Multi-Agent Reinforcement Learning Algorithms

Natalia Akchurina

Dissertation
in Computer Science

submitted to the

Faculty of Electrical Engineering,
Computer Science and Mathematics

University of Paderborn

in partial fulfillment of the requirements for the degree of

doctor rerum naturalium
(Dr. rer. nat.)

Paderborn, February 2010

Abstract

Multi-agent reinforcement learning is an extension of reinforcement learning concept to multi-agent environments. Reinforcement learning allows to program agents by reward and punishment without specifying how to achieve the task. Formally agent-environment interaction in multi-agent reinforcement learning is presented as a discounted stochastic game. The task the agents are facing is formalized as the problem of finding Nash equilibria.

This thesis is devoted to development of multi-agent reinforcement learning algorithms. We propose an algorithm converging in self-play to Nash equilibria for high percentage of general-sum discounted stochastic games. The approach is based on generalization of replicator dynamics for discounted stochastic games. Before there was no algorithm that converged to Nash equilibria for general-sum discounted stochastic games (only for particular cases). The approach also outperforms the methods for solving general-sum discounted stochastic games: nonlinear optimization and stochastic tracing procedure. These algorithms function under the assumption that the games are known from the very beginning in contrast to reinforcement learning where the agent's task is to learn an optimal behavior in unknown environment. Another contribution is an algorithm that always converges to stationary policies and to best-response strategies against stationary opponents. Unlike the first algorithm it doesn't require that the opponents' rewards are observable. We give theoretical foundations for the convergence of the algorithms proposed in this thesis.

The possible application areas include traditional reinforcement learning tasks in multi-agent environments like robot soccer and development of trading agents along with numerous economic problems as a rule modeled as differential games in the field of capital accumulation, advertising, pricing, macroeconomics, warfare and resource economics. We propose to approximate the differential games with stochastic games and apply the developed solver.

Acknowledgements

It is a pleasure to express my deepest gratitude to all the people who supported me during my PhD studies. First of all, I would like to thank my scientific supervisor, Professor Hans Kleine Büning for his assistance and guidance at the University of Paderborn. The discussions we had were very fruitful. I admire his ability at once to see the core of the problem and to pinpoint the drawbacks of the solution. I am very grateful to Professor Leena Suhl, Professor Burkhard Monien and Professor Michael Dellnitz for their interest in my work.

I feel much obliged to the former and current members of the group *Knowledge-Based Systems*. I would like to thank Isabela Anciutti, Heinrich Balzer, Dr. Andreas Goebels, Dr. Elina Hotman, Thomas Kemmerich, Dr. Oliver Kramer, Christina Meyer, Dr. Steffen Priesterjahn, Professor Benno Stein, Alexander Weimer and Yuhan Yan. In particular I would like to thank Dr. Theodor Lettmann and Uwe Bubeck.

My sincere thanks go to all the people who were so friendly and made my stay in Germany very pleasant, especially to Markus Eberling and Patrizia Höfer. Simone Auinger, Astrid Canisius, Gerd Brakhane and Professor Eckhard Steffen were always very helpful in organizational and technical matters.

Last but not least, I thank my mother, Dr. Tatiana Yagodkina, for initializing my Q function correctly — so correctly that after 27 years of parallel exploration of the environment I am still sticking to the old strategy. I owe to her a lot of thanks for rash words she never failed to find whenever my strategy seemed to get biased.

Paderborn, February 2010 *Natalia Akchurina*

Contents

1

Introduction

1.1 Motivation

Reinforcement learning turned out a technique that allowed robots to ride bicycles, computers to play backgammon on the level of human world masters and solve such complicated tasks of high dimensionality as elevator dispatching. Can it come to rescue in the next generation of challenging problems like playing football or bidding at virtual markets? Reinforcement learning that provides a way of programming agents without specifying how the task is to be achieved could be again of use here but the convergence of reinforcement learning algorithms to optimal policies is only guaranteed under the conditions of stationarity of the environment that are violated in multi-agent systems. For reinforcement learning in multi-agent environments general-sum discounted stochastic games become a formal framework instead of Markov decision processes. Also the optimal policy concept in multi-agent systems is different — we can't speak anymore about optimal policy (policy that provides the maximum cumulative reward) without taking into account the policies of other agents that influence our payoffs. In the environment where every agent tries to maximize its cumulative reward it is the most natural to accept Nash equilibrium as the optimal solution concept. In Nash equilibrium each agent's policy is the best-response to the other agents' policies. Thus no agent can gain from unilateral deviation.

A number of algorithms were proposed to extend reinforcement learning approach to multi-agent systems. When the model (general-sum discounted stochastic game) is known two approaches: nonlinear optimization and stochastic tracing procedure are proved to find Nash equilibrium in the general case. In case of unknown model, the convergence to Nash equilibria was proved for very restricted class of environments: strictly competitive, strictly cooperative and 2-agent 2-action iterative game. Nash-Q algorithm has achieved convergence to Nash equilibrium in self-play for strictly competitive and strictly cooperative games under additional very restrictive condition that all equilibria encountered during learning stage are unique.

The main contribution of this thesis is an approach that allows to calculate Nash equilibria of general-sum discounted stochastic games with a given accuracy. We claim that it is the first approach that finds Nash equilibrium for the general case when the model is unknown. The experiments have shown that with the use of our approach much higher percentage of general-sum discounted stochastic games could be solved when the model is known. Its convergence to Nash equilibria is formally proved under certain assumptions.

The application areas of the developed approaches include traditional reinforcement learning tasks in multi-agent environments like robot soccer, development of trading agents along with many economic problems in the field of capital accumulation, advertising, pricing, macroeconomics, warfare and resource economics.

1.2 Organization of the Thesis

Chapter II

In chapter II we introduce reinforcement learning concept for one- and multi-agent environments along with formal definitions of Markov decision process, general-sum discounted stochastic game, Nash equilibrium, etc. The difficulty of computing a Nash equilibrium for multi-agent reinforcement learning is examined in detail. The overview of the proposed methods to solve this problem is given.

Chapter III

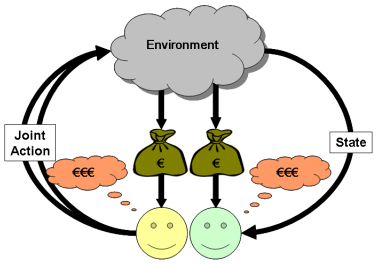
Theoretical basis for Nash-RD approach (the main contribution of the thesis) is developed.

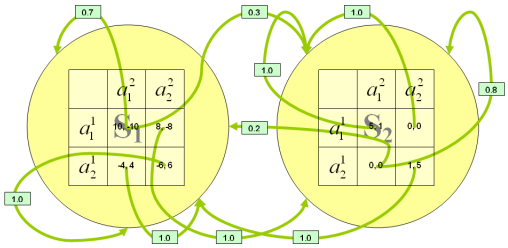
Chapter IV

In chapter IV we develop the algorithms based on Nash-RD approach both for the cases when the games are known and when the games are being learned. The algorithms are compared with the existing methods. The computational complexity is examined theoretically as well as experimentally and the reasons for an unexpected success of the proposed approach are listed.

Chapter V

In this chapter we analyze the existing multi-agent reinforcement learning algorithms from decision making perspective. Algorithm based on variable Hurwicz's optimistic-pessimistic criterion for choosing the best strategy in games against nature is developed. We formally prove the convergence of the algorithm to stationary policies.





(a) *Chapter II*

Theorem 3.14. *From* $1 \Rightarrow 2$

1. *For each state* $s \in S$*, the vector* $(x_s^1, x_s^2, \ldots, x_s^n)$ *constitutes an* ϵ*-equilibrium in the n-matrix game* $(B_s^1, B_s^2, \ldots, B_s^n)$ *with equilibrium payoffs* $(v_s^1, v_s^2, \ldots, v_s^n)$*, where for* $k \in K$ *and* $(a^1, a^2, \ldots, a^n) \in A^1 \times A^2 \times \ldots \times A^n$ *entry* $(a^1, a^2, \ldots, a^n)$ *of* B_s^k *equals*

$$b^k(s, a^1, a^2, \ldots, a^n) = r^k(s, a^1, a^2, \ldots, a^n) + \gamma \sum_{s' \in S} (p(s'|s, a^1, a^2, \ldots, a^n) + \varsigma(s'|s, a^1, a^2, \ldots, a^n))(v_{s'}^k + \sigma_{s'}^k)$$

where $-\sigma < \sigma_s^k < \sigma$*,* $-\varsigma < \varsigma(s'|s, a^1, a^2, \ldots, a^n) < \varsigma$ *for all* $s' \in S$*.*

2. x *is an* ε*-equilibrium in the discounted stochastic game* Γ *and*

$$v^k(x) - \frac{\omega}{1-\gamma}\mathbf{1} < v^k < v^k(x) + \frac{\omega}{1-\gamma}\mathbf{1}$$

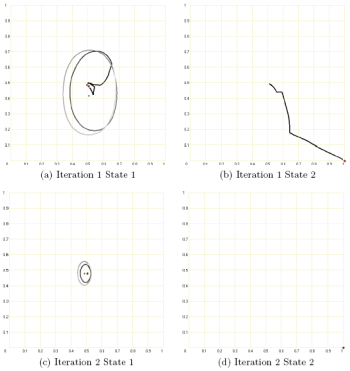
for all $k \in K$ *where*

$$\omega = \gamma\sigma + \gamma\varsigma N \max_{k \in K, s \in S} |v_s^k| + \gamma N \varsigma \sigma$$

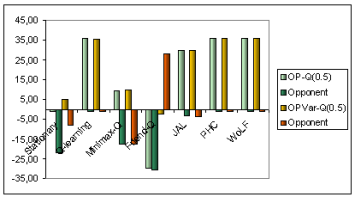
and

$$\varepsilon = \frac{2\omega + \epsilon}{1-\gamma}$$

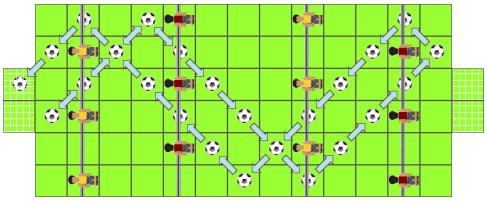
(b) *Chapter III*



(c) *Chapter IV*



(d) *Chapter V*



(e) *Chapter VI*

Fig. 1.1. Overview of the Thesis

Chapter VI

The applications of Nash-RD approach to chocolate duopoly, table soccer and double auctions are studied in detail. The potential of stochastic game representation of economic problems in the field of capital accumulation, advertising, pricing, macroeconomics, warfare and resource economics that are traditionally represented as differential games is investigated.

Chapter VII

The chapter presents final conclusions and the ideas for future work in this area.

Appendix A

The appendix is devoted to introduction of theory of ordinary differential equations that can be useful for analysis of some proofs in chapter III.

Appendix B

In this appendix we enumerate the transitions for stochastic game model of table soccer — one of applications analyzed in chapter VI.

Part I

Foundations of Reinforcement Learning

2

Reinforcement Learning

This chapter presents the foundations of reinforcement learning in one- and multi-agent environments. It is organized as follows. In section 2.1 we introduce reinforcement learning concept for one-agent environment and formal definitions of Markov decision process and optimal policy. In section 2.2 we extend reinforcement learning idea to multi-agent environment as well as recall some definitions from game theory such as discounted stochastic game, Nash equilibrium, etc. In section 2.3.1 the difficulty of calculating Nash equilibria for general-sum discounted stochastic games is considered in detail. In sections 2.3.2 and 2.3.3 the existing approaches to this problem are presented for the cases when the corresponding games are known at the beginning and are being learned by interaction with the environment[1].

2.1 Reinforcement Learning in One-Agent Environments

Inspired by related research in animal psychology [151], [72], reinforcement learning in computer science is a technique of programming agents by reward and punishment without specifying how the task is to be achieved[2]. The goal of the agent is to accumulate as much reward as possible by interacting with the environment. In ideal the agent must learn a behavior that will maximize its expected cumulative reward over a long run from indirect, delayed rewards.

Agent-environment interaction in reinforcement learning is presented in figure 2.1. Agent and environment interact at discrete time steps $t = 0, 1, 2, \ldots$ Agent observes the current state s_t at step t, chooses an action a_t and the environment provides it with a reward r_t (feedback) that reflects how well the agent is functioning in the environment and changes its state to s_{t+1} (in general non-deterministically).

[1] Here we mean not only model-based methods that literally learn the games, but also reinforcement learning methods that directly learn optimal behavior by interaction with the environment.

[2] In contrast to supervised learning [111], [12], [157], [11].

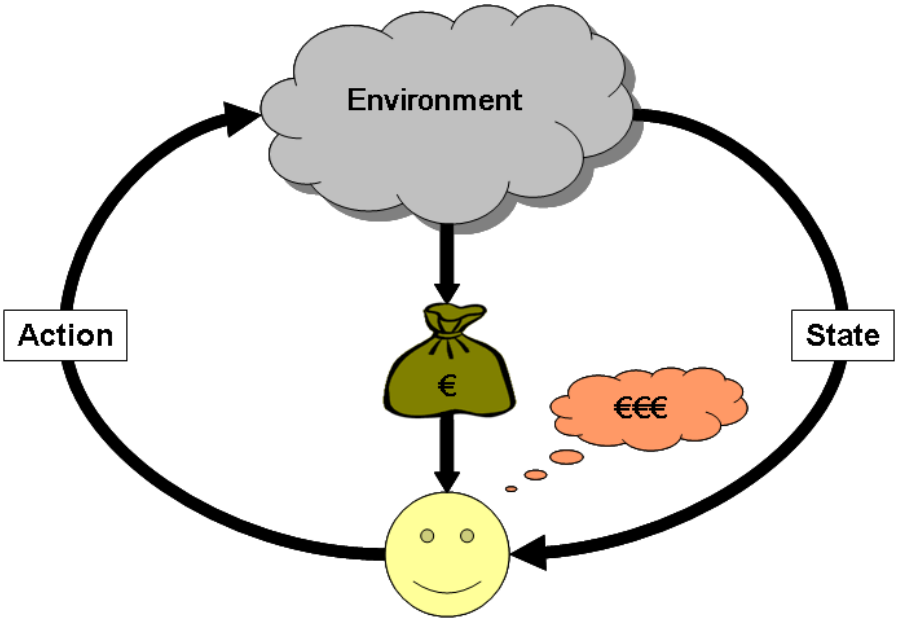


Fig. 2.1. Agent-Environment Interaction in One-Agent Environments

The task of the agent is to learn a policy that maps the states of the environment to the actions the agent should take to maximize its cumulative reward.

In spite of the simplicity of the idea, reinforcement learning has a number of very successful applications:

1. **Board Games:** At the time when Deep Blue beat the reigning World Chess Champion, Garry Kasparov principally due to its computational power [76], [116], TD-Gammon became a world-class backgammon player owing to high efficiency of reinforcement learning. The program used backgammon configuration function approximation based on neural network, trained against itself [148], [149].
2. **Learning to Ride a Bicycle:** To balance on a bicycle and to ride it are tasks that it is difficult to solve by traditional methods. With the use of reinforcement learning they have been solved [124]. After falling down a lot, a robot provided only with the information about the state of the bicycle, its performance (already on the ground or not yet), able to apply torque to handle bars and displace its center of mass, itself with the use of reinforcement learning has managed to balance on the bicycle and to ride it.
3. **Elevator Dispatching** is a task of high dimensionality that can't be solved by traditional methods [43]. But it has been solved with the use of reinforcement learning methods [44], [43]!

As a rule, the environment is formulated as a Markov decision process (MDP).

Definition 2.1. *A Markov decision process is a tuple* $\langle S, A, \gamma, r, p \rangle$*, where* $S = \{s_1, s_2, \ldots, s_N\}$ *is the discrete finite state space,* $A = \{a_1, a_2, \ldots, a_m\}$ *is the discrete finite action space,* $\gamma \in [0, 1)$ *is the discount factor,* $r : S \times A \to \mathbb{R}$ *is the reward function of the agent, and* $p : S \times A \to \Delta$ *is the transition function, where* Δ *is the set of probability distributions over state space* S*.*

Discount factor $\gamma \in [0, 1)$ reflects the notion that rewards depreciate by factor $\gamma < 1$ every time unit.

It is assumed that for every $s, s' \in S$ and for every action $a \in A$, transition probabilities $p(s'|s, a)$ are stationary for all $t = 0, 1, 2, \ldots$ and

$$\sum_{s' \in S} p(s'|s, a) = 1$$

Transition probabilities must satisfy Markov property — they should depend only on state and action chosen and be independent of the history:

$$p(s_{t+1} = s'|s_t, a_t, r_{t-1}, s_{t-1}, a_{t-1}, \ldots, r_0, s_0, a_0) = p(s_{t+1} = s'|s_t, a_t)$$

Example 2.2. In figure 2.2 a Markov decision process is presented. The agent can choose between two actions in each state s_1 and s_2. When the agent chooses the first action in state s_1 it will get 10 as immediate reward and the environment will change to state s_2 with probability 0.8 and stay at state s_1 with probability 0.2.

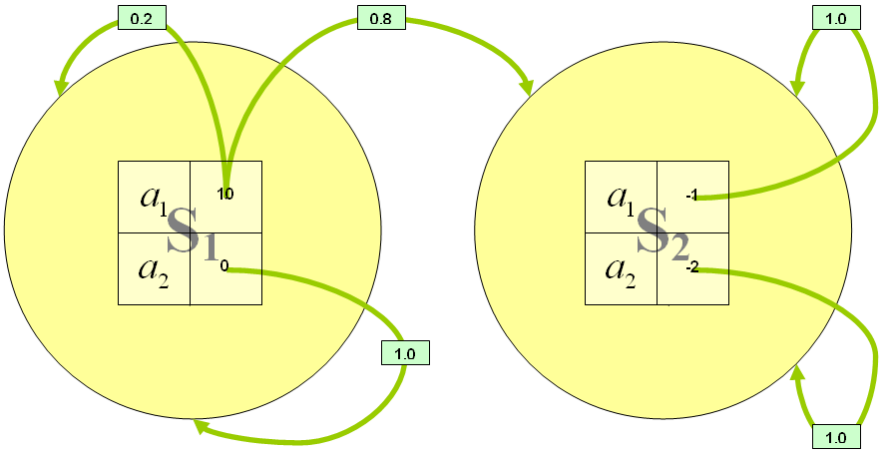


Fig. 2.2. A Markov Decision Process

A policy π is a mapping from each state $s \in S$ to action $a \in A$, the agent should take under this policy.

Accumulating as much reward as possible being the goal of the agent, the quality of the policy can be measured by its value.

Definition 2.3. *The value of policy π in state s, denoted $V^\pi(s)$ is the expected discounted cumulative reward the agent gets when starting in s and following π thereafter:*

$$V^\pi(s) = \sum_{k=0}^{\infty} \gamma^k \mathrm{E}\left[r_{t+k} | \pi, s_t = s\right]$$

Definition 2.4. *The value of taking action a in state s and then following policy π, denoted $Q^\pi(s,a)$ is the expected discounted cumulative reward the agent gets when starting in s, taking action a, and following π thereafter:*

$$Q^\pi(s,a) = \sum_{k=0}^{\infty} \gamma^k \mathrm{E}\left[r_{t+k} | \pi, s_t = s, a_t = a\right]$$

Definition 2.5. *Optimal policy*

$$\pi^* \equiv \arg\max_\pi \sum_{k=0}^{\infty} \gamma^k \mathrm{E}\left[r_{t+k} | \pi, s_t = s\right]$$

for every $s \in S$ [3].

Optimal policies share the same optimal state-value function $V^*(s)$ and the same optimal state-action-value function $Q^*(s,a)$:

$$V^*(s) = \max_\pi V^\pi(s)$$

$$Q^*(s,a) = \max_\pi Q^\pi(s,a)$$

For the state-action pair (s,a), function $Q^*(s,a)$ returns the expected discounted cumulative reward for taking action a in state s and thereafter following an optimal policy. Thus,

$$Q^*(s,a) = \mathrm{E}\left[r_t + \gamma V^*(s_{t+1}) | s_t = s, a_t = a\right]$$

Let us infer Bellman optimality equation for V^* [20]:

[3] According to [141], [117] such policy exists.

$$
\begin{aligned}
V^*(s) &= \max_{a \in A} Q^{\pi^*}(s,a) = \\
&= \max_{a \in A} \sum_{k=0}^{\infty} \gamma^k \mathrm{E}\left[r_{t+k} | \pi^*, s_t = s, a_t = a\right] = \\
&= \max_{a \in A} \mathrm{E}\left[r_t + \gamma \sum_{k=0}^{\infty} \gamma^k r_{t+1+k} | \pi^*, s_t = s, a_t = a\right] = \\
&= \max_{a \in A} \mathrm{E}\left[r_t + \gamma V^*(s_{t+1}) | s_t = s, a_t = a\right] = \\
&= \max_{a \in A} \sum_{s' \in S} p(s'|s,a) \left[r(s,a) + \gamma V^*(s')\right]
\end{aligned}
$$

The Bellman optimality equation for Q^* [141]:

$$
\begin{aligned}
Q^*(s,a) &= \mathrm{E}\left[r_t + \gamma \max_{a' \in A} Q^*(s_{t+1}, a') | s_t = s, a_t = a\right] = \\
&= \sum_{s' \in S} p(s'|s,a) \left[r(s,a) + \gamma \max_{a' \in A} Q^*(s', a')\right] = \\
&= r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V^*(s')
\end{aligned}
$$

$$
V^*(s) = \max_{a \in A} Q^*(s,a)
$$

There are a number of reinforcement learning algorithms that are proved to converge to optimal policies in one-agent environments [22], [23], [24], [163], [164], [47].

In presence of other agents we can't speak anymore about optimal policy (policy that provides the maximum cumulative reward) without taking into account the policies of other agents that influence our payoffs. Moreover, the convergence of reinforcement learning algorithms to optimal policies is only guaranteed under the conditions of stationarity of the environment that are violated in multi-agent systems. We need another framework for multi-agent environments that will explicitly take into account the impact of the other agents.

2.2 Reinforcement Learning in Multi-Agent Environments

In presence of several agents the environment in general case changes its state as a result of their joint action. The immediate payoffs are also the result of the joint action. Agent-environment interaction for multi-agent case is presented in figure 2.3.

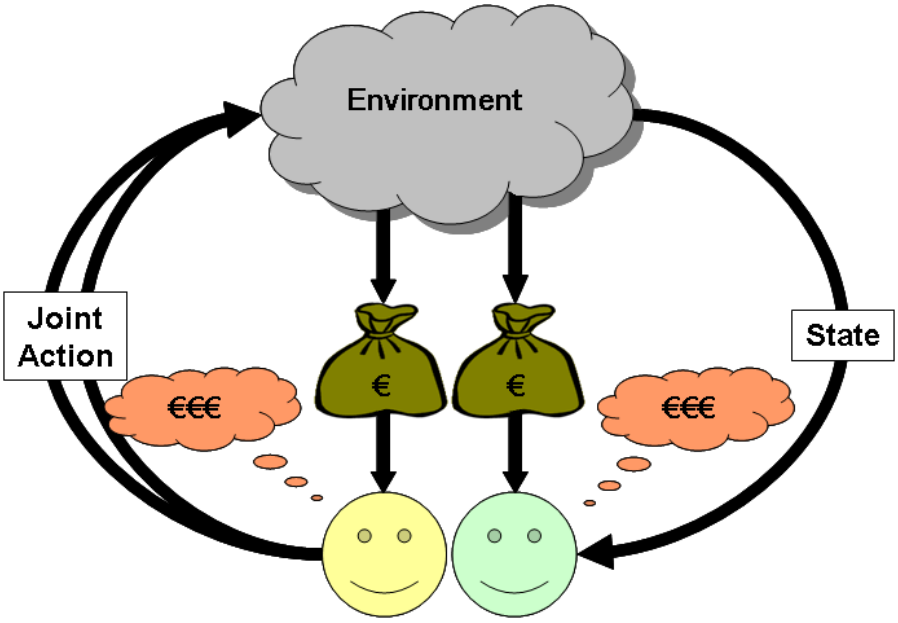


Fig. 2.3. Agent-Environment Interaction in Multi-Agent Environments

For reinforcement learning in multi-agent environments general-sum discounted stochastic games[4] become a formal framework instead of Markov decision processes. Before giving the formal definition we would like first to recall some fundamental concepts from game theory [119], [113].

Definition 2.6. *A pair of matrices* (M^1, M^2) *constitute a bimatrix game* G*, where* M^1 *and* M^2 *are of the same size. The rows of* M^k *correspond to actions of player* 1*,* $a^1 \in A^1$*. The columns of* M^k *correspond to actions of player* 2*,* $a^2 \in A^2$*.* $A^1 = \left\{a_1^1, a_2^1, \ldots, a_{m^1}^1\right\}$ *and* $A^2 = \left\{a_1^2, a_2^2, \ldots, a_{m^2}^2\right\}$ *are the finite sets of discrete actions of players* 1 *and* 2 *respectively. The payoff* $r^k(a^1, a^2)$ *to player* k *can be found in the corresponding entry of the matrix* M^k*,* $k = 1, 2$*.*

Definition 2.7. *A pure* ε*-equilibrium of bimatrix game* G *is a pair of actions* (a_*^1, a_*^2) *such that*

$$r^1(a_*^1, a_*^2) \geq r^1(a^1, a_*^2) - \varepsilon \; for \; all \; a^1 \in A^1$$

$$r^2(a_*^1, a_*^2) \geq r^2(a_*^1, a^2) - \varepsilon \; for \; all \; a^2 \in A^2$$

Definition 2.8. *A mixed* ε*-equilibrium of bimatrix game* G *is a pair of vectors* (ρ_*^1, ρ_*^2) *of probability distributions over action spaces* A^1 *and* A^2*, such that*

[4] Further on, we will use concepts *player* and *agent*, terms *strategy* and *policy* and *reward* and *payoff* interchangeably.

$$\rho_*^1 M^1 \rho_*^2 \geq \rho^1 M^1 \rho_*^2 - \varepsilon \; for \; all \; \rho^1 \in \sigma(A^1)$$

$$\rho_*^1 M^2 \rho_*^2 \geq \rho_*^1 M^2 \rho^2 - \varepsilon \; for \; all \; \rho^2 \in \sigma(A^2)$$

where $\sigma(A^k)$ is the set of probability distributions over action space A^k, such that for any $\rho^k \in \sigma(A^k)$, $\sum_{a \in A^k} \rho_a^k = 1$.

$$\rho^1 M^k \rho^2 = \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \rho_{a^1}^1 r^k(a^1, a^2) \rho_{a^2}^2 =$$

$$= \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} r^k(a^1, a^2) \prod_{i=1}^{2} \rho_{a^i}^i$$

is the expected reward of agent k induced by (ρ^1, ρ^2).

Let us denote the expected reward of agent k induced by (ρ^1, ρ^2) by $r^k(\rho^1, \rho^2)$:

$$r^k(\rho^1, \rho^2) = \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} r^k(a^1, a^2) \prod_{i=1}^{2} \rho_{a^i}^i$$

Definition 2.9. *Nash equilibrium of bimatrix game G is ε-equilibrium with $\varepsilon = 0$.*

Theorem 2.10. *[115] There exists a mixed Nash equilibrium for any bimatrix game.*

Definition 2.11. *A correlated equilibrium of bimatrix game G is a probability distribution ϱ_* over $A^1 \times A^2$, such that*

$$\sum_{a^2 \in A^2} \varrho_*(a^1, a^2) r^1(a^1, a^2) \geq \sum_{a^2 \in A^2} \varrho_*(a^1, a^2) r^1(\mathfrak{a}^1, a^2) \; for \; all \; \mathfrak{a}^1 \in A^1$$

$$\sum_{a^1 \in A^1} \varrho_*(a^1, a^2) r^2(a^1, a^2) \geq \sum_{a^1 \in A^1} \varrho_*(a^1, a^2) r^2(a^1, \mathfrak{a}^2) \; for \; all \; \mathfrak{a}^2 \in A^2$$

The interpretation of correlated equilibrium is the following: all players observe the same public random signal that prescribes to each of them an action (the agents can't get access to the signals of each other). If it is not profitable for any player to unilaterally deviate from the recommended action, the distribution is called a correlated equilibrium.

Example 2.12. In table 2.1 a bimatrix game is presented in a short form (the first payoffs of each entry correspond to payoff matrix M^1 of player 1 and the second ones — to M^2).

This game possesses two Nash equilibria in pure strategies $\left(a_1^1, a_1^2\right)$ and $\left(a_2^1, a_2^2\right)$ and one Nash equilibrium in mixed strategies $\left(\left(\frac{5}{6}, \frac{1}{6}\right), \left(\frac{1}{6}, \frac{5}{6}\right)\right)$. Apparently, no agent will gain from unilateral deviation.

A correlated equilibrium for this game is presented in table 2.2. Obviously, no agent will gain by ignoring the recommendations.

Table 2.1. A Bimatrix Game

	a_1^2	a_2^2
a_1^1	5,1	0,0
a_2^1	0,0	1,5

Table 2.2. A Correlated Equilibrium

Outcome	*Probability*
(a_1^1, a_1^2)	0.5
(a_1^1, a_2^2)	0
(a_2^1, a_1^2)	0
(a_2^1, a_2^2)	0.5

Definition 2.13. *An adversarial Nash equilibrium of bimatrix game G is a Nash equilibrium (ρ_*^1, ρ_*^2) such that*

$$r^1(\rho_*^1, \rho_*^2) \leq r^1(\rho_*^1, \rho^2) \; for \; all \; \rho^2 \in \sigma(A^2)$$

$$r^2(\rho_*^1, \rho_*^2) \leq r^2(\rho^1, \rho_*^2) \; for \; all \; \rho^1 \in \sigma(A^1)$$

where $\sigma(A^k)$ is the set of probability distributions over action space A^k, such that for any $\rho^k \in \sigma(A^k)$, $\sum_{a \in A^k} \rho_a^k = 1$.

Definition 2.14. *A coordination Nash equilibrium of bimatrix game G is a Nash equilibrium (ρ_*^1, ρ_*^2) such that*

$$r^1(\rho_*^1, \rho_*^2) = \max_{a^1 \in A^1, a^2 \in A^2} r^1(a^1, a^2)$$

$$r^2(\rho_*^1, \rho_*^2) = \max_{a^1 \in A^1, a^2 \in A^2} r^2(a^1, a^2)$$

Definition 2.15. *An n-player matrix game is a tuple $\langle K, A^1, \ldots, A^n, r^1, \ldots, r^n \rangle$, where $K = \{1, 2, \ldots, n\}$ is the player set, $A^k = \left\{a_1^k, a_2^k, \ldots, a_{m^k}^k\right\}$ is the finite discrete action space of player k for $k \in K$ ($|A^k| = m^k$) and $r^k : A^1 \times A^2 \times \ldots \times A^n \to \mathbb{R}$ is the reward function for player k.*

Definitions 2.7, 2.8, 2.9, 2.11, 2.13, 2.14 and theorem 2.10 can be generalized for arbitrary number of players.

Definition 2.16. *A 2-player discounted stochastic game Γ is a 7-tuple $\langle S, A^1, A^2, \gamma, r^1, r^2, p\rangle$, where $S = \{s_1, s_2, \ldots, s_N\}$ is discrete finite set of states ($|S| = N$), $A^k = \{a_1^k, a_2^k, \ldots, a_{m^k}^k\}$ is the discrete finite action space of player k for $k = 1, 2$ ($|A^k| = m^k$), $\gamma \in [0, 1)$ is the discount factor, $r^k : S \times A^1 \times A^2 \to \mathbb{R}$ is the reward function for player k bounded in absolute value by $R_{\max}$, $p : S \times A^1 \times A^2 \to \Delta$ is the transition probability map, where Δ is the set of probability distributions over state space S.*

Discount factor γ as in MDP reflects the notion that a reward at time $t+1$ is worth only $\gamma < 1$ of what it is worth at time t.

It is assumed that for every $s, s' \in S$ and for every action $a^1 \in A^1$ and $a^2 \in A^2$, transition probabilities $p(s'|s, a^1, a^2)$ are stationary for all $t = 0, 1, 2, \ldots$ and $\sum_{s' \in S} p(s'|s, a^1, a^2) = 1$.

Every state s of a 2-player stochastic game can be regarded as a bimatrix game $(R^1(s), R^2(s))$, where for $k = 1, 2$:

$$R^k(s) = \left[r^k(s, a^1, a^2)\right]_{a^1 \in A^1, a^2 \in A^2}$$

Example 2.17. In figure 2.4 a 2-agent stochastic game is presented. The agents can choose between two actions in each state s_1 and s_2. Their immediate payoffs are the result of their joint action. So when the first and the second agents choose the first actions in state s_1 the first agent will get 10 and the second agent -10 as immediate rewards and the environment will change to state s_2 with probability 0.3 and stay at state s_1 with probability 0.7.

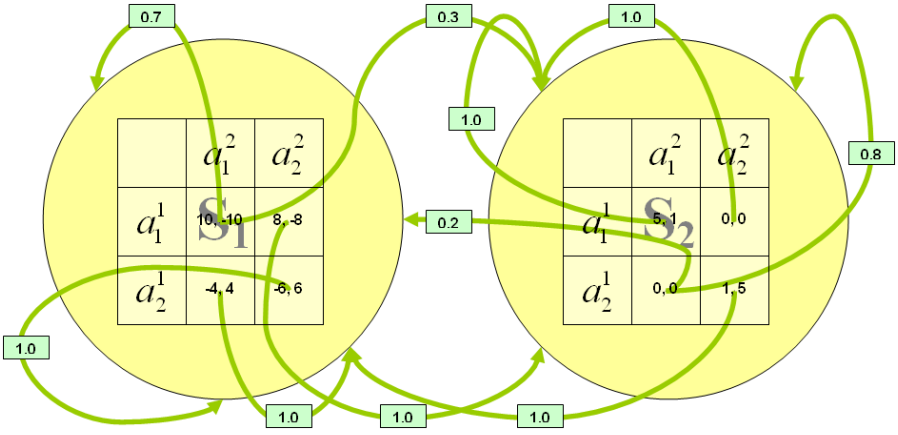


Fig. 2.4. A Stochastic Game

Definition 2.18. *One state discounted stochastic games are called repeated (iterative) games.*

Definition 2.19. *A 2-player discounted stochastic game Γ is called zero-sum when $r^1(s,a^1,a^2)+r^2(s,a^1,a^2)=0$ for all $s\in S$, $a^1\in A^1$ and $a^2\in A^2$, otherwise general-sum.*

Definition 2.20. *A 2-player discounted stochastic game Γ is called common-payoff game (or game of pure coordination) when $r^1(s,a^1,a^2)=r^2(s,a^1,a^2)$ for all $s\in S$, $a^1\in A^1$ and $a^2\in A^2$.*

Policy of agent $k=1,2$ is a vector $x^k=\left(x^k_{s_1},x^k_{s_2},\ldots,x^k_{s_N}\right)$, where $x^k_s=\left(x^k_{sa^k_1},x^k_{sa^k_2},\ldots,x^k_{sa^k_{m^k}}\right)$, $x^k_{sh}\in\mathbb{R}$ being the probability assigned by agent k to its action $h\in A^k$ in state s. A policy x^k is called a *stationary policy* if it is fixed over time. Since all probabilities are nonnegative and their sum is equal to one, the vector $x^k_s\in\mathbb{R}^{m^k}$ belongs to the unit simplex Δ^k:

$$\Delta^k=\left\{x^k_s\in\mathbb{R}^{m^k}_+:\sum_{a^k\in A^k}x^k_{sa^k}=1\right\}$$

.

The policy x^k will belong then to policy space of agent k[5]:

$$\Theta^k=\times_{s\in S}\Delta^k$$

Each player k ($k=1,2$) strives to learn policy maximizing its expected discounted cumulative reward:

$$v^k(s,x^1,x^2)=\sum_{i=0}^{\infty}\gamma^i\mathrm{E}\left[r^k_{t+i}|x^1,x^2,s_t=s\right]$$

where x^1 and x^2 are the policies of players 1 and 2 respectively and s is the initial state.

$v^k(s,x^1,x^2)$ is called the *discounted value* of policies (x^1,x^2) in state s to player k.

Definition 2.21. *An ε-equilibrium of 2-player discounted stochastic game Γ is a pair of policies (x^1_*,x^2_*) such that for all $s\in S$ and for all policies $x^1\in\Theta^1$ and $x^2\in\Theta^2$:*

$$v^1(s,x^1_*,x^2_*)\geq v^1(s,x^1,x^2_*)-\varepsilon$$
$$v^2(s,x^1_*,x^2_*)\geq v^2(s,x^1_*,x^2)-\varepsilon$$

Definition 2.22. *Nash equilibrium of 2-player discounted stochastic game Γ is ε-equilibrium with $\varepsilon=0$.*

[5] Policy space is defined as the Cartesian product of N unit simplexes Δ^k.

Definition 2.23. *An n-player discounted stochastic game is a tuple* $\langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, p \rangle$*, where* $K = \{1, 2, \ldots, n\}$ *is the player set,* $S = \{s_1, s_2, \ldots, s_N\}$ *is the discrete finite state space (*$|S| = N$*),* $A^k = \left\{a_1^k, a_2^k, \ldots, a_{m^k}^k\right\}$ *is the discrete finite action space of player* k *for* $k \in K$ *(*$|A^k| = m^k$*),* $\gamma \in [0, 1)$ *is the discount factor,* $r^k : S \times A^1 \times A^2 \times \ldots \times A^n \to \mathbb{R}$ *is the reward function for player* k *bounded in absolute value by* $R_{\max}$*,* $p : S \times A^1 \times A^2 \times \ldots \times A^n \to \Delta$ *is the transition probability map, where* Δ *is the set of probability distributions over state space* S*.*

Definitions 2.18, 2.19, 2.20, 2.21 and 2.22 can be generalized for n-player stochastic game.

Theorem 2.24. *[57] Every general-sum discounted stochastic game possesses at least one Nash equilibrium in stationary strategies.*

Definition 2.25. *A profile is a vector* $x = (x^1, x^2, \ldots, x^n)$*, where each component* $x^k \in \Theta^k$ *is a policy for player* $k \in K$*. The space of all profiles* $\Phi = \times_{k \in K} \Theta^k$*.*

Let's define the probability transition matrix induced by x:

$$p(s'|s, x) = \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \ldots \sum_{a^n \in A^n} p(s'|s, a^1, a^2, \ldots, a^n) \prod_{i=1}^{n} x_{sa^i}^i$$

$$P(x) = \left(p(s'|s, x)\right)_{s, s' \in S}$$

The immediate expected reward of player k in state s induced by x will be:

$$r^k(s, x) = \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \ldots \sum_{a^n \in A^n} r^k(s, a^1, a^2, \ldots, a^n) \prod_{i=1}^{n} x_{sa^i}^i$$

Then the immediate expected reward matrix induced by profile x will be:

$$r(x) = \left(r^k(s, x)\right)_{s \in S, k \in K}$$

The discounted value matrix of x will be [56]:

$$v(x) = \left[I - \gamma P(x)\right]^{-1} r(x)$$

where I is $N \times N$ identity matrix.

Note that the following recursive formula will hold for the discounted value matrix [56]:

$$v(x) = r(x) + \gamma P(x) v(x)$$

The kth columns of $r(x)$ and $v(x)$ (the immediate expected reward of player k induced by profile x and the discounted value of x to agent k) let us respectively denote $r^k(x)$ and $v^k(x)$.

Let

$$\Gamma = \langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, p \rangle$$

be an n-player discounted stochastic game.

Theorem 2.26. *[56]* $1 \Leftrightarrow 2$

1. *For each state $s \in S$, the vector $(x_s^1, x_s^2, \ldots, x_s^n)$ constitutes a Nash equilibrium in the n-matrix game $(B_s^1, B_s^2, \ldots, B_s^n)$ with equilibrium payoffs $(v_s^1, v_s^2, \ldots, v_s^n)$, where for $k \in K$ and $(a^1, a^2, \ldots, a^n) \in A^1 \times A^2 \times \ldots \times A^n$ entry $(a^1, a^2, \ldots, a^n)$ of B_s^k equals*

$$b^k(s, a^1, a^2, \ldots, a^n) = r^k(s, a^1, a^2, \ldots, a^n) + \gamma \sum_{s' \in S} p(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k$$

2. *x is an Nash equilibrium with discounted values $(v^1, v^2, \ldots, v^n)$ in the discounted stochastic game Γ.*

For arbitrary vectors $v^1, \ldots, v^n \in \mathbb{R}^N$, *Shapley's auxiliary matrix games* for $s \in S$ have the form $(R^1(s, v^1), \ldots, R^n(s, v^n))$, where

$$R^k(s, v^k) = R^k(s) + \gamma T(s, v^k)$$

and

$$T(s, v^k) = \left[\sum_{s' \in S} p(s'|s, a^1, \ldots, a^n) v^k(s') \right]_{a^1 \in A^1, \ldots, a^n \in A^n}$$

If we assume, that $v^1, \ldots, v^n \in \mathbb{R}^N$ are the values of Nash equilibrium, then finding Nash equilibrium policies for Shapley's auxiliary matrix games is exactly the problem we must solve at the current stage.

2.3 Multi-Agent Reinforcement Learning Algorithms

The general overview of reinforcement learning methods is given in figure 2.5[6]. The methods are divided in four classes:

- Reinforcement learning methods for one-agent environments:
  - when the corresponding Markov decision processes are known from the very beginning
  - when the corresponding Markov decision processes are being learned by interaction with the environment
- Reinforcement learning methods for multi-agent environments:

[6] Strictly speaking, methods that function under the assumption that the corresponding models are known from the very beginning are not reinforcement learning methods, though they are inseparably linked with and can be used for reinforcement learning.

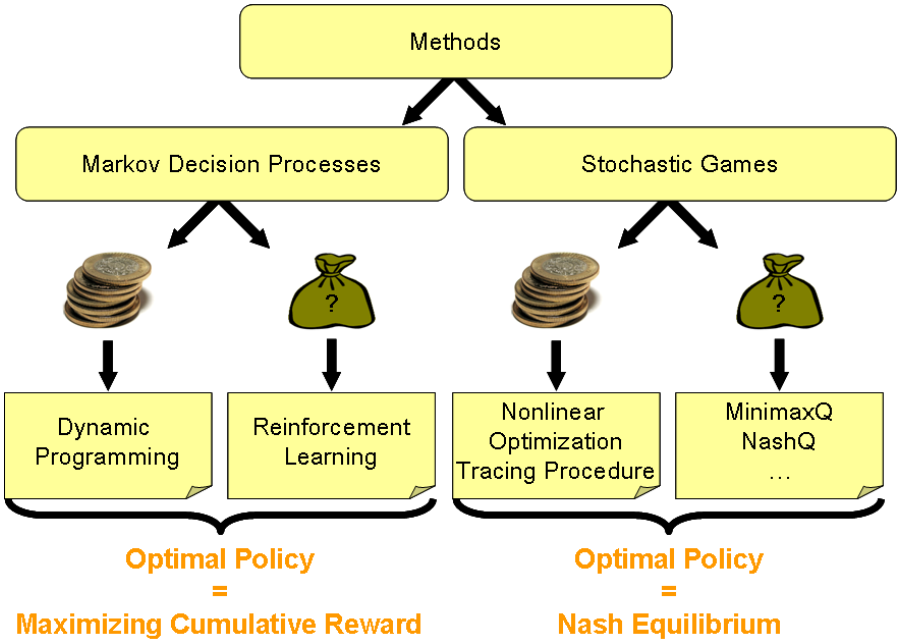


Fig. 2.5. Reinforcement Learning Methods

- when the corresponding discounted stochastic games are known from the very beginning
- when the corresponding discounted stochastic games are being learned by interaction with the environment

2.3.1 Problem of Calculating a Nash Equilibrium

In example 2.12 we have found Nash equilibria for a bimatrix game. For stochastic games (which are multi-stage games) the problem is much more complicated because we have to take into account not only immediate rewards but also cumulative discounted rewards that depend on policies in all the states. We must find Nash equilibrium simultaneously for all the states.

According to theorem 2.26 in order to find a Nash equilibrium for a stochastic game considered in example 2.17 we have to find policies that constitute Nash equilibria simultaneously for two bimatrix games presented in tables 2.3 and 2.4. The peculiarity of these bimatrix games that makes the problem especially complicated is that the payoffs in these games depend on the policies we are looking for.

Table 2.3. State 1

s_1	a_1^2	a_2^2
a_1^1	$10+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^1 \mid x^1, x^2, s_t = s_1)$, $-10+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^2 \mid x^1, x^2, s_t = s_1)$	$8+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^1 \mid x^1, x^2, s_t = s_1)$, $-8+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^2 \mid x^1, x^2, s_t = s_1)$
a_2^1	$-4+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^1 \mid x^1, x^2, s_t = s_1)$, $4+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^2 \mid x^1, x^2, s_t = s_1)$	$-6+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^1 \mid x^1, x^2, s_t = s_1)$, $6+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^2 \mid x^1, x^2, s_t = s_1)$

Table 2.4. State 2

s_2	a_1^2	a_2^2
a_1^1	$5+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^1 \mid x^1, x^2, s_t = s_2)$, $1+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^2 \mid x^1, x^2, s_t = s_2)$	$0+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^1 \mid x^1, x^2, s_t = s_2)$, $0+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^2 \mid x^1, x^2, s_t = s_2)$
a_2^1	$0+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^1 \mid x^1, x^2, s_t = s_2)$, $0+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^2 \mid x^1, x^2, s_t = s_2)$	$1+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^1 \mid x^1, x^2, s_t = s_2)$, $5+\sum_{i=1}^{\infty}\gamma^i \mathrm{E}(r_{t+i}^2 \mid x^1, x^2, s_t = s_2)$

2.3.2 When the Games are Known

Nonlinear Optimization

Let

$$z^T = ((v^1)^T, \ldots, (v^n)^T, x)$$

be a $(nN + N\sum_{k=1}^{n} m^k)$-dimensional vector of variables.

Let's consider the following nonlinear optimization problem:

$$\min\left\{\sum_{k\in K}\sum_{s\in S}\left[v_s^k - r_s^k(x) - \gamma\sum_{s'\in S} p(s'|s,x)v_{s'}^k\right]\right\}$$

subject to:

1.

$$\sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\cdots\sum_{a^n\in A^n}$$
$$R^k(s,v^k)[a^1,a^2,\ldots,a^{k-1},h,a^{k+1},\ldots,a^n]\prod_{i=1,i\neq k}^{n} x_{sa^i}^i \le v_s^k$$

$$s\in S,\ k\in K,\ h\in A^k$$

2.

$$x \in \Phi$$

Let's denote the objective function of the above nonlinear optimization problem by $\phi(z)$.

Theorem 2.27. *[55] The strategy part x of z forms a Nash equilibrium of the general-sum discounted stochastic game Γ if and only if z is the global minimum of the corresponding nonlinear optimization problem with $\phi(z) = 0$.*

Stochastic Tracing Procedure

Stochastic tracing procedure was proposed in [69], [71] as an extension of linear tracing procedure [65], [70] to the class of general-sum discounted stochastic games. It consists in tracing a path [170] from best-response strategies against certain expectations to an actual Nash equilibrium of discounted stochastic game. The mutual expectations are expressed by vector $z = (z^1, z^2, \ldots, z^n) \in \Phi$, where z^k is the policy the other agents believe player k is following, $k \in K$. The vector of expectations z is called a *prior*.

Let us consider an n-player discounted stochastic game

$$\Gamma = \langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, p \rangle$$

The stochastic tracing procedure will be based on a family of auxiliary games

$$\Gamma^{(t)} = \langle K, S, A^1, \ldots, A^n, \gamma, r^{1(t)}, \ldots, r^{n(t)}, p^{(t)} \rangle$$

$$0 \leq t \leq 1$$

such that for all $k \in K$:

$$r^{k(t)}(s, x) = tr^k(s, x) + (1 - t)r^k(s, z^1, \ldots, z^{k-1}, x^k, z^{k+1}, \ldots, z^n)$$

$$p^{(t)}(s'|s, x) = tp(s'|s, x) + (1 - t)p(s'|s, z^1, \ldots, z^{k-1}, x^k, z^{k+1}, \ldots, z^n)$$

The auxiliary game $\Gamma^{(0)}$ corresponds to a stochastic game, where all agents play according to the prior and decomposes into n separate Markov decision processes. The best-response policies of game $\Gamma^{(0)}$ will also constitute its Nash equilibria. The auxiliary game $\Gamma^{(1)}$ coincides with the original game Γ.

Let E^t denote the set of all Nash equilibria of $\Gamma^{(t)}$. According to theorem 2.24 E^t will be nonempty for all $0 \leq t \leq 1$.

Let $\mathcal{Q} = \mathcal{Q}(\Gamma, z)$ be the graph of the correspondence between t $(0 \leq t \leq 1)$ and the set of Nash equilibria E^t:

$$\mathcal{Q}(\Gamma, z) = \left\{(t, x) \in [0, 1] \times \Phi | x \text{ is a stationary Nash equilibrium of } \Gamma^t\right\}$$

A path $\mathcal{L}$ in graph $\mathcal{Q}$ connecting a point $q^0 = (0, x^0)$, where x^0 is a best-response against the prior z with a point $q^1 = (1, x_*)$, corresponding to a stationary Nash equilibrium of the stochastic game $\Gamma^{(1)} = \Gamma$ is called a *feasible path*. The existence of such path is proved in [69] for almost all general-sum discounted stochastic games.

