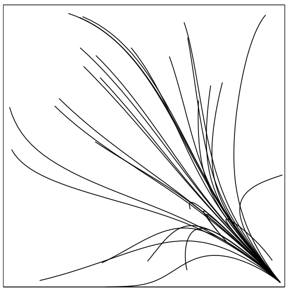
Games and Learning in Auctions

Michael Kaisers

Games and Learning in Auctions

An evolutionary game theory analysis



Michael Kaisers

Thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science in Artificial Intelligence

Maastricht University
Faculty of Humanities and Sciences

August 22, 2008

The unit square on the title page is the result of an experiment of learning in the Prisoners' Dilemma. It shows several average policy trajectories for Q-learners, that start at random points in the policy space and apply a decreasing temperature. Section 6.3.1 elaborates on this result.

About the author

Michael Kaisers graduated from Maastricht University with a B.Sc. in *Knowledge Engineering* in 2007. He earned the honor *suma cum laude* and abbreviated the three years program to two years.
This thesis concludes his M.Sc. in *Artificial Intelligence* at Maastricht University, upon which he is appointed a TopTalent grant for a Ph.D. study at Eindhoven University of Technology, starting September 2008.

To my grandfather

Acknowledgments

> *"We are like dwarfs on the shoulders of giants, so that we can see more than they, and things at a greater distance, not by virtue of any sharpness of sight on our part, or any physical distinction, but because we are carried high and raised up by their giant size."*
>
> Bernard of Chartres, 12^{th} century

This thesis is another stage on the journey that started in the final phase of my Bachelor program. It was a journey of learning to which many affiliates, collaborators and friends contributed. It has prepared my future ways and coined me on an intellectual and personal level.

First, I would like to express my gratitude to my colleagues and collaborators Jinzhong Niu, Daniel Hennes and Steve Phelps. Their support and helpful discussions have contributed to the findings of this thesis, and some of their tools have facilitated quicker progress.

I am also indebted to my supervisors Frank Thuijsman, Simon Parsons and Karl Tuyls. I am looking forward to continue this exciting journey with you as a Ph.D. student. Frank, from our exciting collaboration I learned that science can and should be told through vivid stories. Simon, you have introduced me to the field of auctions and you are a role model of time management and personal atmosphere for me. Thank you for the exciting stay in New York City and the guidance through concise comments that usually hit the very core of the issue. Karl, you enabled much of my development, guiding the Bachelor's and Master's thesis, organizing a stay abroad and supporting opportunities like my PhD position, that I acquired through the TopTalent competition. Besides your technical input, you provide motivation and an extraordinary working atmosphere. Thank you!

Special thanks go to my parents, grandparents, my sister and the rest of my family who are always there when I need them. Thank you for keeping me grounded and stable. Furthermore, I owe much of my endurance to my dearest Rena. Rena, having you lets me regain strength in hard times.

Eventually, thanks to all those who are not listed above but who made remarks on this thesis or inspired me in any other way. *Thank you all.*

Abstract

Auctions are pervasive in today's society and provide a variety of markets, ranging from government-to-business auctions for licenses to consumer-to-consumer online auctions. The success of trading strategies in auctions is highly dependent on the present competitors, hence traders are forced to adapt to the competition to maintain a high level of performance. This adaptation may be modeled by reinforcement learning algorithms, which have a proven relation to evolutionary game theory.

This thesis facilitates a strategic choice between a set of predefined trading strategies. It is based on previous work, which suggests to capture the payoff of trading strategies in a heuristic payoff table. A new methodology to approximate heuristic payoff tables by normal form games is introduced, and it is evaluated by a case study of a 6-agent clearing house auction. Learning models of exploration and exploitation, that link to selection and mutation in an evolutionary perspective, are subsequently applied to compare three common automated trading strategies.

The information loss in the normal form approximation is shown to be reasonably small, such that the concise normal form representation can be used to derive strategic decisions in auctions. Furthermore, the learning model shows that learners with exploration may converge to different strategies than learners of pure exploitation.

The devised methodology establishes a bridge between empirical data in heuristic payoff tables and the means from classical game theory. It might therefore become the basis for a more general framework to analyze strategic interactions in complex multi-agent systems.

Keywords: Auctions, Multi-agent learning, Evolutionary game theory

Contents

List of figures

List of tables

Chapter 1

Introduction

Auctions are deployed in a variety of real markets to foster highly efficient trading. They range from consumer-to-consumer markets like eBay and business-to-business stock exchanges to government-to-business auctions for mineral rights or government licenses for the telecommunications spectrum [16, 22]. Furthermore, auction mechanisms have been transferred successfully to solve other resource allocation problems, e.g. in the domain of efficient internet traffic routing [26]. Despite this diversity, even single markets of enormous scale may have profound impact on society, e.g. the New York Stock Exchange reported a trading volume of $11,060$ billion USD in 2000 [8]. Hence, auctions are pervasive in today's society which motivates strong research interests.

The Santa Fe trading competition has pitted a variety of trading strategies against each other. It has demonstrated that the performance of trading strategies within auctions depends largely on the present competitors [27]. A strategy called KAPLAN won the Santa Fe tournament by withholding from the auction until the price has almost converged or the time is running out and then placing bids that are just high enough to outperform its competitors. Despite its high performance against many other strategies, its advantage dwindles the more agents adopt it [39].

A generally *best* strategy does not exist. Hence, traders cannot rely on finding one perfect trading strategy and apply it at all times. They rather need to adapt to changing circumstances and apply a trading strategy that is a best reply to the current competition. In other words, they need to learn which strategy gives them the highest payoff under the current conditions.

It is realistic to assume that the participants in auctions do not know which trading strategies are used by other traders. Unless the competitors' trading strategies can be identified from the bid and shout patterns that the traders exhibit, the learning process needs to be solely based on the payoff a trader perceives by applying a trading strategy in an auction. The trading agents' learning processes may therefore be modeled by reinforcement learning, for which a variety of algorithms has been devised [18, 40, 32]. The average learning behavior of these algorithms can be described by the replicator dynamics, a mathematical model from the field of evolutionary game theory. While reinforcement learning studies learning on an individual level, evolutionary game theory gives an intuition on the high level learning dynamics. It allows to analyze under which circumstances agents switch or should switch their strategies. Furthermore, the behavior of agents who learned for a sufficiently long time can be predicted. The relative strength of trading strategies can then be estimated from the probability of being used in the long run. In this way, the learning model facilitates a comparison of a set of trading strategies in auctions.

1.1 Related work

A variety of authors have studied auctions due to their economical relevance [12, 24, 38, 39]. Early laboratory experiments conducted by Smith have shown that even "about three active [human] buyers and sellers" converge to equilibrium prices by which high efficiency is attained [31]. This disproofs the previous assumption that large scale markets are required for efficient trading. For simplicity, the case study presented in this thesis is restricted to an auction with 6 traders.

Vickrey pioneered in the approach of modeling auctions as games of incomplete information. He showed that the second price sealed-bid auctions are *strategy proof*, i.e. rational traders reveal their true valuation [38]. On the one hand, it is not clear how to model more sophisticated auctions in order to apply a game theoretical analysis which computes Nash equilibria [39]. On the other hand, mechanisms that can be modeled may yield many equilibria. The contemporary theory does not determine which of them will eventually be played or how to coordinate on one [17]. Theoretical solution concepts like the Nash equilibrium may also be inapplicable because real traders may not be rational or fail to compute the equilibrium adequately. As a result, research increasingly turns to empirical simulation to evaluate properties of trading strategies and auctions [33].

The empirical approach provides powerful means to analyze, predict and control the behavior of self-interested agents [6]. The study of simulation allows insights into realistic and complex auction problems that would render many theoretical approaches intractable. A variety of trading strategies has been formalized in the form of computer algorithms [4, 7, 11]. The available algorithms allow to study the learning behavior by simulating auctions with traders who choose between a set of available trading strategies and learn which one to apply. In contrast to experiments with human subjects, this methodology is cost-effective, more reliable (stochastic influences can be eliminated by high numbers of iterations), deterministic (using seeds for pseudo-random numbers) and thereby reproducible. For these benefits, it is used in this thesis to analyze strategic behavior in auctions.

The payoff information that is gathered from auction simulation needs to be captured for further analysis. When n agents decide between k strategies, there are k^n possible combined decisions. This number grows exponentially in the number of agents and storing all payoffs is intractable for realistic values of n. Fortunately, it is reasonable to assume that the trading agents are interchangeable, i.e. all trading strategies are equally good for each agent. This allows to exploit the symmetric property of auctions. It is not important who uses which strategy but merely how many opponents use which strategies. Walsh proposed to capture the average payoff to each type of trading strategy, given the different compositions of the competition, in a heuristic payoff table [39]. Several authors have adopted these tables as a basis for the analysis of auctions [15, 25] and it will also be the starting point for the analysis in this thesis.

A heuristic payoff table captures the payoffs of trading strategies in auctions, but it requires further mathematical analysis to predict strategic behavior of trading agents. Evolutionary game theory assumes that a population of traders will adopt successful trading strategies and drop less successful ones. It has been used to compare market mechanisms [25] and to automatically devise new trading strategies [24]. A theoretical link between reinforcement learning and evolutionary game theory has been established for Cross learning, which maps to a selection model [2], and has been extended to Q-learning, which maps to a selection-mutation model [35]. This implies that means from evolutionary game theory can be applied to study multi-agent learning [34]. As human learning features phases of exploration and exploitation, a selection-mutation model is used in this thesis to model strategic behavior of trading agents.

1.2 Problem definition and research questions

The payoffs defined by the heuristic payoff tables need an interpretation. Previous research has used evolutionary selection models to model adaptive interactions of trading agents in auctions. Although the toolbox of game theory is much richer than that, other means can not easily be applied to heuristic payoff tables. These methods are either too computationally expensive or not defined for this representation of payoffs. This leads to the following two research questions:

- How can the payoff information captured by heuristic payoff tables be made more accessible for a deeper analysis by means of game theory?
- How does adding mutation to the replicator dynamics influence the strategic behavior in auctions?

Many means from game theory are best studied for the normal form representation of payoffs, suggesting the question whether heuristic payoff tables can be approximated by normal form games. This thesis presents two methods to find such an approximation: Using linear programming (minimizing the maximum absolute deviation of the approximation) and a standard least mean squares algorithm (minimizing the mean squared error). In order to answer the problem statement, the following refined research questions are answered by performing an empirical study:

- Calculate the heuristic payoff table for an auction with 6 agents who may choose between several trading strategies.
- Which approximations result from the proposed methodology?
- How do these approximations differ quantitatively, in the payoffs they define, and qualitatively, in the predictions that result from the learning model?
- How does mutation in the evolutionary analysis alter the learning dynamics?
- What happens when learning is applied with decreasing exploration?

Results of the performed empirical study show a reasonably small error in the approximation which justifies using the approximation for strategic considerations and an intuitive grasp of the game in auctions.

1.3 Outline

The remainder of this thesis is divided into two parts: The first part lays the theoretical foundation for the analysis of games based on heuristic payoff tables while the second part describes an application to the auction domain.

The theoretical part presents the required background for the empirical study of the research questions. It comprises game theoretical background in Chapter 2, concepts from reinforcement learning in Chapter 3 and the proposed methodology to approximate heuristic payoff tables by normal form games, which is given in Chapter 4.

The empirical part first presents the relevant auction type and trading strategies in Chapter 5. These are used in the experiments, for which setup and results are described in Chapter 6. The results are interpreted and discussed in Chapter 7, which also concludes this thesis.

Part I

Theory of games and learning

Chapter 2

Game theoretical background

Game theory introduces games as formal models to study strategic interactions. It emerged from the investigation of strategic conflicts, e.g. in economics and war, and was founded by John von Neumann and Oskar Morgenstern who published the first important book in this discipline in 1944 [1, 37]. Since then, the theory has been enriched by many contributors. Among them John Nash, who introduced what is now known as the Nash Equilibrium (NE) in 1951 [19] and John Maynard Smith, who contributed the notion of evolutionary stable strategies in 1973 [30].

