

Chapter 2

Background and Previous Work

In this chapter, we first review some concepts and terminologies from game theory and multiagent systems areas that will be used throughout the book in Sect. 2.1. After that, we explore previous work in multiagent learning literature by dividing them into two major categories: cooperative multiagent systems and competitive multiagent systems, distinguished by the underlying intentional stance of the agents within the system. Within each category, we review previous work in three different parts distinguished by the different goal that the work targets at. In Sect. 2.2, we focus on reviewing previous work whose goal falls into one of the following major solution concepts: Nash equilibrium, fairness, or social optimality. In Sect. 2.3, we focus on investigating previous work targeting at either of the following major solution concepts: Nash equilibrium, maximizing individual benefits, and Pareto-optimality.

2.1 Background

Game theory has been the most commonly adopted mathematical tool for us to model the strategic interactions among different agents in different interaction scenarios. We usually model the one-time interaction among agents as single-shot normal-form game and finite/infinite repeated games to model the repeated interaction among agents.

2.1.1 Single-Shot Normal-Form Game

In single-shot games, each agent is allowed to choose one action from its own action set simultaneously, and each agent will receive its own payoff based on the joint

action of all agents involved in the interaction. Formally a n -player normal-form game G is a tuple $\langle N, (A_i), (u_i) \rangle$ where

- $N = \{1, 2, \dots, n\}$ is the set of players.
- A_i is the set of actions available to player $i \in N$.
- u_i is the utility function of each player $i \in N$, where $u_i(a_i, a_j)$ corresponds to the payoff player i receives when the outcome (a_i, a_j) is achieved.

We call each possible combination of all players' action (joint action) an outcome. There are a number of important solution concepts defined in normal-form games including Nash equilibrium, Pareto-optimality, and social optimality, which can be understood as a subset of outcomes satisfying special properties. These solution concepts also have been adopted as the learning goals when designing different learning strategies in multiagent learning area. In the following, we will introduce each of them in the context of two-player normal-form games which is the most commonly investigated case in the multiagent learning literature. The most important solution concept in game theory is Nash equilibrium. Under a Nash equilibrium, each agent is making its best response to the strategy of the other agent, and thus no agent has the incentive to unilaterally deviate from its current strategy.

Definition 2.1 A *pure strategy Nash equilibrium* for a two-player normal-form game is a pair of strategies (a_1^*, a_2^*) such that

1. $u_1(a_1^*, a_2^*) \geq u_1(a_1, a_2^*), \forall a_1 \in A_1$
2. $u_2(a_1^*, a_2^*) \geq u_2(a_1^*, a_2), \forall a_2 \in A_2$

For example, considering the two-player normal-form game in Fig. 2.1, we can easily check that there are two pure strategy Nash equilibria in this game: (C, D) and (D, C) . Under each of these two outcomes, no agent has the incentive to unilaterally switch its current action to another one.

If the agents are allowed to use mixed strategy, then we can naturally define the concept of *mixed strategy Nash equilibrium* similarly.

Fig. 2.1 An example of two-player conflicting-interest game

1's payoff, 2's payoff		Player 2's action	
		C	D
Player 1's action	C	0, 0	1, 2
	D	2, 1	0, 0

Definition 2.2 A *mixed strategy Nash equilibrium* for a two-player normal-form game is a pair of strategies (π_1^*, π_2^*) such that

1. $\bar{U}_1(\pi_1^*, \pi_2^*) \geq \bar{U}_1(\pi_1, \pi_2^*), \forall \pi_1 \in \Pi(A_1)$
2. $\bar{U}_2(\pi_1^*, \pi_2^*) \geq \bar{U}_2(\pi_1^*, \pi_2), \forall \pi_2 \in \Pi(A_2)$

where $\bar{U}_i(\pi_1^*, \pi_2^*)$ is player i 's expected payoff under the strategy profile (π_1^*, π_2^*) , and $\Pi(A_i)$ is the set of probability distributions over player i 's action space A_i . A mixed strategy Nash equilibrium (π_1^*, π_2^*) is degenerated to a *pure strategy Nash equilibrium* if both π_1^* and π_2^* are pure strategies.

If we again use the same game in Fig. 2.1 as an example, we can verify that there also exists one mixed strategy Nash equilibrium: $\pi_i^*(C) = \frac{1}{3}, \pi_i^*(D) = \frac{2}{3}$, for $i = 1, 2$.

Another important solution concept in game theory is Pareto-optimality. An outcome is *Pareto-optimal* if and only if there does not exist another outcome under which no player's payoff is decreased and also at least one player's payoff is strictly increased. We can formalize it in the following definition:

Definition 2.3 An outcome s is *Pareto-optimal* if and only if there does not exist another outcome s' such that $\forall i \in N, u_i(s') \geq u_i(s)$, and there exists some $j \in N$, for which $u_j(s') > u_j(s)$.

For example, considering the two-player normal-form game in Fig. 2.1, both the outcomes (C, D) and (D, C) are Pareto-optimal.

The last solution concept is social optimality, which refers to those outcomes under which the sum of all agents' payoffs is maximized. For example, considering the two-player normal-form game in Fig. 2.1, both the outcomes (C, D) and (D, C) are socially optimal, which coincide with the set of Pareto-optimal outcomes.

2.1.2 Repeated Games

Repeated games can be used to model the repeated interactions among the same set of agents for finite or infinite number of times. In repeated games, usually a given normal-form game is played among the same set of players. If a strategic game is infinitely repeated, it is called an infinitely repeated game.

Definition 2.4 An infinitely repeated game of G is $\langle N, (A_i), (u_i), (\succeq_i) \rangle$ where

- N is the set of n players.
- A_i is the set of actions available to player i .
- u_i is the payoff function of each player i , where $u_i(a_i, \dots, a_n) = p_i$, the payoff player i receives when the outcome $(a_i, \dots, a_n)$ is achieved. Here a_i is the actions player i chooses, and $(a_i, \dots, a_n)$ is called an action profile.