The stochastic tracing procedure consists in finding a Nash equilibrium x_* for any general-sum discounted stochastic game Γ by following a feasible path $\mathcal{L}$ from its initial point $q^0 = (0, x^0)$ to its final point $q^1 = (1, x_*)$.

2.3.3 When the Games are Being Learned

Q-learning

One of the most important breakthroughs in reinforcement learning for one-agent environments was the development of Q-learning algorithm [163] (see algorithm 1). Before methods for solving MDPs were based on Bellman optimality equations applied simultaneously over all $s \in S$. Q-learning performs the updates asynchronously and doesn't need explicit knowledge of the transition probabilities p.

The essence of Q-learning algorithm, Q function is updated incrementally.

The general form for incremental update [141]:

$$NewEstimate \leftarrow OldEstimate + LearningRate\,[Target - OldEstimate]$$

where expression $[Target - OldEstimate]$ is the error of the estimate. It is reduced by $LearningRate$ in the direction indicated by the $Target$. In our case the target is $r + \gamma V(s')$ and the correction is performed by an agent each time it receives a reward r on taking action a and getting from s to s'. The probability with which this takes place is precisely $p(s'|s, a)$ and therefore the agent performs the appropriate update without explicit knowledge of p. Q function updated in this way has been shown to converge with probability 1 to the optimal action-value function Q^* under the following assumptions [164]:

1. Exploration: All state action pairs are updated infinitely often.
2. Learning rates $0 \leq \alpha_k < 1$[7]:
   a) $\sum_{k=1}^{\infty} \alpha_k = \infty$
   b) $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$

The condition on exploration can be satisfied in the following way — most of the time we choose an action a that has maximal estimated action value $Q(s, a)$, but with probability ϵ we instead select an action at random.

As we know from section 2.1 we can get an optimal policy from $Q^*(s, a)$ values using the following property: $Q^*(s, a)$ returns the greatest value for the

[7] In algorithm 1 the learning rate is decayed so as to satisfy the conditions.

Algorithm 1 Q-learning

Input: learning rate α, discount factor γ
for all $s \in S$ and $a \in A$ **do**
  $Q(s,a) \leftarrow 0$
  $V(s) \leftarrow 0$
end for
Observe the current state s
loop
  Choose action a for state s using policy π (with proper exploration)
  Take action a, observe reward r and succeeding state s' provided by the environment
  $Q(s,a) \leftarrow Q(s,a) + \alpha\left[r + \gamma V(s') - Q(s,a)\right]$
  $\pi(s) \leftarrow \arg\max_{\mathfrak{a}} Q(s,\mathfrak{a})$
  $V(s) \leftarrow Q(s,\pi(s))$
  decay α
  $s \leftarrow s'$
end loop

action a that should be taken in each particular state s so as to maximize expected discounted cumulative reward.

Q-learning is widely used in multi-agent environments [39], [145], [131] in spite of the fact that such its application can be theoretically justified only in case the opponents' strategies are stationary. In this case, we deal with MDPs with stationary transitions and Q-learning is bound to converge to the optimal policies.

Minimax-Q

Let

$$\Gamma = \langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, p \rangle$$

be n-player discounted stochastic game and let x_* denote a Nash equilibrium of Γ.

According to theorem 2.26:

$$v_s^k(x_*) = Nash_k(R^1(s, v^1(x_*)), R^2(s, v^2(x_*)), \ldots, R^n(s, v^n(x_*)))$$

where $(R^1(s, v^1(x_*)), R^2(s, v^2(x_*)), \ldots, R^n(s, v^n(x_*)))$ is Shapley's auxiliary matrix game and $Nash_k$ denotes the Nash equilibrium payoff to player k.

One can't help noticing a certain resemblance with Bellman optimality equation.

Minimax-Q [103] exploits this analogy, adopting the idea of asynchronous updates first proposed in Q-learning algorithm. It considers though only zero-sum discounted stochastic game case where finding Nash equilibrium very comfortably reduces to solving the following linear program (for player 1):

Algorithm 2 Minimax-Q for player 1

Input: learning rate α, discount factor γ
for all $s \in S$, $a^1 \in A^1$ and $a^2 \in A^2$ **do**
  $Q(s, a^1, a^2) \leftarrow 0$
  $V(s) \leftarrow 0$
  $\pi(s, a^1) \leftarrow 1/|A^1|$
end for
Observe the current state s
loop
  Choose action a^1 for state s using policy π (with proper exploration)
  Take action a^1, observe opponent's action a^2, reward r and succeeding state s' provided by the environment
  $Q(s, a^1, a^2) \leftarrow Q(s, a^1, a^2) + \alpha \left[r + \gamma V(s') - Q(s, a^1, a^2) \right]$
  $\pi(s, \cdot) \leftarrow \arg\max_{\pi'(s,\cdot)} \min_{\mathfrak{a}^2} \sum_{\mathfrak{a}^1} \pi'(s, \mathfrak{a}^1) Q(s, \mathfrak{a}^1, \mathfrak{a}^2)$
  $V(s) \leftarrow \min_{\mathfrak{a}^2} \sum_{\mathfrak{a}^1} \pi(s, \mathfrak{a}^1) Q(s, \mathfrak{a}^1, \mathfrak{a}^2)$
  decay α
  $s \leftarrow s'$
end loop

$$Nash_1 = \max_{\pi'(s,\cdot)} \min_{\mathfrak{a}^2} \sum_{\mathfrak{a}^1} \pi'(s, \mathfrak{a}^1) Q(s, \mathfrak{a}^1, \mathfrak{a}^2)$$

Minimax-Q's convergence to a Nash equilibrium in case of zero-sum discounted stochastic games was formally proved under similar assumptions as Q-learning in [107].

Joint-Action Learners

Joint-action learner (JAL) treats the past relative empirical frequencies of each opponents' action as the actual probability of corresponding action under opponents' current strategy. The idea is similar to fictitious play [32] and rational learning in game theory [59].

Nash-Q

Based on the same analogy as Minimax-Q, Nash-Q was introduced as an algorithm generalizing Q-learning to multi-agent case, capable of finding Nash equilibrium for arbitrary general-sum discounted stochastic game [88] under a certain condition.

Arbitrary stochastic game possessing in general case several equilibria, Nash function ($Nash$) doesn't return the same equilibrium to all the agents. That hinders the convergence in general case.

The assumption under which the algorithm is guaranteed to converge, turned out very restrictive [26], [104]: all equilibria encountered during learning stage must be unique or all be either of adversarial or coordination type. Besides zero-sum and common-payoff games, for which convergent learning

Algorithm 3 JAL for player k

Input: learning rate α, discount factor γ
for all $s \in S$, $a^1 \in A^1$, ..., $a^n \in A^n$ **do**
$Q(s, a^1, \ldots, a^n) \leftarrow 0$
$V(s) \leftarrow 0$
$\pi(s, a^k) \leftarrow 1/|A^k|$
$C(s) \leftarrow 0$
$n(s) \leftarrow 0$
end for
Observe the current state s
loop
Choose action a^k for state s using policy π (with proper exploration)
Take action a^k, observe other agents' actions $a^1, \ldots, a^{k-1}, a^{k+1}, \ldots, a^n$, reward r and succeeding state s' provided by the environment
$Q(s, a^1, \ldots, a^n) \leftarrow Q(s, a^1, \ldots, a^n) + \alpha \left[r + \gamma V(s') - Q(s, a^1, \ldots, a^n) \right]$
$C(s, a^1, \ldots, a^{k-1}, a^{k+1}, \ldots, a^n) \leftarrow C(s, a^1, \ldots, a^{k-1}, a^{k+1}, \ldots, a^n) + 1$
$n(s) \leftarrow n(s) + 1$
$\pi(s, \cdot) \leftarrow \arg\max_{\pi'(s,\cdot)} \sum_{a^1,\ldots,a^n} \pi'(s, a^k) \frac{C(s,a^1,\ldots,a^{k-1},a^{k+1},\ldots,a^n)}{n(s)}$.
$Q(s, a^1, \ldots, a^n)$
$V(s) = \sum_{a^1,\ldots,a^n} \pi(s, a^k) \frac{C(s,a^1,\ldots,a^{k-1},a^{k+1},\ldots,a^n)}{n(s)} Q(s, a^1, \ldots, a^n)$
decay α
$s \leftarrow s'$
end loop

algorithms already existed (see algorithms 2 and 3), no general-sum game is known to possess unique intermediate Nash equilibrium.

Nash-Q makes an additional assumption that the agents can observe other agents' immediate rewards and actions.

Friend- and Foe-Q

The assumption of the algorithm [104] reflected in the name is that the agents in the environment could be divided into two confronting groups having the opposite aims. The agents inside the groups are supposed to pursue the same goals.

Let us choose some agent k and let A^1 through A^l be the actions available to the l agents that belong to the same group as agent k — friends of player k and A^{l+1} through A^n be the actions available to its $n - l$ foes (agents with the opposite goal). Then the value of a state s to agent k is calculated as

$$V^k(s) = \max_{\pi^1(s,\cdot),\ldots,\pi^l(s,\cdot)} \min_{a^{l+1},\ldots,a^n} \sum_{a^1,\ldots,a^l} Q^k(s, a^1, \ldots, a^n) \prod_{j=1}^{l} \pi^j(s, a^j)$$

According to Friend- and Foe-Q's supposition, friends of agent k work together to maximize k's value, while k's foes cooperate to minimize k's value. Friend-

Algorithm 4 Nash-Q for player k

Input: learning rate α, discount factor γ
for all $s \in S$, $i \in K$, $a^1 \in A^1$, ..., $a^n \in A^n$ **do**
  $Q^i(s, a^1, \ldots, a^n) \leftarrow 0$
  $(\pi^1(s, \cdot), \ldots, \pi^n(s, \cdot)) \leftarrow Nash(Q^1(s), \ldots, Q^n(s))$
  $V^i(s) \leftarrow \sum_{a^1, \ldots, a^n} Q^i(s, a^1, \ldots, a^n) \prod_{j=1}^{n} \pi^j(s, a^j)$
end for
Observe the current state s
loop
  Choose action a^k for state s using policy π^k (with proper exploration)
  Take action a^k, observe other agents' actions $a^1, \ldots, a^{k-1}, a^{k+1}, \ldots, a^n$, reward r^k, other agents' rewards $r^1, \ldots, r^{k-1}, r^{k+1}, \ldots, r^n$ and succeeding state s' provided by the environment
  for all $i \in K$ **do**
    $Q^i(s, a^1, \ldots, a^n) \leftarrow Q^i(s, a^1, \ldots, a^n) + \alpha \left[r^i + \gamma V^i(s') - Q^i(s, a^1, \ldots, a^n) \right]$
  end for
  $(\pi^1(s, \cdot), \ldots, \pi^n(s, \cdot)) \leftarrow Nash(Q^1(s), \ldots, Q^n(s))$
  for all $i \in K$ **do**
    $V^i(s) \leftarrow \sum_{a^1, \ldots, a^n} Q^i(s, a^1, \ldots, a^n) \prod_{j=1}^{n} \pi^j(s, a^j)$
  end for
  decay α
  $s \leftarrow s'$
end loop

and Foe-Q regards any game as a two-player zero-sum game with an extended action set.

In case when only foes are present in the environment and the corresponding game has an adversarial equilibrium Friend- and Foe-Q learns values of a Nash equilibrium policy. In this case the algorithm learns a Q function whose corresponding policy will achieve at least the learned values, regardless of the opponents' selected policies. In case when only friends are present and the game has a coordination equilibrium Friend- and Foe-Q learns values of a Nash equilibrium policy. This is true regardless of the other agents' behavior. In presence of friends the algorithm might not get its learned value because of the possibility of incompatible coordination equilibria.

CE-Q

CE-Q [63] is based on calculating correlated equilibria [16], [60] (CE function in algorithm 6) instead of Nash equilibria. The set of correlated equilibria of general-sum matrix game contains its set of Nash equilibria. Thus CE-Q generalizes both Nash-Q and Minimax-Q.

CE-Q faces the same problem as Nash-Q — general-sum games could possess multiple equilibria with different payoff values. The algorithm deals with it by explicitly introducing four correlated equilibrium selection mechanisms.

Algorithm 5 Friend- and Foe-Q for player k

Input: learning rate α, discount factor γ
for all $s \in S$, $a^1 \in A^1$, ..., $a^n \in A^n$ **do**
$Q(s, a^1, \ldots, a^n) \leftarrow 0$
$(\pi^1(s,\cdot), \ldots, \pi^l(s,\cdot)) \leftarrow \arg\max_{\pi^1(s,\cdot),\ldots,\pi^l(s,\cdot)} \min_{a^{l+1},\ldots,a^n} \sum_{a^1,\ldots,a^l} Q(s, a^1, \ldots, a^n) \prod_{j=1}^{l} \pi^j(s, a^j)$
$V(s) \leftarrow \min_{a^{l+1},\ldots,a^n} \sum_{a^1,\ldots,a^l} Q(s, a^1, \ldots, a^n) \prod_{j=1}^{l} \pi^j(s, a^j)$
end for
Observe the current state s
loop
Choose action a^k for state s using policy π^k (with proper exploration)
Take action a^k, observe other agents' actions $a^1, \ldots, a^{k-1}, a^{k+1}, \ldots, a^n$, reward r^k and succeeding state s' provided by the environment
$Q(s, a^1, \ldots, a^n) \leftarrow Q(s, a^1, \ldots, a^n) + \alpha \left[r^k + \gamma V(s') - Q(s, a^1, \ldots, a^n) \right]$
$(\pi^1(s,\cdot), \ldots, \pi^l(s,\cdot)) \leftarrow \arg\max_{\pi^1(s,\cdot),\ldots,\pi^l(s,\cdot)} \min_{a^{l+1},\ldots,a^n} \sum_{a^1,\ldots,a^l} Q(s, a^1, \ldots, a^n) \prod_{j=1}^{l} \pi^j(s, a^j)$
$V(s) \leftarrow \min_{a^{l+1},\ldots,a^n} \sum_{a^1,\ldots,a^l} Q(s, a^1, \ldots, a^n) \prod_{j=1}^{l} \pi^j(s, a^j)$
decay α
$s \leftarrow s'$
end loop

In [63] empirical convergence to correlated equilibria is showed for all four CE-Q algorithms on a testbed of stochastic games.

Note, that in order to calculate correlated equilibrium the agents must possess copies of each other's Q-tables. As in Nash-Q this problem is resolved by assuming that the actions as well as rewards are observable in the environment so that the agents could follow changes in each others Q-tables.

Hyper-Q

Hyper-Q was introduced as a more general and as a more practical extension of Q-learning algorithm to multi-agent systems aiming at avoiding the following restricting assumptions:

- the agents get the information on each other's rewards
- the agents are based on the same learning principle

In Hyper-Q Q-function is used to evaluate entire mixed strategies[8], rather than joint actions, and estimates of the other agents' mixed policies are considered to be part of the states. These estimates are obtained from observation.

Since opponents' learning forms are unrestricted, it is unrealistic to expect Hyper-Q algorithm to converge in general case. Hyper-Q will converge to best-

[8] Obviously, the choice of function approximation scheme [112], [136], [156] is crucial for efficiency of Hyper-Q algorithm.

Algorithm 6 CE-Q for player k

Input: learning rate α, discount factor γ
for all $s \in S$, $i \in K$, $a^1 \in A^1$, ..., $a^n \in A^n$ **do**
  $Q^i(s, a^1, \ldots, a^n) \leftarrow 0$
  $\pi(s) \leftarrow CE(Q^1(s), \ldots, Q^n(s))$
  $V^i(s) \leftarrow \sum_{a^1,\ldots,a^n} Q^i(s, a^1, \ldots, a^n)\pi(s, a^1, \ldots, a^n)$
end for
Observe the current state s
loop
  Choose action a^k for state s according to the signal (with proper exploration)
  Take action a^k, observe other agents' actions $a^1, \ldots, a^{k-1}, a^{k+1}, \ldots, a^n$, reward r^k, other agents' rewards $r^1, \ldots, r^{k-1}, r^{k+1}, \ldots, r^n$ and succeeding state s' provided by the environment
  for all $i \in K$ **do**
    $Q^i(s, a^1, \ldots, a^n) \leftarrow Q^i(s, a^1, \ldots, a^n) + \alpha \left[r^i + \gamma V^i(s') - Q^i(s, a^1, \ldots, a^n)\right]$
  end for
  $\pi(s) \leftarrow CE(Q^1(s), \ldots, Q^n(s))$
  for all $i \in K$ **do**
    $V^i(s) \leftarrow \sum_{a^1,\ldots,a^n} Q^i(s, a^1, \ldots, a^n)\pi(s, a^1, \ldots, a^n)$
  end for
  decay α
  $s \leftarrow s'$
end loop

response policies facing the opponents following stationary strategies. In this case we deal again with MDPs with stationary transitions.

Policy Hill Climbing

Policy hill climbing algorithm (PHC) [28][9] (see algorithm 8[10]) is an extension of Q-learning that is capable of learning mixed strategies. The only difference is that PHC corrects its policy gradually by δ in the direction of the action that yields the best value.

On facing agents that stick to some stationary strategies, this algorithm, like Q-learning, will converge to an optimal policy. But despite its ability to learn mixed policies, it doesn't converge to a Nash equilibrium in general case.

WoLF Policy Hill Climbing

In [28] a specific principle (WoLF, "Win or Learn Fast") for varying the learning rate is introduced (see algorithm 9[11]). The idea is to learn fast when

[9] The algorithm executes hill-climbing in the space of mixed policies.
[10] We preserved the original representation though the policy updating steps could be definitely formalized more elegantly.
[11] We preserved the original representation though the policy updating steps could be definitely formalized more elegantly.

Algorithm 7 Hyper-Q for player k

Input: learning rates α, μ, discount factor γ
for all $s \in S$, $i \in K$, $a^1 \in A^1$, ..., $a^n \in A^n$ **do**
  $Q(s, \pi^1, \ldots, \pi^n) \leftarrow 0$
  $\pi^i(s, a^i) \leftarrow 1/|A^i|$
  $V(s, \pi^1, \ldots, \pi^{k-1}, \pi^{k+1}, \ldots, \pi^n) \leftarrow Q(s, \pi^1, \ldots, \pi^n)$
end for
Observe the current state s
loop
  Choose action a^k for state s using policy π^k (with proper exploration)
  Take action a^k, observe other agents' actions $a^1, \ldots, a^{k-1}, a^{k+1}, \ldots, a^n$, reward r and succeeding state s' provided by the environment
  for all $i = 1$ to n, $i \neq k$ **do**
    for all $\mathfrak{a} \in A^i$ **do**
      $$\pi^{i\prime}(s, \mathfrak{a}) \leftarrow \begin{cases} (1-\mu)\pi^i(s, \mathfrak{a}) + \mu & \mathfrak{a} = a^i \\ (1-\mu)\pi^i(s, \mathfrak{a}) & \text{otherwise} \end{cases}$$
    end for
  end for
  $Q(s, \pi^1, \ldots, \pi^n) \leftarrow Q(s, \pi^1, \ldots, \pi^n) + \alpha \left[r + \gamma V(s', \pi^{1\prime}, \ldots, \pi^{k-1\prime}, \pi^{k+1\prime}, \ldots, \pi^{n\prime}) - Q(s, \pi^1, \ldots, \pi^n) \right]$
  $\pi^k(s, \cdot) \leftarrow \arg\max_{\pi^{k\prime}(s,\cdot)} Q(s, \pi^1, \ldots, \pi^{k-1}, \pi^{k\prime}, \pi^{k+1}, \ldots, \pi^n)$
  $V(s, \pi^1, \ldots, \pi^{k-1}, \pi^{k+1}, \ldots, \pi^n) \leftarrow Q(s, \pi^1, \ldots, \pi^n)$
  decay α
  for all $i = 1$ to n, $i \neq k$ **do**
    $\pi^i \leftarrow \pi^{i\prime}$
  end for
  $s \leftarrow s'$
end loop

losing and with caution when winning. Theoretical analysis was done under the assumption that the agent determines whether it is winning or losing by comparing its expected payoff with the reward it would get if it followed some Nash equilibrium policy. The convergence to Nash equilibrium was only proved in self-play for very restricted class of environments: 2-agent 2-action iterative game.

WoLF-PHC requires two learning rates $\delta_l > \delta_w$. Which parameter to use to update the policy is determined by comparing the expected payoff of the current policy with the expected payoff of the average policy. If the latter is larger, the agent is losing and δ_l is used, otherwise δ_w. In algorithm WoLF-PHC the average policy is taken instead of the unknown equilibrium policy. It is well-founded since in many games average policies do approximate the Nash equilibrium (e.g., in fictitious play [161]).

Algorithm 8 PHC for player k

Input: learning rates α, δ, discount factor γ
for all $s \in S$ and $a^k \in A^k$ **do**
  $Q(s, a^k) \leftarrow 0$
  $V(s) \leftarrow 0$
  $\pi(s, a^k) \leftarrow 1/|A^k|$
end for
Observe the current state s
loop
  Choose action a for state s using policy π (with proper exploration)
  Take action a, observe reward r and succeeding state s' provided by the environment
  $Q(s, a) \leftarrow Q(s, a) + \alpha\left[r + \gamma V(s') - Q(s, a)\right]$
  for all $\mathfrak{a} \in A^k$ **do**
    $\pi(s, \mathfrak{a}) \leftarrow \pi(s, \mathfrak{a}) + \begin{cases} \delta & \text{if } \mathfrak{a} = \arg\max_{\mathfrak{a}'} Q(s, \mathfrak{a}') \\ \frac{-\delta}{|A^k|-1} & \text{otherwise} \end{cases}$
  end for
  Constrain $\pi(s, \cdot)$ to a legal probability distribution
  $V(s) \leftarrow \max_{\mathfrak{a}} Q(s, \mathfrak{a})$
  decay α
  $s \leftarrow s'$
end loop

Other Works

As a result of futile effort to develop an algorithm converging to Nash equilibrium for multi-agent environments in general case, a number of works have appeared [133], [122], [73] that reconsidered the agendas and reasoned in the following vein: "If we are not able to find Nash equilibrium, then we don't need it".

A row of papers proposing new criteria for estimating multi-agent reinforcement learning algorithms have followed [28], [150], [40], [122], [123], [167], [27], [38]. Besides dwelling on the merits of their criteria, each researcher also introduced an algorithm satisfying them.

The two criteria proposed in [28] are:

- *Rationality*: "If the other players' policies converge to stationary policies then the learning algorithm will converge to a policy that is a best-response to the other players' policies."
- *Convergence*: "The learner will necessarily converge to a stationary policy against agents using an algorithm from some class of learning algorithms."

Then they introduce an algorithm WoLF-IGA based on infinitesimal gradient ascent [135] with varying learning rate and prove that it meets the above criteria under the following conditions:

- the environment can be represented as a 2-player 2-action repeated game with known rewards

Algorithm 9 WoLF-PHC for player k

Input: learning rates α, δ_l, δ_w, discount factor γ
for all $s \in S$ and $a^k \in A^k$ **do**
  $Q(s, a^k) \leftarrow 0$
  $V(s) \leftarrow 0$
  $\pi(s, a^k) \leftarrow 1/|A^k|$
  $C(s) \leftarrow 0$
end for
loop
  Choose action a for state s using policy π (with proper exploration)
  Take action a, observe reward r and succeeding state s' provided by the environment
  $Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma V(s') - Q(s, a) \right]$
  $C(s) \leftarrow C(s) + 1$
  for all $\mathfrak{a} \in A^k$ **do**
    $\overline{\pi}(s, \mathfrak{a}) \leftarrow \overline{\pi}(s, \mathfrak{a}) + \frac{1}{C(s)} (\pi(s, \mathfrak{a}) - \overline{\pi}(s, \mathfrak{a}))$
  end for
  $\delta = \begin{cases} \delta_w \text{ if } \sum_{\mathfrak{a}} \pi(s, \mathfrak{a}) Q(s, \mathfrak{a}) > \sum_{\mathfrak{a}} \overline{\pi}(s, \mathfrak{a}) Q(s, \mathfrak{a}) \\ \delta_l \qquad\qquad\qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$
  for all $\mathfrak{a} \in A^k$ **do**
    $\pi(s, \mathfrak{a}) \leftarrow \pi(s, \mathfrak{a}) + \begin{cases} \delta & \text{if } \mathfrak{a} = \arg\max_{\mathfrak{a}'} Q(s, \mathfrak{a}') \\ \frac{-\delta}{|A^k| - 1} & \text{otherwise} \end{cases}$
  end for
  Constrain $\pi(s, \cdot)$ to a legal probability distribution
  $V(s) \leftarrow \max_{\mathfrak{a}} Q(s, \mathfrak{a})$
  decay α
  $s \leftarrow s'$
end loop

- the opponent's policy is observable
- infinitesimally small step sizes are used for gradient ascent

Another algorithm, ReDVaLeR, [17] achieves the two properties in known general-sum repeated games with arbitrarily numbers of actions and agents, but still requires that the opponents' policies are observable and the step sizes could be arbitrarily small. AWESOME introduced in [40] adopts the same criteria and is proved to satisfy them in arbitrary known repeated games. It also doesn't make any of the above assumptions. However, neither WoLF-IGA nor AWESOME algorithm make any guarantee in the presence of non-stationary opponents and can be exploited by adaptive opponents [38].

In order to avoid the possibility of being exploited the agent in [27] is required to have zero average regret. This guarantees that an agent doesn't perform worse than any stationary policy at any time against the opponent's whole history. Several algorithms have been proved to achieve at most zero regret in the limit [66], [93], [171], [15], [58].

In [122], [123] the above criteria were extended in the following way:

- "Against any member of the target set of opponents, the algorithm achieves within ϵ of the expected value of the best-response to the actual opponent."
- "During self-play, the algorithm achieves at least within ϵ of the payoff of some Nash equilibrium that is not Pareto dominated by another Nash equilibrium."
- "Against any opponent, the algorithm always receives at least within ϵ of the security value for the game."

They introduce an algorithm meeting these criteria in known, fully observable two-player repeated games.

As a further work in this direction an algorithm converging to the best-response against a class of non-stationary agents in general-sum stochastic games with arbitrary number of players is proposed in [167]. The algorithm learns best-response policy when facing opponents whose policies are non-stationary but have limits.

There are a number of works whose contribution to the state of the art can't be overlooked. In [18] the attention is drawn to the fact that under some circumstances the adaptive opponents developed in [38] themselves could be exploited. In [105] the adaptation aspect of Q-learning came into consideration. Since Q-learning algorithm regards other agents as a part of the environment and strives to adapt to this environment, it can be thought as a follower. In [105] two "leader" algorithms urging Q-learning to better performance via stubbornness and threats are introduced. Such leader-follower behavior is considered in the context of repeated, two-player, general-sum games.

For better understanding of opponents' policies evolution that directly influence the agent's performance through rewards, it might be profitable to learn not only the superficial changes but to model the processes underneath. In [89] interaction of agents with different levels of recursive models in double auction market is considered.

In [106] a polynomial-time algorithm is proposed for computing a Nash equilibrium of average payoffs in non-stationary strategies for repeated bimatrix games.

Replicator dynamics [166], [74] was also used as an approach for developing multi-agent reinforcement learning algorithms [17], [68] as well as understanding classical reinforcement learning techniques for one-agent environments [153], [25], [155], [154], [59].

2.4 Conclusion

Reinforcement learning is a promising technique of programming agents in multi-agent systems. The environment in this case takes the form of general-sum discounted stochastic game. It is the most natural to accept Nash equilibrium as the optimal solution concept.

A number of algorithms have been proposed to extend reinforcement learning approach to multi-agent systems. When the general-sum discounted

stochastic games are known from the very beginning nonlinear optimization and stochastic tracing procedure are proved to find Nash equilibrium in the general case. In case when the games are being learned by interaction with the environment, a number of algorithms: Minimax-Q, JAL, Nash-Q, Friend- and For-Q, CE-Q, Hyper-Q, PHC, WoLF-PHC were developed. The convergence to Nash equilibria was proved for very restricted class of environments: strictly competitive (Minimax-Q), strictly cooperative (JAL) and 2-agent 2-action iterative games (WoLF-PHC). Nash-Q algorithm has achieved convergence to Nash equilibrium in self-play for strictly competitive and strictly cooperative games under additional very restrictive condition that all equilibria encountered during learning stage are unique.

Part II

Replicator Dynamics Based Multi-Agent Reinforcement Learning Algorithms

3

Nash-RD Approach: Theoretical Basis

In the next chapter we are introducing an approach that allows to compute stationary Nash equilibria of general-sum discounted stochastic games with a given accuracy. The goal of this chapter is to present some theoretical results necessary for a formal proof of the convergence of the developed approach to Nash equilibrium under some assumptions.

As we discussed in section 2.2, stochastic games are a generalization of both Markov decision processes and matrix games. Since for the last 15 years all the attempts to extend Bellman optimality equation (the base of many reinforcement learning algorithms in isolated environments) for multi-agent environments failed, our idea was to extend approaches to calculating Nash equilibria of matrix games for multi-state case (see figure 3.1). The approach presented in the next chapter was inspired by multi-population replicator dynamics [146].

The results of this chapter were partly published in [9] and [7].

The chapter is organized as follows. Multi-population replicator dynamics for matrix games is presented in section 3.1. In section 3.2 we develop multi-population replicator dynamics for discounted stochastic games as well as prove some useful theorems. Nash equilibrium approximation theorem — the theorem that serves as the main theoretical basis for the developed approach is proved in section 3.3. Section 3.4 is devoted to discussion and necessary experimental estimations of the conditions of the Nash equilibrium approximation theorem.

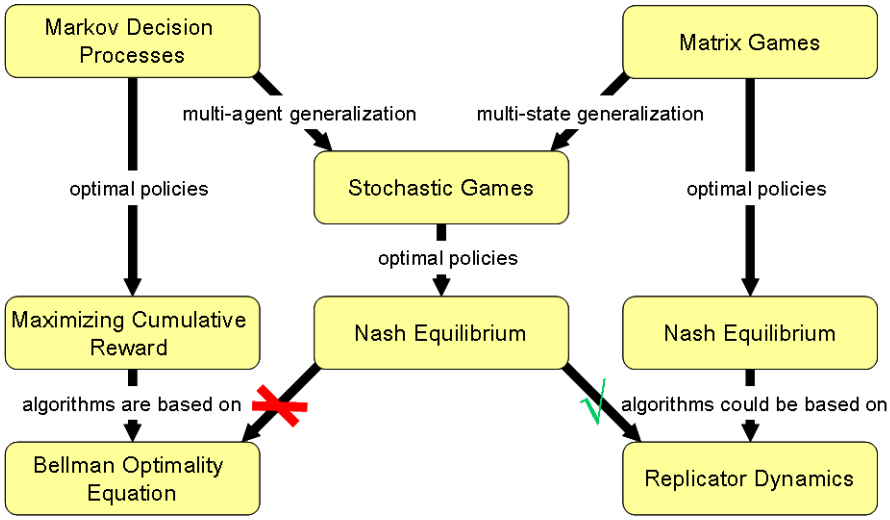


Fig. 3.1. The Idea

3.1 Replicator Dynamics in Matrix Games

The approach presented in this chapter was inspired by multi-population replicator dynamics[1] [147] whose behavior is a subject of investigation of evolutionary game theory [109], [80].

Evolutionary game theory [75] analyzes connections between phenomena arising during the evolution process and the concepts of traditional non-cooperative game theory (e.g., Nash equilibria). Such analysis is of interest to several areas like evolutionary biology and economics. In contrast to evolutionary algorithms where the environment is externally fixed, evolutionary game theory presumes that the environment of an individual consists of other individuals who themselves undergo the process of natural selection.

Replicator dynamics formalizes the following evolutionary process: n finite populations of individuals compete in the environment represented as an n-player matrix game $G = \langle K, A^1, \ldots, A^n, r^1, \ldots, r^n \rangle$. The game is repeated infinitely many times and at each instant of time the players are being randomly chosen from the contending populations. Each individual follows one pure policy determined by inherited replicator. Let $p_h^k(t) \geq 0$ be the number of individuals who follow pure strategy e_h^k at time t, $k \in K$, $h \in A^k$.

$$e_h^k = \left(e^k_{ha_1^k}, e^k_{ha_2^k}, \ldots, e^k_{ha^k_{m^k}} \right)$$

[1] Replicators [46] are entities that are passed on from generation to generation without modifications. The term was first introduced by Peter Schuster and Karl Sigmund in [128].

$$e^k_{ha} = \begin{cases} 0 \; a \neq h \\ 1 \; a = h \end{cases}$$

The size of the whole population $p^k(t)$, $k \in K$ being equal to $p^k(t) = \sum_{h \in A^k} p^k_h(t) > 0$. Thus, a mixed strategy $x^k(t) = (x^k_{a^k_1}(t), \ldots, x^k_{a^k_{m^k}}(t))$ played by the kth population is determined by the proportion of individuals who are programmed to pure policy e^k_h at time t: $x^k_h(t) = \frac{p^k_h(t)}{p^k(t)}$, $h \in A^k$.

The expected reward of an individual of population k programmed to pure strategy e^k_h at instant t will be:

$$r^k(x^1(t), \ldots, x^{k-1}(t), e^k_h, x^{k+1}(t), \ldots, x^n(t)) = \sum_{a^1 \in A^1} \cdots \sum_{a^{k-1} \in A^{k-1}}$$

$$\sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} r^k(a^1, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \prod_{i=1, i \neq k}^{n} x^i_{a^i}(t)$$

The average expected reward of population k at time t will be:

$$r^k(x(t)) = \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^n \in A^n} r^k(a^1, a^2, \ldots, a^n) \prod_{i=1}^{n} x^i_{a^i}(t)$$

The expected reward corresponds to the fitness of the individual. The higher the fitness, the more offsprings (copies of itself) the individual can produce.

The birthrate at instant t of individuals inheriting replicator e^k_h, is $\beta_k + r^k(x^1(t), \ldots, x^{k-1}(t), e^k_h, x^{k+1}(t), \ldots, x^n(t))$, where $\beta_k \geq 0$ is the initial fitness of individuals in the population k. The death rate $\delta_k \geq 0$ is assumed to be the same for all the individuals in population k.

Such population dynamics can be formally presented as the following system of ordinary differential equations[2].

$$\frac{dp^k_h(t)}{dt} = \left[\beta_k + r^k(x^1(t), \ldots, x^{k-1}(t), e^k_h, x^{k+1}(t), \ldots, x^n(t)) - \delta_k\right] p^k_h(t)$$

$$\frac{dp^k(t)}{dt} = \frac{d}{dt} \sum_{h \in A^k} p^k_h(t) = \sum_{h \in A^k} \frac{dp^k_h(t)}{dt} =$$

$$= \sum_{h \in A^k} \left[\beta_k + r^k(x^1(t), \ldots, x^{k-1}(t), e^k_h, x^{k+1}(t), \ldots, x^n(t)) - \delta_k\right] p^k_h(t)$$

[2] Introduction to theory of ordinary differential equations could be found in appendix A.

$$\frac{dp^k(t)}{dt} = \beta_k \sum_{h\in A^k} p_h^k(t) + \sum_{h\in A^k} r^k(x^1(t),\dots,x^{k-1}(t),e_h^k,x^{k+1}(t),\dots,x^n(t))p_h^k(t)$$
$$- \delta_k \sum_{h\in A^k} p_h^k(t) = \beta_k p^k(t) +$$
$$+ \sum_{h\in A^k} r^k(x^1(t),\dots,x^{k-1}(t),e_h^k,x^{k+1}(t),\dots,x^n(t))x_h^k(t)p^k(t) - \delta_k p^k(t)$$
$$= \beta_k p^k(t) + r^k(x(t))p^k(t) - \delta_k p^k(t) = \left[\beta_k + r^k(x(t)) - \delta_k\right] p^k(t)$$

Since

$$p^k(t)x_h^k(t) = p_h^k(t)$$

and

$$\frac{dp_h^k(t)}{dt} = \frac{dp^k(t)}{dt}x_h^k(t) + p^k(t)\frac{dx_h^k(t)}{dt}$$

we get

$$p^k(t)\frac{dx_h^k(t)}{dt} = \frac{dp_h^k(t)}{dt} - \frac{dp^k(t)}{dt}x_h^k(t) =$$
$$= \left[\beta_k + r^k(x^1(t),\dots,x^{k-1}(t),e_h^k,x^{k+1}(t),\dots,x^n(t)) - \delta_k\right] p_h^k(t) -$$
$$- \left[\beta_k + r^k(x(t)) - \delta_k\right] p^k(t)x_h^k(t) =$$
$$= \left[\beta_k + r^k(x^1(t),\dots,x^{k-1}(t),e_h^k,x^{k+1}(t),\dots,x^n(t)) - \delta_k\right] p^k(t)x_h^k(t)$$
$$- \left[\beta_k + r^k(x(t)) - \delta_k\right] p^k(t)x_h^k(t) =$$
$$= \left[r^k(x^1(t),\dots,x^{k-1}(t),e_h^k,x^{k+1}(t),\dots,x^n(t)) - r^k(x(t))\right] p^k(t)x_h^k(t)$$

Dividing by $p^k(t)$ we get the formal representation of replicator dynamics:

$$\frac{dx_h^k}{dt} = \left[r^k(x^1,\dots,x^{k-1},e_h^k,x^{k+1},\dots,x^n) - r^k(x)\right] x_h^k \tag{3.1}$$

The population share of individuals using strategy e_h^k grows proportionally to the gain the pure strategy e_h^k allows to obtain over the mixed strategy x^k. The subpopulation whose corresponding pure policy yields worse than average result diminishes.

Theorem 3.1. *The system of nonlinear differential equations 3.1 has a unique global solution.*

Proof. Directly follows from theorem A.9, A.8, A.10.

□

Example 3.2. Let us consider a bimatrix game $G = \left(M^1, M^2\right)$, where M^1 is the payoff matrix of player 1 and M^2 denotes the payoff matrix of player 2.

The corresponding replicator dynamics will have the following form:

$$\frac{dx_1^1}{dt} = \left[\left(M^1 x^2\right)_1 - x^{1\mathrm{T}} M^1 x^2\right] x_1^1$$
$$\frac{dx_2^1}{dt} = \left[\left(M^1 x^2\right)_2 - x^{1\mathrm{T}} M^1 x^2\right] x_2^1$$
$$\frac{dx_1^2}{dt} = \left[\left(x^{1\mathrm{T}} M^2\right)_1 - x^{1\mathrm{T}} M^2 x^2\right] x_1^2$$
$$\frac{dx_2^2}{dt} = \left[\left(x^{1\mathrm{T}} M^2\right)_2 - x^{1\mathrm{T}} M^2 x^2\right] x_2^2$$

Theorem 3.3. *[166] If interior state x of the system of nonlinear differential equations 3.1 is stationary, then x constitutes a Nash equilibrium of matrix game G.*

Theorem 3.4. *[166] If x is the limit of some interior solution of the system of nonlinear differential equations 3.1, then x constitutes a Nash equilibrium of matrix game G.*

Theorem 3.5. *[166] If state x of the system of nonlinear differential equations 3.1 is Lyapunov stable, then x constitutes a Nash equilibrium of matrix game G.*

Theorem 3.6. *[129] Suppose that $G = \langle A^1, A^2, r^1, r^2\rangle$ is a bimatrix game and let $x(t)$ be some interior solution of corresponding replicator dynamics, then $y_h^k = \lim_{T\to\infty} \frac{1}{T} \int_0^T x_h^k(t)dt$, $k = 1, 2$, $h \in A^k$ constitutes a Nash equilibrium of G.*

3.2 Replicator Dynamics in Stochastic Games

Let now n finite populations compete in the environment presented as n-player discounted stochastic game $\mathit{\Gamma} = \langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, p\rangle$. Let replicator code not only the information on the policy to be followed but also on state of the game the individual is to play in and let discounted value take the place of expected reward and represent the fitness of individuals.

Repeating the inference of section 3.1 we will get the following formal representation of replicator dynamics for discounted stochastic games:

$$\frac{dx_{sh}^k}{dt} = \left[\vartheta_{sh}^k(x) - v_s^k(x)\right] x_{sh}^k \qquad k \in K,\ s \in S,\ h \in A^k \tag{3.2}$$

where

$$\vartheta_{sh}^k(x) = r^k(s, \chi) + \gamma \sum_{s' \in S} p(s'|s, \chi) v_{s'}^k(x)$$

and χ denotes the profile equal to x but where player k always plays action h in state s.

Let $x(t)$ be the solution of the system of differential equations 3.2 satisfying some initial conditions[3]:

$$x_{sh}^k(0) = x_{sh}^{k(0)} \in (0,1)$$
$$\sum_{h \in A^k} x_{sh}^k(0) = 1$$
$$k \in K,\ s \in S,\ h \in A^k$$

Proposition 3.7. *At any instant t, $x(t)$ will be a valid profile*

$$x(t) \in \Phi$$

Proof.

$$\begin{aligned}\sum_{h \in A^k} \frac{dx_{sh}^k}{dt} &= \sum_{h \in A^k} \left[\vartheta_{sh}^k(x) - v_s^k(x)\right] x_{sh}^k = \\ &= \sum_{h \in A^k} \vartheta_{sh}^k(x) x_{sh}^k - \sum_{h \in A^k} v_s^k(x) x_{sh}^k = \\ &= v_s^k(x) - v_s^k(x) = 0\end{aligned}$$

□

Theorem 3.8. *The system of nonlinear differential equations 3.2 has a unique global solution.*

Proof. Directly follows from theorems A.9, A.8, 3.7, A.10.

□

Theorem 3.9. *If x is a Nash equilibrium of discounted stochastic game Γ, then x is a stationary state of the system of nonlinear differential equations 3.2.*

Proof. Since x is a Nash equilibrium of discounted stochastic game Γ according to theorem 2.26:

For every state $s \in S$ the vector $(x_s^1, x_s^2, \ldots, x_s^n)$ constitutes a Nash equilibrium in the n-matrix game $(B_s^1, B_s^2, \ldots, B_s^n)$ with equilibrium payoffs $(v_s^1(x), v_s^2(x), \ldots, v_s^n(x))$, where for $k \in K$ and $(a^1, a^2, \ldots, a^n) \in A^1 \times A^2 \times \ldots \times A^n$ entry $(a^1, a^2, \ldots, a^n)$ of B_s^k equals

$$b^k(s, a^1, a^2, \ldots, a^n) = r^k(s, a^1, a^2, \ldots, a^n) + \gamma \sum_{s' \in S} p(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k(x)$$

[3] $x(t) = \xi(t, x^{(0)})$ according to the notation introduced in appendix A.

That means that for every state $s \in S$ and for every agent $k \in K$ there is no $y_s^k = \left(y_{sa_1^k}^k, y_{sa_2^k}^k, \ldots, y_{sa_{m^k}^k}^k\right) \in \Delta^k$:

$$\sum_{h \in A^k} y_{sh}^k \vartheta_{sh}^k(x) > \sum_{h \in A^k} x_{sh}^k \vartheta_{sh}^k(x)$$

Let us consider $x_s^k = \left(x_{sa_1^k}^k, x_{sa_2^k}^k, \ldots, x_{sa_{m^k}^k}^k\right)$ for some $s \in S$ and $k \in K$.

Without loss of generality let us suppose, that

$$x_{sa_1^k}^k > 0,\ x_{sa_2^k}^k > 0,\ \ldots,\ x_{sa_l^k}^k > 0$$

and

$$x_{sa_{l+1}^k}^k = x_{sa_{l+2}^k}^k = \ldots = x_{sa_{m^k}^k}^k = 0$$

Since $\left(x_s^1, x_s^2, \ldots, x_s^n\right)$ constitutes a Nash equilibrium in corresponding n-matrix game

$$\vartheta_{sa_1^k}^k(x) = \vartheta_{sa_2^k}^k(x) = \ldots = \vartheta_{sa_l^k}^k(x) = v_s^k(x)$$

and for all $h \in \{a_{l+1}^k, a_{l+2}^k, \ldots, a_{m^k}^k\}$:

$$\vartheta_{sh}^k(x) \leq v_s^k(x)$$

Otherwise there would exist a policy y_s^k yielding higher payoffs.

Thus,

$$\left[\vartheta_{sh}^k(x) - v_s^k(x)\right] x_{sh}^k = 0 \quad \text{for all} \quad h \in A^k$$

Since this is valid for all $s \in S$ and $k \in K$, the conclusion follows.

□

Theorem 3.10. *If interior state x of the system of nonlinear differential equations 3.2 is stationary, then x constitutes a Nash equilibrium of discounted stochastic game Γ.*

Proof. (partly by analogy with theorem 3.3)

Suppose interior state x is stationary, then

$$\left[\vartheta_{sh}^k(x) - v_s^k(x)\right] x_{sh}^k = 0$$

for all $k \in K,\ s \in S,\ h \in A^k$.

Since x is an interior state, $x_{sh}^k \neq 0$ and $\vartheta_{sh}^k(x) = v_s^k(x)$ for all $k \in K,\ s \in S,\ h \in A^k$.

Thus, for any arbitrary $y_s^k = \left(y_{sa_1^k}^k, y_{sa_2^k}^k, \ldots, y_{sa_{m^k}^k}^k\right) \in \Delta^k$ for all $k \in K$ and $s \in S$:

$$\sum_{h \in A^k} y_{sh}^k \vartheta_{sh}^k(x) = v_s^k(x) \sum_{h \in A^k} y_{sh}^k = v_s^k(x)$$

Hence, no strategy $y_s^k \in \Delta^k$ will yield higher rewards than x_s^k.

As k is arbitrary, the vector $(x_s^1, x_s^2, \ldots, x_s^n)$ constitutes a Nash equilibrium in the n-matrix game $(B_s^1, B_s^2, \ldots, B_s^n)$ with equilibrium payoffs $(v_s^1(x), v_s^2(x), \ldots, v_s^n(x))$, where for $k \in K$ and $(a^1, a^2, \ldots, a^n) \in A^1 \times A^2 \times \ldots \times A^n$ entry $(a^1, a^2, \ldots, a^n)$ of B_s^k equals

$$b^k(s, a^1, a^2, \ldots, a^n) = r^k(s, a^1, a^2, \ldots, a^n) + \gamma \sum_{s' \in S} p(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k(x)$$

Applying theorem 2.26 we get that x constitutes a Nash equilibrium of discounted stochastic game Γ.

□

Theorem 3.11. *If x is the limit of some interior solution of the system of nonlinear differential equations 3.2, then x constitutes a Nash equilibrium of discounted stochastic game Γ.*

Proof. (partly by analogy with theorem 3.4)

Suppose $x_{sh}^k(0) = x_{sh}^{k(0)} \in (0,1)$ for all $k \in K,\ s \in S,\ h \in A^k$ and $\xi(t, x^0)_{t\to\infty} \to x$.

Then it directly follows from theorems A.8, 3.7 and propositions A.10 and A.14, that x is stationary:

$$\left[\vartheta_{sh}^k(x) - v_s^k(x)\right] x_{sh}^k = 0$$

for all $k \in K,\ s \in S,\ h \in A^k$.

If x does not constitute a Nash equilibrium then according to theorem 2.26 $\vartheta_{sh}^k(x) > v_s^k(x)$ for some $k \in K,\ s \in S,\ h \in A^k$. Because of stationarity of x, $x_{sh}^k = 0$ will hold for such $k \in K,\ s \in S,\ h \in A^k$.

Since $\vartheta_{sh}^k(x) - v_s^k(x)$ is continuous, there is a $\delta > 0$ and a neighborhood U of x such that $\vartheta_{sh}^k(y) - v_s^k(y) \geq \delta$ for all $y \in U$. The condition that $\xi(t, x^0)_{t\to\infty} \to x$ implies that there exists a time $T > 0$ such that $\xi(t, x^0) \in U$ for all $t \geq T$. Since $x_{sh}^k = 0$, there must be some $t \geq T$ such that $\frac{d\xi_{sh}^k(t,x^0)}{dt} < 0$, a contradiction to $\vartheta_{sh}^k(x) - v_s^k(x) > 0$ on U. Thus x constitutes a Nash equilibrium of Γ.

□

Theorem 3.12. *If state x of the system of nonlinear differential equations 3.2 is Lyapunov stable, then x constitutes a Nash equilibrium of discounted stochastic game Γ.*

Proof. (partly by analogy with theorem 3.5)

According to proposition A.16 x is stationary.

Then

$$\left[\vartheta_{sh}^k(x) - v_s^k(x)\right] x_{sh}^k = 0$$

for all $k \in K,\ s \in S,\ h \in A^k$.

If x does not constitute a Nash equilibrium then according to theorem 2.26

$$\vartheta^k_{sh}(x) > v^k_s(x)$$

for some $k \in K$, $s \in S$, $h \in A^k$.

By stationarity of x, $x^k_{sh} = 0$ will hold for such $k \in K$, $s \in S$, $h \in A^k$.

Since $\vartheta^k_{sh}(x) - v^k_s(x)$ is continuous, there is a $\delta > 0$ and a neighborhood U of x such that $\vartheta^k_{sh}(y) - v^k_s(y) \geq \delta$ for all $y \in U$.

Thus $\xi^k_{sh}(t, x^{(0)}) \geq x^k_{sh}(0)e^{\delta t}$ for any $x^{(0)} \in U$ and all $t > 0$ such that $\xi^k_{sh}(t, x^{(0)}) \in U$.