This chapter first takes the perspective of classical game theory and introduces the normal form game, the solution concept of the Nash equilibrium, optimal strategies and the matrix game value. Selected examples are given and used to illustrate the concepts. Subsequently, the point of view is shifted to evolutionary game theory. The replicator dynamics and evolutionary stable strategies are explained and an example is given by the game Rock-Paper-Scissors. The chapter is concluded by a description of heuristic payoff tables, which are compressed representations for symmetric games. The definitions of this chapter are based on [9, 10, 14, 39, 41].

2.1 Classical game theory

Classical game theory is the mathematical study of strategic interactions between rational agents. There are two predominant representations for games, the extensive and the normal form. The extensive form describes how the game is played over time in a game tree. The outcome is captured in a single value for each player, the utility that denotes the preference of that player for that outcome. In this thesis, the terms reward or payoff are used as synonyms of utility or preference. Any extensive form game can be transformed into a normal form game and the considered stateless games are more naturally modeled in normal form. Therefore, only normal form games are considered in this thesis.

2.1.1 Normal form games

In normal form games, the players are assumed to choose their actions, or *pure strategies*, simultaneously and independently. Let $I = \{1, \ldots, n\}$ be the set of n players, where n is a positive integer. For each player i, let S_i denote the set of available pure strategies. For notational convenience, every players' pure strategies will be labeled with positive integers, i.e. $S_i = \{1, 2, \ldots, k_i\}$ for some integer $k_i \geq 2$. Let s_i take on the value of a particular pure strategy j for player i. A pure strategy profile is an n-tuple $s = (s_1, \ldots, s_n)$ that associates one pure strategy with each player. Furthermore, let $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ denote the same profile without the strategy of player i, so that (s_i, s_{-i}) forms a complete profile of strategies. The *pure strategy space* is the cartesian product of the players' pure strategy sets $S = \times_i S_i$.

Let $u_i : S \mapsto \Re$ denote the payoff function of player i, i.e. for any strategy profile $s \in S$ and player $i \in I$, let $u_i(s)$ be the associated payoff to player i. The combined payoff function $u : S \mapsto R^n$ assigns to each pure strategy profile s the full payoff vector $u(s) = (u_1(s), \ldots, u_n(s))$.

An n-player normal form game G may be summarized as the tuple $G = \langle I, S, u \rangle$, where I is the set of players, S is the pure strategy space and u is the combined payoff function. Two-agent games are often labeled with the number of actions for each player, e.g. a 2 x 2 game refers to a two-player game with two actions for both players.

Policy

As normal form games are stateless, the behavior of each player can be described by a probability vector π_i, that assigns a probability to each pure strategy. This probability vector is also called policy or mixed strategy. Let $\pi_{i,j}$ denote the probability of player i to play the pure strategy j.

$$\pi_i : S_i \longrightarrow [0, 1] \text{ such that } \sum_{j \in S_i} \pi_{i,j} = 1$$

Let $\pi = (\pi_1, \ldots, \pi_n)$ denote the mixed strategy profile. Furthermore, let $\pi_{-i} = (\pi_1, \ldots, \pi_{i-1}, \pi_{i+1}, \ldots, \pi_n)$ denote the same profile without player i's policy.

Expected payoff

Let $v_i(\pi)$ denote the expected payoff for playing policy π_i against the set of opponents' mixed strategies π_{-i}. It can be computed from the sum over the utilities of all possible pure strategy profiles, multiplied by their probability:

$$v_i(\pi) = E(u_i|\pi) = \sum_{s \in S} u_i(s) \prod_{m \in I} \pi_{m,s_m}$$

2.1.2 Solution concepts

The solution concepts of game theory prescribe the behavior of rational agents and provide a more specific characterization of normal form games.

Best response

The best response is the set of policies that have the maximal possible reward given all other players' policies. Due to rationality, all players are assumed to pick the best action available to them. A mixed strategy π is a best response of player i if there is no other mixed strategy π' that would lead to a higher reward for this player, given that all other players' strategies π_{-i} remain the same.

$$BR(\pi_{-i}) = \pi_i \text{ iff } \forall \pi'_i : v_i(\pi_i|\pi_{-i}) \geq v_i(\pi'_i|\pi_{-i})$$

Nash equilibrium

A Nash equilibrium is a strategy profile for which no player can improve his payoff by changing his policy while the other players keep their policies fixed. It is a tuple of policies $\pi^* = (\pi^*_1, \ldots, \pi^*_n)$ such that no player has an incentive for unilateral deviation, that is every strategy π^*_i is a best response to π^*_{-i}:

$$\pi^*_i = \arg\max_{\pi_i} v_i(\pi_i|\pi^*_{-i})$$

Nash equilibria are the primary concept to derive rational behavior in competitive games. For cooperative games, Pareto optimality is of primary interest.

Pareto optimality

A strategy profile π Pareto dominates π' if and only if all players obtain at least the same reward and at least one player receives a strictly higher reward when π is played.

$$\begin{aligned} &\pi \textit{ Pareto dominates } \pi' \\ &\quad \text{iff } \forall i \exists j : v_i(\pi) \geq v_i(\pi') \wedge v_j(\pi) > v_j(\pi') \end{aligned}$$

A strategy profile π is Pareto optimal if it is not Pareto dominated.

Optimal strategies and the matrix game value

An optimal strategy π^+ assures a certain payoff against any possible opponent. For any other policy than π^+, he may encounter an opponent that gives him a lower expected payoff. In other words, an optimal strategy is a best reply to a malicious opponent, that tries to minimize one's payoff:

$$\pi^+_i = \arg\max_{\pi_i} \min_{\pi_{-i}} v(\pi_i|\pi_{-i})$$

The value of a matrix game is defined as the payoff that the optimal strategy guarantees:

$$value = \max_{\pi_i} \min_{\pi_{-i}} v(\pi_i | \pi_{-i})$$

The optimal strategy is not generally a best reply against the opponents' policies but rather guarantees a best worst case. Therefore, an agent could even obtain more payoff than with the optimal strategy if he would know the actual current mix of strategies he faces.

2.1.3 Examples

Three representative examples of 2 x 2 games are discussed in this section. Based on the general representation given in Figure 2.1, 2 x 2 games can be divided into the following three subclasses [35]:

Subclass 1

If $(a_{11} - a_{21})(a_{12} - a_{22}) > 0$ or $(b_{11} - b_{21})(b_{12} - b_{22}) > 0$ there exists at least one dominant strategy and therefore only one pure equilibrium. The only exception: Let player i have a dominant strategy s_i and the other player j obtain $u(s_j | s_i) = x \ \forall s_j$, then there are infinitely many equilibria where player j mixes arbitrarily between his actions.

Subclass 2

If $(a_{11} - a_{21})(a_{12} - a_{22}) < 0$, $(b_{11} - b_{21})(b_{12} - b_{22}) < 0$ and $(a_{11} - a_{21})(b_{11} - b_{12}) > 0$ there are two pure and one mixed equilibrium.

Subclass 3

If $(a_{11} - a_{21})(a_{12} - a_{22}) < 0$, $(b_{11} - b_{21})(b_{12} - b_{22}) < 0$ and $(a_{11} - a_{21})(b_{11} - b_{12}) < 0$ there is just one mixed equilibrium.

Next, one example of each subclass will be discussed.

	$Left$	$Right$
Top	a_{11}, b_{11}	a_{12}, b_{12}
$Bottom$	a_{21}, b_{21}	a_{22}, b_{22}

Figure 2.1: General payoff bi-matrix (A, B) for two-agent two-action games, where A and B define the payoff to player 1 and 2 respectively. The first player chooses a row, the second player chooses a column.

	D	C
D	$3,3$	$0,5$
C	$5,0$	$1,1^*$

Figure 2.2: Payoff matrices for the Prisoners' Dilemma (*Deny* or *Confess*). The less the agents like the outcome, the lower the payoff.

The Prisoners' Dilemma

Two criminals are interrogated for a crime they committed together. The police keeps them in separate rooms such that they have no means of communication. Both are offered the same choice; they may either confess the crime or deny testimony. If both criminals deny, they will be charged for illegal possession of weapons and go to jail for a short time. If they confess while their partner denies, they are promised to go free while their partner has to serve a long sentence. However, if both confess, they will serve a mediocre sentence together. What will the criminals do assuming they are rational?

The Prisoners' Dilemma is a symmetric game with one Nash equilibrium (C, C). This example demonstrates that not every Nash equilibrium is Pareto optimal. In particular, (C, C) with utilities $(1, 1)$ is Pareto dominated by (D, D) with utilities $(3, 3)$, which is not a Nash equilibrium.

Battle of Sexes

A couple decided to go out together at night. However, they forgot to agree whether they go to the football match or watch a play at the theater. They have no means of communication before the event and need to decide independently where to spend their evening. The man prefers to meet at the stadium while the woman would prefer the theater, but both will only enjoy their evening if they meet their partner.

	F	T
F	$2,1^*$	$0,0$
T	$0,0$	$1,2^*$

Figure 2.3: Payoff matrices for Battle of Sexes (*Football* or *Theater*).

Battle of Sexes yields two pure equilibria at (B, B) with payoffs $(2, 1)$ and (S, S) with payoffs $(1, 2)$ and one mixed equilibrium where player 1 mixes between the actions $(\frac{2}{3}, \frac{1}{3})$ and player two mixes $(\frac{1}{3}, \frac{2}{3})$ which leads to expected payoffs $(\frac{2}{3}, \frac{2}{3})$. All three Nash equilibria are Pareto optimal.

Matching Pennies

	H	T
H	$1, -1$	$-1, \;\; 1$
T	$-1, \;\; 1$	$1, -1$

Figure 2.4: Payoff matrices for Matching Pennies (*Head* or *Tail*).

Matching Pennies originates from a gambling game. Two players simultaneously reveal a coin each, either showing head or tail. If they reveal the same side of the coin the first player gets both coins, otherwise the second player wins.

In the mixed Nash equilibrium of Matching Pennies both players mix both actions equally and obtain expected rewards $(0, 0)$ which is Pareto optimal.

2.2 Evolutionary game theory

Classical game theory assumes rationality. This implies, that each player is assumed to be absolutely self interested, capable and willing to consider all possible outcomes of the game and to select a strategy that maximizes his expected payoff. One of the main criticisms against classical game theory is the surrealism of that assumption, because this hyper-rationality does not always apply, especially not to humans.

Evolutionary game theory takes a rather descriptive perspective, replacing hyper-rationality from classical game theory by the concept of natural selection from biology [29]. The two central concepts of evolutionary game theory are the replicator dynamics and evolutionary stable strategies. The replicator dynamics presented in the next section describe the evolutionary change in the population. They are a set of differential equations that are derived from biological operators such as selection, mutation and cross-over. The evolutionary stable strategy describe the resulting asymptotic behavior of this population. For a detailed discussion, we refer the interested reader to [13, 14].

2.2.1 Replicator dynamics

A population comprises a set of individuals, where the species that an individual can belong to represent the pure strategies. The utility function can be interpreted as the Darwinian fitness of each species. The distribution of the individuals on the different strategies can be described by a probability vector that is equivalent to a policy. Hence, there is a second view on the evolutionary process: The population may also represent competing strategies within the mind of one agent, who learns which one to apply.

The evolutionary pressure by natural selection can be modeled by the replicator equations. They assume this population to evolve such that successful strategies with higher payoffs grow while less successful ones decay.