- $\succsim_i$ is the preference relation for player i which satisfies the following property: $O_1 \succsim_i O_2$ if and only if $\lim_{t \rightarrow \infty} \sum_{k=1}^t (p_1^k - p_2^k)/t \geq 0$, where $O_1 = (a_{i,t}^1, \dots, a_{n,t}^1)_{t=1}^\infty$ and $O_2 = (a_{i,t}^2, \dots, a_{n,t}^2)_{t=1}^\infty$ are the outcomes of the infinitely repeated game, and p_1^k and p_2^k are the corresponding payoffs player i receives in round k of outcomes O_1 and O_2 , respectively.

The preference relation we adopt here is the limit of means preference relation which is the one we adopt in the book. Notice that there are also other different ways to define preference relation such as discounting and overtaking, and interested reader may refer to the book [1] for details. Also we usually assume that the action set of all agents are the same, i.e., $A_1 = \dots = A_n = A$.

In infinitely repeated games, the strategy of an agent specifies the action choice of the agent for each round. Considering the large space of the agents' possible strategies, it is very difficult for us to identify all the Nash equilibria of a given infinitely repeated game. Thus, usually we turn to characterize the set of all possible payoff profiles that correspond to Nash equilibria of the infinitely repeated game. First, we need to introduce the concepts of *enforceable* and *feasible* payoff profiles which would be useful in the characterization of all possible Nash equilibrium payoff profiles of the infinitely repeated games.

Definition 2.5 A payoff profile (r_1, r_2) is enforceable if and only if $r_i \geq v_i, \forall i \in N$, where v_i is agent i 's minimax payoff.

Definition 2.6 A payoff profile (r_1, r_2) is enforceable if and only if there exist nonnegative rational values α_a such that we have $\sum_{a \in A} \alpha_a u_i(a) = r_i, \forall i \in N$, and also $\sum_{a \in A} \alpha_a = 1$.

Based on the previous concepts, we can have the following theorem—folk theorem, which characterizes the set of all possible Nash equilibrium payoff profiles of the infinitely repeated games with limit of means criterion.

Theorem 2.1 *If a payoff profile r of game G is both feasible and enforceable, then it is the payoff profile for some Nash equilibrium of the limit-of-means infinitely repeated game of G .*

2.2 Cooperative Multiagent Systems

2.2.1 Achieving Nash Equilibrium

Convergence to Nash equilibrium has been the most commonly adopted goal to pursue within different multiagent environments in the multiagent learning literature. We review the representative work within this direction and also point out the limitations.

The interactions among agents are usually modeled as two-player repeated (or stochastic) games. In the work of Claus and Boutilier [2], two different types of

learners are distinguished based on Q -learning algorithm, independent learner and joint-action learner; and their performance in the context of two-agent repeated cooperative games was investigated. An independent learner simply learns its Q -values for its individual actions by ignoring the existence of the other agent, while a joint-action learner learns the Q -values for the joint actions. Empirical results show that both types of learners can successfully coordinate on the optimal joint actions in simple cooperative games without significant performance difference. However, both of them fail to coordinate on optimal joint actions when the game structure becomes more complex, i.e., the climbing game (Fig. 4.1a) and the penalty game (Fig. 4.1b). For example, considering the penalty game in which $k = -100$, it has three Nash equilibria of which two of them ((a, a) and (c, b)) are the optimal equilibria to achieve. Due to initial random explorations, both agents will find both actions a and c unattractive since mis-coordination on (a, c) or (c, a) results in significant payoff loss for both agents. In contrast choosing action b is the more preferred option for both agents since nonnegative payoffs can always be guaranteed. After both agents have learned the policy of choosing action b , the outcome will be converged to the nonoptimal equilibrium (b, b) due to the stable nature of equilibria.

A number of improved learning algorithms have been proposed afterward. Lauer and Rienmiller [3] propose the distributed Q -learning algorithm base on the optimistic assumption. Specifically, the agents' Q -values for each action are updated in such a way that only the maximum payoff received by performing this action is considered. Besides, the agents also need to follow an additional coordination mechanism to avoid mis-coordination on suboptimal joint actions. It is proved that optimal joint actions can be guaranteed to achieve if the cooperative game is deterministic; however, it fails when dealing with stochastic environments. The main difficulty dealing with stochastic domains can be intuitively explained as follows: in stochastic environments, apart from the effects of other agents' actions, the stochastic feature of the environment also influences the way of learning the optimal policy. However, the behaviors of the agents are expected to maximize based on the optimistic assumption, while the stochastic influence of the environment has to be taken into consideration with the expected values. When these two aspects of uncertainties are combined together, the agents may wrongly mix the estimated rewards of different actions, thus cannot learn the real maximal reward of each action.

Kapetanakis and Kudenko [4] propose the FMQ heuristic to alter the Q -value estimation function to handle the stochasticity of the games. Under FMQ heuristic, the original Q -value for each individual action is modified by incorporating the additional information of how frequent the action receives its corresponding maximum payoff. Experimental results show that FMQ agents can successfully coordinate on an optimal joint action in partially stochastic climbing games, but fail for fully stochastic climbing games. An improved version of FMQ (recursive FMQ) is proposed in [5]. The major difference with the original FMQ heuristic is that the Q -function of each action $Q(a)$ is updated using a linear interpolation based on the occurrence frequency of the maximum payoff by performing this action and

bounded by the values of $Q(a)$ and $Q_{\max}(a)$. This improved version of FMQ is more robust and less sensitive to the parameter changes; however, it still cannot achieve satisfactory performance in fully stochastic cooperative games.

Matignon et al. [6] introduce the concept of hysteretic learning to facilitate better coordination on optimal Nash equilibrium in stochastic cooperative games. The general idea behind hysteretic learning is that agents should be cautious when they update their Q -values of their actions based on optimistic assumption. Under optimistic assumption, the Q -value of each action is updated only based on the maximal payoff obtained by taking the action, while all lower payoffs are considered as the results of other agents' random explorations and ignored. However, in stochastic environments, the lower payoffs of different actions may also be caused by the stochastic nature of the environment itself. To this end, under hysteretic learning, the authors propose that each agent should adopt two different learning rates to balance between the updates based on optimistic assumption and the update handling stochasticity of the environment. Simulation results show that hysteretic learners are able to successfully cover to optimal Nash equilibrium in partially stochastic climbing games over 80 % of the runs.

