Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence

The Dynamics of Reinforcement Social Learning in Cooperative Multiagent Systems

Jianye Hao and Ho-fung Leung
Department of Computer Science and Engineering
The Chinese University of Hong Kong
{jyhao,lhf}@cse.cuhk.edu.hk

Abstract

Coordination in cooperative multiagent systems is an important problem in multiagent learning literature. In practical complex environments, the interactions between agents can be sparse, and each agent's interacting partners may change frequently and randomly. To this end, we investigate the multiagent coordination problems in cooperative environments under the *social learning* framework. We consider a large population of agents where each agent interacts with another agent randomly chosen from the population in each round. Each agent learns its policy through repeated interactions with the rest of agents via social learning. It is not clear a priori if all agents can learn a consistent optimal coordination policy in such a situation. We distinguish two types of learners: *individual action learner* and *joint action learner*. The learning performance of both learners are evaluated under a number of challenging cooperative games, and the influence of the information sharing degree on the learning performance is investigated as well.

1 Introduction

In multiagent systems (MASs), one important ability of an agent is to be able to coordinate effectively with other agents towards desirable outcomes, since the outcome not only depends on the action it takes but also the actions taken by other agents that it interacts with. One central and widely studied class of problem is how to coordinate within cooperative MASs in which the agents share common interests [Panait and Luke, 2005; Matignon *et al.*, 2012]. In cooperative MASs, the agents share common interests and the same reward function; the increase in individual satisfaction also results in the increase in the satisfaction of the group.

There are a number of challenges the agents have to face when learning in cooperative MASs. One major difficulty is the *equilibrium selection problem* [Fulda and Ventura, 2007], i.e., multiple optimal joint actions exist under which coordinations between the agents are required in selecting among multiple optimal joint actions. Another issue is the *stochasticity problem* [Matignon *et al.*, 2012], i.e., the game itself can be non-deterministic. In this case, the challenge is that the agents may need to distinguish whether the different payoffs received by performing an action are caused by the explorations of the other agent or the stochasticity of the game itself.

Until now, various multiagent reinforcement learning algorithms [Claus and Boutilier, 1998; Lauer and Riedmiller, 2000; Kapetanakis and Kudenko, 2002; Wang and Sandholm, 2002; Brafman and Tennenholtz, 2004; Panait *et al.*, 2006; Matignon *et al.*, 2012] have been proposed to solve the coordination problem in cooperative MASs. Most of previous works heavily rely on the Q-learning algorithm [Watkins and Dayan, 1992] as the basis, and can be considered as various modifications of single agent Q-learning algorithms to cooperative multiagent environments. The commonly adopted learning framework for studying the coordination problem within cooperative MASs is to consider two (or more) players playing a repeated (stochastic) game, in which the agents learn their optimal coordination policies through repeated interactions with the same opponent(s) [Panait and Luke, 2005; Matignon *et al.*, 2012]. However, in practical complex environments, the interactions between agents can be sparse, i.e., it is highly likely that each agent may not have the opportunity to always interact with the same partner, and its interacting partners may change frequently and randomly. In this situation, an agent's policy that achieved coordination on an optimal joint action with one partner may fail when it comes to a different partner next time. Each agent learns its policy through repeated interactions with different opponents, and this kind of learning is termed as *social learning*[1] [Sen and Airiau, 2007] to distinguish from the case of learning from repeated interactions with the same partner(s). Previous work [Sen and Airiau, 2007; Villatoro *et al.*, 2011] has investigated the emergence of consistent coordination policies (norms) in MASs (e.g., conflicting-interest games) under this social learning framework, however, little work has been done in studying the coordination problem in cooperative environments under such a framework. Apart from the difficulties previously mentioned in the cooperative environments where the interacting partners are fixed, achieving optimal coordinations under the social learning framework can be more

[1] It's worth mentioning that there also exists a huge body of literature from the area of economics and social science (e.g., [Rendell *et al.*, 2010; Sigmund *et al.*, 2010]), in which the definition of social learning is a little different from what we use here.

challenging due to the additional stochasticity introduced by non-fixed interacting partners. It is not clear a priori if all the agents can still learn an optimal coordination policy in such a situation.