Single-population replicator dynamics

Let the individuals of the population be distributed over the pure strategies according to the probability vector π_1. As there is only one population, the index 1 from $\pi_{1,j}$ will be dropped for the sake of clarity, i.e. $\pi = (\pi_1, \dots, \pi_k)$, where π_j denotes the probability of strategy j to be played. The single-population replicator dynamics assume that two individuals are randomly drawn from the same population. These two individuals then play a game, which determines their fitness. Let $S_1 = S_2 = \{1, \dots, k\}$ be the available strategies and $S = S_1 \times S_2$ denote the joint strategy space. Similar to the expected payoff for a two-player game, the expected payoff for policy π against population $\bar{\pi}$ is computed as $v(\pi|\bar{\pi}) = \sum_{s \in S} u(s)\pi_{s_1}\bar{\pi}_{s_2}$. The evolution under natural selection can be modeled by the following system of equations, where e_j denotes the j'th unit vector:

$$\frac{d\pi_j}{dt} = \pi_j \Big[\underbrace{v(e_j|\pi)}_{\text{fitness of j}} - \underbrace{v(\pi|\pi)}_{\text{average fitness}} \Big] \tag{2.1}$$

When the payoff function u is given in form of a matrix A, i.e. $v(\pi|\bar{\pi}) = \pi A \bar{\pi}^T$, this equation simplifies to:

$$\frac{d\pi_j}{dt} = \pi_j \left[e_j A \pi^T - \pi A \pi^T \right] \tag{2.2}$$

Hofbauer and Sigmund have extended the selection model to account for mutation [14]. Let Q be the mutation matrix where Q_{jh} denotes the probability of an agent of species j to mutate to species h, i.e. he switches his pure strategy from j to h. The dynamics can be computed as:

$$\frac{d\pi_j}{dt} = \pi_j \left[v(e_j|\pi) - v(\pi|\pi) \right] + \underbrace{\pi Q e_j^T}_{\text{incoming}} - \underbrace{e_j Q \pi^T}_{\text{leaving}} \tag{2.3}$$

For any symmetric matrix Q, i.e. $Q = Q^T$, we have $Qe_j = (e_j Q)^T$ and the mutation terms cancel out. Equation (2.3) reduces to the pure selection model defined by Equation (2.1).

While the mutation defined by matrix Q describes imitation learning, the selection-mutation model derived in [35] describes the average learning behavior of Q-learning. Let α denote the learning rate and let τ denote the temperature parameter of the Q-learning algorithm. The temperature in the replicator equations balances selection and mutation, mapping to exploitation and exploration in the learning algorithm:

$$\frac{d\pi_j}{dt} = \alpha\pi_j \left[\tau^{-1} \underbrace{\left[v(e_j|\pi) - v(\pi|\pi)\right]}_{\text{selection}} + \underbrace{\pi \log \pi^T - \log \pi_j}_{\text{mutation}}\right] \tag{2.4}$$

This selection-mutation model of evolution is used in this thesis to model learning with exploration, as it allows to tune the amount of exploration with a single temperature parameter. In particular, it is used in Section 6.3.2 to describe strategic behavior in auctions.

Multi-population replicator dynamics

The replicator dynamics can also model several independent learning processes. Therefore, they need to be extended to a multi-population model, where each population can be regarded as a policy of an agent. The multi-population replicator dynamics assumes that two individuals, i.e. pure strategies, are randomly drawn from two different populations. These two individuals then play a game, which determines their fitness. Let π_1 and π_2 denote the distribution of individuals over the pure strategies in population 1 and 2. Here, the expected payoff $v_i(\pi|\bar{\pi}) = \sum_{s \in S} u_i(s)\pi_{s_1}\bar{\pi}_{s_2}$ may differ for the two populations. The extension of the selection model to two populations reads:

$$\begin{aligned} \frac{d\pi_{1,j}}{dt} &= \pi_{1,j}\left[v_1(e_j|\pi_2) - v_1(\pi_1|\pi_2)\right] \\ \frac{d\pi_{2,j}}{dt} &= \pi_{2,j}\left[v_2(e_j|\pi_1) - v_2(\pi_2|\pi_1)\right] \end{aligned} \tag{2.5}$$

For the two-population selection-mutation model, an extension of Equation 2.4, we have:

$$\begin{aligned} \frac{d\pi_{1,j}}{dt} &= \alpha\pi_{1,j}\left[\tau^{-1}\left[v_1(e_j|\pi_2) - v_1(\pi_1|\pi_2)\right] + \pi_1 \log {\pi_1}^T - \log \pi_{1,j}\right] \\ \frac{d\pi_{2,j}}{dt} &= \alpha\pi_{2,j}\left[\tau^{-1}\left[v_2(e_j|\pi_1) - v_2(\pi_2|\pi_1)\right] + \pi_2 \log {\pi_2}^T - \log \pi_{2,j}\right] \end{aligned} \tag{2.6}$$

The two-population dynamics are used in Section 6.3.1 to model strategic behavior in 2 x 2 games.

2.2.2 Evolutionary stable strategies

One of the core concepts from evolutionary game theory is the notion of an evolutionary stable strategy. It is a refinement of the Nash equilibrium, i.e. every evolutionary stable strategy is a Nash equilibrium, but not vice versa. Nash equilibria appear as rest points of the selection dynamics defined by Equation (2.1), i.e. $\frac{d\pi_j}{dt} = 0$. However, stochastic deviations through mutation may lead evolution out of such a rest point. A strategy is called evolutionary stable, if natural selection counters small deviations from this strategy and pushes the population back towards the stable strategy.

Assume, the whole population plays according to some mixed strategy π. At some point in time, a mutant μ appears and is played by a small number of individuals. This mutant strategy may grow and establish in the population, or it may go extinct due to evolutionary pressure of natural selection. A strategy is evolutionary stable, if it cannot be invaded by any mutant strategy μ. Formally, the following two conditions must hold:

$$\begin{aligned} v(\pi|\pi) &\geq v(\mu|\pi) \\ \text{and if } v(\pi|\pi) = v(\mu|\pi) \text{ then } v(\pi|\mu) &> v(\mu|\mu) \end{aligned} \tag{2.7}$$

Thus, a mutant must not gain more profit against π than π against itself. Furthermore, whenever μ obtains an equally high profit as π, the evolutionary stable strategy must do better against the mutant than the mutant against itself.

2.2.3 Example: Rock-Paper-Scissors

The matrix given in Figure 2.5 describes both players' payoff matrix for a symmetric two-player normal form game, where they may choose between the three strategies Rock, Paper and Scissors. The payoff matrix for the second player equals the transposed of the first player's payoffs in symmetric games.

The game Rock-Paper-Scissors yields one mixed Nash equilibrium with the profile $\pi^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. This equilibrium point is not evolutionary stable. To prove this, we regard the two conditions for evolutionary stable strategies for a matrix game A. Let π^* denote the mixed Nash equilibrium and let μ

	Rock	*Paper*	*Scissors*
Rock	0	−1	1
Paper	1	0	−1
Scissors	−1	1	0

Figure 2.5: Payoffs matrix $A = B^T$ for both players in the symmetric two-player normal form game 'Rock-Paper-Scissors'.

denote an invading mutant. The strategy π^* is evolutionary stable, if the following two conditions hold:

$$\begin{aligned} \pi^* A \pi^* &\geq \pi^* A \mu \\ \text{and if } \pi^* A \pi^* = \pi^* A \mu \text{ then } \pi^* A \mu &> \mu A \mu \end{aligned}$$

For the example of Rock-Paper-Scissors, the first condition always holds as $\pi^* A \pi^* = \pi^* A \mu$ for all μ. Consider the example mutant $\mu' = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2})$, for which the second condition does not hold, as $\pi^* A \mu' = 0 = \mu' A \mu'$. Therefore, the strategy π^* can be invaded by mutants and is not evolutionary stable.

The replicator dynamics for this symmetric three-strategy game can be visualized in a simplex. Each corner $c_j = (x_j, y_j)$ of the simplex represents one pure strategy j. The position of a policy π^t at time t in the simplex is the weighted average of the corner points.

$$position(\pi^t) = \sum_{j=1}^{3} \pi_j^t c_j$$

The learning dynamics that can be observed in the next section show cyclic behavior. This explains why the Nash equilibrium is not evolutionary stable. Invading mutants do not go extinct, but rather start a perpetual cycling around the equilibrium point.

Directional field plots

Directional field plots are created by computing the replicator dynamics at a set of grid points and plotting arrows from each of these grid points in the direction of $\frac{d\pi}{dt}$. This captures the local dynamics of the learning behavior and supplies an overview over basins of attraction. The directional field plot for the game Rock-Paper-Scissors with replicator dynamics according to Equation 2.2 is displayed in Figure 2.6. It shows the cyclic evolution in this selection model.

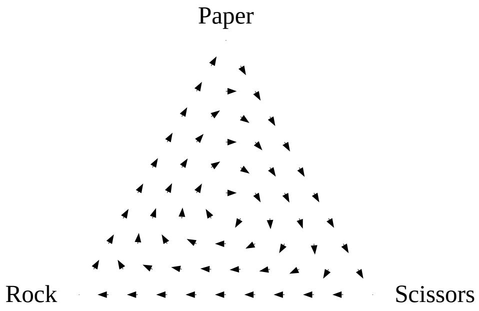


Figure 2.6: Directional field plot for the game Rock-Paper-Scissors.

Force field plots

Similar to the directional field plot, the force field plot uses arrows in the direction of $\frac{d\pi}{dt}$. Additionally, the length of the arrows is proportional to $||\frac{d\pi}{dt}||$. The force field plot for the Rock-Paper-Scissors with replicator dynamics according to Equation 2.2 is given in Figure 2.7. In contrast to the directional field plot, it shows that the learning speed changes less rapidly around the central rest point and the corner points, which are also always rest points in the selection model.

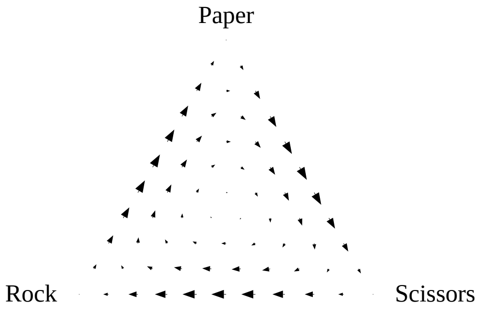


Figure 2.7: Force field plot for the game Rock-Paper-Scissors.

The replicator equations define a dynamical system which may feature a number of repellers and attractors. The latter are of particular importance to the analysis of asymptotic behavior. The strategy space can be partitioned into a set of basins of attraction. Each basin specifies a set of initial strategy profiles, for which the learning process eventually converges to the same attractor [13]. Assuming that an evolutionary process may start uniformly at any point in the strategy space, the size of the basin of attraction may be used to estimate the practical importance of an attractor. This can be achieved by inspection of the directional field plots or by analyzing the convergences of trajectories with initial policies that are uniformly sampled from the mixed strategy space.

2.3 Heuristic payoff tables

In the context of auctions, each pure strategy corresponds to a trading strategy. Furthermore, the utility function is proportional to the profit that a trader makes with each strategy, given the pure strategy profile.

Auctions with n trading agents and k available trading strategies provide a complex symmetric game. The full representation of this game requires to define the payoff for each of the k^n possible pure strategy profiles. This makes it intractable for practical purposes. Fortunately, the symmetric property of auctions can be exploited to compress the payoff representation.

A heuristic payoff table is proposed in [39] and adopted by several authors to capture the average profit of each type of trading strategy for all pure strategy profiles in a finite population [15, 24]. For the domain of auctions, the required profits can only be computed in simulation, where the private valuation of each trader is known.

The heuristic payoff table is defined for a finite population of identical agents who play a symmetric game. Because the agents are identical, it is not important who plays which strategy, but merely how many agents are playing each of the different strategies. So, given a pure strategy profile $s = (s_1, \ldots, s_n)$, we can derive that there are N_1 agents playing strategy 1, N_2 agents playing strategy 2, etc.. This will be denoted by a *discrete profile* $N = (N_1, \ldots, N_k)$, telling exactly how many agents play each strategy. The average profit for playing a strategy can then be denoted by the payoff vector $U(N) = (U_1(N), \ldots, U_k(N))$, indicating that strategy $j \in \{1, 2, \ldots, k\}$ would yield an average payoff of $U_j(N)$ for the discrete profile N. The distribution of n agents on k pure strategies is a combination with repetition, hence the number of discrete profiles of a heuristic payoff table is given by:

$$\binom{n+k-1}{n}$$

Let D denote a matrix that yields all discrete profiles as rows. The payoffs of these discrete profiles can be measured in many practical domains, including poker and auctions. However, measurements do not allow to capture the payoff to strategies that are not present in marginal strategy profiles, i.e. whenever $N_j = 0$ then $U_j(N)$ is unknown. Let U denote a matrix, where each row r corresponds to the payoff vector of the r'th row in D. The full heuristic payoff table is the compound $H = (D, U)$ of the profiles in D, and the payoff matrix U. The next section illustrates this by an example.