Panait et al. [7] propose the lenient multiagent learning algorithm to overcome the noise introduced by the stochasticity of the games. The basic idea is that at the beginning each lenient learner has high leniency (being optimistic), i.e., update the utility of each action based on the maximum payoff received from choosing this action and ignore those penalties occurred by choosing this action. Gradually, the learners will decrease their leniency degrees. Simulation results show that the agents can achieve coordination on the optimal joint action in the fully stochastic climbing game in more than 93.5 % of the runs compared with only around 40 % of the runs under FMQ heuristic.

Matignon et al. [8] review most of the existing independent multiagent reinforcement learning algorithms including decentralized Q -learning [2], distributed Q -learning [3], recursive FMQ [4], lenient Q -learning [7], and hysteretic Q -learning [6] in cooperative Markov games. The strength and weakness of these algorithms are evaluated and discussed in a number of multiagent domains including matrix games, Boutillier's coordination games, predator pursuit domains, and a special multistate game. Their evaluation results show that all of them fail to achieve coordination for fully stochastic games and only recursive FMQ can achieve coordination for 58 % of the runs.

To summarize, most previous work [2–5, 7, 9–11] studies the coordination problem within the framework of two players iteratively playing a single-stage cooperative game. However, little work has been done in an alternative important setting involving a large population of agents in which each agent interacts with another agent randomly chosen from the population each round. The interaction between each pair of agents is still modeled as a single-stage cooperative game, and the payoff each agent receives in each round only depends on the joint action of itself and its interacting partner. Under this framework, each agent learns its policy through repeated interactions with multiple agents, which is termed as *social learning* [12], in contrast to the case of learning from repeated interactions against

the same opponent in the two-agent case. This social learning framework has been adopted to investigate the norm emergence problem in conflict-interest game (e.g., anti-coordination game) [12, 13], and it has been found that the agents can finally learn to follow an optimal norm through social learning. However, it is still not clear a priori whether the agents can learn to coordinate on an optimal joint action in cooperative games under such a social learning framework, which is the major question we are going to investigate in Sect. 4.1.

2.2.2 Achieving Fairness

2.2.2.1 Fairness in Behavioral Economics

The theory of inequity aversion has been used as one of the most popular theories for modeling fairness in behavioral economics [14] with the premise that people not only care about their own payoffs but also the relative payoffs with other people [15, 16]. Based on this theory, Fehr and Schmidt [15] propose an inequity aversion model to explain the phenomena that irrational actions instead of fully self-interested decisions can be made by people to prevent inequitable outcomes. This model is based on the assumption that people show a weak aversion toward advantageous inequity and a strong aversion toward disadvantageous inequity. In their fairness model, for a set of n agents, the utility function of agent i can be formally expressed as follows:

$$U_i(u) = u_i - \alpha_i \frac{1}{n-1} \sum_{j \neq i} \max\{u_j - u_i, 0\} - \beta_i \frac{1}{n-1} \sum_{j \neq i} \max\{u_i - u_j, 0\}. \quad (2.1)$$

Here $U_i(u)$ is the perceived utility of agent i , and $u = \{u_1, u_2, \dots, u_n\}$ with u_i the payoff received by agent i . The second term describes the utility cost when the payoff it receives is lower than that of other agents, and the third term represents the utility cost when the payoff it receives is higher than that of others. The parameters α_i and β_i refer to the weighting factors representing the agent's suffering degree when it receives lower and higher payoff than others, respectively. Usually $\alpha_i \geq \beta_i$ is assumed in this model indicating that the agent suffers more from the inequity when it receives lower payoff than others. Moreover, $0 \leq \beta_i < 1$ is assumed here. By assuming $\beta_i \geq 0$, it is guaranteed that no agent is willing to receive higher payoff than others to increase inequity, and if $\beta_i \geq 1$, then the agent will be ready to throw away part of its own payoff to increase the equity among the agents. The authors have shown that this model can accurately predict human behavior in various games such as ultimatum game and public goods game with punishments.

Another line of research on modeling fairness in behavioral economics is based on the theory of reciprocity [17, 18]. Reciprocity means people tend to show kindness to those people that are kind to themselves and show unkindness to those people that are unkind to themselves. The key distinction with inequity-averse model is that this model also takes the kindness/unkindness intention of each action into account, and fairness is achieved by reciprocal behavior which is regarded as a response to kindness/unkindness instead of a desire for reducing inequity among payoffs.

Both types of models are designed targeting at analyzing single-shot simultaneous (or sequential) games such as ultimatum game and dictator game and have their own merits in explaining certain aspects of human behaviors that are inconsistent with the theory of individual rationality. However both types of models cannot be directly applied to repeated games. We believe that it is equally important to take into account fairness in the context of repeated game since it is a suitable framework for modeling multiagent repeated interactions. This issue will be further discussed and investigated in Chap. 3 (Sect. 3.2.2). Another limitation of previous work is that previous fairness models [15, 17] only take one fairness factor (either inequity aversion or reciprocity) into consideration, which thus only reflects a partial view of the motivations behind fairness. This issue will be further discussed and handled in Chap. 3 (Sect. 3.2.1).

2.2.2.2 Achieving Fairness Through Multiagent Reinforcement Learning

A lot of approaches have been proposed for solving the multiagent learning problem based on reinforcement learning [19–22]. However, the criterion of fairness has not been given much attention in the literature of multiagent reinforcement learning, and we review two representative works in this direction.

