Collaborative Reinforcement Learning Model for Sustainability of Cooperation in Sequential Social Dilemmas

Extended Abstract

Ritwik Chaudhuri
IBM Research, Delhi, India
charitwi@in.ibm.com

Kushal Mukherjee
IBM Research, Delhi, India
kushmukh@in.ibm.com

Ramasuri Narayanam
IBM Research, Bangalore, India
ramasurn@in.ibm.com

Rohith Dwarakanath Vallam
IBM Research, Bangalore, India
rovallam@in.ibm.com

Ayush Kumar
IIT Delhi, India
me2150727@iitd.ac.in

Antriksh Mathur
IIT Delhi, India
me2150722@iitd.ac.in

Shweta Garg
IBM Research, Delhi, India
shgarg87@in.ibm.com

Sudhanshu Singh
IBM Research, Delhi, India
sudsing3@in.ibm.com

Gyana Parija
IBM Research, Delhi, India
gyana.parija@in.ibm.com

ABSTRACT

Learning the emergence of cooperation in conflicting scenarios such as social dilemmas is a centrepiece of research. Many reinforcement learning based theories exist in the literature to address this problem. The well-known fact about RL based model's very slow learning capabilities coupled with large state space exhibit significant negative effects especially in repeated version of social dilemma settings such as repeated Public Goods Game (PGG) and thereby making them ineffective to model sustainability of cooperation. In this paper, we address this research challenge by augmenting the reinforcement learning based models with a notion of collaboration among the agents, motivated by the fact that humans learn not only through their own actions but also by following the actions of other agents who also continuously learn about the environment. In particular, we propose a novel model, which we refer to as Collaborative Reinforcement Learning (CRL), wherein we define collaboration among the agents as the ability of agents to fully follow other agent's actions/decisions. This is also termed as social learning. The proposed CRL model significantly influences the speed of individual learning, which eventually has a large effect on the collective behavior as compared to that of RL only models and thereby effectively explaining the sustainability of cooperation in repeated PGG settings. We also extend the CRL model for PGGs over different generations where agents die out and new agents are born following a birth-death process.

KEYWORDS

Multi-agent learning; Learning agent capabilities; Reinforcement learning

ACM Reference Format:
Ritwik Chaudhuri, Kushal Mukherjee, Ramasuri Narayanam, Rohith Dwarakanath Vallam, Ayush Kumar, Antriksh Mathur, Shweta Garg, Sudhanshu Singh, and Gyana Parija. 2019. Collaborative Reinforcement Learning Model for Sustainability of Cooperation in Sequential Social Dilemmas. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 3 pages.

1 INTRODUCTION

Understanding the emergence of cooperation in conflicting scenarios such as social dilemmas has been an important topic of interest in the research community [2, 7]. In particular, why and under what circumstances speculatively selfish individuals cooperate in repeated versions of social dilemma settings, such as repeated Public Goods Game (PGG), has been a long-standing research question. Towards this end, several reinforcement learning (RL) based approaches exist in the literature such as multi-agent reinforcement learning [3, 10, 12], simple reinforcement learning [5], influences of social networks [8], enforcement of laws and rewards for altruism and punishment [1] , emotions giving rise to direct reciprocity [11] or even indirect reciprocity [9] to achieve stable cooperation among the agents in repeated multi-agent social dilemma settings. In repeated social dilemmas (including repeated PGGs), note that it takes a reasonably long time for the autonomous agents to learn whether cooperation is the best policy to get a long-term aggregated reward [4]. Further, it takes a longer period of time to achieve stability with respect to cooperation in such repeated social dilemmas as some agents being fully autonomous always have parasitic tendencies to free ride [6]. Building on this, the well-known fact about RL based model's very slow learning capabilities coupled with large state space exhibit significant negative effects especially in modelling the emergence of cooperation in the repeated version of public goods games and thereby making them inadequate to explain sustainability of cooperation in such scenarios. Then, an *interesting research question would be how to modify the reinforcement learning framework making them effective not only for faster agent learning but also for explaining sustainability of cooperation in repeated social dilemma setting such as repeated PGGs?*. We address this research gap by augmenting the reinforcement learning based models with a notion of *collaboration* among the agents, motivated by the fact that humans often learn not only through their own actions but also by keeping a track or following the actions of other agents who also continuously learn about the environment. Our

proposed models are very different from an imitation learning paradigm which hinges upon the fact that there is an expert/teacher whom an agent follows. In our proposed model even though each agent keeps a track of the actions of other agents, all the agents co-evolve together in the environment without the presence of any expert. In particular, we propose a novel model, which we refer to as *Collaborative Reinforcement Learning (CRL)*, wherein we define collaboration among the agents as the ability of agents to fully keep a track of other agent's actions/decisions but autonomously decide whether to take an action based on the actions taken by the peers in the past.

2 PROPOSED CRL MODEL FOR REPEATED PGGS

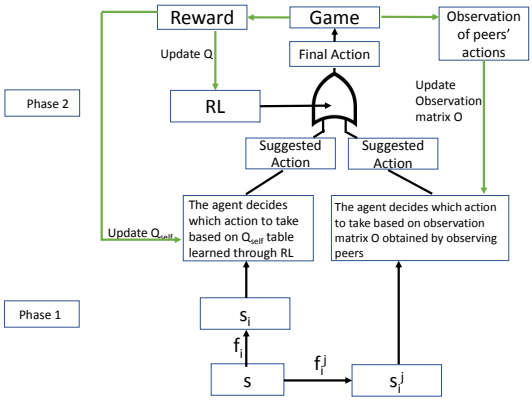

Figure 1: CRL methodology followed by each agent in every round of an iterated PGG

Consider n agents in a repeated PGG setting.In each round of the PGG, the reward of agent i is given by $r_i = 1 - e_i + \beta \sum_{j=1}^{n} e_j$ where contribution level for agent i is $e_i \in [0, 1]$. The range $[0, 1]$ of e_i is discretized into M possible different contribution levels, which are represented as $C = \{c_0, c_1, ..., c_{M-1}\}$.