2.3.1 Example: Rock-Paper-Scissors

Table 2.1 shows a heuristic payoff table for the example game Rock-Paper-Scissors. Unknown payoffs are indicated with a dash. Let us assume the heuristic payoff table for Rock-Paper-Scissors was obtained from observations: There are two individuals who play a policy that is proportional to N, and for whom the payoff function is defined in Figure 2.5. Trivially, when all agents choose Rock, an average payoff of 0.0 is observed for Rock, while the other payoffs remain unknown. When the two players mix $(0.5, 0.5, 0)$ between Rock, Paper and Scissors, Rock and Paper will draw half of the time, and Rock will loose against Paper sometimes. The actual calculations to obtain this example are given in Section 4.1.

Table 2.1: The heuristic payoff table of Rock-Paper-Scissors with 2 agents. The first three columns give the discrete profiles N over the strategies and the last three columns give the corresponding payoff vectors $U(N)$.

N_{Rock}	N_{Paper}	$N_{Scissors}$	U_{Rock}	U_{Paper}	$U_{Scissors}$
2	0	0	0.0	-	-
1	1	0	-0.5	0.5	-
1	0	1	0.5	-	-0.5
0	2	0	-	0.0	-
0	1	1	-	-0.5	0.5
0	0	2	-	-	0.0

2.3.2 Replicator dynamics from heuristic payoff tables

A heuristic payoff table captures the payoff for a finite population of traders. Consequently, the payoff is only defined for the mixed strategy profiles that are feasible in this finite population. This section defines the expected payoff for any mixed strategy, which is required to compute the replicator dynamics directly from the heuristic payoff table.

Assuming each agent i independently chooses his pure strategy $s_i \in \{1, \ldots, k\}$ according to the same policy $\bar{\pi} = (\bar{\pi}_1, \ldots, \bar{\pi}_k)$, the probability of each strategy profile $s = (s_1, \ldots, s_n)$ equals $\prod_i^n \bar{\pi}_{s_i}$. The probability of a discrete profile can be computed as the product of the number of strategy profiles that are subsumed by this discrete profile and their probability. It is a multinomial, for which $Pr(N|\bar{\pi})$ is the probability of the discrete profile N given the mixed strategy $\bar{\pi}$.

$$Pr(N|\bar{\pi}) = \binom{n}{N_1, \ldots, N_k} \prod_{j=1}^{k} \bar{\pi}^{N_j}$$

The expected payoff for the mixed strategy $\bar{\pi}$ can then be computed as the weighted average over the payoffs received in all profiles. However, not all payoffs are known. When the payoff $U_j(N)$ to strategy j in the discrete profile N is not known, we define $U_j(N) = 0$. This implies, that the weight on a marginal strategy profile will not contribute to the weighted payoff average for the non-occurring strategy j. A correction term accounts for the missing payoffs by scaling the weighted average up by $\frac{1}{1-Pr(unknown|\bar{\pi})}$, where $Pr(unknown|\bar{\pi})$ denotes the accumulated weight that falls on unknown payoffs.

$$U_{average,j}(\bar{\pi}) = \frac{\sum_N Pr(N|\bar{\pi}) \cdot U_j(N)}{1 - Pr(unknown|j)}$$

The combined weighted payoff function gives the payoff vector $U_{average}(\bar{\pi}) = (U_{average,1}(\bar{\pi}), \ldots, U_{average,k}(\bar{\pi}))$ against the policy $\bar{\pi}$. This weighted payoff function is the equivalent of $A\bar{\pi}$ given a matrix game A. Now, the expected profit for policy π against policy $\bar{\pi}$ can be computed from the heuristic payoff table:

$$v(\pi|\bar{\pi}) = \pi U_{average}(\bar{\pi})^T$$

Using this definition of expected payoff, the replicator dynamics can be computed from the heuristic payoff table.

Chapter 3

Reinforcement learning

This chapter provides a concise introduction to single-agent and multi-agent reinforcement learning. The fundamentals and general principles are explained, as far as they contribute to the understanding of the analysis that is applied in this thesis. The learning schemes of Cross learning and Q-learning are introduced and challenges of multi-agent learning are sketched. In contrast to evolutionary game theory, which describes the learning process on a population level, the algorithms describe learning on an individual level.

Reinforcement learning has originally been studied in the context of single-agent environments. An agent receives a numerical reward signal, which it seeks to maximize in the long run. The environment provides this signal as a feedback on the sequence of actions that has been executed by the agent. Figure 3.1 depicts the environment-agent interaction schematically. Learners relate the reward signal to previously executed actions to learn a policy that maximizes the expected future reward [32].

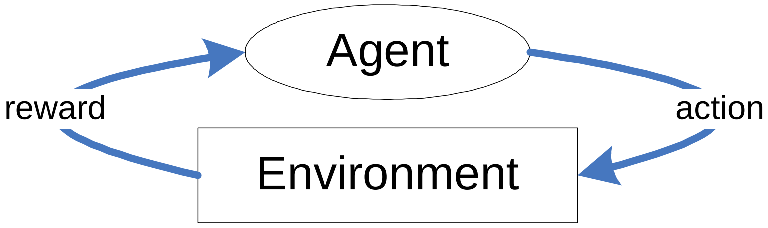


Figure 3.1: Agent-environment interaction, a feedback is given as a response to each action and may depend on the complete sequence of executed actions.

3.1 Models of reinforcement learning

Reinforcement learning models can be divided into policy and value iterators [36]. This section explains the learning models that are described by the evolutionary analysis. It first presents the most simple policy iterator *Cross learning* and then elaborates on Q-learning, an example of value based iteration.

3.1.1 Cross learning

Cross learning is a mathematical model for learning that originates from the field of psychology and has been first considered by Cross [5]. It assumes several agents who repeatedly play the same normal form game. At each iteration, the behavior of each agent can be described by his policy $\pi_i = (\pi_{i1}, \ldots, \pi_{ik})$, which indicates how likely any available action is to be played. One pure strategy is drawn according to the probabilities and the policy is updated based on the experienced reward.

In order to apply Cross learning, the normal form game needs to be normalized such that the payoffs are positive. Let us assume $u(s) \in [0,1]^k$. Upon execution of the strategy profile s at time t, the policy π_i is updated according to the following scheme:

$$\begin{aligned} \pi_{i,s_i}^{t+1} &\leftarrow u_i(s) - (1 - u_i(s))\, \pi_{i,s_i}^{t} \\ \text{and for all } j \neq s_i : \pi_{i,j}^{t+1} &\leftarrow (1 - u_i(s))\, \pi_{i,j}^{t} \end{aligned}$$

At each iteration, the probability of the played strategy s_i is pushed towards its utility. The term $1 - u_i(s)$ maintains the probability vector by scaling the previous policy down, such that $u_i(s)$ can be added to the played strategy.

There is an analogy between learning and biological evolution [3]. The learning agent updates, or evolves, his policy over the population of pure strategies. At each time step, a pure strategy is evaluated and subjected to evolutionary pressure, i.e. the learning process increases its probability if it performs well and decreases the probability if it performs poorly. The average learning behavior of Cross learning in the continuous time limit converges to the replicator dynamics with selection, defined by Equation 2.1 [2]. Cross learning belongs to the family of learning automata, for which a complete discussion is given in [18]

3.1.2 Q-learning

Q-learning was initially introduced for single-agent environments [40]. Each learning step refines a utility-estimation function for state-action pairs and generates a new policy from the estimated values to draw the next action to execute. Q-learning has been proven to converge to the optimal policy under appropriate parameter settings and additional assumptions like a suitable environment [40].

The idea originates from the Bellman optimality equation which is given in Equation 3.1. Let $R(p)$ be the reward for being in state p, $R(p) = 0$ for all states p except the goal state p^* for which $R(p^*) > 0$. $P(p'|p, a)$ denotes the probability to be in state p' given that action a is executed in state p and $\gamma \in [0, 1]$ is the discount factor for future rewards. V^* estimates the value of a state by taking into account the immediate reward and the discounted, expected future rewards.

$$V^*(p) = R(p) + \max_a \gamma \cdot \sum_{p'} P(p'|p, a) V^*(p') \tag{3.1}$$

The optimal action a is given by

$$a = \arg\max_{a'} \gamma \cdot \sum_{p'} P(p'|p, a') V^*(p')$$

Q-learning leverages the state value estimation to relate rewards to state-action pairs. A full discussion of Q-learning with states is given in [40, 32]. This thesis only considers stateless multi-agent games and consequently stateless Q-learning is described in the following section.

Multi-agent Q-learning without states

Multi-agent learning can be approached in different ways, depending on the information that each agent perceives. The approaches vary from joint action space learners to independent learners. When the joint action is known to an agent, he can learn in the joint strategy space. However, we assume that the opponents' strategies are unknown to each trading agent. Therefore, this thesis takes the perspective of independent learners, where each agent maintains an independent learning process that only depends on his own action and the perceived payoff.

The games under consideration do not feature states hence the Q-function plainly estimates utilities of the available actions. Furthermore, each agent has an independent Q-value estimation function. Equation 3.2 shows the Q-update rule for stateless Q-learning using the following terms:

- $Q_i^t(s_i)$ Q-value estimation function of player i at iteration t for action s_i
- s_i^t Action of player i played in iteration t
- r_i^t Reward for player i obtained in iteration t, in the context of games defined by the utility function
- $\alpha \in [0, 1]$ Learning rate

The new estimation is the weighted sum of the old estimation and the observed reward.

$$Q_i^t(s_i) \leftarrow (1-\alpha)\cdot Q_i^{t-1}(s_i) + \alpha \cdot r_i^t \tag{3.2}$$

In each iteration, an action needs to be chosen based on the current knowledge. This step is essential to balance exploitation versus exploration. A dynamic trade-off between exploration and exploitation can be implemented using the idea of temperature from physics. The Boltzmann distribution allows a probability generation from arbitrary parameters. This approach is also often used for simulated annealing where an initially high temperature promotes exploration and decreasing temperature over time leads to strong exploitation in the final phase. The policy π is generated by:

$$\pi_j = \frac{e^{Q_i(j)\cdot\tau^{-1}}}{\sum\limits_{k\in S} e^{Q_i(k)\cdot\tau^{-1}}}$$

By tuning the temperature parameter τ, the balance between exploration and exploitation can be adjusted while exploration is still directed toward promising actions.

The average learning behavior of Q-learning can be described by Equation (2.4) [35], which consequently features the Q-learning's parameters α and τ.

3.2 The challenge of multi-agent learning

Multi-agent learning is inherently more challenging than single-agent learning. Consider the example of two agents that learn to play soccer. Initially, both are amateurs and they learn to pass the ball in a very safe manner. They can handle more and more difficult situations the more they advance. Now, the safe pass is not actually a good pass anymore, because the agents could do better with a more aggressive forward pass that an amateur would not be able to get. The best action has changed by the learning process and is not only dependent on the state of the environment, but also on the complete history of actions that has been played before. In other words, the Markov property does not hold. This implies that many proofs from classical single-agent learning theory do not apply anymore and convergence to optimal strategies is not guaranteed. Fortunately, the link to evolutionary game theory provides a new powerful theoretical framework to analyze the learning behavior in multi-agent games.

Chapter 4

Normal form approximation

This chapter elaborates on the approximation of heuristic payoff tables by normal form games and defines the transitions between the two representations. Heuristic payoff tables are a compressed representation for symmetric games of arbitrarily many agents, e.g. they can represent a simple symmetric two-agent normal form game or auctions with many traders. Section 4.1 shows that creating a heuristic payoff table for a symmetric two-agent game is straight forward. This direct relation suggests that it is also possible to create a normal form game representation for any heuristic payoff table, which is a potentially lossy compression. Despite the possible information loss, the normal form approximation is attractive, as normal form games are more intuitive and therefore easier to analyze. They are well studied and allow to apply means from game theory with less computational effort such that more complex models can be used to derive strategic behavior.