Jong et al. [23] argue that it is important for a software agent to exhibit humanlike fairness behaviors since it is inevitable that an agent has to interact with human beings. To this end, they transform the descriptive inequity-averse fairness model from behavioral economic area into a computational model driven by continuous action learning automata for specifying agents' behaviors. A continuous action learning automata can be considered as a reinforcement learner with infinite actions. Each time learning automata select an action and receive a feedback from the environment and update the probability of choosing each action based on the feedback. In this work, the authors incorporate the inequity-averse fairness model into the decision-making process of the continuous action learning automata by mapping the original feedback from the environment to the utility value based on the inequity-averse fairness model each time. The authors apply this computational model into two representative game settings: the ultimatum game and Nash bargaining game. Simulation results show that the agents adopting this computational model can successfully exhibit fair behaviors that are better aligned with actual human behaviors observed in human experiments.

Nowé et al. [24, 25] firstly investigate the problem of how agents are able to coordinate on both fair and efficient outcomes through repeated interactions. To this end, they propose a periodical policy for multiple agents to achieve fair outcomes by incorporating the descriptive inequity-averse fairness model. Their policy involves two periods: reinforcement learning period and communication period. In the reinforcement learning period, the agents adopt Q -learning approach to choose actions with learning rate η . In the communication period, the agent who receives highest accumulative payoff and periodical payoff removes its current action out of its action space. The purpose is to give other agents the opportunity to reach their preferred outcomes. Simulation results show that each agent can obtain approximately equal average payoffs by adopting this periodical policy. However, this policy suffers from two major problems, i.e., stochastic exploration and non-robustness as follows, and this topic will be further investigated in Chap. 3 (Sect. 3.1).

Stochastic exploration Under the periodical policy, the agent whose performance is currently the best will exclude its current action from its action space, and then the agents will begin a new exploration phase using reinforcement learning method at the beginning of the next noncommunication period. The purpose of doing this is to give other agents the opportunity to achieve higher payoffs by achieving different Nash equilibria preferred by different agents. However, this also brings in the side effect that the agents will always explore their action space stochastically at the beginning of each noncommunication period, and thus it is inefficient in terms of the rate of convergence to another new Nash equilibrium. Additionally, it is possible that the agents will converge to an old Nash equilibrium which has been achieved before, and thus this will slow down the rate of achieving fairness between agents.

Non-robustness In the periodical policy, it is assumed that all agents obey the same coordination principle (i.e., excluding the current action if it is the best agent and reexploring its new action subspace at the beginning of each noncommunication period). However, this policy is vulnerable to exploitations of malicious agents when the system is open to outside agents, since the malicious agent can always choose its preferred action without following the coordination principle and its preferred Nash equilibrium will be achieved at the end of each noncommunication period.¹ In other words, the malicious agent always has the incentive to exploit others instead of obeying the coordination rule of the periodical policy. Therefore, fairness between agents cannot be guaranteed to be achieved when this kind of malicious agent exists.

¹The rate of convergence to the Nash equilibrium preferred by the malicious agent depends on the learning rate of the agents, and this Nash equilibrium can always be achieved as long as the length of the noncommunication period is large enough.

2.2.3 Achieving Social Optimality

Learning to achieve coordination on (socially) optimal outcomes in the setting of a population of agents has received a lot of attention in multiagent learning literature [26–31]. Previous work mainly focuses on the coordination problem in two different games, prisoner’s dilemma game (Fig. 2.2) and anti-coordination games (Fig. 2.3), which are representatives of two different types of coordination scenarios. For the first type of games, the agents need to coordinate on the outcomes with identical actions to achieve socially optimal outcomes, while in the second type of games, the achievement of socially optimal outcomes requires the agents to coordinate on outcomes with complementary actions. In previous work, tag has been extensively used as an effective mechanism to bias the interactions among agents in the population. Using tags to bias the interaction among agents in prisoner’s dilemma game is first proposed in [32, 33], and a number of further investigations on the tag mechanism in one-shot or iterated prisoner’s dilemma game are conducted in [34–36].

Hales and Edmonds [26] firstly introduce the tag mechanism originated in other fields such as artificial life and biological science into multiagent systems research to design effective interaction mechanism for autonomous agents. In their model for prisoner’s dilemma game, each agent is represented by $L + 1$ bits. The first bit indicates the agent’s strategy (i.e., playing C or D), and the remaining L bits are the tag bits, which are used for biasing the interaction among the agents and are assumed

Fig. 2.2 An instance of the prisoner’s dilemma game

1’s payoff, 2’s payoff		Player 2’s action	
		C	D
Player 1’s action	C	3, 3	0, 5
	D	5, 0	1, 1

Fig. 2.3 An instance of the anti-coordination game

1’s payoff, 2’s payoff		Player 2’s action	
		C	D
Player 1’s action	C	0, 0	1, 2
	D	2, 1	0, 0

to be observable by all agents. In each generation, each agent is allowed to play the prisoner's dilemma game with another agent owning the same tag string. If an agent cannot find another agent with the same tag string, then it will play the prisoner's dilemma game with a randomly chosen agent in the population. The agents in the next generation are formed via fitness (i.e., the payoff received in the current generation) proportional reproduction scheme. Besides, low level of mutation is applied on both the agents' strategies and tags. This mechanism is demonstrated to be effective in promoting high level of cooperation among agents when the length of tag is large enough. They also apply the same technique to other more complex task domains which shares the same dilemma with prisoner's dilemma game (i.e., the socially rational action cannot be achieved if the agents are making purely individually rational decisions) and show that the tag mechanism can motivate the agents toward choosing socially rational behaviors instead of individually rational ones. However, there are some limitations of this tag mechanism. Since the agents mimic the strategy and tags of other more successful agents and the agents choose the interaction partner based on self-matching scheme, the agents can only play the same strategy with the interacting agents. Thus, socially optimal outcomes can be obtained only in the cases where the achievement of socially optimal outcomes requires the agents to coordinate on identical actions.