So there exists some neighborhood $B \subset U$ of x that the system abandons in finite time if started in any neighborhood $B^0 \subset B$ of x, what contradicts the assumption that x is Lyapunov stable.

□

3.3 Nash Equilibrium Approximation Theorem

This section is devoted to the proof of Nash equilibrium approximation theorem — the main theoretical basis for the developed approach. Nash equilibrium approximation theorem is proved for a general case — when the transition probabilities are being learned.

First we will present some our theoretical results for an arbitrary n-player discounted stochastic game $\Gamma = \langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, p \rangle$.

Lemma 3.13. *If $k \in K$, $x \in \Phi$ and $v, \epsilon \in \mathbb{R}^N$ are such that*

$$v \geq r^k(x) + \gamma P(x)v - \epsilon$$

then $v \geq v^k(x) - \sum_{t=0}^{\infty} \gamma^t P^t(x)\epsilon$.

Proof.

$$\begin{aligned} v &\geq r^k(x) + \gamma P(x)\left[r^k(x) + \gamma P(x)v - \epsilon\right] - \epsilon \\ &= r^k(x) + \gamma P(x)r^k(x) + \gamma^2 P^2(x)v - \\ &- \epsilon - \gamma P(x)\epsilon \end{aligned}$$

If we substitute the above inequality into itself $i-1$ times we get:

$$\begin{aligned} v \geq\ & r^k(x) + \gamma P(x)r^k(x) + \gamma^2 P^2(x)r^k(x) + \\ &+ \ldots + \\ &+ \gamma^{i-1} P^{i-1}(x)r^k(x) + \gamma^i P^i(x)v - \\ &- \epsilon - \gamma P(x)\epsilon - \gamma^2 P^2(x)\epsilon - \ldots - \gamma^{i-1} P^{i-1}(x)\epsilon \end{aligned}$$

Upon taking the limit at infinity ($i \to \infty$) we will obtain:

$$v \geq v^k(x) - \sum_{t=0}^{\infty} \gamma^t P^t(x)\epsilon$$

□

Theorem 3.14. *From* $1 \Rightarrow 2$

1. *For each state* $s \in S$*, the vector* $(x_s^1, x_s^2, \ldots, x_s^n)$ *constitutes an* ϵ*-equilibrium in the n-matrix game* $(B_s^1, B_s^2, \ldots, B_s^n)$ *with equilibrium payoffs* $(v_s^1, v_s^2, \ldots, v_s^n)$*, where for* $k \in K$ *and* $(a^1, a^2, \ldots, a^n) \in A^1 \times A^2 \times \ldots \times A^n$ *entry* $(a^1, a^2, \ldots, a^n)$ *of* B_s^k *equals*

$$b^k(s, a^1, a^2, \ldots, a^n) = r^k(s, a^1, a^2, \ldots, a^n) + \gamma \sum_{s' \in S} (p(s'|s, a^1, a^2, \ldots, a^n) + \\ + \varsigma(s'|s, a^1, a^2, \ldots, a^n))(v_{s'}^k + \sigma_{s'}^k)$$

   where $-\sigma < \sigma_s^k < \sigma$*,* $-\varsigma < \varsigma(s'|s, a^1, a^2, \ldots, a^n) < \varsigma$ *for all* $s' \in S$*.*
2. x *is an* ε*-equilibrium in the discounted stochastic game* Γ *and*

$$v^k(x) - \frac{\omega}{1-\gamma}\mathbf{1} < v^k < v^k(x) + \frac{\omega}{1-\gamma}\mathbf{1}$$

   for all $k \in K$ *where*

$$\omega = \gamma\sigma + \gamma\varsigma N \max_{k \in K, s \in S} |v_s^k| + \gamma N \varsigma \sigma$$

   and

$$\varepsilon = \frac{2\omega + \epsilon}{1-\gamma}$$

Proof.

$$\begin{aligned} b^k(s, a^1, a^2, \ldots, a^n) &= r^k(s, a^1, a^2, \ldots, a^n) + \\ &+ \gamma \sum_{s' \in S} p(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k + \\ &+ \gamma \sum_{s' \in S} \varsigma(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k + \\ &+ \gamma \sum_{s' \in S} p(s'|s, a^1, a^2, \ldots, a^n) \sigma_{s'}^k + \\ &+ \gamma \sum_{s' \in S} \varsigma(s'|s, a^1, a^2, \ldots, a^n) \sigma_{s'}^k = \\ &= r^k(s, a^1, a^2, \ldots, a^n) + \xi^k(s, a^1, a^2, \ldots, a^n) + \\ &+ \gamma \sum_{s' \in S} p(s'|s, a^1, a^2, \ldots, a^n) v_{s'}^k \end{aligned}$$

where

$$\xi^k(s,a^1,a^2,\ldots,a^n) = \gamma \sum_{s'\in S} p(s'|s,a^1,a^2,\ldots,a^n)\sigma^k_{s'} + \\ + \gamma \sum_{s'\in S} \varsigma(s'|s,a^1,a^2,\ldots,a^n)v^k_{s'} + \\ + \gamma \sum_{s'\in S} \varsigma(s'|s,a^1,a^2,\ldots,a^n)\sigma^k_{s'}$$

Let's estimate the worst case:

$$- \gamma \sum_{s'\in S} p(s'|s,a^1,a^2,\ldots,a^n)\sigma - \\ - \gamma \sum_{s'\in S} \varsigma \max_{k\in K} |v^k_{s'}| - \gamma \sum_{s'\in S} \varsigma\sigma < \\ < \gamma \sum_{s'\in S} p(s'|s,a^1,a^2,\ldots,a^n)\sigma^k_{s'} + \\ + \gamma \sum_{s'\in S} \varsigma(s'|s,a^1,a^2,\ldots,a^n)v^k_{s'} + \\ + \gamma \sum_{s'\in S} \varsigma(s'|s,a^1,a^2,\ldots,a^n)\sigma^k_{s'} < \\ < \gamma \sum_{s'\in S} p(s'|s,a^1,a^2,\ldots,a^n)\sigma + \\ + \gamma \sum_{s'\in S} \varsigma \max_{k\in K} |v^k_{s'}| + \gamma \sum_{s'\in S} \varsigma\sigma$$

Let's denote $\omega = \gamma\sigma + \gamma\varsigma N \max_{k\in K, s\in S} |v^k_s| + \gamma N\varsigma\sigma$

$$-\omega < \xi^k(s,a^1,a^2,\ldots,a^n) < \omega$$

Let's take some arbitrary $f \in \Theta^1$. If (1) is true, then for each state $s \in S$ by definition of ϵ-equilibrium:

$$r^1(s,f,x^2,\ldots,x^n)+\zeta^1(s,f,x^2,\ldots,x^n)+\gamma \sum_{s'\in S} p(s'|s,f,x^2,\ldots,x^n)v^1_{s'} \le v^1_s+\epsilon$$

where

$$\zeta^k(s,x) = \sum_{a^1\in A^1} \sum_{a^2\in A^2} \cdots \sum_{a^n\in A^n} \xi^k(s,a^1,a^2,\ldots,a^n) \prod_{i=1}^{n} x^i_{sa^i}$$

In the worst case:

$$-\sum_{a^1\in A^1}\sum_{a^2\in A^2}\dots\sum_{a^n\in A^n}\omega\prod_{i=1}^{n}x^i_{sa^i}<$$

$$<\sum_{a^1\in A^1}\sum_{a^2\in A^2}\dots\sum_{a^n\in A^n}\xi^k(s,a^1,a^2,\dots,a^n)\prod_{i=1}^{n}x^i_{sa^i}<$$

$$<\sum_{a^1\in A^1}\sum_{a^2\in A^2}\dots\sum_{a^n\in A^n}\omega\prod_{i=1}^{n}x^i_{sa^i}$$

$$-\omega<\zeta^k(s,x)<\omega$$

$$r^1(s,f,x^2,\dots,x^n)+\gamma\sum_{s'\in S}p(s'|s,f,x^2,\dots,x^n)v^1_{s'}\le v^1_s+\omega+\epsilon$$

Applying lemma 3.13 we get

$$v^1(f,x^2,\dots,x^n)-(\omega+\epsilon)\sum_{t=0}^{\infty}\gamma^t\mathbf{1}\le v^1$$

for all $f\in\Theta^1$.

Since

$$\sum_{t=0}^{\infty}\gamma^t=\frac{1}{1-\gamma}$$

we obtain

$$v^1(f,x^2,\dots,x^n)-\frac{\omega+\epsilon}{1-\gamma}\mathbf{1}\le v^1$$

holds for all $f\in\Theta^1$.

And by symmetrical arguments it follows that

$$v^k(x^1,\dots,x^{k-1},f,x^{k+1},\dots,x^n)-\frac{\omega+\epsilon}{1-\gamma}\mathbf{1}\le v^k$$

holds for all $k\in K$ and for all $f\in\Theta^k$.

But

$$v^k=r^k(x)+\gamma P(x)v^k+\zeta^k(x)$$

$$\begin{aligned}v^k&=r^k(x)+\gamma P(x)[r^k(x)+\gamma P(x)v^k+\zeta^k(x)]+\zeta^k(x)=\\&=r^k(x)+\gamma P(x)r^k(x)+\gamma^2P^2(x)v^k+\\&+\zeta^k(x)+\gamma P(x)\zeta^k(x)\end{aligned}$$

If we substitute the above inequality into itself and take the limit, we get:

$$v^k = v^k(x) + \sum_{t=0}^{\infty} \gamma^t P^t(x) \zeta^k(x)$$

$$v^k(x) - \omega \sum_{t=0}^{\infty} \gamma^t \mathbf{1} < v^k < v^k(x) + \omega \sum_{t=0}^{\infty} \gamma^t \mathbf{1}$$

$$v^k(x) - \frac{\omega}{1-\gamma}\mathbf{1} < v^k < v^k(x) + \frac{\omega}{1-\gamma}\mathbf{1}$$

$$v^k(x^1, \ldots, x^{k-1}, f, x^{k+1}, \ldots, x^n) - \frac{\omega + \epsilon}{1-\gamma}\mathbf{1} \leq v^k(x) + \frac{\omega}{1-\gamma}\mathbf{1}$$

for all $k \in K$ and for all $f \in \Theta^k$.

$$v^k(x^1, \ldots, x^{k-1}, f, x^{k+1}, \ldots, x^n) \leq v^k(x) + \frac{2\omega + \epsilon}{1-\gamma}\mathbf{1}$$

for all $k \in K$ and for all $f \in \Theta^k$.

So the condition of theorem 3.14 is satisfied with

$$\varepsilon = \frac{2\omega + \epsilon}{1-\gamma}$$

where

$$\omega = \gamma\sigma + \gamma\varsigma N \max_{k \in K, s \in S} |v_s^k| + \gamma N \varsigma \sigma$$

and we get (2).

□

Now we are ready to prove Nash equilibrium approximation theorem for a general case — when the transition probabilities are being learned.

Let

$$\Gamma = \langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, p \rangle$$

and

$$\widetilde{\Gamma} = \langle K, S, A^1, \ldots, A^n, \gamma, r^1, \ldots, r^n, \widetilde{p} \rangle$$

be n-player discounted stochastic games such that for all $s' \in S$ and $(s, a^1, a^2, \ldots, a^n) \in S \times A^1 \times A^2 \times \ldots \times A^n$:

$$\widetilde{p}(s'|s, a^1, a^2, \ldots, a^n) = p(s'|s, a^1, a^2, \ldots, a^n) + \varsigma(s'|s, a^1, a^2, \ldots, a^n)$$

where $-\varsigma < \varsigma(s'|s, a^1, a^2, \ldots, a^n) < \varsigma$.

Since the agents don't know transition probabilities they have only an approximation $\widetilde{\Gamma}$ of the actual game Γ at each learning stage. We suppose though that the agents have already learned the reward functions.

Further on $\widetilde{}$ will indicate that we are using an approximation of transition probabilities $\widetilde{p}$ instead of the actual transition probabilities p for calculation of the corresponding values.

Let's consider replicator dynamics in game $\widetilde{\Gamma}$:

$$\frac{dx_{sh}^k}{dt} = \left[\vartheta_{sh}^k(x) - \widetilde{v}_s^k(x)\right] x_{sh}^k \qquad k \in K,\ s \in S,\ h \in A^k \tag{3.3}$$

where

$$\vartheta_{sh}^k(x) = r^k(s,\chi) + \gamma \sum_{s' \in S} \widetilde{p}(s'|s,\chi)\widetilde{v}_{s'}^k(x)$$

and χ denotes the profile equal to x but where player k always plays action h in state s.

Let $x(t)$ be the solution of the system of differential equations 3.3 satisfying some initial conditions:

$$x_{sh}^k(0) = x_{sh}^{k(0)} \in (0,1)$$

$$\sum_{h \in A^k} x_{sh}^k(0) = 1$$

$$k \in K,\ s \in S,\ h \in A^k$$

Let y denote the profile where $y_{sh}^k = \frac{1}{T}\int_0^T x_{sh}^k(t)dt$ for some T, for all $k \in K,\ s \in S$ and $h \in A^k$.

And let $\upsilon = (\upsilon_s^k)_{s \in S, k \in K}$ denote the matrix where $\upsilon_s^k = \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt$.

Let

$$\nu_{sh}^k(y) = r^k(s,\varphi) + \gamma \sum_{s' \in S} \widetilde{p}(s'|s,\varphi)\upsilon_{s'}^k$$

where φ stands for the profile equal to y but where player k always plays action h in state s.

Theorem 3.15. *If $T, \epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4 \in \mathbb{R}$ are such that:*

1. *for all $k \in K$ and $s \in S$ $\exists C1_s^k$ and $C2_s^k$:*

$$C1_s^k \cup C2_s^k = A^k$$

$$C1_s^k \cap C2_s^k = \emptyset$$

and such that

a) for all $h \in C1_s^k$: $\left|\frac{1}{T}\ln x_{sh}^k(T) - \frac{1}{T}\ln x_{sh}^k(0)\right| < \epsilon_1$

b) for all $h \in C2_s^k$: $\frac{1}{T}\ln x_{sh}^k(T) - \frac{1}{T}\ln x_{sh}^k(0) \leq -\epsilon_1$

c) $\sum_{h \in C2_s^k} y_{sh}^k(\max_{i \in A^k} \nu_{si}^k - \nu_{sh}^k) < \epsilon_2$

2. *for all $k \in K,\ s, s' \in S$ and $(a^1, a^2, \ldots, a^n) \in A^1 \times A^2 \times \ldots \times A^n$ the following holds:*

a) $\left|\frac{1}{T}\int_0^T \prod_{i=1, i \neq k}^n x_{sa^i}^i(t)dt - \prod_{i=1, i \neq k}^n y_{sa^i}^i\right| < \epsilon_3$

b) $\left|\frac{1}{T}\int_0^T \widetilde{v}^k_{s'}(t)\prod_{i=1,i\neq k}^n x^i_{sa^i}(t)dt - \upsilon^k_{s'}\prod_{i=1,i\neq k}^n y^i_{sa^i}\right| < \epsilon_4$

then

$y^k_{sh} = \frac{1}{T}\int_0^T x^k_{sh}(t)dt$, $k \in K$, $s \in S$, $h \in A^k$ *constitutes an* ε*-equilibrium of discounted stochastic game* Γ *and*

$$\left|v^k_s(y) - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt\right| < \sigma + \frac{\omega}{1-\gamma}$$

where

$$\varepsilon = \frac{2\omega + \epsilon}{1-\gamma}$$

$$\omega = \gamma\sigma + \gamma\varsigma N \max_{k\in K, s\in S}\left|\sum_{h\in A^k}\nu^k_{sh}(y)y^k_{sh}\right| + \gamma N\varsigma\sigma$$

$$\epsilon = 2\epsilon_1 + 2R_{\max}\epsilon_3\prod_{k=1}^n m^k + 2\gamma\epsilon_4\prod_{k=1}^n m^k + \epsilon_2$$

$$\sigma = 3\epsilon_1 + 3R_{\max}\epsilon_3\prod_{k=1}^n m^k + 3\gamma\epsilon_4\prod_{k=1}^n m^k + \epsilon_2$$

Proof. Let

$$b^k_{sh} = \sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\cdots\sum_{a^n\in A^n} r^k(s,a^1,a^2,\ldots,a^{k-1},h,a^{k+1},\ldots,a^n)\prod_{i=1,i\neq k}^n y^i_{sa^i} +$$
$$+\gamma\sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\cdots\sum_{a^n\in A^n}\sum_{s'\in S}\widetilde{p}(s'|s,a^1,a^2,\ldots,a^{k-1},h,a^{k+1},\ldots,a^n)\upsilon^k_{s'}\prod_{i=1,i\neq k}^n y^i_{sa^i}$$

and

$$b^k_s = \sum_{h\in A^k} b^k_{sh}y^k_{sh}$$

If we could show that for some ϵ and σ and for all $k \in K$, $s \in S$ and $h \in A^k$:

1. $b^k_{sh} \leq b^k_s + \epsilon$ (in other words that $y^k_{sh} = \frac{1}{T}\int_0^T x^k_{sh}(t)dt$ constitute ϵ-equilibrium of corresponding matrix games with equilibrium payoffs $(b^1_s, b^2_s, \ldots, b^n_s)$)
2. $\left|b^k_s - \upsilon^k_s\right| < \sigma$

then by applying the theorem 3.14 we could get the implication in question.

Let's consider arbitrary agent $k \in K$ and state $s \in S$.

Without losing generality let $C1_s^k = \{a_1^k, \ldots, a_l^k\}$ and $C2_s^k = \{a_{l+1}^k, \ldots, a_{m^k}^k\}$.

Evidently, $x_{sh}^k(t) > 0$ on $[0, T]$ and

$$(\ln x_{sh}^k)' = \frac{x_{sh}^{k\,\prime}}{x_{sh}^k} = \vartheta_{sh}^k(x) - \widetilde{v}_s^k(x)$$

If we integrate the above equality from 0 to T and then divide by T, we obtain:

$$\frac{1}{T}\ln x_{sh}^k(T) - \frac{1}{T}\ln x_{sh}^k(0) = \frac{1}{T}\int_0^T \vartheta_{sh}^k(x(t))dt - \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt$$

Thus

$$\left|\frac{1}{T}\int_0^T \vartheta_{sa_1^k}^k(x(t))dt - \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt\right| < \epsilon_1$$

$$\vdots$$

$$\left|\frac{1}{T}\int_0^T \vartheta_{sa_l^k}^k(x(t))dt - \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt\right| < \epsilon_1$$

Upon using the properties of integral we get:

$$\begin{aligned}\frac{1}{T}\int_0^T \vartheta_{sh}^k(x(t))dt = &\sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\cdots\sum_{a^n\in A^n} \\ &r^k(s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n)\cdot \\ &\cdot\ \frac{1}{T}\int_0^T \prod_{i=1, i\neq k}^{n} x_{sa^i}^i(t)dt + \\ &+\gamma\sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\cdots\sum_{a^n\in A^n}\sum_{s'\in S} \\ &\widetilde{p}(s'|s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n)\cdot \\ &\cdot\ \frac{1}{T}\int_0^T \widetilde{v}_{s'}^k(x(t)) \prod_{i=1, i\neq k}^{n} x_{sa^i}^i(t)dt\end{aligned}$$

Let

$$\frac{1}{T}\int_0^T \prod_{i=1, i\neq k}^{n} x_{sa^i}^i(t)dt - \prod_{i=1, i\neq k}^{n}\frac{1}{T}\int_0^T x_{sa^i}^i(t)dt = \epsilon_{3_{sq^k}^k}$$

$$q^k = 1, \ldots, \prod_{i=1, i \neq k}^{n} m^i$$

and

$$\frac{1}{T} \int_0^T \widetilde{v}_{s'}^k(x(t)) \prod_{i=1, i \neq k}^{n} x_{sa^i}^i(t) dt - \frac{1}{T} \int_0^T \widetilde{v}_{s'}^k(x(t)) dt$$
$$\cdot \prod_{i=1, i \neq k}^{n} \frac{1}{T} \int_0^T x_{sa^i}^i(t) dt = \epsilon_{4_{ss'jk}^k}$$

$$j^k = 1, \ldots, \prod_{i=1, i \neq k}^{n} m^i$$

And according to prerequisite 2:

$$-\epsilon_3 < \epsilon_{3_{sq^k}^k} < \epsilon_3$$

and

$$-\epsilon_4 < \epsilon_{4_{ss'jk}^k} < \epsilon_4$$

Thus we get:

$$\begin{aligned} \frac{1}{T} \int_0^T \vartheta_{sh}^k(x(t)) dt = & \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} \\ & r^k(s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \cdot \\ & \cdot \left(\prod_{i=1, i \neq k}^{n} \frac{1}{T} \int_0^T x_{sa^i}^i(t) dt + \epsilon_{3_{sq^k}^k} \right) + \\ & + \gamma \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} \sum_{s' \in S} \\ & \widetilde{p}(s' | s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \cdot \\ & \cdot \left(\frac{1}{T} \int_0^T \widetilde{v}_{s'}^k(x(t)) dt \cdot \prod_{i=1, i \neq k}^{n} \frac{1}{T} \int_0^T x_{sa^i}^i(t) dt + \epsilon_{4_{ss'jk}^k} \right) \end{aligned}$$

$$\frac{1}{T}\int_0^T \vartheta^k_{sh}(x(t))dt = \sum_{a^1\in A^1}\sum_{a^2\in A^2}\dots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\dots\sum_{a^n\in A^n} r^k(s,a^1,a^2,\dots,a^{k-1},h,a^{k+1},\dots,a^n)\cdot \cdot \prod_{i=1,i\neq k}^{n}\frac{1}{T}\int_0^T x^i_{sa^i}(t)dt + + \gamma\sum_{a^1\in A^1}\sum_{a^2\in A^2}\dots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\dots\sum_{a^n\in A^n}\sum_{s'\in S} \widetilde{p}(s'|s,a^1,a^2,\dots,a^{k-1},h,a^{k+1},\dots,a^n)\cdot \cdot \frac{1}{T}\int_0^T \widetilde{v}^k_{s'}(x(t))dt\cdot\prod_{i=1,i\neq k}^{n}\frac{1}{T}\int_0^T x^i_{sa^i}(t)dt+\epsilon_{5^k_{sh}}$$

where

$$-\epsilon_5 < \epsilon_{5^k_{sh}} < \epsilon_5$$

$$\epsilon_5 = m^1m^2\dots m^n(R_{\max}\epsilon_3+\gamma\epsilon_4) = (R_{\max}\epsilon_3+\gamma\epsilon_4)\prod_{k=1}^{n}m^k$$

Apparently $\frac{1}{T}\int_0^T \vartheta^k_{sh}(x(t))dt = b^k_{sh}+\epsilon_{5^k_{sh}}$.
Hence we will have the following inequalities:

$$-\epsilon_1 < b^k_{sa^k_1}+\epsilon_{5^k_{sa^k_1}}-\frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt<\epsilon_1$$

$$\vdots$$

$$-\epsilon_1 < b^k_{sa^k_l}+\epsilon_{5^k_{sa^k_l}}-\frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt<\epsilon_1$$

And finally we get

$$-\epsilon_6 < b^k_{sa^k_1}-\frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt<\epsilon_6$$

$$\vdots$$

$$-\epsilon_6 < b^k_{sa^k_l}-\frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt<\epsilon_6$$

where $\epsilon_6=\epsilon_1+\epsilon_5$.

So the difference between b^k_{sh} for $h\in C1^k_s$ and $\frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt$ won't exceed ϵ_6. Hence the difference between any two $b^k_{sh_1}$ and $b^k_{sh_2}$, $h_1,h_2\in C1^k_s$ won't be more than $2\epsilon_6$.

For all $h\in C2^k_s$ the following condition holds:

$$\frac{1}{T}\ln x^k_{sh}(T) - \frac{1}{T}\ln x^k_{sh}(0) \le -\epsilon_1$$

$$\frac{1}{T}\int_0^T \vartheta^k_{sh}(x(t))dt - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt = \frac{1}{T}\ln x^k_{sh}(T) - \frac{1}{T}\ln x^k_{sh}(0) \le -\epsilon_1$$

$$b^k_{sh} + \epsilon_{5^k_{sh}} - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt \le -\epsilon_1$$

$$b^k_{sh} - \frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt < \epsilon_5 - \epsilon_1 < \epsilon_5 + \epsilon_1 = \epsilon_6$$

Let $b^k_{sh_*} = \max_{h\in A^k} b^k_{sh}$.

If $h_* \in C1^k_s$ then for any $h \in C1^k_s$ the difference between corresponding b^k_{sh} and $b^k_{sh_*}$ won't exceed $2\epsilon_6$ (as we have already demonstrated the difference between any two $b^k_{sh_1}$ and $b^k_{sh_2}$, $h_1, h_2 \in C1^k_s$ won't be more than $2\epsilon_6$). If $h_* \in C2^k_s$ then for any $h \in C1^k_s$ the difference between corresponding b^k_{sh} and $b^k_{sh_*}$ also won't exceed $2\epsilon_6$ (b^k_{sh} from $C2^k_s$ that deviates from $\frac{1}{T}\int_0^T \widetilde{v}^k_s(x(t))dt$ by more than ϵ_6 can't for sure be the maximal for the whole A^k because it will be less than any b^k_{sh} for $h \in C1^k_s$).

The condition

$$\sum_{h\in C2^k_s} y^k_{sh}(\max_{i\in A^k}\nu^k_{si} - \nu^k_{sh}) < \epsilon_2$$

we can rewrite as

$$\sum_{h\in C2^k_s} b^k_{sh_*}y^k_{sh} - \sum_{h\in C2^k_s} b^k_{sh}y^k_{sh} < \epsilon_2$$

$$\sum_{h\in C2^k_s} b^k_{sh}y^k_{sh} > \sum_{h\in C2^k_s} b^k_{sh_*}y^k_{sh} - \epsilon_2$$

$$\begin{aligned} b^k_s = \sum_{h\in A^k} b^k_{sh}y^k_{sh} &= \sum_{h\in C1^k_s} b^k_{sh}y^k_{sh} + \sum_{h\in C2^k_s} b^k_{sh}y^k_{sh} > \\ &> \sum_{h\in C1^k_s} (b^k_{sh_*} - 2\epsilon_6)y^k_{sh} + \sum_{h\in C2^k_s} b^k_{sh_*}y^k_{sh} - \epsilon_2 = \\ &= b^k_{sh_*}\sum_{h\in A^k} y^k_{sh} - 2\epsilon_6 \sum_{h\in C1^k_s} y^k_{sh} - \epsilon_2 > \\ &> b^k_{sh_*} - 2\epsilon_6 - \epsilon_2 \end{aligned}$$

Thus the first inequality $b^k_{sh} \le b^k_s + \epsilon$ that must be proved will hold with $\epsilon = 2\epsilon_6 + \epsilon_2$ for all $h \in A^k$.

As we have just demonstrated for all $h \in A^k$:

$$b_{sh}^k - \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt < \epsilon_6$$

If we multiply each inequality by y_{sh}^k accordingly and sum up we will get:

$$\sum_{h\in A^k} b_{sh}^k y_{sh}^k - \sum_{h\in A^k} \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt \cdot y_{sh}^k < \sum_{h\in A^k} \epsilon_6 y_{sh}^k$$

$$b_s^k - \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt < \epsilon_6$$

From the first inequality that we have already proved and the estimations of b_{sh}^k for $h \in C1_s^k$ we can derive:

$$-\epsilon_6 < b_{sh}^k - \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt \le b_s^k + \epsilon - \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt$$

$$-\epsilon_6 - \epsilon < b_s^k - \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt$$

The second inequality:

$$\left| b_s^k - \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt \right| < \sigma$$

will hold with $\sigma = 3\epsilon_6 + \epsilon_2$.

By theorem 3.14 we get:

$$v_s^k(y) - \frac{\omega}{1-\gamma} < b_s^k < v_s^k(y) + \frac{\omega}{1-\gamma}$$

$$-\sigma < \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt - b_s^k < \sigma$$

$$b_s^k - \sigma < \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt < \sigma + b_s^k$$

$$v_s^k(y) - \frac{\omega}{1-\gamma} - \sigma < b_s^k - \sigma < \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt < \sigma + b_s^k < \sigma + v_s^k(y) + \frac{\omega}{1-\gamma}$$

Thus

$$-\frac{\omega}{1-\gamma} - \sigma < \frac{1}{T}\int_0^T \widetilde{v}_s^k(x(t))dt - v_s^k(y) < \sigma + \frac{\omega}{1-\gamma}$$

Let's calculate ϵ and σ.

$$\epsilon = 2\epsilon_6 + \epsilon_2 = 2(\epsilon_1 + \epsilon_5) + \epsilon_2 =$$
$$= 2(\epsilon_1 + (R_{\max}\epsilon_3 + \gamma\epsilon_4)\prod_{k=1}^{n} m^k) + \epsilon_2 =$$
$$= 2\epsilon_1 + 2R_{\max}\epsilon_3 \prod_{k=1}^{n} m^k + 2\gamma\epsilon_4 \prod_{k=1}^{n} m^k + \epsilon_2$$

$$\sigma = 3\epsilon_6 + \epsilon_2 = 3(\epsilon_1 + \epsilon_5) + \epsilon_2 =$$
$$= 3(\epsilon_1 + (R_{\max}\epsilon_3 + \gamma\epsilon_4)\prod_{k=1}^{n} m^k) + \epsilon_2 =$$
$$= 3\epsilon_1 + 3R_{\max}\epsilon_3 \prod_{k=1}^{n} m^k + 3\gamma\epsilon_4 \prod_{k=1}^{n} m^k + \epsilon_2$$

Applying the theorem 3.14 we get the implication in question.

□

3.4 Discussion and Experimental Estimations

Let's consider the conditions of the theorem 3.15 in detail.

For each component of the solution $x^k_{sh}(t)$ there are only two possibilities:

1. for any $t \in [0, \infty)$ $x^k_{sh}(t)$ remains bounded from 0 on some value $\delta > 0$
2. $x^k_{sh}(t)$ comes arbitrarily close to 0

In the first case we can reduce ϵ_1 arbitrarily by increasing T (h belongs to $C1^k_s$ in this case).

In the second case if the condition on ϵ_1 for class $C1^k_s$ holds[4] — h belongs to $C1^k_s$ otherwise to $C2^k_s$.

On fixing T and ϵ_1 we get ϵ_2 estimations automatically.

ϵ_3 and ϵ_4 are much more difficult to deal with...

In general the systems of differential equations can be solved:

1. analytically (solution in explicit form)
2. qualitatively (with the use of vector fields)
3. numerically (numerical methods, e.g., Runge-Kutta [33])

It is hopeless to try to solve the system of such complexity as 3.2 by the first two approaches and therefore a proof that its solutions satisfy the prerequisites of the theorem 3.15 seems to us non-trivial. Till now we have managed to obtain ϵ_3 and ϵ_4 estimations only experimentally.

In table 3.1 estimations of average relative $\overline{\epsilon_{5^k_{sh}}}$ and average relative $\overline{\epsilon_5}$ are presented for different game classes (with different number of states, agents and actions). The averages are calculated for 100 games of each class and

[4] We assume that $\left|\frac{1}{T}\ln x^k_{sh}(0)\right| < \epsilon_1$ here.

$T = 1000$. Initially all actions were assigned equal probabilities. The games are generated with the use of Gamut [77] with uniformly distributed payoffs from interval $[-100, 100]$. Transition probabilities were also derived from uniform distribution. Discount factor $\gamma = 0.9$. As we can see the preconditions of the theorem 3.15 hold with a quite acceptable accuracy for all the classes.

Table 3.1. Experimental Estimations

States	*Agents*	*Actions*	$\overline{\epsilon_{5^k_{sh}}}$	$\overline{\epsilon_5}$
2	2	2	0.08%	0.24%
2	2	3	0.20%	0.36%
2	2	5	0.16%	0.25%
2	2	10	0.48%	0.73%
2	3	2	0.18%	0.85%
2	3	3	0.68%	1.74%
2	5	2	1.80%	4.36%
5	2	2	0.00%	0.04%
5	2	3	0.14%	0.22%
5	2	5	0.10%	0.14%
5	3	3	0.35%	1.58%
10	2	2	0.02%	0.06%

3.5 Conclusion

This chapter provides the necessary theoretical basis for Nash-RD approach developed in chapter 4. Nash-RD allows to calculate stationary Nash equilibria of general-sum discounted stochastic games with a given accuracy and is based on multi-population replicator dynamics for discounted stochastic games introduced in this chapter.

4

Nash-RD Approach: Algorithms

In this chapter we are introducing an approach Nash-RD[1] that allows to compute stationary Nash equilibria of general-sum discounted stochastic games with a given accuracy. The algorithms are proposed for the case when the corresponding discounted stochastic games are known from the very beginning as well as for the case when the games are being learned during the interaction with the environment. The developed algorithms are compared with the existing methods. The experiments have shown that with the use of our approach much higher percentage of general-sum discounted stochastic games could be solved.

The results of this chapter were partly published in [9], [7] and [8].

The chapter is organized as follows. In section 4.1 we dwell on the assumptions we made in order to propose the algorithms for the case when the games are known in section 4.2.1 and for the case when the games are being learned in section 4.3.1. The analysis of the results of experiments is presented in sections 4.2.2 and 4.3.2 correspondingly. The complexity of the approach is studied in section 4.4. Section 4.5 is devoted to the analysis of the unexpected success of the algorithms. Few cases when the approach failed to converge are also examined in section 4.5.

4.1 Assumptions

In sections 4.2.1 and 4.3.1 we propose algorithms based on theorem 3.15 for the cases, when the corresponding discounted stochastic games are known from the very beginning and being learned through interaction with the environment.

To propose the algorithms we have to make an assumption that till now we have managed to confirm only experimentally in most of the cases, namely:

[1] Maintaining the tradition the name reflects the result — the approximation of a Nash equilibrium as well as the approach — replicator dynamics.

Assumption 1 *The more accurate approximation of Nash equilibrium we choose as an initial condition for our system 3.2 the more precisely the prerequisites of the theorem 3.15 hold and the more accurate approximation of Nash equilibrium we get*[2]*.*

Remark 4.1. In the neighborhood of a Nash equilibrium (that corresponds according to theorem 3.9 to a stationary state[3] of the system of nonlinear differential equations 3.2) we could give some arguments for validity of the above assumption.

Let $\mathbf{x}_0$ be an equilibrium (a stationary state) of the system of nonlinear differential equations

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x})$$

and let $\mathbf{z}(t) = \mathbf{x}(t) - \mathbf{x}_0$.

Then

$$\frac{d\mathbf{z}}{dt} = \frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}_0 + \mathbf{z})$$

Obviously, $\mathbf{z}(t) = 0$ is an equilibrium.

Lemma 4.2. *[29] Let* $\mathbf{f}(\mathbf{x})$ *have two continuous partial derivatives with respect to each of its variables* $x_1, \ldots, x_n$*. Then* $\mathbf{f}(\mathbf{x}_0 + \mathbf{z})$ *can be written in the form*

$$\mathbf{f}(\mathbf{x}_0 + \mathbf{z}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{A}\mathbf{z} + \mathbf{g}(\mathbf{z})$$

where[4]

$$\mathbf{A} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x}_0)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_n(\mathbf{x}_0)}{\partial x_n} \end{pmatrix}$$

and $\frac{\mathbf{g}(\mathbf{z})}{\max\{|z_1|,\ldots,|z_n|\}}$ *is a continuous function of* $\mathbf{z}$ *which vanishes for* $\mathbf{z} = 0$*.*

Proof. [29] Lemma directly follows from Taylor's theorem according to which each component $f_j(\mathbf{x}_0 + \mathbf{z})$ of $\mathbf{f}(\mathbf{x}_0 + \mathbf{z})$ can be represented in the following form

$$f_j(\mathbf{x}_0 + \mathbf{z}) = f_j(\mathbf{x}_0) + \frac{\partial f_j(\mathbf{x}_0)}{\partial x_1} z_1 + \ldots + \frac{\partial f_j(\mathbf{x}_0)}{\partial x_n} z_n + g_j(\mathbf{z})$$

where $\frac{g_j(\mathbf{z})}{\max\{|z_1|,\ldots,|z_n|\}}$ is a continuous function of $\mathbf{z}$ equal to 0 at $\mathbf{z} = 0$.

[2] In other words we demand that the resulting sequence of ε-equilibria converges to a Nash equilibrium.

[3] Stationary state of system of differential equations is also called an equilibrium of this system.

[4] $\mathbf{A}$ is called the Jacobian matrix of $\mathbf{f}$.

Therefore,

$$\mathbf{f}(\mathbf{x}_0 + \mathbf{z}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{A}\mathbf{z} + \mathbf{g}(\mathbf{z})$$

where

$$\mathbf{A} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_1(\mathbf{x}_0)}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_n(\mathbf{x}_0)}{\partial x_1} & \cdots & \frac{\partial f_n(\mathbf{x}_0)}{\partial x_n} \end{pmatrix}$$

□

Thus, if $\mathbf{f}(\mathbf{x})$ has two continuous partial derivatives with respect to each of its variables $x_1, \ldots, x_n$, then

$$\frac{d\mathbf{z}}{dt} = \mathbf{A}\mathbf{z} + \mathbf{g}(\mathbf{z})$$

where $\frac{\mathbf{g}(\mathbf{z})}{\max\{|z_1|,\ldots,|z_n|\}}$ is a continuous function of $\mathbf{z}$ that can be omitted in the neighborhood of equilibrium $\mathbf{z} = 0$.

So in some neighborhood of equilibrium $\mathbf{z} = 0$ our system will behave as system of linear differential equations [143]:

$$\frac{d\mathbf{z}}{dt} = \mathbf{A}\mathbf{z}$$

Every solution of such system is of the following form [29], [67]:

$$\begin{aligned}
\mathbf{x}(t) = {} & c_1 e^{\lambda_1 t}\mathbf{v}^1 + c_2 e^{\lambda_2 t}\mathbf{v}^2 + \\
& + \ldots + \\
& + c_m e^{\alpha_m t}(\mathbf{v}^m \cos\beta_m t - \mathbf{v}^{m+1}\sin\beta_m t) + \\
& + c_{m+1} e^{\alpha_m t}(\mathbf{v}^m \sin\beta_m t + \mathbf{v}^{m+1}\cos\beta_m t) + \\
& + c_{m+2} e^{\alpha_{m+2} t}(\mathbf{v}^{m+2} \cos\beta_{m+2} t - \mathbf{v}^{m+3}\sin\beta_{m+2} t) + \\
& + c_{m+3} e^{\alpha_{m+2} t}(\mathbf{v}^{m+2} \sin\beta_{m+2} t + \mathbf{v}^{m+3}\cos\beta_{m+2} t) + \\
& + \ldots + \\
& + c_l e^{\lambda_l t}(\mathbf{v}^l + t(\mathbf{A} - \lambda_l \mathbf{I})\mathbf{v}^l) + \\
& + \ldots + \\
& + c_k e^{\lambda_k t}(\mathbf{v}^k + t(\mathbf{A} - \lambda_k \mathbf{I})\mathbf{v}^k + \frac{t^2}{2!}(\mathbf{A} - \lambda_k \mathbf{I})^2 \mathbf{v}^k) + \\
& + \ldots
\end{aligned}$$

where $\lambda_1, \lambda_2, \ldots$ are distinct real eigenvalues of $\mathbf{A}$ with corresponding eigenvectors $\mathbf{v}^1, \mathbf{v}^2, \ldots$, $\lambda_m = \alpha_m \pm i\beta_m, \lambda_{m+2} = \alpha_{m+2} \pm i\beta_{m+2}, \ldots$ are complex eigenvalues of $\mathbf{A}$ with corresponding eigenvectors $\mathbf{v}^m + i\mathbf{v}^{m+1}, \mathbf{v}^{m+2} +$

$i\mathbf{v}^{m+3}, \ldots$ and where the rest components[5] appear in case the number of linear independent eigenvectors is less than n.

The solution $e^{\lambda_i t}\mathbf{v}^i$ converges to (diverges from) the equilibrium along vector $\mathbf{v}^i$ when $\lambda_i < 0$ ($\lambda_i > 0$).

The solution

$$c_m e^{\alpha_m t}(\mathbf{v}^m \cos\beta_m t - \mathbf{v}^{m+1}\sin\beta_m t) + c_{m+1}e^{\alpha_m t}(\mathbf{v}^m \sin\beta_m t + \mathbf{v}^{m+1}\cos\beta_m t)$$

depicts a spiral (spanned by vectors $\mathbf{v}^m$ and $\mathbf{v}^{m+1}$) along which the solution converges to (diverges from) the equilibrium in accordance with the sign of α_m.

In case of converging orbits (as well as in case of some diverging spirals) the averages will be nearer to the equilibrium than the initial point.

4.1.1 Jacobian Matrix of Replicator Dynamics in Stochastic Games

Let us first calculate the Jacobian matrix of the following system of nonlinear differential equations:

$$\frac{dx_k}{dt} = f_k(\mathbf{x})x_k \qquad\qquad k = 1, 2, \ldots, n$$

$$\mathbf{A} = \begin{pmatrix} \frac{\partial f_1(\mathbf{x})x_1}{\partial x_1} & \frac{\partial f_1(\mathbf{x})x_1}{\partial x_2} & \cdots & \frac{\partial f_1(\mathbf{x})x_1}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x})x_2}{\partial x_1} & \frac{\partial f_2(\mathbf{x})x_2}{\partial x_2} & \cdots & \frac{\partial f_2(\mathbf{x})x_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n(\mathbf{x})x_n}{\partial x_1} & \frac{\partial f_n(\mathbf{x})x_n}{\partial x_2} & \cdots & \frac{\partial f_n(\mathbf{x})x_n}{\partial x_n} \end{pmatrix} =$$

$$= \begin{pmatrix} f_1(\mathbf{x}) + x_1\frac{\partial f_1(\mathbf{x})}{\partial x_1} & x_1\frac{\partial f_1(\mathbf{x})}{\partial x_2} & \ldots & x_1\frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ x_2\frac{\partial f_2(\mathbf{x})}{\partial x_1} & f_2(\mathbf{x}) + x_2\frac{\partial f_2(\mathbf{x})}{\partial x_2} & \ldots & x_2\frac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ x_n\frac{\partial f_n(\mathbf{x})}{\partial x_1} & x_n\frac{\partial f_n(\mathbf{x})}{\partial x_2} & \ldots & f_n(\mathbf{x}) + x_n\frac{\partial f_n(\mathbf{x})}{\partial x_n} \end{pmatrix} =$$

$$= \begin{pmatrix} f_1(\mathbf{x}) & 0 & \ldots & 0 \\ 0 & f_2(\mathbf{x}) & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & f_n(\mathbf{x}) \end{pmatrix} + \begin{pmatrix} x_1 & 0 & \ldots & 0 \\ 0 & x_2 & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & x_n \end{pmatrix} \cdot \begin{pmatrix} \frac{\partial f_1(\mathbf{x})}{\partial x_1} & \frac{\partial f_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\mathbf{x})}{\partial x_n} \\ \frac{\partial f_2(\mathbf{x})}{\partial x_1} & \frac{\partial f_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\mathbf{x})}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n(\mathbf{x})}{\partial x_1} & \frac{\partial f_n(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial f_n(\mathbf{x})}{\partial x_n} \end{pmatrix}$$

$$= \mathbf{F} + \mathbf{X} \cdot \mathbf{J}$$

Since replicator dynamics in stochastic games 3.2 can be written in the above form:

[5] $\lambda_l, \lambda_k, \ldots$ can be complex eigenvalues here.

$$\frac{dx^k_{sh}}{dt} = f^k_{sh}(\mathbf{x})x^k_{sh} = \left[\vartheta^k_{sh}(\mathbf{x}) - v^k_s(\mathbf{x})\right] x^k_{sh} \qquad k \in K,\ s \in S,\ h \in A^k$$

its Jacobian matrix will be equal to:

$$\mathbf{A} = \mathbf{F} + \mathbf{X} \cdot \mathbf{J}$$

where $\mathbf{F}$, $\mathbf{X}$ and $\mathbf{J}$ are represented by matrices 4.1, 4.2 and 4.3.

The following statements hold:

If $s \neq o$

$$\frac{\partial f^k_{sh}(\mathbf{x})}{\partial x^l_{op}} = \gamma \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} \sum_{s' \in S} p(s'|s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \cdot \frac{\partial v^k_{s'}(\mathbf{x})}{\partial x^l_{op}} \prod_{i=1, i \neq k}^{n} x^i_{sa^i} - \frac{\partial v^k_s(\mathbf{x})}{\partial x^l_{op}}$$

If $s = o$ and $k = l$

$$\frac{\partial f^k_{sh}(\mathbf{x})}{\partial x^l_{op}} = \gamma \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^n \in A^n} \sum_{s' \in S} p(s'|s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^n) \cdot \frac{\partial v^k_{s'}(\mathbf{x})}{\partial x^l_{op}} \prod_{i=1, i \neq k}^{n} x^i_{sa^i} - \frac{\partial v^k_s(\mathbf{x})}{\partial x^l_{op}}$$

If $s = o$ and $k \neq l$[6]

$$\begin{aligned}\frac{\partial f^k_{sh}(\mathbf{x})}{\partial x^l_{op}} = & \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^{l-1} \in A^{l-1}} \sum_{a^{l+1} \in A^{l+1}} \cdots \sum_{a^n \in A^n} \\ & r^k(s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^{l-1}, p, a^{l+1}, \ldots, a^n) \cdot \prod_{i=1, i \neq k, i \neq l}^{n} x^i_{sa^i} \\ & + \gamma \sum_{a^1 \in A^1} \sum_{a^2 \in A^2} \cdots \sum_{a^{k-1} \in A^{k-1}} \sum_{a^{k+1} \in A^{k+1}} \cdots \sum_{a^{l-1} \in A^{l-1}} \sum_{a^{l+1} \in A^{l+1}} \cdots \sum_{a^n \in A^n} \sum_{s' \in S} \\ & p(s'|s, a^1, a^2, \ldots, a^{k-1}, h, a^{k+1}, \ldots, a^{l-1}, p, a^{l+1}, \ldots, a^n) \cdot v^k_{s'}(\mathbf{x}) \prod_{i=1, i \neq k, i \neq l}^{n} x^i_{sa^i}\end{aligned}$$

[6] Here we assume that $k < l$.

$$+\gamma \sum_{a^1\in A^1}\sum_{a^2\in A^2}\cdots\sum_{a^{k-1}\in A^{k-1}}\sum_{a^{k+1}\in A^{k+1}}\cdots\sum_{a^n\in A^n}\sum_{s'\in S}$$
$$p(s'|s,a^1,a^2,\ldots,a^{k-1},h,a^{k+1},\ldots,a^n)\cdot\frac{\partial v^k_{s'}(\mathbf{x})}{\partial x^l_{op}}\prod_{i=1,i\neq k}^{n}x^i_{sa^i}$$
$$-\frac{\partial v^k_s(\mathbf{x})}{\partial x^l_{op}}$$

4.2 When the Games are Known

4.2.1 Nash-RD Algorithm

Having made the above assumption we can propose an iterative algorithm for calculating Nash equilibria of discounted stochastic games with some given accuracy ε (see algorithm 10).

Algorithm 10 Nash-RD algorithm

Input: accuracy ε, T, initial policy profile x^0
$x(0) \leftarrow x^0$
while $x(0)$ doesn't constitute ε-equilibrium **do**
Numerically find the solution of the system 3.2 (see below) through the point $x(0)$ on the interval $[0,T]$:

$$\frac{dx^k_{sh}}{dt} = \left[\vartheta^k_{sh}(x) - v^k_s(x)\right]x^k_{sh} \qquad k\in K,\ s\in S,\ h\in A^k$$

where

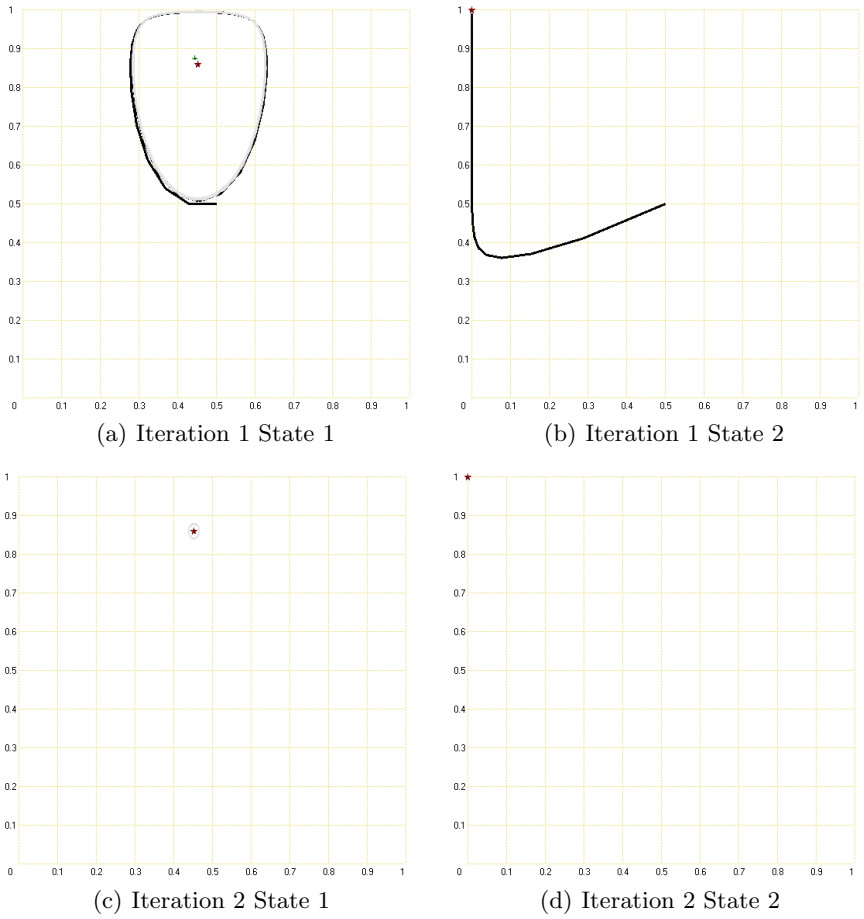
$$\vartheta^k_{sh}(x) = r^k(s,\chi) + \gamma\sum_{s'\in S}p(s'|s,\chi)v^k_{s'}(x)$$

and χ denotes the profile equal to x but where player k always plays action h in state s
Let the initial point be $x^k_{sh}(0) = \frac{1}{T}\int_0^T x^k_{sh}(t)dt$
end while

An example of its performance on a 2-state 2-agent 2-action discounted stochastic game is presented in the figure 4.1. In each figure the probabilities assigned to the first actions of the first and the second agents are presented as xy-plot (it is quite descriptive since the probabilities of the second actions are equal to one minus probabilities of the first ones). The solutions are lighter at the end of $[0,T]$ interval. The precise Nash equilibrium is designated by a star and the average $\frac{1}{T}\int_0^T x^k_{sh}(t)dt$ for each iteration — by a cross[7]. The algorithm

[7] In figures 4.1(b), 4.1(c) and 4.1(d) the cross and the star coincide.

converges to the Nash equilibrium with the given accuracy $\varepsilon = 0.001$ in two iterations.