4.1 From normal form games to heuristic payoff tables

The heuristic payoff table lists all possible discrete profiles and their associated payoff vectors. The combinations of k strategies and n agents provide the discrete profiles. Corresponding payoff vectors for each profile N against a mixed strategy $\pi = \frac{1}{n}N$ can be computed from the matrix game A as $A\pi^T$. Considering the example Rock-Paper-Scissors given in Section 2.3.1, the payoff to the profile $N = (1, 1, 0)$ can be computed as $U(N) = A\frac{1}{2}(1, 1, 0)^T = (-0.5, 0.5, 0)^T$, where A is the matrix given in Figure 2.5. The heuristic payoff table yields the payoff vector $(-0.5, 0.5, -)$ for consistency reasons with data from observations, where the payoff for non-occuring strategies cannot be determined.

Let D be the matrix where each row corresponds to a discrete profile N of n agents. Furthermore, let the matrix $P = \frac{1}{n} \cdot D$ yield policies that are proportional to the discrete profiles. The matrix U, that yields the corresponding payoff vectors $U(N)$ as rows, can then be computed as the product of P and A.

$$U = P \cdot A^T \tag{4.1}$$

The heuristic payoff table $H = (D, U)$ is the composition of the discrete profiles and the corresponding payoffs.

4.2 From heuristic payoff tables to normal form games

This section reverses the step of the previous section and shows the transition from a heuristic payoff table to a normal form game approximation. However, equation (4.1) cannot simply be solved for A as the values in the heuristic payoff table may be noise-prone due to stochasticity in the simulation experiments and may also be more complex than the normal form can capture. This leads to an over-constrained system of equations which can only be approximated, e.g. by minimizing the mean squared error or the maximal absolute deviation.

Minimizing mean squared error

A normal form game A that approximates the heuristic payoff table $H = (D, U)$ can be determined incrementally for each row A_i by finding a least mean squared error fit between the j'th column of U, denoted as U_j, and the reconstructed payoff vector $\tilde{U}_j = P \cdot A_j^T$ from the normal form game, where $P = \frac{1}{n} \cdot D$ as above, by solving the minimization problem:

$$\min_{A_j} \left|\left|U_j - \tilde{U}_j\right|\right|_2$$

A standard linear least square fitting algorithm can be used to solve this system for each row to compose the normal form game matrix.

Minimizing maximum absolute deviation

Linear programming optimizes a linear goal function subject to a system of linear inequalities. Using the same definitions of the profile matrix D, the probability matrix P, the game A and the payoff matrix U as above, the following program can be formulated.

$$\begin{aligned} \text{minimize} \quad & \epsilon \\ \text{variables} \quad & \epsilon, A_{jh}, \text{ for } j, h \in \{1, \ldots, k\} \\ \text{subject to} \quad & P \cdot A^T \leq U + \epsilon \\ & P \cdot A^T \geq U - \epsilon \end{aligned}$$

This program needs to be transformed to standard notation in order to apply algorithms from linear programming. For sake of convenience, each row A_j is determined separately. Let $c = (1, 0, \ldots, 0)$ and $x = (\epsilon, A_j)$ such that the goal function $c \cdot x^T$ minimizes epsilon. Furthermore, let $M = \begin{pmatrix} -1 & \\ \vdots & P \\ \vdots & -P \\ -1 & \end{pmatrix}$ and $b = \begin{pmatrix} U_j \\ -U_j \end{pmatrix}$ where U_j is the j'th column of the payoff matrix. Then, this linear program can be solved in standard notation:

$$\min_x \; c \cdot x^T \quad \text{subject to } \; M \cdot x^T \leq b, \; x \geq 0$$

In total, k linear programs need to be solved to compute the complete normal form matrix that approximates the heuristic payoff table with a minimal maximum absolute deviation.

Part II

Application to the auction domain

Chapter 5

Auctions

"Computer science is no more about computers than astronomy is about telescopes."

Edsger Dijkstra

This chapter specifies the auction type that is used in the experiments presented in this thesis, gives an intuition for the available trading strategies and explains the application of evolutionary game theory in auctions.

5.1 The clearing house auction

The traders that participate in an auction agree to subject to a set of market rules in order to exchange goods for money. This thesis considers a commodity market, i.e. a single type of an abstract good is traded. Each trader is assumed to have a *private valuation* of the good which is only known to himself. Buyers and sellers place offers to indicate their intention to trade at a certain price. The here considered *clearing house auction* proceeds in rounds and polls offers from each trader each round. When all offers are collected, an equilibrium price is established based on the available offers such that demand meets supply at this price. It is set to the average of the two offers that define the range of possible equilibrium prices, i.e. the lowest bid and the highest ask that can be matched in the equilibrium. Each buyer with an offer above that price is matched with a seller having an offer below that price. The *profit* of a transaction can be computed as the difference between the transaction price and the private value, assuming that buyers will not buy above their private value and sellers will not sell below their private value. For a more complete discussion and the relation of a clearing house auction to other auction types, we refer to [21, 22].

5.2 Trading strategies

A multitude of trading strategies has been devised to derive a good offer, possibly exploiting the knowledge about offers and transactions that were observed in previous rounds. The most trivial one is *Truth Telling* (TT) which just reveals the private value by placing offers exactly at that value. Despite its simplicity, it may be optimal in some situations [28]. The experiment of this thesis considers three more sophisticated trading strategies. Roth and Erev devised a reinforcement learning model of human trading behavior in [7] which is modified to perform in a clearing house auction as *Modified Roth-Erev* (MRE) [20]. MRE is evaluated in competition to *Gjerstad and Dickhaut* (GD) and *Zero Intelligence Plus* (ZIP). GD maximizes the expected profit, which is computed for a set of relevant prices as the product of profit and probability of successful matching [11]. ZIP places stochastic bids within a certain profit margin, which is lowered when a more competitive offer was rejected and increased when a less competitive offer was accepted [4].

Each of the trading strategy may have several parameters. The profit margin of ZIP is updated by a learner that can be tuned by its learning rate and momentum parameter. MRE learns a policy over a set of k discrete prices, where the learning behavior can be tuned by a recency and an exploration parameter. GD evaluates a price range and requires to specify the price interval to consider. This interval is commonly set to zero up to a maximum of relevant prices.

5.3 Evolutionary game theory in auctions

Given a set of available trading strategies, it is of high interest to find out which strategy is *best* in the sense that it yields the highest expected payoff. However, this question cannot be answered in general, as the performance of a trading strategy is highly dependent on the competition it faces [27]. Let us assume an auction, where traders only choose between the trading strategies described above. This means, the trading strategies that define the decisions within the auction and that may be adaptive themselves are now considered atomic strategies. Each iteration corresponds to one auction, where one of these strategies is played by each agent. The profit of each trader is dependent on the overall mix of strategies and traders may choose to change their strategy from auction to auction, e.g. applying a reinforcement learning algorithm to improve their expected payoff. This adaptation can be modeled by the replicator dynamics from evolutionary game theory which are formally connected to reinforcement learning [34] and which have been introduced in Chapter 2.

Chapter 6

Experiments

This chapter presents the experimental setup and results for the empirical investigation of the research questions listed in 1.2. In particular, the proposed methodology is tested on an example from the auction domain.

The general setup is summarized in Figure 6.1 and can be described as follows: Section 6.1 describes the simulation setup and calculation of a heuristic payoff table for a clearing house auction with the three trading strategies ZIP, GD and MRE. The resulting table is approximated in Section 6.2 using the methods described above and subsequently compared to its approximations. The comparison uses difference plots and replicator dynamics, visualized by directional or force field plots. Section 6.3 applies a learning model of selection and mutation to 2x2 games and to a normal form game auction approximation. The learning is illustrated by directional and force field plots for the replicator dynamics and by example trajectories.

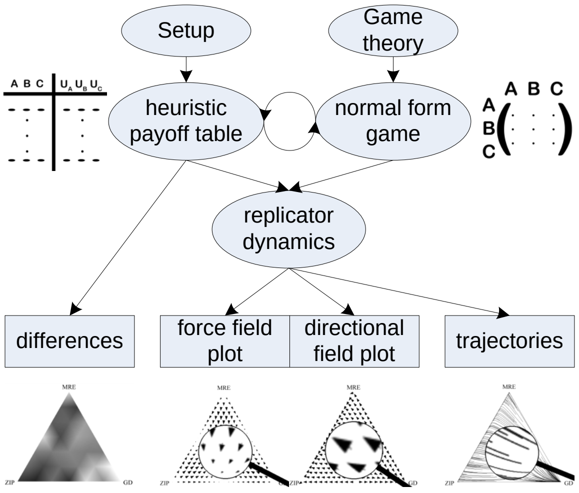


Figure 6.1: A scheme of the general experimental setup.

6.1 Calculating the heuristic payoff table

The heuristic payoff table given in Table 6.1 is obtained by simulating auctions with the *Java Auction Simulator API* (JASA) [23]. This empirical platform contains the trading strategies ZIP, MRE and GD which are described in Chapter 5, according to [4, 11, 20]. They are setup with the following parameters: ZIP uses a learning rate of 0.3, a momentum of 0.05 and a JASA specific scaling of 0.2. MRE chooses between 40 discrete prices using a recency parameter of 0.1, an exploration of 0.2 and scaling of 9. GD evaluates prices in the interval $[0, 360]$.

The heuristic payoff table is obtained from an average of 2000 iterations of a clearing house auction, populated by 6 traders. On the start of each

Table 6.1: The heuristic payoff table of a clearing house auction with 6 agents and the three strategies ZIP, MRE and GD. The first three columns give the discrete profiles N over the trading strategies and the last three columns give the corresponding payoff vectors $U(N)$.

N_{ZIP}	N_{MRE}	N_{GD}	U_{ZIP}	U_{MRE}	U_{GD}
6	0	0	99	-	-
5	1	0	97	100	-
5	0	1	89	-	69
4	2	0	96	94	-
4	1	1	90	88	65
4	0	2	85	-	69
3	3	0	97	92	-
3	2	1	87	90	64
3	1	2	85	80	73
3	0	3	76	-	73
2	4	0	97	96	-
2	3	1	91	91	66
2	2	2	84	83	67
2	1	3	78	70	76
2	0	4	62	-	80
1	5	0	97	97	-
1	4	1	93	89	62
1	3	2	86	84	69
1	2	3	73	71	75
1	1	4	73	57	77
1	0	5	56	-	80
0	6	0	-	94	-
0	5	1	-	91	62
0	4	2	-	84	67
0	3	3	-	75	71
0	2	4	-	65	76
0	1	5	-	43	79
0	0	6	-	-	79

auction, all traders are initialized without knowledge of previous auctions and with a private value drawn from the same distribution as in [39], i.e. an integer lower bound b is drawn uniformly from $[61, 160]$ and the upper bound from $[b + 60, b + 209]$ for each buyer. The sellers' private values are initialized similarly. These private values then remain fixed over the course of the auction, which runs 300 rounds on each of 5 trading days, where each trader is entitled to trade one item per day.

The further analysis of trading strategies in this auction is based on the heuristic payoff table. Besides applying evolutionary game theory directly to it, the next section also tests the newly proposed methodology of approximating the heuristic payoff table by normal form games and applying the analysis subsequently.

6.2 Normal form approximation

In order to test the proposed methodology, the heuristic payoff table is approximated as described in Section 4.2. This leads to the normal form game representations given in Figure 6.2. The two methods generate similar but not identical games. In both cases, ZIP against MRE yields the heighest payoff while MRE against GD yields the lowest one. Considering the full ranking of payoffs, only the two joint strategies ZIP-ZIP and MRE-ZIP are switched. Both games feature the same pure and symmetric Nash equilibrium $(0, 0, 1)$, but the least mean squared error approximation features another pure equilibrium at $(1, 0, 0)$ while a mixed symmetric equilibrium at $(0.73, 0.27, 0)$ is present in the least maximum absolute deviation game. The latter is used for further analysis. It has a matrix game value of 73.1, which can be guaranteed by the optimal mixed trading strategy $\pi^+ = (0.3, 0, 0.7)$. This means that a trader who plays ZIP with probability 0.3 and GD with probability 0.7 will get an expected payoff of at least 73.1 against any combination of ZIP, MRE and GD.

