$Z_{t\pi\sigma, V_t|U_t}^m(h^m) \geq Z_{t\pi\sigma, V_t|U_t}^n(h^n) \quad \forall m \leq n. \quad (5.19)$$

Therefore

$$\sup_{h^m \in V_t^m} Z_{t\pi\sigma, V_t|U_t}^m(h^m) \geq \sup_{h^n \in V_t^n} Z_{t\pi\sigma, V_t|U_t}^n(h^n) \quad \forall m \leq n. \quad (5.20)$$

Notice that (5.19) implies

$$(a) \quad h^n \in W_t^n \text{ implies } h^m \in W_t^m \text{ for all } m \leq n$$

and also that (5.20) and (5.18) yield

$$(b) \quad W_t^n \neq \emptyset \text{ for all } n \in \mathbb{N}.$$

Based on properties (a) and (b) we are going to construct a history  $\tilde{h} \in V_t$  with the property that

$$\tilde{h}^n \in W_t^n \text{ for all } n \in \mathbb{N}. \quad (5.21)$$

By property (b), the finiteness of  $W_t^1$  implies that there exists a  $\tilde{h}^1 \in W_t^1$  such that

$$|\{n \geq 1 \mid W_t^n[\tilde{h}^1] \neq \emptyset\}| = \infty.$$

Then using property (a) we obtain

$$(a') \quad h^n \in W_t^n[\tilde{h}^1] \text{ implies } h^m \in W_t^m[\tilde{h}^1] \text{ for all } m \leq n,$$

hence

$$(b') \quad W_t^n[\tilde{h}^1] \neq \emptyset \text{ for all } n \geq 1.$$

This means that the sets  $W_t^n[\tilde{h}^1]$ ,  $n \in \mathbb{N}$ , have similar properties as the sets  $W_t^n$ ,  $n \in \mathbb{N}$ . Applying the same arguments for the sets  $W_t^n[\tilde{h}^1]$ ,  $n \in \mathbb{N}$ , there exists a history  $\tilde{h}^2 \in W_t^2[\tilde{h}^1]$  such that  $W_t^n[\tilde{h}^2] \neq \emptyset$  for all  $n \geq 2$ . Repeating the steps above we find an infinite history  $\tilde{h}$  with the property that  $\tilde{h}^n \in W_t^n[\tilde{h}^{n-1}] \subset W_t^n$  for all  $n \in \mathbb{N}$ , thus the infinite history  $\tilde{h}$  satisfies (5.21).

By (5.21) we have for all  $n \in \mathbb{N}$  that

$$Z_{t\pi\sigma, V_t|U_t}^n(\tilde{h}^n) \geq \rho + d.$$

Now taking the limit yields

$$Z_{t\pi\sigma, V_t|U_t}(\tilde{h}) \geq \rho + d = \sup_{h \in V_t} Z_{t\pi\sigma, V_t|U_t}(h) + d,$$

which is a contradiction. Therefore (5.17) must hold, which completes the proof of (5.16).

**Step 5.** Finally, we show the validity of the upper-bound in the lemma. Using (5.14) we obtain

$$\begin{aligned} \mathcal{P}_{t\pi\sigma}(\theta \in V_t \mid \theta \in U_t) &= \lim_{n \rightarrow \infty} \sum_{h^n \in V_t^n} Z_{t\pi\sigma, V_t|U_t}^n(h^n) \cdot T_{t\pi\sigma, V_t|U_t}^n(h^n) \\ &\leq \limsup_{n \rightarrow \infty} \left[ \left( \sup_{h^n \in V_t^n} Z_{t\pi\sigma, V_t|U_t}^n(h^n) \right) \cdot \sum_{h^n \in V_t^n} T_{t\pi\sigma, V_t|U_t}^n(h^n) \right] \\ &= \sup_{h \in V_t} Z_{t\pi\sigma, V_t|U_t}(h), \end{aligned}$$

where the last equality follows from (5.13a) and (5.16).  $\square$

## Chapter 6

# Almost stationary $\varepsilon$ -equilibria

### 6.1 Introduction

So far in the literature of stochastic games, existence of  $(\varepsilon)$ -equilibria has been frequently established in terms of almost stationary strategy pairs (see for example Vrieze & Thuijsman [1989], Vieille [1993]?? or, in a more general fashion, Thuijsman & Vrieze [1996]). Intuitively, a pair of strategies is called almost stationary if, with probability almost 1, the players always use the same mixed actions in any state of the game; so the strategies behave as if they were simple stationary strategies. Formally, the concept of almost stationary  $\varepsilon$ -equilibria is defined as follows.

**Definition 77** *An  $\varepsilon$ -equilibrium  $(\pi, \sigma)$  is an almost stationary  $\varepsilon$ -equilibrium,  $\varepsilon \geq 0$ , if there exists a pair of stationary strategies  $(x^*, y^*)$  such that*

$$\mathcal{P}_{s\pi\sigma}(\pi_{s^n}(\theta^{n-1}) = x_{s^n}^*, \sigma_{s^n}(\theta^{n-1}) = y_{s^n}^* \quad \forall n \in \mathbb{N}) \geq 1 - \varepsilon \quad \forall s \in S,$$

where  $\theta^n$  denotes the random variable for the history up to stage  $n$  and  $s^n$  the random variable for the state at stage  $n$ .

Note that although  $\varepsilon$  has two different roles in this definition, it will lead to no confusion.

The main reason for dealing with almost stationary  $\varepsilon$ -equilibria is the fact that they can usually be treated easier than general  $\varepsilon$ -equilibria due to the simple structure of stationary strategies. However, they are also more appealing, as the players settle on playing simple stationary strategies and only need to switch to other strategies probabilities close to zero.

It is an interesting fact that, in zero-sum stochastic games, on the one hand  $\varepsilon$ -equilibria are known to exist for all  $\varepsilon > 0$  (as we have mentioned in section 2.6, any pair of  $\varepsilon$ -optimal strategies yields a  $2\varepsilon$ -equilibrium), but on the other hand the existence of almost stationary  $\varepsilon$ -equilibria has remained an open problem. In this chapter, which is based on Flesch et al. [1998,II], we will answer this question in the affirmative, namely we will construct almost stationary  $\varepsilon$ -equilibria, for all  $\varepsilon > 0$ , in all zero-sum stochastic games.

**Main Theorem 6** *In every zero-sum stochastic game, for any  $\varepsilon > 0$ , there exists an almost stationary  $\varepsilon$ -equilibrium.*

In view of lemma 24-(d),(e), it is clear that, in zero-sum stochastic games, 0-equilibria do not necessarily exist and also that history dependent strategies are indispensable for obtaining  $\varepsilon$ -equilibria,  $\varepsilon > 0$ , so the result is sharp in this sense.

The proof will be based on a construction for specific stationary strategy pairs with rewards equal to the value. It is clear that the value is acceptable for both players, as neither player is able to guarantee a better reward in his favour. However, in order to force the players to play these stationary strategies, as a standard tool, the players will use statistical tests on the past action frequencies of their opponents to detect deviations with probability almost 1. If a deviation is detected then a history dependent  $\delta$ -optimal strategy has to be played in the future, where  $\delta > 0$  is sufficiently small. The role of these  $\delta$ -optimal strategies is to rule out the profitability of possible deviations of the players. To illustrate the issue, we now briefly discuss the Big Match.

**Example 78**

	$L$	$R$
$T$	0	1
$B$	1 *	0 *
	1	

This game is the Big Match (cf. example 23). In lemma 24-(b),(c) we showed an  $\varepsilon$ -optimal strategy  $\pi^\varepsilon$  for player 1 and a stationary optimal strategy  $y = (1/2, 1/2)$  for player 2. This pair of strategies  $(\pi^\varepsilon, y)$  would be an  $\varepsilon$ -equilibrium that yields precisely  $1/2$  to player 1. Instead of achieving this  $1/2$  through this complicated strategy  $\pi^\varepsilon$ , the players could play the almost stationary  $\varepsilon$ -equilibrium  $(\xi^\varepsilon, y)$ , where  $\xi^\varepsilon$  is the strategy defined, roughly speaking, by: play action  $T$  unless at some stage in the far future you notice that player 2's action frequencies are not sufficiently close to  $(1/2, 1/2)$ , in that case immediately start playing  $\pi^\varepsilon$ . Notice that if player 2 truly plays  $y$  then the probability that player 2's action frequencies are not close enough to  $(1/2, 1/2)$  in the far future is very small (by the law of large numbers). Hence, with probability almost 1, the players play stationary strategies forever.

## 6.2 Preliminaries

For  $s \in S$ ,  $x_s \in X_s$ ,  $y_s \in Y_s$  let  $L_s(x_s, y_s)$  be the probability that, after transition from state  $s$  with respect to  $(x_s, y_s)$ , the new value  $v_t$  is different from  $v_s$ , so

$$L_s(x_s, y_s) := \sum_{t \in S, v_t \neq v_s} p_s(t|x_s, y_s)$$

(if  $v_t = v_s$  for all  $t \in S$  then  $L_s(x_s, y_s)$  is defined to equal 0). Obviously,  $V_s(x_s, y_s) \neq v_s$  implies  $L_s(x_s, y_s) > 0$  (recall definition 34). The next lemma states that if, with

respect to  $(x, y)$ , the value does not change in expectation under transitions then the value is a constant on each set of states that is ergodic with respect to  $(x, y)$ .

**Lemma 79** *Let  $(x, y) \in X \times Y$  satisfy  $V(x, y) = v$ . Suppose that  $E$  is an ergodic set with respect to  $(x, y)$ . Then  $v_s = v_t$  for all  $s, t \in E$ , and therefore  $L_s(x_s, y_s) = 0$  for all  $s \in E$ .*

**Proof.** Let  $\bar{E} := \{s \in E \mid v_s = \max_{t \in E} v_t\}$ . Using  $V(x, y) = v$  and the fact that  $E$  is an ergodic set for  $(x, y)$  we obtain

$$v_s = V_s(x_s, y_s) = \sum_{t \in S} p_s(t \mid x_s, y_s) v_t = \sum_{t \in E} p_s(t \mid x_s, y_s) v_t \quad \forall s \in \bar{E},$$

thus  $\bar{E} \subset E$  is a closed set of states for  $(x, y)$ . Therefore  $\bar{E} = E$ , which implies  $v_s = v_t$  for all  $s, t \in E$ . Now  $L_s(x_s, y_s) = 0$  for all  $s \in E$  follows from the definition of  $L_s(x_s, y_s)$ .  $\square$

For  $s \in S$  let

$$\begin{aligned} X'_s &:= \{x_s \in X_s \mid V_s(x_s, y_s) \geq v_s \quad \forall y_s \in Y_s\}, & X' &:= \times_{s \in S} X'_s, \\ Y'_s &:= \{y_s \in Y_s \mid V_s(x_s, y_s) \leq v_s \quad \forall x_s \in X_s\}, & Y' &:= \times_{s \in S} Y'_s, \\ \bar{X}_s &:= \{x_s \in X_s \mid V_s(x_s, y_s) = v_s \quad \forall y_s \in Y'_s\}, & \bar{X} &:= \times_{s \in S} \bar{X}_s, \\ \bar{Y}_s &:= \{y_s \in Y_s \mid V_s(x_s, y_s) = v_s \quad \forall x_s \in X'_s\}, & \bar{Y} &:= \times_{s \in S} \bar{Y}_s. \end{aligned}$$

By lemmas 35 and 33, the above sets are nonempty polytopes. Let  $\bar{I}_s$  and  $\bar{J}_s$  denote the sets of extreme points of  $\bar{X}_s$  and  $\bar{Y}_s$ , respectively. Recall that the relative interior of the polytope  $Z$ , denoted by  $\text{Relint}(Z)$ , 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. Due to lemma 33 again, for all  $s \in S$ ,  $x_s \in \text{Relint}(X'_s)$ ,  $y_s \in \text{Relint}(Y'_s)$ , we have

$$\bar{I}_s = \{i_s \in I_s \mid x_s(i_s) > 0\}, \quad \bar{J}_s = \{j_s \in J_s \mid y_s(j_s) > 0\}. \quad (6.1)$$

The next lemma provides sufficient conditions for  $\bar{X}_s = X'_s$  and for  $\bar{Y}_s = Y'_s$  in some state  $s \in S$ .

**Lemma 80** *Let  $s \in S$ . If  $L_s(x_s, j_s) > 0$  implies  $V_s(x_s, j_s) > v_s$  for all  $(x_s, j_s) \in X'_s \times J_s$ , then  $\bar{X}_s = X'_s$ . Similarly, if  $L_s(i_s, y_s) > 0$  implies  $V_s(i_s, y_s) < v_s$  for all  $(i_s, y_s) \in I_s \times Y'_s$ , then  $\bar{Y}_s = Y'_s$ .*

**Proof.** We only show the first part; the proof of the second part is similar. By (6.1) we have  $\bar{X}_s \supset X'_s$ . It remains to verify that  $\bar{X}_s \subset X'_s$ . Since  $X'_s$  is convex, it is sufficient to show that  $i_s \in X'_s$  for all  $i_s \in \bar{I}_s$ . Take an arbitrary  $i_s \in \bar{I}_s$ . Using the compactness of  $X'_s$ , there exists an  $\hat{x}_s \in X'_s$  satisfying  $\hat{x}_s(i_s) \geq x_s(i_s)$  for all  $x_s \in X'_s$ . By the condition we have that  $L_s(\hat{x}_s, j_s) > 0$  implies  $V_s(\hat{x}_s, j_s) > v_s$  for all  $j_s \in J_s$ , therefore, by using  $\hat{x}_s \in X'_s$ , we obtain that  $((1 - \lambda) \cdot \hat{x}_s + \lambda \cdot i_s) \in X'_s$  for small  $\lambda > 0$ .

By the choice of  $\hat{x}_s$  we must have  $\hat{x}_s = i_s$ , thus  $i_s \in X'_s$ .  $\square$

Thuijsman & Vrieze [1993] showed that in every zero-sum game there exists an initial state  $s_1$  in  $S^{\max} := \{s \in S \mid v_s = \max_{t \in S} v_t\}$  for which player 1 has a stationary optimal strategy  $x^1$ , and similarly, there exists an initial state  $s_2$  in  $S^{\min} := \{s \in S \mid v_s = \min_{t \in S} v_t\}$  for which player 2 has a stationary optimal strategy  $y^2$ . Obviously, the strategy  $x^1$  must keep the play in  $S^{\max}$  with probability 1 when starting in  $s^1$ , and  $y^2$  must keep the play in  $S^{\min}$  with probability 1 when starting in  $s^2$ . Hence if we take stationary best replies  $y^1$  against  $x^1$  and  $x^2$  against  $y^2$  then we obtain the following result.

**Lemma 81** *There exist stationary strategy pairs  $(x^1, y^1)$ ,  $(x^2, y^2)$  and corresponding ergodic sets  $E^1, E^2$  such that*

$$E^1 \subset S^{\max} := \{s \in S \mid v_s = \max_{t \in S} v_t\}, \quad \gamma_s(x^1, y^1) = v_s \quad \forall s \in E^1,$$

$$E^2 \subset S^{\min} := \{s \in S \mid v_s = \min_{t \in S} v_t\}, \quad \gamma_s(x^2, y^2) = v_s \quad \forall s \in E^2.$$

Suppose that  $E \subset S$  is an ergodic set with respect to  $(x', y') \in \text{Relint}(X') \times \text{Relint}(Y')$  and also that  $\bar{X}_s = X'_s, \bar{Y}_s = Y'_s$  for all  $s \in E$ . Then we may define a restricted game  $\bar{\Gamma}_E$ , as in chapter 3, where the state space is  $E$  and the players are restricted to use strategies that only prescribe actions in  $\bar{I}_s$  and  $\bar{J}_s$ , if the play is in state  $s \in E$ . Obviously, this restricted game  $\bar{\Gamma}_E$  is a well-defined stochastic game as well. Let  $\bar{v}$  denote the value of the restricted game  $\bar{\Gamma}_E$ . Observe that for the original value, by lemma 79, we have  $v_s = v_t =: v_E$  for all  $s, t \in E$ . It follows from the first concluding remark in chapter 3 that if either  $\bar{v}_s \geq v_E$  for all  $s \in E$  or  $\bar{v}_s \leq v_E$  for all  $s \in E$  then there exists a state  $s \in E$  such that  $\bar{v}_s = v_E$ . Recall that the value of the restricted game  $\bar{v}_s, s \in E$ , does not need to be equal to the original value  $v_E$  in all states in  $E$ , not even under the above condition.

### 6.3 The construction

Fix arbitrary  $x' \in \text{Relint}(X')$  and  $y' \in \text{Relint}(Y')$ . We keep  $x'$  and  $y'$  fixed for the rest of this section. Let  $T$  denote the set of transient states and  $\mathcal{R}$  the set of ergodic sets with respect to  $(x', y')$ . Since any stationary strategy pair induces at least one ergodic set, we have  $\mathcal{R} \neq \emptyset$ . Now we divide  $\mathcal{R}$  into three parts. Let

$$\mathcal{R}^1 := \{E \in \mathcal{R} \mid \exists s \in E, \exists (i_s, y_s) \in I_s \times Y'_s : V_s(i_s, y_s) = v_s, L_s(i_s, y_s) > 0\}$$

$$\mathcal{R}^2 := \{E \in \mathcal{R} \setminus \mathcal{R}^1 \mid \exists s \in E, \exists (x_s, j_s) \in X'_s \times J_s : V_s(x_s, j_s) = v_s, L_s(x_s, j_s) > 0\}$$

$$\mathcal{R}^3 := \mathcal{R} \setminus (\mathcal{R}^1 \cup \mathcal{R}^2).$$

Note that all the sets  $T, \mathcal{R}^1, \mathcal{R}^2, \mathcal{R}^3$  are independent of the particular choices of  $x' \in \text{Relint}(X')$  and  $y' \in \text{Relint}(Y')$ . Here  $\mathcal{R}^1$  is the set of ergodic sets  $E$  with respect to  $(x', y')$  for which there exists a pair of mixed actions in some state  $s \in E$  such that player 1 plays a “pure” action, player 2 plays a mixed action in  $Y'_s$ , and the expected

value after transition equals the original value, but with a positive probability a transition occurs to a state where the value is different. The intuition behind  $\mathcal{R}^2$  is analogous. The partition of  $\mathcal{R}$  naturally induces the following partition of  $S \setminus T$ :

$$S^1 := \cup_{E \in \mathcal{R}^1} E, \quad S^2 := \cup_{E \in \mathcal{R}^2} E, \quad S^3 := \cup_{E \in \mathcal{R}^3} E.$$

If  $\mathcal{R}^1 \cup \mathcal{R}^2 \neq \emptyset$ , then by the definitions of  $\mathcal{R}^1$  and  $\mathcal{R}^2$  there exists a nonempty set  $S^* \subset S^1 \cup S^2$ , which contains precisely one state from each ergodic set in  $\mathcal{R}^1 \cup \mathcal{R}^2$ , such that, for all  $s \in S^* \cap S^1$ , there exists a pair  $(i_s^*, y_s^*) \in I_s \times Y_s'$  satisfying  $V_s(i_s^*, y_s^*) = v_s$ ,  $L_s(i_s^*, y_s^*) > 0$  and, for all  $s \in S^* \cap S^2$ , there exists a pair  $(x_s^*, j_s^*) \in X_s' \times J_s$  satisfying  $V_s(x_s^*, j_s^*) = v_s$ ,  $L_s(x_s^*, j_s^*) > 0$ . In fact, these states and pairs of mixed actions provide the possibility to leave all the ergodic sets belonging to  $\mathcal{R}^1$  and  $\mathcal{R}^2$  in such a way that the value does not change in expectation.