(a) Iteration 1 State 1

(b) Iteration 1 State 2

(c) Iteration 2 State 1

(d) Iteration 2 State 2

Fig. 4.1. Convergence of Algorithm 10

4.2.2 Experimental Results

When the models (discounted stochastic games) are known from the very beginning two approaches: nonlinear optimization [56] and stochastic tracing procedure [69] have been proved to find Nash equilibria in the general case.

Table 4.1. Results of Experiments When the Games are Known

States	*Agents*	*Actions*	*CONOPT*	*KNITRO*	*MINOS*	*SNOPT*	*TP*	*Nash-RD*
2	2	2	58%	65%	66%	67%	83%	100%
2	2	3	39%	39%	41%	46%	86%	98%
2	2	5	16%	30%	19%	20%	79%	90%
2	2	7	12%	18%	12%	12%	67%	93%
2	2	10	8%	10%	5%	2%	—	90%
2	3	2	44%	47%	51%	43%	82%	92%
2	3	3	22%	33%	28%	27%	81%	92%
2	3	5	21%	25%	17%	13%	—	90%
2	3	7	7%	13%	5%	5%	—	92%
2	5	2	34%	44%	27%	39%	82%	93%
2	5	3	20%	26%	11%	21%	—	94%
2	7	2	21%	31%	15%	33%	—	87%
5	2	2	36%	37%	41%	40%	83%	100%
5	2	3	17%	15%	15%	20%	59%	97%
5	2	5	1%	5%	2%	1%	44%	91%
5	2	7	1%	7%	0%	0%	—	82%
5	3	2	18%	20%	11%	12%	77%	85%
5	3	3	2%	4%	4%	6%	66%	79%
5	5	2	9%	13%	9%	8%	—	72%
10	2	2	12%	16%	24%	23%	68%	100%
10	2	3	2%	3%	3%	1%	35%	98%
10	3	2	5%	7%	1%	1%	70%	82%

Four nonlinear optimization algorithms: CONOPT [81], [49], KNITRO [82], [34], MINOS [83], SNOPT [84], [61] are compared with stochastic tracing procedure (TP)[8] and the developed Nash-RD algorithm.

We chose the following settings for our algorithm: $T = 1000$. The solution of 3.2 was numerically approximated with classic Runge-Kutta fourth-order method [102] where the step size h was set to 0.01. We used Gauss-Jordan elimination [140] to find inverse matrix for $v(x)$ computation. First we initialized x so that each action of every agent had equal probability. The number of iterations was restricted to 500 and after every 100 iterations we started with a new random initial policy profile x. The dependency graphics of the games solved with given accuracy $\varepsilon = 0.001$ on the number of iterations are presented in figures 4.3 – 4.5. The internal parameter of stochastic tracing procedure was set to 10^{-8}. We increased the maximum processing time for nonlinear optimization algorithms but left default values for other parameters [85].

The percentage of games for which we managed to find Nash equilibria with the use of the above approaches with given accuracy $\varepsilon = 0.001$ (relative

[8] We are infinitely grateful to P. Jean-Jacques Herings and Ronald Peeters who were so kind as to render their original stochastic tracing procedure.

accuracy $\varepsilon = 10^{-5}\%$) is presented in the corresponding columns of table 4.1. The percentage is calculated for 100 games of each class that differs in the number of states, agents and actions. The games are generated with the use of Gamut [118] with uniformly distributed payoffs from interval $[-100, 100]$. Transition probabilities were also derived from uniform distribution. Discount factor $\gamma = 0.9$. As it can be seen from the table 4.1, the developed algorithm showed the best results for all game classes. We guess, that the main reason is that nonlinear optimizers are inclined to get stuck in local optima, whereas only global optima constitute Nash equilibria. As for stochastic tracing procedure, if we had set internal parameter to less than 10^{-8}, we would very probably have got solutions to higher percentage of stochastic games with given accuracy $\varepsilon = 10^{-3}$ but there were some games ("—" in the table) whose processing has taken us 5 hours already[9].

4.3 When the Games are Being Learned

4.3.1 Nash-RD Algorithm

In reinforcement learning case we assume that the agents can observe each others immediate rewards and thus modeling each other get access to precise policies. Each agent finds a Nash equilibrium by itself with the use of algorithm 11. Discordance being impossible, the agents converge to the same Nash equilibrium.

An example of its performance on a 2-state 2-agent 2-action discounted stochastic game is presented in the figure 4.2. In each figure the probabilities assigned to the first actions of the first and the second agents are presented as xy-plot (it is quite descriptive since the probabilities of the second actions are equal to one minus probabilities of the first ones). The solutions are lighter at the end of $[0, T]$ interval. The precise Nash equilibrium is designated by a star and the average $\frac{1}{T}\int_0^T x_{sh}^k(t)dt$ for each iteration — by a cross[10]. Since the agents in reinforcement learning don't know either transition probabilities or reward functions and they learn them online the first policies are quite random. The algorithm converges in self-play to the Nash equilibrium with the given relative accuracy $\varepsilon = 1\%$ in two iterations.

4.3.2 Experimental Results

Since the agents in reinforcement learning don't know either transition probabilities or reward functions they have to approximate the model somehow. We tested our algorithm as a model-based version with epsilon greedy exploration (the agents learn the model and pursue the best learned policy so far in

[9] Intel Celeron, 1.50GHz, 504 MB of RAM.
[10] In figures 4.2(b) and 4.2(d) the cross and the star coincide.

Algorithm 11 Nash-RD algorithm for the player i

Input: accuracy ε, T, initial policy profile x^0
$x(0) \leftarrow x^0$
while $x(0)$ doesn't constitute ε-equilibrium **do**
Numerically find the solution of the system 3.3 (see below) through the point $x(0)$ on the interval $[0, T]$ (updating model of the game in parallel):

$$\frac{dx^k_{sh}}{dt} = \left[\vartheta^k_{sh}(x) - \widetilde{v}^k_s(x)\right] x^k_{sh} \qquad k \in K,\ s \in S,\ h \in A^k$$

where

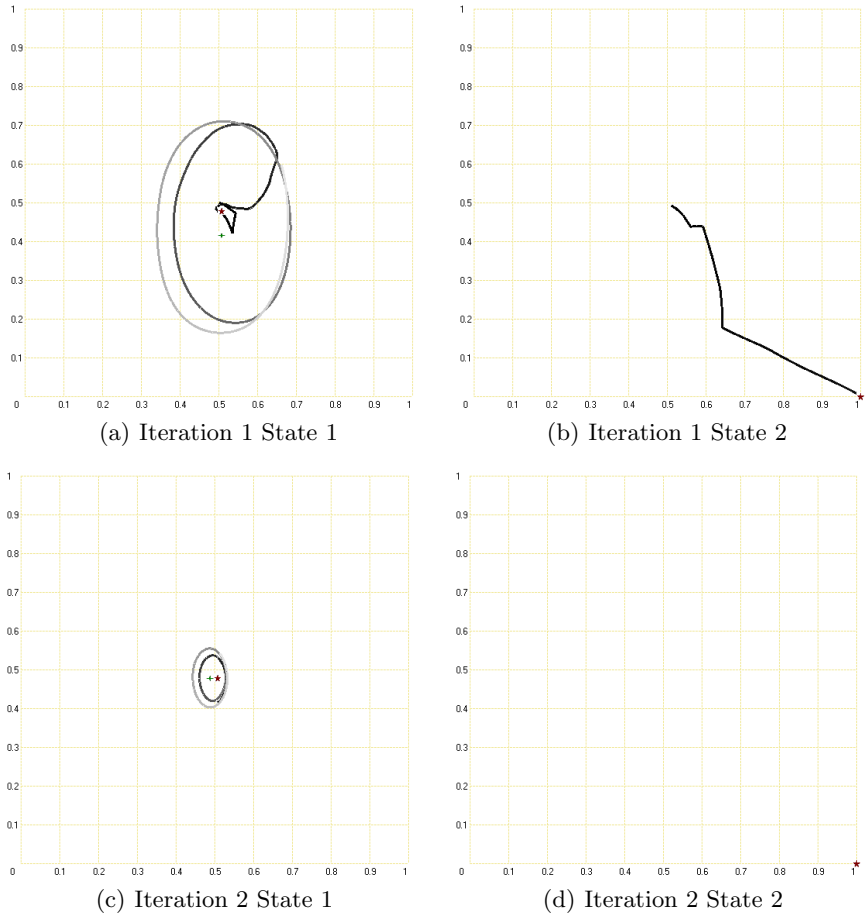
$$\vartheta^k_{sh}(x) = r^k(s, \chi) + \gamma \sum_{s' \in S} \widetilde{p}(s'|s, \chi)\widetilde{v}^k_{s'}(x)$$

and χ denotes the profile equal to x but where player k always plays action h in state s and $\widetilde{}$ indicates that the agent uses an approximation of transition probabilities $\widetilde{p}$ instead of the actual transition probabilities p for calculation of the corresponding values
Let the initial point be $x^k_{sh}(0) = \frac{1}{T}\int_0^T x^k_{sh}(t)dt$
end while

Table 4.2. Results of Experiments When the Games are Being Learned

States	*Agents*	*Actions*	*Tr*	*Iterations*	*Nash-RD*
2	2	2	8	11.23	98%
2	2	3	18	9.43	95%
2	2	5	50	18.60	90%
2	2	10	200	38.39	94%
2	3	2	16	16.03	87%
2	3	3	54	30.64	91%
2	5	2	64	27.79	87%
5	2	2	80	31.60	83%
5	2	3	180	52.26	93%
5	2	5	500	62.74	91%
5	3	3	540	85.83	75%
10	2	2	360	69.68	82%

the most of cases (we chose — 90% of cases) and explore the environment in 10% of cases). Other internal settings and testbed were the same as in section 4.2.2. The results of the experiments are presented in table 4.2. The number of independent transitions to be learned can be calculated by the formula $Tr = N(N-1)\prod_{k=1}^{n} m^k$ and is presented in the corresponding column for each game class. In column "Iterations" the average number of iterations necessary to find a Nash equilibrium with relative accuracy $\varepsilon = 1\%$ is presented. And in the last column — the percentage of games for which we managed to find a Nash equilibrium with the given relative accuracy $\varepsilon = 1\%$ in less than 500 iterations.



(a) Iteration 1 State 1 (b) Iteration 1 State 2

(c) Iteration 2 State 1 (d) Iteration 2 State 2

Fig. 4.2. Convergence of Algorithm 11

In general one can see the following trend: the larger is the model the more iterations the agents require to find a 1%-equilibrium, and the oftener they fail to converge to this equilibrium in less than 500 iterations.

In addition to the reasons analyzed in section 4.5, the failures are also caused by the agents' incapability to approximate large models to the necessary accuracy (their approximations of transition probabilities are too imprecise — they explore the environment only in 10% of cases each and the transition probabilities of some combinations of actions remain very poorly estimated) and as a result they can't find an equilibrium or converge to it more slowly (let us not forget that the accuracy of transition probabilities acts as a relative factor and comes to ε estimation of theorem 3.15 multiplied by the maximal discounted value). In order to decrease the average number of

iterations and to increase the percentage of solved games it appears promising to test a version of the algorithm with a more intensive exploration stage (first learn the model to some given precision and only then act according to the policy found by the algorithm and keep on learning in parallel). For instance, it can be achieved by setting exploration rate to higher values at the beginning.

4.4 Complexity

Under our settings the complexity of solution approximation step is obviously $O\left(\frac{T}{h} \cdot N^2 \cdot \max\left(N, n \cdot \prod_{k=1}^{n} m^k\right)\right)$[11]. It is impossible to estimate the complexity of the whole algorithm till we single out classes of stochastic games for which the validity of the assumption 1 could be proved and the upper bound for the number of iterations could be found for a given accuracy. The experimental estimations of the necessary number of iterations are shown in figures 4.3 through 4.5 where y-values correspond to the percentage of the games solved in less than x number of iterations. The execution time of one iteration could be found in table 4.3.

4.5 Discussion

In this section we would like to dwell on two main questions:

- Why doesn't our approach converge in 100% cases?
- Why does our approach find Nash equilibria whereas all the attempts to propose an algorithm that converges to a Nash equilibrium in reinforcement learning community during the last 15 years failed?[12]

First, we must notice, that the remark 4.1 is totally useless for the analysis of the first question, since the calculated Nash equilibria are never near enough to the initial point.

As a result of thorough examination we discovered two reasons Nash-RD failed to find Nash equilibrium:

- The assumption 1 doesn't hold — the game and the initial point are such that the conditions 2.a) and 2.b) of theorem 3.15 do not hold with necessary accuracy and the averages of solution components do not sequentially converge to a Nash equilibrium.

[11] See [3] or [41] for introduction into algorithm analysis techniques.

[12] To be precise, the success was claimed from time to time, but there was in general no (very weak) theoretical foundation and no thorough testing (the authors declared the convergence of their algorithms on a pair of games). Nash-Q algorithm has been proved to converge to a Nash equilibrium in self-play [90] for strictly competitive and strictly cooperative games under additional very restrictive condition that all equilibria encountered during learning stage are unique.

Table 4.3. Execution Time

States	*Agents*	*Actions*	*One Iteration (seconds)*	*Nash-RD*
2	2	2	0.131	100%
2	2	3	0.263	98%
2	2	5	0.753	90%
2	2	7	1.616	93%
2	2	10	3.891	90%
2	3	2	0.250	92%
2	3	3	0.856	92%
2	3	5	4.688	90%
2	3	7	16.416	92%
2	5	2	1.269	93%
2	5	3	12.022	94%
2	7	2	7.044	87%
5	2	2	0.644	100%
5	2	3	1.153	97%
5	2	5	2.959	91%
5	2	7	5.825	82%
5	3	2	1.138	85%
5	3	3	3.163	79%
5	5	2	4.553	72%
10	2	2	2.753	100%
10	2	3	4.269	98%
10	3	2	4.197	82%
15	2	2	6.747	100%

- The solution gets out of the border of Φ as a result of Runge-Kutta approximation. The numerical solution is exposed to two types of errors: round-off error and discretization error. Round-off errors are caused by floating point arithmetics on a computer and increase in the number of arithmetical operations. The total accumulated discretization error of fourth-order Runge-Kutta method with step size h is equal to $O(h^4)$ [36].

As regards the second question, for the last 15 years all the attempts were directed to extending Bellman optimality equation (the base of many reinforcement learning algorithms in isolated environments) for multi-agent environments, and not to extending approaches to calculating Nash equilibria of matrix games for multi-state case (see figure 3.1). The algorithms that rashly changed pure stationary strategies (e.g., Q-learning and JAL) were bound to fail in general case since discounted stochastic games are not guaranteed to possess Nash equilibria in pure stationary strategies. The algorithms that gradually changed the policies (e.g., PHC and WoLF-PHC) must have passed the Nash equilibria (see [50], [35], [114] and [74] to get a notion how careful one must be discretizing the continuous replicator dynamics).

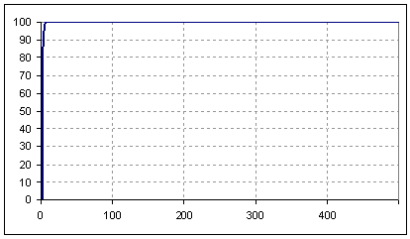
4.6 Conclusion

This chapter is devoted to an actual topic of extending reinforcement learning approach for multi-agent systems. An approach based on multi-population replicator dynamics for discounted stochastic games introduced in chapter 3 is developed. A formal proof of its convergence with a given accuracy to a Nash equilibrium is given under some assumptions. Thorough testing showed that the assumptions necessary for the formal convergence hold in quite many cases. We claim that it is the first algorithm that converges to a Nash equilibrium for high percentage of general-sum discounted stochastic games when the latter are being learned by interaction with the environment. When the games are known from the very beginning the experiments have shown that with the use of the developed approach higher percentage of general-sum discounted stochastic games could be solved.

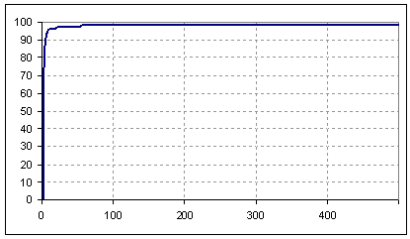
$$
\mathrm{F} = \begin{pmatrix}
f^1_{s_1 a_1^1}(\mathbf{x}) & \dots & 0 & \dots & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & f^1_{s_1 a_{m^1}^1}(\mathbf{x}) & \dots & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & \dots & f^n_{s_1 a_1^n}(\mathbf{x}) & \dots & 0 & \dots & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & \dots & 0 & \dots & f^n_{s_1 a_{m^n}^n}(\mathbf{x}) & \dots & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & \dots & 0 & \dots & 0 & \dots & f^1_{s_N a_1^1}(\mathbf{x}) & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & \dots & 0 & \dots & 0 & \dots & 0 & \dots & f^1_{s_N a_{m^1}^1}(\mathbf{x}) \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & \dots & 0 & \dots & 0 & \dots & 0 & \dots & 0 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
0 & \dots & 0 & \dots & 0 & \dots & 0 & \dots & 0 & \dots & f^n_{s_N a_{m^n}^n}(\mathbf{x})
\end{pmatrix} \tag{4.1}
$$

$$
\mathrm{X} = \begin{pmatrix}
x^1_{s_1 a^1_1} & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\
\cdots & \cdots & x^1_{s_1 a^1_{m^1}} & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & x^n_{s_1 a^n_1} & \cdots & x^n_{s_1 a^n_{m^n}} & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & \cdots & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & x^1_{s_N a^1_1} & \cdots & x^1_{s_N a^1_{m^1}} & \cdots & \cdots & \cdots & \cdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & x^n_{s_N a^n_1} & \cdots & x^n_{s_N a^n_{m^n}}
\end{pmatrix}
\tag{4.2}
$$

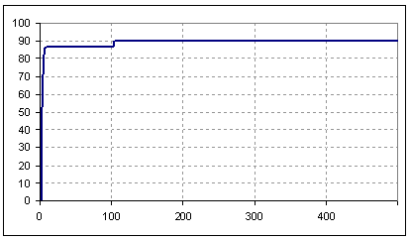
$$
\mathrm{J} = \begin{pmatrix}
\frac{\partial f^1_{s_1 a^1_1}(\mathbf{x})}{\partial x^1_{s_1 a^1_1}} & \cdots & \frac{\partial f^1_{s_1 a^1_1}(\mathbf{x})}{\partial x^1_{s_1 a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^1_{s_1 a^1_1}(\mathbf{x})}{\partial x^n_{s_1 a^n_1}} & \cdots & \frac{\partial f^1_{s_1 a^1_1}(\mathbf{x})}{\partial x^n_{s_1 a^n_{m^n}}} & \cdots\cdots\cdots\cdots & \frac{\partial f^1_{s_1 a^1_1}(\mathbf{x})}{\partial x^1_{s_N a^1_1}} & \cdots & \frac{\partial f^1_{s_1 a^1_1}(\mathbf{x})}{\partial x^1_{s_N a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^1_{s_1 a^1_1}(\mathbf{x})}{\partial x^n_{s_N a^n_1}} & \cdots & \frac{\partial f^1_{s_1 a^1_1}(\mathbf{x})}{\partial x^n_{s_N a^n_{m^n}}} \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{\partial f^1_{s_1 a^1_{m^1}}(\mathbf{x})}{\partial x^1_{s_1 a^1_1}} & \cdots & \frac{\partial f^1_{s_1 a^1_{m^1}}(\mathbf{x})}{\partial x^1_{s_1 a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^1_{s_1 a^1_{m^1}}(\mathbf{x})}{\partial x^n_{s_1 a^n_1}} & \cdots & \frac{\partial f^1_{s_1 a^1_{m^1}}(\mathbf{x})}{\partial x^n_{s_1 a^n_{m^n}}} & \cdots\cdots\cdots\cdots & \frac{\partial f^1_{s_1 a^1_{m^1}}(\mathbf{x})}{\partial x^1_{s_N a^1_1}} & \cdots & \frac{\partial f^1_{s_1 a^1_{m^1}}(\mathbf{x})}{\partial x^1_{s_N a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^1_{s_1 a^1_{m^1}}(\mathbf{x})}{\partial x^n_{s_N a^n_1}} & \cdots & \frac{\partial f^1_{s_1 a^1_{m^1}}(\mathbf{x})}{\partial x^n_{s_N a^n_{m^n}}} \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{\partial f^n_{s_1 a^n_1}(\mathbf{x})}{\partial x^1_{s_1 a^1_1}} & \cdots & \frac{\partial f^n_{s_1 a^n_1}(\mathbf{x})}{\partial x^1_{s_1 a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^n_{s_1 a^n_1}(\mathbf{x})}{\partial x^n_{s_1 a^n_1}} & \cdots & \frac{\partial f^n_{s_1 a^n_1}(\mathbf{x})}{\partial x^n_{s_1 a^n_{m^n}}} & \cdots\cdots\cdots\cdots & \frac{\partial f^n_{s_1 a^n_1}(\mathbf{x})}{\partial x^1_{s_N a^1_1}} & \cdots & \frac{\partial f^n_{s_1 a^n_1}(\mathbf{x})}{\partial x^1_{s_N a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^n_{s_1 a^n_1}(\mathbf{x})}{\partial x^n_{s_N a^n_1}} & \cdots & \frac{\partial f^n_{s_1 a^n_1}(\mathbf{x})}{\partial x^n_{s_N a^n_{m^n}}} \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{\partial f^n_{s_1 a^n_{m^n}}(\mathbf{x})}{\partial x^1_{s_1 a^1_1}} & \cdots & \frac{\partial f^n_{s_1 a^n_{m^n}}(\mathbf{x})}{\partial x^1_{s_1 a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^n_{s_1 a^n_{m^n}}(\mathbf{x})}{\partial x^n_{s_1 a^n_1}} & \cdots & \frac{\partial f^n_{s_1 a^n_{m^n}}(\mathbf{x})}{\partial x^n_{s_1 a^n_{m^n}}} & \cdots\cdots\cdots\cdots & \frac{\partial f^n_{s_1 a^n_{m^n}}(\mathbf{x})}{\partial x^1_{s_N a^1_1}} & \cdots & \frac{\partial f^n_{s_1 a^n_{m^n}}(\mathbf{x})}{\partial x^1_{s_N a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^n_{s_1 a^n_{m^n}}(\mathbf{x})}{\partial x^n_{s_N a^n_1}} & \cdots & \frac{\partial f^n_{s_1 a^n_{m^n}}(\mathbf{x})}{\partial x^n_{s_N a^n_{m^n}}} \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{\partial f^1_{s_N a^1_1}(\mathbf{x})}{\partial x^1_{s_1 a^1_1}} & \cdots & \frac{\partial f^1_{s_N a^1_1}(\mathbf{x})}{\partial x^1_{s_1 a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^1_{s_N a^1_1}(\mathbf{x})}{\partial x^n_{s_1 a^n_1}} & \cdots & \frac{\partial f^1_{s_N a^1_1}(\mathbf{x})}{\partial x^n_{s_1 a^n_{m^n}}} & \cdots\cdots\cdots\cdots & \frac{\partial f^1_{s_N a^1_1}(\mathbf{x})}{\partial x^1_{s_N a^1_1}} & \cdots & \frac{\partial f^1_{s_N a^1_1}(\mathbf{x})}{\partial x^1_{s_N a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^1_{s_N a^1_1}(\mathbf{x})}{\partial x^n_{s_N a^n_1}} & \cdots & \frac{\partial f^1_{s_N a^1_1}(\mathbf{x})}{\partial x^n_{s_N a^n_{m^n}}} \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{\partial f^1_{s_N a^1_{m^1}}(\mathbf{x})}{\partial x^1_{s_1 a^1_1}} & \cdots & \frac{\partial f^1_{s_N a^1_{m^1}}(\mathbf{x})}{\partial x^1_{s_1 a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^1_{s_N a^1_{m^1}}(\mathbf{x})}{\partial x^n_{s_1 a^n_1}} & \cdots & \frac{\partial f^1_{s_N a^1_{m^1}}(\mathbf{x})}{\partial x^n_{s_1 a^n_{m^n}}} & \cdots\cdots\cdots\cdots & \frac{\partial f^1_{s_N a^1_{m^1}}(\mathbf{x})}{\partial x^1_{s_N a^1_1}} & \cdots & \frac{\partial f^1_{s_N a^1_{m^1}}(\mathbf{x})}{\partial x^1_{s_N a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^1_{s_N a^1_{m^1}}(\mathbf{x})}{\partial x^n_{s_N a^n_1}} & \cdots & \frac{\partial f^1_{s_N a^1_{m^1}}(\mathbf{x})}{\partial x^n_{s_N a^n_{m^n}}} \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{\partial f^n_{s_N a^n_1}(\mathbf{x})}{\partial x^1_{s_1 a^1_1}} & \cdots & \frac{\partial f^n_{s_N a^n_1}(\mathbf{x})}{\partial x^1_{s_1 a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^n_{s_N a^n_1}(\mathbf{x})}{\partial x^n_{s_1 a^n_1}} & \cdots & \frac{\partial f^n_{s_N a^n_1}(\mathbf{x})}{\partial x^n_{s_1 a^n_{m^n}}} & \cdots\cdots\cdots\cdots & \frac{\partial f^n_{s_N a^n_1}(\mathbf{x})}{\partial x^1_{s_N a^1_1}} & \cdots & \frac{\partial f^n_{s_N a^n_1}(\mathbf{x})}{\partial x^1_{s_N a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^n_{s_N a^n_1}(\mathbf{x})}{\partial x^n_{s_N a^n_1}} & \cdots & \frac{\partial f^n_{s_N a^n_1}(\mathbf{x})}{\partial x^n_{s_N a^n_{m^n}}} \\
\vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots & & \vdots \\
\frac{\partial f^n_{s_N a^n_{m^n}}(\mathbf{x})}{\partial x^1_{s_1 a^1_1}} & \cdots & \frac{\partial f^n_{s_N a^n_{m^n}}(\mathbf{x})}{\partial x^1_{s_1 a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^n_{s_N a^n_{m^n}}(\mathbf{x})}{\partial x^n_{s_1 a^n_1}} & \cdots & \frac{\partial f^n_{s_N a^n_{m^n}}(\mathbf{x})}{\partial x^n_{s_1 a^n_{m^n}}} & \cdots\cdots\cdots\cdots & \frac{\partial f^n_{s_N a^n_{m^n}}(\mathbf{x})}{\partial x^1_{s_N a^1_1}} & \cdots & \frac{\partial f^n_{s_N a^n_{m^n}}(\mathbf{x})}{\partial x^1_{s_N a^1_{m^1}}} & \cdots\cdots & \frac{\partial f^n_{s_N a^n_{m^n}}(\mathbf{x})}{\partial x^n_{s_N a^n_1}} & \cdots & \frac{\partial f^n_{s_N a^n_{m^n}}(\mathbf{x})}{\partial x^n_{s_N a^n_{m^n}}}
\end{pmatrix} \tag{4.3}
$$



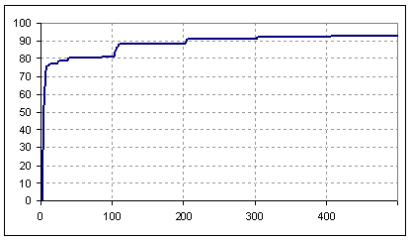
(a) 2-state 2-agent 2-action games



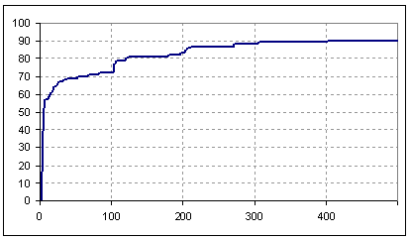
(b) 2-state 2-agent 3-action games



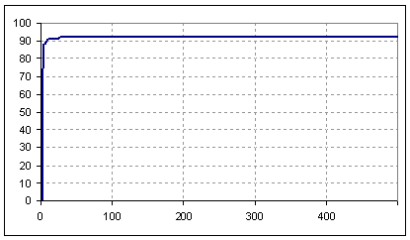
(c) 2-state 2-agent 5-action games



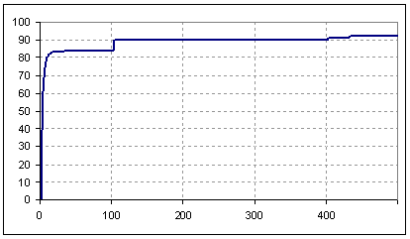
(d) 2-state 2-agent 7-action games



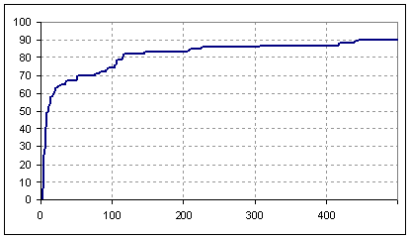
(e) 2-state 2-agent 10-action games



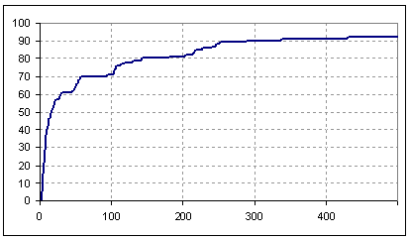
(f) 2-state 3-agent 2-action games



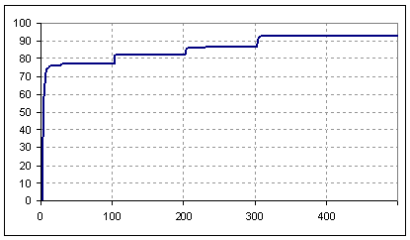
(g) 2-state 3-agent 3-action games



(h) 2-state 3-agent 5-action games

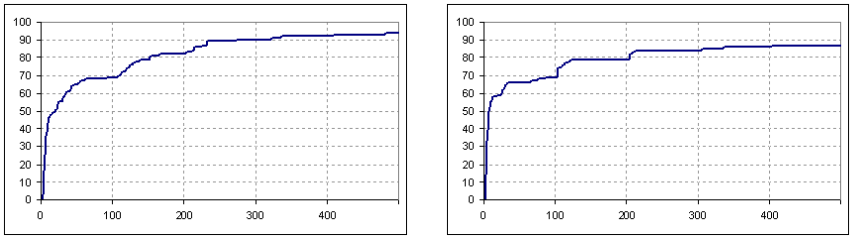


(i) 2-state 3-agent 7-action games



(j) 2-state 5-agent 2-action games

Fig. 4.3. Nash-RD: Experimental Estimations of Number of Iterations



(a) 2-state 5-agent 3-action games (b) 2-state 7-agent 2-action games

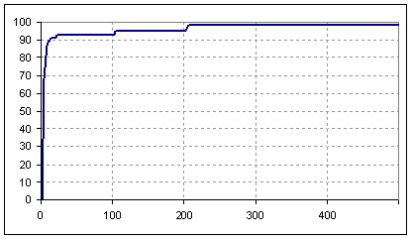
(c) 5-state 2-agent 2-action games (d) 5-state 2-agent 3-action games

(e) 5-state 2-agent 5-action games (f) 5-state 2-agent 7-action games

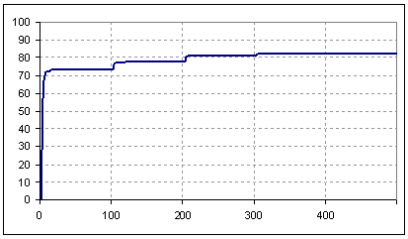
(g) 5-state 3-agent 2-action games (h) 5-state 3-agent 3-action games

(i) 5-state 5-agent 2-action games (j) 10-state 2-agent 2-action games

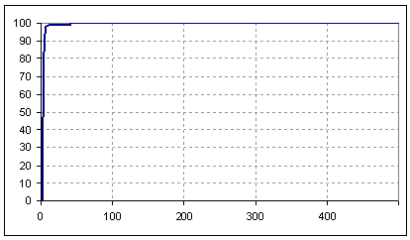
Fig. 4.4. Nash-RD: Experimental Estimations of Number of Iterations



(a) 10-state 2-agent 3-action games



(b) 10-state 3-agent 2-action games



(c) 15-state 2-agent 2-action games

Fig. 4.5. Nash-RD: Experimental Estimations of Number of Iterations

Part III

Bellman Optimality Equation Based Multi-Agent Reinforcement Learning Algorithms

5

Optimistic-Pessimistic Q-learning Algorithm with Variable Criterion

In this chapter we consider the case when the agents neither know the other players' payoff functions in advance nor get the information on their immediate rewards in the process of the play. From decision making perspective every state of stochastic game under such assumptions can be regarded as a game against nature. In such a game states of nature correspond to the other players' joint actions and the agent's decision problem is a typical problem of decision under uncertainty. A number of criteria for making decision under uncertainty have been developed, each reflecting some special aspect of rational behavior and attitude to risk.

The idea was to make use of the achievements of decision making theory in coming to decision under uncertainty and to develop an algorithm satisfying the criteria formulated in [28] (see section 2.3.3):

- *Rationality*: "If the other players' policies converge to stationary policies then the learning algorithm will converge to a policy that is a best-response to the other players' policies."
- *Convergence*: "The learner will necessarily converge to a stationary policy against agents using an algorithm from some class of learning algorithms."

The results of this chapter were partly published in [5] and [6].

In this chapter we introduce a reinforcement learning algorithm for multi-agent systems OPVar-Q based on variable Hurwicz's[1] optimistic-pessimistic criterion. Hurwicz's criterion allows to embed initial knowledge of how friendly the environment in which the agent is supposed to function will be. We formally prove that the developed algorithm always converges to stationary policies. In addition the variability of Hurwicz's criterion causes the rationality of the algorithm — the convergence to best-response against opponents with stationary policies.

[1] Leonid Hurwicz (1917-2008) was an American economist and mathematician of Russian origin. In 2007 he was awarded the Nobel Prize in Economic Sciences for mechanism design theory.

Thorough testing of the developed algorithm against well-known multi-agent reinforcement learning algorithms has shown that OPVar-Q can function on the level of its opponents. In self play for all types (according to Rapoport's classification [125]) of repeated 2×2 games the proposed algorithm has converged to a pure Nash equilibrium when the latter existed.

The chapter is organized as follows. Different criteria for making decision under uncertainty are examined in section 5.1. Sections 5.2, 5.3 and 5.4 present the theorems that we will use in the proof of the convergence of our method in section 5.5. Section 5.6 is devoted to the analysis of the results of thorough testing of our algorithm against other multi-agent reinforcement learning algorithms.

5.1 Criteria for Choosing Strategy in Games against Nature

In games against nature the agents do not know the payoff matrices of each other and treat all other agents present in the environment as a part of the environment (a nature). The agent's goal is to choose the strategy maximizing its gain in one-time game against the nature. The agent's reward in games against nature is usually presented as a matrix (see table 5.1), where $A_1, A_2, \ldots, A_n$ are actions of the agent and $N_1, N_2, \ldots, N_m$ are the states of the nature.

Table 5.1. Payoff Matrix

	N_1	N_2	...	N_m
A_1	a_{11}	a_{12}	...	a_{1m}
A_2	a_{21}	a_{22}	...	a_{2m}
...	...	...	...	...
A_n	a_{n1}	a_{n2}	...	a_{nm}

5.1.1 Laplace's Criterion

Laplace's criterion functions under the hypothesis: "Since we don't know anything about the states of the nature, it is reasonable to regard them as equiprobable" and chooses the strategy with the highest average value of the outcomes [21]:

$$\max_{1\leq i\leq n}\frac{1}{m}\sum_{j=1}^{m}a_{ij}$$

5.1.2 Wald's Criterion

Wald's criterion [162] is criterion of extreme pessimism. The agent using it believes that the circumstances will be against it and tries to make its best being prepared for the worst:

$$\max_{1\leq i\leq n}\min_{1\leq j\leq m}a_{ij}$$

Wald's criterion is also called max min-criterion.

5.1.3 Optimistic Criterion

Optimistic criterion sticks to quite the opposite hypothesis. The agent using it believes that the circumstances will always favor it and chooses the strategy with maximum gain:

$$\max_{1\leq i\leq n}\max_{1\leq j\leq m}a_{ij}$$

5.1.4 Hurwicz's Criterion

Hurwicz's criterion [14] is based on optimistic-pessimistic parameter λ. The agent believes that with λ probability the circumstances will be favorable and the agents will act so as to maximize its reward and in $1-\lambda$ cases will force it to achieve the minimum value and it chooses the strategy that will maximize its gain under the above described circumstances:

$$\max_{1\leq i\leq n}\left[(1-\lambda)\min_{1\leq j\leq m}a_{ij}+\lambda\max_{1\leq j\leq m}a_{ij}\right]$$

5.2 Convergence Theorem

Theorem 5.1. *[142] Let $\mathcal{X}$ be an arbitrary finite set and $\mathcal{B}$ be the space of bounded, real-valued functions over $\mathcal{X}$ and $T:\mathcal{B}(\mathcal{X})\to\mathcal{B}(\mathcal{X})$ is an arbitrary mapping with fixed point v^*. Let $U_0\in\mathcal{B}(\mathcal{X})$ be an arbitrary function and $\mathcal{T}=(T_0,T_1,\ldots)$ be a sequence of random operators $T_t:\mathcal{B}(\mathcal{X})\times\mathcal{B}(\mathcal{X})\to\mathcal{B}(\mathcal{X})$ such that $U_{t+1}=T_t(U_t,v^*)$ converges to Tv^* with probability 1. Let $V_0\in\mathcal{B}(\mathcal{X})$ be an arbitrary function, and define $V_{t+1}=T_t(V_t,V_t)$. If there exist random functions $0\leq F_t(x)\leq 1$ and $0\leq G_t(x)\leq 1$ satisfying the conditions below with probability 1, then V_t converges to v^* with probability 1:*

1. for all $U_1 \in \mathcal{B}(\mathcal{X})$ *and* $U_2 \in \mathcal{B}(\mathcal{X})$*, and all* $x \in \mathcal{X}$*,*

$$|T_t(U_1, v^*)(x) - T_t(U_2, v^*)(x)| \leq G_t(x)|U_1(x) - U_2(x)|$$

2. for all $U \in \mathcal{B}(\mathcal{X})$ *and* $V \in \mathcal{B}(\mathcal{X})$*, and all* $x \in \mathcal{X}$*,*

$$|T_t(U, v^*)(x) - T_t(U, V)(x)| \leq F_t(x)\|v^* - V\|$$

3. $\sum_{t=1}^{n}(1 - G_t(x))$ *converges to infinity as* $n \to \infty$
4. there exists $0 \leq \gamma < 1$ *such that for all* $x \in \mathcal{X}$ *and for all* t

$$F_t(x) \leq \gamma(1 - G_t(x))$$

5.3 Stochastic Approximation

Stochastic approximation procedure allows to estimate the solution $x = \theta$ of the equation $M(x) = \alpha$, where $M(x)$ is the expected value of random variable $Y = Y(x)$ for given x.

The distribution function of Y has the form:

$$Pr[Y(x) \leq y] = H(y|x)$$

$$M(x) = \int_{-\infty}^{\infty} y dH(y|x)$$

It is assumed that neither the exact nature of $H(y|x)$ nor that of $M(x)$ is known.

Let's define a (non-stationary) Markov chain $\{x_n\}$

$$x_{n+1} - x_n = \alpha_n(\alpha - y_n) \tag{5.1}$$

where x_1 is an arbitrary constant and y_n is an observation on the random variable $Y(x_n)$.

Theorem 5.2. *[126] If* $\{\alpha_n\}$ *is a fixed sequence of positive constants such that*
1. $0 < \sum_{n=1}^{\infty} \alpha_n^2 = A < \infty$
2. $\sum_{n=1}^{\infty} \alpha_n = \infty$
if $\exists C > 0 : \ Pr[|Y(x)| \leq C] = \int_{-C}^{C} dH(y|x) = 1$ *for all* x *and*
1. $M(x)$ *is nondecreasing*
2. $M(\theta) = \alpha$
3. $M'(\theta) > 0$
then $\lim_{n\to\infty} \mathrm{E}(x_n - \theta)^2 = 0$

5.4 Dvoretzky's Theorem

Theorem 5.3. *[51] Let α_n, β_n and γ_n, $n = 1, 2, \ldots$, be nonnegative real numbers satisfying*

$$\lim_{n\to\infty} \alpha_n = 0$$

$$\sum_{n=1}^{\infty} \beta_n < \infty$$

$$\sum_{n=1}^{\infty} \gamma_n = \infty$$

Let θ be a real number and T_n, $n = 1, 2, \ldots$, be measurable transformations satisfying

$$|T_n(r_1, \ldots, r_n) - \theta| \le \max[\alpha_n, (1 + \beta_n)|r_n - \theta| - \gamma_n]$$

for all real $r_1, \ldots, r_n$. Let X_1 and Z_n, $n = 1, 2, \ldots$, be random variables and define

$$X_{n+1}(\omega) = T_n[X_1(\omega), \ldots, X_n(\omega)] + Z_n(\omega)$$

for $n \ge 1$.

Then the conditions

$$\mathrm{E}\left\{X_1^2\right\} < \infty$$

$$\sum_{n=1}^{\infty} \mathrm{E}\left\{Z_n^2\right\} < \infty$$

and

$$\mathrm{E}\left\{Z_n | x_1, \ldots, x_n\right\} = 0$$

with probability 1 for all n, imply

$$\lim_{n\to\infty} \mathrm{E}\left\{(X_n - \theta)^2\right\} = 0$$

and

$$P\left\{\lim_{n\to\infty} X_n = \theta\right\} = 1$$

Convergence with probability 1 of sequence 5.1 was inferred from Dvoretzky's theorem under additional assumptions:

Theorem 5.4. *[51] If the following conditions are fulfilled:*

1. *Y has uniformly bounded variance*
2. *$M(x)$ is measurable*
3. *$|M(x)| < A|x| + B < \infty$ for all x and suitable A and B*
4. *$\inf_{1/k<x-\theta<k} M(x) > 0$ for all $k > 0$*
5. *$\sup_{1/k<\theta-x<k} M(x) < 0$ for all $k > 0$*

then sequence 5.1 converges to θ both in mean square and with probability 1.

5.5 OPVar-Q Algorithm

If we examine JAL algorithm (algorithm 3) from decision making perspective we will notice that the agent does nothing more than applies Laplace's criterion to its approximation of discounted cumulative payoff matrix. The only difference is that playing the game repeatedly the agent can use better estimations of probability distribution of the states of the nature. Minimax-Q (algorithm 2) turns out to use Wald's criterion for repeated games at every time step and Friend-Q (algorithm 5) — optimistic criterion. It is easily explainable: Minimax-Q was developed for competitive environments where the agents pursue the opposite goals, the win of the one agent being the loss of the other, the agents will try to minimize each other's reward. Friend-Q was proposed for cooperative environments where all the agents have the same payoff matrices and maximizing their own rewards they maximize the rewards of all the agents present in the environment. Competitive or cooperative environments are just extreme cases. In most cases the environment where our agent will function is competitive / cooperative to some degree. In this section we are proposing a multi-agent reinforcement learning algorithm (algorithm 12) based on Hurwicz's optimistic-pessimistic criterion that allows us to embed preliminary knowledge of how friendly the environment will be. For example, parameter $\lambda = 0.3$ means that we believe that with 30% probability the circumstances will be favourable and the agents will act so as to maximize OPVar-Q's reward and in 70% will force it to achieve the minimum value and we choose the strategy in each state that will maximize our gain under the above described circumstances (OPVar-Q with $\lambda = 0.3$ tries more often to avoid low rewards than to get high rewards in comparison with OPVar-$Q(0.5)$). When the other agent is recognized to play stationary policies, the algorithm switches to best-response strategy. The algorithm is presented for 2-player stochastic game but without difficulty can be extended for arbitrary number of players. The learning rate in the algorithm is decayed so as to satisfy the conditions of theorem 5.9.