Least mean squared error

	ZIP	MRE	GD
ZIP	97.4	98.8	52.3
MRE	96.8	98.6	42.6
GD	64.8	59.1	83.4

Least maximum absolute deviation

	ZIP	MRE	GD
ZIP	93.8	102.7	52.9
MRE	94.9	100.0	38.3
GD	66.2	60.5	81.8

Figure 6.2: The symmetric two-player normal form game approximations of the heuristic payoff table for a clearing house auction with the three strategies ZIP, MRE and GD.

Comparison

The normal form game can also be converted back into a heuristic payoff table as described in Section 4.1. This allows a direct comparison of the payoffs in the different models and can be used to evaluate the information loss in the normal form. To understand the difference plots, which facilitate the comparison, it is useful to see which points of the simplex are actually defined by the heuristic payoff table. Figure 6.3 indicates the different positions of profiles from the heuristic payoff table with asterisks.

Two kinds of differences between the heuristic payoff table and its approximations are visualized in Figure 6.4. Both use a heatmap in form of a simplex, where the brightness is proportional to the differences. The first dif-

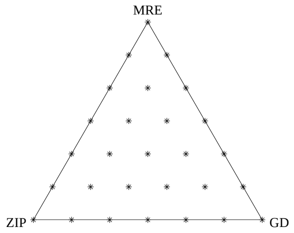
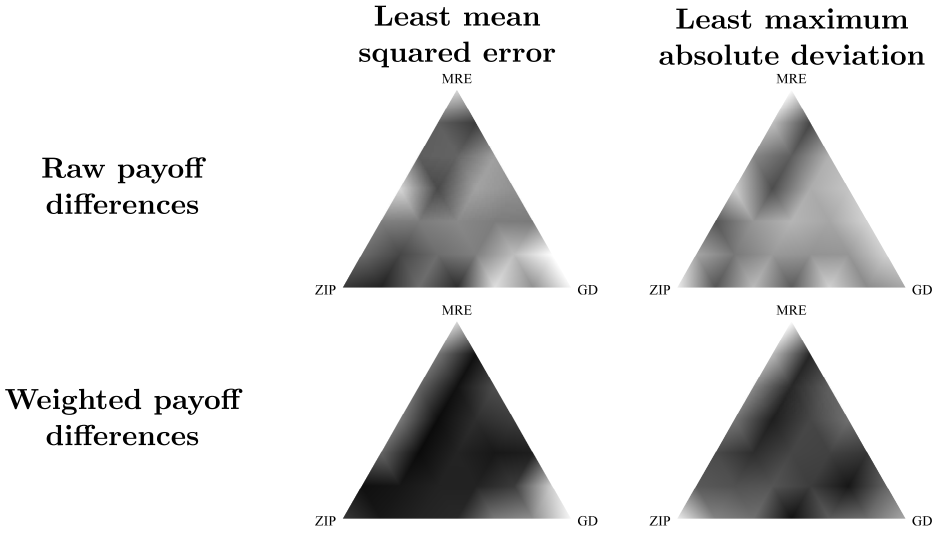


Figure 6.3: Profiles that are defined in a heuristic payoff table for 6 agents are marked with a star.


Least mean squared error

Least maximum absolute deviation

Raw payoff differences

Weighted payoff differences

Figure 6.4: Payoff differences between the heuristic payoff table and its normal form approximations. The brightness is proportional to the length of the difference vectors at the defined profile positions and the simplex is filled by interpolation.

ference compares the raw payoffs, i.e. $brightness(N) \sim \left|\left|U(N) - \tilde{U}(N)\right|\right|_2$, where $U(N)$ denotes the payoff vector in the heuristic payoff table and $\tilde{U}(N)$ denotes the one in its approximation. However, the raw payoffs are only given for exactly six agents who use a specific mix of trading strategies. When a large population is assumed that makes independent choices according to the probability vectors defined by the profiles, the payoff vector for such a population needs to be computed by weighting the payoffs as in Section 2.3.2. The second difference is hence computed from the weighted payoffs of both models, i.e. $brightness(N) \sim \left|\left|U_{average}(N) - \tilde{U}_{average}(N)\right|\right|_2$.

The maximum absolute deviation in the raw payoff differences amounts to 15.82% of the interval length of the payoffs in the heuristic payoff table for the least mean squared error and 9.95% in the least maximum absolute deviation approximation. The root mean squared error however is much lower at 4.92% and 5.51% respectively. The least mean squared error naturally generates a better average fit while the least maximum absolute deviation yields a better worst fit.

The weighted payoff differences are significantly lower than the raw payoff differences and relatively high only in marginal profiles. This may be ascribed to the fact that the marginal payoffs are not altered by the weighting. An arbitrary resolution can be applied for the weighted payoff comparison and the bright areas are expected to diminish further when the differences are computed with higher resolution in future analysis. The maximum absolute deviation is reduced to 9.72% for the least mean squared error approximation but remains the same in the least maximum absolute deviation fitting. The root mean squared error decreases to 2.84% and 3.40% respectively.

A small error in the weighted payoffs is a good condition for the replicator dynamics, which are computed from the weighted payoffs. The replicator dynamics describe the evolutionary change of a population which maps to a learning process, that is the actual subject of interest. In particular, the learning model allows to derive strategic behavior in auctions and therefore tackles the research questions. An approximation should not alter this analysis qualitatively, hence another evaluation of the normal form game approximations is based on the comparison of the replicator dynamics in the different models.

The selection dynamics of Equation (2.1) are derived from the heuristic payoff table and the normal form game representations and compared in Figure 6.5. There is a clear qualitative correspondence of the dynamics that arise from the three models. Differences are very small and hard to identify from the force field plots. Therefore, directional field plots are given as well, which allows to find the attractors and basins of attraction by inspection.

Analyzing the convergence of 10000 trajectories with uniformly sampled starting points, the position of the attractors and their corresponding basin

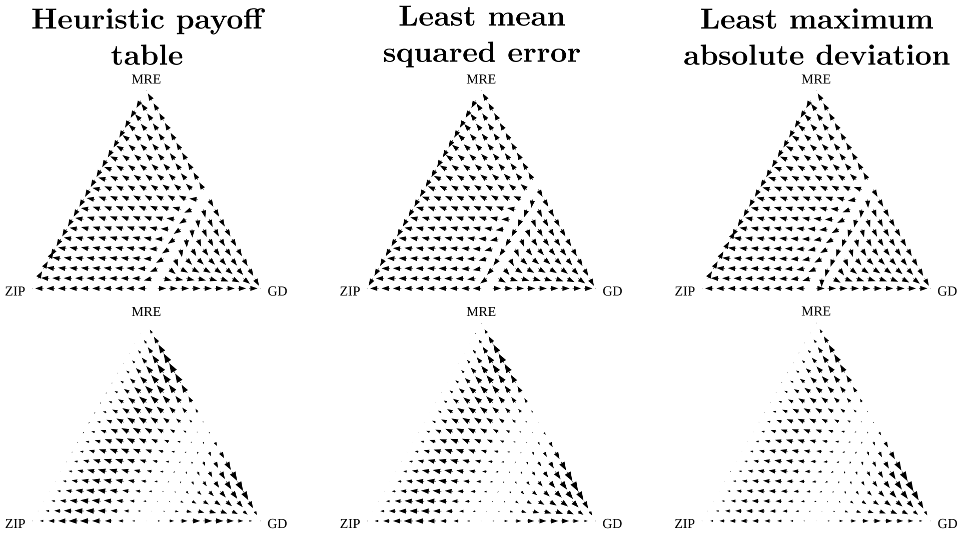


Figure 6.5: Comparison of the original replicator dynamics from the heuristic payoff table (left) to those from the normal form game approximations by least mean squared error (center) and minimized maximum absolute deviation (right) in the clearing house auction with 6 agents.

size can be estimated. A mixed attractor can be found at $(0.82, 0.18, 0.0)$ for the heuristic payoff table, at $(1, 0, 0)$ in least mean squared error fitting and at $(0.73, 0.27, 0.0)$ in minimized maximum absolute deviation. Despite these differences in the location of the attractor, the strategy space is partitioned into very similar basins of attraction. The pure attractor at $(0, 0, 1)$ is present in all dynamics and consumes 26.0% of the strategy space in the heuristic payoff table in comparison to 26.4% and 27.3% in the approximations. Learning processes that start in the remaining part of the strategy space converge to the mixed attractor.

In the context of evolutionary game theory, evolutionary stable strategies provide a concept to find stable solutions in symmetric normal form games. The attractors are evolutionary stable in the normal form game approximations and predict the attractors that are observed in the auction game dynamics. The differences in the basins of attraction are very small in both approximations, but the least maximum absolute deviation approximation is closer in the position of the mixed attractor. Therefore the latter is used for the further analysis. The concise representation allows to apply means from classical game theory and facilitates computing the game dynamics with less computational effort. Therefore, a more complex model of learning can be applied to the auction game, leveraging the newly obtained normal form representations in the next section.

6.3 Selection-mutation model

One of the criticisms on current research in auctions is the lacking account for exploration as it occurs in human learning. This section tackles this issue by applying the evolutionary selection-mutation model given in Equation 2.4. This model has a free temperature parameter to balance selection and mutation, which map to exploitation and exploration in the Q-learning algorithm [35].

First, this model is applied to 2 x 2 games to facilitate understanding of the model. Then, it is used to analyze the least maximum absolute deviation approximation of the auction. The selection-mutation equations converge to the selection model when temperature τ approaches zero. Therefore, the results presented below use $\tau = 0.01$ for a learning model of almost pure selection and increasing temperatures for different mutation rates. The selection-mutation model is not defined for marginal profiles. Hence, only non-marginal profiles are considered, where each strategy is played with a probability of at least 10^{-6}.

6.3.1 2 x 2 games

Three 2 x 2 games are investigated and represent the three subclasses that 2 x 2 games can be divided into. Figure 6.6 shows the two-population dynamics according to Equation (2.6) under different temperatures in the Prisoners' Dilemma, Battle of Sexes and Matching Pennies. The plots are insensitive to the learning rate α. Though α scales $\frac{d\pi_i}{dt}$, the length of the arrows is normalized and only shows relative differences in $\left|\left|\frac{d\pi_i}{dt}\right|\right|_2$.

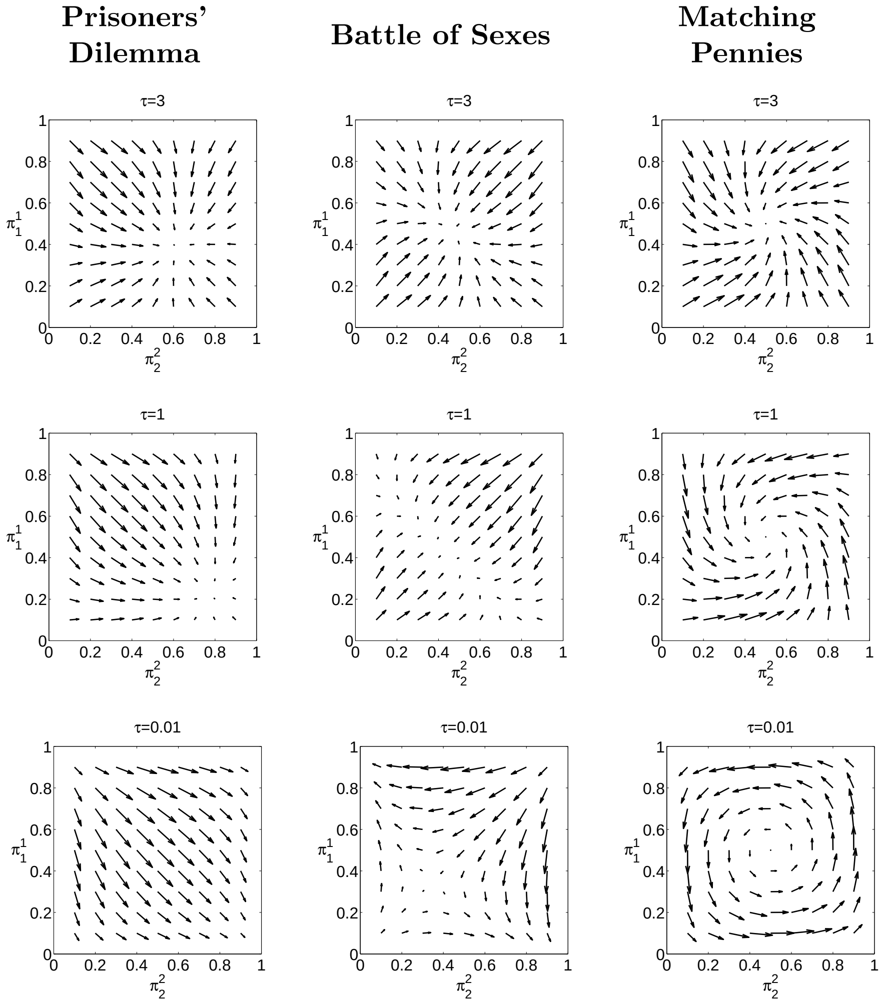


Figure 6.6: Force field plots of the replicator dynamics in a representative selection of normal form game examples.