To this end, in this paper, we study the multiagent coordination problem within a large population of agents, where each agent interacts with another agent randomly selected from the population during each round. The interactions between each pair of agents are modeled as two-player cooperative games. Each agent learns its policy concurrently over repeated interactions with randomly selected agents from the population. Different from the fixed agents repeated interactions framework, the agents can have more information at their disposal under the social learning framework, i.e., the learning experience of other pairs of agents from the population. Therefore the agents can learn from both their own interacting experience and the experience from other pairs of agents in each round. We distinguish two different types of learners depending on the amount of information the agents can perceive on the basis of Q-learning algorithm: individual action learners (IALs) and joint action learners (JALs). IALs learn the values of each individual action by ignoring the existence of their interacting partners, while JALs learn the values of the joint actions of themselves and their interacting partners. Both optimistic assumption and the FMQ heuristic are incorporated into IALs and JALs by utilizing the learning experience of their own and other pairs of interacting agents. We investigate the learning performance of both types of learners under the testbed of Claus and Boutilier's coordination games and also the more challenging stochastic variants. Our experimental results and analysis throw light on the learning dynamics of both types of learners via social learning. Specifically it is shown that IALs can learn to fully coordinate on optimal joint action(s) in most of the games, while the JALs' performance is even better, i.e., full coordination on optimal joint action can be achieved even in the fully stochastic game. We also investigate the influence of the amount of information from other pairs of agents available to each agent on the overall performance of the system, which is the key difference from the framework involving repeated interactions among fixed agents.

The remainder of the paper is organized as follows. In Section 2, we give an overview of previous work of multiagent reinforcement learning in cooperative MASs. In Section 3, the social learning framework and both IALs and JALs are described. In Section 4, we present the learning performance of both types of learners in different cooperative games. Lastly conclusion and future work are given in Section 5.

2 Related Work

There has been a significant amount of research in the multiagent reinforcement learning literature on solving coordination problem in cooperative MASs modeled as two-player cooperative repeated (or stochastic) games. In [Claus and Boutilier, 1998], two different types of learners (without optimistic exploration) are distinguished based on Q-learning algorithm: independent learner and joint-action learner, and investigate their performance in the context of two-agent repeated cooperative games. 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 and the penalty game.

A number of improved learning algorithms have been proposed afterwards. [Lauer and Riedmiller, 2000] 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.

Kapetanakis and Kudenko [2002] 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 [Matignon *et al.*, 2008]. 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.

Panait et al. [2006] propose the lenient multiagent learning algorithm to overcome the noise introduced by the stochasticity of the games. The basic idea is that initially each learner has high leniency (optimism) by updating each action's utility based on the maximum payoff received from choosing this action while ignore those penalties occurred by choosing this action. Gradually, the learners 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 around 40% of the runs under FMQ heuristic.

Matignon et al. [2012] review all existing independent multiagent reinforcement learning algorithms in cooperative MASs, and evaluate and discuss their strength and weakness. 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. In this work, we consider the coordination problem in cooperative MASs under the more practical *social learning* framework, where we show that, by introducing *social learning*, full co-

ordinations in fully stochastic games can be achieved as well.

3 Social Learning Framework

Under the social learning framework, there are a population of n agents and each agent learns its policy through repeated pairwise interactions with the rest of agents in the population. The interaction between each pair of agents is modeled as a two-player cooperative game. During each round, each agent interacts with a randomly chosen agent from the population, and one agent is randomly assigned as the row player and the other agent as the column player. The agents are assumed to know their roles, i.e., either as row player or column player, during each interaction. At the end of each round, each agent updates its policy based on the learning experience it receives from the current round. The overall interaction protocol under the social learning framework is presented in Algorithm 1.