Lemma 5.5. *[142] Let $\mathcal{Z}$ be a finite set, $f_1 : \mathcal{Z} \to \mathbb{R}$ and $f_2 : \mathcal{Z} \to \mathbb{R}$. Then*

$$\left|\min_{z\in\mathcal{Z}} f_1(z) - \min_{z\in\mathcal{Z}} f_2(z)\right| \leq \max_{z\in\mathcal{Z}} |f_1(z) - f_2(z)|$$

Lemma 5.6. *[142] Let $\mathcal{Z}$ be a finite set, $f_1 : \mathcal{Z} \to \mathbb{R}$ and $f_2 : \mathcal{Z} \to \mathbb{R}$. Then*

$$\left|\max_{z\in\mathcal{Z}} f_1(z) - \max_{z\in\mathcal{Z}} f_2(z)\right| \leq \max_{z\in\mathcal{Z}} |f_1(z) - f_2(z)|$$

Lemma 5.7. *Let $Q, Q_1, Q_2 : S \times A^1 \times A^2 \to \mathbb{R}$ then for Hurwicz's criterion:*

$$H(Q(s)) = \max_{a^1\in A^1} \left[(1-\lambda) \min_{a^2\in A^2} Q(s,a^1,a^2) + \lambda \max_{a^2\in A^2} Q(s,a^1,a^2)\right]$$

Algorithm 12 OPVar-Q for player 1

Input: learning rate α (see theorem 5.9), discount factor γ, Hurwicz's criterion parameter λ, exploration probability ϵ
for all $s \in S$, $a^1 \in A^1$, and $a^2 \in A^2$ **do**
  $Q(s, a^1, a^2) \leftarrow 0$
  $V(s) \leftarrow 0$
  $\pi^1(s, a^1) \leftarrow 1/|A^1|$
end for
Observe the current state s
loop
  Choose action a^1 for state s using policy π^1 with probability $1 - \epsilon$ and with probability ϵ select an action at random
  Take action a^1, observe the other agent's action a^2, reward r^1 and succeeding state s' provided by the environment
  $Q(s, a^1, a^2) \leftarrow Q(s, a^1, a^2) + \alpha \left[r^1 + \gamma V(s') - Q(s, a^1, a^2)\right]$
  if the other agent's policy π^2 has become stationary **then**
    $\pi^1(s, a^1) \leftarrow \begin{cases} 1 & a^1 = \arg\max_{\mathfrak{a}^1 \in A^1} \sum_{\mathfrak{a}^2 \in A^2} \pi^2(s, \mathfrak{a}^2) Q(s, \mathfrak{a}^1, \mathfrak{a}^2) \\ 0 & \text{otherwise} \end{cases}$
    $V(s) \leftarrow \max_{\mathfrak{a}^1 \in A^1} \sum_{\mathfrak{a}^2 \in A^2} \pi^2(s, \mathfrak{a}^2) Q(s, \mathfrak{a}^1, \mathfrak{a}^2)$
  else
    $\pi^1(s, a^1) \leftarrow \begin{cases} 1 & a^1 = \arg\max_{\mathfrak{a}^1 \in A^1} \left[(1-\lambda) \min_{\mathfrak{a}^2 \in A^2} Q(s, \mathfrak{a}^1, \mathfrak{a}^2) + \right. \\ & \left. + \lambda \max_{\mathfrak{a}^2 \in A^2} Q(s, \mathfrak{a}^1, \mathfrak{a}^2)\right] \\ 0 & \text{otherwise} \end{cases}$
    $V(s) \leftarrow \max_{\mathfrak{a}^1 \in A^1} \left[(1-\lambda) \min_{\mathfrak{a}^2 \in A^2} Q(s, \mathfrak{a}^1, \mathfrak{a}^2) + \lambda \max_{\mathfrak{a}^2 \in A^2} Q(s, \mathfrak{a}^1, \mathfrak{a}^2)\right]$
  end if
  decay α
  $s \leftarrow s'$
end loop

where $0 \leq \lambda \leq 1$ *the following inequality holds:*

$$|H(Q_1(s)) - H(Q_2(s))| \leq \max_{a^1 \in A^1} \max_{a^2 \in A^2} \left|Q_1(s, a^1, a^2) - Q_2(s, a^1, a^2)\right|$$

Proof.

$$\begin{aligned} & |H(Q_1(s)) - H(Q_2(s))| = \\ = & \left| \max_{a^1 \in A^1} \left[(1-\lambda) \min_{a^2 \in A^2} Q_1(s, a^1, a^2) + \lambda \max_{a^2 \in A^2} Q_1(s, a^1, a^2)\right] \right. \\ - & \left. \max_{a^1 \in A^1} \left[(1-\lambda) \min_{a^2 \in A^2} Q_2(s, a^1, a^2) + \lambda \max_{a^2 \in A^2} Q_2(s, a^1, a^2)\right] \right| \end{aligned}$$

$$
\begin{aligned}
&\leq \max_{a^1\in A^1}\left|(1-\lambda)\left(\min_{a^2\in A^2} Q_1(s,a^1,a^2)-\min_{a^2\in A^2} Q_2(s,a^1,a^2)\right)\right.\\
&\left.+\ \lambda\left(\max_{a^2\in A^2} Q_1(s,a^1,a^2)-\max_{a^2\in A^2} Q_2(s,a^1,a^2)\right)\right|\\
&\leq \max_{a^1\in A^1}\left[\left|(1-\lambda)\left(\min_{a^2\in A^2} Q_1(s,a^1,a^2)-\min_{a^2\in A^2} Q_2(s,a^1,a^2)\right)\right|\right.\\
&\left.+\ \left|\lambda\left(\max_{a^2\in A^2} Q_1(s,a^1,a^2)-\max_{a^2\in A^2} Q_2(s,a^1,a^2)\right)\right|\right]\\
&\leq \max_{a^1\in A^1}\left[(1-\lambda)\max_{a^2\in A^2}\left|Q_1(s,a^1,a^2)-Q_2(s,a^1,a^2)\right|\right.\\
&\left.+\ \lambda\max_{a^2\in A^2}\left|Q_1(s,a^1,a^2)-Q_2(s,a^1,a^2)\right|\right]\\
&= \max_{a^1\in A^1}\max_{a^2\in A^2}\left|Q_1(s,a^1,a^2)-Q_2(s,a^1,a^2)\right|
\end{aligned}
$$

We used triangle inequality, lemma 5.5 and lemma 5.6 during the inference. □

Lemma 5.8. *Let $Q, Q_1, Q_2 : S \times A^1 \times A^2 \to \mathbb{R}$ and $\pi^2 \in \Theta^2$ be the policy of player 2 then for*

$$BR(Q(s),\pi^2(s)) = \max_{a^1\in A^1}\sum_{a^2\in A^2}\pi^2(s,a^2)Q(s,a^1,a^2)$$

the following inequality holds:

$$\left|BR(Q_1(s),\pi^2(s)) - BR(Q_2(s),\pi^2(s))\right| \leq \max_{a^1\in A^1}\max_{a^2\in A^2}\left|Q_1(s,a^1,a^2)-Q_2(s,a^1,a^2)\right|$$

Proof.

$$
\begin{aligned}
&\left|BR(Q_1(s),\pi^2(s)) - BR(Q_2(s),\pi^2(s))\right| =\\
&= \left|\max_{a^1\in A^1}\sum_{a^2\in A^2}\pi^2(s,a^2)Q_1(s,a^1,a^2) - \max_{a^1\in A^1}\sum_{a^2\in A^2}\pi^2(s,a^2)Q_2(s,a^1,a^2)\right|\\
&\leq \max_{a^1\in A^1}\left|\sum_{a^2\in A^2}\pi^2(s,a^2)Q_1(s,a^1,a^2) - \sum_{a^2\in A^2}\pi^2(s,a^2)Q_2(s,a^1,a^2)\right|\\
&= \max_{a^1\in A^1}\left|\sum_{a^2\in A^2}\pi^2(s,a^2)\left[Q_1(s,a^1,a^2)-Q_2(s,a^1,a^2)\right]\right|\\
&\leq \max_{a^1\in A^1}\left|\sum_{a^2\in A^2}\pi^2(s,a^2)\max_{a^2\in A^2}\left|Q_1(s,a^1,a^2)-Q_2(s,a^1,a^2)\right|\right|\\
&= \max_{a^1\in A^1}\max_{a^2\in A^2}\left|Q_1(s,a^1,a^2)-Q_2(s,a^1,a^2)\right|
\end{aligned}
$$

The above holds due to lemma 5.5 and lemma 5.6.

□

Theorem 5.9. *If $\{\alpha_t\}$ is a sequence, such that:*
1. $0 \leq \alpha_t < 1$
2. $\sum_{t=1}^{\infty} \chi(s_t = s, a_t^1 = a^1, a_t^2 = a^2)\alpha_t = \infty$ [2]
3. $\sum_{t=1}^{\infty} \chi(s_t = s, a_t^1 = a^1, a_t^2 = a^2)\alpha_t^2 < \infty$
with probability 1 *over* $S \times A^1 \times A^2$ *then OPVar-Q algorithm converges to the stationary policy* π^1 *determined by fixed point of operator*[3]*:*

$$[TQ](s, a^1, a^2) = r^1(s, a^1, a^2) + \gamma \sum_{s' \in S} p(s'|s, a^1, a^2) BR(Q(s'), \pi^2(s'))$$

$$\pi^1(s, a^1) \leftarrow \begin{cases} 1 \; a^1 = \arg\max_{\mathfrak{a}^1 \in A^1} \sum_{\mathfrak{a}^2 \in A^2} \pi^2(s, \mathfrak{a}^2) Q(s, \mathfrak{a}^1, \mathfrak{a}^2) \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{otherwise} \end{cases}$$

against opponent with stationary policy π^2*, and to the stationary policy* π^1 *determined by fixed point of operator*

$$[TQ](s, a^1, a^2) = r^1(s, a^1, a^2) + \gamma \sum_{s' \in S} p(s'|s, a^1, a^2) H(Q(s'))$$

$$\pi^1(s, a^1) \leftarrow \begin{cases} 1 \; a^1 = \arg\max_{\mathfrak{a}^1 \in A^1} \big[(1-\lambda) \min_{\mathfrak{a}^2 \in A^2} Q(s, \mathfrak{a}^1, \mathfrak{a}^2) + \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad + \lambda \max_{\mathfrak{a}^2 \in A^2} Q(s, \mathfrak{a}^1, \mathfrak{a}^2)\big] \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{otherwise} \end{cases}$$

against other classes of opponents.

Proof. (partly by analogy with [107])

Let further on $V(Q(s)) = BR(Q(s), \pi^2(s))$ when the other agent plays stationary policy π^2 and $V(Q(s)) = H(Q(s))$ otherwise.

Let Q^* be fixed point of operator T[4] and

$$M(Q)(s, a^1, a^2) = Q(s, a^1, a^2) - r^1(s, a^1, a^2) - \gamma \sum_{s' \in S} p(s'|s, a^1, a^2) V(Q^*(s'))$$

It's evident that conditions of theorem 5.2 on M are fulfilled:
1. $M(Q)$ is nondecreasing
2. $M(Q^*) = \alpha = 0$
3. $M'(Q) = 1 \;\Rightarrow\; M'(Q^*) = 1$

---

[2] $\chi(s_t = s, a_t^1 = a^1, a_t^2 = a^2) = \begin{cases} 1 \text{ if } s_t = s \text{ and } a_t^1 = a^1 \text{ and } a_t^2 = a^2 \\ 0 \qquad\qquad\qquad\qquad \text{otherwise} \end{cases}$

[3] We assume here that OPVar-Q plays for the first agent.

[4] The existence and uniqueness of Q^* follow from Banach fixed point theorem.

The random approximating operator:

$$Q_{t+1}(s,a^1,a^2) = T_t(Q_t,Q^*)(s,a^1,a^2)$$

$$T_t(Q_t,Q^*)(s,a^1,a^2) = \begin{cases} Q_t(s_t,a_t^1,a_t^2) + \alpha_t[r^1(s_t,a_t^1,a_t^2) + \gamma V(Q^*(s_t')) \\ -Q_t(s_t,a_t^1,a_t^2)] \\ \text{if } s = s_t \text{ and } a^1 = a_t^1 \text{ and } a^2 = a_t^2 \\ \\ Q_t(s,a^1,a^2) \text{ otherwise} \end{cases}$$

where

$$y_t(s,a^1,a^2) = \begin{cases} Q_t(s_t,a_t^1,a_t^2) - r^1(s_t,a_t^1,a_t^2) - \gamma V(Q^*(s_t')) \\ \text{if } s = s_t \text{ and } a^1 = a_t^1 \text{ and } a^2 = a_t^2 \\ \\ 0 \text{ otherwise} \end{cases}$$

It is evident that the other conditions will be satisfied if s_t' is randomly selected according to the probability distribution defined by $p(\cdot|s_t,a_t^1,a_t^2)$ and the actions a_t^1, a_t^2 are chosen with appropriate exploration.

Then according to theorem 5.2 T_t approximates in mean square the solution of the equation $M(Q) = 0$. In other words, $Q_{t+1} = T_t(Q_t,Q^*)$ converges to TQ^* in mean square.

Convergence with probability 1 of sequence 5.1 follows from theorem 5.4.

Let's make sure that the conditions of theorem 5.4 are satisfied.

Y has uniformly bounded variance

$$\exists\sigma < \infty : \ \forall Q \ \mathrm{E}\{(Y(Q) - M(Q))^2\} \le \sigma^2 < \infty$$

and

$$\forall(s,a_1,a_2) \in \mathcal{X}$$

$$|M(Q)(s,a_1,a_2)| < A|Q(s,a_1,a_2)| + B < \infty$$

for $A = 1$ and $B < \infty$ since the reward functions are bounded in absolute value by $R_{\max}$ and discount factor $0 \le \gamma < 1$.

Obviously, other conditions will be fulfilled since $M(Q^*) = 0$ and $M(Q)$ is a strictly increasing function.

Thus T_t approximates the solution of the equation $M(Q) = 0$ with probability 1. In other words, $Q_{t+1} = T_t(Q_t,Q^*)$ converges to TQ^* with probability 1.

Let $G_t(s,a^1,a^2) = \begin{cases} 1-\alpha_t & \text{if } s = s_t \text{ and } a^1 = a_t^1 \text{ and } a^2 = a_t^2 \\ 1 & \text{otherwise} \end{cases}$

and $F_t(s,a^1,a^2) = \begin{cases} \gamma\alpha_t & \text{if } s = s_t \text{ and } a^1 = a_t^1 \text{ and } a^2 = a_t^2 \\ 0 & \text{otherwise} \end{cases}$

Let's check up the conditions of theorem 5.1:

1. when $s = s_t$ and $a^1 = a_t^1$ and $a^2 = a_t^2$:

$$\begin{aligned}
&|T_t(Q_1, Q^*)(s, a^1, a^2) - T_t(Q_2, Q^*)(s, a^1, a^2)| = \\
&= |(1 - \alpha_t)Q_1(s_t, a_t^1, a_t^2) + \\
&+ \alpha_t(r^1(s_t, a_t^1, a_t^2) + \gamma V(Q^*(s_t'))) - \\
&- (1 - \alpha_t)Q_2(s_t, a_t^1, a_t^2) - \\
&- \alpha_t(r^1(s_t, a_t^1, a_t^2) + \gamma V(Q^*(s_t')))| = \\
&= G_t(s, a^1, a^2)|Q_1(s, a^1, a^2) - Q_2(s, a^1, a^2)|
\end{aligned}$$

   when $s \neq s_t$ or $a^1 \neq a_t^1$ or $a^2 \neq a_t^2$ it is evident that the condition holds.
2. when $s = s_t$ and $a^1 = a_t^1$ and $a^2 = a_t^2$:

$$\begin{aligned}
&|T_t(Q_1, Q^*)(s, a^1, a^2) - T_t(Q_1, Q_2)(s, a^1, a^2)| = \\
&= |(1 - \alpha_t)Q_1(s_t, a_t^1, a_t^2) + \\
&+ \alpha_t(r^1(s_t, a_t^1, a_t^2) + \gamma V(Q^*(s_t'))) - \\
&- (1 - \alpha_t)Q_1(s_t, a_t^1, a_t^2) - \\
&- \alpha_t(r^1(s_t, a_t^1, a_t^2) + \gamma V(Q_2(s_t')))| = \\
&= F_t(s_t, a_t^1, a_t^2)|V(Q^*(s_t')) - V(Q_2(s_t'))| \leq \\
&\leq F_t(s, a^1, a^2) \max_{a^1 \in A^1} \max_{a^2 \in A^2} |Q^*(s', a^1, a^2) - Q_2(s', a^1, a^2)|
\end{aligned}$$

   The last inequality holds due to lemmas 5.7 and 5.8.
   when $s \neq s_t$ or $a^1 \neq a_t^1$ or $a^2 \neq a_t^2$ it is evident that the condition holds.
3. $\sum_{t=1}^{n}(1 - G_t(x))$ converges to infinity as $n \to \infty$ (see the assumption of the theorem)
4. the fourth condition evidently holds.

Consequently, OPVar-Q algorithm converges with probability 1 to the stationary policies determined by fixed points of corresponding operators.

□

5.6 Experimental Results

We tested OPVar-Q algorithm on 14 classes of 10-state 2×2 stochastic games (derived with the use of Gamut [118] with uniformly distributed transition probabilities) whose immediate payoff matrices at all the states belong to one of the following types of bimatrix games (see Gamut classification [118]):

- battle of the sexes game
- chicken game
- collaboration game
- coordination game
- covariant game
- dispersion game

- grab the dollar game
- hawk and dove game
- matching pennies game
- prisoners' dilemma game
- random compound game
- random game
- random zero-sum game
- two by two game

For the sake of reliability we derived 100 instances of each game class and made 20000 iterations. The agent plays as both the row agent and the column agent. We tested OPVar-Q algorithm against the following well-known algorithms for multi-agent reinforcement learning (see section 2.3.3)[5][6]:

- *Stationary* opponent plays the first action in 75% cases and the second action in 25% cases.
- *Q-learning* [163] (see algorithm 1) was initially developed for single-agent environments. The algorithm learns by immediate rewards a tabular function $Q(s, a)$ that returns the largest value for the action a that should be taken in each particular state s so as to maximize expected discounted cumulative reward. When applied to multi-agent systems Q-learning algorithm ignores totally the presence of other agents though the latter naturally influence its immediate rewards.
- *Minimax-Q* [103] (see algorithm 2) was developed for strictly competitive games and chooses the policy that maximizes its notion of the expected discounted cumulative reward convinced that the circumstances will be against it.
- *Friend-Q* [104] (see algorithm 5) was developed for strictly cooperative games and chooses the action that will bring the highest possible expected discounted cumulative reward when the circumstances will favor it.
- *JAL* [39] (see algorithm 3) believes that the average opponent's strategy very well approximates the opponent's policy in the future and takes it into account while choosing the action that is to maximize its expected discounted cumulative reward.
- *PHC* [28] (see algorithm 8) in contrast to Q-learning algorithm changes its policy gradually in the direction of the highest Q values.
- *WoLF* [28] (see algorithm 9) differs from PHC only in that it changes its policy faster when losing and more slowly when winning.

The results of the experiments (see tables 5.1 and 5.2) showed that the developed algorithm can function on the level (sometimes better) of its opponents which though don't possess both properties: rationality (convergence to best-response against opponents with stationary policies) and convergence to stationary policies against all types of opponents. OPVar-Q forces the opponent to play the strategy that is more profitable for it and at the same time tunes its policy facing stationary opponent.

[5] We have implemented Nash-Q algorithm (see algorithm 4) as well. But because of high complexity of this method we failed to make enough iterations to recognize any interesting trends.

[6] We didn't test Nash-RD approach since neither its assumptions on the available information nor on the opponent's behavior hold in this experiment.

We present the analysis only of a few game classes that though allows to gain the general notion of interaction between the developed OPVar-Q and the above presented multi-agent reinforcement learning algorithms. The test classes are presented in general form, where A, B, C, D are uniformly distributed in the interval $[-100, 100]$ payoffs and $A > B > C > D$. Discount factor $\gamma = 0.9$. We will analyze the results as though OPVar-Q played for the row agent. For all games we chose neutral parameter $\lambda = 0.5$ for OPVar-Q, except random zero-sum games and matching pennies. For these two classes we embedded our preliminary knowledge and set parameter λ to a more cautious value 0.3. To illustrate the gain of variable Hurwicz's optimistic-pessimistic criterion against stationary opponents we compare OPVar-Q with the version (OP-$Q(0.5)$) of our algorithm that doesn't make a difference between opponents with stationary and non-stationary policies.

Q-learning, PHC, WoLF, JAL turned out to have very similar final behavior. Small difference in the performance of these algorithms is due to a bit different manner of tuning the policy and underlying mechanism.

5.6.1 Battle of the Sexes

We analyze OPVar-Q's (OP-Q's) interaction with the above presented algorithms after a short exploration phase. Because of slow incremental update (see algorithm 12) the mutual influence of the states is small in comparison with immediate rewards and can be omitted. The states that are either of type 1 (see tables 5.2) or of type 2 (see table 5.3) can be examined separately.

Type 1

Table 5.2. Battle of the sexes: type 1

A,B	C,C
C,C	B,A

After a short exploration phase OP-Q (and OPVar-Q at first) chooses the first strategy in battle of the sexes type 1. Indeed Hurwicz's criterion for the first and the second actions are:

$$H_1 = 0.5 \cdot A + 0.5 \cdot C$$

$$H_2 = 0.5 \cdot C + 0.5 \cdot B$$

- *Stationary* opponent gets $0.75 \cdot B + 0.25 \cdot C$ as OP-Q (OPVar-Q) plays the first strategy. OP-Q gets in average $0.75 \cdot A + 0.25 \cdot C$ in this state. After noticing that its opponent is stationary OPVar-Q also plays the first strategy for $0.75 \cdot A + 0.25 \cdot C > 0.75 \cdot C + 0.25 \cdot B$ and gets in average $0.75 \cdot A + 0.25 \cdot C$.
- *Q-learning*, *PHC*, *WoLF* get the impression that in their environment (where OP-Q (OPVar-Q) agent is constantly playing the first strategy) the first strategy is much more profitable than the second one (B against C, where $B > C$) and play it. As a result OP-Q gets A as average reward in this state after exploration stage and Q-learning, PHC, WoLF only — B. On realizing that the opponent's strategy has become stationary $(1, 0)$, OPVar-Q also plays the first strategy ($A > C$) and gets A as average reward.
- *Minimax-Q* strives to maximize its expected discounted cumulative reward in the worst case. But battle of the sexes is not strictly competitive. Therefore OP-Q and OPVar-Q show better results.
- *Friend-Q* developed for cooperative environments believes that when it gets the best reward so do the other agents in the environment and therefore it is the most profitable for them to play the other part of the joint action that results in the largest reward to Friend-Q. In battle of the sexes it is constantly playing the second action. As a result OP-Q and Friend-Q both get very low C immediate reward. After realizing that its opponent plays the second strategy OPVar-Q also plays the second strategy for $B > C$ and this results in A to Friend-Q and B to OPVar-Q as average rewards in this state.
- *JAL* taking into account OP-Q's (OPVar-Q's) stationary $(1, 0)$ policy chooses also the first more profitable for it action ($B > C$). OP-Q and JAL respectively get A and B as average rewards. As JAL's policy becomes stationary OPVar-Q also plays the first strategy ($A > C$) and gets A as average reward.

Type 2

Table 5.3. Battle of the sexes: type 2

B,A	C,C
C,C	A,B

After a short exploration phase OP-Q (and OPVar-Q at first) chooses the second action in battle of the sexes type 2.

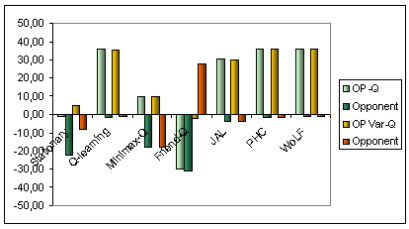
- *Stationary* opponent gets $0.75 \cdot C + 0.25 \cdot B$ as OP-Q (OPVar-Q) plays the second strategy. OP-Q gets in average $0.75 \cdot C + 0.25 \cdot A$. OPVar-$Q$ results are higher because it chooses the action that will maximize its cumulative reward against stationary opponent.
- *Q-learning*, *PHC*, *WoLF*, *JAL* and OP-Q (OPVar-Q) play the second strategies and get B and A as average rewards correspondingly.
- *Minimax-Q* the same as for type 1.
- *Friend-Q* plays the first strategy while OP-Q chooses the second action. They both get low C average reward. On getting to know that opponent permanently plays policy $(1, 0)$ OPVar-Q chooses the first action and gets B as average reward in this state while Friend-Q gets A.

5.6.2 Self Play

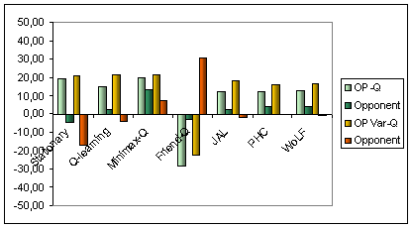
In self play OPVar-Q converged to one of pure Nash equilibria for every class of 2×2 repeated games (out of 78 according to Rapoport's classification [125]) where the latter exist.

5.7 Conclusion

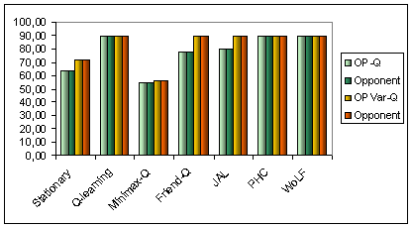
Multi-agent reinforcement learning has been for the first time considered from decision making perspective. It turned out that under the assumptions that the agents neither know the other players' payoff functions in advance nor get the information on their immediate rewards during the play each state of the stochastic game can be represented as a game against nature. A number of criteria for choosing the best strategy in games against nature have been analyzed. An algorithm based on variable Hurwicz's optimistic-pessimistic criterion was developed. Hurwicz's criterion allows us to embed initial knowledge of how friendly the environment in which the agent is supposed to function will be. A formal proof of the algorithm convergence to stationary policies is given. The variability of Hurwicz's criterion allowed it to converge to best-response strategies against opponents with stationary policies. Thorough testing of the developed algorithm against Q-learning, PHC, WoLF, Minimax-Q, Friend-Q and JAL showed that OPVar-Q functions on the level of its non-convergent (on irrational) opponents in the environments of different level of amicability.



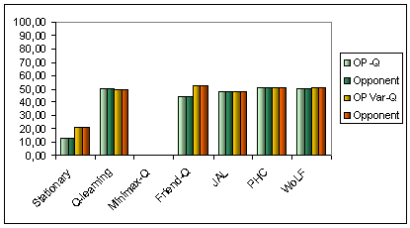
(a) Battle of the Sexes Game



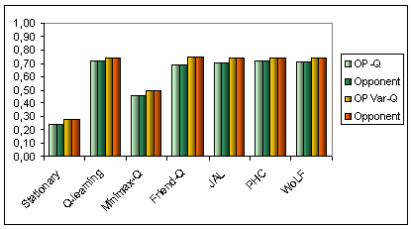
(b) Chicken Game



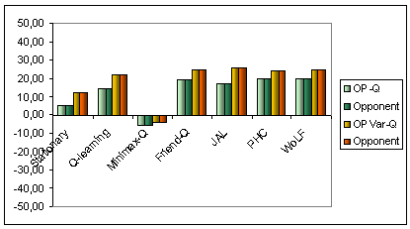
(c) Collaboration Game



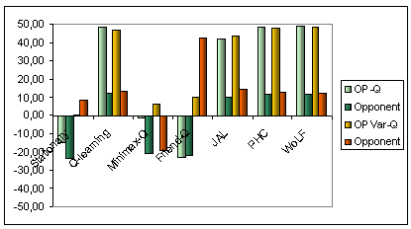
(d) Coordination Game



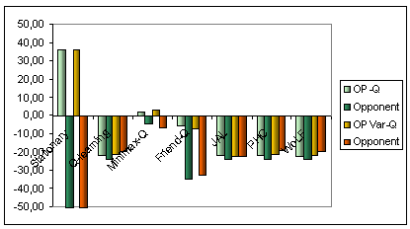
(e) Covariant Game



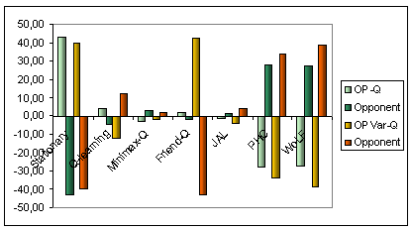
(f) Dispersion Game



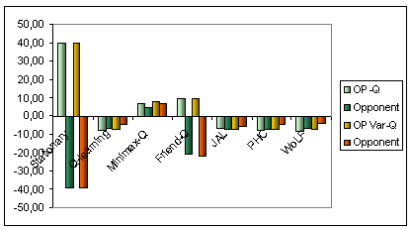
(g) Grab the Dollar Game



(h) Hawk and Dove Game

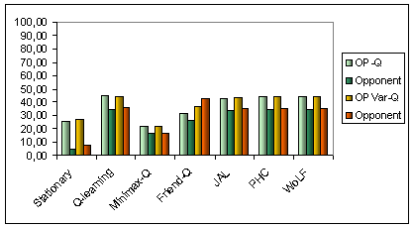


(i) Matching Pennies Game

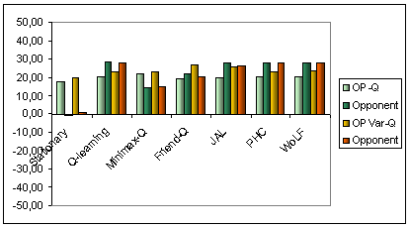


(j) Prisoners' Dilemma Game

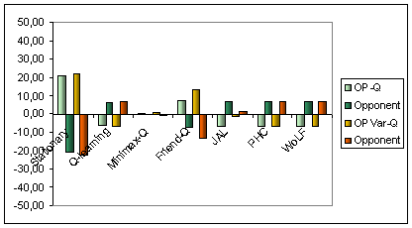
Fig. 5.1. Results of Experiments



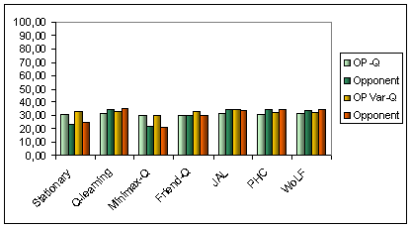
(a) Random Compound Game



(b) Random Game



(c) Random Zero-Sum Game



(d) Two by Two Game

Fig. 5.2. Results of Experiments

Part IV

Applications

6

Applications

The application area of the developed approaches is very broad. It includes traditional reinforcement learning tasks in multi-agent environments along with many economic problems in the field of capital accumulation, advertising, pricing, macroeconomics, warfare and resource economics that are traditionally represented as differential games.

The models examined in the context of differential games were rather dictated by mathematical tractability than by practical plausibility.

Stochastic game representation may shed light on the initial problems as it could allow to take into consideration the interdependences that were omitted in differential game representation in order to make them solvable.

The chapter is organized as follows. In section 6.1 we are examining a chocolate duopoly model [56] inspired by a differential game advertising model. We are applying the developed methods to table soccer problem in section 6.2. In section 6.3 we investigate the behavior of Nash-RD based agents trading at a double auction. In section 6.4 an introduction to differential game theory could be found. Section 6.5 is devoted to overview of differential game models of capital accumulation, advertising, pricing, marketing channels, macroeconomics, warfare and arms race, exploitation of renewable and nonrenewable resources and pollution. Applications of differential games being the topic of dozens books ([110], [92], [48], [96] and [158] to name only a few), our overview can't claim completeness in any sense. We tried though to choose the most (from our point of view) illustrative examples.

6.1 Chocolate Duopoly

6.1.1 Problem

In a middle-sized town the assortment of dark 70+% chocolate bars is represented by the production of two companies. The market for such chocolate is very stable since dark chocolate connoisseurs will never switch to other

ignoble chocolate sorts or plain sweets. Chocolate being considered a weak drug, they won't also stop consuming chocolate. The task the two companies $k = 1, 2$ are facing is to optimize the level of advertisement a^k and the chocolate price p^k and quality q^k so as to maximize their cumulative profits. We assume that advertisement level, price and quality can undergo changes only in the beginning of the month. The product being on the market for a certain time the firms are no more free to set the prices arbitrary[1]. But the regulation of the prices could be achieved to the customers' delight by discount policy. The firms are considering 10% and 25% reductions of prices (see table 6.2). In the war for customers the companies could also extend the production time to obtain chocololate of higher quality what though will lead to somewhat higher costs (see table 6.3). As to the advertisement levels here the firms have two options — whether not to advertise at all or to advertise at the price of 7500€ (see table 6.1).

Let the customers be in 67% price-sensitive who would prefer the chocolate at the lowerest price despite the difference in qualities and 33% be quality-sensitive who would rather buy chocolate of the highest quality regardless of the price. Of course price-sensitive buyers will choose the chocolate bar of a higer quality, prices being equal. Similar preference is true for quality-sensitive customers facing two chocolate bars of the same quality. Let 18% buyers correspond to those credulous customers who will buy whatever is being advertised.

There is also some natural inertness in customers' behavior — they will react to the companies' current offers only in the next month and merely in 30% of cases.

Market volume N of a middle-sized German town (140000 people) is approximately 134000 bars per month[2].

Table 6.1. Advertisement

	Company 1	*Company* 2
no	0€	0€
yes	7500€	7500€

The profits in the next month are worth 99% of what they are worth now.

This problem can be formally represented as the following general-sum discounted stochastic game $\Gamma = \langle S, A^1, A^2, \gamma, r^1, r^2, p \rangle$.

[1] Such price policies can only arouse indignation and distrust.

[2] According to [87] and [130], annual average consumption of chocolate in Germany is 8.21 kg, 14% of which fall to the share of dark chocolate [86].

Table 6.2. Prices

	Company 1	*Company* 2
discounted	1.79€ (10% off)	1.79€ (25% off)
normal	1.99€	2.39€

Table 6.3. Quality Costs $c^k(q^k)$

	Company 1	*Company* 2
normal	1.60€	1.75€
supreme	1.90€	2.15€

The states correspond to market shares of the companies (m^1, m^2):

$$\begin{aligned}
s_1: &\quad m^1 \in [0\%, 10\%) \quad m^2 \in [90\%, 100\%] \\
s_2: &\quad m^1 \in [10\%, 20\%) \quad m^2 \in [80\%, 90\%) \\
s_3: &\quad m^1 \in [20\%, 30\%) \quad m^2 \in [70\%, 80\%) \\
s_4: &\quad m^1 \in [30\%, 40\%) \quad m^2 \in [60\%, 70\%) \\
s_5: &\quad m^1 \in [40\%, 50\%) \quad m^2 \in [50\%, 60\%) \\
s_6: &\quad m^1 \in [50\%, 60\%) \quad m^2 \in [40\%, 50\%) \\
s_7: &\quad m^1 \in [60\%, 70\%) \quad m^2 \in [30\%, 40\%) \\
s_8: &\quad m^1 \in [70\%, 80\%) \quad m^2 \in [20\%, 30\%) \\
s_9: &\quad m^1 \in [80\%, 90\%) \quad m^2 \in [10\%, 20\%) \\
s_{10}: &\quad m^1 \in [90\%, 100\%] \quad m^2 \in [0\%, 10\%)
\end{aligned}$$

Each month the companies can choose their actions (a^k, p^k, q^k), $k = 1, 2$. As a result they will get[3]

$$r^k(s, (a^1, p^1, q^1), (a^2, p^2, q^2)) = (p^k - c^k(q^k)) \cdot m^k \cdot N - a^k$$

[3] The state s reveals only the interval the actual market share m^k belongs to. Further on we will use the average $\overline{m^k}$ over the corresponding interval in our calculations.

The discount factor will be $\gamma = 0.99$.

Let us calculate transition probabilities:

Let m_*^1 and m_*^2 denote new market shares (the result of the chosen actions) if there were no inertness.

Since customers will react to the current offers with $p = 30\%$ probability, the firms with market shares m^1 and m^2 will reach market shares $m^{1\prime}$ and $m^{2\prime}$ in the next month:

$$m^{1\prime} = (1-p) \cdot m^1 + p \cdot m_*^1$$
$$m^{2\prime} = (1-p) \cdot m^2 + p \cdot m_*^2$$

The corresponding probability will be equal to 1.0.

Example 6.1. Let us calculate the transition probability:

$$p(s_5|s_3, (a_2^1, p_2^1, q_1^1), (a_2^2, p_2^2, q_1^2))$$

Since both companies chose to advertise, it won't have any influence on the customers.

As the price $p_2^1 < p_2^2$, all price-sensitive buyers will prefer the production of the first firm.

The qualities being equal, the quality-sensitive buyers will also prefer the cheaper chocolate bars of the first firm:

$$m_*^1 = 67\% + 33\% = 100\%$$
$$m^1 \in [20\%, 30\%)$$
$$\overline{m^1} = 25\%$$
$$m^{1\prime} = (1-p) \cdot \overline{m^1} + p \cdot m_*^1 = (1 - 0.3) \cdot 25\% + 0.3 \cdot 100\% = 47.5\%$$

Thus,

$$p(s_5|s_3, (a_2^1, p_2^1, q_1^1), (a_2^2, p_2^2, q_1^2)) = 1.0$$

6.1.2 Solution

With the use of Nash-RD approach we found a Nash equilibrium with accuracy 1€ presented in tables 6.5, 6.7, 6.8, 6.9, 6.10, 6.11, 6.12, 6.14, 6.15, 6.16[4]. The values of the Nash equilibrium to each company for every market distribution can be found in table 6.4.

Let's interpret the policies that constitute the Nash equilibrium for every market distribution.

Table 6.4. Values of Market Shares to Companies

	Company 1	*Company* 2
$s_1:\quad m^1 \in [0\%, 10\%) \quad m^2 \in [90\%, 100\%]$	2 320 696€	2 973 460€
$s_2:\quad m^1 \in [10\%, 20\%) \quad m^2 \in [80\%, 90\%)$	2 333 422€	2 957 384€
$s_3:\quad m^1 \in [20\%, 30\%) \quad m^2 \in [70\%, 80\%)$	2 333 053€	2 930 307€
$s_4:\quad m^1 \in [30\%, 40\%) \quad m^2 \in [60\%, 70\%)$	2 349 074€	2 921 200€
$s_5:\quad m^1 \in [40\%, 50\%) \quad m^2 \in [50\%, 60\%)$	2 351 040€	2 894 936€
$s_6:\quad m^1 \in [50\%, 60\%) \quad m^2 \in [40\%, 50\%)$	2 354 326€	2 894 400€
$s_7:\quad m^1 \in [60\%, 70\%) \quad m^2 \in [30\%, 40\%)$	2 374 629€	2 867 863€
$s_8:\quad m^1 \in [70\%, 80\%) \quad m^2 \in [20\%, 30\%)$	2 369 978€	2 866 512€
$s_9:\quad m^1 \in [80\%, 90\%) \quad m^2 \in [10\%, 20\%)$	2 375 204€	2 858 760€
$s_{10}:\quad m^1 \in [90\%, 100\%] \quad m^2 \in [0\%, 10\%)$	2 400 530€	2 839 452€

Market shares $m^1 \in [0\%, 10\%)$ and $m^2 \in [90\%, 100\%]$

When almost all the customers buy chocolate bars of the second company, the best policy for the first company will be to set prices to the maximum level 1.99€ and sell the product of the normal quality (the profit per bar will be maximum in this case — 0.39€) and in 16.8% of cases to try to attract credulous customers by advertisement (see table 6.5). The corresponding policy for the second firm will be to set maximal prices and normal quality and advertize the bars in 47.3% of cases.

Since our customers are inert to a high degree, they will start reacting at the offers only in the next month and the immediate rewards will be:

$$r^1(s_1, (a_1^1, p_2^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = (p_2^1 - c^1(q_1^1)) \cdot \overline{m^1} \cdot N - a_1^1 = \\ = (1.99 - 1.60) \cdot 0.05 \cdot 134000 - 0 = 2613$$

$$r^1(s_1, (a_1^1, p_2^1, q_1^1), (a_2^2, p_2^2, q_1^2)) = r^1(s_1, (a_1^1, p_2^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 2613$$

[4] The values are rounded to three fractional digits.

Table 6.5. Policies for Market Shares $m^1 \in [0\%, 10\%)$, $m^2 \in [90\%, 100\%]$

			Company 1	*Company* 2
a_1	p_1	q_1	0	0
a_1	p_1	q_2	0	0
a_1	p_2	q_1	0.832	0.527
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0.168	0.473
a_2	p_2	q_2	0	0

$$r^1(s_1,(a_2^1,p_2^1,q_1^1),(a_1^2,p_2^2,q_1^2)) = (p_2^1 - c^1(q_1^1)) \cdot \overline{m^1} \cdot N - a_2^1 = \\ = (1.99 - 1.60) \cdot 0.05 \cdot 134000 - 7500 = -4887$$

$$r^1(s_1,(a_2^1,p_2^1,q_1^1),(a_2^2,p_2^2,q_1^2)) = r^1(s_1,(a_2^1,p_2^1,q_1^1),(a_1^2,p_2^2,q_1^2)) = -4887$$

$$r^2(s_1,(a_1^1,p_2^1,q_1^1),(a_1^2,p_2^2,q_1^2)) = (p_2^2 - c^2(q_1^2)) \cdot \overline{m^2} \cdot N - a_1^2 = \\ = (2.39 - 1.75) \cdot 0.95 \cdot 134000 - 0 = 81472$$

$$r^2(s_1,(a_2^1,p_2^1,q_1^1),(a_1^2,p_2^2,q_1^2)) = r^2(s_1,(a_1^1,p_2^1,q_1^1),(a_1^2,p_2^2,q_1^2)) = 81472$$

$$r^2(s_1,(a_1^1,p_2^1,q_1^1),(a_2^2,p_2^2,q_1^2)) = (p_2^2 - c^2(q_1^2)) \cdot \overline{m^2} \cdot N - a_2^2 = \\ = (2.39 - 1.75) \cdot 0.95 \cdot 134000 - 7500 = 73972$$

$$r^2(s_1,(a_2^1,p_2^1,q_1^1),(a_2^2,p_2^2,q_1^2)) = r^2(s_1,(a_1^1,p_2^1,q_1^1),(a_2^2,p_2^2,q_1^2)) = 73972$$

The expected rewards will be:

$$r^1(s_1,x) = \sum_{\mathfrak{a}^1 \in A^1} \sum_{\mathfrak{a}^2 \in A^2} r^1(s_1,\mathfrak{a}^1,\mathfrak{a}^2) \cdot x^1_{s_1\mathfrak{a}^1} \cdot x^2_{s_1\mathfrak{a}^2} = 2613 \cdot 0.832 \cdot 0.527 + \\ + 2613 \cdot 0.832 \cdot 0.473 + -4887 \cdot 0.168 \cdot 0.527 + \\ - 4887 \cdot 0.168 \cdot 0.473 = 1353$$

$$r^2(s_1, x) = \sum_{\mathfrak{a}^1 \in A^1} \sum_{\mathfrak{a}^2 \in A^2} r^2(s_1, \mathfrak{a}^1, \mathfrak{a}^2) \cdot x^1_{s_1 \mathfrak{a}^1} \cdot x^2_{s_1 \mathfrak{a}^2} = 81472 \cdot 0.832 \cdot 0.527 +$$
$$+ 73972 \cdot 0.832 \cdot 0.473 + 81472 \cdot 0.168 \cdot 0.527 +$$
$$+ 73972 \cdot 0.168 \cdot 0.473 = 77924.5$$

Though advertising reduces the immediate rewards and therefore seems quite unreasonable, it is a good investment in the future as we will see.

In $83.2\% \cdot 52.7\%$ cases the first and the second firm set the maximum price, the normal quality and won't resort to advertising. But the maximum price of the first firm is lower than the one of the second and thus all price-sensitive customers will come to the first firm. The quality-sensitive buyers prefer quality to price. Since the offered qualities are equal, they will also buy production of the first firm.

Consequently,

$$m^1_* = 100\% \quad m^2_* = 0\%$$

Reiterating the calculations of example 6.1, we get:

$$p(s_4 | s_1, (a^1_1, p^1_2, q^1_1), (a^2_1, p^2_2, q^2_1)) = 1.0$$

In $83.2\% \cdot 47.3\%$ cases the firms will follow the same price-quality policies but the second firm will use the services of the advertising agency that will cost it 7500€. In such a way it will manage to attract 18% of customers and

$$m^1_* = 82\% \quad m^2_* = 18\%$$

$$p(s_3 | s_1, (a^1_1, p^1_2, q^1_1), (a^2_2, p^2_2, q^2_1)) = 1.0$$

In $16.8\% \cdot 52.7\%$ the first company will use the services of the advertising agency and would attract all the customers but for inertness.

$$m^1_* = 100\% \quad m^2_* = 0\%$$

$$p(s_4 | s_1, (a^1_2, p^1_2, q^1_1), (a^2_1, p^2_2, q^2_1)) = 1.0$$

And in $16.8\% \cdot 47.3\%$ cases both firms will resort to the services of the advertising agency that will nullify the effect of such services and

$$m^1_* = 100\% \quad m^2_* = 0\%$$

$$p(s_4 | s_1, (a^1_2, p^1_2, q^1_1), (a^2_2, p^2_2, q^2_1)) = 1.0$$

The decision not to advertise will bring the following profit in the long run:

$$\begin{aligned}\vartheta^1_{s_1(a_1^1,p_2^1,q_1^1)}(x) = &\sum_{\mathfrak{a}^2\in A^2}[r^1(s_1,(a_1^1,p_2^1,q_1^1),\mathfrak{a}^2)+\\ &+\gamma\sum_{s'\in S}p(s'|s_1,(a_1^1,p_2^1,q_1^1),\mathfrak{a}^2)v^1_{s'}]\cdot x^2_{s_1\mathfrak{a}^2}=\\ &=[2613+0.99\cdot 2349074]\cdot 0.527+\\ &+[2613+0.99\cdot 2333053]\cdot 0.473\approx 2320694\end{aligned}$$

The decision to advertise will yield the following profit in the long run:

$$\begin{aligned}\vartheta^1_{s_1(a_2^1,p_2^1,q_1^1)}(x) = &\sum_{\mathfrak{a}^2\in A^2}[r^1(s_1,(a_2^1,p_2^1,q_1^1),\mathfrak{a}^2)+\\ &+\gamma\sum_{s'\in S}p(s'|s_1,(a_2^1,p_2^1,q_1^1),\mathfrak{a}^2)v^1_{s'}]\cdot x^2_{s_1\mathfrak{a}^2}=\\ &=[-4887+0.99\cdot 2349074]\cdot 0.527+\\ &+[-4887+0.99\cdot 2349074]\cdot 0.473\approx 2320696\end{aligned}$$

So, though the decision to use expensive advertisement seems an irrational one, it is generously rewarded in the long run.

All other price-quality-advertisement combinations will bring smaller profits (see table 6.6[5]). The prerequisites of theorem 3.14 are satisfied with accuracy $\epsilon = 0.01$€ that will allow to get the final values of Nash equilibrium approximation with accuracy $\varepsilon = 1$€.

Market shares $m^1 \in [10\%, 20\%)$ and $m^2 \in [80\%, 90\%)$

As could be seen from table 6.7, the companies will set the maximum price and normal quality. The second firm will advertise, the first firm — not. Hence, the first firm will be content with future market share $m^1_* = 82\%$ ideal and $m^1 \in [30, 40)$ owing to customers' inertness and the second company will be satisfied with $m^2_* = 18\%$ and $m^2 \in [60, 70)$. It will charge the maximum price for normal quality, thus getting high immediate profits.

Market shares $m^1 \in [20\%, 30\%)$ and $m^2 \in [70\%, 80\%)$

The companies won't order any advertisement but will sell the chocolate bars of normal quality at the highest and lowest prices (see table 6.8).

[5] For computation of entries of table 6.6 we used precise policies (not their rounded values presented in tables 6.5, 6.7, 6.8, 6.9, 6.10, 6.11, 6.12, 6.14, 6.15, 6.16) in contrast to other calculations.

Table 6.6. Cumulative Profits of Company 1

			Cumulative Profits
a_1	p_1	q_1	2 319 356.141
a_1	p_1	q_2	2 317 346.141
a_1	p_2	q_1	2 320 696.141
a_1	p_2	q_2	2 318 686.141
a_2	p_1	q_1	2 319 356.131
a_2	p_1	q_2	2 317 346.131
a_2	p_2	q_1	2 320 696.131
a_2	p_2	q_2	2 318 686.131

In $4.3\% \cdot 43.2\%$ of cases the first and the second firms set the minimum price — 1.79€ and the customers would divide equally — 50% would buy at the first company, 50% — at the second one.

$$p(s_4|s_3, (a_1^1, p_1^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1.0$$

The firms will get immediate rewards:

$$r^1(s_3, (a_1^1, p_1^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 6365$$
$$r^2(s_3, (a_1^1, p_1^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 4020$$

When the first firm sets the minimum price and the second — the maximum ($4.3\% \cdot 56.8\%$ cases), all the buyers would move to the first firm but for inertness, so

$$p(s_5|s_3, (a_1^1, p_1^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 1.0$$
$$r^1(s_3, (a_1^1, p_1^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 6365$$
$$r^2(s_3, (a_1^1, p_1^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 64320$$

When the first firm raises the price and the second sticks to minimum one ($95.7\% \cdot 43.2\%$ cases), all the buyers would buy at the second firm in the next month but for inertness

$$p(s_2|s_3, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1.0$$

Table 6.7. Policies for Market Shares $m^1 \in [10\%, 20\%)$, $m^2 \in [80\%, 90\%)$

			Company 1	*Company* 2
a_1	p_1	q_1	0	0
a_1	p_1	q_2	0	0
a_1	p_2	q_1	1	0
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0	1
a_2	p_2	q_2	0	0

$$r^1(s_3, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 13065$$
$$r^2(s_3, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 4020$$

In $95.7\% \cdot 56.8\%$ of cases they both raise the prices, the buyers will go to the first firm, because its maximum price is lower.

$$p(s_5|s_3, (a_1^1, p_2^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 1.0$$
$$r^1(s_3, (a_1^1, p_2^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 13065$$
$$r^2(s_3, (a_1^1, p_2^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 64320$$

Market shares $m^1 \in [30\%, 40\%)$ and $m^2 \in [60\%, 70\%)$

The companies in this case prefer not to spend money on advertisement and charge maximal price for minimal quality (see table 6.9).

If not inertness, all the buyers would transfer to the first company.

$$p(s_6|s_4, (a_1^1, p_2^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 1.0$$

The second company will have large immediate profits in this month before our inert customers realize the offers:

$$r^1(s_4, (a_1^1, p_2^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 18291$$
$$r^2(s_4, (a_1^1, p_2^1, q_1^1), (a_1^2, p_2^2, q_1^2)) = 55744$$

Table 6.8. Policies for Market Shares $m^1 \in [20\%, 30\%)$, $m^2 \in [70\%, 80\%)$

			Company 1	*Company* 2
a_1	p_1	q_1	0.043	0.432
a_1	p_1	q_2	0	0
a_1	p_2	q_1	0.957	0.568
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0	0
a_2	p_2	q_2	0	0

Market shares $m^1 \in [40\%, 50\%)$ and $m^2 \in [50\%, 60\%)$

The first firm will use no advertisement, charge maximum price for minimum quality in 61.2% of cases and spend part of the profits on advertisement to attract customers in 38.8% cases (see table 6.10).

The second company won't spend money on advertisement, offer normal chocolate at special price in 62.7% of cases and resort to the help of advertising agencies and charge high price for the normal quality chocolate in 37.3% of cases.

In cases when the first firm chooses action (a_1^1, p_2^1, q_1^1) and the second firm — (a_1^2, p_1^2, q_1^2) the price-sensitive customers will choose the second firm, the quality-sensitive buyers will also choose the second firm since the quality is the same but the price is lower, thus $m_*^1 = 0\%$ and $m_*^2 = 100\%$.

But since the customers are inert

$$p(s_4|s_5, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1.0$$

The immediate rewards will be

$$r^1(s_5, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 23517$$
$$r^2(s_5, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 2948$$

The second firm acts in this case strategically — the larger market share it will have in the next month the more profit it will get (see table 6.4).

Table 6.9. Policies for Market Shares $m^1 \in [30\%, 40\%)$, $m^2 \in [60\%, 70\%)$

			Company 1	*Company* 2
a_1	p_1	q_1	0	0
a_1	p_1	q_2	0	0
a_1	p_2	q_1	1	1
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0	0
a_2	p_2	q_2	0	0

When the first company chooses (a_1^1, p_2^1, q_1^1) and the second — (a_2^2, p_2^2, q_1^2), the second firm could count only on 18% of customers but for inertness.

$$p(s_6|s_5, (a_1^1, p_2^1, q_1^1), (a_2^2, p_2^2, q_1^2)) = 1.0$$

Nevertheless immediate rewards of both companies are large.

$$r^1(s_5, (a_1^1, p_2^1, q_1^1), (a_2^2, p_2^2, q_1^2)) = 23517$$
$$r^2(s_5, (a_1^1, p_2^1, q_1^1), (a_2^2, p_2^2, q_1^2)) = 39668$$

When the first company chooses to charge maximum price for minimum quality and spend part of the profits on advertisement (a_2^1, p_2^1, q_1^1) and the second firm — to save money on advertisement but make discounts for normal quality chocolate (a_1^2, p_1^2, q_1^2), the first company should be content with only 18% percent of buyers in the next month.