We will now turn our attention to  $\mathcal{R}^3$ . By using  $S^3 \cap (S^1 \cup S^2) = \emptyset$ , in light of lemma 80, we have  $\bar{X}_s = X_s'$  and  $\bar{Y}_s = Y_s'$  for all  $s \in S^3$ . Assume that  $E \in \mathcal{R}^3$  (in fact, later we will show that  $\mathcal{R}^3$  is always nonempty). Consider the restricted game  $\bar{\Gamma}_E$ . Clearly, in this restricted game the respective stationary strategy spaces are  $\bar{X}_E := \times_{s \in E} \bar{X}_s$  and  $\bar{Y}_E := \times_{s \in E} \bar{Y}_s$ . We use  $\bar{v}_s$ ,  $s \in E$ , for the value of the restricted game  $\bar{\Gamma}_E$ . Recall that for the original value we have  $v_s = v_t =: v_E$  for all  $s, t \in E$ . In the restricted game  $\bar{\Gamma}_E$ , in order to avoid confusion, we use  $\bar{q}$  and  $\bar{\gamma}$  instead of  $q$  and  $\gamma$ . In the next important lemma, we show the existence of stationary strategy pairs in  $\bar{\Gamma}_E$  with rewards equal to the original value  $v_E$ .

**Lemma 82** *Let  $E \in \mathcal{R}^3$ . There exists a stationary strategy pair  $(\bar{x}, \bar{y}) \in \bar{X}_E \times \bar{Y}_E$  such that  $\bar{\gamma}_s(\bar{x}, \bar{y}) = v_E$  for all  $s \in E$ .*

**Proof.** We distinguish two essentially different cases.

**Part 1:** Assume that  $\bar{v}_s \geq v_E$  for all  $s \in E$  (if  $\bar{v}_s \leq v_E$  for all  $s \in E$ , then an analogous proof can be applied).

As mentioned in section 6.2, there must exist a state  $s \in E$  such that  $\bar{v}_s = v_E$ . Let  $E^{\min} := \{t \in E \mid \bar{v}_t = v_E\}$ . Let  $(x^2, y^2) \in \bar{X}_E \times \bar{Y}_E$  and let  $E^2 \subset E^{\min}$  in  $\bar{\Gamma}_E$  as in Lemma 81. So we have  $\bar{\gamma}_s(x^2, y^2) = v_E$  for all  $s \in E^2$ . For  $s \in E$  let

$$\bar{x}_s := \begin{cases} x_s^2 & \text{if } s \in E^2 \\ x_s' & \text{if } s \in E \setminus E^2 \end{cases}, \quad \bar{y}_s := \begin{cases} y_s^2 & \text{if } s \in E^2 \\ y_s' & \text{if } s \in E \setminus E^2. \end{cases}$$

The only ergodic set for  $(\bar{x}, \bar{y}) \in \bar{X}_E \times \bar{Y}_E$  in the restricted game  $\bar{\Gamma}_E$  is clearly  $E^2$ , hence for any  $s, t \in E$  we have that  $\bar{q}_s(t \mid \bar{x}, \bar{y}) > 0$  only holds if  $t \in E^2$ , thus lemma 9-(c) yields  $\bar{\gamma}_s(\bar{x}, \bar{y}) = v_E$  for all  $s \in E$ .  $\diamond$

**Part 2:** Assume that  $\min_{s \in E} \bar{v}_s < v_E < \max_{s \in E} \bar{v}_s$ .

Take  $(x^1, y^1) \in \bar{X}_E \times \bar{Y}_E$ ,  $E^1 \subset E^{\max} := \{s \in E \mid \bar{v}_s = \max_{t \in E} \bar{v}_t\}$  and  $(x^2, y^2) \in \bar{X}_E \times \bar{Y}_E$ ,  $E^2 \subset E^{\min} := \{s \in E \mid \bar{v}_s = \min_{t \in E} \bar{v}_t\}$  in  $\bar{\Gamma}_E$  as in lemma 81. By the assumption we have  $E^1 \cap E^2 = \emptyset$ . For  $a, b \in (0, 1)$  and  $s \in E$  let

$$(x_s^{ab}, y_s^{ab}) := \begin{cases} (a \cdot x_s^1 + (1-a) \cdot x_s', a \cdot y_s^1 + (1-a) \cdot y_s') & \text{if } s \in E^1 \\ (b \cdot x_s^2 + (1-b) \cdot x_s', b \cdot y_s^2 + (1-b) \cdot y_s') & \text{if } s \in E^2 \\ (x_s', y_s') & \text{if } s \in E \setminus (E^1 \cup E^2). \end{cases}$$

(Recall that we have fixed  $x' \in \text{Relint}(X')$  and  $y' \in \text{Relint}(Y')$ , and also that  $\bar{X}_s = X'_s$ ,  $\bar{Y}_s = Y'_s$  for all  $s \in E$ .) Notice that  $x_s^{ab} \in \text{Relint}(\bar{X}_s)$  and  $y_s^{ab} \in \text{Relint}(\bar{Y}_s)$  for all  $s \in E$  and  $a, b \in (0, 1)$ , hence the set  $E$  is ergodic for  $(x^{ab}, y^{ab})$  for all  $a, b \in (0, 1)$ . Notice also that  $a$  and  $b$  control the respective expected lengths of periods when staying in  $E^1$  and  $E^2$ . Since  $E$  is ergodic for  $(x^{ab}, y^{ab})$  for all  $a, b \in (0, 1)$ , lemma 9-(d) implies that  $\bar{\gamma}_s(x^{ab}, y^{ab}) = \bar{\gamma}_t(x^{ab}, y^{ab}) := \bar{\gamma}_E^{ab}$  for all  $s, t \in E$ ,  $a, b \in (0, 1)$ .

We show that there are  $a, b \in (0, 1)$  such that  $\bar{\gamma}_E^{ab} = v_E$ . Take arbitrary  $a', b' \in (0, 1)$ . If  $\bar{\gamma}_E^{a'b'} = v_E$  then we are done. So assume without loss of generality that  $\bar{\gamma}_E^{a'b'} > v_E$  and consider  $(x^{a'b}, y^{a'b})$ . Observe that the larger  $b$  we take, the more time the play spends in  $E^2$ . Thus one can show that

$$\lim_{b \uparrow 1} \bar{\gamma}_E^{a'b} = \min_{t \in E} \bar{v}_t < v_E.$$

By lemma 10, we have that  $\bar{\gamma}_E^{a'b}$  is continuous in  $b \in (0, 1)$ , hence there is a  $b$  such that  $\bar{\gamma}_E^{a'b} = v_E$ .  $\square$

Now we are ready to complete the construction based on the previously derived results. Recall that we have already fixed a pair of stationary strategies  $(x', y') \in \text{Relint}(X') \times \text{Relint}(Y')$ . For all ergodic sets  $E \in \mathcal{R}^3$  let  $(\bar{x}_s, \bar{y}_s) \in \bar{X}_s \times \bar{Y}_s$ ,  $s \in E$ , be as in lemma 82. We define a stationary strategy pair for all  $\tau \in (0, 1)$ :

$$(x_s^\tau, y_s^\tau) := \begin{cases} (\tau \cdot x'_s + (1 - \tau) \cdot i_s^*, y_s^*) & \text{if } s \in S^* \cap S^1 \\ (x_s^*, \tau \cdot y'_s + (1 - \tau) \cdot j_s^*) & \text{if } s \in S^* \cap S^2 \\ (\bar{x}_s, \bar{y}_s) & \text{if } s \in S^3 \\ (x'_s, y'_s) & \text{otherwise.} \end{cases}$$

The next lemma shows that for these stationary strategy pairs the recurrent states all belong to  $S^3$  and the reward equals the value for all initial states.

**Lemma 83** *For all  $\tau \in (0, 1)$ , we have  $\gamma(x^\tau, y^\tau) = v$  and, if  $U$  is an ergodic set with respect to  $(x^\tau, y^\tau)$ , then  $U \subset S^3$ .*

**Proof.** Let  $\tau \in (0, 1)$ . By the definitions, we have  $V_s(x_s^\tau, y_s^\tau) = v_s$  for all  $s$ . In view of lemma 79, the value is a constant on each ergodic set for  $(x^\tau, y^\tau)$ . By the construction of  $(x^\tau, y^\tau)$ , in each state in  $S^*$ , a transition occurs to a state with a different value with a positive probability, so all recurrent states must belong to  $S^3$ .

The equality  $V(x^\tau, y^\tau) = v$  implies  $P(x^\tau, y^\tau)v = v$ . By using induction, we have for all  $n \in \mathbb{N}$  that  $P^n(x^\tau, y^\tau)v = v$ , hence the definition of  $Q(x^\tau, y^\tau)$  yields

$$Q(x^\tau, y^\tau)v = v.$$

For any  $s \in S$ , if  $q_s(t|x^\tau, y^\tau) > 0$  then  $t$  belongs to an ergodic set with respect to  $(x^\tau, y^\tau)$ , so we have  $t \in S^3$ . Now the choice of  $(\bar{x}_z, \bar{y}_z)$ ,  $z \in S^3$ , implies by Lemma 82 that  $\gamma_t(x^\tau, y^\tau) = v_t$  for all  $t \in S^3$ , so applying lemma 9-(c) gives

$$\begin{aligned} \gamma_s(x^\tau, y^\tau) &= \sum_{t \in S} q_s(t|x^\tau, y^\tau) \cdot \gamma_t(x^\tau, y^\tau) \\ &= \sum_{t \in S^3} q_s(t|x^\tau, y^\tau) \cdot v_t \\ &= v_s \quad \forall s \in S, \end{aligned}$$

which completes the proof.  $\square$

Finally, we turn to the proof of Main Theorem 6. We show that, for any  $\varepsilon > 0$ , the stationary strategy pair  $(x^\tau, y^\tau)$ , for sufficiently large  $\tau \in (0, 1)$ , can be supplemented with history dependent  $\delta$ -optimal strategies, for small  $\delta > 0$ , in order to obtain an almost stationary  $\varepsilon$ -equilibrium.

**Proof of Main Theorem 6:**

We only give an outline of the proof, since the tools used are standard (see for example Vrieze & Thuijsman [1989], Vieille [1993] or, in a more general fashion, Thuijsman & Vrieze [1996]). Let  $\varepsilon > 0$ . We will define strategy pairs  $(\pi^\tau, \sigma^\tau)$  for all  $\tau \in (0, 1)$  so that  $(\pi^\tau, \sigma^\tau)$  is an almost stationary  $\varepsilon$ -equilibrium for sufficiently large  $\tau \in (0, 1)$ . These strategy pairs will be constructed in such a way that if neither player deviates then the mixed actions according to the stationary strategy pair  $(x^\tau, y^\tau)$  are played forever with probability at least  $\tau$ . In view of lemma 83, this means that the corresponding rewards are converging to the value  $v$  as  $\tau$  tends to 1. Hence, when verifying the  $\varepsilon$ -equilibrium conditions, it suffices to show that, for any initial state  $s \in S$ , player 1 cannot get more than  $v_s + \frac{\varepsilon}{2}$  and player 2 cannot decrease the reward below  $v_s - \frac{\varepsilon}{2}$  by unilateral deviations. The strategies  $\pi^\tau$  and  $\sigma^\tau$  will be analogously defined, so we only focus on player 1's strategy  $\pi^\tau$  and on the possible deviations of player 2.

Let  $\tau \in (0, 1)$  be close to 1. Player 1's strategy  $\pi^\tau$  will use the mixed actions according to  $x^\tau$  unless, on condition that player 2 should play the mixed actions according to  $y^\tau$ , player 1 detects with probability almost 1 that player 2 has deviated from  $y^\tau$ . If player 1 detects such a deviation then player 1 starts playing a  $(1 - \tau)$ -optimal strategy. Player 2's possible deviations are detected by means of employing statistical tests on player 2's behavior during the past history. Such a statistical test, with respect to some arbitrary stationary strategy pair  $(\tilde{x}, \tilde{y})$ , is based on the observations that, if player 2 truly uses his stationary strategy  $\tilde{y}$ , then: (1) player 2 never chooses an action which has probability zero with regard to  $\tilde{y}$ ; (2) if the play remains in the same ergodic set (ergodic with respect to  $(\tilde{x}, \tilde{y})$ ), then the empirical action frequencies of player 2 should converge to the weights of the mixed actions corresponding to  $\tilde{y}$  (by the law of large numbers); (3) from any transient state (transient with respect to  $(\tilde{x}, \tilde{y})$ ), the probability of remaining in the transient states longer than  $n$  stages, converges to zero as  $n$  tends to infinity. So, if player 2 chooses an action with probability zero according to  $\tilde{y}$ , then player 1 knows for sure that player 2 has deviated; if after some specified number of stages within an ergodic set, player 2's action frequencies are not within some specified range from the theoretical ones, then player 2 has deviated with probability close to 1; if the play remains in the set of transient states for longer than some specified number of stages, then player 2 has deviated with probability close to 1. Note that all these probabilities are conditioned on the initially given stationary strategies.

First we consider the case when player 2 chooses an action  $j_s \in J_s$  in state  $s \in S$  with  $y_s^\tau(j_s) = 0$  (see observation (1) above). Then, clearly, player 1 immediately notices the deviation, so the inequalities  $\lim_{\tau \uparrow 1} V_s(x_s^\tau, j_s) \geq v_s$  for all  $j_s \in J_s$ ,  $s \in S$ , assure

that by choosing any action  $j_s \in J_s$  in any state  $s \in S$  with  $y_s^\tau(j_s) = 0$ , the reward is at least  $V_s(x_s^\tau, j_s) - (1 - \tau) \geq v_s - \frac{\varepsilon}{2}$  if  $\tau$  is large enough, using that  $\pi^\tau$  prescribes a  $(1 - \tau)$ -optimal strategy afterwards.

Now we assume the other case, namely that player 2 only prescribes actions which have positive probabilities with respect to  $y^\tau$ . We divide the set of stages up to the current stage into blocks  $B^k$  of consecutive stages as follows: a new block starts at each stage the play enters  $T$ , or a set  $E \in \mathcal{R}$ , or an ergodic set  $U$  with respect to  $(x^\tau, y^\tau)$  (we must have  $U \subset S^3$  in view of lemma 83). In block  $B^k$  the probability that player 1 detects a deviation of player 2 although player 2 truly used  $y^\tau$  will be at most  $d^k$ , where  $d^k \in (0, 1)$  for all  $k \in \mathbb{N}$  and  $\sum_{k=1}^{\infty} d^k \leq 1 - \tau$ . The latter inequality will guarantee that the total probability of making this mistake is at most  $1 - \tau$ .

Assume that the play enters some ergodic set  $U \subset S^3$  (with respect to  $(x^\tau, y^\tau)$ ) and the new block is  $B^k$ . In this ergodic set, player 1 checks the action frequencies of player 2, and if the empirical action frequencies are not close enough to the theoretical ones then player 1 detects a deviation (see observation (2) above with  $(\tilde{x}, \tilde{y}) = (x^\tau, y^\tau)$ ) and starts playing a  $(1 - \tau)$ -optimal strategy. If the number of stages in this block  $B^k$  is large enough, then the probability that player 1 detects a deviation although player 2 used  $y^\tau$  is at most  $d^k$ . If the empirical action frequencies are close to the theoretical ones, then the corresponding reward is close to the value. Notice that the play never leaves  $U$  if the players only use actions which are chosen with positive probabilities with respect to the pair  $(x^\tau, y^\tau)$ .

Assume that the play enters  $T$ , or a set  $E^2 \in \mathcal{R}^2$ , or a set  $E^3 \in \mathcal{R}^3$  but not an ergodic set  $U \subset S^3$  (ergodic with respect to  $(x^\tau, y^\tau)$ ), and that the new block is  $B^k$ . Then, by lemma 83, if player 2 uses  $y^\tau$ , the play should leave  $T$ , or  $E^2$ , or enter an ergodic set  $U \subset E^3 \subset S^3$ ,  $U \neq E^2$  (ergodic with respect to  $(x^\tau, y^\tau)$ ) within  $N^k$  stages, for large  $N^k$ , with probability at least  $1 - d^k$  (see observation (3) above with  $(\tilde{x}, \tilde{y}) = (x^\tau, y^\tau)$ ). If the play does not leave  $T$  or  $E$  within  $N^k$  stages then player 1 detects a deviation of player 2, with probability at least  $1 - d^k$ , so he starts playing a  $(1 - \tau)$ -optimal strategy afterwards. Notice that  $x_s^\tau \in X'_s$  for all  $s \in T \cup S^2 \cup S^3$ , hence the play can only leave in such a way that the value does not decrease in expectation.

Finally, assume that the play enters some  $E \in \mathcal{R}^1$  and the new block is  $B^k$ . Notice that, if  $\tau$  is large, then  $x^\tau$  almost equals  $x'$  in all states in  $E$ , thus the set  $E$  is “almost” ergodic for  $(x^\tau, y^\tau)$ . Therefore, player 1 has enough time to check the action frequencies of player 2 in  $E$  (see observation (2) above with  $(\tilde{x}, \tilde{y}) = (x', y^\tau)$ ). This way player 1 can make sure that the unique state  $s$  in  $S^* \cap E$  is visited frequently enough and also that the play leaves  $E$  via  $i_s^*$  and the new value does not differ “much” from  $v_s$  (recall that  $V_s(i_s^*, y_s^*) = v_s$ ). If player 2 truly uses  $y^\tau$  then player 1 detects no deviation with probability at least  $1 - d^k$ .

We have described how player 1 makes sure that the reward is not much less than the value once the play reaches an ergodic set  $U \subset S^3$  (ergodic with respect to  $(x^\tau, y^\tau)$ ), and also that the play eventually reaches such an ergodic set in such a way that the value does not drop “much” in expectation, so, by taking a sufficiently large  $\tau \in (0, 1)$ , the proof is complete.  $\square$

## 6.4 Examples

We provide two examples to illustrate the construction of almost stationary  $\varepsilon$ -equilibria.

### Example 84

	$L$	$R$		
$T$	0	1		
	1	1		
$B$	1	0	1	
	2	3	2	3
	1		2	3

We reexamine the Big Match (cf. example 78). This example shows how the ergodic sets in  $\mathcal{R}^1$  and  $\mathcal{R}^2$  can be left in such a way that the value does not change “much” in expectation. In view of lemma 24-(a), the value is known to be  $v = (1/2, 1, 0)$ . Following the construction above, we have

$$X'_1 = \{(1, 0)\}, \quad X'_2 = X'_3 = \{(1)\},$$

$$Y'_1 = \text{conv} \{(1/2, 1/2), (0, 1)\}, \quad Y'_2 = Y'_3 = \{(1)\},$$

$$\mathcal{R}^1 = \{\{1\}\}, \quad S^1 = \{1\}, \quad \mathcal{R}^2 = \emptyset, \quad S^2 = \emptyset, \quad \mathcal{R}^3 = \{\{2\}, \{3\}\}, \quad S^3 = \{2, 3\},$$

where  $\text{conv}$  stands for the convex hull of a set. To see that  $S^1 = \{1\}$  take  $S^* = \{1\}$ ,  $i_1^* = B$ ,  $y_1^* = (1/2, 1/2)$ . As  $X'$  is a singleton and states 2 and 3 are trivial, for  $\tau \in (0, 1)$  we have

$$x^\tau = ((\tau, 1 - \tau), (1), (1)), \quad y^\tau = ((1/2, 1/2), (1), (1)).$$

Clearly,  $\gamma(x^\tau, y^\tau) = v$  for all  $\tau \in (0, 1)$ . Note that player 1 has no incentive to deviate from  $x^\tau$  when playing against  $y^\tau$ , because any strategy of player 1 would give reward  $1/2$  against  $y^\tau$ ,  $\tau \in (0, 1)$ . On the other hand, if  $\tau$  is large then player 1 is able to check the action frequencies of player 2 in state 1 with a high precision, thus if the initial state is state 1 then player 1 can make sure that the eventual transitions to state 2 and state 3 will have almost equal probabilities, so, by the choice of a sufficiently large  $\tau$ , player 2 cannot gain more than an arbitrarily small  $\varepsilon$  by any deviation from  $y^\tau$ .