Algorithm 1 Overall interaction protocol of the social learning framework

```
for a number of rounds do
    repeat
        two agents are randomly chosen from the population, and one of them is assigned as the row player and the other one as the column player.
        both agents play a two-player cooperative game by choosing their actions independently and simultaneously.
    until all agents in the population have been selected
    for each agent in the population do
        update its policy based on its experience in the current round
    end for
end for
```

Under the social learning framework, we identify two classes of learners, *individual action learners* (IALs) and *joint action learners* (JALs), depending on the amount of information each agent can have access to. Under the social learning framework, the agents can have more information at their disposal since they can also learn from the peers from other groups. Accordingly we will first introduce the two different learning settings in terms of the amount of information each agent can perceive in Section 3.1. Following that, both types of learners will be described in details in Section 3.2.

3.1 Observation

Under the social learning framework, each agent interacts with another agent randomly chosen within the population during each round. We define each pair of interacting agents as being in the same group. For the framework involving two (multiple) agents repeatedly playing a cooperative game (within the same group), the learning experience of each agent comes from its own group only (local interaction). However, under the social learning framework, since all agents interact with their own interacting partners concurrently, different agents may be exposed to experience of those agents from other groups. Allowing agents to observe the information of other agents outside their direct interactions may result in a faster learning rate and facilitate coordination on optimal solutions. One extreme case would be that each agent can have access to the experience of all the other agents in the population (global observation). However, this assumption could be unrealistic in practical scenarios, since the agents are usually physically distributed and the cost of communication among agents is often significant. To trade off between global observation and local interaction, here we allow each agent to have access to the information of other M groups randomly chosen in the system at the end of each round. If the value of M is equal to the total number of groups existing in the population, it becomes equivalent with the case of global observation; if the value of M is zero, it reduces to the case of local interaction. The value of M represents the connectivity degree in terms of information sharing among different groups. By spreading this information into the population with $M \ll N$, it serves as an alternative biased exploration mechanism to accelerate the agents' convergence to optimal joint actions while keeping communication overhead at a low level at the same time. Similar mechanism has recently been proposed as a useful social instrument for facilitating the emergence of convention in MASs [Villatoro *et al.*, 2011], but it is still unknown whether it can be beneficial for coordinating on optimal joint action in cooperative games.

We identify two different learning settings depending on the amount of information that each agent can perceive under the social learning framework. In the first setting, apart from its own action and payoff, each agent can also observe the actions and payoffs of all agents with the same role as itself from other M groups. Formally, the information that each agent i can perceive at the end of each round t can be represented as the set $S_i^t = \{\langle a_i^t, r^t\rangle, \langle b_1^t, r_1^t\rangle, \langle b_2^t, r_2^t\rangle, \ldots, \langle b_M^t, r_M^t\rangle\}$. Here $\langle a_i^t, r^t\rangle$ is the action and payoff of agent i itself and the rest are the actions and payoffs of those agents with the same role from other M groups in the system. This setting is parallel to the independent learning setting in the fixed-agent repeated interaction framework [Claus and Boutilier, 1998], and can be considered as its extension to the social learning framework.

The second setting is a natural extension of the joint action learning setting from fixed-agent repeated interaction framework to the social learning framework. Apart from the same information available in the first setting, each agent is also assumed to able to perceive the action choices of its interacting partner and those agents with opposite role from other M groups. Formally, the experience for each agent i at the end of each round t can be denoted as the set $P_i^t = \{\langle (a_i^t, a_j^t), r^t\rangle, \langle (b_1^t, c_1^t), r_1^t\rangle, \langle (b_2^t, c_2^t), r_2^t\rangle, \ldots, \langle (b_M^t, c_M^t), r_M^t\rangle\}$. Here $\langle (a_i^t, a_j^t), r^t\rangle$ is the joint action and payoff of agent i and its interacting partner and the rest are the joint actions and payoffs of other M groups.