But since the buyers are inert:

$$p(s_4|s_5, (a_2^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1.0$$

Immediate rewards are 16017 and 2948 correspondingly.

When the chosen actions are (a_2^1, p_2^1, q_1^1) and (a_2^2, p_2^2, q_1^2), all the customers would move to the first firm but for inertness.

So the market shares when inertness is not taken into account will be:

$$m_*^1 = 100\% \quad m_*^2 = 0\%$$

Table 6.10. Policies for Market Shares $m^1 \in [40\%, 50\%)$, $m^2 \in [50\%, 60\%)$

			Company 1	*Company* 2
a_1	p_1	q_1	0	0.627
a_1	p_1	q_2	0	0
a_1	p_2	q_1	0.612	0
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0.388	0.373
a_2	p_2	q_2	0	0

Inertness calculated:

$$p(s_7|s_5, (a_2^1, p_2^1, q_1^1), (a_2^2, p_2^2, q_1^2)) = 1.0$$

Immediate rewards are:

$$r^1(s_5, (a_2^1, p_2^1, q_1^1), (a_2^2, p_2^2, q_1^2)) = 16017$$
$$r^2(s_5, (a_2^1, p_2^1, q_1^1), (a_2^2, p_2^2, q_1^2)) = 39668$$

As before the policy of the first firm is the best reply (up to accuracy $\epsilon = 0.01$) in the long run to the policy of the second company and vice versa.

Market shares $m^1 \in [50\%, 60\%)$ and $m^2 \in [40\%, 50\%)$

Now the first firm possesses large enough market share and can count on immediate rewards and exploit the loyalty of the customers — it will charge the maximum price for normal quality chocolate bars (see table 6.11). Advertisement is of no advantage for it now.

The second company lost its position in the market and must win the customers — it tries to achieve it by reducing the price.

Since the price will be the lowest and the quality will be the same — the offer of the second company would attract all the customers but for inertness.

$$p(s_4|s_6, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1.0$$

The profits in this month will be 28743€ and 2412€ correspondingly.

Table 6.11. Policies for Market Shares $m^1 \in [50\%, 60\%)$, $m^2 \in [40\%, 50\%)$

			Company 1	*Company* 2
a_1	p_1	q_1	0	1
a_1	p_1	q_2	0	0
a_1	p_2	q_1	1	0
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0	0
a_2	p_2	q_2	0	0

Market shares $m^1 \in [60\%, 70\%)$ and $m^2 \in [30\%, 40\%)$

The first firm will charge maximum price for normal quality and use no advertisement, the second firm will try to attract customers by charging minimum price for normal quality chocolate bars in 30% of cases and will charge maximum price in 70% of cases (see table 6.12).

In the first case (30% of cases)

$$m^1_* = 0\% \quad m^2_* = 100\%$$
$$p(s_5|s_7,(a^1_1,p^1_2,q^1_1),(a^2_1,p^2_1,q^2_1)) = 1.0$$
$$r^1(s_7,(a^1_1,p^1_2,q^1_1),(a^2_1,p^2_1,q^2_1)) = 33969$$
$$r^2(s_7,(a^1_1,p^1_2,q^1_1),(a^2_1,p^2_1,q^2_1)) = 1876$$

In the second case (70% of cases)

$$m^1_* = 100\% \quad m^2_* = 0\%$$
$$p(s_8|s_7,(a^1_1,p^1_2,q^1_1),(a^2_1,p^2_2,q^2_1)) = 1.0$$

and immediate rewards

$$r^1(s_7,(a^1_1,p^1_2,q^1_1),(a^2_1,p^2_2,q^2_1)) = 33969$$
$$r^2(s_7,(a^1_1,p^1_2,q^1_1),(a^2_1,p^2_2,q^2_1)) = 30016$$

Table 6.12. Policies for Market Shares $m^1 \in [60\%, 70\%)$, $m^2 \in [30\%, 40\%)$

			Company 1	*Company* 2
a_1	p_1	q_1	0	0.3
a_1	p_1	q_2	0	0
a_1	p_2	q_1	1	0.7
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0	0
a_2	p_2	q_2	0	0

Since the above policies constitute an approximation of a Nash equilibrium of the corresponding bimatrix game, actions (a_1^2, p_1^2, q_1^2) and (a_1^2, p_2^2, q_1^2) must bring almost equal profits in the long run. Moreover these profits must be the highest over all the actions. Otherwise there would exist a better strategy (the one with maximum rewards) — a contradiction to the fact that the above policies constitute an approximation of a Nash equilibrium.

Let's check it up:

$$\begin{aligned}\vartheta^2_{s_7(a_1^2,p_1^2,q_1^2)}(x) = &\sum_{\mathfrak{a}^1 \in A^1} [r^2(s_7, \mathfrak{a}^1, (a_1^2, p_1^2, q_1^2)) + \\ &+ \gamma \sum_{s' \in S} p(s'|s_7, \mathfrak{a}^1, (a_1^2, p_1^2, q_1^2)) v^2_{s'}] \cdot x^1_{s_7 \mathfrak{a}^1} = \\ &= 1876 + 0.99 \cdot 2894936 \approx 2867862.64\end{aligned}$$

$$\begin{aligned}\vartheta^2_{s_7(a_1^2,p_2^2,q_1^2)}(x) = &\sum_{\mathfrak{a}^1 \in A^1} [r^2(s_7, \mathfrak{a}^1, (a_1^2, p_2^2, q_1^2)) + \\ &+ \gamma \sum_{s' \in S} p(s'|s_7, \mathfrak{a}^1, (a_1^2, p_2^2, q_1^2)) v^2_{s'}] \cdot x^1_{s_7 \mathfrak{a}^1} = \\ &= 30016 + 0.99 \cdot 2866512 \approx 2867862.88\end{aligned}$$

As we can see they are equal with an acceptable error.

Table 6.13. Cumulative Profits of Company 2

			Cumulative Profits
a_1	p_1	q_1	2 867 862.641
a_1	p_1	q_2	2 849 102.641
a_1	p_2	q_1	2 867 862.640
a_1	p_2	q_2	2 850 440.014
a_2	p_1	q_1	2 860 362.641
a_2	p_1	q_2	2 841 602.641
a_2	p_2	q_1	2 860 362.642
a_2	p_2	q_2	2 842 940.014

All other actions will result in lower cumulative rewards (see table 6.13[6]). The prerequisites of theorem 3.14 are satisfied with accuracy $\epsilon = 0.01$€ that will allow to get the final values of Nash equilibrium approximation with accuracy $\varepsilon = 1$€.

Market shares $m^1 \in [70\%, 80\%)$ and $m^2 \in [20\%, 30\%)$

The first company will spend money neither on advertisement nor on improvement of chocolate quality (see table 6.14). In 1.1% cases it will make discounts and in 98.9% cases offer no special prices.

The second company will try to attract customers by low prices.

In the first case (a_1^1, p_1^1, q_1^1) and (a_1^2, p_1^2, q_1^2).

The buyers would divide equally between two companies but because of inertness:

$$p(s_7|s_8, (a_1^1, p_1^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1.0$$

The immediate profits will be correspondingly 19095€ and 1340€.

In the second case (a_1^1, p_2^1, q_1^1) and (a_1^2, p_1^2, q_1^2).

The buyers would all choose the product of the second firm but for inertness.

[6] For computation of entries of table 6.13 we used precise policies (not their rounded values presented in tables 6.5, 6.7, 6.8, 6.9, 6.10, 6.11, 6.12, 6.14, 6.15, 6.16) in contrast to other calculations.

Table 6.14. Policies for Market Shares $m^1 \in [70\%, 80\%)$, $m^2 \in [20\%, 30\%)$

			Company 1	*Company* 2
a_1	p_1	q_1	0.011	1
a_1	p_1	q_2	0	0
a_1	p_2	q_1	0.989	0
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0	0
a_2	p_2	q_2	0	0

$$p(s_6|s_8, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1.0$$

In this case the companies will get the corresponding profits:

$$r^1(s_8, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 39195$$
$$r^2(s_8, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1340$$

Market shares $m^1 \in [80\%, 90\%)$ and $m^2 \in [10\%, 20\%)$

The first company will charge the highest price possible for average quality chocolate bars and spend no money on advertisement in 71.5% of cases (see table 6.15). In 28.5% of cases it will advertise its products.

The second company will offer chocolate bars of normal quality at the lowest price in 37.3% of cases and use services of no advertising agency. In 62.7% of cases the company will resort to advertisement.

In $71.5\% \cdot 37.3\%$ cases

$$m_*^1 = 0\% \quad m_*^2 = 100\%$$

But because of inertness

$$p(s_6|s_9, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1.0$$

Table 6.15. Policies for Market Shares $m^1 \in [80\%, 90\%)$, $m^2 \in [10\%, 20\%)$

			Company 1	*Company* 2
a_1	p_1	q_1	0	0.373
a_1	p_1	q_2	0	0
a_1	p_2	q_1	0.715	0
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0.627
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0.285	0
a_2	p_2	q_2	0	0

The companies in this case will get correspondingly 44421€ and 804€ immediate profits.

In $71.5\% \cdot 62.7\%$ cases

$$m^1_* = 0\% \quad m^2_* = 100\%$$

But because of inertness

$$p(s_6|s_9, (a^1_1, p^1_2, q^1_1), (a^2_2, p^2_1, q^2_1)) = 1.0$$

And the companies will earn 44421€ and −6696€ correspondingly.

In $28.5\% \cdot 37.3\%$ cases the first company by spending money on advertisements would attract only credulous customers (18% of the market).

But owing to inertness the next month market shares will be

$$m^{1\prime} \in [60\%, 70\%) \quad m^{2\prime} \in [30\%, 40\%)$$
$$p(s_7|s_9, (a^1_2, p^1_2, q^1_1), (a^2_1, p^2_1, q^2_1)) = 1.0$$

The immediate rewards of the firms in this case will be 36921€ and 804€.

In $28.5\% \cdot 62.7\%$ cases

$$p(s_6|s_9, (a^1_2, p^1_2, q^1_1), (a^2_2, p^2_1, q^2_1)) = 1.0$$

The firms will get correspondingly 36921€ and −6696€ in this case.

In this state the second firm is trying to attract the customers and gain higher market share even when it incurs losses.

Table 6.16. Policies for Market Shares $m^1 \in [90\%, 100\%]$, $m^2 \in [0\%, 10\%)$

			Company 1	*Company* 2
a_1	p_1	q_1	0	1
a_1	p_1	q_2	0	0
a_1	p_2	q_1	1	0
a_1	p_2	q_2	0	0
a_2	p_1	q_1	0	0
a_2	p_1	q_2	0	0
a_2	p_2	q_1	0	0
a_2	p_2	q_2	0	0

Market shares $m^1 \in [90\%, 100\%]$ and $m^2 \in [0\%, 10\%)$

The first firm will try to earn as much as possible from loyalty of almost 100% of the market by spending no money on advertisement and charging the maximum price for average quality chocolate (see table 6.16).

The second firm tries to gain a larger share of customers by offering the chocolate at the minimum price.

The profits in this month will be 49647€ and 268€ correspondingly.

$$m_*^1 = 0\% \quad m_*^2 = 100\%$$

$$p(s_7|s_{10}, (a_1^1, p_2^1, q_1^1), (a_1^2, p_1^2, q_1^2)) = 1.0$$

It won't be profitable to change the policy simultaneously in any of the states and when the firm choses to play its part of the Nash equilibrium there is no better policy for the other firm as to play the above strategy[7].

Since the customers can't leave the market in this model, the companies will behave monopolistically — they will never spend money on additional quality, instead the firms will force the customers to buy average quality bars at maximum price, offering discounts or resorting to the help of advertising agencies when they want to gain larger market shares.

A number of extensions could be proposed for this basic model. For example, we can include any stochastic customer behavior.

[7] This is though true only for two agent stochastic games.

6.2 Table Soccer

6.2.1 Problem

In this section we will consider a table soccer game presented in figure 6.1. The aim of the players is to maximize the difference between the goals scored and the goals conceded by controlling 6 foosmen. The player in possession of the ball can kick it straightly or diagonally by moving the corresponding stick. The opponent at this moment can move her foosmen to the best position to intercept the ball and prevent a goal. The players are allowed to change the positions of their foosmen only when the ball is in possession of one of the foosmen[8]. When scored the ball passes to an arbitrary foosman.

As in the real soccer, the foosmen can score a goal (see figures 6.2(a) and 6.2(b)) and an own goal (see figures 6.2(c) and 6.2(d)), pass the ball to a team member (see figures 6.3(a) and 6.3(b)) and lose the ball (see figures 6.3(c) and 6.3(d)).

This problem can be represented as the following general-sum discounted stochastic game.

The states correspond to the number of the foosman f in possession of the ball (see figure 6.1) as well as its position $left$ (the foosmen being firmly fastened to the stick, their positions are determined by the distance between the first foosman in the row and the border of the table soccer field)[9]:

$$\begin{aligned} s_1: &\quad f = 1 \quad left = 0 \\ s_2: &\quad f = 1 \quad left = 1 \\ s_3: &\quad f = 2 \quad left = 0 \\ s_4: &\quad f = 2 \quad left = 1 \\ &\quad \vdots \\ s_{24}: &\quad f = 12 \quad left = 1 \end{aligned}$$

As soon as the ball is intercepted by one of the foosmen, the players can choose actions (see table 6.17) in the resulting state.

Since the players would prefer to score sooner rather than later, the discount factor will be $\gamma = 0.9$.

On scoring the player gets 1 as immediate reward and its opponent gets -1, otherwise the rewards will be equal to 0.

Deterministic transitions for this game are presented in appendix B.

---

[8] Though such assumption seems a serious restriction, it is very credible for a fast game when the players have time for any deliberate actions only in the pauses between rash flights of the ball.

[9] Because the foosmen are symmetric in this game, the number of the states could still be reduced.

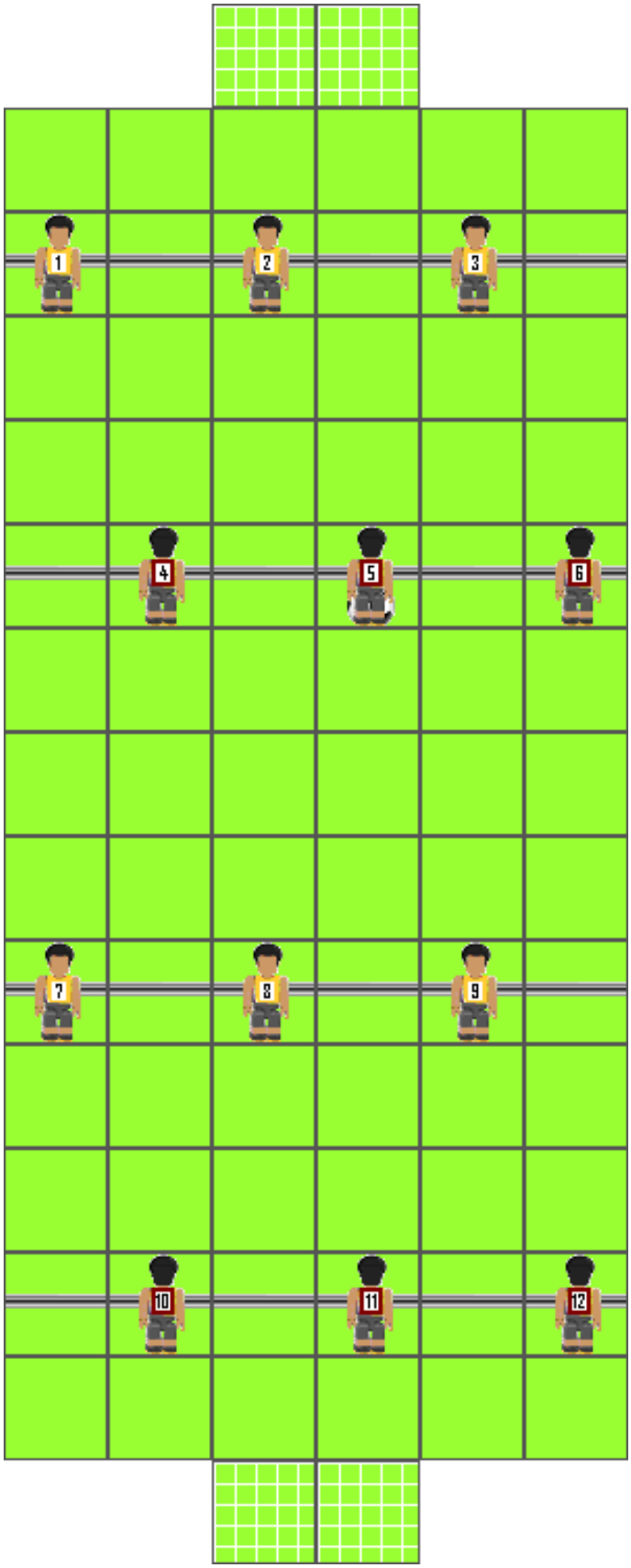
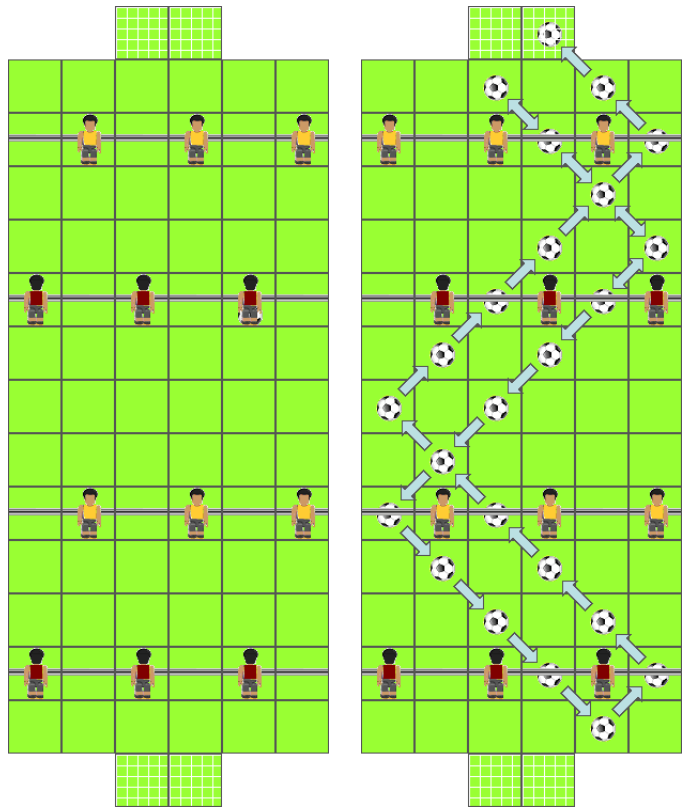
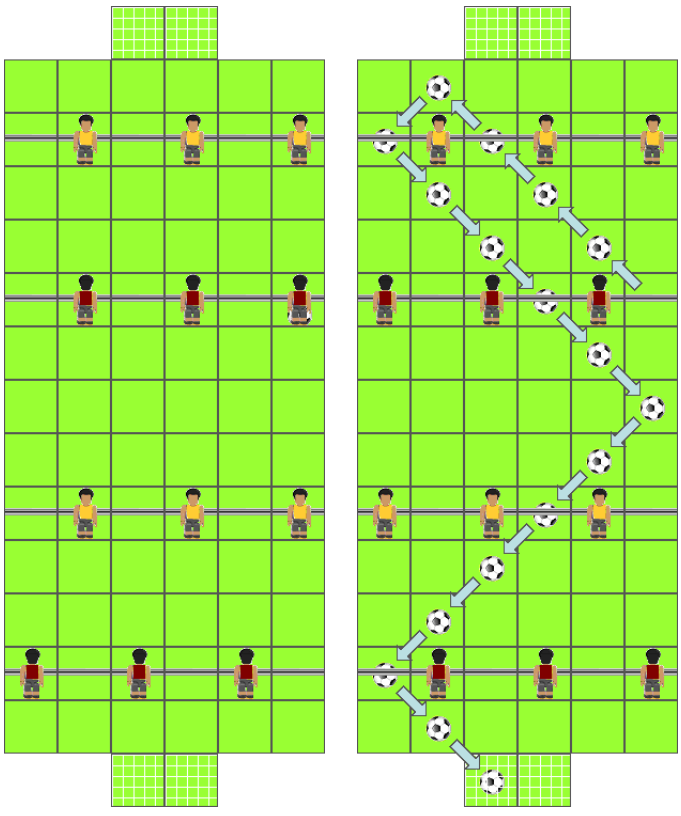


Fig. 6.1. Table Soccer Game

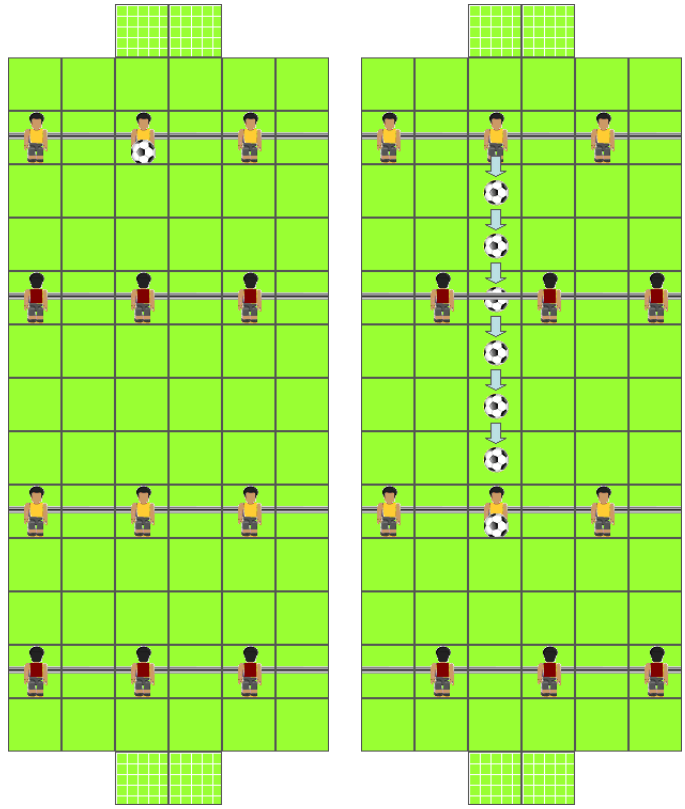


(a) Goal: Initial State (b) Goal: Ball Movement

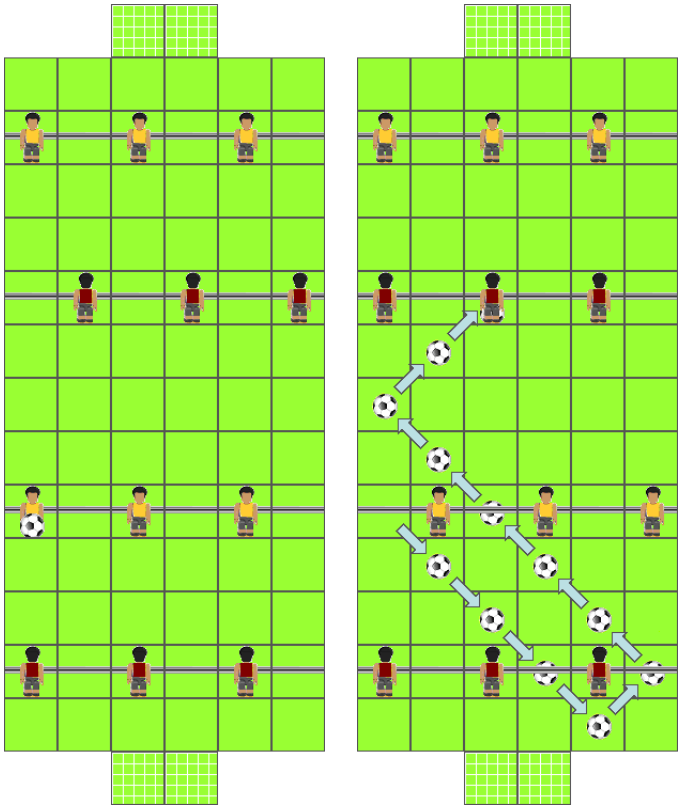


(c) Own Goal: Initial State (d) Own Goal: Ball Movement

Fig. 6.2. Goal and Own Goal



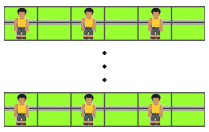
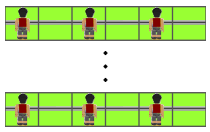
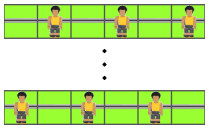
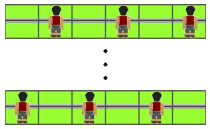
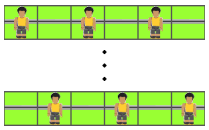
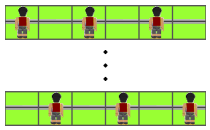
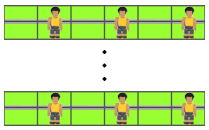
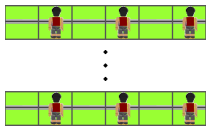
(a) Pass: Initial State (b) Pass: Ball Movement



(c) Loss: Initial State (d) Loss: Ball Movement

Fig. 6.3. Pass and Loss

Table 6.17. Players' Actions

a_1^1		a_1^2	
a_2^1		a_2^2	
a_3^1		a_3^2	
a_4^1		a_4^2	

6.2.2 Solution

With the use of Nash-RD approach we found a Nash equilibrium with accuracy 10^{-5} presented in tables 6.19 – 6.30 (the policies when foosmen 7 – 12 are about to kick the ball are the same for symmetric configurations).

The values of the Nash equilibrium to each player for every state can be found in table 6.18.

Let's interpret the policies that constitute the Nash equilibrium for some states.

Foosman 1 is in Possession of the Ball, its Position $left = 0$

As it can be seen from table 6.19, the player 1 will avoid kicking the ball diagonally as in 94.6% cases the ball could be intercepted by the fifth foosman of the opponent (see figures B.2(b), B.2(d), B.4(b) and B.4(d)), that scores with 50% probability as we will see.

Foosman 1 is in Possession of the Ball, its Position $left = 1$

This state is absolutely symmetric to the state when the third foosman is in possession of the ball and its position $left = 0$ that we are examining below.

Table 6.18. Values of States to Players in Table Soccer

	Player 1	*Player* 2
$s_1: \quad f=1 \quad left=0$	0	0
$s_2: \quad f=1 \quad left=1$	0	0
$s_3: \quad f=2 \quad left=0$	0	0
$s_4: \quad f=2 \quad left=1$	0	0
$s_5: \quad f=3 \quad left=0$	0	0
$s_6: \quad f=3 \quad left=1$	0	0
$s_7: \quad f=4 \quad left=0$	0	0
$s_8: \quad f=4 \quad left=1$	0	0
$s_9: \quad f=5 \quad left=0$	-0.5	0.5
$s_{10}: \quad f=5 \quad left=1$	-0.5	0.5
$s_{11}: \quad f=6 \quad left=0$	0	0
$s_{12}: \quad f=6 \quad left=1$	0	0
$s_{13}: \quad f=7 \quad left=0$	0	0
$s_{14}: \quad f=7 \quad left=1$	0	0
$s_{15}: \quad f=8 \quad left=0$	0.5	-0.5
$s_{16}: \quad f=8 \quad left=1$	0.5	-0.5
$s_{17}: \quad f=9 \quad left=0$	0	0
$s_{18}: \quad f=9 \quad left=1$	0	0
$s_{19}: \quad f=10 \quad left=0$	0	0
$s_{20}: \quad f=10 \quad left=1$	0	0
$s_{21}: \quad f=11 \quad left=0$	0	0
$s_{22}: \quad f=11 \quad left=1$	0	0
$s_{23}: \quad f=12 \quad left=0$	0	0
$s_{24}: \quad f=12 \quad left=1$	0	0

Foosman 2 is in Possession of the Ball, its Position $left = 0$

As it can be seen from table 6.21, player 1 will kick the ball only diagonally to prevent ball interception by the fifth foosman (see figures B.5(a), B.5(c), B.7(a) and B.7(c)). The goal in figure B.7(b) will be prevented by the second player. The configuration under which this goal is possible will be entirely avoided by her.

Foosman 2 is in Possession of the Ball, its Position $left = 1$

Symmetric to the previous state where the second foosman is at position $left = 0$ and possesses the ball.

Foosman 3 is in Possession of the Ball, its Position $left = 0$

In this state for the similar reasons player 1 will avoid kicking the ball diagonally (see figures B.10(b), B.10(d), B.12(b) and B.12(d)).

Foosman 3 is in Possession of the Ball, its Position $left = 1$

Symmetric to the state where foosman 1 is in possession of the ball and $left = 0$.

Foosman 4 is in Possession of the Ball, its Position $left = 0$

The fourth foosman possessing the ball, the player 2 will avoid kicking the ball diagonally since it could be intercepted by the eighth foosman of the first player (see figures B.14(a) and B.16(a)), that scores with 50% probability. A possible own goal in figure B.14(c) will be also thus precluded.

Foosman 4 is in Possession of the Ball, its Position $left = 1$

This state is symmetric to the state where the sixth foosman is about to kick the ball and is at position $left = 0$.

Foosman 5 is in Possession of the Ball, its Position $left = 0$

In 50% of cases player 1 will defend the left part of the goal and in half of the cases the right part. If player 2 chose to kick the ball always straightly, it would score in 50% of cases. If it chose to kick it diagonally, it would succeed also in 50% cases. The same reflections are correct for player 1 when its opponent follows policy presented in table 6.27. Thus, such policies constitute a Nash equilibrium of the corresponding bimatrix game. The values of this state are $(-0.5, 0.5)$ correspondingly.

Foosman 5 is in Possession of the Ball, its Position $left = 1$

Symmetric to the previous state where the fifth foosman is at position $left = 0$ and kicks the ball.

Foosman 6 is in Possession of the Ball, its Position $left = 0$

All passes and losses of the ball are neither beneficial nor dangerous for any player. The only situation to be prevented is a possible goal in figure B.22(c). As it can be seen from table 6.29 the action when this goal could be scored has zero probability under this Nash equilibrium.

Foosman 6 is in Possession of the Ball, its Position $left = 1$

The state is symmetric to the one where foosman 4 kicks the ball at the initial position $left = 0$.

As discussed above and could be seen from table 6.18 the beneficial (dangerous) situations are when the fifth and the eighth foosmen come into possession of the ball and the players are successfully trying to avoid passing the ball to these foosmen as well as configurations when the goals could be scored.

Table 6.19. Foosman 1 is in Possession of the Ball, $left = 0$

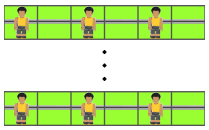
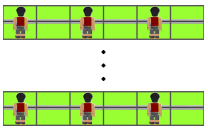
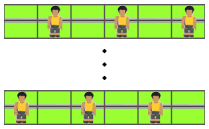
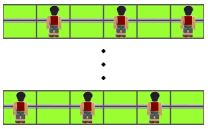
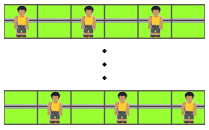
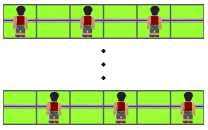
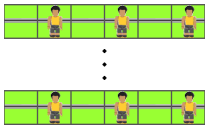
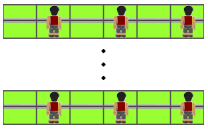
	0.431		0.042
	0		0.341
	0.569		0.012
	0		0.605

Table 6.20. Foosman 1 is in Possession of the Ball, $left = 1$

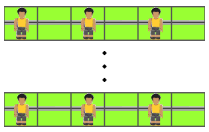
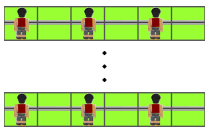
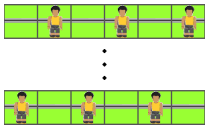
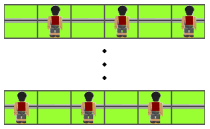
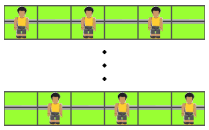
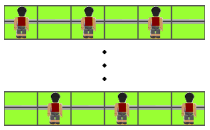
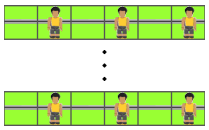
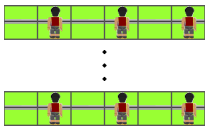
	0		0.353
	0.42		0.287
	0		0.221
	0.58		0.139

Table 6.21. Foosman 2 is in Possession of the Ball, $left = 0$

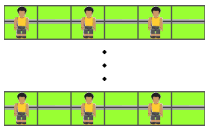
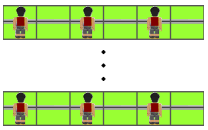
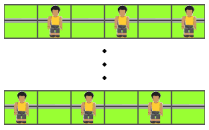
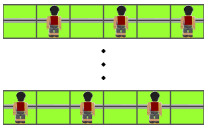
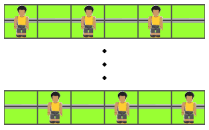
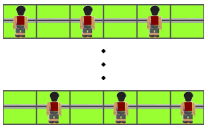
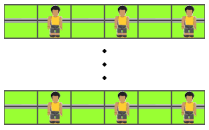
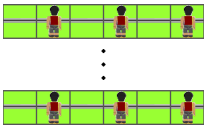
	0		0.823
	0.662		0.134
	0		0
	0.338		0.043

Table 6.22. Foosman 2 is in Possession of the Ball, $left = 1$

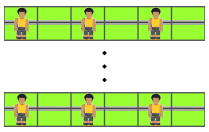
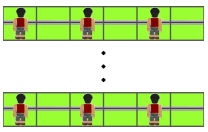
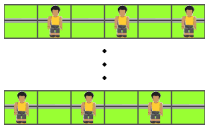
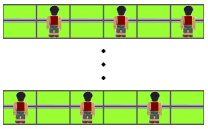
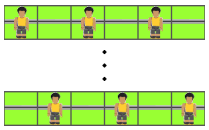
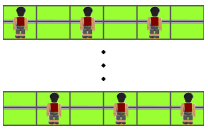
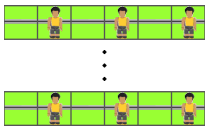
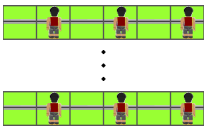
	0.338		0.043
	0		0
	0.662		0.134
	0		0.823

Table 6.23. Foosman 3 is in Possession of the Ball, $left = 0$

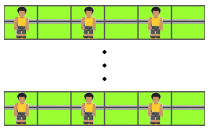
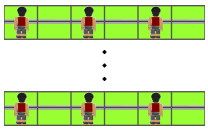
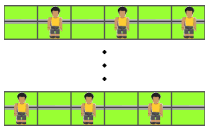
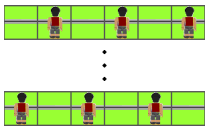
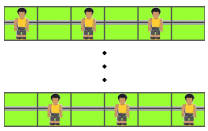
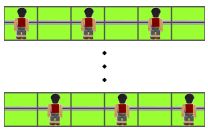
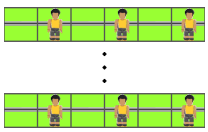
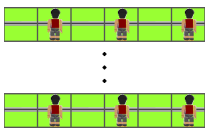
	0.58		0.139
	0		0.221
	0.42		0.287
	0		0.353

Table 6.24. Foosman 3 is in Possession of the Ball, $left = 1$

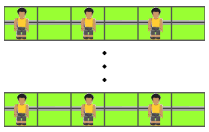
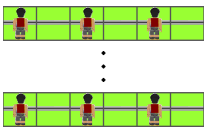
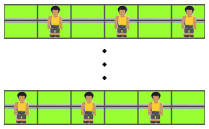
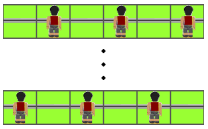
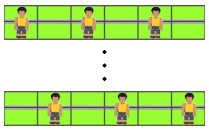
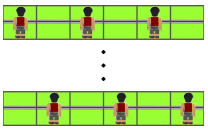
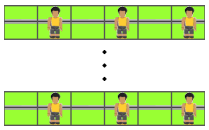
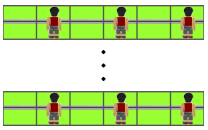
	0		0.605
	0.569		0.012
	0		0.341
	0.431		0.042

Table 6.25. Foosman 4 is in Possession of the Ball, $left = 0$

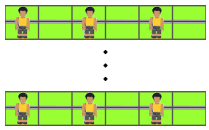
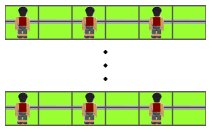
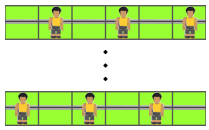
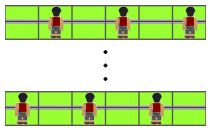
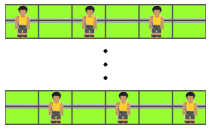
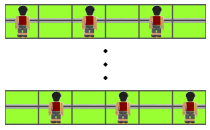
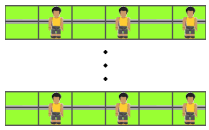
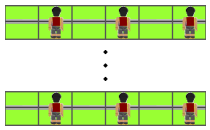
	0.348		0.5
	0.22		0
	0.213		0.5
	0.22		0

Table 6.26. Foosman 4 is in Possession of the Ball, $left = 1$

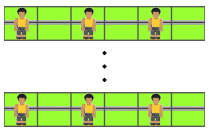
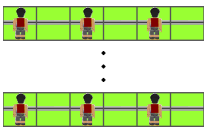
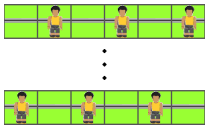
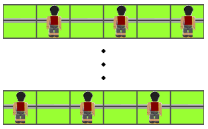
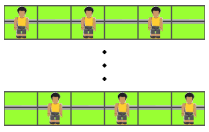
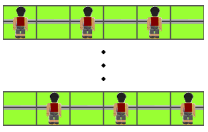
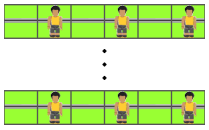
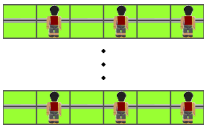
	0.378		0.145
	0		0.253
	0.378		0.348
	0.245		0.253

Table 6.27. Foosman 5 is in Possession of the Ball, $left = 0$

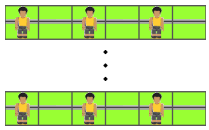
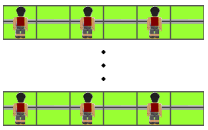
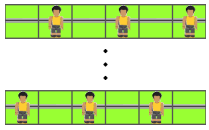
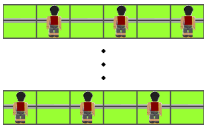
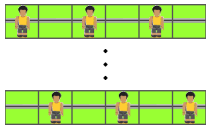
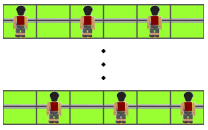
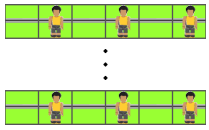
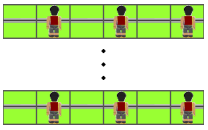
	0.25		0.25
	0.25		0.25
	0.25		0.25
	0.25		0.25

Table 6.28. Foosman 5 is in Possession of the Ball, $left = 1$

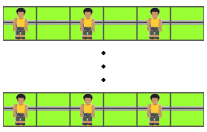
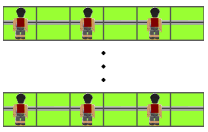
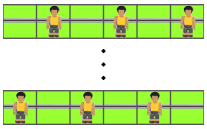
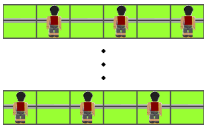
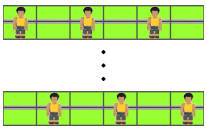
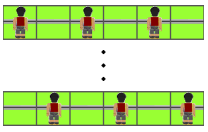
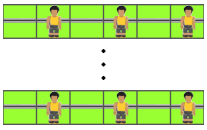
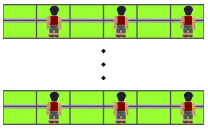
	0.25		0.25
	0.25		0.25
	0.25		0.25
	0.25		0.25

Table 6.29. Foosman 6 is in Possession of the Ball, $left = 0$

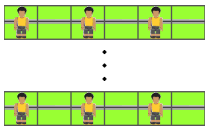
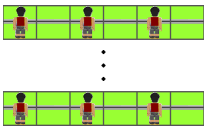
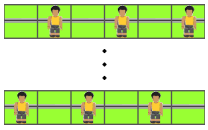
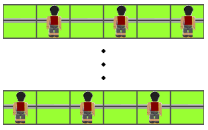
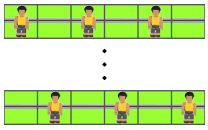
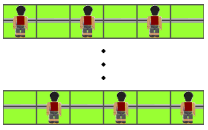
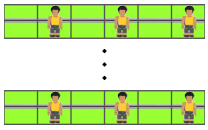
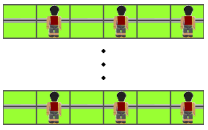
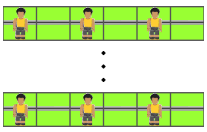
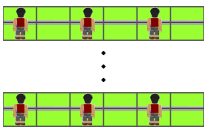
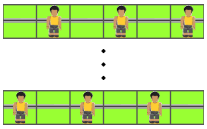
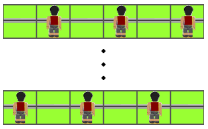
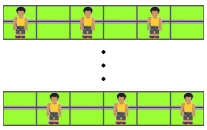
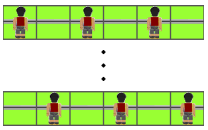
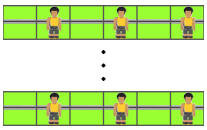
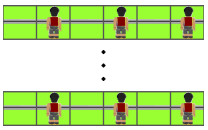
	0.245		0.253
	0.378		0.348
	0		0.253
	0.378		0.145

Table 6.30. Foosman 6 is in Possession of the Ball, $left = 1$

	0.22		0
	0.213		0.5
	0.22		0
	0.348		0.5

6.3 Double Auction

6.3.1 Problem

With the appearance of electronic commerce [127], [152], the development of the automatic trading agents became a separate important research area [120], [139], [62], [168]. In this section we will examine the Nash-RD based agents dealing at double auction. Double auction is a form of trading mechanism where several buyers and several sellers participate simultaneously. In each round the buyers name their bids and the sellers call their asks. In case the maximum bid b is higher or equal to minimum ask a, all the sellers who ask no more than clearing price $p = \frac{a+b}{2}$ have the possibility to sell at this price while all buyers whose bids were not lower than p can buy. Otherwise, the next round starts, in which the buyers' are allowed to name bids higher or equal to b of the last round or leave and the remaining sellers' asks must not exceed a.

Double auction can be very easily represented in the form of a stochastic game. The payoff will correspond to the difference between the good valuation and its final price, the states — to the states of the auction: the maximum bid and minimum ask of the last round, the actions — to the possible bids, the decision to leave will be encoded by the unacceptable bids and the discount

factor will reflect the impatience of the participants of the auction to make the deal. The resulting game will be deterministic.

Let us consider a double auction where the object to be traded is a new digital camera. Let us also assume that the number of cameras traded is restricted to one per auction. In case several participants can claim the camera according to the rules of the auction, it passes randomly to one of them. The initial price of the camera is 559€. The minimum raise is set to 10€[10]. Three buyers and two sellers participate in the auction and have quite different notions of the camera's real value (see tables 6.31). The highest valuation being equal to 600€, our participants can choose among five level of prices 559€, 569€, 579€, 589€, 599€. Let us imagine that our participants do not exaggerate the value of time and let the discount factor γ be 0.99.

6.3.2 Solution

Being equipped with Nash-RD algorithm, our agents converged to a Nash equilibrium with accuracy 1€ presented in table 6.32. The peculiarity of this Nash equilibrium is that the same bidding strategies are prescribed to the participants for all maxbid-minask combinations. As a result the first buyer purchases the camera from the first or the second seller at price 584€ in the first round of the auction. Let us check up that these policies really constitute a good approximation of Nash equilibrium. The buyer with camera valuation of 600€ must bid the highest price 599€. In case it reduces its bid to 589€, the clearing price will be $p = \frac{a+b}{2} = \frac{589+569}{2} = 579$€ and it will lose a part of its profit because of the increased probability that the third buyer will get the camera. The second and the third buyers will never get some profit (the difference between their camera valuations and the clearing price) and can't do any better (otherwise they'll get negative rewards). Each seller asks 569€ for the camera. If any of them tries to lower their ask to 559€, it will reduce the final price to 579€ but thus won't get rid of the competitor (the final price is still higher than the ask price of the rival). If it tries to raise its ask, it either will get the same profit (ask 579€) or will sell no camera at all (in case ask $>$ 579€). Hence, no participant will get higher profits (at least 1€ higher) if it unilaterally deviates from this approximation of Nash equilibrium.

Table 6.31. Participants' Valuations of the Camera

Buyer 1	*Buyer* 2	*Buyer* 3	*Seller* 1	*Seller* 2
600€	565€	580€	540€	525€

[10] At eBay the minimum bid for the price range 500.00€–999.99€ is exactly 10€.

Table 6.32. Double Auction Strategies

Bids	*Buyer* 1	*Buyer* 2	*Buyer* 3	*Seller* 1	*Seller* 2
559€	0	≈ 0.56	≈ 0	0	0
569€	0	≈ 0.44	≈ 0.02	1	1
579€	0	≈ 0	≈ 0.98	0	0
589€	0	0	0	0	0
599€	1	0	0	0	0

6.4 Foundations of Differential Game Theory

Definition 6.2. *A differential game problem is formulated as follows [110]:*

$$\max_{u_i \in U_i \subseteq \mathbb{R}^{m_i}} \left\{ \int_0^T e^{-\rho_i t} g_i(t, x(t), u_1(t), \ldots, u_N(t)) dt + e^{-\rho_i T} S_i(T, x(T)) \right\}$$

$$i = 1, \ldots, N$$

subject to

$$\dot{x}(t) = \frac{dx}{dt}(t) = f(t, x(t), u_1(t), \ldots, u_N(t))$$

$$x(0) = x_0$$

$$x \in X \subseteq \mathbb{R}^n$$

In case the horizon T is a fixed finite number the corresponding games are called *finite horizon differential games*. When $T = \infty$ we deal with *infinite horizon differential games*. T could also express some special terminating condition.

At every moment t the state of the system is determined by an n-vector of *state variables*

$$x(t) = (x_1(t), \ldots, x_n(t)) \in X$$

X denotes a *state space* here and x_0 is the system's *initial state*.

One or two state variables are usually sufficient to describe most applications [96]. For instance, $x_i(t)$, $i = 1, 2$, could stand for the market shares of two companies as in a chocolate duopoly model in section 6.1. In this case

differential game representation will be valid only if the companies' payoffs totally depend on their market shares.

At every instant t the players $i = 1, \ldots, N$ can take action $u_i(t)$:

$$u_i(t) = (u_{i1}(t), \ldots, u_{im_i}(t))$$

$$m_i \geq 1$$

$u_i(t)$ is called the *control variable* of player i. The set U_i designates the *control space* of player i. m_i is also restricted to two in most applications [96].

In the framework of differential games the system dynamics is described by a system of differential equation[11]:

$$\dot{x}(t) = \frac{dx}{dt}(t) = f(t, x(t), u_1(t), \ldots, u_N(t))$$

$$x(0) = x_0$$

It should be noted that the system evolves as a result of joint controls of all N players.

The choice of controls $u_1(t), \ldots, u_N(t)$ will yield the following instant reward rate to player i at time t and state $x(t)$:

$$g_i(t, x(t), u_1(t), \ldots, u_N(t))$$

The players' goal is to maximize their cumulative rewards

$$J_i(u_1(\cdot), \ldots, u_N(\cdot)) = \int_0^T e^{-\rho_i t} g_i(t, x(t), u_1(t), \ldots, u_N(t))dt + e^{-\rho_i T} S_i(T, x(T))$$

Discount rate $\rho_i = const$ and $\rho_i \geq 0$ determines the relative value of delayed versus immediate rewards.

If $\rho_i = 0$, rewards in the future are as valuable for player i as current ones. If $\rho_i \to \infty$, the player i doesn't care for the future at all.

$S_i(T, x(T))$ is a *salvage value*, the profit player i will get at the end of $[0, T]$ time interval.

Definition 6.3. *A strategy is a decision rule φ_i that associates some information with a control variable u_i, $i = 1, \ldots, N$.*

Definition 6.4. *A Markovian strategy is a decision rule φ_i of the form:*

$$u_i(t) = \varphi_i(t, x(t))$$

$$i = 1, \ldots, N$$

[11] Usually the evolution of the system is described by a system of ordinary differential equations, but the theory for games with stochastic differential equations has been also developed [169].

