Learning By Demonstration in Repeated Stochastic Games (Extended Abstract)

Jacob W. Crandall
Masdar Institute of Science and Technology
Abu Dhabi, UAE
jcrandall@masdar.ac.ae

Malek H. Altakrori
Masdar Institute of Science and Technology
Abu Dhabi, UAE
maltakrori@masdar.ac.ae

Yomna M. Hassan
Masdar Institute of Science and Technology
Abu Dhabi, UAE
yhassan@masdar.ac.ae

ABSTRACT

Despite much research in recent years, newly created multi-agent learning (MAL) algorithms continue to have one or more fatal weaknesses. These weaknesses include slow learning rates, failure to learn non-myopic solutions, and inability to scale up to domains with many actions, states, and associates. To overcome these weaknesses, we argue that fundamentally different approaches to MAL should be developed. One possibility is to develop methods that allow people to teach learning agents. To begin to determine the usefulness of this approach, we explore the effectiveness of *learning by demonstration* (LbD) in repeated stochastic games.

Categories and Subject Descriptors

H.4 [**Information Systems**]: Miscellaneous

General Terms

Algorithms

Keywords

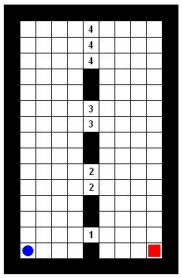
Multiagent learning, learning by demonstration

1. INTRODUCTION

Despite high research emphasis over the last few decades, newly created multi-agent learning (MAL) algorithms continue to learn slowly, fail to learn non-myopic solutions, or are unable to scale up to domains with many actions, states, and associates. To overcome these repeated shortcomings, we believe that fundamentally new approaches to MAL must be developed. One potential solution is to augment the learning process with intermittent interactions with a human teacher. In this paper, we study the effectiveness of learning by demonstration (LbD) [1], wherein the teacher intermittently demonstrates the actions that he or she believes the agent should perform, in repeated stochastic games.

LbD has been studied and applied to many problems, particularly in the robotics domain [1]. Most of this research has pertained to situations in which the human teacher knows successful behavior. However, in repeated games, information about learning associates, their tendencies, behaviors, and goals, and even the game itself is lacking. Thus, a human teacher may not know how the agent should behave to be successful. Since the teacher will also likely learn throughout the repeated game, demonstrations provided by the human are likely to be noisy and to change over time.

Cite as: Learning By Demonstration in Repeated Stochastic Games (Extended Abstract), J. W. Crandall, M. H. Altakrori and Y. M. Hassan, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. XXX-XXX.



(a)

	Defect G1	**Coop** G2, G3, G4
Defect G1	-25, -25	-10, -32
Coop G2, G3, G4	-32, -10	-16, -16

(b)

Figure 1: (a) A multi-stage prisoner's dilemma game. (b) High-level payoff matrix.

2. MULTI-STAGE PRISONERS' DILEMMA

To begin to investigate the effectiveness of LbD in repeated stochastic games, we consider the game shown in Fig. 1(a) [2]. In this game, two players begin each round in opposite corners of the world, and seek to move across the world through one of four gates to the other player's start position in as few moves as possible. If both agents seek to go through gate 1, then gates 1 and 2 close and the agents must go through gate 3. However, if only one agent goes through gate 1, gates 1-3 close and the other agent must go through gate 4. When both agents seek to go through gate 2 they are both allowed passage.

When a player attempts to move through gate 1, it is said to have *defected*. Otherwise, it is said to have *cooperated*. Viewed in this way, the *high-level* game is the prisoner's dilemma matrix game shown in Fig. 1(b). Each cell specifies the negative cost, based on the minimum number of steps it takes to reach the goal, of the row player (first number) and the column player (second number), respectively. We refer to this game as the multi-step prisoners' dilemma (MSPD).

3. PREVIOUS LEARNERS IN THE MSPD

Existing MAL algorithms for repeated stochastic games fall into two categories: followers and leaders [3]. Follower algorithms typically attempt to learn a best response to associates' strategies using only their own payoffs. We represent

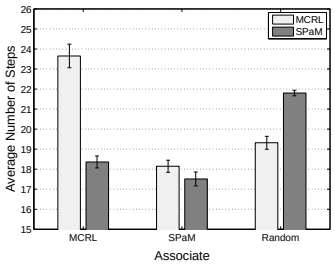


Figure 2: Average number of steps taken by MCRL and SPaM against various associates in the MSPD.

the performance of follower algorithms in the MSPD with a Monte Carlo reinforcement learning (MCRL) algorithm that uses k-nearest neighbor function approximation. So-called leader algorithms coax associates to learn less-myopic strategies. We represent follower algorithms with SPaM [2], a leader algorithm designed for stochastic games that encourages associates to cooperate in the MSPD.