### Example 85

	$L$	$R$		
$T$	1	0		
	1	3		
$B$	0	1	0	0
	2	4	1	2
	1		2	
	0		1	
	3		4	

This example clarifies how stationary strategy pairs with rewards equal to the value can be constructed in ergodic sets in  $\mathcal{R}^3$  (cf. Lemma 82). Notice that states 3 and 4 are trivial. The value of the game is  $v = (0, 0, 0, 1)$ . To see that  $v_1 = v_2 = 0$ , take the stationary strategy  $y^\delta = ((1 - \delta, \delta), (0, 1), (1), (1))$  for player 2, where  $\delta \in (0, 1)$ . One can easily check that  $\gamma_1(i, y^\delta) \leq \delta$  and  $\gamma_2(i, y^\delta) = 0$  for all  $i \in I$ , so using that against a stationary strategy there always exists a pure stationary best reply (cf. lemma 16-(b)) and the fact that the smallest payoff in the game is zero we have  $v_1 = v_2 = 0$  indeed. Note that the strategy  $y^\delta$  is  $\delta$ -optimal for player 2. Following the construction above, we have

$$\begin{aligned} X' &= X, & Y_1' &= \{(1, 0)\}, & Y_s' &= Y_s \quad \forall s = 2, 3, 4, \\ \mathcal{R}^1 &= \emptyset, & S^1 &= \emptyset, & \mathcal{R}^2 &= \emptyset, & S^2 &= \emptyset, \\ \mathcal{R}^3 &= \{\{1, 2\}, \{3\}, \{4\}\}, & S^3 &= \{1, 2, 3, 4\}. \end{aligned}$$

We only focus on the ergodic set  $E = \{1, 2\}$ , as states 3 and 4 are trivial. Consider the restricted game  $\bar{\Gamma}_E$ . Let  $\bar{v}_s$ ,  $s = 1, 2$ , denote the value of  $\bar{\Gamma}_E$ . Clearly,

$$\bar{v}_1 = 1 > 0 = v_1, \quad \bar{v}_2 = 0 = v_2.$$

Note that  $\bar{v}_s \geq v_s$  for all  $s \in E$ , thus, as mentioned in section 6.2, we have that  $\bar{v}_t = v_t$  for some  $t \in E$  (take  $t = 2$  here). Now the strategies

$$\bar{x} = ((1/2, 1/2), (1)) \in \bar{X}_E, \quad \bar{y} = ((1, 0), (0, 1)) \in \bar{Y}_E$$

satisfy  $\gamma_s(\bar{x}, \bar{y}) = v_s$  for all  $s \in E$  (cf. part 1 of the proof of Lemma 82). Now for all  $\tau \in (0, 1)$  we have

$$x^\tau = ((1/2, 1/2), (1), (1), (1)), \quad y^\tau = ((1, 0), (0, 1), (1), (1)).$$

Clearly,  $\gamma(x^\tau, y^\tau) = v$  for all  $\tau \in (0, 1)$ . Note that player 2 has no incentive to deviate from  $y^\tau$  when playing against  $x^\tau$ . On the other hand, player 2 can check the action frequencies of player 1 in state 1. So if player decides to play action  $T$  at each stage, which is the only way for player 1 to get a reward higher than 0 when playing against  $y^\tau$  from initial state 1, then after finitely many stages player 2 detects the deviation of player 1 with probability almost 1 and starts using the  $\delta$ -optimal strategy  $y^\delta$ , where  $\delta$  is small. This assures that player 1 cannot improve more than an arbitrary small  $\varepsilon > 0$ . Note that the probability that player 1 truly uses  $x^\tau$ ,  $\tau \in (0, 1)$ , but accidentally chooses action  $T$  for a very long time is small.

## Part II

# General-sum stochastic games



## Chapter 7

# Recursive repeated games with absorbing states

### 7.1 Introduction

In this chapter, which is based on Flesch et al. [1996], we deal with stochastic games where all the states but one are absorbing, and in the non-absorbing state all the payoffs are equal to zero. Since these games are precisely those games which belong to the classes of recursive games and repeated games with absorbing states (cf. definition 29-(g),(h)), we call them recursive repeated games with absorbing states.

The main result, which will follow from theorem 91, is the following one.

**Main Theorem 7** *In every recursive repeated game with absorbing states, for any  $\varepsilon > 0$ , there exists a stationary  $\varepsilon$ -equilibrium.*

Several examples prove the sharpness of the result. Examples 94 and 28 demonstrate that stationary  $\varepsilon$ -equilibria, for small  $\varepsilon > 0$ , do not necessarily exist in recursive games and in repeated games with absorbing states. Moreover, example 93 shows that recursive repeated games with absorbing states do not always have stationary equilibria. Finally, example 102 clarifies why the above result fails to extend to games of this kind with more than two players.

### 7.2 Preliminaries

Without loss of generality we may assume that the absorbing states are of size  $1 \times 1$  (clearly, in any absorbing state stationary equilibria exist, hence we may replace any absorbing state by another absorbing state with only one cell in which the payoffs equal to the rewards of such a stationary equilibrium of the original state). Suppose that non-absorbing state is state 1 and has size  $m \times n$ . Thus state 1 is the only non-trivial state with action spaces  $I := \{1, \dots, m\}$  and  $J := \{1, \dots, n\}$ . Therefore the

stationary strategy spaces  $X$  and  $Y$  have the form:

$$X = \left\{ x = x(i)_{i \in I} \mid \sum_{i \in I} x(i) = 1, x(i) \geq 0 \quad \forall i \in I \right\},$$

$$Y = \left\{ y = y(j)_{j \in J} \mid \sum_{j \in J} y(j) = 1, y(j) \geq 0 \quad \forall j \in J \right\}.$$

Assume that the initial state is state 1. For the sake of simplicity, we suppress state 1 in the notations, so we write  $\gamma$  and  $p$  instead of  $\gamma_1$  and  $p_1$ . We also assume that there is at least one absorbing state, otherwise the analysis becomes trivial.

If entry  $(i, j)$  of state 1 is chosen, then with probability

$$p_{ij}^* := \sum_{s \in S \setminus \{1\}} p(s|i, j) = 1 - p(1|i, j)$$

absorption occurs in the set of absorbing states with expected absorption payoff  $a_{ij}^k$  for player  $k \in \{1, 2\}$ , and with probability  $1 - p_{ij}^*$  the play stays in the initial state with payoffs zero. For completeness we define  $a_{ij}^k := 0$  if  $p_{ij}^* = 0$ .

**Definition 86** Let  $x \in X$  and  $y \in Y$ . Let

$$p_{xy}^* := \sum_{i \in I} \sum_{j \in J} x(i) y(j) p_{ij}^*,$$

$$T^1(y) := \{i \in I \mid p_{iy}^* > 0\},$$

$$B^1(y) := \{i \in I \mid \gamma(i, y) \geq \gamma(\pi, y) \quad \forall \pi \in \Pi\}.$$

The sets  $T^2(x)$  and  $B^2(x)$  are analogously defined. For the strategy pair  $(x, y)$  we have that  $p_{xy}^*$  is the one step absorption probability. If  $p_{xy}^* > 0$  then  $(x, y)$  eventually leads to absorption, thus we say that  $(x, y)$  is absorbing. If  $p_{xy}^* = 0$  then absorption never occurs with regard to  $(x, y)$ ; in this case we call  $(x, y)$  non-absorbing. Now  $T^1(y)$  consists of the pure stationary strategies (or actions) that are absorbing against  $y$  and  $B^1(y)$  is the set of pure stationary best replies against  $y$ . Similar definitions and interpretations apply for  $T^2(x)$  and  $B^2(x)$ . In view of theorem 16-(b), the sets  $B^1(y)$  and  $B^2(x)$  are always non-empty.

The next lemma provides explicit expressions for the reward when stationary strategies  $x$  and  $y$  are used. If  $(x, y)$  is non-absorbing then the reward is 0, while if  $(x, y)$  is absorbing then the reward equals a convex combination of the rewards for  $(i, y)$ ,  $i \in I$ , where the coefficients are precisely the probabilities that the absorption occurs when player 1 plays action  $i$ ,  $i \in I$ .

**Lemma 87** Let  $(x, y) \in X \times Y$ . Let  $k \in \{1, 2\}$ . If  $p_{xy}^* = 0$  then  $\gamma^k(x, y) = 0$ , while if  $p_{xy}^* > 0$

$$\gamma^k(x, y) = \frac{\sum_{i \in I} \sum_{j \in J} x(i) y(j) a_{ij}^k}{p_{xy}^*}$$

and

$$\gamma^k(x, y) = \frac{\sum_{i \in I} x(i) p_{iy}^* \gamma^k(i, y)}{\sum_{i \in I} x(i) p_{iy}^*} = \frac{\sum_{i \in T^1(y)} x(i) p_{iy}^* \gamma^k(i, y)}{\sum_{i \in T^1(y)} x(i) p_{iy}^*}.$$

**Proof.** Let  $(x, y) \in X \times Y$  and  $k \in \{1, 2\}$ . If  $p_{xy}^* = 0$  then the play remains in state 1 forever with probability 1, so by using that all the payoffs in state 1 equal zero, we obtain  $\gamma^k(x, y) = 0$ .

Assume now that  $p_{xy}^* > 0$ . As  $p_{xy}^* > 0$ , the pair  $(x, y)$  is absorbing, thus  $\gamma^k(x, y)$  equals the expected absorption payoff:

$$\gamma^k(x, y) = \frac{\sum_{i \in I} \sum_{j \in J} x(i) y(j) p_{ij}^* a_{ij}^k}{p_{xy}^*}$$

(one can also show it with the help of lemma 9-(c)). Similarly, if  $p_{iy}^* > 0$  for some  $i \in I$  then

$$\gamma^k(i, y) = \frac{\sum_{j \in J} y(j) p_{ij}^* a_{ij}^k}{p_{iy}^*}.$$

Hence

$$\begin{aligned} \frac{\sum_{i \in I} x(i) p_{iy}^* \gamma^k(i, y)}{\sum_{i \in I} x(i) p_{iy}^*} &= \frac{\sum_{i \in I} x(i) \sum_{j \in J} y(j) p_{ij}^* a_{ij}^k}{\sum_{i \in I} x(i) \sum_{j \in J} y(j) p_{ij}^*} \\ &= \frac{\sum_{i \in I} \sum_{j \in J} x(i) y(j) p_{ij}^* a_{ij}^k}{p_{xy}^*} \\ &= \gamma^k(x, y). \end{aligned}$$

Since  $p_{xy}^* > 0$ , we must have  $T^1(y) \neq \emptyset$ . The definition of  $T^1(y)$  implies

$$\gamma^k(x, y) = \frac{\sum_{i \in T^1(y)} x(i) p_{iy}^* \gamma^k(i, y)}{\sum_{i \in T^1(y)} x(i) p_{iy}^*},$$

so the proof is complete.  $\square$

For our main result we introduce proper and  $\delta$ -proper strategy pairs, where  $\delta \in (0, 1)$ .

**Definition 88** A pair of strategies  $(x_\delta, y_\delta) \in X \times Y$  is called  $\delta$ -proper for  $\delta > 0$ , if

- (a)  $(x_\delta, y_\delta)$  is completely mixed, namely  $x_\delta(i) > 0$  for all  $i \in I$  and  $y_\delta(j) > 0$  for all  $j \in J$ ,
- (b)  $\gamma^1(i, y_\delta) > \gamma^1(i', y_\delta)$  implies  $\delta \cdot x_\delta(i) \geq x_\delta(i')$  for all  $i, i' \in I$ ,
- (c)  $\gamma^2(x_\delta, j) > \gamma^2(x_\delta, j')$  implies  $\delta \cdot y_\delta(j) \geq y_\delta(j')$  for all  $j, j' \in J$ .

A pair of strategies  $(x, y)$  is called proper, if  $(x, y) = \lim_{n \rightarrow \infty} (x_n, y_n)$  for some sequence of  $\delta_n$ -proper strategy pairs  $(x_n, y_n)$ , where  $\delta_n$  is a positive and monotonously decreasing sequence converging to 0.

The notions of proper and  $\delta$ -proper strategy pairs are very similar to those of so-called proper and  $\delta$ -proper equilibria in normal form games (cf. Myerson [1978] and van Damme [1991]). However, here proper and  $\delta$ -proper strategy pairs do not necessarily correspond to ( $\varepsilon$ -)equilibria, for small  $\varepsilon > 0$ , as shown in the following example.

**Example 89**

	$L$	$R$	
$T$	0,0	2,-2	*
$B$	1,-1	0,0	*
		1	

In this game, entry  $(T, L)$  is non-absorbing and all other entries are absorbing with probability 1, indicated by \*, giving the corresponding absorption payoffs to players 1 and 2 respectively (although, formally, all the payoffs in state 1 should equal to 0, it makes no difference for the average reward whether the payoffs in the absorbing cells differ from 0). Here  $((1 - \delta^2, \delta^2), (1 - \delta^2, \delta^2))$  is  $\delta$ -proper for small  $\delta > 0$ , so  $((1, 0), (1, 0))$  is proper, but neither one is an  $\varepsilon$ -equilibrium for small  $\varepsilon > 0$ . Indeed,  $((1, 0), (1, 0))$  is no  $\varepsilon$ -equilibrium for  $\varepsilon \in [0, 1)$ , since player 1 can improve his reward by 1 if he chooses action  $B$ , while  $((1 - \delta^2, \delta^2), (1 - \delta^2, \delta^2))$  is no  $\varepsilon$ -equilibrium for  $\varepsilon \in [0, 1/2]$ , because using lemma 87

$$\begin{aligned} \gamma^1((1 - \delta^2, \delta^2), (1 - \delta^2, \delta^2)) &= \frac{(1 - \delta^2)\delta^2 \cdot 2 + \delta^2 [(1 - \delta^2) \cdot 1 + \delta^2 \cdot 0]}{(1 - \delta^2)\delta^2 + \delta^2} \\ &= \frac{3(1 - \delta^2)}{2 - \delta^2} \\ &< \frac{3}{2}, \end{aligned}$$

while action  $T$  gives 2 against  $(1 - \delta^2, \delta^2)$ .  $\triangleleft$

**Theorem 90** *In any recursive repeated game with absorbing states, there exists a proper strategy pair.*

**Proof.** As  $X \times Y$  is compact, any sequence in  $X \times Y$  must have a convergent subsequence. Therefore it suffices to show that, for  $\delta > 0$  sufficiently small, there exists a  $\delta$ -proper pair. For  $\delta > 0$  let

$$X_\delta := \{x \in X \mid x(i) \geq \delta^m \quad \forall i \in I\}, \quad Y_\delta := \{y \in Y \mid y(j) \geq \delta^n \quad \forall j \in J\}$$

(recall that  $|I| = m$  and  $|J| = n$ ). Consider the following correspondence  $\Psi$  from  $X_\delta \times Y_\delta$  to the set of all subsets of  $X_\delta \times Y_\delta$ :  $\Psi(x, y) = (X_\delta(y), Y_\delta(x))$ , where

$$X_\delta(y) := \{x \in X_\delta \mid \gamma^1(i, y) > \gamma^1(i', y) \text{ implies } \delta \cdot x(i) \geq x(i') \quad \forall i, i' \in I\},$$

$$Y_\delta(x) := \{y \in Y_\delta \mid \gamma^2(x, j) > \gamma^2(x, j') \text{ implies } \delta \cdot y(j) \geq y(j') \quad \forall j, j' \in J\}.$$

Note that, for small  $\delta > 0$ , the sets  $X_\delta$ ,  $Y_\delta$ , and  $X_\delta(y)$ ,  $Y_\delta(x)$  are nonempty. Now  $\Psi$  has a fixed point, since all the conditions of Kakutani's fixed point theorem (cf. Kakutani [1941]) are satisfied. Because every fixed point is a  $\delta$ -proper pair, the proof is complete.  $\square$

### 7.3 The construction

Whenever we deal with a set of strategy pairs  $(x_\delta, y_\delta)$ ,  $\delta > 0$ , in  $X \times Y$ , by the finiteness of the action spaces  $I$  and  $J$ , there must exist a countable subset of  $\mathcal{D}$  of positive real numbers such that 0 is a limit point of  $\mathcal{D}$ ; the sets  $T^1(y_\delta)$ ,  $T^2(x_\delta)$ ,  $B^1(y_\delta)$ , and  $B^2(x_\delta)$  are independent of  $\delta \in \mathcal{D}$ ; and the following sequences have limits as  $\delta$  tends to zero in  $\mathcal{D}$ : (i)  $x_\delta(i^1)/x_\delta(i^2)$  for all  $i^1, i^2 \in I$ , (ii)  $y_\delta(j^1)/y_\delta(j^2)$  for all  $j^1, j^2 \in J$ , (iii)  $(x_\delta, y_\delta)$ . In the sequel each time that we are dealing with limits when  $\delta$  converges to zero, we will have such a subset  $\mathcal{D}$  in mind.

The following theorem provides a construction for stationary  $\varepsilon$ -equilibria for all  $\varepsilon > 0$  in recursive repeated games with absorbing states.

**Theorem 91** *In a recursive repeated game with absorbing states, let  $(\tilde{x}, \tilde{y})$  be a proper pair, where*

$$(\tilde{x}, \tilde{y}) = \lim_{\delta \downarrow 0, \delta \in \mathcal{D}} (x_\delta, y_\delta),$$

where  $(x_\delta, y_\delta)$  is  $\delta$ -proper for all  $\delta \in \mathcal{D}$ . Then for any  $\varepsilon > 0$

- (a) if  $(\tilde{x}, \tilde{y})$  is absorbing, then  $(x_\delta, y_\delta)$  is an  $\varepsilon$ -equilibrium for small  $\delta \in \mathcal{D}$ ,
- (b) if  $(\tilde{x}, \tilde{y})$  is non-absorbing, then  $(\tilde{x}, \tilde{y})$ , or  $(x_\delta, \tilde{y})$ , or  $(\tilde{x}, y_\delta)$  is an  $\varepsilon$ -equilibrium for small  $\delta \in \mathcal{D}$ .

The following example provides an illustration for the above construction.

#### Example 92

	$L$	$R$
$T$	0,0	4,-3 *
$M$	3,-2 *	1,-4 *
$B$	1,-4 *	3,-2 *
	1	

Here entry  $(T, L)$  is non-absorbing while all the other entries lead to absorption with probability 1 (recall that, although formally all the payoffs in state 1 should equal to 0, it makes no difference for the average reward whether the payoffs in the absorbing cells differ from 0).

We will now illustrate the above construction for stationary  $\varepsilon$ -equilibria, where  $\varepsilon > 0$ . Notice that the stationary strategy pair

$$(x_\delta, y_\delta) = ((1 - \delta^2 - \delta^4, \delta^4, \delta^2), (\delta^2, 1 - \delta^2))$$

is  $\delta$ -proper for small  $\delta > 0$ , hence

$$(\tilde{x}, \tilde{y}) = ((1, 0, 0), (0, 1))$$

is proper. Here  $(\tilde{x}, \tilde{y})$  is absorbing, and one can easily check that, for any  $\varepsilon > 0$ , the stationary strategy pair  $(x_\delta, y_\delta)$  is an  $\varepsilon$ -equilibrium for small  $\delta > 0$ . Note that  $(\tilde{x}, \tilde{y})$  is not an  $\varepsilon$ -equilibrium in this game for small  $\varepsilon > 0$ , because player two would be better off by choosing action  $L$  with probability 1.