Definition 6.5. *A stationary strategy is a decision rule φ_i of the form:*

$$u_i(t) = \varphi_i(x(t))$$

$$i = 1, \ldots, N$$

It remains only to introduce Nash equilibrium solution concept for differential games. Informally, a tuple of strategies constitute a Nash equilibrium if it is not profitable for any player to deviate from its strategy when the other players stick to their strategies.

Definition 6.6. *A Nash equilibrium is an N-tuple of strategies $(\varphi_1, \ldots, \varphi_N)$, such that*

$$J_i(\varphi_1, \ldots, \varphi_N) \geq J_i(\varphi_1, \ldots, \varphi_{i-1}, u_i, \varphi_{i+1}, \ldots, \varphi_N)$$

$$\forall u_i \in U_i$$

$$\forall i = 1, \ldots, N$$

6.5 Application Fields of Differential Game Theory

In this section we will examine a number of differential game models of capital accumulation, advertising, pricing, marketing channels, macroeconomics, warfare and arms race, exploitation of renewable and nonrenewable resources and pollution. Studying these examples one can't help noticing that differential game models are rough simplifications of the problems encountered in these fields in reality. The models examined in the context of differential games are rather dictated by mathematical tractability than by practical plausibility [19]. At the current state of art the differential game theory allows to gain insights rather than to solve real problems. Stochastic game representation may shed light on the initial problems as it could allow to take into consideration the interdependences that were omitted in differential game representation in order to make them solvable. Of course we must sacrifice a certain accuracy discretizing the state space and control spaces but we gain freedom in incorporating any dependencies in reward function. We can introduce more control variables (actions) into the model as we have done in section 6.1 as well as reflect more sophisticated (even stochastic) transitions between the states and analyze the environment with more than two competing entities. Which model will induce more valuable insights should be considered in every particular case separately!

6.5.1 Capital Accumulation

In this section we introduce a general capital accumulation model [138].

N companies are competing in oligopolistic market by regulating their investment $I_i(t)$ and production $Q_i(t)$ rates, $i = 1, \ldots, N$.

The capital stock $K_i(t)$ of each company i evolves according to the following differential equation:

$$\dot{K}_i(t) = I_i(t) - \delta_i K_i(t)$$

where $\delta_i \geq 0$ is the depreciation rate.

The production rates are bounded by the available capital stock:

$$Q_i(t) \leq K_i(t)$$

The price $P(Q(t))$ is determined by inverse demand function $P(\cdot)$ of total production rate $Q(t) = Q_1(t) + Q_2(t) + \ldots + Q_N(t)$.

The cost function of company i is denoted by $m_i(\cdot)$.

The profit rate of each company depends on the production rates of all the firms in the market.

$$\pi_i(Q_1(t), \ldots, Q_N(t), K_1(t), \ldots, K_N(t)) = P(Q(t))Q_i(t) - m_i(Q_i(t))$$

The companies regulate their investment and production rates so as to maximize the cumulative profits

$$J_i = \int_0^\infty e^{-\rho_i t} \left[\pi_i(Q_1(t), \ldots, Q_N(t), K_1(t), \ldots, K_N(t)) - C_i(I_i(t))\right] dt$$

where $C_i(I_i(t))$ are the investment costs, $i = 1, \ldots, N$.

6.5.2 Advertising

In this section we will introduce several differential game advertising models [94]. The payoff functional to be optimized is the same for all the models and has the form:

$$J_i = \int_0^\infty e^{-\rho_i t} \left[q_i x_i(t) - u_i(t)\right] dt$$

where x_i denotes the sales rate of the i company, u_i is the rate of its advertising expenditure and q_i is profit brought by each sold unit, $i = 1, \ldots, N$.

A Lanchester Model

In general Lanchester model the system dynamics is described by the following differential equations [98]:

$$\dot{x}_i(t) = [m - x_i(t)]\, f_i(u_i(t)) - x_i(t) \sum_{j=1, j\neq i}^{N} f_j(u_j(t)) =$$
$$= m f_i(u_i(t)) - x_i(t) \sum_{j=1}^{N} f_j(u_j(t))$$

where $f_i(u)$ is an increasing advertisement response function and m is fixed sales volume.

A Vidale-Wolfe Model

In the initial task formulation [160] the sales rate x_i depends on the rate of advertising expenditure u_i at time t in such an unsophisticated manner:

$$\dot{x}_i(t) = \gamma u_i(t)\, [m - x_i(t)] - \delta x_i(t)$$
$$i = 1, \ldots, N$$

where $\gamma > 0$ is the level of response, $m > 0$ — maximum sales volume and $\delta > 0$ is the level of decay.

6.5.3 Pricing

Sticky-Price Oligopoly

Duopoly version of the model was first introduced in [134].

Oligopoly differential game will have the following form:

$$\max_{u_i \geq 0} \left\{ J_i = \int_0^{\infty} e^{-\rho_i t} \left[x(t) u_i(t) - c u_i(t) - \frac{1}{2} u_i(t)^2 \right] dt \right\}$$

$$i = 1, \ldots, N$$

subject to

$$\dot{x}(t) = k \left[a - 2b \frac{u_1(t) + u_2(t) + \ldots + u_N(t)}{N} - x(t) \right]$$
$$x(0) = x_0$$
$$x(t) \geq 0 \text{ for all } t \geq 0$$

In this model N companies are assumed to sell an identical product. Its price determined by inverse demand depends on the firms' production rates u_i linearly:

$$\tilde{x}(t) = a - 2b\frac{u_1(t) + u_2(t) + \ldots + u_N(t)}{N}$$
$$a, b > 0$$

The actual price changes gradually towards $\tilde{x}$:

$$\dot{x}(t) = k\left[\tilde{x}(t) - x(t)\right]$$

where k is the adjustment coefficient.

The players are to choose the production rates

$$u_i(t) \geq 0$$

so as to maximize their discounted cumulative profits J_i.

The cost function is quadratic here:

$$C(u_i(t)) = cu_i(t) + \frac{1}{2}u_i(t)^2$$
$$c > 0$$

Production-Advertising Oligopoly

In production-advertising game analyzed in [37] N companies offer an identical good and thus constitute an oligopoly.

The companies choose the production $q_i(t)$ as well as advertising expenditure $a_i(t)$ rates, $i = 1, \ldots, N$.

By spending money on advertisement, the companies increase the maximum price the customers are ready to pay for the good:

$$\dot{r}(t) = \sum_{i=1}^{N} a_i(t) - \delta r(t)$$

$$r(0) = r_0 > 0$$

$\delta > 0$ is a decay rate, that expresses customers' forgetfulness of former advertising campaigns.

The price $p(t)$ is determined by the inverse demand function:

$$p(t) = \left[r(t) - Q(t)\right]^{\frac{1}{\alpha}}$$

where $\alpha > 0$ and $Q(t) = \sum_{i=1}^{N} q_i(t)$ is the total production rate.

Each firm chooses its production rate $q_i(t)$ and the advertising rate $a_i(t)$ so as to maximize their discounted cumulative profits.

$$J_i = \int_0^\infty e^{-\rho_i t} \left[p(t) q_i(t) - c_i(q_i(t)) \right] dt$$

where $c_i(q_i(t))$ are cost functions.

6.5.4 Marketing Channels

In this section we will study a goodwill accumulation model in the context of marketing channels [95]. Our marketing channel is composed of two independent firms: a manufacturer and a retailer.

The control variables of the retailer and manufacturer consist of the consumer advertising expenditure rates $a_R(t)$ and $a_M(t)$ as well as the consumer $p_R(t)$ and the transfer $p_M(t)$ prices.

The advertising efforts increase the stock of goodwill G which in its turn reflects the customers' willingness to buy the product.

The goodwill dynamics is described thus:

$$\dot{G}(t) = a_M(t) + a_R(t) - \delta G(t)$$

$$G(0) = G_0$$

where δ is decay rate.

The sales rate depends on the stock of goodwill as well as on the consumer price:

$$S(p_R(t), G(t)) = [\alpha - \beta p_R(t)] \left[g_1 G(t) - \frac{g_2}{2} G(t)^2 \right]$$

where $\alpha > 0$, $\beta > 0$, $g_1 > 0$ and $g_2 > 0$.

The firms as usual strive to maximize their discounted cumulative pure profits:

$$J_M = \int_0^\infty e^{-\rho_M t} \left[(p_M(t) - c) S(p_R(t), G(t)) - \frac{w}{2} a_M(t)^2 \right] dt$$

$$J_R = \int_0^\infty e^{-\rho_R t} \left[(p_R(t) - p_M(t)) S(p_R(t), G(t)) - \frac{w}{2} a_R(t)^2 \right] dt$$

where $w > 0$ and c is the production cost.

6.5.5 Macroeconomics

In this section we will consider a differential game model of trade-off between unemployment rate and inflation faced by the government and central bank [53].

Let $\pi(t)$ be the price inflation and $H(t)$ denote the aggregate excess demand in the goods and labor markets at time t.

Excess demand $H(t)$ is assumed to depend on the growth rate of the nominal money supply $m(t)$ and the growth rate of real public expenditures for good and services $g(t)$ in the following way:

$$H(t) = \beta\left[m(t) - \pi(t)\right] + \gamma g(t)$$

where $\beta > 0$ and $\gamma > 0$.

Inflation rate $\pi(t)$ is equal to linear combination of excess demand $H(t)$ and the expected inflation rate $\pi^*(t)$:

$$\pi(t) = \lambda H(t) + \pi^*(t)$$
$$\lambda > 0$$

It is supposed that the companies' inflation expectation rate $\pi^*(t)$ changes gradually.

$$\dot{\pi}^*(t) = \eta\left[\pi(t) - \pi^*(t)\right]$$
$$\eta > 0$$

Let the normal rate of unemployment be u_N.

The unemployment rate $u(t)$ is assumed to be equal to:

$$u(t) = u_N - \delta H(t)$$

where $\delta > 0$.

The government controls the growth rate of real public expenditures, and the central bank is responsible for the growth rate of the nominal money supply.

By regulating these control variables they are trying to minimize:

$$J_i = \int_0^\infty e^{-\rho_i t}\left[a_i(u(t) - u_N) + b_i\pi(t)\right] dt$$

where $a_i, b_i > 0$, $i = 1, 2$.

6.5.6 Warfare and Arms Race

War of Attrition and Attack

In the war of attrition and attack [92], the confronting armies must distribute their air forces for direct battle and air raids on enemy's plane supply.

The amount of aircraft in confronting armies x_1 and x_2 evolves according to the following differential equations.

$$\dot{x}_1(t) = m_1(t) - c_2\phi_2(t)x_2(t) - l_1\phi_1(t)x_1(t) - L_1(1 - \phi_1(t))x_1(t)$$

$$\dot{x}_2(t) = m_2(t) - c_1\phi_1(t)x_1(t) - l_2\phi_2(t)x_2(t) - L_2(1 - \phi_2(t))x_2(t)$$

where $m_1(t)$ and $m_2(t)$ denote the aircraft production rates, $\phi_1(t)$ and $\phi_2(t)$ are the shares of planes destroying the enemy's aircraft supply at moment t of confrontation, c_1, c_2 are the rates of their effectiveness l_i, L_i — airplane loss rates during the enemy's supply attack caused by actions of the defense and as a result of active combat.

Only the shares of aircraft that participate in active war are included in the payoffs.

$$J_1 = \int_0^\infty e^{-\rho_1 t} \left[(1-\phi_1(t))x_1(t) - (1-\phi_2(t))x_2(t)\right] dt$$

$$J_2 = \int_0^\infty e^{-\rho_2 t} \left[(1-\phi_2(t))x_2(t) - (1-\phi_1(t))x_1(t)\right] dt$$

Discount factor reflects the notion that the airplanes are more worth in the beginning of the war.

Missile War

In the model introduced in [91] the conflicting nations target their missiles at the missile supply of each other as well as civil towns.

Control variables α and β correspond to the intensity of launching missiles. The distribution of missile attack between the two goals can be regulated by α' and β' control variables.

The reduction of missile supply of two conflicting nations $M_A(t)$ and $M_B(t)$ as well as demolition of the cities $C_A(t)$ and $C_B(t)$ are described by the following differential equations.

$$\begin{aligned}
\dot{M}_A(t) &= -\alpha(t)M_A(t) - \beta'(t)\beta(t)M_B(t)f_B \\
\dot{M}_B(t) &= -\beta(t)M_B(t) - \alpha'(t)\alpha(t)M_A(t)f_A \\
\dot{C}_A(t) &= -(1-\beta'(t))\beta(t)M_B(t)v_B \\
\dot{C}_B(t) &= -(1-\alpha'(t))\alpha(t)M_A(t)v_A
\end{aligned}$$

where f_A and f_B correspond to the effectiveness of missiles against missile supply and v_A and v_B express the destructive power of the rockets against enemy's civil population.

The nation whose civil population will be first destroyed to a certain capitulation level loses the war.

Arms Race

In arms race model [30] the subject of investigation is the trade-off between the level of consumption and the feeling of security induced by the accumulated weapon stock.

Let $C_i(t)$ be the consumption rate of country i, $i = 1, 2$, $Z_i(t)$ be its expenditure on maintaining and extending its weapon supply at time t and Y_i be the net national product of country i.

$$Y_i = C_i(t) + Z_i(t)$$

Let $\beta_i w_i$ be the expenditure on maintaining the weapon supply. Then its increase is described by the following differential equation:

$$\dot{w}_i(t) = Z_i(t) - \beta_i w_i(t)$$

Each country strives to maximize its cumulative discounted utility that depends on the level of consumption C_i and the feeling of security $D_i(w_i, w_j)$:

$$J_i = \int_0^\infty e^{-\rho_i t} U_i \left[C_i(t), D_i(w_i(t), w_j(t))\right] dt$$

$$i, j = 1, 2$$

$$i \neq j$$

6.5.7 Resource Economics

Nonrenewable Resources

In this section we present a model of extraction of nonrenewable natural resources (e.g., an oil field) by N players [48].

Let $c_i(t)$ correspond to the intensity of oil extraction by player i at time t, $i = 1, \ldots, N$.

The dynamics of decrease of oil stock $x(t)$ has the following form:

$$\dot{x}(t) = -\sum_{i=1}^{N} c_i(t)$$

The cumulative reward depends on the strictly concave and increasing utility $u(c_i)$ of the extracted oil.

$$J_i = \int_0^\infty e^{-\rho_i t} u(c_i(t)) dt$$

Fishery Games

Differential game formulation of fishery problem was first proposed in [108]. The developed model is general enough to describe exploitation of any renewable resource like hunting, felling and fishery by N autonomous players.

The players must decide on exploitation rate u_i of the renewable resource x, $i = 1, \ldots, N$.

The players face the inverse demand for their offered resource $p(\cdot)$ satisfying the following conditions

$$\frac{\partial p}{\partial u_i} < 0$$

$$u_i \frac{\partial^2 p}{\partial u_i^2} < -2\frac{\partial p}{\partial u_i}$$

The cost $c(\cdot)$ is a decreasing convex function of resource stock x.
$G(x)$ is a natural growth function that is increasing and concave.
The resource stock evolves according to the equation.

$$\dot{x}(t) = G(x(t)) - \sum_{i=1}^{N} u_i(t)$$

The players aspire to control their rates of exploitation so as to maximize their cumulative profits.

$$\max_{u_i \geq 0} \left\{ J_i = \int_0^\infty e^{-\rho_i t} \left[p\left(\sum_{j=1}^{N} u_j(t) \right) - c(x(t)) \right] u_i(t) dt \right\}$$
$$i = 1, \ldots, N$$

Pollution

This pollution differential game [48] models a pollution-consumption trade-off faced by two countries.

Let $E_i(t)$ denote the amount of polluting substances emitted in the atmosphere as a side effect of production of $Y_i(t)$ quantity of good.

$$E_i = G_i(Y_i)$$

The pollution level increases as a result of emissions of both countries and reduces owing to natural clarification.

$$\dot{S}(t) = E_1(t) + E_2(t) - kS(t)$$

where $k > 0$ is the nature recovering rate.

The countries strive to maximize their cumulative comfort that has the following form

$$J_i = \int_0^\infty e^{-\rho_i t} \left[aY_i(t) - \frac{1}{2} Y_i(t)^2 - \frac{b}{2} S(t)^2 \right] dt$$

where $a, b > 0$.

6.6 Conclusion

In this chapter we applied Nash-RD approach to chocolate duopoly model, table soccer and double auction.

The chocolate duopoly model serves as an illustration of the proposed in this chapter idea — to represent the economic problems traditionally modeled by differential games as stochastic games and solve them with the use of the developed approach. The certain accuracy will be sacrificed, but we gain freedom in setting transition and reward functions. The loss of accuracy would be fatal if differential games were not rough simplifications themselves. Stochastic games could be a powerful mechanism to understand many economic processes.

Creating a team of robots that will beat humans in football is a challenge to the whole field of artificial intelligence [100], [99]. In this chapter we apply the developed Nash-RD approach to a simplified version of soccer — a table football.

With the appearance of electronic commerce, the development of the automatic trading agents became a separate important research area. In this chapter we examining Nash-RD equipped agents trading at double auction. It would be also interesting to study the behavior of the Nash-RD based trading agents in the context of combinatorial auctions [42], multi-stage auctions [144], FCC spectrum auctions [45], [165], simultaneous auctions [101] (e.g., at eBay) as well as use the developed approach as a basis for shopbots and pricebots [10], [120], [97].

Other traditional reinforcement learning tasks in multi-agent environments include numerous and diverse problems of task execution by several robots.

Part V

Summary

7

Conclusion and Future Work

7.1 Contributions of the Thesis

Let us enumerate the main contributions of this thesis:

- **Nash-RD approach** (introduced in chapter 4) that allows to compute stationary Nash equilibria of general-sum discounted stochastic games with a given accuracy. For the case when the games are known from the very beginning (algorithmic game theoretic case), much higher percentage of general-sum discounted stochastic games could be solved than by the existing methods: stochastic tracing procedure and nonlinear optimization. When the games are being learned by interaction with the environment (reinforcement learning case), the developed approach is the first and the only approach that allows to find Nash equilibria for high percentage of general-sum discounted stochastic games.
- **Formal proof** of Nash-RD convergence to a Nash equilibrium under certain assumptions and several theoretical results for replicator dynamics in stochastic games — the basis of Nash-RD approach (in chapter 3).
- **Decision making analysis of multi-agent reinforcement learning** (carried out in chapter 5). Multi-agent reinforcement learning has been for the first time considered from decision making perspective. It turned out that under some assumptions each state of the stochastic game can be represented as a game against nature. A number of multi-agent reinforcement learning algorithms have been analyzed from this perspective.
- **OPVar-Q algorithm** (proposed in chapter 5) based on variable Hurwicz's optimistic-pessimistic criterion for choosing the best strategy in games against nature was developed. Hurwicz's criterion allows us to embed initial knowledge of how friendly the environment in which the agent is supposed to function will be. A formal proof of the convergence of the algorithm to stationary policies is given. The variability of Hurwicz's criterion allowed it to converge to best-response strategies against opponents with stationary policies. Thorough testing of the developed algorithm against

other multi-agent reinforcement learning algorithms showed that OPVar-Q functions on the level of its non-convergent (or irrational) opponents in environments of different level of amicability.
- The applications of Nash-RD to chocolate duopoly model, table soccer and double auction (in chapter 6).
- The investigation of **potential stochastic game representation of economic problems** (conducted in chapter 6) in the field of capital accumulation, advertising, pricing, macroeconomics, warfare and resource economics that are traditionally represented as differential games.

7.2 Future Work

The following directions for future research seem to us most promising:

1. Analyze the behavior of the resulting system of nonlinear differential equations 3.2 for different game classes (different number of states, agents, actions, different conditions on reward functions and transitions):
   a) in the neighborhood of a Nash equilibrium:
      i. statistically analyze eigenvalues (positive, negative, imaginary, pure imaginary) in the neighborhood of the Nash equilibrium depending on conditions on reward functions and transition probabilities
   b) globally:
      i. with program system AUTO [78]
      ii. with global methods
      iii. existence of periodic orbits
      iv. existence of spirals
      v. analysis of periodic orbits
      vi. analysis of spirals
2. Single out classes for which Nash-RD approach will always converge:
   a) as a result of research goal 1
   b) tree like stochastic games
3. Prove the assumptions under which Nash-RD converges to a Nash equilibrium formally for game classes singled out in goal 2.
4. Research the potential of development of multi-agent reinforcement learning algorithms for continuous environments on the basis of differential games (at present continuous problems are discretized and then solved by reinforcement learning algorithms and there are no theoretical foundations for convergence of algorithms even in one-agent case, differential game theory is very well theoretically founded).
5. Analyze the Nash-RD results from the common good perspective (the common good measure reflects how good the found Nash equilibrium is for the whole society of agents).

6. Research the potential of development of multi-agent reinforcement learning algorithms for finite horizon games (inspired by results in game theory: for finite horizon stochastic games there are known approaches that allow to find non-stationary Nash equilibria efficiently for large games).
7. Find possible application areas of finite horizon stochastic games.
8. Research the potential of development of multi-agent reinforcement learning algorithms on the basis of efficient linear programming techniques for special stochastic game classes:
   a) single-controller discounted games [121]
   b) separable reward state independent transition (SER-SIT) discounted stochastic games [137]
   c) switching controller discounted stochastic games [54]
   d) find possible areas of applications for the developed multi-agent reinforcement learning algorithms
9. Development of stochastic game solver (stochastic games have a broad area of application in economics, experimentally we showed that the developed Nash-RD approach can solve higher percentage of stochastic games than the existing approaches, therefore it seems promising to try to develop faster versions of our approach):
   a) speed up by development of parallel version of the algorithm
   b) speed up by policy evaluation method [141]
10. Use reinforcement learning techniques to deal with large state / action spaces (a number of methods were developed to enable reinforcement learning algorithms (one-agent case) to deal with large state / action spaces, it seems promising to use them for our multi-agent reinforcement learning algorithm).
11. Analyze the possibilities to develop discrete form of Nash-RD algorithm (from evolutionary game theory we know that discretization of replicator dynamics in matrix games sometimes leads to loss of all useful properties [166]), research the loss of useful properties when discretized by:
    a) OLG dynamics [166]
    b) numerical methods for solving systems of differential equations (e.g., Runge-Kutta: implicit, explicit)
12. Investigate the potential of stochastic game representation of economic problems traditionally modeled as differential games in detail.
13. Single out application areas for which approximation of Nash equilibrium will be sufficient.
14. Develop learning by observation approaches for multi-agent environments (see [4] for one-agent environments).

Part VI

Appendices

A

Foundations of Theory of Ordinary Differential Equations

In this appendix we are quoting basic definitions and theorems of the theory of ordinary differential equations that could be found in [64], [166], [79], [159], [13].

Definition A.1. *A differential equation is an equation that involves the derivatives of a function of time as well as the function itself.*

Definition A.2. *The order of a differential equation is the order of the highest derivative of the function that appears in the equation.*

Definition A.3. *A system of differential equations that does not depend on time is called autonomous (or time homogeneous).*

Definition A.4. *If only ordinary derivatives are involved, the equation is called an ordinary differential equation (in contrast to partial differential equation).*

A system of k autonomous, first-order, ordinary differential equations could be written in vector form as

$$\dot{\mathbf{x}} = \varphi(\mathbf{x}) \tag{A.1}$$

where

$$\dot{\mathbf{x}} = (\dot{x_1}, \ldots, \dot{x_k}) = \frac{d\mathbf{x}}{dt} = \left(\frac{dx_1}{dt}, \ldots, \frac{dx_k}{dt} \right)$$

and $\varphi : X \to \mathbb{R}^k$ is continuous, X is an open set and $X \subset \mathbb{R}^k$.

The function φ is a vector field and defines at each state $\mathbf{x}$ the direction and velocity of the change.

Definition A.5. *A (local) solution through a point $\mathbf{x}^0 \in X$ to a system A.1 is a function $\xi(\cdot, \mathbf{x}^0) : T \to X$, where T is an open interval containing $t = 0$, such that*

$$\xi(0, \mathbf{x}^0) = \mathbf{x}^0$$

$$\frac{d}{dt}\xi(t, \mathbf{x}^0) = \varphi\left[\xi(t, \mathbf{x}^0)\right]$$

holds for all $t \in T$. The solution is global if $T = \mathbb{R}$.

Definition A.6. *A function $\varphi : X \to \mathbb{R}^k$, where $X \subset \mathbb{R}^k$, is (locally) Lipschitz continuous if for every closed bounded set $C \subset X$ there exists some real number λ such that*

$$\|\varphi(\mathbf{x}) - \varphi(\mathbf{y})\| \leq \lambda\|\mathbf{x} - \mathbf{y}\|$$

for all $\mathbf{x}, \mathbf{y} \in C$.

Theorem A.7. *[166], [64] If $X \subset \mathbb{R}^k$ is open and the vector field $\varphi : X \to \mathbb{R}^k$ is Lipschitz continuous, then the system A.1 has a unique solution $\xi(\cdot, \mathbf{x}^0)$ through every state $\mathbf{x}^0 \in X$. Moreover $\xi(t, \mathbf{x}^0)$ is continuous in t and $\mathbf{x}^0$.*

Proposition A.8. *[13], [166] If $X \subset \mathbb{R}^k$ is open and the vector field $\varphi : X \to \mathbb{R}^k$ has continuous first partial derivatives, then it is Lipschitz continuous.*

Theorem A.9. *[29] If the vector field $\varphi : X \to \mathbb{R}^k$ has continuous first partial derivatives with respect to $x_1, \ldots, x_k$, then the system A.1 has a unique solution $\xi(\cdot, \mathbf{x}^0)$ through every state $\mathbf{x}^0 \in X$. Moreover $\xi(t, \mathbf{x}^0)$ is continuous in t and $\mathbf{x}^0$.*

Proposition A.10. *[166], [64] Suppose that $X \subset \mathbb{R}^k$ is open, that $\varphi : X \to \mathbb{R}^k$ is Lipschitz continuous, and that C is a compact subset of X such that $\xi(t, \mathbf{x}^0) \in C$ for all $\mathbf{x}^0 \in C$ and $t \in T(\mathbf{x}^0)$. Then $T(\mathbf{x}^0)$ can be taken to be $\mathbb{R}$, and the induced solution mapping $\xi : \mathbb{R} \times C \to C$ will meet the three conditions:*

1. $\xi(0, \mathbf{x}) = \mathbf{x}\ \forall \mathbf{x} \in C$
2. $\xi\left[t, \xi(s, \mathbf{x})\right] = \xi(t + s, \mathbf{x})\ \forall \mathbf{x} \in C,\ \forall s, t \in \mathbb{R}$
3. *ξ is continuous*

Definition A.11. *The (solution) trajectory (or path) $\tau(\mathbf{x}^0)$ through a state $\mathbf{x}^0 \in C$ is the graph of the solution $\xi(\cdot, \mathbf{x}^0)$:*

$$\tau(\mathbf{x}^0) = \left\{(t, \mathbf{x}) \in \mathbb{R} \times C :\ \mathbf{x} = \xi(t, \mathbf{x}^0)\right\}$$

Definition A.12. *The orbit $\gamma(\mathbf{x}^0)$ through an initial state $\mathbf{x}^0$ is the image of the whole time axis under the solution mapping $\xi(\cdot, \mathbf{x}^0)$:*

$$\gamma(\mathbf{x}^0) = \left\{\mathbf{x} \in C :\ \mathbf{x} = \xi(t, \mathbf{x}^0)\ \textit{for}\ t \in \mathbb{R}\right\}$$

Definition A.13. *A stationary state (or equilibrium) under a solution mapping ξ is a state $\mathbf{x} \in C$ such that $\xi(t, \mathbf{x}) = \mathbf{x}$ for all $t \in \mathbb{R}$.*

Proposition A.14. *[166] If $\mathbf{x}, \mathbf{y} \in C$ and*

1. $\lim_{t \to +\infty} \xi(t, \mathbf{x}) = \mathbf{y}$

2. $\xi\left[t, \xi(s, \mathbf{x})\right] = \xi(t+s, \mathbf{x})\ \forall \mathbf{x} \in C,\ \forall s, t \in \mathbb{R}$
3. ξ *is continuous*
then $\mathbf{y}$ *is stationary.*

Definition A.15. *A state* $\mathbf{x} \in C$ *is Lyapunov stable if every neighborhood* B *of* $\mathbf{x}$ *contains a neighborhood* B^0 *of* $\mathbf{x}$ *such that* $\xi(t, \mathbf{x}^0) \in B$ *for all* $\mathbf{x}^0 \in B^0 \cap C$ *and* $t \geq 0$*. A state* $\mathbf{x} \in C$ *is asymptotically stable if it is Lyapunov stable and there exists a neighborhood* B^* *such that*

$$\lim_{t \to \infty} \xi(t, \mathbf{x}^0) = \mathbf{x}$$

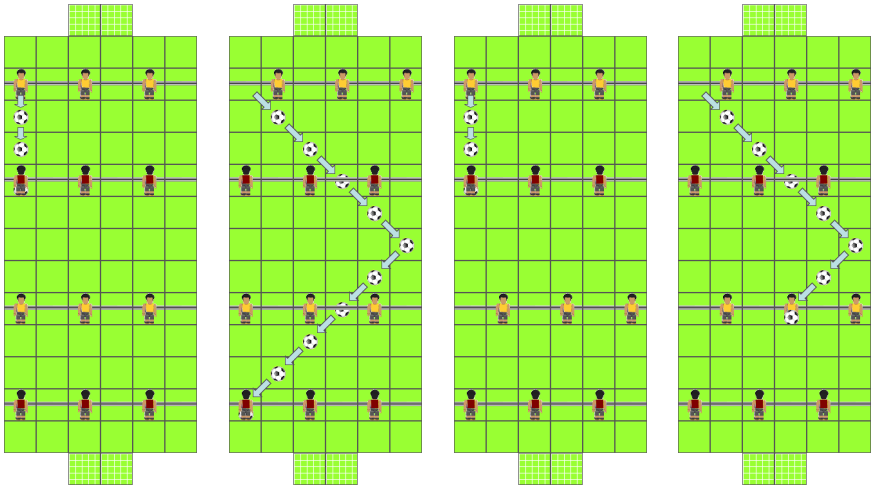
holds for all $\mathbf{x}^0 \in B^* \cap C$*.*

Proposition A.16. *[166] If a state is Lyapunov stable, then it is stationary.*

B

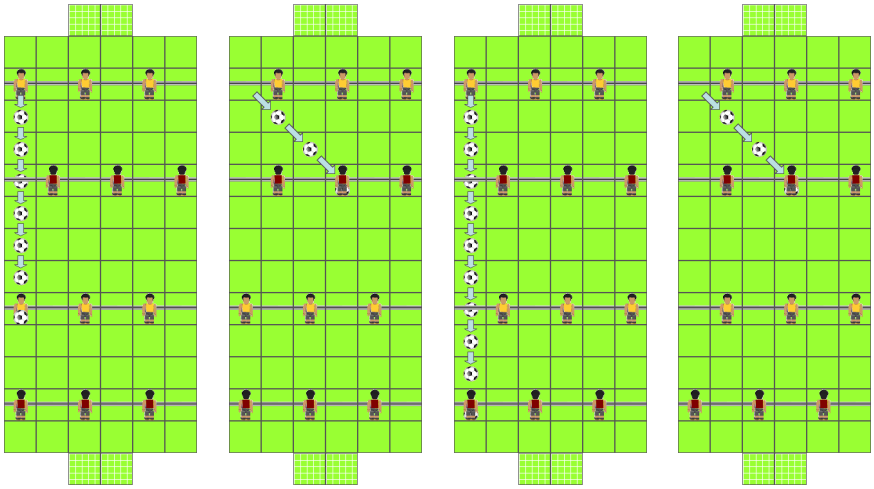
Table Soccer Transitions

In this appendix the deterministic transitions of table soccer game examined in section 6.2 are presented. Though the illustrations are not exhaustive, due to symmetry the rest transitions are obvious. The initial state is the same for all transitions under consideration and presented in figure B.1(a). The actions of the players are characterized by the final distances between the first foosmen in each controllable row and the left border of the table soccer field. Figures B.1 – B.4 illustrate the movements of the ball and the final states as a result of corresponding joint actions when the first foosman kicks the ball, figures B.5 – B.8 — when the second foosman was in possession of the ball and so forth.



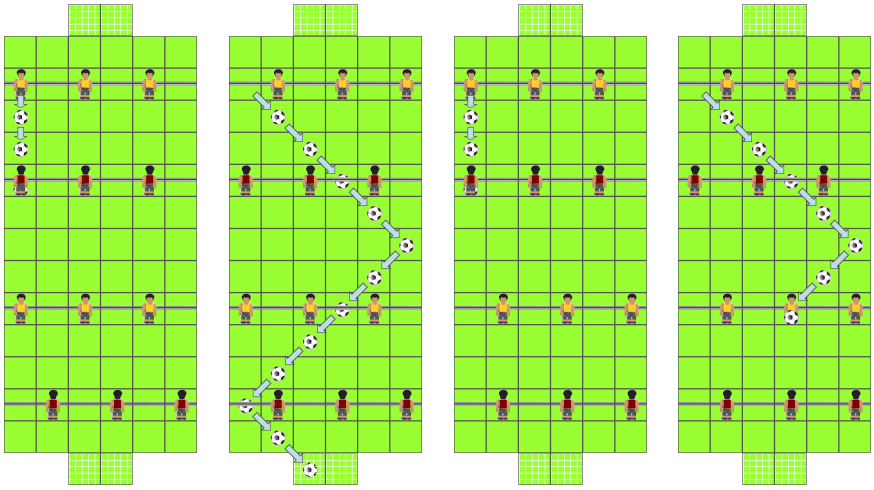
(a) $a_1^1 = (0,0)$ $a_1^2 = (0,0)$ (b) $a_2^1 = (1,0)$ $a_1^2 = (0,0)$ (c) $a_3^1 = (0,1)$ $a_1^2 = (0,0)$ (d) $a_4^1 = (1,1)$ $a_1^2 = (0,0)$

Fig. B.1. Foosman 1 Kicks the Ball



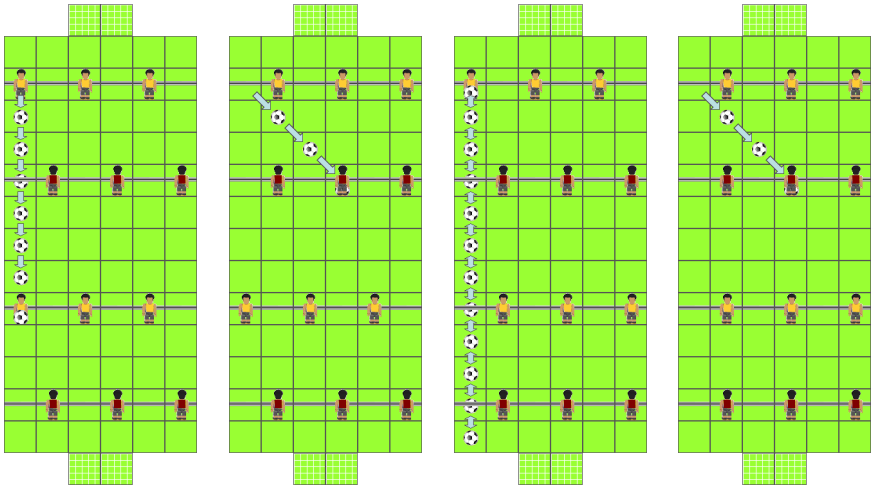
(a) $a_1^1 = (0,0)$ $a_2^2 = (1,0)$ (b) $a_2^1 = (1,0)$ $a_2^2 = (1,0)$ (c) $a_3^1 = (0,1)$ $a_2^2 = (1,0)$ (d) $a_4^1 = (1,1)$ $a_2^2 = (1,0)$

Fig. B.2. Foosman 1 Kicks the Ball



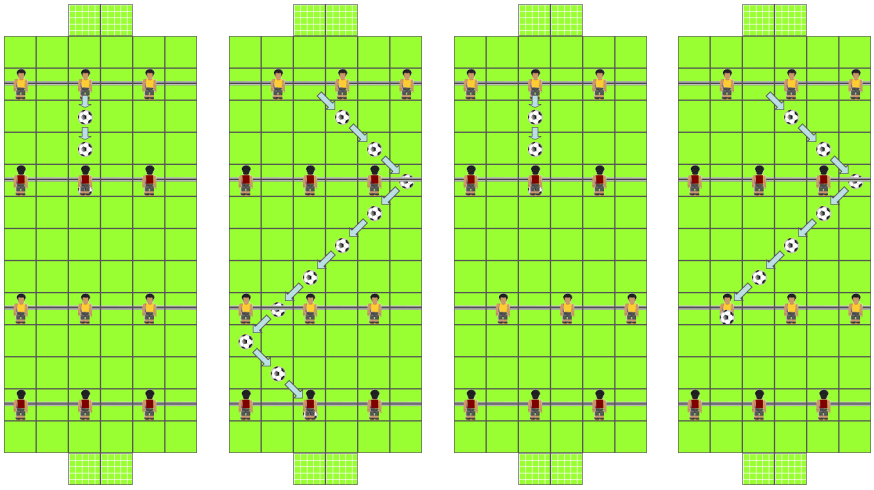
(a) $a_1^1 = (0,0)$ $a_3^2 = (0,1)$ (b) $a_2^1 = (1,0)$ $a_3^2 = (0,1)$ (c) $a_3^1 = (0,1)$ $a_3^2 = (0,1)$ (d) $a_4^1 = (1,1)$ $a_3^2 = (0,1)$

Fig. B.3. Foosman 1 Kicks the Ball



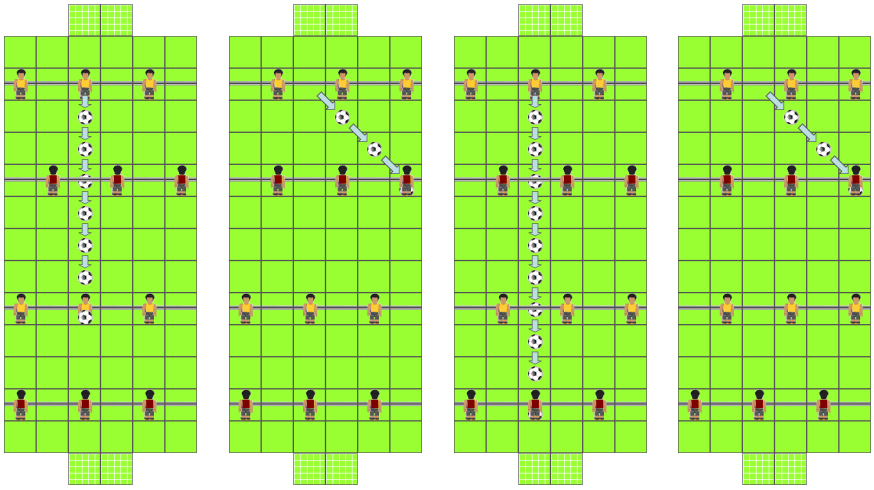
(a) $a_1^1 = (0,0)$ $a_4^2 = (1,1)$ (b) $a_2^1 = (1,0)$ $a_4^2 = (1,1)$ (c) $a_3^1 = (0,1)$ $a_4^2 = (1,1)$ (d) $a_4^1 = (1,1)$ $a_4^2 = (1,1)$

Fig. B.4. Foosman 1 Kicks the Ball



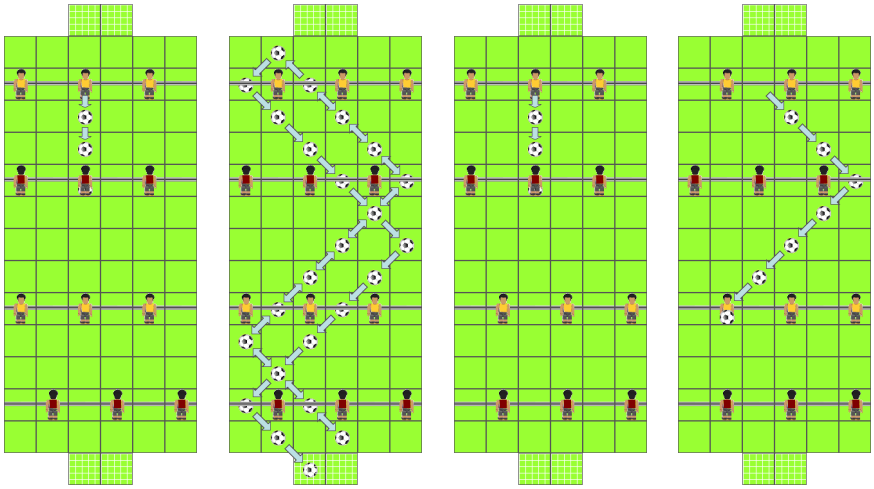
(a) $a_1^1 = (0,0)$ $a_1^2 = (0,0)$ (b) $a_2^1 = (1,0)$ $a_1^2 = (0,0)$ (c) $a_3^1 = (0,1)$ $a_1^2 = (0,0)$ (d) $a_4^1 = (1,1)$ $a_1^2 = (0,0)$

Fig. B.5. Foosman 2 Kicks the Ball



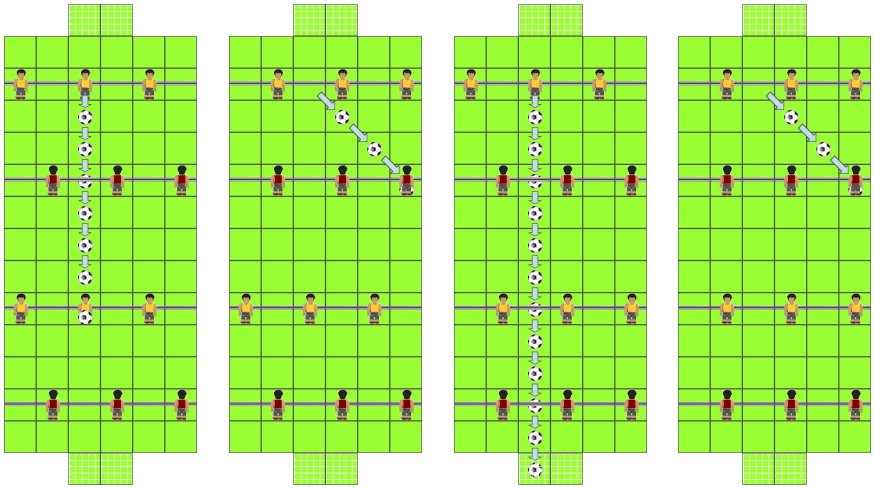
(a) $a_1^1 = (0,0)$ $a_2^2 = (1,0)$ (b) $a_2^1 = (1,0)$ $a_2^2 = (1,0)$ (c) $a_3^1 = (0,1)$ $a_2^2 = (1,0)$ (d) $a_4^1 = (1,1)$ $a_2^2 = (1,0)$

Fig. B.6. Foosman 2 Kicks the Ball



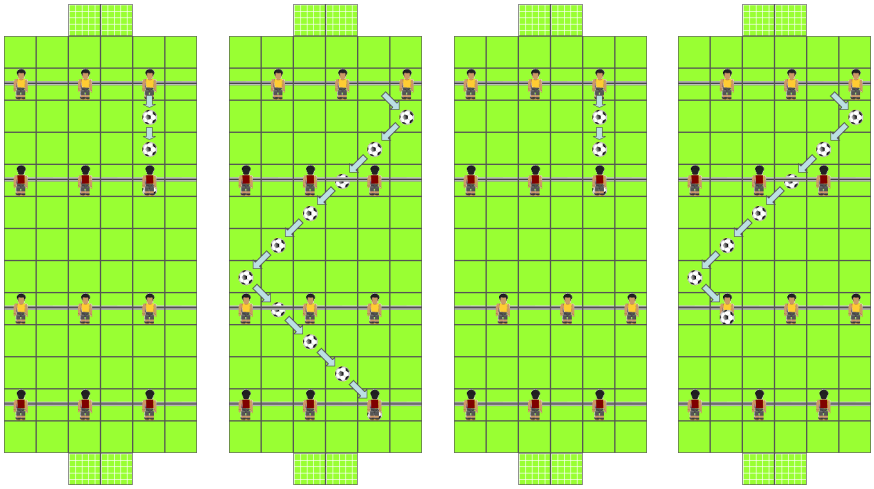
(a) $a_1^1 = (0,0)$ $a_3^2 = (0,1)$ (b) $a_2^1 = (1,0)$ $a_3^2 = (0,1)$ (c) $a_3^1 = (0,1)$ $a_3^2 = (0,1)$ (d) $a_4^1 = (1,1)$ $a_3^2 = (0,1)$

Fig. B.7. Foosman 2 Kicks the Ball



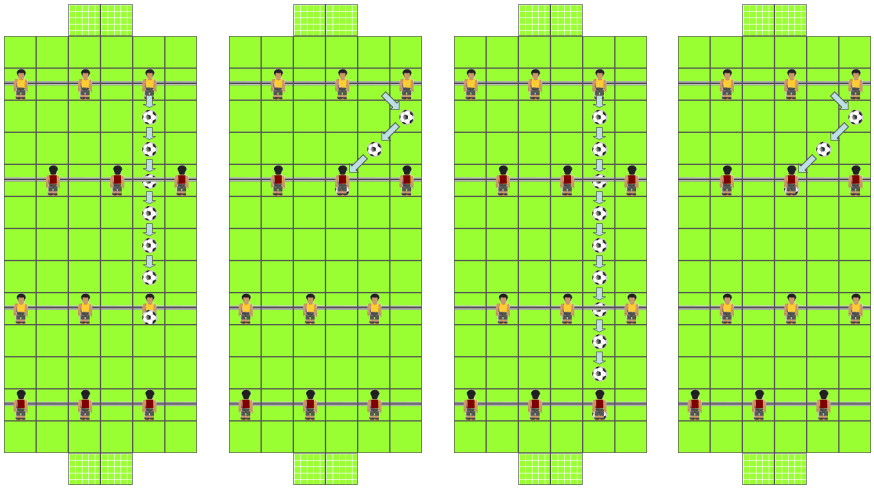
(a) $a_1^1 = (0,0)$ $a_4^2 = (1,1)$ (b) $a_2^1 = (1,0)$ $a_4^2 = (1,1)$ (c) $a_3^1 = (0,1)$ $a_4^2 = (1,1)$ (d) $a_4^1 = (1,1)$ $a_4^2 = (1,1)$

Fig. B.8. Foosman 2 Kicks the Ball



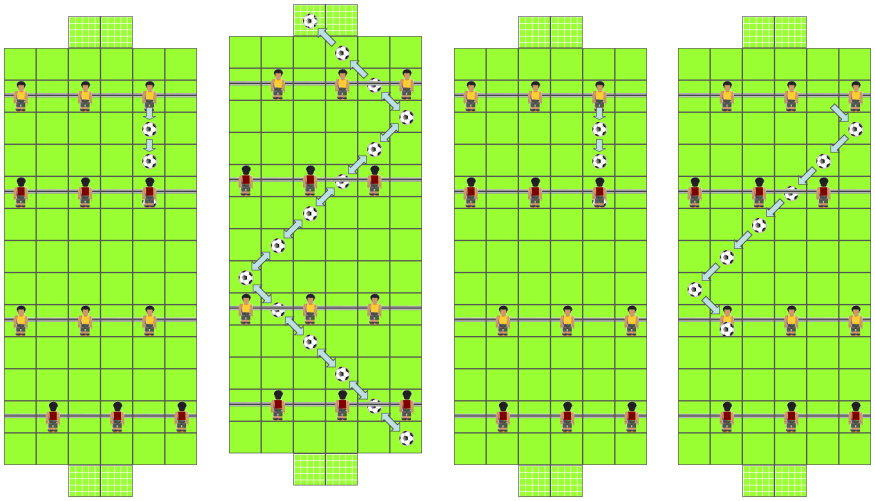
(a) $a_1^1 = (0,0)$ $a_1^2 = (0,0)$ (b) $a_2^1 = (1,0)$ $a_1^2 = (0,0)$ (c) $a_3^1 = (0,1)$ $a_1^2 = (0,0)$ (d) $a_4^1 = (1,1)$ $a_1^2 = (0,0)$

Fig. B.9. Foosman 3 Kicks the Ball



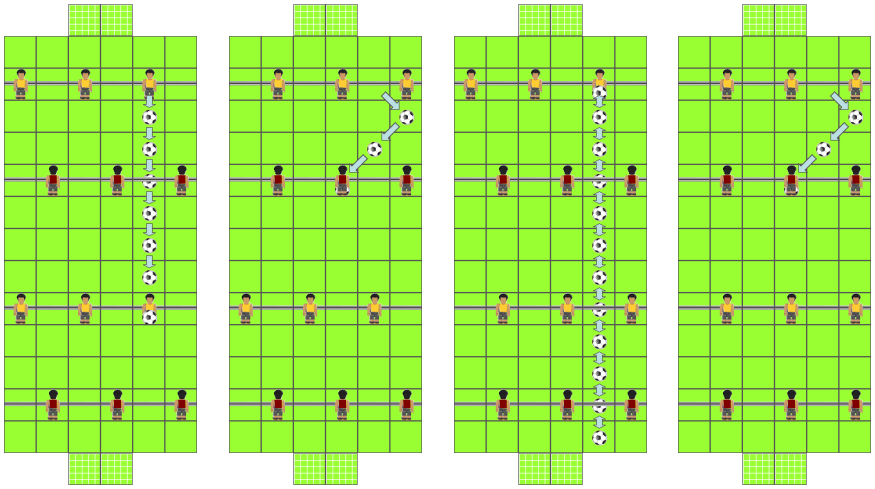
(a) $a_1^1 = (0,0)$ $a_2^2 = (1,0)$ (b) $a_2^1 = (1,0)$ $a_2^2 = (1,0)$ (c) $a_3^1 = (0,1)$ $a_2^2 = (1,0)$ (d) $a_4^1 = (1,1)$ $a_2^2 = (1,0)$

Fig. B.10. Foosman 3 Kicks the Ball



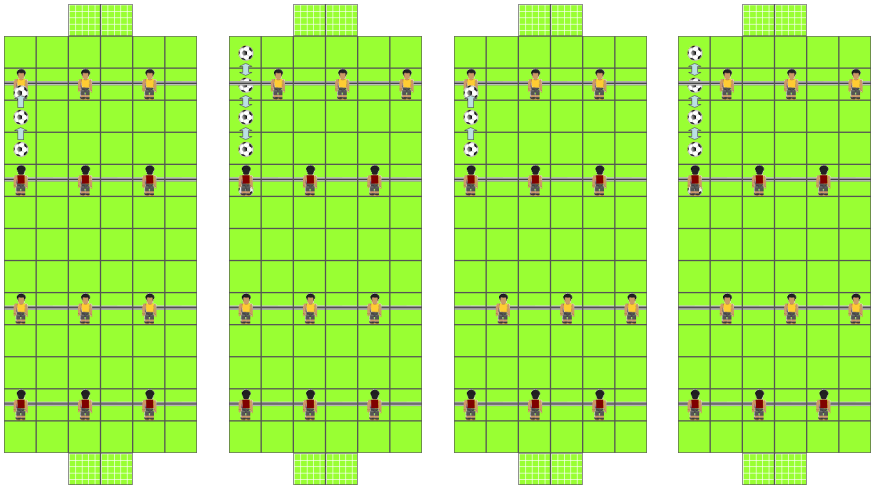
(a) $a_1^1 = (0,0)$ $a_3^2 = (0,1)$ (b) $a_2^1 = (1,0)$ $a_3^2 = (0,1)$ (c) $a_3^1 = (0,1)$ $a_3^2 = (0,1)$ (d) $a_4^1 = (1,1)$ $a_3^2 = (0,1)$

Fig. B.11. Foosman 3 Kicks the Ball.