Fig. 2 shows the asymptotic performance of MCRL and SPaM in the MSPD against several associates. SPaM learns effectively when playing both itself and MCRL, reaching mutual cooperation in both cases. On the other hand, MCRL performs effectively when it associates with SPaM, but learns mutual defection in self play. However, MCRL scores better when associating with Random than does SPaM. The best thing to do against Random in the MSPD is to always defect, which MCRL learns to do. SPaM on the other hand, continues to try to teach Random to cooperate. Thus, it cooperates when it believes that Random will cooperate and defects when it believes that Random will defect.

These results indicate that, in general, neither follower nor leader algorithms perform well against all kinds of agents in the MSPD. Additionally, both MCRL and SPaM require domain-specific knowledge in order to learn effectively in the MSPD, which limits the generalizability of these algorithms.

4. LBD IN THE MSPD

We next consider the potential of two LbD algorithms in repeated stochastic games. These algorithms receive periodic demonstrations from a human teacher throughout the repeated game. In rounds in which the teacher provides demonstrations, the agent follows the demonstrations. Otherwise, the agent follows the strategy it has derived.

The first algorithm, called Imitator, uses a k-nearest neighbor classifier to imitate the teacher's demonstrations. We anticipate that this algorithm will perform well when the teacher provides good demonstrations, but that it will not perform well when demonstrations are not well informed. The second algorithm, called MCRL-LbD, uses reinforcement learning to distinguish between effective and ineffective demonstrations. Initially, MCRL-LbD imitates the teacher's demonstrations. However, as it gains experiences, it acts so as to maximize its expected payoffs. Ideally, this algorithm would eventually learn effective behavior even when the teacher's demonstrations are not well informed.

We ran simulations using three forms of teacher demonstrations: tit-for-tat (TFT), random demonstrations (Random), and demonstrations that transitioned from random to always defect to TFT as the game progressed (Learner).

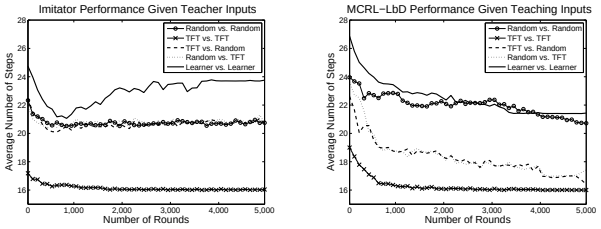


Figure 3: Performance of Imitator and MCRL-LbD in self play given various demonstration.

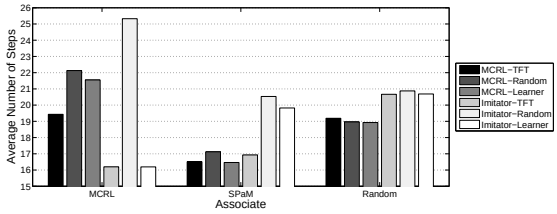


Figure 4: Performance of Imitator and MCRL-LBD against MCRL, SPaM, and Random.

The combination of the two algorithms with the three forms of human demonstrations form six algorithms. The average performances of these algorithms in self play and against other learners are shown in Figs. 3 and 4. Imitator is able to learn effective behavior when the teacher's demonstrations are well informed, but does not learn effectively when demonstrations are not well informed. MCRL-LbD typically learns successful behavior when demonstrations are well informed. It also sometimes learns effective behavior when demonstrations are not well informed. For example, it learns effectively against Random (defects) and SPaM (cooperates) regardless of the demonstrations given (Fig. 4), but produces mixed results against MCRL.

5. CONCLUSIONS

These results show the potential of LbD in repeated games. When teachers provide well informed demonstrations, LbD is successful. Moreover, MCRL-LbD is also sometimes effective when demonstrations are not well informed. This indicates that interactive learning algorithms can potentially be developed that allow agents to learn successfully even when human input is not well informed. Improvements can likely be made by altering the learning algorithm itself, the interactions between the teacher and the learner, or both.

6. REFERENCES

[1] B. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and Autonomous Systems*, 57(5):469–483, 2009.

[2] J. W. Crandall and M. A. Goodrich. Establishing reputation using social commitment in repeated games. In *AAMAS workshop on Learning and Evolution in Agent Based Systems*, New York City, NY, 2004.

[3] M. L. Littman and P. Stone. Leading best-response strategies in repeated games. In *IJCAI workshop on Economic Agents, Models, and Mechanisms*, 2001.