Chao et al. [30] propose a tag mechanism similar with the one proposed in [26], and the only difference is that mutation is applied on the agents' tags only. They evaluate the performance of their tag mechanism with a number of learning mechanisms commonly used in the field of multiagent system such as generalized tit-for-tat [30], WSLs (win-stay, loose-shift) [37], basic reinforcement learning algorithm [38], and evolutionary learning. They conduct their evaluations using two games: prisoner's dilemma game and coordination game. Simulation results show that their tag mechanism can have comparable or better performance than other learning algorithms they compared under both games. However, this tag mechanism suffers from similar limitations with Hales and Edmonds' model [26]: It only works when the agents are only needed to learn the identical strategies in order to achieve socially optimal outcomes, and this is indeed the case in both games they evaluated. It cannot handle the case when complementary strategies are needed to achieve coordination on socially optimal outcomes.

McDonald and Sen [27] reinvestigate the Hales and Edmonds' model [26] and the reason why large tags are required to sustain cooperation. The main explanation they come up with is that in order to keep high level of cooperation, sufficient number of cooperative groups is needed to prevent mutation from infecting most of the cooperative groups with defectors at the same time. There are two ways to guarantee the coexistence of sufficient number of cooperative groups, i.e., adopting high tag mutation rate or using large tags, which are all verified by simulation results. The limitation of Hales and Edmonds' model [26] is also pointed out in this paper, and they also provide an example (i.e., anti-coordination game in Fig. 2.3) in which Hales and Edmonds' model fails to promote the agents toward socially optimal behaviors. In anti-coordination game, complementary actions (either (C, D) or (D, C)) are required for agents to coordinate on socially optimal outcomes.

However, under the self-matching tag mechanism in Hales and Edmonds' model, each agent only interacts with others with the same tag strings, and each interacting pair of agents usually shares the same strategy. Therefore, their mechanism only works when identical actions are needed to achieve socially optimal outcomes, while it fails when it comes to games like anti-coordination game.

Considering the limitation of Hales and Edmonds' model, McDonald and Sen [28, 29] propose three new tag mechanisms to tackle the limitation. The first tag mechanism is called tag-matching patterns (one and two sided). Each agent is equipped with both a tag and a tag-matching string, and agent i is allowed to interact with agent j if its tag-matching string matches agent j 's tag in one-sided matching. For two-sided matching, it also requires agent j 's tag-matching string to match agent i 's tag in order to allow the interaction between agent i and j . By introducing this tag mechanism, the agents are allowed to play different strategy with their interacting partners. The second mechanism is payoff sharing mechanism, which requires each agent to share part of its payoff with its opponent. This mechanism is shown to be effective in promoting socially optimal outcomes in both prisoner's dilemma game and anti-coordination game; however, it can be applied only when side payment is allowed. The last mechanism they propose is called paired reproduction mechanism. It is a special reproduction mechanism which makes copies of matching pairs of individuals with mutation at corresponding place on the tag of one and the tag-matching string of the other at the same time. The purpose of this mechanism is to preserve the matching between this pair of agents after mutation in order to promote the survival rate of cooperators. Simulation results show that this mechanism can help in sustaining the percentage of agents coordinating on socially optimal outcomes at a high level in both the prisoner dilemma and anti-coordination games, and this mechanism can be applied in more general multiagent settings where payoff sharing is not allowed. However, similar to Hales and Edmonds' model [26], these mechanisms all heavily depend on mutation to sustain the diversity of groups in the system; thus, this leads to the undesired result that the variation of the percentage of coordination on socially optimal outcomes is very high all the time. Besides, the paired reproduction mechanism is shown to be effective in two types of symmetric games (prisoner's dilemma game and anti-coordination game) only, and it cannot be applied to the case of asymmetric games. The reason is that in asymmetric games, given a pair of interacting agents and their corresponding strategies, each agent may receive different payoffs depending on its current role (row or column agent). For example, considering the asymmetric game in Fig. 2.4, suppose a pair of interacting agents (agent A and agent B) choosing actions D and C , respectively. In this situation, agent A 's payoff is different when it acts as the row (payoff of 2) or column player (payoff of 4), even though both agents' strategies keep unchanged. However, in McDonald and Sen's learning framework, the role of each agent during interaction is not distinguished. They implicitly assume that all agents' roles are always the same in accordance with the definition of symmetric game. Thus, the payoffs of the interacting agents become nondeterministic when it comes to asymmetric games. Therefore, their learning framework is only applicable to those

Fig. 2.4 Asymmetric anti-coordination game: the agents need to coordinate on the outcome (D, C) to achieve socially rational solution

A's payoff, B's payoff		Agent B's action	
		C	D
Agent A's action	C	0, 0	5, 4
	D	2, 15	0, 0

games that are agent symmetric (prisoner's dilemma game and anti-coordination game), i.e., the roles of the agents have no influence on their payoffs received.

To summarize, previous work [26, 28, 29] mainly focuses on modifying the interaction protocol from random interaction to tag-based interaction to evolve coordination on socially optimal outcomes. However, the agents' decision-making processes in the previous work are based on evolutionary learning (i.e., imitation and mutation) following evolutionary game theory, and thus the deviation of percentage of coordination is significant even though the average coordination rate is good. Another limitation of previous approaches is that it is explicitly assumed that each agent has access to all other agents' information (i.e., their payoffs and actions) in previous rounds, which may be unrealistic in many practical situations due to communication limitations. Finally, all previous work only focuses on two symmetric games, prisoner's dilemma game and anti-coordination game, and cannot be directly applied in the settings of asymmetric games. We will further discuss how to handle these limitations in Chap. 4 (Sect. 4.2).

2.3 Competitive Multiagent Systems

2.3.1 Achieving Nash Equilibrium

Littman [39] considers the strictly competitive environment in which there are two agents (agent i and j) and they have diametrically opposed goals. In this work, Markov game formalism is introduced as the mathematical framework for reasoning about the two-agent interaction environments. Markov games can be considered as the natural generalization of Markov decision process (MDP) and repeated game. Within this framework, the minimax- Q learning algorithm is proposed by extending the original Q -learning algorithm in MDP to the framework of Markov games. The main difference between minimax- Q learning and Q -learning is that the "max" operator used in the update step of Q -learning is replaced with the "minimax" operator, which is justified by the underlying assumption that the opponent agent j always behaves to cause the greatest harm to agent i . The agents

adopting minimax- Q learning algorithm are theoretically guaranteed to converge to Nash equilibrium strategy profile in any two-player zero-sum Markov games under self-play, provided that every action and state are sampled infinitely often. However, when it comes to general-sum games, each agent may not be interested in minimizing its opponent(s)'s payoffs; thus, pursuing the solution concept of Nash equilibrium becomes less justified. An alternative learning goal is to maximize each agent's individual payoffs, which has been commonly adopted in the literature, and its related work will be reviewed next section.