3.2 Learning Strategy

In general, to achieve coordination on optimal joint actions, an agent's behaviors as the row or column player can be the same or different depending on the characteristics of the game and its opponent. Accordingly we propose that each agent owns a pair of strategies, one used as the row player and the other used as the column player, to play with any other agent from the population. The strategies we develop here are natural extensions of the Q-learning techniques from fixed-agent

repeated interaction framework to the social learning framework. There are two distinct ways of applying Q-learning to the social learning framework depending on the setting that the agents are situated in.

Individual Action Learner

In the first setting, each agent can only perceive the actions and payoffs of itself and those agent with the same role from other M groups. Thus it is reasonable for each agent to ignore the existence of its interacting partner and those agents with the opposite role from other M groups and consider them as part of the environment. Each agent has two possible states corresponding to its roles as the row player or column player, and each agent knows its current role during each interaction. Naturally each agent holds a Q-value $Q(s, a)$ for each action a under each state $s \in \{Row, Column\}$, which keeps record of action a's past performance and serves as the basis for making decisions. At the end of each round t, each agent i picks an action (randomly choosing one action in case of a tie) with the highest payoff from the set S_i^t, and updates this action's Q-value following Equation 1,

$$Q_i^{t+1}(s, a) = Q_i^t(s, a) + \alpha^t(s)[r(a) * freq(a) - Q_i^t(s, a)] \quad (1)$$

where $r(a)$ is the highest payoff of action a among all elements in set S_i^t, $freq(a)$ is the frequency that action a occurs with the highest reward $r(a)$ among set S_i^t, and $\alpha^t(s)$ is the current learning rate in state s.

Intuitively, the above update rule incorporates both the optimistic assumption and the FMQ heuristic. On one hand, this update rule is optimistic since we only update the Q-values of those actions that receive the highest payoff based on the current round's experience. On the other hand, similar to the FMQ heuristic, the update rule also takes into consideration the information of how frequent each action can receive the highest payoff based on the current round's experience. However, it is worth noting that under the social learning framework, both heuristics are applied on each agent's single-round learning experience, while in previous work, both heuristics are applied on each agent's multiple-round learning experience in the history.

Each agent chooses its action based on the corresponding set of Q-values during each interaction according to the ϵ-greedy mechanism. Specifically each agent chooses its action with the highest Q-value with probability $1 - \epsilon$ to exploit the action with best performance currently (randomly selection in case of a tie), and makes random choices with probability ϵ for the purpose of exploring new actions with potentially better performance.

Joint Action Learner

In the second setting, each agent can have access to the joint actions of its own group and other M groups as well. Accordingly, each agent can learn the Q-values for each joint action in contrast to learning Q-values for individual actions only in the first setting. Specifically, at the end of each round t, each agent i updates its Q-values for each joint action $\overrightarrow{a}$ belonging to the set P_i^t as follows,

$$Q_i^{t+1}(s, \overrightarrow{a}) = Q_i^t(s, \overrightarrow{a}) + \alpha^t(s)[r(\overrightarrow{a}) - Q_i^t(s, \overrightarrow{a})] \quad (2)$$

where $r(\overrightarrow{a})$ is the payoff of agent i or an agent with the same role as agent i under the joint action $\overrightarrow{a}$ and $\alpha^t(s)$ is its current learning rate under state s.