The pair

$$(x_\delta, y_\delta) = ((1 - \delta^2 - \delta^4, \delta^2, \delta^4), (1 - \delta^2, \delta^2))$$

is also  $\delta$ -proper for small  $\delta > 0$ , so

$$(\tilde{x}, \tilde{y}) = ((1, 0, 0), (1, 0))$$

is proper. Here  $(\tilde{x}, \tilde{y})$  is non-absorbing, and action  $M$  of player 1 is a profitable best reply against  $\tilde{y}$  and leads to absorption in entry  $(M, L)$ . In order to make player 1 satisfied, we let player 1 play  $x_\delta$  with a sufficiently small  $\delta > 0$ . Observe that, for small  $\delta > 0$ , the pair  $(x_\delta, \tilde{y})$  also leads to absorption in entry  $(M, L)$  with probability close to 1, so for any  $\varepsilon > 0$ , the strategy  $x_\delta$  is an  $\varepsilon$ -best reply against  $\tilde{y}$  if  $\delta > 0$  is small. On the other hand,  $\tilde{y}$  is obviously a best reply against  $x_\delta$ , so for any  $\varepsilon > 0$ , the stationary strategy pair  $(x_\delta, \tilde{y})$  is an  $\varepsilon$ -equilibrium for small  $\delta > 0$  indeed.  $\triangleleft$

The next example demonstrates that stationary equilibria need not necessarily exist in recursive repeated games with absorbing states.

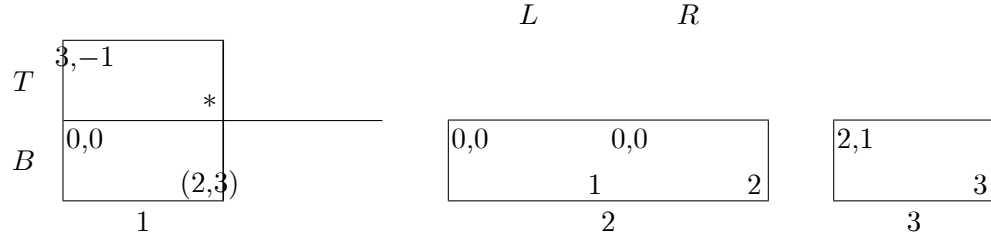
### Example 93

	$L$	$R$	
$T$	0,0	1,-1	*
$B$	1,-1	0,0	*
		1	

We show that there are no stationary equilibria in this game. Assume by way of contradiction that  $(x, y) \in X \times Y$  is an equilibrium. Then if  $y$  puts a positive probability on action  $R$  then  $x$  has to choose action  $T$  with probability 1, which contradicts the fact that  $y$  does not play action  $L$  with probability 1. On the other hand, if  $y$  takes action  $L$  with probability 1, then  $x$  must choose action  $B$  with a positive probability, which is in contradiction with the fact that  $y$  does not choose action  $R$  with probability 1. So we may conclude that no stationary equilibrium exists in this game.  $\triangleleft$

With the help of the following example, we illustrate that Main Theorem 7 does not extend to recursive games (cf. definition 29-(h)).

**Example 94**



Here  $(2, 3)$  stands for the transition vector which brings the play to state 2 with probability  $\frac{1}{2}$  and to state 3 with probability  $\frac{1}{2}$ . Note that we may assume that, if player 1 chooses action  $T$ , then the players receive payoffs zero in state 1 and payoffs  $3, -1$  after absorption; so this game may be viewed as a recursive game without loss of generality. We will now prove that there are no stationary equilibria; one can similarly show that this game does not possess stationary  $\varepsilon$ -equilibria for small  $\varepsilon > 0$  either. We represent stationary strategies of the players by the probabilities on actions  $T$  and  $L$ , respectively. Suppose by way of contradiction that  $(x, y)$  is an equilibrium. If  $y > 0$  then  $x = 0$  must hold, which contradicts  $y > 0$ ; while if  $y = 0$  then we must have  $x = 1$ , which is in contradiction with  $y = 0$ .  $\triangleleft$

**7.4 The proof**

The first lemma deals with stationary  $\varepsilon$ -best replies,  $\varepsilon > 0$ , of the players against stationary strategies of the opponent.

**Lemma 95** *Let  $\varepsilon > 0$ . Let  $(x_\delta, y_\delta) \in X \times Y$  for all  $\delta \in \mathcal{D}$ , and let  $\tilde{y} := \lim_{\delta \downarrow 0, \delta \in \mathcal{D}} y_\delta$ . Suppose that there exists an action  $i^* \in B^1(y_\delta) \cap T^1(\tilde{y})$  such that  $x_\delta(i^*) > 0$ . If*

$$\lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \frac{x_\delta(i)}{x_\delta(i^*)} = 0 \quad \forall i \in T^1(y_\delta) \setminus B^1(y_\delta),$$

*then, for sufficiently small  $\delta \in \mathcal{D}$ , the strategy  $x_\delta$  is an  $\varepsilon$ -best reply against  $y_\delta$ . A similar statement holds for player 2 as well.*

**Proof.** We only show the statement for player 1. Notice that  $i^* \in T^1(\tilde{y})$  implies  $i^* \in T^1(y_\delta)$ . Lemma 87 yields that for sufficiently small  $\delta \in \mathcal{D}$  we have

$$\begin{aligned}
\gamma^1(x_\delta, y_\delta) &= \frac{\sum_{i \in T^1(y_\delta)} x_\delta(i) p_{iy_\delta}^* \gamma^1(i, y_\delta)}{\sum_{i \in T^1(y_\delta)} x_\delta(i) p_{iy_\delta}^*} \\
&= \frac{\sum_{i \in T^1(y_\delta) \cap B^1(y_\delta)} x_\delta(i) p_{iy_\delta}^* \gamma^1(i, y_\delta) + \sum_{i \in T^1(y_\delta) \setminus B^1(y_\delta)} x_\delta(i) p_{iy_\delta}^* \gamma^1(i, y_\delta)}{\sum_{i \in T^1(y_\delta) \cap B^1(y_\delta)} x_\delta(i) p_{iy_\delta}^* + \sum_{i \in T^1(y_\delta) \setminus B^1(y_\delta)} x_\delta(i) p_{iy_\delta}^*} \\
&= \frac{\sum_{i \in T^1(y_\delta) \cap B^1(y_\delta)} \frac{x_\delta(i)}{x_\delta(i^*)} p_{iy_\delta}^* \gamma^1(i, y_\delta) + \sum_{i \in T^1(y_\delta) \setminus B^1(y_\delta)} \frac{x_\delta(i)}{x_\delta(i^*)} p_{iy_\delta}^* \gamma^1(i, y_\delta)}{\sum_{i \in T^1(y_\delta) \cap B^1(y_\delta)} \frac{x_\delta(i)}{x_\delta(i^*)} p_{iy_\delta}^* + \sum_{i \in T^1(y_\delta) \setminus B^1(y_\delta)} \frac{x_\delta(i)}{x_\delta(i^*)} p_{iy_\delta}^*} \\
&\geq \frac{\sum_{i \in T^1(y_\delta) \cap B^1(y_\delta)} \frac{x_\delta(i)}{x_\delta(i^*)} p_{iy_\delta}^* \gamma^1(i, y_\delta)}{\sum_{i \in T^1(y_\delta) \cap B^1(y_\delta)} \frac{x_\delta(i)}{x_\delta(i^*)} p_{iy_\delta}^*} - \varepsilon \\
&= \gamma^1(i^*, y_\delta) - \varepsilon.
\end{aligned}$$

As  $i^* \in B^1(y_\delta)$ , the proof is complete.  $\square$

Now we are ready to prove theorem 91.

**Proof of theorem 91.**

(a) Let  $\varepsilon > 0$ . Since  $(\tilde{x}, \tilde{y})$  is absorbing, the  $\delta$ -properness of  $(x_\delta, y_\delta)$  implies the existence of  $i^* \in B^1(y_\delta) \cap T^1(\tilde{y})$ . By the  $\delta$ -properness of  $(x_\delta, y_\delta)$

$$\lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \frac{x_\delta(i)}{x_\delta(i^*)} = 0 \quad \forall i \in T^1(y_\delta) \setminus B^1(y_\delta),$$

so the conditions of lemma 95 are fulfilled and therefore  $x_\delta$  is an  $\varepsilon$ -best reply against  $y_\delta$  for sufficiently small  $\delta \in \mathcal{D}$ . One can similarly show that  $y_\delta$  is also an  $\varepsilon$ -best reply against  $x_\delta$  for sufficiently small  $\delta \in \mathcal{D}$ . Therefore  $(x_\delta, y_\delta)$  is an  $\varepsilon$ -equilibrium on condition that  $\delta \in \mathcal{D}$  is sufficiently small.

(b) Let  $\varepsilon > 0$  and assume that  $(\tilde{x}, \tilde{y})$  is non-absorbing. If  $(\tilde{x}, \tilde{y})$  is an equilibrium, then we are done. Otherwise, at least one of the players has a profitable deviation with respect to  $(\tilde{x}, \tilde{y})$ . Without loss of generality suppose that player 1 has a profitable deviation. Then we show that  $(x_\delta, \tilde{y})$  must be an  $\varepsilon$ -equilibrium for sufficiently small  $\delta \in \mathcal{D}$ .

Let  $i^* \in B^1(\tilde{y})$  be a profitable best reply of player 1 against  $\tilde{y}$ . Then, since  $(\tilde{x}, \tilde{y})$  is non-absorbing and  $i^*$  is a profitable deviation, we must have  $i^* \in T^1(\tilde{y})$ . Since  $i^* \in T^1(\tilde{y})$ , we also have  $i^* \in T^1(y_\delta)$ . Assume  $i \in T^1(\tilde{y}) \setminus B^1(\tilde{y})$ . Then  $i \in T^1(y_\delta)$  as well and by lemma 10 (or by lemma 87)

$$\lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \gamma^1(i^*, y_\delta) = \gamma^1(i^*, \tilde{y}) > \gamma^1(i, \tilde{y}) = \lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \gamma^1(i, y_\delta).$$

Hence  $\gamma^1(i^*, y_\delta) > \gamma^1(i, y_\delta)$  for sufficiently small  $\delta \in \mathcal{D}$ . By the  $\delta$ -properness of  $(x_\delta, y_\delta)$  we now have

$$\lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \frac{x_\delta(i)}{x_\delta(i^*)} = 0$$

and, by lemma 95 with  $(x_\delta, \tilde{y})$  instead of  $(x_\delta, y_\delta)$ , we obtain that  $x_\delta$  is an  $\varepsilon$ -best reply against  $\tilde{y}$  for small  $\delta \in \mathcal{D}$ .

On the other hand,  $\tilde{y}$  is a best reply against  $x_\delta$ . One can argue as follows. Let  $j^* \in J$  satisfy  $\tilde{y}(j^*) > 0$ . Then

$$\lim_{\delta \downarrow 0, \delta \in \mathcal{D}} \frac{y_\delta(j^*)}{y_\delta(j)} > 0 \quad \text{for all } j \in J, \quad (7.1)$$

so, by the  $\delta$ -properness of  $(x_\delta, y_\delta)$ , we must have  $\gamma^2(x_\delta, j^*) \geq \gamma^2(x_\delta, j)$  for all  $j \in J$  and for small  $\delta \in \mathcal{D}$ . Hence  $j^* \in B^2(x_\delta)$  for all  $j^* \in J$  with  $\tilde{y}(j^*) > 0$ . This yields in view of lemma 87 that the stationary strategy  $\tilde{y}$  is also a best reply against  $x_\delta$ . Therefore we may conclude that  $(x_\delta, \tilde{y})$  is an  $\varepsilon$ -equilibrium, whenever  $\delta \in \mathcal{D}$  is sufficiently small.  $\square$

## 7.5 Concluding remarks

There is another way to establish equilibria having similar properties as  $\delta$ -proper pairs by defining the following restricted strategy spaces for  $\delta > 0$

$$\bar{X}_\delta := \left\{ x \in X \mid \sum_{i \in U} x(i) \geq \delta^{m-|U|} \quad \forall \emptyset \neq U \subset I \right\},$$

$$\bar{Y}_\delta := \left\{ y \in Y \mid \sum_{j \in V} y(j) \geq \delta^{n-|V|} \quad \forall \emptyset \neq V \subset J \right\},$$

and by defining “linearized” rewards

$$\bar{\gamma}^1(x, y) := \sum_{i \in I} x(i) \gamma^1(i, y), \quad \bar{\gamma}^2(x, y) := \sum_{j \in J} y(j) \gamma^2(x, j).$$

By applying Kakutani’s fixed point theorem, one can show the existence of stationary equilibria  $(\bar{x}_\delta, \bar{y}_\delta)$  in  $\bar{X}_\delta \times \bar{Y}_\delta$  with respect to the rewards  $(\bar{\gamma}^1, \bar{\gamma}^2)$ . Such equilibria have similar properties as  $\delta$ -proper pairs, and the existence of average  $\varepsilon$ -equilibria can be established analogously.

Notice that a stationary equilibrium  $(z_\delta, w_\delta)$  would also exist in  $\bar{X}_\delta \times \bar{Y}_\delta$  with respect to the original rewards  $(\gamma^1, \gamma^2)$ , but for such an equilibrium  $\gamma^1(i, w_\delta) > \gamma^1(i', w_\delta)$  would not necessarily imply  $\delta \cdot z_\delta(i) \geq z_\delta(i')$ . This causes a discontinuity in the best reply structures when approaching  $X \times Y$  by  $\bar{X}_\delta \times \bar{Y}_\delta$ .



# Chapter 8

## Average-discounted equilibria

### 8.1 Introduction

In this chapter, which is based on Flesch et al. [1998,III], we investigate existence of equilibria in stochastic games where the players use different evaluations. We assume that player 1 uses the average reward, while player 2 is interested in his  $\beta$ -discounted reward,  $\beta \in (0, 1)$ . We will call these games average-discounted games. By the nature of these rewards, the players are interested in different time periods of the play, which may lead to a natural cooperation between them. First we define what we mean by equilibria in these games.

**Definition 96** *A strategy pair  $(\pi, \sigma)$  is called an average- $\beta$ -discounted  $\varepsilon$ -equilibrium, where  $\varepsilon \geq 0$  and  $\beta \in (0, 1)$ , if for all  $s \in S$ ,  $\bar{\pi} \in \Pi$ ,  $\bar{\sigma} \in \Sigma$ , we have*

$$\gamma_s^1(\bar{\pi}, \sigma) \leq \gamma_s^1(\pi, \sigma) + \varepsilon \quad \text{and} \quad \gamma_{\beta s}^2(\pi, \bar{\sigma}) \leq \gamma_{\beta s}^2(\pi, \sigma) + \varepsilon.$$

The main results of this chapter, which will follow from theorems 99 and 100, can be summarized as follows.

**Main Theorem 8**

- (a) In any stochastic game, for any  $\varepsilon > 0$  and  $\beta \in (0, 1)$ , there exists a stationary average- $\beta$ -discounted  $\varepsilon$ -equilibrium.
- (b) In any stochastic game, for any  $\varepsilon > 0$  and  $\beta \in (0, 1)$ , there exist average- $\beta$ -discounted  $\varepsilon$ -equilibria such that up to some stage  $N$  the players play Markov strategies, from stage  $N + 1$  they play stationary strategies, and if the play is at stage  $N + 1$  in state  $s$ , then player 1 receives  $\rho_s := \sup_{\pi \in \Pi, \sigma \in \Sigma} \gamma_s^1(\pi, \sigma)$ .

In view of (a), stationary strategies are sufficient for establishing  $\varepsilon$ -equilibria,  $\varepsilon > 0$ , in these average-discounted games; recall that, in classical discounted games, stationary strategies are even sufficient to obtain 0-equilibria, as stated in theorem 26. The existence of  $\varepsilon$ -equilibria in terms of stationary strategies is appealing, since stationary strategies are rather simple strategies. On the other hand, however, these stationary  $\varepsilon$ -equilibria have the draw-back that they do not make use of the special nature of these games, namely, they do not use that different time periods interest the players.

Therefore, in (b), we also prove the existence of  $\varepsilon$ -equilibria, where, after a large stage when the discounted game is not interesting any longer, the players cooperate to guarantee the highest feasible reward to player 1. These  $\varepsilon$ -equilibria are formed by only slightly more complex Markov strategies, which we will call “ultimately stationary” (after finitely many stages stationary strategies are played forever).

Example 101 will demonstrate that average- $\beta$ -discounted 0-equilibria do not always exist, not even in terms of history dependent strategies, so the result is sharp. Finally, we examine the existence of average-discounted  $\varepsilon$ -equilibria,  $\varepsilon > 0$ , in special classes of stochastic games.

We now briefly discuss the following game to clarify the issues.

**Example 97**

	<i>L</i>	<i>R</i>
<i>T</i>	0,0 *	1,2
<i>B</i>	2,1 *	0,0 *
	1	

Take an arbitrary discount factor  $\beta \in (0, 1)$ . There are two really simple stationary average- $\beta$ -discounted equilibria. One of them is playing entry  $(B, L)$  at stage 1, yielding absorption in state 2 and reward  $(2, 1)$ , while the other one is playing entry  $(T, R)$  at each stage, which gives reward  $(1, 2)$ . These stationary equilibria, however, are not really in the spirit of the game. The players could also decide to play entry  $(T, R)$  sufficiently long so that player 2’s reward, which is rather determined by the near future payoffs, becomes almost 2, and then, when the rest of the play does not really interest player 2 any longer, to play entry  $(B, L)$  so as to give player 1 his highest feasible payoff (namely payoff 2) at each further stages. This plan, yielding a reward close to  $(2, 2)$ , can be realized by ultimately stationary strategies (after finitely many stages stationary strategies are played forever). Note that rewards close to  $(2, 2)$  cannot be guaranteed by stationary  $\varepsilon$ -equilibria, with small  $\varepsilon \geq 0$ .

## 8.2 Stationary $\varepsilon$ -equilibria

This section is devoted to the analysis of the existence of stationary  $\varepsilon$ -equilibria,  $\varepsilon > 0$ , in these average-discounted games. First we introduce a restricted strategy space for player 2. Let

$$\bar{\delta} := \min_{s \in S} \frac{1}{|J_s|}.$$

For  $\delta \in [0, \bar{\delta}]$  let

$$Y(\delta) := \{y \in Y \mid y_s(j_s) \geq \delta \quad \forall s \in S, \forall j_s \in J_s\};$$

in words,  $Y(\delta)$  is the set of stationary strategies of player 2 which use each action in each state with probability at least  $\delta$ . Obviously,  $Y(\delta)$  is a polytope, and by the

choice of  $\bar{\delta}$  it is nonempty. The following lemma summarizes some properties of the rewards and sets of best replies.

**Lemma 98**

(a) The function  $\gamma_s^1(x, \cdot)$  is continuous on  $Y(\delta)$  for any  $s \in S$ ,  $x \in X$ ,  $\delta \in (0, \bar{\delta}]$ .

(b) Let  $\bar{y} \in Y$ . Then the set

$$\mathcal{B}^1(\bar{y}) := \{x \in X \mid \gamma_s^1(x, \bar{y}) \geq \gamma_s^1(\hat{x}, \bar{y}) \quad \forall s \in S, \forall \hat{x} \in X\}$$

is nonempty and convex.