2.3.2 *Maximizing Individual Benefits*

In competitive multiagent systems, agents usually aim at maximizing their individual benefits only. One of the most common and important practical multiagent competitive interaction scenarios is multiagent negotiation, which is our focus in this book. During negotiation, different negotiating parties usually have conflicting interests with each other and are only interested in obtaining as much utility as possible through negotiation. Until now great efforts have been devoted to develop efficient negotiation strategies for automated negotiating agents in the literature. The most typical approach existing in the literature is concession-based strategy such as ABMP strategy [40]. The ABMP strategy agent decides on the next move based on its own utility space only and makes concession to its negotiating partners according to certain concession pattern. However, since the dynamics and influences of the negotiating partners are not taken into consideration, the weakness of this type of strategies is that it is difficult to reach those mutually beneficial outcomes and thus is inefficient in complex negotiation scenarios.

To overcome the aforementioned limitation, various techniques have been proposed to model certain aspects of the negotiation scenario to improve the efficiency of the negotiation outcome such as the opponent's preference profile and the knowledge of the negotiation domain [41–44]. Saha et al. [42] propose a learning mechanism using Chebychev's polynomials to approximately model the negotiating opponent's decision function in the context of repeated two-player negotiations. They prove that their algorithm is guaranteed to converge to the actual probability function of the negotiating partner under infinite sampling. Experiments also show that the agent using their learning mechanism can outperform other simple learning mechanisms and also be robust to noisy data. However, in their approach, it is assumed that the agents negotiate over one indivisible item (price) only, thus is not applicable to more general multi-issue negotiation scenarios.

Hindriks and Tykhonov [43] propose a Bayesian learning-based technique to model the negotiating opponent's private preference in the context of bilateral multi-issue negotiations. They test this technique on several negotiation domains and show that it can improve the efficiency of the bidding process by incorporating it into a negotiation strategy. One application of this modeling technique is that it

is integrated into a negotiation strategy used by agent *TheNegotiator* [45] which participated in ANAC 2011 [45].

Brzostowski and Kowalczyk [46] propose a mechanism for predicting the negotiating partner's future behaviors based on the difference method. Based on the prediction results, the negotiation can be modeled as multi-stage control process, and the task of determining the optimal next-step offer is equivalent to the problem of determining the sequence of optimal control. Simulation results show that the agents using their mechanism can greatly outperform the classical approach in terms of utilities. However, their mechanism is only applicable in the single-item negotiation scenario. Besides, the underlying assumption of their mechanism is that the negotiation partner's strategy is the combination of time-dependent and behavior-dependent tactics, and thus may not work well against other types of negotiation partners.

To summarize, the major difficulty in designing automated negotiation agent is how to achieve optimal negotiation results given incomplete information on the negotiating partner. The negotiation partner usually keeps its negotiation strategy and its preference as its private information to avoid exploitations. A lot of research efforts have been devoted to better understand the negotiation partner by either estimating the negotiation partner's preference profile [43, 47, 48] or predicting its decision function [46, 49]. On one hand, with the aid of different preference profile modeling techniques, the negotiating agents can get a better understanding of their negotiating partners and thus increase their chances of reaching mutually beneficial negotiation outcomes. On the other hand, effective strategy prediction techniques enable the negotiating agents to maximally exploit their negotiating partners and thus receive as much benefit as possible from negotiation.

However, in most of previous work, the above two aspects are often investigated separately, and little efforts have been devoted to combine them together and evaluate the negotiation performance of various combinations of different techniques. To this end, in recent years a number of negotiation strategies, which take advantage of existing techniques from both aspects as previously mentioned, have been proposed, and agents employing these strategies have participated in *Automated Negotiating Agents Competition (ANAC)* [45, 50]. During the competitions, their performance has been extensively evaluated in a variety of multi-issue negotiation scenarios, and valuable insights have been obtained in terms of the advantages and disadvantages of different techniques, e.g., the efficacy of different acceptance conditions [44]. It is still an open and interesting problem to design more efficient automated negotiation strategies against a variety of negotiating opponents in different negotiation domains. We will further discuss this topic in Chap. 5.

2.3.3 Achieving Pareto-Optimality

Compared with the goal of Nash equilibrium and maximizing individual benefits, adopting Pareto-optimal solution has the advantage that it can prevent the agents

from achieving loss-loss outcomes and also can improve all agents' utilities and the system level's utility most of the times. One well-known example is the prisoner's dilemma game (Fig. 2.2), in which if both agents only seek to maximize their individual payoffs, it will result in the inefficient Nash equilibrium (i.e., mutual defection). However, there exists a Pareto-optimal outcome under which both agents' utilities can be increased (mutual cooperation).

A number of approaches [51–53] have been proposed targeting at the prisoner's dilemma game only, and the learning goal is to achieve Pareto-optimal solution of mutual cooperation instead of Nash equilibrium solution of mutual defection. Besides, there also exists some work [54, 55] which addresses the problem of achieving Pareto-optimal solution in the context of general-sum games as follows. Sen et al. [54] propose an innovative expected utility probabilistic learning strategy by incorporating action revelation mechanism. Simulation results show that agents using action revelation strategy under self-play can achieve Pareto-optimal outcomes which dominate Nash equilibrium in certain games, and also the average performance with action revelation is significantly better than Nash equilibrium solution over a large number of randomly generated game matrices. Banerjee et al. [55] propose the conditional joint action learning (CJAL) strategy under which each agent takes into consideration the probability of an action taken by its opponent given its own action and utilizes this information to make its own decision. Simulation results show that agents adopting this strategy under self-play can learn to converge to the Pareto-optimal solution of mutual cooperation in prisoner's dilemma game when the game structure satisfies certain condition. However, all previous strategies are based on the assumption of self-play, and there is no guarantee of the performance against the opponents using other strategies.