After enough explorations, we can see that the above Q-values can reflect the expected performance of each joint action, but each agent still needs to determine the relative performance its individual actions to make wise decisions. At the end of each round t, for each action a, let us define $r_a^{max}(s) = max\{Q_i^{t+1}(s, (a, b)) \mid b \in A_i\}$, and denote the corresponding opponent's action as $b^{max}(a)$. The value of $r_a^{max}(s)$ reflects the maximum possible expected payoff that agent i can obtain from performing action a under the current state s. However, agent i's actual expected payoff of performing action a also closely depends on the action choices of its interacting partners. To take this factor into consideration, each agent i also maintains the belief of the frequency of its interacting partners performing action b when it chooses action a, which is denoted as $freq_i(a, b)$. The value of $freq_i(a, b)$ can be easily estimated from agent i's current round experience P_i^t as follows,

$$freq_i(a, b) = \frac{|\{\langle (a, b), r\rangle \mid \langle (a, b), r\rangle \in P_i^t\}|}{|\{\langle (a, y), r\rangle \mid \langle (a, y), r\rangle \in P_i^t, y \in A_i\}|} \quad (3)$$

Finally, each agent i assesses the relative performance $EV(s, a)$ of an action a under the current state s as follows,

$$EV(s, a) = r_a^{max}(s) \times freq_i(a, b^{max}(a)) \quad (4)$$

Similar to the case of IALs, the way of evaluating the relative performance of each individual action incorporates not only the optimistic assumption but also the information of the frequency that the maximum payoff can be received by performing this action based on the current round experience. Based on the EV-values $EV(s, \cdot)$ of its individual actions, each agent chooses its action in a similar way as it would use Q-values in Section 3.2 following the ϵ-greedy mechanism.

4 Experimental Results

In this section, we evaluate the learning performance of both IALs and JALs under the social learning framework. All the experiments are conducted in a population of 100 agents[2], and the default value of M is set to 5. We first evaluate the performance of both IALs and JALs in deterministic environments by focusing on two difficult deterministic cooperative games, which captures all the difficulties (i.e., high miscoordination penalty and the existence of multiple optimal joint actions) for coordination in deterministic cooperative games. Following that, we turn to the performance evaluation in stochastic environments, and specifically we consider two variants of the climbing games discussed in [Kapetanakis and Kudenko, 2002]. To better understand the influence of social learning, we also investigate the influences of the value of M on the coordination performance.

4.1 Performance Evaluation

Deterministic games

We consider two particularly difficult coordination problems: the climbing game (see Figure 1(a)) and the penalty game

[2]For all agents, the initial learning rate α is 0.8 and te exploration rate ϵ is initially set to 0.4. Both values are decreased gradually.

with $k = -50$ (see Figure 1(c)). The climbing game has one optimal joint action (a, a) and two joint actions (a, b) and (b, a) with high penalties. The high penalty induced by (a, b) or (b, a) can make the agents find action a very unattractive, which thus usually results in convergence to the suboptimal outcome (b, b). Figure 1(b) shows the percentage of agents reaching the optimal joint action (a, a) of the climbing game as a function of the number of rounds for both IALs and JALs. We can see that both IALs and JALs can successfully learn to coordinate on the optimal joint action (a, a) without significant performance difference. Since there is no stochasticity in the climbing game, there is no significant advantage for the JALs by having the additional joint action information at their disposal.

In the penalty game, apart from the existence of two joint actions with heavy penalties (a, c) and (c, a) ($k = -50$), there also exist two optimal joint actions (a, a) and (c, c) in contrast to only one optimal joint action in the climbing game. Thus it is more difficult for the agents to reach an optimal joint action in the penalty game since the agents also need to agree on which optimal joint action to choose. Figure 1(d) shows the frequency of coordinating on optimal joint actions (a, a) or (c, c) as a function of the number of rounds for both IALs and JALs. All agents in the system can finally learn to converge on one optimal joint action for both IALs and JALs. Note that the results are averaged over 500 runs, which means that half of times all agents learn to adopt the policy of coordinating on the optimal joint action (a, a), and learn to converge to another optimal joint action (c, c) the other half of times. This is expected since the game itself is symmetric and the agents' preferences on both optimal joint actions are the same. Besides, similar to the climbing game, since the penalty game is deterministic, there is no significant difference in the learning performance between IALs and JALs even though JALs have more information at their disposal.