The attractor in the Prisoners' Dilemma converges to the pure strategy *confess* when the temperature decreases and the learner ceases exploration. Similarly, the two attractors of the Battle of Sexes appear mixed under exploration and converge to the pure strategies. The mixed attractor of Matching Pennies remains at the same mixed strategy but the convergence behavior changes gradually. Under exploration the learners converge quickly while they tend to cyclic behavior the more they exploit. In the selection-only limit, the rest point is not evolutionary stable anymore and the dynamics are cyclic.

The model of investigation describes the average learning behavior of Q-learning, which may be applied with a decreasing temperature to overcome local optima [32]. This can be illustrated by example learning trajectories. A trajectory is started at some point π^0 in the strategy space and evolves according to the learning dynamics with $\pi^{t+1} = \pi^t + \delta \cdot \frac{d\pi^t}{dt}$. Figure 6.7 shows 30 policy trajectories. They are computed with $t_{max} = 1000$ iterations, a step size of $\delta = 0.01$ and learning rate $\alpha = 1$, where the temperature decreased from $\tau_{max} = 3$ to $\tau_{min} = 0.25$ according to Figure 6.8, which plots the temperature τ over time t.

$$\tau(t) = (\tau_{max} - \tau_{min}) \cdot (1 - \frac{t}{t_{max}})^3 + \tau_{min} \tag{6.1}$$

The force field plots in Figure 6.6 help to interpret these trajectories. When the temperature decreases, the attractors in the Prisoners' Dilemma and Battle of Sexes move from mixed strategies to pure ones. It can be observed that the trajectories follow the attractor like a moving target. The speed of that target is determined by the temperature function, given in Equation 6.1, while the speed of the trajectories is determined by the learning rate α and the step size parameter δ. This is an interesting observation because it means that the temperature decrease needs to be tuned with respect to α and δ.

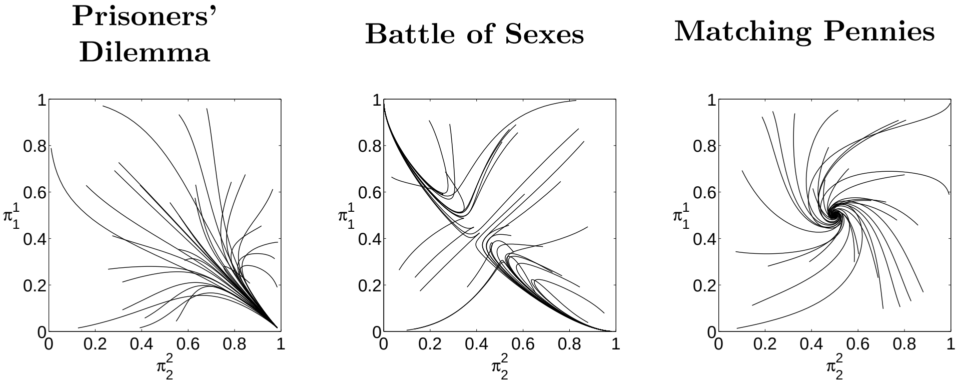


Figure 6.7: Normal form game trajectories with decreasing temperature.

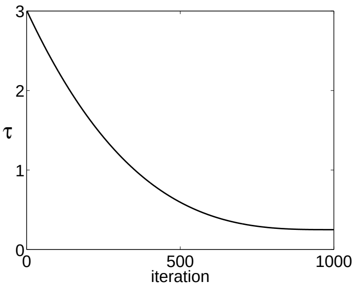


Figure 6.8: Temperature function for the normal form game trajectories.

6.3.2 Auctions

The previous section has investigated 2 x 2 game dynamics under varying temperatures. This section applies the same analysis to the least maximum absolute deviation normal form game approximation of the auction game, but it uses the single population dynamics defined in Equation 2.4. Figure 6.9 displays the directional field and force field plots under different temperatures. A mixed attractor converges to the Pareto optimal evolutionary stable strategy when the temperature approaches zero. The pure attractor for GD only appears when exploration is very low. This suggests, that the learning process of all initial policies may converge to the Pareto optimal solution, given a sufficiently long time of exploration.

Figure 6.10 shows 200 learning trajectories of $t_{max} = 1000$ iterations with uniformly sampled starting points, a step size of $dt = 0.01$ and a learning rate $\alpha = 1$. The temperature is decreased from $\tau_{max} = 3$ to $\tau_{min} = 0.001$, according to Equation (6.1). As expected, all trajectories converge to the Pareto optimal attractor $(0.73, 0.27, 0)$.

Similar to the trajectories of 2 x 2 games, the trajectories of the auction game follow a moving attractor. The results show that the whole strategy space may converge to the Pareto optimal evolutionary stable strategy, given a sufficiently long phase of exploration in this auction. This result was correctly predicted by the replicator dynamics given in Figure 6.9. It differs qualitatively from the selection-only model, showing less convergence to the local optimum. Therefore, the conclusions about the asymptotic behavior of agents depend on the assumed amount of exploration in the learning process under investigation. It is realistic to assume that humans explore in their learning process, therefore the selection-mutation model is essential for a realistic description of strategic behavior in auctions.

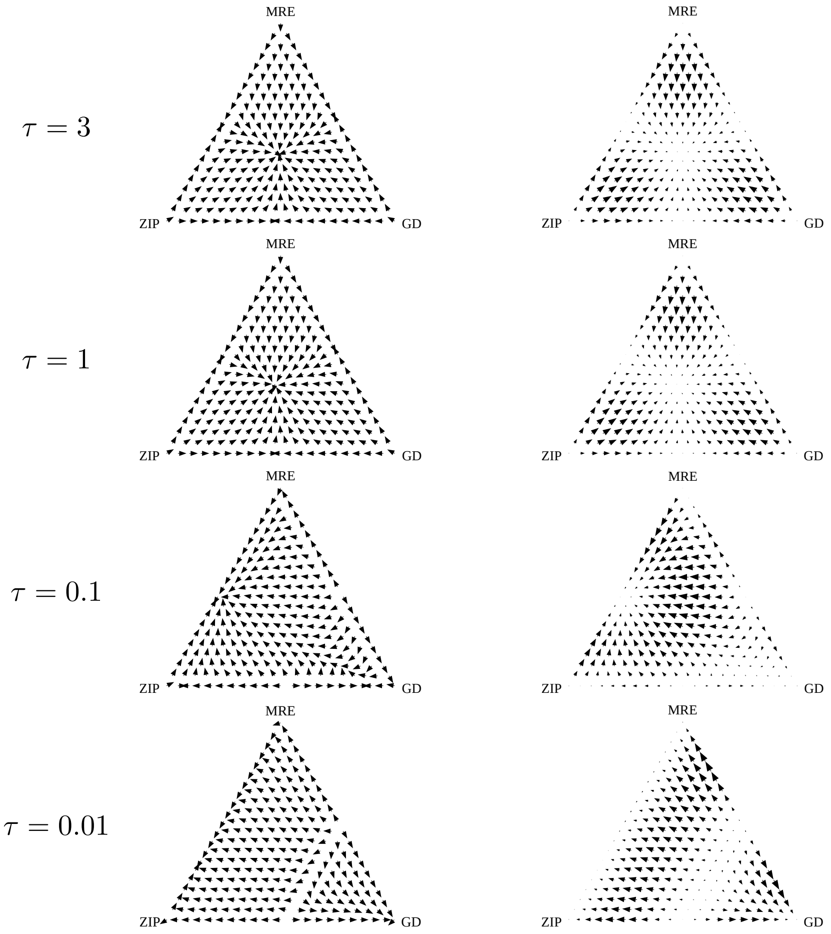


Figure 6.9: Replicator dynamics of the least maximum absolute deviation approximation given in Figure 6.2.

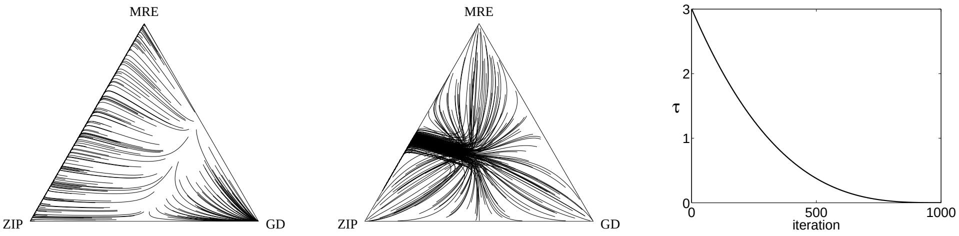


Figure 6.10: Learning trajectories in the least maximum absolute deviation approximation using a selection model (left) and a mutation model (center), where the temperature decreases according to the function that is displayed on the right.

Chapter 7

Discussion and conclusions

This chapter first relates the methodology and experiments of this thesis to previous work and then discusses limitations of the proposed approach. Conclusions and directions for future research close this thesis.

7.1 Discussion

The heuristic payoff table is a simplification of a symmetric game. In fact, the methodology that is introduced in this thesis extends this simplification by the approximation to an even smaller model. The full payoff function for the symmetric game would map each joint strategy to a payoff for each agent and require k^n entries. In contrast to this, the heuristic payoff table maps discrete strategy profiles to payoffs for each strategy, using $\binom{n+k-1}{n}$ rows and k times as many payoff entries. The normal form game approximation maps probability distributions over the strategies to payoffs for each strategy and only requires k^2 entries. The example used in the case study of this thesis features 6 agents and 3 strategies. It would require $3^6 = 729$ entries for the full representation, $\binom{6+3-1}{6} = 28$ rows or $28 \cdot 3 = 84$ payoff entries in the heuristic payoff table, and only $3^2 = 9$ entries for the normal form game approximation. The latter actually reduces the multi-player game to a two-player game. The reduced size makes it easier to process the numerical representation computationally, but also makes it more accessible to manual inspection. This is of particular importance when the number of strategies is increased. The matrix game lists how much each strategy gains against each of the alternatives, i.e. all pairwise comparisons. Even for large numbers of strategies, the normal form gives an intuition on their relative strength.

The methodology proposed in this thesis extends previous work, which has captured the payoff in auctions in heuristic payoff tables and used evolutionary game theory to perform a comparative analysis of trading strategies. A bridge to classical game theory has been devised and is used to apply a more sophisticated game theoretical analysis, i.e. computing optimal strategies and using a selection-mutation perspective from evolutionary game theory. The case study of the proposed methodology investigates the performance regarding a 6-trader clearing house auction. It has demonstrated the feasibility of this approach and has allowed to apply a computationally more complex analysis. This analysis shows significant differences to the previously deployed selection-only model. It emphasizes that explorative learning yields significantly different learning dynamics. As we consider humans explorative, it suggests that the selection-mutation perspective should be applied to obtain a realistic model of human learning.