In competitive environments, agents may not have the incentive to follow the same strategy specified by the system designer. To this end, a number of work [1, 56–58] assumes that the opponent may adopt different rational strategies and investigate the problem of how these agents can be incentivized to coordinate on Pareto-optimal outcomes. One representative work is folk theorem [1] in the literature of game theory. The basic idea of folk theorem is that there are some strategies based on punishment mechanism which can enforce desirable outcomes and are also in Nash equilibrium, assuming that all players are perfectly rational. From the multiagent learning perspective, the focus is to utilize the ideas in folk theorem to design efficient strategy against adaptive best-response opponents. Also the strategies we explore need not be in equilibrium in the strict sense, since it is very difficult to construct a strategy which is the best response to a particular learning strategy such as Q -learning. A number of teacher strategies [56, 57] have been proposed to induce better performance from the opponents via punishment mechanism, assuming that the opponents adopt best-response strategies such as Q -learning. Based on the teacher strategy, Goldfather++ [57], Crandall, and Goodrich [58] propose the strategy SPaM employing both teaching and following strategies and show its better performance in the context of two-player games against a number of best-response learners. The goal of the teaching component is to encourage its opponent to behave in the way desirable for the SPaM agent, and the goal of the

following component is to guarantee that the SPaM agent can still achieve as high payoff as possible by exploiting its opponent while teaching its opponent at the same time.

Following this direction, we target at a more refined solution concept—socially optimal outcome, which represents those Pareto-optimal outcomes that are also socially optimal, i.e., maximizing the sum of all agents' payoffs involved. We investigate the problem of achieving socially optimal outcomes in competitive environments within two different types of learning contexts: two-player repeated interaction framework and a population of agents interacting with each other, which will be introduced in details in Chap. 6.

References

1. Osborne MJ, Rubinstein A (1994) A course in game theory. MIT, Cambridge
2. Claus C, Boutilier C (1998) The dynamics of reinforcement learning in cooperative multiagent systems. In: Proceedings of AAAI'98, Madison, pp 746–752
3. Lauer M, Riemmiller M (2000) An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In: Proceedings of ICML'00, Stanford, pp 535–542
4. Kapetanakis S, Kudenko D (2002) Reinforcement learning of coordination in cooperative multiagent systems. In: Proceedings of AAAI'02, Edmonton, pp 326–331
5. Matignon L, Laurent GJ, Le For-Piat N (2008) A study of FMQ heuristic in cooperative multi-agent games. In: AAMAS'08 workshop: MSDM, Estoril, pp 77–91
6. Matignon L, Laurent GJ, Le Fort-Piat N (2007) Hysteretic Q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In: Proceedings of IEEE/RSJ international conference on intelligent robots and systems (IROS), San Diego, pp 64–69
7. Panait L, Sullivan K, Luke S (2006) Lenient learners in cooperative multiagent systems. In: Proceedings of AAMAS'06, Hakodate, pp 801–803
8. Matignon L, Laurent GJ, Le For-Piat N (2012) Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *Knowl Eng Rev* 27:1–31
9. Panait L, Luke S (2005) Cooperative multi-agent learning: the state of the art. *Auton Agents Multi-agent Syst* 11(3):387–434
10. Wang X, Sandholm T (2002) Reinforcement learning to play an optimal nash equilibrium in team markov games. In: Proceedings of NIPS'02, Vancouver, pp 1571–1578
11. Brafman RI, Tennenholtz M (2004) Efficient learning equilibrium. *Artif Intell* 159:27–47
12. Sen S, Airiau S (2007) Emergence of norms through social learning. In: Proceedings of IJCAI'07, Hyderabad, pp 1507–1512
13. Villatoro D, Sabater-Mir J, Sen S (2011) Social instruments for robust convention emergence. In: Proceedings of IJCAI'11, Barcelona, pp 420–425
14. Camerer CF, Loewenstein G, Rabin M (2003) Advances in behavioral economics. Russell Sage Foundation Press/Princeton University Press, New York/Princeton
15. Fehr E, Schmidt KM (1999) A theory of fairness, competition and cooperation. *Q J Econ* 114:817–868
16. Bolton GE, Ockenfels A (2000) Erc-a theory of equity, reciprocity and competition. *Am Econ Rev* 90:166–193
17. Rabin M (1993) Incorporating fairness into game theory and economics. *Am Econ Rev* 83:1281–1302
18. Falka A, Fischbacher U (2006) A theory of reciprocity. *Games Econ Behav* 54:293–315