(c) The function  $\gamma_{\beta s}^2(\cdot, \cdot)$  is continuous on  $X \times Y$  for any  $s \in S$ ,  $\beta \in (0, 1)$ .

(d) Let  $\beta \in (0, 1)$ ,  $\bar{x} \in X$ ,  $\delta \in [0, \bar{\delta}]$ . Then the set

$$\mathcal{B}_\beta^2(\delta, \bar{x}) := \{y \in Y(\delta) \mid \gamma_{\beta s}^2(\bar{x}, y) \geq \gamma_{\beta s}^2(\bar{x}, \hat{y}) \quad \forall s \in S, \forall \hat{y} \in Y(\delta)\}$$

is nonempty, convex, and closed.

**Proof.** Property (a) follows from lemma 10, (c) is the same as lemma 13 #.  $\square$

Notice that the set  $\mathcal{B}^1(\bar{y})$  is not necessarily closed, which was clarified for instance by example 11.

The main result of this section is the following theorem.

**Theorem 99** *In any stochastic game, for any  $\varepsilon > 0$  and  $\beta \in (0, 1)$ , there exists a stationary average- $\beta$ -discounted  $\varepsilon$ -equilibrium.*

**Proof.** Take arbitrary  $\varepsilon > 0$  and  $\beta \in (0, 1)$ . For a strategy  $y \in Y$  let  $\bar{\mathcal{B}}^1(y)$  denote the closure of  $\mathcal{B}^1(y)$ . By lemma 98-(c), the function  $\gamma_{\beta s}^2(\cdot, \cdot)$  is continuous on the compact space  $X \times Y$ , for any  $s \in S$ , hence it is uniformly continuous as well. Therefore there exists a  $\delta \in (0, \bar{\delta}]$  such that for all  $s \in S$  we have

$$\sup_{x \in X} \left[ \sup_{y \in Y} \gamma_{\beta s}^2(x, y) - \sup_{y \in Y(\delta)} \gamma_{\beta s}^2(x, y) \right] \leq \frac{\varepsilon}{2}. \quad (8.1)$$

Now consider the following set-valued map:

$$\Psi : (x, y) \in X \times Y(\delta) \longmapsto \bar{\mathcal{B}}^1(y) \times \mathcal{B}_\beta^2(\delta, x) \subset X \times Y(\delta).$$

In view of lemma 98, this correspondence  $\Psi$  satisfies the conditions of Kakutani's fixed point theorem (cf. Kakutani [1941]). Therefore  $\Psi$  has a fixed point, namely, there exists a pair  $(x, y) \in X \times Y(\delta)$  such that  $(x, y) \in (\bar{\mathcal{B}}^1(y), \mathcal{B}_\beta^2(\delta, x))$ .

By using this fixed point  $(x, y)$ , we construct a stationary average- $\beta$ -discounted  $\varepsilon$ -equilibrium in the game. Since  $x \in \bar{\mathcal{B}}^1(y)$  and  $y \in \mathcal{B}_\beta^2(\delta, x)$ , by the uniform continuity

of  $\gamma_{\beta_s}^2(\cdot, \cdot)$  on  $X \times Y$  for all  $s \in S$ , there exists an  $x' \in \mathcal{B}^1(y)$  such that, for all  $s \in S$ , all the following inequalities (and equality) hold:

$$\gamma_{\beta_s}^2(x', y) + \frac{\varepsilon}{2} \geq \gamma_{\beta_s}^2(x, y) + \frac{\varepsilon}{4} = \sup_{\bar{y} \in Y(\delta)} \gamma_{\beta_s}^2(x, \bar{y}) + \frac{\varepsilon}{4} \geq \sup_{\bar{y} \in Y(\delta)} \gamma_{\beta_s}^2(x', \bar{y}). \quad (8.2)$$

We show that  $(x', y)$  is an average- $\beta$ -discounted  $\varepsilon$ -equilibrium. Using theorem 16-(b) and  $x' \in \mathcal{B}^1(y)$ , we have for all  $s \in S$  that

$$\gamma_s^1(\pi, y) \leq \sup_{\bar{x} \in X} \gamma_s^1(\bar{x}, y) = \gamma_s^1(x', y) \quad \forall \pi \in \Pi.$$

For player 2, applying (8.1) and (8.2), we obtain for all  $s \in S$  that

$$\gamma_{\beta_s}^2(x', \sigma) \leq \sup_{\bar{y} \in Y} \gamma_{\beta_s}^2(x', \bar{y}) \leq \sup_{\bar{y} \in Y(\delta)} \gamma_{\beta_s}^2(x', \bar{y}) + \frac{\varepsilon}{2} \leq \gamma_{\beta_s}^2(x', y) + \varepsilon \quad \forall \sigma \in \Sigma.$$

Therefore  $(x', y)$  is an average- $\beta$ -discounted  $\varepsilon$ -equilibrium indeed.  $\square$

### 8.3 Ultimately stationary $\varepsilon$ -equilibria

In the previous section we showed the existence of stationary average- $\beta$ -discounted  $\varepsilon$ -equilibria for all  $\varepsilon > 0$  and  $\beta \in (0, 1)$ . These stationary  $\varepsilon$ -equilibria are appealing, because simple strategies are used. In this section, however, we also prove the existence of  $\varepsilon$ -equilibria in terms of ultimately stationary strategies (after finitely many stages stationary strategies are played forever), where the players naturally cooperate, by making use of the different nature of their rewards. The idea is that, after a large stage  $N$ , player 2 becomes uninterested in the game due to the large powers of the discount factor  $\beta$ , so after stage  $N$  the players can cooperate to guarantee the highest feasible reward for player 1 in the future. During the first  $N$  stages, obviously, player 1 has to be careful not to ruin his future perspectives after stage  $N$ .

**Theorem 100** *In any stochastic game, for any  $\varepsilon > 0$  and  $\beta \in (0, 1)$ , there exist average- $\beta$ -discounted  $\varepsilon$ -equilibria such that up to some stage  $N$  the players play Markov strategies, from stage  $N + 1$  they play stationary strategies, and if the play is at stage  $N + 1$  in state  $s$ , then player 1 receives  $\rho_s := \sup_{\pi \in \Pi, \sigma \in \Sigma} \gamma_s^1(\pi, \sigma)$ .*

**Proof.** Consider a stochastic game  $\Gamma$ . Take arbitrary  $\varepsilon > 0$  and  $\beta \in (0, 1)$ . Let  $N \in \mathbb{N}$  be so large that

$$\beta^N \cdot \left[ \max_{s, i_s, j_s} r_s^2(i_s, j_s) - \min_{s, i_s, j_s} r_s^2(i_s, j_s) \right] \leq \varepsilon,$$

so after stage  $N$  player 2 can only improve his  $\beta$ -discounted reward by at most  $\varepsilon$ . It is known that there exists a pure stationary strategy pair  $(i^*, j^*) \in I \times J$  such that for  $\rho_s$ , as defined in the theorem, it holds that

$$\rho_s = \gamma_s^1(i^*, j^*) \quad \forall s \in S.$$

(The existence of such a pure stationary strategy pair  $(i^*, j^*) \in I \times J$  stems from the theory of Markov decision processes. In fact, such pairs  $(i^*, j^*) \in I \times J$  are exactly

the pure optimal solutions of the Markov decision process which is derived from the stochastic game by assuming that there is only one player with action space  $I_s \times J_s$ , payoff function  $r_s^1$ , and transition map  $p_s$  in states  $s \in S$ .)

Consider the game  $\Gamma^N$  which is played up to stage  $N$  and in which player 1 maximizes the expected value of  $\rho_{s^{N+1}}$  on condition the play up to stage  $N$  (here  $s^{N+1}$  denotes the random variable for the state at stage  $N + 1$ ) and player 2 maximizes his  $N$ -stage  $\beta$ -discounted reward. Using backwards induction, one can construct an  $N$ -stage Markov average- $\beta$ -discounted 0-equilibrium  $(f^N, g^N)$  in the game  $\Gamma^N$ .

Let  $f$  denote the Markov strategy which coincides with  $f^N$  for the first  $N$  stages and which prescribes the pure stationary strategy  $i^*$  afterwards. The definition of  $g$  is analogous. Thus by their definitions,  $f$  and  $g$  satisfy the requirements of the theorem. We only have to show that  $(f, g)$  is an  $\varepsilon$ -equilibrium. Observe that player 1's average reward  $\gamma^1$  is completely determined by the value of  $\rho$  in the state at stage  $N + 1$ , which is exactly what he maximizes during the first  $N$  stages, so player 1 cannot improve at all. Player 2 can only improve his reward by  $\varepsilon$  after stage  $N$ , because of the choice of  $N$ ; while during the first  $N$  stages, by his reward function in the game  $\Gamma^N$ , he cannot improve it at all. So  $(f, g)$  is an average- $\beta$ -discounted  $\varepsilon$ -equilibrium indeed.  $\square$

### 8.4 A game without average-discounted 0-equilibria

In the previous two sections we showed the existence of  $\varepsilon$ -equilibria, for all  $\varepsilon > 0$ , in terms of stationary and ultimately stationary strategies. The following interesting example will demonstrate that, in these average-discounted games, 0-equilibria do not always exist, not even in history dependent strategies. So as it might be expected, the solutions of average-discounted games are on the one hand more complex than that of discounted games, where stationary 0-equilibria always exist as mentioned above, but on the other hand simpler than that of average games, where stationary  $\varepsilon$ -equilibria do not generally exist for small  $\varepsilon \geq 0$ .

**Example 101**

	$L$	$R$
$T$	$1, -1$	$-1, 1$
$B$	$-1, 1$	$0, 0$
	$*$	$*$
	$*$	$1$
	$1$	

We show that in the above game there are no 0-equilibria for initial state 1 with respect to  $(\gamma^1, \gamma_\beta^2)$  for any  $\beta \in (0, 1)$ .

Notice that strategies only need to be defined for histories where the initial state is 1 and no absorption has occurred. As the only information carried by these histories is the current stage, all history dependent strategies are simply Markov strategies. Since any mixed action in state 1 can be represented by the probability assigned to the first action, any Markov strategy for any player is an element of the set  $\times_{n=1}^\infty [0, 1]$ .

Suppose by way of contradiction that  $(f, g) = (f(n), g(n))_{n=1}^{\infty}$  is a Markov 0-equilibrium with respect to  $(\gamma^1, \gamma^2_{\beta})$ , where  $\beta \in (0, 1)$ ; here  $f(n)$  and  $g(n)$  denote the probabilities of playing action  $T$  and  $L$ , respectively, at stage  $n$ . Let  $f^k := (f(n))_{n=k}^{\infty}$  and  $g^k := (g(n))_{n=k}^{\infty}$  for any  $k \in \mathbb{N}$ , so  $f^k$  and  $g^k$  are the Markov strategies  $f$  and  $g$  starting from stage  $k$ . Let  $\xi^k$  denote player 1's limiting average reward when using  $(f^k, g^k)$  for initial state 1.

Based on the assumption that  $(f, g)$  is a 0-equilibrium, we subsequently derive that we should have

- (1)  $\xi^1 > -1$ ;
- (2)  $0 < f(1) < 1$  and  $0 < g(1) < 1$ ;
- (3)  $(f^n, g^n)$  is a 0-equilibrium,  $0 < f(n) < 1$ , and  $0 < g(n) < 1$  for all  $n \in \mathbb{N}$ ;
- (4)  $\xi^n < \xi^{n+1}$  and  $g(n) < g(n+1)$  for all  $n \in \mathbb{N}$ .

Next we show that these properties lead to a contradiction.

**Proof of (1).** Since  $(f, g)$  is a 0-equilibrium, it suffices to define a strategy  $\bar{f}$  for player 1 which guarantees a reward larger than  $-1$  when playing against  $g$ . For  $n \in \mathbb{N}$  let

$$\bar{f}(n) := \begin{cases} 1 & \text{if } g(n) > 0 \\ 0 & \text{if } g(n) = 0. \end{cases}$$

Now with respect to  $(\bar{f}, g)$ , whenever the play is in state 1, either the cell  $(B, R)$  is played with probability 1 or the cell  $(T, L)$  is played with a positive probability, hence  $\gamma^1(1, \bar{f}, g) > -1$ .

**Proof of (2).** If  $f(1) = 1$  then  $g(1) = 0$ , since it yields absorption in entry  $(T, R)$  giving the highest possible reward 1 for player 2. However, this contradicts  $\xi^1 > -1$  (cf. (1)), hence  $f(1) < 1$  must hold. If  $f(1) = 0$  then  $g(1) = 1$ , which also contradicts  $\xi^1 > -1$ ; hence  $f(1) > 0$ .

If  $g(1) = 1$  then  $f(1) = 1$  has to hold because  $f$  is a best reply against  $g$ , which contradicts  $0 < f(1) < 1$ . Hence  $g(1) < 1$ . Now suppose that  $g(1) = 0$ . Using (1) we have

$$-1 < \xi^1 = f(1) \cdot (-1) + (1 - f(1)) \cdot \xi^2,$$

thus by  $f(1) > 0$  we obtain  $\xi^2 > \xi^1$ , which means that player 1 would be better off by playing action  $B$  at stage 1 and playing  $f^2$  from stage 2 on assuring reward  $\xi^2$ . This is in contradiction with the fact that  $f(1) > 0$ . Hence  $g(1) > 0$  must hold.

**Proof of (3).** By (2), the probability of no absorption at stage 1 has a positive probability, therefore, clearly,  $(f^2, g^2)$  must be a 0-equilibrium as well. Using that  $(f^2, g^2)$  is a 0-equilibrium, one can show similarly that  $0 < f(2) < 1$  and  $0 < g(2) < 1$ . Now repeating this argument yields the statement.

**Proof of (4).** The strategy  $f^1$  is a best reply against  $g^1$  and player 1 plays action  $B$  with a positive probability at stage 1 (cf. (3)), hence

$$\xi^1 = g(1) \cdot (-1) + (1 - g(1)) \cdot \xi^2.$$

Now using (1) and  $g(1) > 0$  (cf. (3)), we have  $\xi^1 < \xi^2$  indeed. Repeating this argument leads to  $\xi^n < \xi^{n+1}$  for all  $n \in \mathbb{N}$ .

At stages  $n$  and  $n+1$ , in view of (3), player 1 plays action  $T$  with positive probabilities, thus

$$\xi^n = g(n) \cdot 1 + (1 - g(n)) \cdot (-1), \quad \xi^{n+1} = g(n+1) \cdot 1 + (1 - g(n+1)) \cdot (-1).$$

Now from  $\xi^n < \xi^{n+1}$  it follows that  $g(n) < g(n+1)$ .

**Deriving a contradiction.** Consider the strategy  $\bar{f}_K$ ,  $K \geq 2$ , which prescribes action  $B$  up to stage  $K-1$  and the strategy  $f^K$  from stage  $K$  on. Then with respect to  $(\bar{f}_K, g)$ , player 1's reward is  $-1$  if absorption occurs during the first  $K-1$  stages and equals  $\xi^K$  otherwise. Thus we have for all  $K \geq 2$  that

$$\begin{aligned} \gamma^1(\bar{f}_K, g) &= \left[ 1 - \prod_{n=1}^{K-1} (1 - g(n)) \right] \cdot (-1) + \left[ \prod_{n=1}^{K-1} (1 - g(n)) \right] \cdot \xi^K \\ &= \left[ 1 - \prod_{n=1}^{K-2} (1 - g(n)) \right] \cdot (-1) + \left[ \prod_{n=1}^{K-2} (1 - g(n)) \right] \cdot [g(K-1) \cdot (-1) + (1 - g(K-1))] \\ &= \left[ 1 - \prod_{n=1}^{K-2} (1 - g(n)) \right] \cdot (-1) + \left[ \prod_{n=1}^{K-2} (1 - g(n)) \right] \cdot \xi^{K-1} \\ &= \dots \\ &= g(1) \cdot (-1) + (1 - g(1)) \cdot \xi^2 \\ &= \xi^1. \end{aligned}$$

However, by properties 2 and 4 we have that  $g(n) > g(1) > 0$  for all  $n \in \mathbb{N}$ . Therefore, if player 1 uses  $\bar{f}_K$  with a large  $K$  then absorption occurs in entry  $(B, L)$  during the first  $K-1$  stages with probability almost 1. Formally,

$$\lim_{K \rightarrow \infty} \left[ 1 - \prod_{n=1}^{K-1} (1 - g(n)) \right] = 1,$$

thus

$$\xi^1 = \lim_{K \rightarrow \infty} \gamma^1(\bar{f}_K, g) = -1,$$

which contradicts (1). Hence the basic assumption that  $(f, g)$  is a 0-equilibrium is false.  $\square$

## 8.5 Special classes of stochastic games

This section is devoted to the study of average-discounted equilibria in special classes of games. We briefly treat several classes of games in which  $(\varepsilon)$ -equilibria can be achieved by using other techniques. Recall definition 29.

*Unichain games.* In unichain stochastic games there is just one ergodic set of states. This condition assures that the average reward  $\gamma_s^1(\cdot, \cdot)$  is also continuous on  $X \times Y$  for

all  $s \in S$  and that the best reply sets  $\mathcal{B}^1(\bar{y})$ ,  $\bar{y} \in Y$ , are closed (cf. lemma 98-(a),(b)). In these games one can establish stationary average- $\beta$ -discounted 0-equilibria, for any  $\beta \in (0, 1)$ , by simply applying Kakutani's fixed point theorem on  $X \times Y$  (cf. Kakutani [1941]).

*Perfect information, switching control and ARAT games.* In perfect information games and ARAT games one can establish average-discounted 0-equilibria, almost analogously as in the proof for the existence of average 0-equilibria for these games in Thuijsman & Raghavan [1997]. The idea is that player 1 has to play a pure stationary average optimal strategy  $i$ , namely  $\inf_{\sigma \in \Sigma} \gamma_s^1(i, \sigma) = \sup_{\pi \in \Pi} \inf_{\sigma \in \Sigma} \gamma_s^1(\pi, \sigma)$  for all  $s \in S$  (cf. theorem 30-(b),(d)), and player 2 has to play a stationary  $\beta$ -discounted best reply  $y$  against the strategy  $i$ . This already implies that player 2 does not have a profitable deviation against  $i$ . Notice that, since the strategy  $i$  prescribes one pure action for each state, player 2 can immediately detect any deviation of player 1. Now in order to eliminate the profitability of deviations of player 1, if player 2 detects a deviation from  $i$  then he has to punish player 1 by switching to a strategy  $\sigma$  satisfying  $\gamma_s^1(\pi, \sigma) \leq \gamma_s^1(i, y) + \delta$  for all  $s$  and  $\pi$ , where  $\delta > 0$  is sufficiently small. Note that these punishments are effective due to the transition structure of these games.

In switching control stochastic games, the proof is somewhat more complicated, because player 1 does not need to have pure stationary average optimal strategies. Nevertheless, by lemma 30-(c) there exist stationary average optimal strategies for player 1. Now the main difference is that player 2 cannot immediately detect deviations of player 1, but, as shown in Thuijsman & Raghavan [1997], player 2 can conduct statistical tests on the action frequencies of player 1, and by doing so he can detect deviations in the long run with probability almost 1. This way we obtain average- $\beta$ -discounted  $\varepsilon$ -equilibria, for all  $\varepsilon > 0$  and  $\beta \in (0, 1)$ , for switching control games as well.