The proposed approach is general in the number of actions and can be transfered to higher dimensions. However, the approximation of heuristic payoff tables by normal form games imposes a linear model on the approximated payoffs. To be precise, let the payoff matrix A denote the normal form approximation. Changing from policy π to policy π', where $\delta = \pi' - \pi$ is the difference vector between the policies, the payoff vector will change by exactly $A\delta$. In contrast to that, the heuristic payoff table may define more complex payoff functions such that the difference between the payoff vectors for π and π' is almost arbitrary.

This implies, that the approximation may be an oversimplification for complex dynamics, which may arise from intricate interactions of real multi-agent systems. However, the multinomial weighting, that is applied to heuristic payoff tables to compute the expected payoff, appears to smoothen the payoff signal such that the qualitative differences are reduced. This may be explained by the fact, that the weighting is a linear combination of multinomial weights of the different profiles. This implies, that the payoff differences between two policies π and π' are also constrained, especially if they are close together. Consider two example policies $\pi = (0.9, 0.1, 0)$ and $\pi' = (0.8, 0.2, 0)$ for the game Rock-Paper-Scissors. Computing the expected payoff from the 2-agent heuristic payoff table given in Table 2.1 would use a high weight on profiles $(2, 0, 0)$ and $(1, 1, 0)$ for both policies. Therefore, their payoff vectors will be similar. Similar reasoning applies for more agents and more strategies, because the multinomial weighting adapts accordingly. The exact theoretical relation between the linearity limitation and the weighting scheme remains to be addressed in future work.

7.2 Conclusions

This thesis has modeled the strategic behavior of trading agents in a clearing house auction. A heuristic payoff table has been obtained from simulation and was approximated by the newly devised methodology. The approximation has been evaluated quantitatively and qualitatively. Furthermore, a selection-mutation model has been applied to 2 x 2 games and an approximation of the heuristic payoff table.

The contributions of this thesis are two-fold: A methodology to approximate heuristic payoff tables by normal form games has been introduced and an empirical case study has been performed by applying an evolutionary selection-mutation model to a game representation of an auction.

The case study has demonstrated the viability of the proposed methodology. Rather than merely participating myopically, a rational agent can now inspect the game strategically. This implies, that means and reasoning from classical game theory can be applied, e.g. to analyze asymptotic properties of the auction. Furthermore, the selection-mutation model has shown that learners may converge to a different strategy mix than in the selection-only model, when a sufficiently long phase of exploration is present. It is reasonable to assume, that humans are not hyper-rational but rather adapt to the circumstances using exploration. Therefore, the selection-mutation model describes human learning more realistically.

The obtained normal form game approximation is more intuitive, computationally less expensive to analyze and fills in a gap of missing payoffs in the blind spots of the heuristic payoff table. In fact, the normal form game can even be constructed from partial heuristic payoff tables, e.g. when a number of profiles could not be observed. This may provide useful insights into the dynamics of incomplete heuristic payoff tables that can be observed in different domains, e.g. in poker. The theoretical contributions of this thesis are not domain-specific. Hence, they may be transfered to any other domain, where the strategic interaction between identical agents can be modeled as a symmetric game.

7.3 Future work

The proposed methodology needs to be tested on other auctions and domains, possibly evaluating it on higher numbers of strategies. An even better approximation may be obtained by changing the goal function. Currently, each row of the normal form approximation is computed separately. However, the relation between the rows is important, because the qualitative differences in the approximation result from changes in the relative rank of payoffs for one profile. Minimizing the length $\left|\left|U(N) - \tilde{U}(N)\right|\right|_2$ of the difference vectors between the heuristic payoff table and its approximation may therefore yield better results than minimizing the maximum absolute deviation or the mean squared error.

Future work will also aim to argue for the described approach on a more theoretical level. Therefore, we will look for structure in the deviation from the linear model, in particular where and why qualitative changes occur. Finally, the approach that is described in this thesis may be established as the basis of a more general framework to analyze strategic behavior in complex multi-agent games.

Bibliography

[1] K. Binmore. *Fun and Games.* D. C. Heath and Company, 1992.

[2] T. Börgers and R. Sarin. Learning through reinforcement and replicator dynamics. *Journal of Economic Theory*, 77(1), 1997.

[3] R. Boyd and P. J. Richerson. *Culture and the Evolutionary Process.* University of Chicago Press, Chicago, 1985.

[4] D. Cliff and J. Bruten. Minimal-intelligence agents for bargaining behaviours in market-based environments. Technical report, Hewlett-Packard Research Laboratories, 1997.

[5] J. G. Cross. A stochastic learning model of economic behavior. *The Quarterly Journal of Economics*, 87(2):239–66, May 1973.

[6] R. Dash, D. Parkes, and N. Jennings. Computational mechanism design: A call to arms. *IEEE Intelligent Systems*, 18 (6):40–47, 2003.

[7] I. Erev and A. E. Roth. Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *The American Economic Review*, 88(4):848–881, 1998.

[8] N. Y. S. Exchange. Stock market activity. Technical report, New York Stock Exchange, 2000. available at http://www.nyse.com/ ... pdfs/02_STOCKMARKETACTIVITY.pdf.

[9] R. Gibbons. *A Primer in Game Theory.* Harvester Wheatsheaf, 1992.

[10] H. Gintis. *Game Theory Evolving.* Princeton University Press, 2000.

[11] S. Gjerstad and J. Dickhaut. Price formation in double auctions. *Games and Economic Behavior*, 22(1):1–29, January 1998.

[12] D. K. Gode, K. Dhananjay, and S. Sunder. Allocative efficiency of markets with zero-intelligence traders: Market as a partial substitute for individual rationality. *Journal of Political Economy*, 101(1):119–137, February 1993.

[13] M. W. Hirsch, S. Smale, and R. Devaney. *Differential Equations, Dynamical Systems, and an Introduction to Chaos.* Academic Press, 2002.

[14] J. Hofbauer and K. Sigmund. *Evolutionary Games and Population Dynamics.* Cambridge University Press, 2002.

[15] T. B. Klos and G. J. van Ahee. Evolutionary dynamics for designing multi-period auctions (short paper). In *Proceedings 7th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2008).* IFAAMAS, 2008.

[16] J. McMillan. Selling spectrum rights. *Journal of Economic Perspectives*, 8(3):145–162, Summer 1994.

[17] P. Mirowski. *Machine Dreams: Economics Becomes a Cyborg Science.* Cambridge University Press, New York, NY, USA, 2001.

[18] K. Narendra and M. Thathachar. *Learning Automata An Introduction.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1989.

[19] J. Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, September 1951.

[20] J. Nicolaisen, V. Petrov, and L. Tesfatsion. Market power and effciency in a computational electricity market with discriminatory double-auction pricing. *IEEE Transactions on Evolutionary Computation*, 5(5):504–523, 2001.

[21] S. Parsons, M. Marcinkiewicz, J. Niu, and S. Phelps. Everything you wanted to know about double auctions, but were afraid to (bid or) ask. Technical report, Brooklyn College, City University of New York, 2900 Bedford Avenue, Brooklyn, NY 11210, USA, 2006.

[22] S. Parsons, J. Rodriguez-Aguilar, and M. Klein. A bluffer's guide to auctions. Technical report, Center for Coordination Science, MIT, 2004.

[23] S. Phelps. Java auction simulator api. http://www.csc.liv.ac.uk/ sphelps/jasa/, 2005.

[24] S. Phelps, M. Marcinkiewicz, and S. Parsons. A novel method for automatic strategy acquisition in n-player non-zero-sum games. In *AAMAS '06: Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 705–712, Hakodate, Japan, 2006. ACM.

[25] S. Phelps, S. Parsons, and P. McBurney. Automated trading agents verses virtual humans: An evolutionary game-theoretic comparison of two double-auction market designs. In *Proceedings of the 6th Workshop on Agent-Mediated Electronic Commerce*, 2004.

[26] T. Roughgarden. The price of anarchy is independent of the network topology. *Journal of Computer and System Sciences*, 67(2):341–364, 2003.

[27] J. Rust, J. Miller, and R. Palmer. Behavior of trading automata in a computerized double auction market. In D. Friedman and J. Rust, editors, *The Double Auction Market: Institutions, Theories, and Evidence.* Addison-Wesley, 1993.

[28] M. A. Satterthwaite and S. R. Williams. The rate of convergence to efficiency in the buyer's bid double auction as the market becomes large. *The Review of Economic Studies*, 56(4):477–498, 1989.

[29] J. M. Smith. *Evolution and the Theory of Games.* Cambridge University Press, Cambridge, UK, 1982.

[30] J. M. Smith and G. R. Price. The logic of animal conflict. *Nature*, 246:15–18, 1973.

[31] V. L. Smith. An experimental study of competitive market behaviour. *Journal of Political Economy*, 70 (2):111–137, April 1962.

[32] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

[33] L. Tesfatsion. Agent-based computational economics: Growing economies from the bottom up. *Artif. Life*, 8(1):55–82, 2002.

[34] K. Tuyls and S. Parsons. What evolutionary game theory tells us about multiagent learning. *Artif. Intell.*, 171(7):406–416, 2007.

[35] K. Tuyls, P. t Hoen, and B. Vanschoenwinkel. An evolutionary dynamical analysis of multi-agent learning in iterated games. *Autonomous Agents and Multi-Agent Systems*, 12:115–153, 2005.

[36] H. J. van den Herik, D. Hennes, M. Kaisers, K. Tuyls, and K. Verbeeck. Multi-agent learning dynamics: A survey. In M. Klusch, K. V. Hindriks, M. P. Papazoglou, and L. Sterling, editors, *CIA*, volume 4676 of *Lecture 3s in Computer Science*, pages 36–56. Springer, 2007.

[37] J. van Neumann and O. Morgenstern. *The Theory of Games and Economic Behavior.* Princeton University Press, 1944.

[38] W. Vickrey. Counterspeculation, auctions, and competitive sealed tenders. *The Journal of Finance*, 16(1):8–37, 1961.

[39] W. E. Walsh, R. Das, G. Tesauro, and J. O. Kephart. Analyzing complex strategic interactions in multi-agent systems. In P. Gmytrasiewicz

and S. Parsons, editors, *Proceedings of the Workshop on Game Theoretic and Decision Theoretic Agents*, pages 109–118, 2002.

[40] Watkins. Learning from delayed rewards. *PhD thesis, King's College, Oxford*, 1989.

[41] J. W. Weibul. *Evolutionary Game Theory.* The MIT Press, 1995.

Glossary

Battle of Sexes
A 2x2 normal form game with two pure equilibria. 11, 12, 39, 41

Cross learning
A reinforcement learning algorithm. 3, 21, 22

directional field plot
A visualization for replicator dynamics. 15, 16, 37

evolutionary game theory
Mathematical study of strategic conflicts. 1, 3, 7, 12, 14, 31, 32, 35, 38, 45, G

evolutionary stable strategy
Solution concept of evolutionary game theory. 7, 12, 14, 38, 41

force field plot
A visualization for replicator dynamics. 15, 33, 37, 41

game
Numerical representation of a strategic conflict. 7, 46

game theory
Mathematical study of strategic conflicts. 3, 7–9, 12, 17, 18, 25, 38, 45, 46

heuristic payoff table
Table that lists discrete profiles and their payoff vectors. 3, 4, 7, 17–19, 25–27, 33–37, 45–47

learning automaton
A reinforcement learning algorithm. 22

Matching Pennies
A 2x2 normal form game with one mixed equilibrium. 12, 39, 41

normal form game

Game of simultaneous action selection. 3, 7, 8, 15, 17, 18, 22, 25, 26, 33, 35, 37, 38, 41, 45, 46

Prisoners' Dilemma

A 2x2 normal form game with one pure equilibrium. 11, 39, 41

Q-learning

A reinforcement learning algorithm. 3, 14, 21–24, 39, 41

reinforcement learning

A model for learning that is solely based on a reward feedback on performed actions. 1, 3, 21, 32, G

replicator dynamics

Describes the evolutionary progress of a population. 1, 12, 15, 32, 33, 37

Rock-Paper-Scissors

A 3x3 normal form game with one mixed equilibrium. 15, 18, 25