Partially stochastic climbing game

In this section, we consider the first variant of the climbing game - the partially stochastic climbing game (shown in Figure 1(e)) [Kapetanakis and Kudenko, 2002]. This version of the climbing game is more difficult than the original one in that a stochastic payoff is introduced for the joint action (b, b). The joint action (b, b) yields the payoff of 14 and 0 with probability 0.5, which makes it easy for the agents to misperceive (b, b) as the optimal joint action. This partially stochastic climbing game is in essence equivalent with the original climbing game, since the expected payoff of each agent by achieving (b, b) remains unchanged.

Figure 2(a) illustrates the dynamics of proportion of agents coordinating on the optimal joint action (a, a) for both IALs and JALs. First we can see that both IALs and JALs can reach full coordination on (a, a) after approximately 2200 rounds. Another observation is that the JALs do perform significantly better than that of IALs in terms of the convergence rate. This is expected since the JALs can distinguish the Q-values of different joint actions and have the ability of quickly identifying which action pair is optimal. In contrast, for the IALs, since they cannot perceive the actions of their interacting partners, they cannot distinguish between the noise from the stochasticity of the game and the explorations of their interacting partners. Thus it takes more time for the IALs to learn the actual Q-values of their individual actions.

Fully stochastic climbing game

Next let us further consider another more difficult variant of the climbing game - the fully stochastic climbing game (see Figure 1(f)). The characteristic of this game is that all joint actions yield two different payoffs with equal probability. However, this version of the climbing game is still functionally equivalent with the originally deterministic one, since the average payoff of each joint action remains the same with the payoff in the deterministic version.

Figure 2(b) illustrates the percentage of agents coordinating on the optimal joint action (a, a) of the fully stochastic climbing game as a function of the number of rounds for both IALs and JALs. We can see that JALs can always successfully learn to coordinate on (a, a), while IALs fail to do that. For IALs, they cannot distinguish whether the uncertainty of each action's payoff is caused by the stochasticity of the game itself or the random explorations of their interacting partners. Even though IALs take into consideration the information of the occurrence frequency of each action corresponding to its highest payoff, they fail to accurately learn the Q-values of each action due to too much noise introduced by the full stochasticity of the game. However, for the JALs, since they learn the Q-values of each joint action first and then evaluate the relative performance of each individual action (EV-values) based on the Q-values of each joint action, it is equivalent that JALs learn over the estimated deterministic version of the fully stochastic game. Thus it is expected that efficient coordinations on optimal joint action can be achieved, and indeed from Figure 2(b) we can see that full coordinations on optimal joint action among JALs can be achieved.

It is worth mentioning that, under this fully stochastic climbing game, previous work [Matignon *et al.*, 2012] shows that both (recursive) FMQ heuristic and optimistic assumption fail to coordinate on the optimal joint action (a, a), and the lenient learners only can coordinate on the optimal joint action (a, a) in approximately 90% of the runs on average. All the above learning strategies are designed for fixed-agent repeated interaction framework only with no *social learning* involved. In contrast, we have shown that 100% of JALs can successfully learn to coordinate on optimal joint action under our *social learning* framework, which also can better reflect the practical interactions among agents.

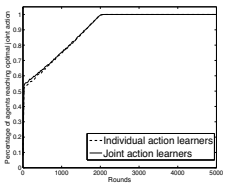
4.2 Effects of degree of information sharing degree among agents

One major difference between the social learning framework and the setting of fixed-agent repeated interactions is that the agents may have more information (i.e., other pairs of agents in the population) at their disposal. Therefore in this section, we evaluate the influences of the amount of information shared among agents (i.e., the value of M) in the social learning framework on the overall learning performance.

