Reinforcement Social Learning of Coordination in Cooperative Multiagent Systems

(Extended Abstract)

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 and has received a lot of attention in multiagent learning literature. Most of previous works study the problem of how two (or more) players can coordinate on Pareto-optimal Nash equilibrium(s) through fixed and repeated interactions in the context of cooperative games. However, 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, in this paper, we investigate the multiagent coordination problems in cooperative environments under the social learning framework, in which there exists a large population of agents and each agent interacts with another agent randomly in each round. Each agent learns its policy through repeated interactions with the rest of agents via social learning. We distinguish two different types of learners depending on the amount of information each agent can perceive: *individual action learner* and *joint action learner*. The learning performance of both types of learners are evaluated under a number of challenging deterministic and stochastic cooperative games.

Categories and Subject Descriptors

I.2 [**ARTIFICIAL INTELLIGENCE**]: Distributed Artificial Intelligence—*Multiagent systems*

Keywords

Multiagent learning, coordination, cooperative games

1. INTRODUCTION

In multiagent systems (MASs), one important and widely studied class of problem is how to coordinate within cooperative MASs. There are a number of challenges the agents have to face when learning in cooperative MASs, e.g., *equilibrium selection problem* [1] and *stochasticity problem* [2].

Until now, various multiagent reinforcement learning algorithms [3] have been proposed in the literature to solve the coordination problem in cooperative MASs. Most of previous works heavily rely on the Q-learning algorithm 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, 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. Each agent learns its policy through repeated interactions with different opponents, and this kind of learning is termed as *social learning* [4] to distinguish from the case of learning from repeated interactions with the same partner(s). 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. 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). We investigate the learning performance of both types of learners under the testbed of Claus and Boutilier's coordination games and the more challenging stochastic variants, and throw light on the learning dynamics of both types of learners via social learning.

2. 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. At the end of each round, each agent updates its policy based on the learning experience it receives from the current round.

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

Appears in: *Proceedings of the 12th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2013)*, Ito, Jonker, Gini, and Shehory (eds.), May, 6–10, 2013, Saint Paul, Minnesota, USA.

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, denoted as S_i^t. In the second 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, denoted as P_i^t.

2.1 Individual Action Learner

In the first setting, each agent holds a Q-value $Q(s,a)$ for each action a under each state $s \in \{Row, Column\}$. At the end of each round t, each agent i picks an action (randomly choosing in case of a tie) with the highest payoff from the set S_i^t, and updates this action's Q-value following Equ. 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 $freq(a)$ is the frequency that action a occurs with the highest reward $r(a)$ among set S_i^t.

Each agent chooses its action based on the corresponding set of Q-values during each interaction according to the ϵ-greedy mechanism: each agent chooses its action with the highest Q-value with probability $1 - \epsilon$ and makes random choices with probability ϵ.

2.2 Joint Action Learner

In the second setting, 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,

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

At the end of each round t, for each action a, let us define $r_a^{max}(s) = max\{Q_i^{t+1}(s,(a,b)) \mid b \in A_i\}$. Finally, each agent i assesses the relative performance $EV(s,a)$ of an action a under the current state s as follows,

$$EV(s,a) = r_a^{max}(s) \times freq_i(a,b') \quad (3)$$

where the joint action pair (a,b') corresponds to the maximum payoff r_a^{max}(s) in $Q_i^{t+1}(s,\overrightarrow{a})$. Based on the EV-values $EV(s,.)$ of its individual actions, each agent chooses its action in the same way as it would use Q-values in Section 2.1 following the ϵ-greedy mechanism.

3. EXPERIMENTAL RESULTS

3.1 Deterministic games

We consider two particularly difficult coordination problems: the climbing game (Figure 1(a)) and the penalty game with $k = -50$ (Figure 1(b)). Simulation resulsts show that both IALs and JALs can successfully learn to coordinate on the optimal joint action (a,a) after approximately 2000 rounds without significant performance difference. However note that for the penalty game, since there exist two different optimal outcomes, half of times all agents learn to coordinate on the optimal joint action (a,a), and learn to converge to another optimal joint action (c,c) the other half of times.

3.2 Stochastic climbing game

We consider two different versions of stochastic climbing games: partially stochastic version (only the payoff for outcome (b,b) is stochastic (-14/0)) and fully stochastic version (the payoffs for all outcomes are stochastic, but the expected payoffs are still unchanged). For the partially stochastic climbing game, we observe 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.

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)

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

(b)

Figure 1: Payoff matrices for (a) the climbing game, (b) the penalty game

For the fullly stochastic climbing game both IALs and JALs, simulation results show that JALs can always successfully learn to coordinate on (a,a), while the IALs fail to do that. We hypothesize that it is because IALs cannot distinguish whether the uncertainty of each action's payoff is caused by the stochasticity of the game itself or the random exploration of their interacting partners.

4. CONCLUSIONS

In this paper, we investigate the multiagent coordination problem 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. Two different types of learners (IALs and JALs) based on the traditional Q-learning algorithm are introduced by incorporating both heuristics of optimal assumption and FMQ strategy. For deterministic cooperative games, both IALs and JALs can effectively learn to coordinate on optimal joint actions without significant performance difference, however, when it comes to stochastic cooperative games, JALs usually can achieve much better performance than IALs, since it can better distinguish between the stochasticity of the game itself and the stochastic explorations of the interacting agents. One interesting direction is to consider the design of alternative interaction mechanisms instead of adopting random interaction mechanism to facilitate more efficient coordinations among agents.

5. REFERENCES

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

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

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

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