19. Littman M (1994) Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of ICML'94, New Brunswick, pp 322–328
20. Kapetanakis S, Kudenko D (2002) Reinforcement learning of coordination in cooperative multi-agent systems. In: Proceedings of AAAI'02, Edmonton, pp 326–331
21. Bowling M, Veloso M (2002) Multiagent learning using a variable learning rate. *Artif Intell* 136:215–250
22. Hu J, Wellman M (1998) Multiagent reinforcement learning: theoretical framework and an algorithm. In: Proceedings of ICML'98, Madison, pp 242–250
23. de Jong S, Tuyls K, Verbeeck K (2008) Artificial agents learning human fairness. In: Proceedings of AAMAS'08, Estoril. ACM, pp 863–870
24. Nowé A, Parent J, Verbeeck K (2001) Social agents playing a periodical policy. In: Proceedings of ECML'01, vol 2176. Springer, Berlin/New York, pp 382–393
25. Verbeeck K, Nowé A, Parent J, Tuyls K (2006) Exploring selfish reinforcement learning in repeated games with stochastic rewards. *Auton Agents Multi-agent Syst* 14:239–269
26. Hales D, Edmonds B (2003) Evolving social rationality for mas using “tags”. In: Proceedings of AAMAS'03, pp 497–503. ACM, New York
27. Matlock M, Sen S (2005) The success and failure of tag-mediated evolution of cooperation. In: Proceedings of the first international workshop on learning and adaption in multi-agent systems. Springer, Berlin/New York, pp 155–164
28. Matlock M, Sen S (2007) Effective tag mechanisms for evolving coordination. In: Proceedings of AAMAS'07, Honolulu, p 251
29. Matlock M, Sen S (2009) Effective tag mechanisms for evolving cooperation. In: Proceedings of AAMAS'09, Budapest, pp 489–496
30. Chao I, Ardaiz O, Sanguesa R (2008) Tag mechanisms evaluated for coordination in open multi-agent systems. In: Proceedings of 8th international workshop on engineering societies in the agents world, Athens, pp 254–269
31. Hao JY, Leung HF (2011) Learning to achieve social rationality using tag mechanism in repeated interactions. In: Proceedings of ICTAI'11, Boca Raton, pp 148–155
32. Holland JH, Holyoak K, Nisbett R, Thagard P (1986) *Induction: processes of inferences, learning, and discovery*. MIT, Cambridge
33. Allison PD (1992) The cultural evolution of beneficent norms. *Soc Forces* 71(2):279–301
34. Riolo R, Cohen MD (2001) Cooperation without reciprocity. *Nature* 414(6862):441–443
35. Hales D (2000) Cooperation without space or memory-tag, groups and the prisoner's dilemma. In: *Multi-agent-based simulation*, Boston
36. Howley E, O'Riordan C (2005) The emergence of cooperation among agents using simple fixed bias tagging. In: *IEEE congress on evolutionary computation*, Edinburgh
37. Nowak M, Sigmund K (1993) A strategy of winstay, lose-shift that outperforms tit-for-tat in the prisoner's dilemma game. *Nature* 364(6432):56–58
38. Wakano JY, Yamamura N (2001) A simple learning strategy that realizes robust cooperation better than pavlov in iterated prisoner's dilemma. *J Ethol* 19:9–15
39. Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of ICML'94, New Brunswick, pp 157–163
40. Jonker C, Robu V, Treur J (2006) An agent architecture for multi-attribute negotiation using incomplete preference information. *Auton Agent Multi-agent Syst* 15(2):221–252
41. Faratin P, Sierra C, Jennings NR (2003) Using similarity criteria to make negotiation trade-offs. *Artif Intell* 142(2):205–237
42. Saha S, Biswas A, Sen S (2005) Modeling opponent decision in repeated one-shot negotiations. In: Proceedings of AAMAS'05, Utrecht, pp 397–403
43. Hindriks K, Tykhonov D (2008) Opponent modeling in automated multi-issue negotiation using Bayesian learning. In: Proceedings of AAMAS'08, Estoril, pp 331–338
44. Baarslag T, Hindriks K, Jonker C (2011) Acceptance conditions in automated negotiation. In: *Proceedings of ACAN'11*, Taipei

45. Baarslag T, Fujita K, Gerding EH, Hindriks K, Ito T, Jennings NR, Jonker C, Kraus S, Lin R, Robu V, Williams CR (2013) Evaluating practical negotiating agents: results and analysis of the 2011 international competition. *Artif Intell* 198:73–103
46. Jakub B, Ryszard K (2006) Predicting partner's behaviour in agent negotiation. In: *Proceedings of AAMAS'06*, Hakodate, pp 355–361
47. Zeng D, Sycara K (1998) Bayesian learning in negotiation. *Int J Hum Comput Syst* 48:125–141
48. Coehoorn RM, Jennings NR (2004) Learning an opponent's preferences to make effective multi-issue negotiation trade-offs. In: *Proceedings of ICEC'04*, Delft, pp 59–68
49. Zeng D, Sycara K (1996) Bayesian learning in negotiation. In: *AAAI symposium on adaptation, co-evolution and learning in multiagent systems*, Portland, pp 99–104
50. Baarslag T, Hindriks K, Jonker C, Kraus S, Lin R (2010) The first automated negotiating agents competition (ANAC 2010). In: Ito T, Zhang M, Robu V, Fatima S, Matsuo T (eds) *New trends in agent-based complex automated negotiations*. Springer, Berlin/Heidelberg, pp 113–135
51. Stimpson JL, Goodrich MA, Walters LC (2001) Satisficing and learning cooperation in the prisoner's dilemma. In: *Proceeding of IJCAI'01*, Seattle, pp 535–540
52. Moriyama K (2007) Utility based Q-learning to maintain cooperation in prisoner's dilemma game. In: *Proceedings of IAT'07*, Silicon Valley, pp 146–152
53. Moriyama K (2008) Learning-rate adjusting Q-learning for prisoner's dilemma games. In: *Proceedings of WI-IAT'08*, Sydney, pp 322–325
54. Sen S, Airiau S, Mukherjee R (2003) Towards a pareto-optimal solution in general-sum games. In: *Proceedings of AAMAS'03*, Melbourne, pp 153–160
55. Banerjee D, Sen S (2007) Reaching pareto optimality in prisoner's dilemma using conditional joint action learning. In: *Proceedings of AAMAS'07*, Honolulu, pp 211–218
56. Littman ML, Stone P (2001) Leading best-response strategies in repeated games. In: *IJCAI workshop on economic agents, models, and mechanisms*, Seattle
57. Littman ML, Stone P (2005) A polynomial time nash equilibrium algorithm for repeated games. *Decis Support Syst* 39:55–66
58. Crandall JW, Goodrich MA (2005) Learning to teach and follow in repeated games. In: *AAAI workshop on multiagent learning*, Pittsburgh

Interactions in Multiagent Systems: Fairness, Social
Optimality and Individual Rationality

Hao, J.; Leung, H.-f.

2016, IX, 178 p. 122 illus., 11 illus. in color., Hardcover

ISBN: 978-3-662-49468-4