*Repeated games with absorbing states.* Here one can establish average- $\beta$ -discounted  $\varepsilon$ -equilibria, for all  $\varepsilon > 0$  and  $\beta \in (0, 1)$ , as follows. As stated in theorem 26, for any  $\alpha \in (0, 1)$ , there exists a stationary equilibrium  $(x^{\alpha\beta}, y^{\alpha\beta})$  with respect to  $(\gamma_\alpha^1, \gamma_\beta^2)$ . Using techniques as in Vrieze & Thuijsman [1989] one can show that either  $(x^{1\beta}, y^{1\beta})$  or  $(x^{1\beta}, y^{\alpha\beta})$  with a large  $\alpha$  can be supplemented with history dependent "punishment" strategies to establish an  $\varepsilon$ -equilibrium with respect to  $(\gamma^1, \gamma_\beta^2)$ ; here  $(x^{1\beta}, y^{1\beta})$  is the limit strategy pair of some sequence  $(x^{\alpha_n\beta}, y^{\alpha_n\beta})$ ,  $n \in \mathbb{N}$ , with  $\alpha_n \uparrow 1$ .

## 8.6 Concluding remarks

We wish to remark that, in the literature of stochastic games and Markov decision processes, games have already been studied where, instead of using the discounted or the average evaluation, the players (or the player) use convex combinations of several discounted rewards with different discount factors and the average reward (cf. for example Filar & Vrieze [1992], Feinberg & Shwartz [1994], Feinberg & Shwartz [1995]). Although the ideas have something in common, the arising problems require a different analysis.

# Chapter 9

## More than two players

### 9.1 Introduction

The model and the solution concepts of two-person general-sum stochastic games naturally extend to stochastic games with more than two players. In two-person stochastic games, many of the usual techniques for establishing equilibria are based on sequences of stationary equilibria in auxiliary games, where either the strategy spaces are specifically restricted or the reward functions are approximated by continuous functions (for example discounted rewards), that approach the original game in a certain sense. However, the analysis of  $K$ -person stochastic games,  $K \geq 3$ , involves several difficulties which do not appear in the case of only two players.

In this chapter, which is based on Flesch et al. [1997,I], we demonstrate that  $K$ -person stochastic games, with  $K \geq 3$ , require an analysis that is substantially different from any analysis used for two-person games. This is done by examining a specific three-person stochastic game.

**Main Theorem 9** Consider the following three-person game  $\Gamma$ :

					$F$
		$N$			
		$L$	$R$		
$T$	0,0,0	0,1,3		3,0,1	1,1,0
			*	*	*
$B$	1,3,0	1,0,1		0,1,1	0,0,0
		*		*	*

*In this cubic three-person game each player has two actions. The actions of the players are denoted as follows: player 1:  $T$  (top),  $B$  (bottom); player 2:  $L$  (left),  $R$  (right); player 3:  $N$  (near),  $F$  (far). The game is represented by taking separately the two layers of the cube that belong to the two actions of player 3 ( $N$  and  $F$ ). Absorbing entries are indicated by  $*$ 's as usual.*

*This game has the following properties:*

(a) Let the Markov strategies  $\kappa, \lambda, \mu$  for player 1,2,3 be respectively given by

$$\kappa = \left( \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, \dots \right)$$

$$\lambda = \left( 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \dots \right)$$

$$\mu = \left( 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, \dots \right),$$

where the  $n$ -th coordinates of the above strategies are the probabilities for the second actions of the players at stage  $n$ . Then  $(\kappa, \lambda, \mu)$  is a Markov equilibrium with equilibrium rewards  $\gamma(\kappa, \lambda, \mu) = (1, 2, 1)$ .

(b) Let  $(\kappa, \lambda, \mu)$  be an equilibrium with the property that, at any stage, at least one of the players plays his second action with a positive probability. Then, at each stage, exactly one of the players plays his second action with a positive probability, and these players appear cyclically in the order 1,2,3.

(c) The set of equilibrium rewards is the triangle

$$\Psi := \{(u, v, w) \in \mathbb{R}^3 \mid u, v, w \geq 1, u + v + w = 4, u = 1 \text{ or } v = 1 \text{ or } w = 1\}.$$

The above theorem will follow from theorems 106, 108, and 109.

To our knowledge, this is the first three-person stochastic game studied in detail. In fact, this game is a three-person recursive repeated game with absorbing states (as we discussed in chapter 7, it makes no difference for the average reward whether or not the payoffs in the absorbing cells differ from 0). In chapter 7, for two-person recursive repeated games with absorbing states we showed the existence of stationary  $\varepsilon$ -equilibria for all  $\varepsilon > 0$ . The above example demonstrates that the two-person result does not extend to stochastic games of this kind with more than two players.

In this game, the gap between two-person and three-person stochastic games also appears in the nature of equilibria. As far as we know, this is the first stochastic game where (cyclic) Markov strategies are indispensable, so the class of almost stationary strategy pairs (cf. definition 77) is too narrow to tackle the equilibrium existence problem for stochastic games with more than two players. A thorough study of the potential possibilities of the use of (cyclic or non-cyclic) Markov strategies is needed.

## 9.2 A cyclic three-person game

We consider the game  $\Gamma$  :

### Example 102

		$N$				$F$	
		$L$	$R$				
$T$		0,0,0	0,1,3	3,0,1		1,1,0	
			*	*			*
$B$		1,3,0	1,0,1	0,1,1		0,0,0	
			*	*			*

In this cubic three-person game each player has two actions. The actions of the players are denoted as follows: player 1: T (top), B (bottom); player 2: L (left), R (right); player 3: N (near), F (far). The game is represented by taking separately the two layers of the cube that belong to the two actions of player 3 (N and F). Absorbing entries are indicated by *\*'s* as usual.

Note that all entries but one are absorbing, so the play absorbs as soon as one of the players chooses his second action, and also that the payoffs and the absorbing entries are cyclically symmetric ( $r^1(i_1, i_2, i_3) = r^2(i_2, i_3, i_1) = r^3(i_3, i_1, i_2)$  for any entry  $(i_1, i_2, i_3) \in \{1, 2\}^3$ ). However we wish to emphasize that we have only introduced this cyclic symmetry to make the analysis of this game clearer. Similar results on the existence of cyclic Markov equilibria and on the non-existence of stationary  $\varepsilon$ -equilibria,  $\varepsilon > 0$ , can also be obtained in non-symmetric games with the very same absorption structure.

In the game  $\Gamma$ , each mixed action can be represented by the probability assigned to the second action, which lets the stationary strategy spaces equal  $[0, 1]$  for each player. For stationary strategies of the players we use the notations  $x, y$  and  $z$  respectively. The spaces of Markov strategies equal  $[0, 1]^\infty$  for each player. Markov strategies are denoted by  $\kappa$  for player 1, by  $\lambda$  for player 2, and by  $\mu$  for player 3.

For this game the only history up to stage  $n \in \mathbb{N}$ , if no absorption has occurred, is the trivial one where all the players have chosen their first action at all stages up to stage  $n$ . Therefore all history dependent strategies are only Markov strategies.

Now we investigate the game  $\Gamma$  in detail.

**Lemma 103** *There is no stationary equilibrium in  $\Gamma$ .*

**Proof.** Suppose by way of contradiction that  $(x, y, z)$  is a stationary equilibrium. First we prove that  $0 < x, y, z < 1$ . Recall that  $x, y, z$  are the probabilities on actions  $B, R$  and  $F$  respectively. If  $x = 0$  then, because of a best reply argument,  $y = 1$  and therefore  $z = 0$ , which contradicts  $x = 0$ . On the other hand  $x = 1$  would imply  $y = 0$ , hence  $z = 1$ , which contradicts  $x = 1$ . So  $0 < x < 1$ , and by symmetry we also have  $0 < y, z < 1$ .

Since  $0 < x < 1$  we have (by applying similar expressions as in lemma 87) that

$$\frac{3(1-y)z + yz}{1 - (1-y)(1-z)} = \gamma^1(0, y, z) = \gamma^1(1, y, z) = 1 - z,$$

thus

$$y = \frac{z^2 + 2z}{z^2 + 1} > z.$$

By symmetry  $z > x$  and  $x > y$ . Hence  $y > z > x > y$ , contradiction.  $\square$

We call a triple  $(x, y, z)$  of stationary strategies in the game  $\Gamma$  absorbing, if  $x > 0$  or  $y > 0$  or  $z > 0$ . Such an absorbing triple eventually leads to absorption with probability 1. On the other hand, a triple  $(x, y, z)$  in the game  $\Gamma$  is called non-absorbing, if  $x = y = z = 0$ ; in this case entry  $(T, L, N)$  is played forever with probability 1.

**Lemma 104** *Let  $(x_n, y_n, z_n)$  be a sequence of stationary strategy triples in the game  $\Gamma$  with  $(\tilde{x}, \tilde{y}, \tilde{z}) := \lim_{n \rightarrow \infty} (x_n, y_n, z_n)$ .*

(a) *Assume that  $(\tilde{x}, \tilde{y}, \tilde{z})$  is absorbing. Then*

$$\gamma(\tilde{x}, \tilde{y}, \tilde{z}) := \lim_{n \rightarrow \infty} \gamma(x_n, y_n, z_n).$$

(b) *Assume that  $(\tilde{x}, \tilde{y}, \tilde{z})$  is non-absorbing,  $(x_n, y_n, z_n)$  are absorbing for all  $n \in \mathbb{N}$ , and the limits of the sequences*

$$w_{(x_n, y_n, z_n)}(T, L, F) := \frac{(1 - x_n)(1 - y_n)z_n}{1 - (1 - x_n)(1 - y_n)(1 - z_n)}$$

$$w_{(x_n, y_n, z_n)}(T, R, N) := \frac{(1 - x_n)y_n(1 - z_n)}{1 - (1 - x_n)(1 - y_n)(1 - z_n)}$$

$$w_{(x_n, y_n, z_n)}(B, L, N) := \frac{x_n(1 - y_n)(1 - z_n)}{1 - (1 - x_n)(1 - y_n)(1 - z_n)}$$

*exist as  $n$  tends to infinity. Then*

$$\sum_{\substack{(i_1, i_2, i_3) \in \{(T, L, F), \\ (T, R, N), (B, L, N)\}}} \left[ \lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(i_1, i_2, i_3) \right] = 1$$

*and*

$$\lim_{n \rightarrow \infty} \gamma(x_n, y_n, z_n) = \sum_{\substack{(i_1, i_2, i_3) \in \{(T, L, F), \\ (T, R, N), (B, L, N)\}}} \left[ \lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(i_1, i_2, i_3) \right] \cdot r(i_1, i_2, i_3).$$

**Proof.** Part (a) follows from lemma 10, so it remains to verify part (b). Notice that  $w_{(x_n, y_n, z_n)}(T, L, F)$  expresses the probability that, with respect to  $(x_n, y_n, z_n)$ , the absorption occurs in entry  $(T, L, F)$ ; in fact  $w_{(x_n, y_n, z_n)}(T, L, F)$  equals entry  $q_1(t|x_n, y_n, z_n)$  of the matrix  $Q(x_n, y_n, z_n)$ , where  $t$  is the absorbing state that entry  $(T, L, F)$  leads to. Similar interpretations hold for the other two expressions  $w_{(x_n, y_n, z_n)}(T, R, N)$  and  $w_{(x_n, y_n, z_n)}(B, L, N)$  as well. It is clear that, with respect to  $(x_n, y_n, z_n)$ , the probability that the eventual absorption occurs in one of the entries  $(T, L, F)$ ,  $(T, R, N)$ , and  $(B, L, N)$  converges to 1 as  $n$  tends to infinity. Now the condition that the above limits exist guarantee that  $Q(x_n, y_n, z_n)$  has a limit as well, so the statement follows from lemma 9-(c).  $\square$

**Theorem 105** *There is no stationary  $\varepsilon$ -equilibrium in  $\Gamma$  for small  $\varepsilon > 0$ .*

**Proof.** Suppose by way of contradiction that  $(x_n, y_n, z_n)$  is a stationary  $\varepsilon_n$ -equilibrium for some positive decreasing sequence  $\varepsilon_n$  converging to 0. Then by taking subsequences, we may assume without loss of generality that  $(x_n, y_n, z_n)$  is absorbing for all  $n \in \mathbb{N}$  (as stationary  $\varepsilon$ -equilibria must be absorbing for small  $\varepsilon > 0$  due to positive payoffs in entries  $(T, L, F)$ ,  $(T, R, N)$ , and  $(B, L, N)$ ),  $(x_n, y_n, z_n)$  is convergent in the compact space  $[0, 1]^3$ , and the expressions in lemma 104-(b) have limits. Let  $(\tilde{x}, \tilde{y}, \tilde{z}) := \lim_{n \rightarrow \infty} (x_n, y_n, z_n)$ . We distinguish two cases.

**Case 1:**  $(\tilde{x}, \tilde{y}, \tilde{z})$  is absorbing, namely either  $\tilde{x} > 0$  or  $\tilde{y} > 0$  or  $\tilde{z} > 0$ .

Suppose without loss of generality that  $\tilde{z} > 0$ . Then  $(\tilde{x}, \tilde{y}, \tilde{z})$  and  $(x, \tilde{y}, \tilde{z})$ , for all  $x \in X$ , are absorbing, hence by lemma 104-(a)

$$\begin{aligned}\gamma^1(\tilde{x}, \tilde{y}, \tilde{z}) &= \lim_{n \rightarrow \infty} \gamma^1(x_n, y_n, z_n) \\ \gamma^1(x, \tilde{y}, \tilde{z}) &= \lim_{n \rightarrow \infty} \gamma^1(x, y_n, z_n) \quad \forall x \in X.\end{aligned}$$

Since  $(x_n, y_n, z_n)$  is an  $\varepsilon_n$ -equilibrium, for any  $n \in \mathbb{N}$ , we have for all  $x \in X$

$$\gamma^1(x, y_n, z_n) \leq \gamma^1(x_n, y_n, z_n) + \varepsilon_n,$$

thus we obtain for all  $x \in X$  that

$$\begin{aligned}\gamma^1(\tilde{x}, \tilde{y}, \tilde{z}) &= \lim_{n \rightarrow \infty} \gamma^1(x_n, y_n, z_n) \\ &\geq \lim_{n \rightarrow \infty} (\gamma^1(x, y_n, z_n) - \varepsilon_n) \\ &= \gamma^1(x, \tilde{y}, \tilde{z}).\end{aligned}$$

For player 2, we have by using analogous arguments that for all  $y \in Y$

$$\gamma^2(\tilde{x}, \tilde{y}, \tilde{z}) \geq \gamma^2(\tilde{x}, y, \tilde{z}).$$

If  $\tilde{x} > 0$  or  $\tilde{y} > 0$  then, similarly for player 3 as well, for all  $z \in Z$

$$\gamma^3(\tilde{x}, \tilde{y}, \tilde{z}) \geq \gamma^3(\tilde{x}, \tilde{y}, z),$$

otherwise for all  $z \in Z$

$$\gamma^3(\tilde{x}, \tilde{y}, \tilde{z}) = 1 \geq \gamma^3(\tilde{x}, \tilde{y}, z).$$

Hence  $(\tilde{x}, \tilde{y}, \tilde{z})$  is a stationary equilibrium, which contradicts lemma 103.

**Case 2:**  $(\tilde{x}, \tilde{y}, \tilde{z})$  is non-absorbing, namely  $\tilde{x} = \tilde{y} = \tilde{z} = 0$ .

By taking a subsequence, due to the symmetrical structure of the game, we may further assume without loss of generality that for all  $n \in \mathbb{N}$

$$w_{(x_n, y_n, z_n)}(T, L, F) \geq \max \{w_{(x_n, y_n, z_n)}(T, R, N), w_{(x_n, y_n, z_n)}(B, L, N)\} \quad (9.1)$$

and  $w_{(0, y_n, z_n)}(T, L, F)$  has a limit as  $n$  tends to infinity. By lemma 104-(b), we have

$$w_{(x_n, y_n, z_n)}(T, L, F) \geq \frac{1}{3}.$$

As

$$3a + b < \frac{3a}{1-b} \quad \forall a \in [\frac{1}{3}, 1], \forall b \in (0, 1),$$

if  $\lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(B, L, N) > 0$  then we obtain

$$\begin{aligned} 3 \lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(T, L, F) + \lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(B, L, N) & \quad (9.2) \\ < 3 \frac{\lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(T, L, F)}{1 - \lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(B, L, N)}. \end{aligned}$$

We have

$$w_{(x_n, y_n, z_n)}(T, L, F) = \frac{(1-x_n)(1-y_n)z_n}{1 - (1-x_n)(1-y_n)(1-z_n)},$$

$$w_{(x_n, y_n, z_n)}(B, L, N) = \frac{x_n(1-y_n)(1-z_n)}{1 - (1-x_n)(1-y_n)(1-z_n)},$$

$$w_{(0, y_n, z_n)}(T, L, F) = \frac{(1-y_n)z_n}{1 - (1-y_n)(1-z_n)}.$$

We now show that

$$\lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(B, L, N) = 0$$

The opposite

$$\lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(B, L, N) > 0$$

would imply, using lemma 104-(b) and (9.2), that

$$\begin{aligned} \lim_{n \rightarrow \infty} \gamma^1(x_n, y_n, z_n) &= \lim_{n \rightarrow \infty} [3 w_{(x_n, y_n, z_n)}(T, L, F) + w_{(x_n, y_n, z_n)}(B, L, N)] \\ &< \lim_{n \rightarrow \infty} \left[ 3 \frac{w_{(x_n, y_n, z_n)}(T, L, F)}{1 - w_{(x_n, y_n, z_n)}(B, L, N)} \right] \\ &= \lim_{n \rightarrow \infty} \left[ 3 \frac{1}{1-x_n} w_{(0, y_n, z_n)}(T, L, F) \right] \\ &= \lim_{n \rightarrow \infty} \gamma^1(0, y_n, z_n), \end{aligned}$$

so  $(x_n, y_n, z_n)$  would not be an  $\varepsilon_n$ -equilibrium for large  $n \in \mathbb{N}$ . Hence,

$$\lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(B, L, N) = 0$$

must hold indeed. But then lemma 104 and (9.1) yield

$$\lim_{n \rightarrow \infty} \gamma^2(x_n, y_n, z_n) = \lim_{n \rightarrow \infty} w_{(x_n, y_n, z_n)}(T, R, N) < 1 = \lim_{n \rightarrow \infty} \gamma^2(x_n, 1, z_n),$$

which contradicts the fact that  $(x_n, y_n, z_n)$  is an  $\varepsilon_n$ -equilibrium for large  $n \in \mathbb{N}$ .  $\square$

Now we turn to the class of Markov strategies. First we present a Markov equilibrium, which has a cyclic nature.

**Theorem 106** *In the game  $\Gamma$ , let the Markov strategies  $\kappa, \lambda, \mu$  for players 1, 2, 3 be respectively given by*

$$\kappa = \left( \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, \dots \right)$$

$$\lambda = \left( 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \dots \right)$$

$$\mu = \left( 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, \dots \right).$$

*Then  $(\kappa, \lambda, \mu)$  is a Markov equilibrium in  $\Gamma$  with equilibrium reward*

$$\gamma(\kappa, \lambda, \mu) = (1, 2, 1).$$

**Proof.** First notice that

$$\begin{aligned} \gamma(\kappa, \lambda, \mu) &= \frac{1}{2} \cdot (1, 3, 0) + \left(\frac{1}{2}\right)^2 \cdot (0, 1, 3) + \left(\frac{1}{2}\right)^3 \cdot (3, 0, 1) \\ &\quad + \left(\frac{1}{2}\right)^4 \cdot (1, 3, 0) + \left(\frac{1}{2}\right)^5 \cdot (0, 1, 3) + \dots \\ &= (1, 2, 1). \end{aligned}$$

We prove that  $\kappa$  is a best reply of player 1 against  $(\lambda, \mu)$ . Similar proofs can be given to show that the other two players have no profitable deviations either, so the proof will then be complete.