Figure 2(c) shows the learning dynamics of the percentages of IALs coordinating on optimal joint action with different values of M in the partially stochastic climbing game

1's payoff / 2's payoff		Agent 2		
		a	b	c
Agent 1	a	11	-30	0
	b	-30	7	6
	c	0	0	5

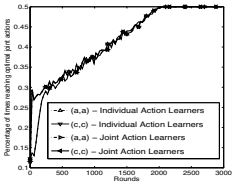
(a)



(b)

1's payoff / 2's payoff		Agent 2		
		a	b	c
Agent 1	a	10	0	k
	b	0	2	0
	c	k	0	10

(c)



(d)

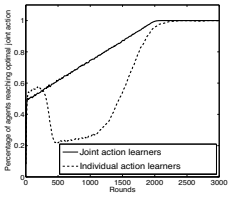
1's payoff / 2's payoff		Agent 2		
		a	b	c
Agent 1	a	11	-30	0
	b	-30	14/0	6
	c	0	0	5

(e)

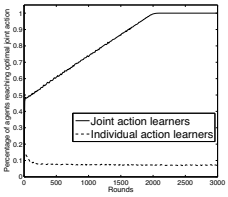
1's payoff / 2's payoff		Agent 2		
		a	b	c
Agent 1	a	10/12	5/-65	8/-8
	b	5/-65	14/0	12/0
	c	5/-5	5/-5	10/0

(f)

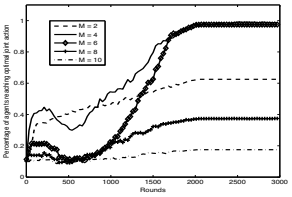
Figure 1: (a) The climbing game, (b) Percentage of agents reaching optimal joint action in climbing game, (c) The penalty game, (d) Percentage of agents reaching optimal joint actions in penalty game, (e) Partially stochastic climbing game ((b, b) yields the payoff of 14 or 0 with equal probability.), and (f) Fully stochastic climbing game (each joint action yields two different payoffs with equal probability.)



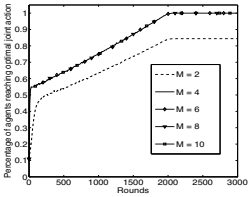
(a)



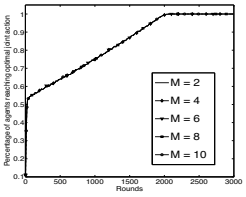
(b)



(c)



(d)



(e)

Figure 2: Percentage of agents coordinating on optimal joint action(s) in : (a) partially stochastic climbing game, (b) fully stochastic climbing game, (c) partially stochastic climbing game when M varies for IALs, (d) deterministic climbing game when M varies for IALs, and (e) fully stochastic climbing game when M varies for JALs (averaged over 500 times)

(see Figure 1(e)). We can see that when the value of M is too small, the agents fail to learn a consistent policy of coordinating on the optimal joint action. As the value of M increases from 2 to 4, better performance (full coordinations) can be achieved in terms of the percentages of agents coordinating on the optimal joint action. However, it is interesting to notice that the coordination performance is gradually decreased when the value of M is further increased ($4 \rightarrow 6 \rightarrow 8 \rightarrow 10$). This implies that either giving too little or too much information to the agents can impede efficient coordinations among them for IALs. We hypothesize that it is due to the reason as follows. In the partially stochastic climbing game, for IALs, the accurate evaluations of each individual action require appropriate consideration of the effects introduced by both the noise from the stochasticity of the game and the explorations of their interacting partners. If the value of M is too small, the agents cannot get an accurate estimation of the performance of each individual action since the sampling size is too small. However, if the value of M becomes too large, the noise introduced by the explorations of the interacting partners will dominate and the stochasticity of the game will be neglected. Accordingly, the IALs will get a biased estimation of the performance of each action when M is too large. On the other hand, when the climbing game is deterministic, the above problem does not exist any more, thus the learning performance of IALs is always improved as M is increased. The above hypothesis can be further verified from the experimental results shown in Figure 2(d): the overall performance is improved when M is increased from 2 to 4 and remains unchanged when M is further increased.