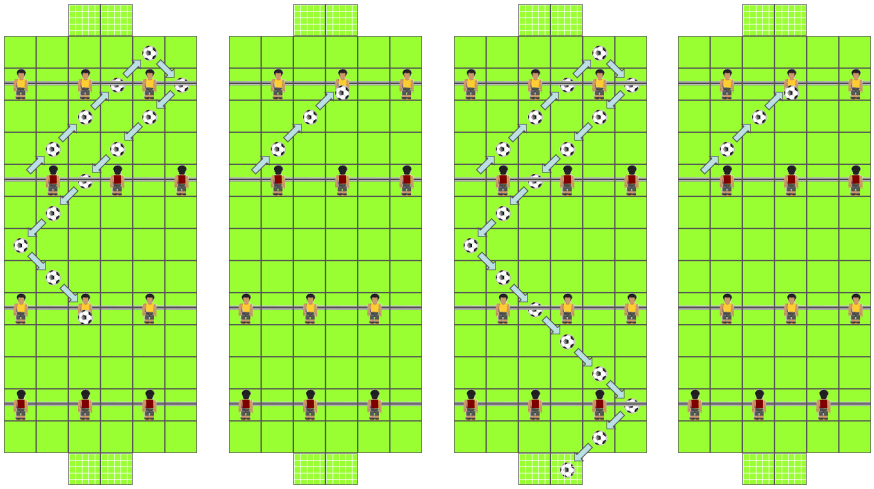
(a) $a_1^1 = (0,0)$ $a_4^2 = (1,1)$ (b) $a_2^1 = (1,0)$ $a_4^2 = (1,1)$ (c) $a_3^1 = (0,1)$ $a_4^2 = (1,1)$ (d) $a_4^1 = (1,1)$ $a_4^2 = (1,1)$

Fig. B.12. Foosman 3 Kicks the Ball



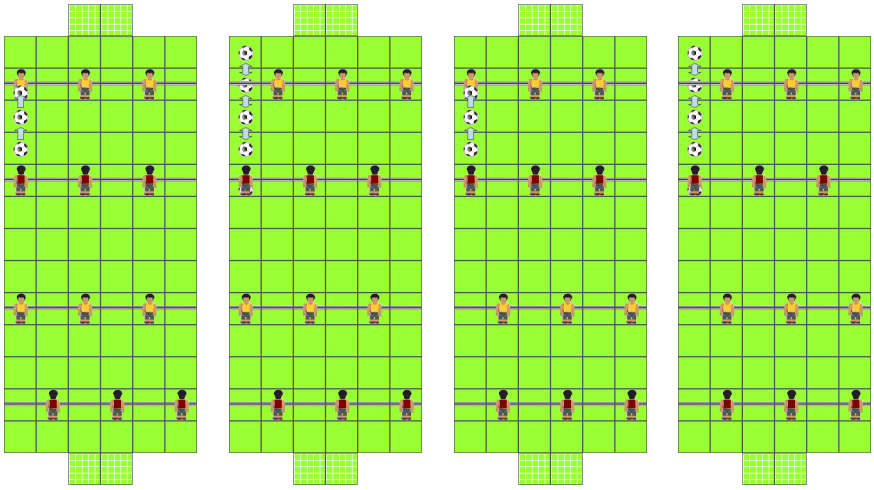
(a) $a_1^1 = (0,0)$ $a_1^2 = (0,0)$ (b) $a_2^1 = (1,0)$ $a_1^2 = (0,0)$ (c) $a_3^1 = (0,1)$ $a_1^2 = (0,0)$ (d) $a_4^1 = (1,1)$ $a_1^2 = (0,0)$

Fig. B.13. Foosman 4 Kicks the Ball



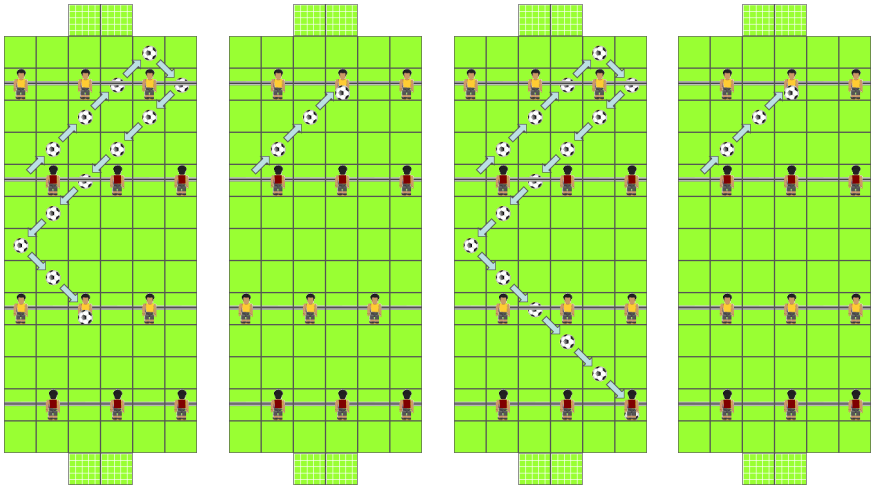
(a) $a_1^1 = (0,0)$ $a_2^2 = (1,0)$ (b) $a_2^1 = (1,0)$ $a_2^2 = (1,0)$ (c) $a_3^1 = (0,1)$ $a_2^2 = (1,0)$ (d) $a_4^1 = (1,1)$ $a_2^2 = (1,0)$

Fig. B.14. Foosman 4 Kicks the Ball



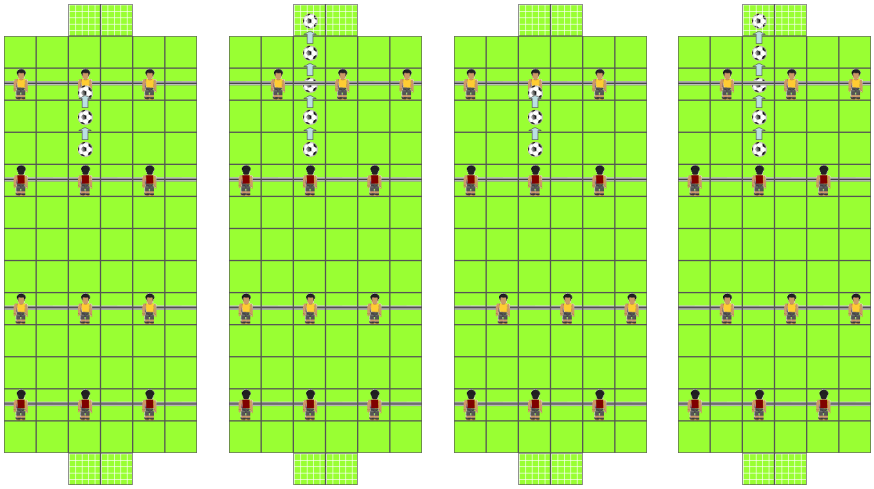
(a) $a_1^1 = (0,0)$ $a_3^2 = (0,1)$ (b) $a_2^1 = (1,0)$ $a_3^2 = (0,1)$ (c) $a_3^1 = (0,1)$ $a_3^2 = (0,1)$ (d) $a_4^1 = (1,1)$ $a_3^2 = (0,1)$

Fig. B.15. Foosman 4 Kicks the Ball



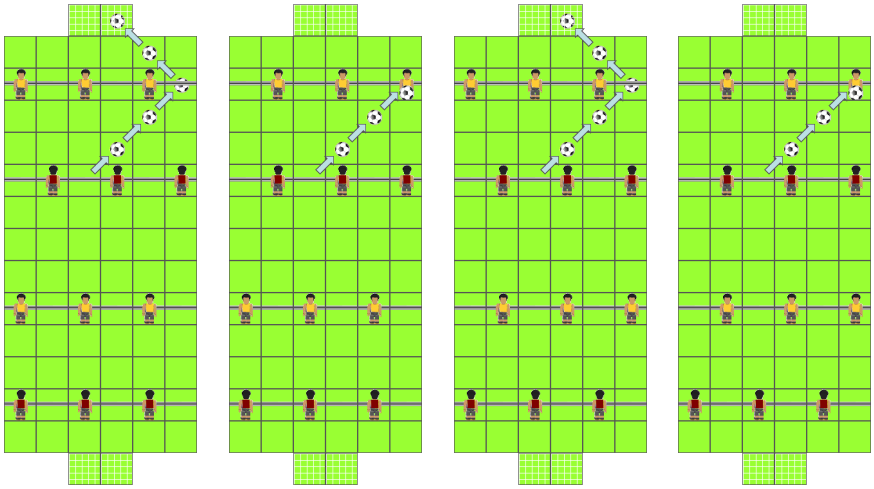
(a) $a_1^1 = (0,0)$ $a_4^2 = (1,1)$ (b) $a_2^1 = (1,0)$ $a_4^2 = (1,1)$ (c) $a_3^1 = (0,1)$ $a_4^2 = (1,1)$ (d) $a_4^1 = (1,1)$ $a_4^2 = (1,1)$

Fig. B.16. Foosman 4 Kicks the Ball



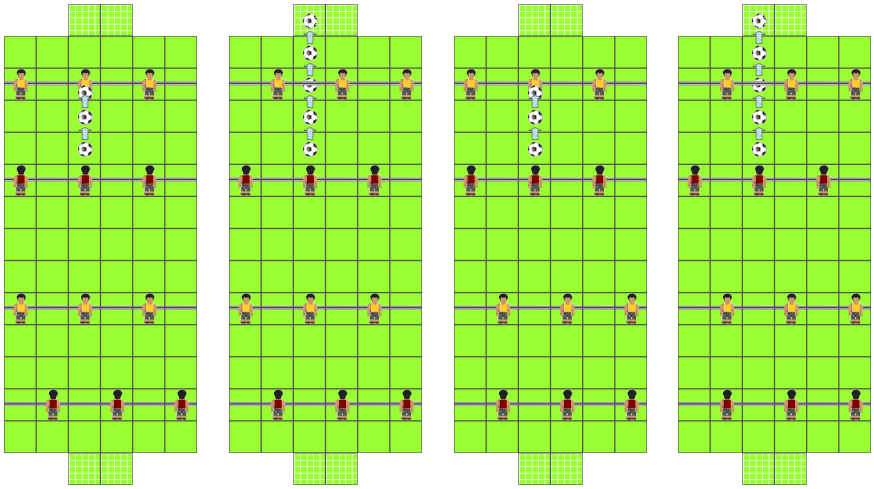
(a) $a_1^1 = (0,0)$ $a_1^2 = (0,0)$ (b) $a_2^1 = (1,0)$ $a_1^2 = (0,0)$ (c) $a_3^1 = (0,1)$ $a_1^2 = (0,0)$ (d) $a_4^1 = (1,1)$ $a_1^2 = (0,0)$

Fig. B.17. Foosman 5 Kicks the Ball



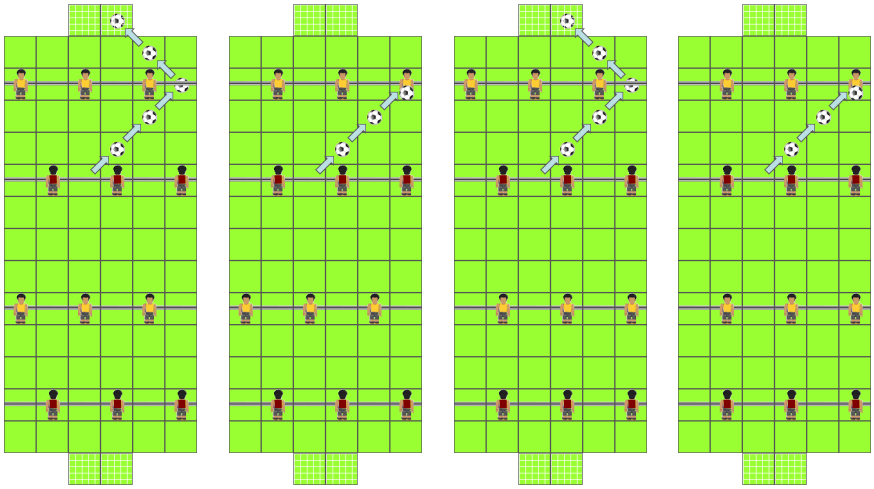
(a) $a_1^1 = (0,0)$ $a_2^2 = (1,0)$ (b) $a_2^1 = (1,0)$ $a_2^2 = (1,0)$ (c) $a_3^1 = (0,1)$ $a_2^2 = (1,0)$ (d) $a_4^1 = (1,1)$ $a_2^2 = (1,0)$

Fig. B.18. Foosman 5 Kicks the Ball



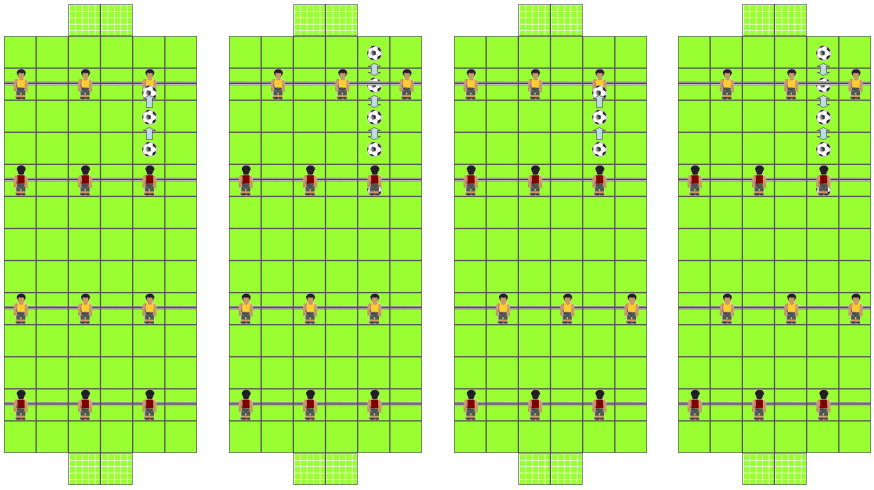
(a) $a_1^1 = (0,0)$ $a_3^2 = (0,1)$ (b) $a_2^1 = (1,0)$ $a_3^2 = (0,1)$ (c) $a_3^1 = (0,1)$ $a_3^2 = (0,1)$ (d) $a_4^1 = (1,1)$ $a_3^2 = (0,1)$

Fig. B.19. Foosman 5 Kicks the Ball



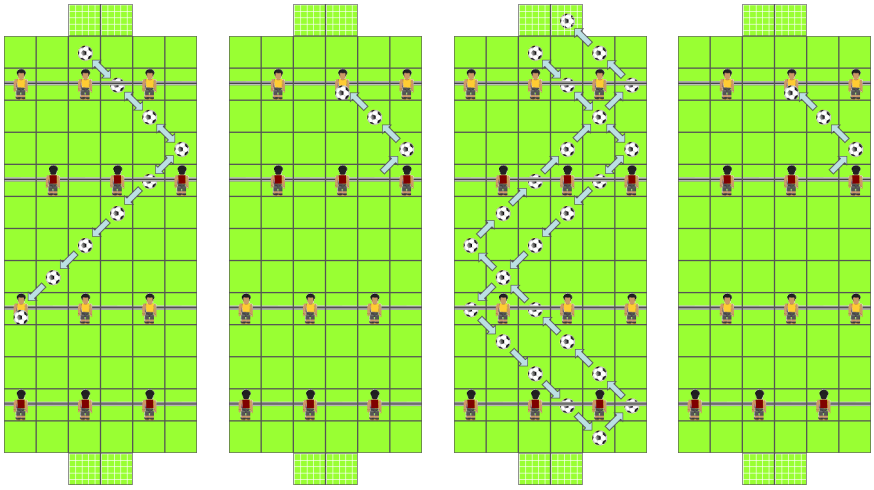
(a) $a_1^1 = (0,0)$ $a_4^2 = (1,1)$ (b) $a_2^1 = (1,0)$ $a_4^2 = (1,1)$ (c) $a_3^1 = (0,1)$ $a_4^2 = (1,1)$ (d) $a_4^1 = (1,1)$ $a_4^2 = (1,1)$

Fig. B.20. Foosman 5 Kicks the Ball



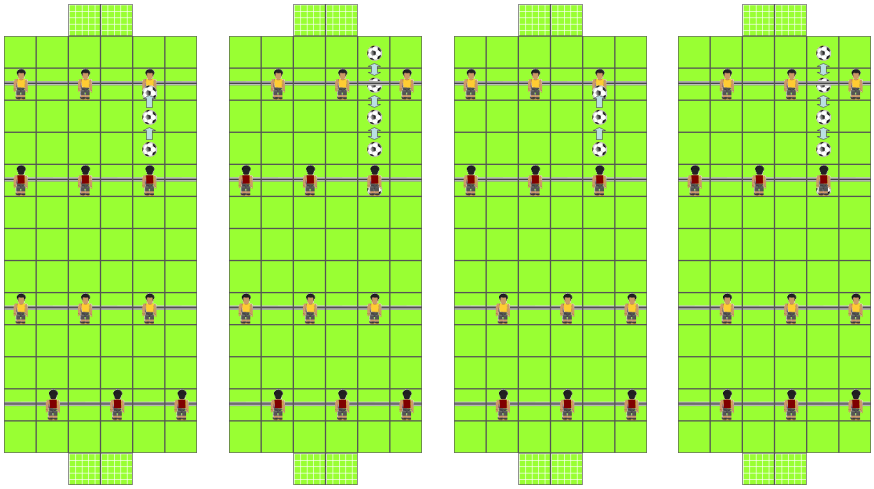
(a) $a_1^1 = (0,0)$, $a_1^2 = (0,0)$ (b) $a_2^1 = (1,0)$, $a_1^2 = (0,0)$ (c) $a_3^1 = (0,1)$, $a_1^2 = (0,0)$ (d) $a_4^1 = (1,1)$, $a_1^2 = (0,0)$

Fig. B.21. Foosman 6 Kicks the Ball



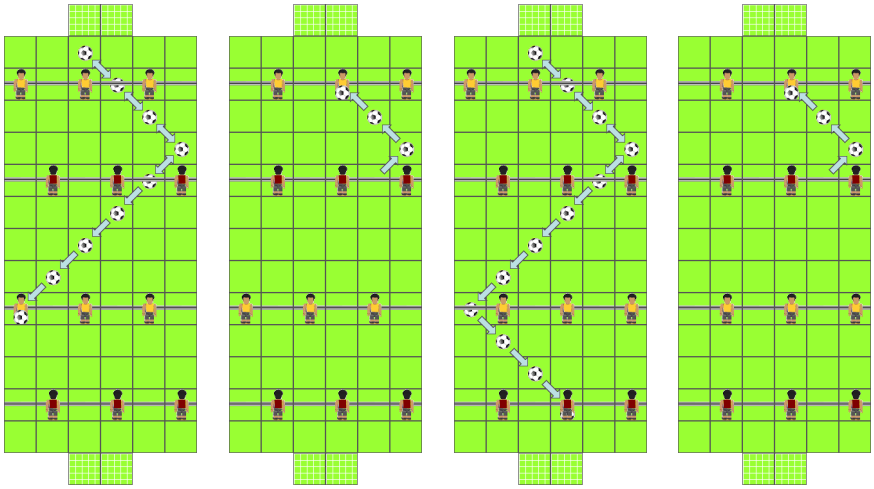
(a) $a_1^1 = (0,0)$, $a_2^2 = (1,0)$ (b) $a_2^1 = (1,0)$, $a_2^2 = (1,0)$ (c) $a_3^1 = (0,1)$, $a_2^2 = (1,0)$ (d) $a_4^1 = (1,1)$, $a_2^2 = (1,0)$

Fig. B.22. Foosman 6 Kicks the Ball



(a) $a_1^1 = (0,0)$ $a_3^2 = (0,1)$ (b) $a_2^1 = (1,0)$ $a_3^2 = (0,1)$ (c) $a_3^1 = (0,1)$ $a_3^2 = (0,1)$ (d) $a_4^1 = (1,1)$ $a_3^2 = (0,1)$

Fig. B.23. Foosman 6 Kicks the Ball



(a) $a_1^1 = (0,0)$ $a_4^2 = (1,1)$ (b) $a_2^1 = (1,0)$ $a_4^2 = (1,1)$ (c) $a_3^1 = (0,1)$ $a_4^2 = (1,1)$ (d) $a_4^1 = (1,1)$ $a_4^2 = (1,1)$

Fig. B.24. Foosman 6 Kicks the Ball

References

1. *3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004), 19-23 August 2004, New York, NY, USA*. IEEE Computer Society, 2004.
2. *Advances in Neural Information Processing Systems 17, Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada*, 2004.
3. Alfred V. Aho, John E. Hopcroft, and Jeffrey Ullman. *Data Structures and Algorithms*. Addison-Wesley Longman Publishing Co., Inc., 1983.
4. Natalia Akchurina. Learning by observation: Comparison of three methods of embedding mentor's knowledge in reinforcement learning algorithms. In Patrick Olivier and Christian Kray, editors, *AISB'07: Proceedings of the AISB Annual Convention*, pages 270–278. The Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2007.
5. Natalia Akchurina. Multi-agent reinforcement learning algorithm with variable optimistic-pessimistic criterion. In Malik Ghallab, Constantine D. Spyropoulos, Nikos Fakotakis, and Nikolaos M. Avouris, editors, *ECAI'08: Proceedings of the Eighteenth European Conference on Artificial Intelligence*, volume 178 of *Frontiers in Artificial Intelligence and Applications*, pages 433–437. IOS Press, 2008.
6. Natalia Akchurina. Optimistic-pessimistic Q-learning algorithm for multi-agent systems. In Ralph Bergmann, Gabriela Lindemann, Stefan Kirn, and Michal Pechoucek, editors, *MATES'08: Proceedings of the Sixth German Conference on Multiagent System Technologies*, volume 5244 of *Lecture Notes in Computer Science*, pages 13–24. Springer, 2008.
7. Natalia Akchurina. Computation of Nash equilibria in general-sum discounted stochastic games. Technical Report tr-ri-09-306, University of Paderborn, 2009.
8. Natalia Akchurina. In search of Nash equilibrium for multiagent reinforcement learning. In *Proceedings des gemeinsamen Workshops der Informatik-Graduiertenkollegs und Forschungskollegs (Dagstuhl 2009)*, pages 169–170, Dagstuhl, Germany, 2009. GITO.
9. Natalia Akchurina. Multiagent reinforcement learning: Algorithm converging to Nash equilibrium in general-sum discounted stochastic games. In Keith S. Decker, Jaime Simão Sichman, Carles Sierra, and Cristiano Castelfranchi, edi-

tors, *AAMAS'09: Proceedings of The Eighth International Conference on Autonomous Agents and Multiagent Systems*, volume 2, pages 725–732. IFAAMAS, 2009.

10. Natalia Akchurina and Hans Kleine Büning. Virtual markets: Q-learning sellers with simple state representation. In Vladimir Gorodetsky, Chengqi Zhang, Victor A. Skormin, and Longbing Cao, editors, *AIS-ADM'07: Proceedings of the Second International Workshop on Autonomous Intelligent Systems: Agents and Data Mining*, volume 4476 of *Lecture Notes in Computer Science*, pages 192–205. Springer, 2007.
11. Natalia Akchurina and Vadim Vagin. Generalized value partition problem: A rough set approach. *Journal of Computer and Systems Sciences International*, 43(2):223–238, 2004.
12. Natalia Akchurina and Vadim Vagin. Parallel preprocessing for classification problems in knowledge discovery systems. In Enn Tyugu and Takahira Yamaguchi, editors, *JCKBSE'06: Proceedings of the Seventh Joint Conference on Knowledge-Based Software Engineering*, volume 140 of *Frontiers in Artificial Intelligence and Applications*, pages 275–284. IOS Press, 2006.
13. Andrey A. Amosov, Yuliy A. Dubinsky, and Natalia V. Kopchenova. *Numerical Methods for Engineers*. Vyshaya shkola, Moscow, Russia, 1994.
14. Kenneth Arrow. Hurwiczs optimality criterion for decision making under ignorance. Technical Report 6, Stanford University, 1953.
15. Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-arm bandit problem. In *FOCS'95: Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 322–331, 1995.
16. Robert J. Aumann. Correlated equilibrium as an expression of Bayesian rationality. *Econometrica*, 55(1):1–18, 1987.
17. Bikramjit Banerjee and Jing Peng. Performance bounded reinforcement learning in strategic interactions. In Deborah L. McGuinness and George Ferguson, editors, *AAAI'04: Proceedings of the 19th National Conference on Artificial Intelligence*, pages 2–7. AAAI Press / The MIT Press, 2004.
18. Bikramjit Banerjee and Jing Peng. The role of reactivity in multiagent learning. In *AAMAS'04: Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems* [1], pages 538–545.
19. Martino Bardi, T.E.S. Raghavan, and Thiruvenkatachari Parthasarathy. *Stochastic and Differential Games: Theory and Numerical Methods (Annals of the International Society of Dynamic Games)*. Birkhäuser, 1999.
20. Richard Bellman. *Dynamic Programming*. Dover Publications, 2003.
21. Jos Luis Bermdez. *Decision Theory and Rationality*. Oxford University Press, 2009.
22. Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume I. Athena Scientific, 2nd edition, 2001.
23. Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control*, volume II. Athena Scientific, 2nd edition, 2001.
24. Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
25. Tilman Borgers and Rajiv Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1):1–14, 1997.

26. Michael H. Bowling. Convergence problems of general-sum multiagent reinforcement learning. In *ICML'00: Proceedings of the 17th International Conference on Machine Learning*, pages 89–94, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
27. Michael H. Bowling. Convergence and no-regret in multiagent learning. In *NIPS'04: Proceedings of the 18th Annual Conference on Neural Information Processing Systems*, pages 209–216, 2004.
28. Michael H. Bowling and Manuela M. Veloso. Multiagent learning using a variable learning rate. *Artificial Intelligence*, 136(2):215–250, 2002.
29. Martin Braun. *Differential Equations and Their Applications: An Introduction to Applied Mathematics*. Springer-Verlag, 1993.
30. Dagobert L. Brito. A dynamic model of an armaments race. *International Economic Review*, 13(2):359–375, 1972.
31. Carla E. Brodley and Andrea Pohoreckyj Danyluk, editors. *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001*. Morgan Kaufmann, 2001.
32. George W. Brown. Iterative solution of games by fictitious play. In Tjalling C. Koopmans, editor, *Activity Analysis of Production and Allocation*, pages 374–376. Wiley, 1951.
33. John C. Butcher. *Numerical Methods for Ordinary Differential Equations*. John Wiley & Sons, Inc., 2003.
34. Richard H. Byrd, Nicholas I. M. Gould, Jorge Nocedal, and Richard A. Waltz. An algorithm for nonlinear optimization using linear programming and equality constrained subproblems. *Mathematical Programming*, 100(1):27–48, 2004.
35. Antonio Cabrales and Joel Sobel. On the limit points of discrete selection dynamics. *Journal of Economic Theory*, 57(2):407–419, 1992.
36. Julyan H. E. Cartwright and Oreste Piro. The dynamics of Runge-Kutta methods. *International Journal of Bifurcation and Chaos*, 2:427–449, 1992.
37. Roberto Cellini and Luca Lambertini. Advertising in a differential oligopoly game. *Journal of Optimization Theory and Applications*, 116(1):61–81, 2003.
38. Yu-Han Chang and Leslie Pack Kaelbling. Playing is believing: The role of beliefs in multi-agent learning. In Thomas G. Dietterich, Suzanna Becker, and Zoubin Ghahramani, editors, *NIPS'01: Proceedings of the 14th Conference on Neural Information Processing Systems*, pages 1483–1490. The MIT Press, 2001.
39. Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI'98/IAAI'98: Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, pages 746–752, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence.
40. Vincent Conitzer and Tuomas Sandholm. AWESOME: A general multiagent learning algorithm that converges in self-play and learns a best response against stationary opponents. *Machine Learning*, 67(1-2):23–43, 2007.
41. Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2001.
42. Peter Cramton, Yoav Shoham, and Richard Steinberg. *Combinatorial Auctions*. The MIT Press, 2006.

43. Robert H. Crites and Andrew G. Barto. Improving elevator performance using reinforcement learning. In David S. Touretzky, Michael Mozer, and Michael E. Hasselmo, editors, *NIPS'95: Proceedings of the 8th Conference on Neural Information Processing Systems*, pages 1017–1023. The MIT Press, 1995.
44. Robert Harry Crites. *Large-scale dynamic optimization using teams of reinforcement learning agents.* PhD thesis, University of Massachusetts Amherst, Massachusetts, USA, 1996.
45. János Csirik, Michael Littman, Satinder P. Singh, and Peter Stone. FAucS: An FCC spectrum auction simulator for autonomous bidding agents. In *WELCOM'01: Proceedings of the Second International Workshop on Electronic Commerce*, pages 139–151, London, UK, 2001. Springer-Verlag.
46. Richard Dawkins. *The Selfish Gene.* Oxford University Press, Oxford, 1976.
47. Peter Dayan. The convergence of TD(λ) for general λ. *Machine Learning*, 8:341–362, 1992.
48. Engelbert J. Dockner, Steffen Jørgensen, Ngo Van Long, and Gerhard Sorger. *Differential Games in Economics and Management Science.* Cambridge University Press, 2000.
49. Arne Stolbjerg Drud. CONOPT—a large-scale GRG code. *INFORMS Journal on Computing*, 6(2):207–216, 1994.
50. Martin Dufwenberg, Jonas Björnerstedt, Peter Norman, and Jörgen Weibull. Evolutionary selection dynamics and irrational survivors. Technical report, Mimeo, Department of Economics, Stockholm University, 1996.
51. Aryeh Dvoretzky. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 39–55, 1956.
52. Tom Fawcett and Nina Mishra, editors. *ICML'03: Proceedings of the Twentieth International Conference on Machine Learning.* AAAI Press, 2003.
53. Gustav Feichtinger, editor. *Optimal Control Theory and Economic Analysis 2: Second Viennese Workshop on Economic Applications of Control Theory.* Elsevier Science Pub., 1985.
54. Jerzy Filar. *Algorithms for solving some undiscounted stochastic games.* PhD thesis, University of Illinois at Chicago, Illinois, USA, 1980.
55. Jerzy Filar, T. A. Schultz, Frank Thuijsman, and Koos Vrieze. Nonlinear programming and stationary equilibria in stochastic games. *Mathematical Programming*, 50(2):227–237, 1991.
56. Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes.* Springer-Verlag, 1996.
57. Arlington M. Fink. Equilibrium in a stochastic n-person game. *Journal of Science of the Hiroshima University*, 28(1):89–93, 1964.
58. Yoav Freund and Robert E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.
59. Drew Fudenberg and David K. Levine. The theory of learning in games. Levine's Working Paper Archive 624, University of California, Los Angeles, 1996.
60. Drew Fudenberg and Jean Tirole. *Game Theory.* The MIT Press, 1991.
61. Philip E. Gill, Walter Murray, and Michael A. Saunders. SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM Review*, 47(1):99–131, 2005.

62. Amy R. Greenwald. The 2002 trading agent competition: An overview of agent strategies. *AI Magazine*, 24(1):83–91, 2003.
63. Amy R. Greenwald and Keith Hall. Correlated Q-learning. In Fawcett and Mishra [52], pages 242–249.
64. Jack K. Hale. *Ordinary Differential Equations*. Robert E. Krieger Rublishing Company, New York, 1980.
65. John C. Harsanyi and Reinhard Selten. *A General Theory of Equilibrium Selection in Games*, volume I. The MIT Press, 1988.
66. Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.
67. Philip Hartman. *Ordinary Differential Equations*. Birkhäuser, 2nd edition, 1982.
68. Daniel Hennes, Karl Tuyls, and Matthias Rauterberg. State-coupled replicator dynamics. In *AAMAS'09: Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems*, pages 789–796, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems.
69. P. Jean-Jacques Herings and Ronald J. A. P. Peeters. Stationary equilibria in stochastic games: structure, selection, and computation. *Journal of Economic Theory*, 118(1):32–60, 2004.
70. P. Jean-Jacques Herings and Ronald J.A.P. Peeters. A differentiable homotopy to compute Nash equilibria of n-person games. *Economic Theory*, 18(1):159–185, 2001.
71. P. Jean-Jacques Herings and Ronald J.A.P. Peeters. Equilibrium selection in stochastic games. Game Theory and Information 0205002, EconWPA, 2001.
72. Ernest R. Hilgard and Gordon H. Bower. *Theories of learning*. Appleton-Century-Crofts, 3rd edition, 1966.
73. Pieter Jan't Hoen, Karl Tuyls, Liviu Panait, Sean Luke, and Johannes A. La Poutré. An overview of cooperative and competitive multiagent learning. In Karl Tuyls, Pieter Jan't Hoen, Katja Verbeeck, and Sandip Sen, editors, *LAMAS'05: Proceedings of the First International Workshop on Learning and Adaption in Multi-Agent Systems*, volume 3898 of *Lecture Notes in Computer Science*, pages 1–46. Springer, 2005.
74. Josef Hofbauer and Karl Sigmund. *The Theory of Evolution and Dynamical Systems*. Cambridge University Press, 1988.
75. Josef Hofbauer and Karl Sigmund. *Evolutionary Games and Population Dynamics*. Cambridge University Press, 1998.
76. Feng-Hsiung Hsu. *Behind Deep Blue: building the computer that defeated the world chess champion*. Princeton University Press, 2002.
77. http://gamut.stanford.edu/.
78. http://indy.cs.concordia.ca/auto/.
79. http://mathworld.wolfram.com/.
80. http://plato.stanford.edu/entries/game evolutionary/.
81. http://www.gams.com/dd/docs/solvers/conopt.pdf.
82. http://www.gams.com/dd/docs/solvers/knitro.pdf.
83. http://www.gams.com/dd/docs/solvers/minos.pdf.
84. http://www.gams.com/dd/docs/solvers/snopt.pdf.
85. http://www.gams.com/solvers/index.htm.
86. http://www.optagro.ru/work/newsanalitics/13819.

87. http://www.sfu.ca/geog351fall03/groups webpages/gp8/references.html.
88. Junling Hu and Michael P. Wellman. Multiagent reinforcement learning: Theoretical framework and an algorithm. In Shavlik [132], pages 242–250.
89. Junling Hu and Michael P. Wellman. Online learning about other agents in a dynamic multiagent system. In *AGENTS'98: Proceedings of the 2nd International Conference on Autonomous Agents*, pages 239–246, New York, NY, USA, 1998. ACM.
90. Junling Hu and Michael P. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of Machine Learning Research*, 4:1039–1069, 2003.
91. Michael D. Intriligator and Dagobert L. Brito. Can arms races lead to the outbreak of war? *Journal of Conflict Resolution*, 28(1):63–84, 1984.
92. Rufus Isaacs. *Differential Games: A Mathematical Theory with Applications to Warfare and Pursuit, Control and Optimization*. John Wiley & Sons, 1965.
93. Amir Jafari, Amy R. Greenwald, David Gondek, and Gunes Ercal. On no-regret learning, fictitious play, and Nash equilibrium. In Brodley and Danyluk [31], pages 226–233.
94. Steffen Jørgensen. A survey of some differential games in advertising. *Journal of Economic Dynamics and Control*, 4(1):341–369, 1982.
95. Steffen Jørgensen and Georges Zaccour. Equilibrium pricing and advertising strategies in a marketing channel. *Journal of Optimization Theory and Applications*, 102(1):111–125, 1999.
96. Steffen Jørgensen and Georges Zaccour. *Differential Games in Marketing*. Springer, 2004.
97. Jeffrey O. Kephart and Amy R. Greenwald. Shopbot economics. *Autonomous Agents and Multi-Agent Systems*, 5(3):255–287, 2002.
98. George E. Kimball. Some industrial applications of military operations research methods. *Operations Research*, 5(2):201–204, 1957.
99. Hiroaki Kitano. Research program of RoboCup. *Applied Artificial Intelligence*, 12(2-3):117–125, 1998.
100. Hiroaki Kitano, Minoru Asada, Yasuo Kuniyoshi, Itsuki Noda, Eiichi Osawa, and Hitoshi Matsubara. RoboCup: A challenge problem for AI. *AI Magazine*, 18(1):73–85, 1997.
101. Paul Klemperer. *Auctions: Theory and Practice (The Toulouse Lectures in Economics)*. Princeton University Press, 2004.
102. John D. Lambert. *Numerical methods for ordinary differential systems: the initial value problem*. John Wiley & Sons, Inc., 1991.
103. Michael Littman. Markov games as a framework for multi-agent reinforcement learning. In *ICML'94: Proceedings of the 11the International Conference on Machine Learning*, pages 157–163, 1994.
104. Michael Littman. Friend-or-Foe Q-learning in general-sum games. In Brodley and Danyluk [31], pages 322–328.
105. Michael Littman and Peter Stone. Leading best-response strategies in repeated games. In *IJCAI'01: Seventeenth Annual International Joint Conference on Artificial Intelligence, Workshop on Economic Agents, Models, and Mechanisms*, 2001.
106. Michael Littman and Peter Stone. A polynomial-time Nash equilibrium algorithm for repeated games. *Decision Support Systems*, 39(1):55–66, 2005.

107. Michael Littman and Csaba Szepesvári. A generalized reinforcement-learning model: Convergence and applications. In *ICML'96: Proceedings of the 13th International Conference on Machine Learning*, pages 310–318, 1996.
108. Pan-Tai Liu. *Dynamic Optimization and Mathematical Economics*. Springer, 1980.
109. John Maynard Smith. *Evolution and the theory of games*. Cambridge University Press, 1982.
110. Alexander Mehlmann. *Applied Differential Games*. Plenum Press, London, 1988.
111. Thomas Mitchell. *Machine Learning*. McGraw Hill Higher Education, 1997.
112. Remi Munos. A convergent reinforcement learning algorithm in the continuous case based on a finite difference method. In *IJCAI'97: Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, pages 826–831, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.
113. Roger B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, 1997.
114. John H. Nachbar. "Evolutionary" selection dynamics in games: convergence and limit properties. *International Journal of Game Theory*, 19(1):59–89, 1990.
115. John Nash. Non-cooperative games. *The Annals of Mathematics*, 54(2):286–295, 1951.
116. Monty Newborn. *Deep Blue*. Springer, 2002.
117. Abraham Neyman and Sylvain Sorin. *Stochastic Games and Applications*. Springer, 2003.
118. Eugene Nudelman, Jennifer Wortman, Yoav Shoham, and Kevin Leyton-Brown. Run the GAMUT: A comprehensive approach to evaluating game-theoretic algorithms. In *AAMAS'04: Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems* [1], pages 880–887.
119. Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. The MIT Press, 1994.
120. Simon Parsons. *Game Theory and Decision Theory in Agent-Based Systems*. Kluwer Academic Publishers, 2002.
121. Thiruvenkatachari Parthasarathy and T. E. S. Raghavan. An orderfield property for stochastic games when one player controls transition probabilities. *Journal of Optimization Theory and Applications*, 33(3):375–392, 1981.
122. Rob Powers and Yoav Shoham. New criteria and a new algorithm for learning in multi-agent systems. In *NIPS* [2], pages 1089–1096.
123. Rob Powers and Yoav Shoham. Learning against opponents with bounded memory. In Leslie Pack Kaelbling and Alessandro Saffiotti, editors, *IJCAI'05: Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 817–822. Professional Book Center, 2005.
124. Jette Randløv and Preben Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In Shavlik [132], pages 463–471.
125. Anatol Rapoport, Melvin J. Guger, and David G. Gordon. *The 2 × 2 Game*. The University of Michigan Press, 1976.
126. Herbert Robbins and Sutton Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22(3):400–407, 1951.
127. Gary Schneider. *Electronic Commerce*. Course Technology, 8th edition, 2008.
128. Peter Schuster and Karl Sigmund. Replicator dynamics. *Journal of Theoretical Biology*, 100(3):533–538, 1983.

129. Peter Schuster, Karl Sigmund, Josef Hofbauer, and Robert Wolff. Selfregulation of behavior in animal societies. *Biological Cybernetics*, 40(1):9–15, 1981.
130. Frances H. Seligson, Debra A. Krummel, and Joan L. Apgar. Patterns of chocolate consumption. *American Journal of Clinical Nutrition*, 60(6):1060–1064, 1994.
131. Sandip Sen, Mahendra Sekaran, and John Hale. Learning to coordinate without sharing information. In *AAAI'94: Proceedings of the 12th national conference on Artificial intelligence*, volume I, pages 426–431, Menlo Park, CA, USA, 1994. American Association for Artificial Intelligence.
132. Jude W. Shavlik, editor. *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconson, USA, July 24-27, 1998*. Morgan Kaufmann, 1998.
133. Yoav Shoham, Rob Powers, and Trond Grenager. Multi-agent reinforcement learning: a critical survey. Technical report, Stanford University, 2003.
134. Marwan A. Simaan and Takashi Takayama. Game theory applied to dynamic duopoly problems with production constraints. *Automatica*, 14(2):161–166, 1978.
135. Satinder P. Singh, Michael J. Kearns, and Yishay Mansour. Nash convergence of gradient dynamics in general-sum games. In Craig Boutilier and Moisés Goldszmidt, editors, *UAI'00: Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 541–548. Morgan Kaufmann, 2000.
136. William D. Smart and Leslie Pack Kaelbling. Practical reinforcement learning in continuous spaces. In *ICML'00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 903–910, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
137. Matthew J. Sobel. Myopic solutions of Markov decision processes and stochastic games. *Operations Research*, 29(5):995–1009, 1981.
138. A. Michael Spence. Investment strategy and growth in a new market. *The Bell Journal of Economics*, 10(1):1–19, 1979.
139. Peter Stone and Amy R. Greenwald. The first international trading agent competition: Autonomous bidding agents. *Electronic Commerce Research*, 5(2):229–265, 2005.
140. Gilbert Strang. *Introduction to Linear Algebra*. Wellesley Cambridge, 3rd edition, 2003.
141. Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. The MIT Press, 1998.
142. Csaba Szepesvári and Michael Littman. Generalized Markov decision processes: Dynamic-programming and reinforcement-learning algorithms. Technical Report CS-96-11, Brown University, 1996.
143. Michael Tabor. *Chaos and integrability in nonlinear dynamics: an introduction*. Wiley, 1989.
144. Kalyan T. Talluri and Garrett J. van Ryzin. *Theory and Practice of Revenue Management: 68 (International Series in Operations Research and Management Science)*. Springer, 3rd edition, 2005.
145. Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *ICML'93: Proceedings of the 10th International Conference on Machine Learning*, pages 330–337. Morgan Kaufmann, 1993.

146. Peter D. Taylor. Evolutionarily stable strategies with two types of player. *Journal of Applied Probability*, 16(1):76–83, 1979.
147. Peter D. Taylor and Leo B. Jonker. Evolutionary stable strategies and game dynamics. *Mathematical Biosciences*, 40(1-2):145–156, 1978.
148. Gerald Tesauro. Practical issues in temporal difference learning. *Machine Learning*, 8:257–277, 1992.
149. Gerald Tesauro. Temporal difference learning and TD-Gammon. *Communications of the ACM*, 38(3):58–68, 1995.
150. Gerald Tesauro. Extending Q-learning to general adaptive multi-agent systems. In Sebastian Thrun, Lawrence K. Saul, and Bernhard Schölkopf, editors, *NIPS'03: Proceedings of the 16th Annual Conference on Neural Information Processing Systems*. The MIT Press, 2003.
151. Edward L. Thorndike. *Animal Intelligence*. Hafner, Darien, Conn, 1911.
152. Efraim Turban, Jae Kyu Lee, Dave King, Judy McKay, and Peter Marshall. *Electronic Commerce 2008*. Prentice Hall, 2007.
153. Karl Tuyls, Dries Heytens, Ann Nowé, and Bernard Manderick. Extended replicator dynamics as a key to reinforcement learning in multi-agent systems. In Nada Lavrac, Dragan Gamberger, Ljupco Todorovski, and Hendrik Blockeel, editors, *ECML'03: Proceedings of the 14th European Conference on Machine Learning*, volume 2837 of *Lecture Notes in Computer Science*, pages 421–431. Springer, 2003.
154. Karl Tuyls, Tom Lenaerts, Katja Verbeeck, Sam Maes, and Bernard Manderick. Towards a relation between learning agents and evolutionary dynamics. In *BNAIC'02: Proceedings of the 14th Belgian-Dutch Conference on Artificial Intelligence*, pages 21–22, 2002.
155. Karl Tuyls, Katja Verbeeck, and Tom Lenaerts. A selection-mutation model for Q-learning in multi-agent systems. In *AAMAS'03: Proceedings of the 2nd International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 693–700. ACM, 2003.
156. William T. B. Uther and Manuela M. Veloso. Tree based discretization for continuous state space reinforcement learning. In *AAAI'98/IAAI'98: Proceedings of the Fifteenth National/Tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence*, pages 769–774, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence.
157. Vadim Vagin and Natalia Akchurina. New techniques for handling missing data and feature extraction for knowledge discovery. In Vadim Stefanuk and Kenji Kaijiri, editors, *JCKBSE'04: Proceedings of the Sixth Joint Conference on Knowledge-Based Software Engineering*, volume 108 of *Frontiers in Artificial Intelligence and Applications*, pages 169–176. IOS Press, 2004.
158. Emmanuil M. Vaisbord. *Introduction to Multi-Player Differential Games and Their Applications (Studies in Cybernetics, Vol 15)*. Routledge, 1988.
159. Ferdinand Verhulst. *Nonlinear differential equations and dynamical systems*. Springer-Verlag, 1990.
160. M. L. Vidale and H. B. Wolfe. An operations-research study of sales response to advertising. *Operations Research*, 5(3):370–381, 1957.
161. Koos Vrieze. *Stochastic Games with Finite State and Action Spaces*. Mathematisch Centrum, Amsterdam, 1987.
162. Abraham Wald. *Statistical decision functions*. John Wiley & Sons, 1950.

163. Chris J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge, England, 1989.
164. Christopher J. C. H. Watkins and Peter Dayan. Technical note Q-learning. *Machine Learning*, 8:279–292, 1992.
165. Robert J. Weber. Making more from less: Strategic demand reduction in the FCC spectrum auctions. *Journal of Economics & Management Strategy*, 6(3):529–548, 1997.
166. Jörgen W. Weibull. *Evolutionary Game Theory*. The MIT Press, 1996.
167. Michael Weinberg and Jeffrey S. Rosenschein. Best-response multiagent learning in non-stationary environments. In *AAMAS'04: Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems* [1], pages 506–513.
168. Michael P. Wellman, Amy Greenwald, Peter Stone, and Peter R. Wurman. The 2001 trading agent competition. In *AAAI/IAAI'02: Proceedings of the Eighteenth National Conference on Artificial Intelligence*, pages 935–941, Menlo Park, CA, USA, 2002. American Association for Artificial Intelligence.
169. David W. K. Yeung and Leon A. Petrosyan. *Cooperative Stochastic Differential Games*. Springer, 2009.
170. Willard I. Zangwill and C. B. Garcia. *Pathways to Solutions, Fixed Points, and Equilibria (Prentice-Hall Series in Computational Mathematics)*. Prentice Hall, 1981.
171. Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In Fawcett and Mishra [52], pages 928–936.