First we clarify why player 1 must have a cyclic Markov best reply against  $(\lambda, \mu)$ . We may represent the cyclic strategies  $\lambda$  and  $\mu$  as stationary strategies  $y$  and  $z$  on three non-absorbing states where each of these non-absorbing states is identical with the non-absorbing state of the original game  $\Gamma$  except for entries  $(T, L, N)$  which make the play visit these three states cyclically. Then the cyclic strategies  $\lambda$  and  $\mu$  become stationary strategies, so the proof of theorem 16-(b) extends to this game with three players. Therefore, player 1 has a stationary best reply  $x$  against  $y$  and  $z$ , which, due to the representation, corresponds to a cyclic Markov best reply  $\tilde{\kappa}$  against  $(\lambda, \mu)$  in the original game  $\Gamma$ .

By way of contradiction suppose now that a cyclic best reply  $\tilde{\kappa} = (\tilde{x}_n)_{n=1}^\infty$  of player 1 against  $(\lambda, \mu)$  is profitable, namely  $\gamma^1(\tilde{\kappa}, \lambda, \mu) > \gamma^1(\kappa, \lambda, \mu)$  and  $\gamma^1(\tilde{\kappa}, \lambda, \mu) \geq \gamma^1(\bar{\kappa}, \lambda, \mu)$  for all  $\bar{\kappa}$ . We show that we need to have  $\tilde{\kappa} = (0, 0, 0, \dots)$ . Let  $\tilde{\kappa}_l := (\tilde{x}_n)_{n=l}^\infty$ , and let  $\kappa_l := (x_n)_{n=l}^\infty$  for all  $l \in \mathbb{N}$ . Let  $\lambda_l, \mu_l$  be defined analogously. We have

$$\gamma^1(\tilde{\kappa}, \lambda, \mu) > \gamma^1(\kappa, \lambda, \mu) = 1.$$

So  $\tilde{x}_1 = 0$ , which implies

$$\gamma^1(\tilde{\kappa}_2, \lambda_2, \mu_2) = \gamma^1(\tilde{\kappa}, \lambda, \mu) > 1 = \gamma^1(\kappa_2, \lambda_2, \mu_2),$$

therefore  $\tilde{x}_2 = 0$ . Using the equality

$$\gamma^1(\tilde{\kappa}_2, \lambda_2, \mu_2) = \frac{1}{2} \gamma^1(\tilde{\kappa}_3, \lambda_3, \mu_3)$$

we obtain

$$\gamma^1(\tilde{\kappa}_2, \lambda_2, \mu_2) > 2 = \gamma^1(\kappa_3, \lambda_3, \mu_3),$$

so  $\tilde{x}_3 = 0$ . Now using the cyclicity of  $\tilde{\kappa}$ , we have  $\tilde{\kappa} = (0, 0, 0, \dots)$  indeed.

But then

$$\begin{aligned} \gamma^1(\tilde{\kappa}, \lambda, \mu) &= \frac{1}{2} \cdot (0, 1, 3) + \left(\frac{1}{2}\right)^2 \cdot (3, 0, 1) + \left(\frac{1}{2}\right)^3 \cdot (0, 1, 3) + \dots \\ &= 1 \\ &= \gamma^1(\kappa, \lambda, \mu), \end{aligned}$$

contradiction.  $\square$

Observe that for all  $l \in \mathbb{N}$  the strategies  $\kappa_l := (x_n)_{n=l}^\infty, \lambda_l := (y_n)_{n=l}^\infty, \mu_l := (z_n)_{n=l}^\infty$  form cyclic Markov equilibria as well, where  $\kappa, \lambda$ , and  $\mu$  are defined as in theorem 106. Also, if for  $x \in [0, 1]$  and  $n \in \mathbb{N}$  the notation  $x(n)$  represents playing  $x$  for  $n$  subsequent stages, then the strategies

$$\pi = (\alpha(n), 0(n), 0(n), \alpha(n), 0(n), 0(n), \alpha(n), \dots)$$

$$\sigma = (0(n), \alpha(n), 0(n), 0(n), \alpha(n), 0(n), 0(n), \dots)$$

$$\tau = (0(n), 0(n), \alpha(n), 0(n), 0(n), \alpha(n), 0(n), \dots)$$

form an equilibrium for each  $n$ , if  $(1 - \alpha)^n = \frac{1}{2}$ . The equality  $(1 - \alpha)^n = 1/2$  makes that in any period  $n$  of stages the play absorbs with probability  $1/2$ .

Notice that, in an equilibrium, any stage where all the players choose their first actions may be skipped without loosing the equilibrium property. The following lemma considers equilibria in terms of strategies where, at any stage, at least one of the players plays his second action with a positive probability.

**Lemma 107** *Let  $(\kappa, \lambda, \mu)$  be an equilibrium in  $\Gamma$  with the property that, at any stage, at least one of the players plays his second action with a positive probability. Then*

- (a)  $x_n, y_n, z_n < 1$  and  $(\kappa_n, \lambda_n, \mu_n)$  is an equilibrium for all  $n \in \mathbb{N}$ ;
- (b) there exists an  $n \in \mathbb{N}$  for which  $x_n = 0$  or  $y_n = 0$  or  $z_n = 0$ ;
- (c) for any  $n \in \mathbb{N}$ , if  $z_n = 0$  then either  $x_n = 0$  or  $y_n = 0$ ;
- (d) if  $x_n > 0, y_n = z_n = 0$  then either  $x_{n+1} > 0, y_{n+1} = z_{n+1} = 0$  or  $y_{n+1} > 0, x_{n+1} = z_{n+1} = 0$ ;
- (e) if  $x_n > 0$  and  $y_n = z_n = 0$  then  $\min\{u_n, v_n, w_n\} = 1$ ;
- (f)  $x_1 = 0$  or  $y_1 = 0$  or  $z_1 = 0$ .

**Proof.** Let

$$\kappa = (x_n)_{n=1}^{\infty}, \quad \lambda := (y_n)_{n=1}^{\infty}, \quad \mu_l := (z_n)_{n=1}^{\infty}.$$

For all  $l \in \mathbb{N}$  let

$$\begin{aligned} \kappa_l &:= (x_n)_{n=l}^{\infty}, & \lambda_l &:= (y_n)_{n=l}^{\infty}, & \mu_l &:= (z_n)_{n=l}^{\infty} \\ (u_l, v_l, w_l) &:= \gamma(\kappa_l, \lambda_l, \mu_l). \end{aligned}$$

**Proof of (a):**  $x_n, y_n, z_n < 1$  and  $(\kappa_n, \lambda_n, \mu_n)$  is an equilibrium for all  $n \in \mathbb{N}$ .

▷ If  $x_1 = 1$  or  $y_1 = 1$  or  $z_1 = 1$ , then  $(x_1, y_1, z_1)$  would be a stationary equilibrium, which would contradict lemma 103. Hence  $x_1, y_1, z_1 < 1$ , which means that stage 2 is reached with a positive probability, so  $(\kappa_2, \lambda_2, \mu_2)$  must be an equilibrium. Therefore  $x_2, y_2, z_2 < 1$  must hold as well. Now repeating this argument implies the statement.

**Proof of (b):** there exists an  $n \in \mathbb{N}$  for which  $x_n = 0$  or  $y_n = 0$  or  $z_n = 0$ .

▷ Suppose by way of contradiction that  $0 < x_n, y_n, z_n$  for all  $n \in \mathbb{N}$ . Then by (a) we have  $0 < x_n, y_n, z_n < 1$  for all  $n \in \mathbb{N}$ . Now for stage 1 we have

$$u_1 = \gamma^1((0, x_2, x_3, \dots), \lambda, \mu) = \gamma^1((1, x_2, x_3, \dots), \lambda, \mu),$$

hence

$$u_1 = 3(1 - y_1)z_1 + y_1z_1 + (1 - y_1)(1 - z_1)u_2 = 1 - z_1.$$

By expressing  $u_2$

$$u_2 = \frac{1 - 4z_1 + 2y_1z_1}{(1 - y_1)(1 - z_1)}.$$

Similar equations hold concerning the other two players. Due to symmetry we may assume for stage 1 that  $u_1 \leq \min\{v_1, w_1\}$ . Then by the equations  $u_1 = 1 - z_1$ ,  $v_1 = 1 - x_1$ ,  $w_1 = 1 - y_1$  we obtain  $z_1 \geq \max\{x_1, y_1\}$ . This implies

$$u_2 \leq 1 - \frac{z_1}{1 - z_1},$$

and then

$$u_1 - u_2 \geq \frac{z_1}{1 - z_1} - z_1 > z_1^2.$$

So we have

$$\begin{aligned} \min\{u_2, v_2, w_2\} &\leq u_2 \\ &< u_1 - z_1^2 \\ &= \min\{u_1, v_1, w_1\} - (\max\{x_1, y_1, z_1\})^2 \\ &< \min\{u_1, v_1, w_1\}. \end{aligned}$$

For stage 2 we have  $u_2 = 1 - z_2$ ,  $v_2 = 1 - x_2$ ,  $w_2 = 1 - y_2$ , which yields

$$\max\{x_2, y_2, z_2\} > \max\{x_1, y_1, z_1\}.$$

Then analogously

$$\begin{aligned} \min\{u_3, v_3, w_3\} &< \min\{u_2, v_2, w_2\} - (\max\{x_2, y_2, z_2\})^2 \\ &< \min\{u_1, v_1, w_1\} - ((\max\{x_1, y_1, z_1\})^2 + (\max\{x_2, y_2, z_2\})^2), \end{aligned}$$

and

$$\max\{x_3, y_3, z_3\} > \max\{x_2, y_2, z_2\}.$$

By using this inductively, we find that as  $n$  increases the number  $\min\{u_n, v_n, w_n\}$  goes below zero, which is a contradiction.

**Proof of (c):** For any  $n \in \mathbb{N}$ , if  $z_n = 0$  then either  $x_n = 0$  or  $y_n = 0$ .

▷ Assume by way of contradiction that  $0 < x_n, y_n$ . By (a) we have  $0 < x_n, y_n < 1$ . Then

$$\begin{aligned} u_n &= 1, \quad u_{n+1} = \frac{u_n}{1 - y_n} > 1 \\ v_n &= 1 - x_n, \quad v_{n+1} = \frac{v_n - 3x_n}{1 - x_n} < 1. \end{aligned}$$

Since  $u_{n+1} > 1$  we obtain  $x_{n+1} = 0$ . But then  $v_{n+1} \geq 1$ , contradiction.

**Proof of (d):** If  $x_n > 0$ ,  $y_n = z_n = 0$  then either  $x_{n+1} > 0$ ,  $y_{n+1} = z_{n+1} = 0$  or  $y_{n+1} > 0$ ,  $x_{n+1} = z_{n+1} = 0$ .

▷ Since  $1 \leq w_n = (1 - x_n)w_{n+1}$  we have  $w_{n+1} > 1$ . The second action of any player cannot give himself more than 1, so  $z_{n+1} = 0$ , and by (c) either  $x_{n+1} = 0$  or  $y_{n+1} = 0$ .

**Proof of (e):** If  $x_n > 0$  and  $y_n = z_n = 0$  then  $\min\{u_n, v_n, w_n\} = 1$ .

▷ Obviously, we have  $u_n = 1$  and  $w_n \geq 1$ . Suppose  $\bar{n}$  is the first stage after stage  $n$  with  $y_{\bar{n}} > 0$  (there must be such a stage, otherwise by (d) we would obtain  $z_n = z_{n+1} = \dots = 0$ , and hence  $\gamma^3(\kappa_n, \lambda_n, \mu_n) = 0 < 1 = \gamma^3(\kappa_n, \lambda_n, 1)$  would hold contradicting the fact that  $(\kappa, \lambda, \mu)$  is an equilibrium). Thus  $\gamma^2(\kappa_{\bar{n}}, \lambda_{\bar{n}}, \mu_{\bar{n}}) = 1$ . By (d) we have  $z_n = \dots = z_{\bar{n}-1} = 0$ . Therefore

$$\begin{aligned} v_n &= 3(x_n + (1 - x_n)x_{n+1} + \dots + (1 - x_n) \dots (1 - x_{\bar{n}-2})x_{\bar{n}-1}) \\ &\quad + (1 - x_n) \dots (1 - x_{\bar{n}-1})\gamma^2(\kappa_{\bar{n}}, \lambda_{\bar{n}}, \mu_{\bar{n}}), \end{aligned}$$

and by (a) we obtain  $(1 - x_n) \dots (1 - x_{\bar{n}-1}) > 0$ , so  $v_n > 1$ .

**Proof of (f):**  $x_1 = 0$  or  $y_1 = 0$  or  $z_1 = 0$ .

▷ Suppose that  $n$  is the first stage when  $x_n = 0$  or  $y_n = 0$  or  $z_n = 0$ . If  $n = 1$  then we are done. Otherwise assume by way of contradiction that  $n > 1$ . Then  $0 < x_1, y_1, z_1, \dots, x_{n-1}, y_{n-1}, z_{n-1} < 1$ , and  $u_1 = 1 - z_1 < 1$ , and therefore just like above we have  $\min\{u_n, v_n, w_n\} < \min\{u_1, v_1, w_1\} \leq u_1 < 1$ , which contradicts (e). ◻

The next theorem says that all Markov equilibria are of the same type as presented in theorem 106.

**Theorem 108** *Let  $(\kappa, \lambda, \mu)$  be an equilibrium in  $\Gamma$  with the property that, at any stage, at least one of the players plays his second action with a positive probability. Then, at each stage exactly one of the players plays his second action with a positive probability, and these players appear cyclically in the order 1,2,3.*

**Proof.** Due to symmetry and lemma 107-(f), we may suppose without loss of generality that  $z_1 = 0$ . Then by lemma 107-(c) we have  $x_1 = 0$  or  $y_1 = 0$ . Assume  $y_1 = 0$ , so  $x_1 > 0$ . By lemma 107-(d) either  $x_2 > 0$ ,  $y_2 = z_2 = 0$  or  $y_2 > 0$ ,  $x_2 = z_2 = 0$ . Now using lemma 107-(d) inductively we obtain that at any stage exactly one of the players plays his second action with positive probability. And finally, the second statement of the theorem is an immediate consequence of lemma 107-(d), so the proof is complete.  $\square$

**Theorem 109** *The set of feasible equilibrium rewards for  $\Gamma$  is the triangle*

$$\Psi := \{(u, v, w) \in \mathbb{R}^3 \mid u, v, w \geq 1, u + v + w = 4, u = 1 \text{ or } v = 1 \text{ or } w = 1\}.$$

**Proof.** Let the strategy triple  $(\kappa, \lambda, \mu)$  be a Markov equilibrium for  $\Gamma$  with rewards  $(u, v, w) = \gamma(\kappa, \lambda, \mu)$ . We show that  $(u, v, w) \in \Psi$ . Suppose  $x_1 > 0$ . Then by theorem 108 we have  $y_1 = z_1 = 0$ . Hence  $u = 1, w \geq 1$ . Let  $n$  be the first stage when  $y_n > 0$ . This implies that  $\gamma^2(\kappa_n, \lambda_n, \mu_n) = 1$  and  $z_1 = \dots = z_{n-1} = 0$ . Thus

$$\begin{aligned} v &= 3(x_1 + (1 - x_1)x_2 + \dots + (1 - x_1) \dots (1 - x_{n-2})x_{n-1}) \\ &\quad + (1 - x_1) \dots (1 - x_{n-1})\gamma^2(\kappa_n, \lambda_n, \mu_n), \end{aligned}$$

and since  $x_1, \dots, x_{n-1} < 1$  (cf. lemma 107-(a)) we have  $(1 - x_1) \dots (1 - x_{n-1}) > 0$ , so  $v \geq 1$ . The equality  $u + v + w = 4$  is trivial.

Now we show that if  $(u, v, w) \in \Psi$  then there exists a Markov equilibrium  $(\kappa, \lambda, \mu)$  with rewards  $(u, v, w)$ . By symmetry it suffices to find a Markov equilibrium with rewards  $(1, 1 + \alpha, 2 - \alpha)$ , where  $\alpha \in [0, 1]$ . Let

$$\begin{aligned} \kappa &= \left( \frac{\alpha}{2}, 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, \dots \right) \\ \lambda &= \left( 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, \dots \right) \\ \mu &= \left( 0, 0, \frac{1}{2}, 0, 0, \frac{1}{2}, 0, 0, \dots \right). \end{aligned}$$

These are almost the strategies defined in theorem 106, but the mixed action of player 1 for the first stage is modified. Now  $\gamma(\kappa, \lambda, \mu) = (1, 1 + \alpha, 2 - \alpha)$ , and it can be verified similarly to the proof of theorem 106 that  $(\kappa, \lambda, \mu)$  is a Markov equilibrium indeed.  $\square$

### 9.3 Concluding remarks

Recently, Solan [1998] proved the existence of  $\varepsilon$ -equilibria, for all  $\varepsilon > 0$ , in three-person repeated games with absorbing states. He showed that, in each three-person repeated games with absorbing states, at least one of three different types of  $\varepsilon$ -equilibria must occur. One of these types is exactly the class of cyclic  $\varepsilon$ -equilibria, with further interesting results concerning them.

## Chapter 10

# Appendix: uniform optimality and equilibria

We wish to mention that there are several alternative rewards, which, just like the average reward (cf. definition 7), are frequently used in stochastic game theory for an evaluation of the long-term average payoffs. Some of the most important ones, for player  $k \in \{1, 2\}$  and with respect to a strategy pair  $(\pi, \sigma) \in \Pi \times \Sigma$  and initial state  $s \in S$ , are the following :

$$\rho_s^{1,k}(\pi, \sigma) := \mathcal{E}_{s\pi\sigma} \left( \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n^k \right),$$

$$\rho_s^{2,k}(\pi, \sigma) := \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \mathcal{E}_{s\pi\sigma} (R_n^k),$$

$$\rho_s^{3,k}(\pi, \sigma) := \mathcal{E}_{s\pi\sigma} \left( \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n^k \right),$$

where  $R_n^k$  denotes the random variable for the payoff of player  $k$  at stage  $n$ .

We now briefly discuss the relationship between the average reward  $\gamma$  and these above mentioned alternative rewards. First of all, it always holds for any player  $k \in \{1, 2\}$  that for any  $s \in S$ ,  $(\pi, \sigma) \in \Pi \times \Sigma$  we have

$$\rho_s^{1,k}(\pi, \sigma) \leq \gamma_s^k(\pi, \sigma) \leq \rho_s^{2,k}(\pi, \sigma) \leq \rho_s^{3,k}(\pi, \sigma). \quad (10.1)$$

Here the first inequality is a simple consequence of Fatou's lemma (cf. Fatou [1906]), the second inequality follows from the fact the limit inferior is always smaller than or equal to the limit superior of any real sequence, finally the last inequality is an implication of Fatou's lemma together with the fact that  $\limsup_{n \rightarrow \infty} a_n = -\liminf_{n \rightarrow \infty} (-a_n)$  holds for any bounded real sequence  $(a_n)_{n \in \mathbb{N}}$ ; note that the boundedness of the payoffs from below is crucial when applying Fatou's lemma above.

We wish to stress that equalities in (10.1) do not hold in general. Nevertheless, it is fortunate to know that all these four rewards are equal when both players use stationary strategies, duely to the fact that stationary strategy pairs induce Markov

processes on the set of states, as discussed in section 2.2.3. In fact, for any of the alternative rewards  $\rho^m$ ,  $m = 1, 2, 3$ , the properties in lemmas 9 and 10 remain valid just as theorem 16-(b) on the best replies against a fixed stationary strategy (since lemma 9 still holds, a stationary strategy is a best reply with respect to any of these rewards if and only if it is a best reply with respect to all of them). Moreover, stationary strategies guarantee the same rewards (cf. definition 22) irrespective of the choice of these alternative rewards (cf. Bewley & Kohlberg [1978]).