For JALs, since they can perceive the joint actions of all M pairs of interacting agents, the larger M is, the more accurate estimations of the expected payoff of each joint action (Q-values) they can get. Given a stochastic game, the JALs actually learn over its corresponding approximated deterministic version where the payoff of each joint action is replaced by its expected value. Therefore, similar to the analysis of IALs in deterministic games, there are no detrimental effects on the learning performance of JALs when the value of M is increased. This analysis also can be verified from the simulation results shown in Figure 2(e) in the fully stochastic climbing game for JALs: full coordinations on optimal joint action can be achieved already with $M = 2$ and the performance remains unchanged when M is further increased.

5 Conclusion and Future Work

In this paper, we investigate the multiagent coordination problem by proposing two types of learners (IALs and JALs) in cooperative environments under the social learning framework, which is complementary to the large body of previous work in the framework of repeated interactions among fixed agents. Selecting the optimal value of M for different practical problems is nontrivial. As future work, we are going to further investigate the relation between the optimal value of M and the characteristic of the problem (game) itself, and also evaluations using larger testbeds are needed to further validate the effectiveness of our social learning framework. Another interesting direction is to explicitly consider the underlying network structure instead of random interaction mechanism to facilitate more efficient coordinations.

References

[Brafman and Tennenholtz, 2004] R. I. Brafman and M. Tennenholtz. Efficient learning equilibrium. *Artificial Intelligence*, 159:27–47, 2004.

[Claus and Boutilier, 1998] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *AAAI'98*, pages 746–752, 1998.

[Fulda and Ventura, 2007] N. Fulda and D. Ventura. Predicting and preventing coordination problems in cooperative learning systems. In *IJCAI'07*, 2007.

[Kapetanakis and Kudenko, 2002] S. Kapetanakis and D. Kudenko. Reinforcement learning of coordination in cooperative multiagent systems. In *AAAI'02*, pages 326–331, 2002.

[Lauer and Riedmiller, 2000] M. Lauer and M. Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *ICML'00*, pages 535–542, 2000.

[Matignon *et al.*, 2008] L. Matignon, G. J. Laurent, and N. Le For-Piat. A study of fmq heuristic in cooperative multi-agent games. In *AAMAS'08 workshop: MSDM*, pages 77–91, 2008.

[Matignon *et al.*, 2012] L. Matignon, G. J. Laurent, and N. Le For-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *Knowledge Engineering Review*, 27:1–31, 2012.

[Panait and Luke, 2005] L. Panait and S. Luke. Cooperative multi-agent learning: The state of the art. *AAMAS*, 11(3):387–434, 2005.

[Panait *et al.*, 2006] L. Panait, K. Sullivan, and S. Luke. Lenient learners in cooperative multiagent systems. In *AAMAS'06*, pages 801–803, 2006.

[Rendell *et al.*, 2010] L. Rendell, R. Boyd, D. Cownden, M. Enquist, K. Eriksson, M. W. Feldman, L. Fogarty, S. Ghirlanda, T. Lillicrap, and K. N. Laland. Why copy others? insights from the social learning strategies tournament. *Science*, 328(5975):208 213, 2010.

[Sen and Airiau, 2007] S. Sen and S. Airiau. Emergence of norms through social learning. In *IJCAI'07*, pages 1507–1512, 2007.

[Sigmund *et al.*, 2010] K. Sigmund, H. D. Silva, A. Traulsen, and C. Hauert. Social learning promotes institutions for governing the commons. *Nature*, 466:7308, 2010.

[Villatoro *et al.*, 2011] D. Villatoro, J. Sabater-Mir, and S. Sen. Social instruments for robust convention emergence. In *IJCAI'11*, pages 420–425, 2011.

[Wang and Sandholm, 2002] X. Wang and T. Sandholm. Reinforcement learning to play an optimal nash equilibrium in team markov games. In *NIPS'02*, pages 1571–1578, 2002.

[Watkins and Dayan, 1992] C. J. C. H. Watkins and P. D. Dayan. Q-learning. *Machine Learning*, pages 279–292, 1992.