## Zero-sum stochastic games

Naturally, with respect to any of the rewards  $\rho^m$ ,  $m = 1, 2, 3$ , we can speak of zero-sum stochastic games, values, and optimality, as in section 2.5 with respect to the average reward  $\gamma$ . Before turning to these issues, we wish to define the very appealing concept of uniform ( $\varepsilon$ -)optimality in zero-sum stochastic games. The idea here is, intuitively, to find strategies that are ( $\varepsilon$ -)optimal in the finite game up to stage  $N$  on the condition that  $N$  is sufficiently large, and at the same time ( $\varepsilon$ -)optimal in the infinite game.

**Definition 110** *In a zero-sum stochastic game, for a strategy  $\pi \in \Pi$  and initial state  $s \in S$ , let  $\underline{w}_s(\pi)$  be the supremum of all the real numbers  $a_s$  with the properties that*

(a) *for all  $\delta > 0$  there exists a stage  $N^\delta$  such that for all  $\sigma \in \Sigma$*

$$\mathcal{E}_{s\pi\sigma} \left( \frac{1}{N} \sum_{n=1}^N R_n \right) \geq a_s - \delta \quad \forall N \geq N^\delta,$$

(b) *for all  $\sigma \in \Sigma$*

$$\mathcal{E}_{s\pi\sigma} \left( \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n \right) \geq a_s;$$

where  $R_n$  denotes the random variable for the payoff at stage  $n$ .

We say that strategy  $\pi$  uniformly guarantees reward  $c_s$  for initial state  $s$ , if  $\underline{w}_s(\pi) \geq c_s$ . Uniformly guaranteed rewards  $\bar{w}_s(\sigma)$  for strategies  $\sigma$  of player 2 are similarly defined. If there exists a real valued vector  $w = (w_s)_{s \in S}$  such that

$$w_s = \sup_{\pi \in \Pi} \underline{w}_s(\pi) = \inf_{\sigma \in \Sigma} \bar{w}_s(\sigma) \quad \forall s \in S,$$

then  $w$  is called the uniform value of the zero-sum stochastic game.

Whenever the uniform value  $w$  exists, we can define uniform optimality and uniform  $\varepsilon$ -optimality,  $\varepsilon > 0$ , as in the case of the average reward in definition 22.

In light of the definition, uniformly ( $\varepsilon$ -)optimal strategies can be applied whenever the zero-sum stochastic game is to be played sufficiently long (even over infinitely many stages). As mentioned in Mertens & Neyman [1981], uniformly ( $\varepsilon$ -)optimal strategies are also ( $\varepsilon$ -)optimal with respect to the  $\beta$ -discounted reward if the discount factor  $\beta \in (0, 1)$  is sufficiently close to 1. So the main motivation for using uniformly

( $\varepsilon$ -)optimal strategies is that their structure is independent of the exact duration of the game or of the exact discount factor (on condition that the game is sufficiently long or the discount factor is large enough).

Mertens & Neyman [1981] showed that, in every zero-sum stochastic game, the values for the alternative rewards  $\rho^m$ ,  $m = 1, 2, 3$  and the uniform value exist and equal to the average value  $v$  (so particularly  $v_s = w_s$  for all  $s \in S$ ); and the players have strategies, for any  $\varepsilon > 0$ , that are  $\varepsilon$ -optimal with regard to the rewards  $\gamma$  and  $\rho^m$ ,  $m = 1, 2, 3$ , and at the same time uniformly  $\varepsilon$ -optimal.

Note that, using (10.1) and that the values are equal, average ( $\varepsilon$ -)optimality implies ( $\varepsilon$ -)optimality for  $\rho^m$ ,  $m = 2, 3$  as well as uniform ( $\varepsilon$ -)optimality yields ( $\varepsilon$ -)optimality for any of the rewards  $\gamma$  and  $\rho^m$ ,  $m = 1, 2, 3$ .

Bewley & Kohlberg [1978] showed that stationary strategies guarantee the same rewards in the uniform sense as for all the rewards  $\gamma$  and  $\rho^m$ ,  $m = 1, 2, 3$ , so particularly  $\underline{v}_s(x) = \underline{w}_s(x)$  for all initial states  $s \in S$  and  $x \in X$ , and similarly for player 2.

Now we would like to discuss how the results in chapters 3, 4, and 5, extend to guaranteed rewards and ( $\varepsilon$ -)optimality with respect to the rewards  $\rho^m$ ,  $m = 1, 2, 3$ , as well as to uniformly guaranteed rewards and uniform ( $\varepsilon$ -)optimality. Based on the previous discussion, the extensions to the rewards  $\rho^m$ ,  $m = 2, 3$ , are immediate. Hence we only focus on reward  $\rho^1$ , uniformly guaranteed rewards, and uniform ( $\varepsilon$ -)optimality. As, in the light of the observations above, the extensions of the results concerning stationary strategies are straightforward, we only need to examine the results in these chapters where no stationary strategies are involved.

First of all, in chapter 3, the Markov strategy  $f$  in the second part of Main Theorem 3 is unfortunately not necessarily optimal for  $\rho^1$  nor uniformly optimal. Note, however, that its construction in the proof of theorem 36-(b) guarantees that for all  $\delta > 0$  there exists a stage  $N^\delta$  such that for all  $\sigma \in \Sigma$

$$\mathcal{E}_{sf\sigma} \left( \frac{1}{N} \sum_{n=1}^N R_n \right) \geq v_s - \delta \quad \forall N \geq N^\delta. \quad (10.2)$$

In order to achieve optimality for  $\rho^1$ , which would now also be sufficient for uniform optimality by (10.2), a somewhat more subtle but rather technical construction can be given. As the techniques in chapter 4 are very similar to those in chapter 3, the same can be said about the Markov strategy in Main Theorem 4.

In chapter 5, for any  $\varepsilon > 0$ , if  $K \in \mathbb{N}$  is sufficiently large then the Markov strategies  $f^K$  constructed for theorem 60 is also  $\varepsilon$ -optimal for reward  $\rho^1$  (see inequalities (5.12) in the proof of lemma 69), hence, by using (10.1) and that the values are equal,  $f^K$  is necessarily  $\varepsilon$ -optimal for  $\rho^m$ ,  $m = 2, 3$ . Although the proofs suggest that  $f^K$  should be uniformly  $\varepsilon$ -optimal as well, it does not immediately follow from the proven results.

## General-sum stochastic games

With respect to the rewards  $\rho^m$ ,  $m = 1, 2, 3$ , we can investigate general-sum stochastic games and we may define ( $\varepsilon$ -)equilibria, as in section 2.6 for the average reward  $\gamma$ . But first we define uniform ( $\varepsilon$ -)equilibria, which, intuitively, are strategy pairs with the property that they form ( $\varepsilon$ -)equilibria in the finite game up to stage  $N$  on condition that  $N$  is sufficiently large, and at the same time they are ( $\varepsilon$ -)equilibria in the infinite game.

**Definition 111** *In a general-sum stochastic game, for initial state  $s \in S$ , a pair of strategies  $(\pi, \sigma) \in \Pi \times \Sigma$  is called a uniform  $\varepsilon$ -equilibrium,  $\varepsilon \geq 0$ , with reward  $a = (a^1, a^2)$ , if for all  $\delta > 0$  there exists a stage  $N^\delta$  with the following properties*

(a) *for both players  $k = 1, 2$*

$$a_s^k - \delta \leq \mathcal{E}_{s\pi\sigma} \left( \frac{1}{N} \sum_{n=1}^N R_n^k \right) \leq a_s^k + \delta \quad \forall N \geq N^\delta,$$

$$\mathcal{E}_{s\pi\sigma} \left( \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n^k \right) = \mathcal{E}_{s\pi\sigma} \left( \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n^k \right) = a_s^k;$$

(b) *for all  $\bar{\pi} \in \Pi$*

$$\mathcal{E}_{s\bar{\pi}\sigma} \left( \frac{1}{N} \sum_{n=1}^N R_n^1 \right) \leq a_s^1 + \varepsilon + \delta \quad \forall N \geq N^\delta,$$

$$\mathcal{E}_{s\bar{\pi}\sigma} \left( \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n^1 \right) \leq a_s^1 + \varepsilon;$$

(c) *for all  $\bar{\sigma} \in \Sigma$*

$$\mathcal{E}_{s\pi\bar{\sigma}} \left( \frac{1}{N} \sum_{n=1}^N R_n^2 \right) \leq a_s^2 + \varepsilon + \delta \quad \forall N \geq N^\delta,$$

$$\mathcal{E}_{s\pi\bar{\sigma}} \left( \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N R_n^2 \right) \leq a_s^2 + \varepsilon,$$

where  $R_n^k$  denotes the random variable for the payoff of player  $k$  at stage  $n$ .

The strategy pair  $(\pi, \sigma)$  is a uniform  $\varepsilon$ -equilibrium, if it is a uniform  $\varepsilon$ -equilibrium for all initial states  $s \in S$ . Uniform 0-equilibria are simply called uniform equilibria.

The motivation for applying uniform ( $\varepsilon$ -)equilibria is quite the same as for uniformly ( $\varepsilon$ -)optimal strategies. Uniform ( $\varepsilon$ -)equilibria can thus be used whenever the stochastic game is to be played sufficiently long (even over infinitely many stages). Moreover, uniform ( $\varepsilon$ -)equilibria are also ( $\varepsilon$ -)optimal with respect to the discounted rewards on condition that the discount factors are sufficiently close to 1.

Note that, using (10.1), uniform ( $\varepsilon$ -)equilibria are necessarily ( $\varepsilon$ -)equilibria for any of the rewards  $\gamma$  and  $\rho^m$ ,  $m = 1, 2, 3$ .

We will now briefly discuss how the results in chapters 6,7,8, and 9 extend to uniform ( $\varepsilon$ -)equilibria, and therefore to ( $\varepsilon$ -)equilibria for the rewards  $\rho^m$ ,  $m = 1, 2, 3$ . Based on the discussion above on the similarities between the rewards  $\gamma$  and  $\rho^m$ ,  $m = 1, 2, 3$ ,

when stationary strategies are used, it is easy to see the validity of Main Theorems 6 and 7 for uniform  $\varepsilon$ -equilibria (in chapter 6, by almost stationary uniform  $\varepsilon$ -equilibria we obviously mean uniform  $\varepsilon$ -equilibria which have the almost stationary property specified in definition 77); note that the existence of uniformly  $\varepsilon$ -optimal strategies,  $\varepsilon > 0$ , in zero-sum games plays a crucial role in the extension to almost stationary uniform  $\varepsilon$ -equilibria.

We may define average-discounted uniform ( $\varepsilon$ -)equilibria as average-discounted ( $\varepsilon$ -)equilibria (cf. definition 96) where uniformity is expected on the side of player 1 (recall that player 1 uses the average reward, while player 2 is interested in his discounted reward). It is not hard to verify that Main Theorem 8 generalizes to the uniform case, based on the structure of the strategies constructed there.

In chapter 9, it is not hard to check that the results of Main Theorem 9 hold in the uniform sense as well, which is due to the transition structure of the game presented there.



# Chapter 11

## References

- Bewley T & Kohlberg E [1976]: The asymptotic theory of stochastic games. *Mathematics of Operations Research* 1, 197-208.
- Bewley T & Kohlberg E [1978]: On stochastic games with stationary optimal strategies. *Mathematics of Operations Research* 3, 104-125.
- Blackwell D [1962]: Discrete dynamic programming. *Annals of Mathematical Statistics* 33, 719-726.
- Blackwell D & Ferguson TS [1968]: The big match. *Annals of Mathematical Statistics* 39, 159-163.
- Bohnenblust HF, Karlin S & Shapley LS [1950]: Solutions of discrete two-person games. In: Kuhn HW & Tucker AW (eds.), *Contributions to the theory of games I* (*Annals of Mathematical Studies* 24, Princeton University Press, Princeton, 51-72.
- Coulomb JM [1992]: Repeated games with absorbing states and no signals. *International Journal of Game Theory* 21, 161-174.
- van Damme E [1991]: *Stability and perfection of Nash equilibria*. Springer Verlag, Berlin.
- Doob JL [1953]: *Stochastic processes*. Wiley, New York.
- Everett H [1957]: Recursive games. In: Dresher M, Tucker AW & Wolfe P (eds.), *Contributions to the Theory of Games III*, *Annals of Mathematical Studies* 39, Princeton University Press, 47-78.
- Fatou PJL [1906]: Séries trigonométriques et séries de Taylor. *Acta Mathematica* 30, 335-400.
- Federgruen A [1978]: On  $n$ -person stochastic games with denumerable state space. *Advances in Applied Probability* 10, 452-471.
- Feinberg EA & Shwartz A [1994]: Markov decision models with weighted discounted criteria. *Mathematics of Operations Research* 19, 152-168.
- Feinberg EA & Shwartz A [1995]: Constrained Markov decision models with weighted discounted rewards. *Mathematics of Operations Research* 20, 302-320.
- Filar JA [1981]: Ordered field property for stochastic games when the player who

- controls transitions changes from state to state. *Journal of Optimization Theory and Applications* 34, 503-515.
- Filar JA & Vrieze OJ [1992]: Weighted reward criteria in competitive Markov decision processes. *Zeitschrift für Operations Research - Methods and models of operations research* 36, 343-358.
- Fink AM [1964]: Equilibrium in a stochastic  $n$ -person game. *Journal of Science of Hiroshima University, Series A-I* 28, 89-93.
- Flesch J, Perea y Monswé A [1997]: Repeated games with exogenous choice of information mechanism. Report.
- Flesch J, Thuijsman F & Vrieze OJ [1996]: Recursive repeated games with absorbing states. *Mathematics of Operations Research* 21, 1016-1022.
- Flesch J, Thuijsman F & Vrieze OJ [1997,I]: Cyclic Markov equilibria in a cubic game. *International Journal of Game Theory* 26, 303-314
- Flesch J, Thuijsman F & Vrieze OJ [1997,II]: Markov strategies are better than stationary strategies. Report M97-09.
- Flesch J, Thuijsman F & Vrieze OJ [1998,I]: Simplifying optimal strategies in stochastic games. *SIAM Journal of Control and Optimization* 36, No. 4, 1331-1347.
- Flesch J, Thuijsman F & Vrieze OJ [1998,II]: Almost stationary  $\varepsilon$ -equilibria in zero-sum stochastic games. *Journal of Optimization Theory and Applications* (to appear).
- Flesch J, Thuijsman F & Vrieze OJ [1998,III]: Average-discounted equilibria in stochastic games. *European Journal of Operational Research* (to appear).
- Flesch J, Thuijsman F & Vrieze OJ [1998,IV]: Improving and non-improving strategies in stochastic games (to appear).
- Gale D & Sherman S [1950]: Solutions of finite two-person games. In: Kuhn HW & Tucker AW (eds.), *Contributions to the theory of games I (Annals of Mathematical Studies* 24, Princeton University Press, Princeton, 37-49.
- 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, 179-187.
- Hoffman AJ & Karp RM [1966]: On nonterminating stochastic games. *Management Science* 12, 359-370.
- Hordijk A, Kallenberg LCM & Wanrooij GL [1983]: Semi-Markov strategies in stochastic games. *International Journal of Game Theory* 12, 81-89.
- Kakutani S [1941]: A generalization of Brouwer's fixed point theorem. *Duke Mathematical Journal* 8, 416-427.
- Kemeny J & Snell J [1960]: *Finite Markov chains*. Van Nostrand, Princeton.
- Kohlberg E [1974]: Repeated games with absorbing states. *Annals of Statistics* 2, 724-738.
- Kolmogorov A [1933]: *Grundbegriffe der wahrscheinlichkeitsrechnung*. Ergebnisse der Mathematik 2, no. 3, Springer Verlag, Berlin.

- Liggett TM & Lippman SA [1969]: Stochastic games with perfect information and time average payoff. *SIAM Review* 11, 604-607.
- Myerson RB [1978]: Refinements of the Nash equilibrium concept. *International Journal of Game Theory* 7, 73-80.
- Mertens JF & Neyman A [1981]: Stochastic games. *International Journal of Game Theory* 10, 53-66.
- Monash CA [1980]: *Stochastic games: the minimax theorem*. Ph.D. thesis, Harvard University, Cambridge, Massachusetts.
- von Neumann J [1928]: Zur theorie der gesellschaftsspiele. *Mathematische Annalen* 100, 295-320.
- Nowak AS & Raghavan TES [1991]: Positive stochastic games and a theorem of Ornstein. In: Raghavan TES, Ferguson TS, Vrieze OJ, Parthasarathy T (eds.), *Stochastic Games and Related Topics*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 127-134.
- Parthasarathy TES, Tijds SH & Vrieze OJ [1984]: Stochastic games with state independent transitions and separable rewards. In: Hammer G & Pallaschke (eds.), *Selected Topics in Operations Research and Mathematical Economics*, Springer Verlag, Berlin 262-271.
- Raghavan TES, Tijds SH & Vrieze OJ [1985]: On stochastic games with additive reward and transition structure. *Journal of Optimization Theory and Applications* 47, 451-464.
- Rogers PD [1969]: *Non-zerosum stochastic games*. Ph.D. thesis, Report ORC 69-8, Operations Research Center, University of California, Berkeley.
- Shapley LS [1953]: Stochastic games. *Proceedings of the National Academy of Sciences U.S.A.* 39, 1095-1100.
- Sobel MJ [1971]: Noncooperative stochastic games. *Annals of Mathematical Statistics* 42, 1930-1935.
- Solan E [1998]: *Stochastic games*. Ph.D. thesis, Hebrew University, Jerusalem.
- Sorin S [1986]: Asymptotic properties of a non-zerosum game. *International Journal of Game Theory*, 15, 101-107.
- Schweitzer PJ [1968]: Perturbation theory and finite Markov chains. *Journal of Applied Probability*, 5, 401-41.
- Takahashi [1964]: Equilibrium points of stochastic noncooperative  $n$ -person games. *Journal of Science of Hiroshima University*, Series A-I 28, 95-99.
- Thuijsman F [1992]: *Optimality and Equilibria in Stochastic Games*. CWI-Tract 82, CWI, Amsterdam.
- Thuijsman F & Raghavan TES [1997]: Perfect information stochastic games and related classes. *International Journal of Game Theory*, 26, 403-408.
- Thuijsman F & Vrieze OJ [1992]: Note on recursive games. In: Dutta B et. al. (eds.): *Game theory and economic applications*, Lecture Notes in Economics and

Mathematical Systems, Springer Verlag, Berlin, 389, 133-145.

Thuijsman F & Vrieze OJ [1993]: Stationary  $\varepsilon$ -optimal strategies in stochastic games., *OR Spektrum*, Springer Verlag, 15, 9-15.

Thuijsman F & Vrieze OJ [1996]: The power of threats in stochastic games. In: Bardi et al. (eds.), *Stochastic Games and Numerical Methods for Dynamic Games*, Birkhauser, Boston (forthcoming).

Vieille N [1993]: Solvable states in stochastic games. *International Journal of Game Theory* 21, 395-404.

Vieille N [1994]: On equilibria in undiscounted stochastic games. D.P. 9446, Ceremade.

Vieille N [1997,I]: 2-person stochastic games I: A reduction, D.P. 9745, Ceremade.

Vieille N [1997,II]: 2-person stochastic games II: The case of recursive games, D.P. 9747, Ceremade.

Vrieze OJ & Thuijsman F [1989]: On equilibria in repeated games with absorbing states. *International Journal of Game Theory* 18, 293-310.