As shown in Figure 1, each agent follows a hierarchical RL model to take an action in each round of the game. The hierarchical RL model is divided into 2 phases. In Phase 1, agent i first finds out the best possible action he could play based on the Q-table learned from self taken actions. Then agent i also finds out the best possible action he could choose based on the observation matrix he obtained watching the actions of the peers. Finally in the Phase 2 he uses a stateless RL model with two actions, to decide whether to use the action based on Q-table or the action based on observing peers (matrix O).

3 EXPERIMENTAL RESULTS

For empirical analysis, we consider an iterated PGG with $n = 5$ agents. We conduct an experiment with 70000 iterated PGG games where each game lasts for 30 rounds. Hence, after each 30 rounds, we restart the game but we retain the learned Q-tables of each of the agents so that in the next game they can start using the retained Q-tables and update those in subsequent rounds of the new game. The contribution levels of each agent are either 0 or 1.

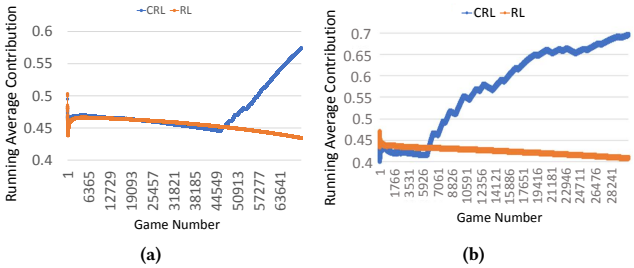

Figure 2: (a)Running average of contribution levels of agents in 70000 games (b)Running average of contribution level of new born agent in 30000 games

Note that, following the proposed CRL model, each agent observes the actions taken by other agents. As mentioned earlier we also recursively initialize the observation matrix to 0 after every 50 episode. In Figure 2(a) we consider 5 agents playing 70000 iterated PGG games in one case using the CRL model and in the other case using the standard RL model. We consider a metric: $C_R = \frac{1}{R}\sum_{T=1}^{R}(\frac{1}{5}\sum_{i=1}^{5}\frac{1}{30}\sum_{r=1}^{30} e_i^{r,T})$, where $e_i^{r,T}$ is the contribution level of the i^{th} agent in the r^{th} round of game T. This metric essentially computes the running average contribution levels of the agents till game R starting from game 1. We plot the running averages for each value of $R = 1, 2, ..., 70000$, that is for each of the 70000 games. It can be observed that agents, when learning through CRL model, reach a significantly higher running average of contribution levels as compared to the same when agents are learning through the RL model. Following CRL model of learning, observing other agents' actions and at times making actions based on the observed behaviour, helps each agent to learn faster about the environment and when there is a shift among a majority of agents towards contributing a higher amount, all the agents also start contributing higher amounts. The plotshown is for 1 of 10 runs.Over 10 runs the average contribution of the 5 agents (70000 episodes) obtained using CRL is significantly more than that obtained using RL by 20% with a p value=0.005(0.5%).

We conduct another experiment with 100000 games of iterated PGG each lasting for 30 rounds with $\beta = 0.7$. In this setup, after 70000 games, we replace the highest average contributing agent with a new agent who takes part in the next 30000 games. The reason for replacing the highest average contributing agent is, he gets the lowest average reward among all 5 agents in the 70000 games. The newly born agent may free ride on those agents contributing high in the system for some period, but eventually, it should learn to contribute due to the very fact that contributing agents will get less reward due to the free-riding behavior of the new agent. In Figure 2(b) we plot the running average of the contribution level of the new agent for each game of the 30000 games. It can be observed that the agent learns to contribute much faster while following CRL model as the agent gets to observe the actions of already high contributing agents. When the learning takes place only through the RL model, the agent learns slowly and in fact learns to contribute less in the 30000 games. The close tracking of the initial performance of CRL with RL(Figure 2(a),2(b)) is due to the initial large exploration factor, which is gradually reduced.

REFERENCES

[1] James Andreoni, William Harbaugh, and Lise Vesterlund. 2003. The Carrot or the Stick: Rewards, Punishments, and Cooperation. *American Economic Review* 93, 3 (2003).

[2] R. Axelrod and W.D. Hamilton. 1996. The evolution of cooperation. *Biosystems* 211, 1-2 (1996), 1390–1396.

[3] Daan Bloembergen, Karl Tuyls, Daniel Hennes, and Michael Kaisers. 2015. Evolutionary dynamics of multi-agent learning: A survey. *Journal of Artificial Intelligence Research* 53 (2015), 659–697.

[4] R. Engelmore. 1978. Prisoner's dilemma-recollections and observations. *Game Theory as a Theory of a Conflict Resolution* (1978), 17–34.

[5] Anna Gunnthorsdottir and Amnon Rapoport. 2006. Embedding social dilemmas in intergroup competition reduces free-riding. *Organizational Behavior and Human Decision Processes* 101, 2 (2006), 184–199.

[6] Paul AM Van Lange, Jeff Joireman, Craig D Parks, and Eric Van Dijk. 2013. The psychology of social dilemmas: A review. *Organizational Behavior and Human Decision Processes* 120, 2 (2013), 125–141.

[7] J. Ledyard. 1995. *A survey of experimental research.* The Handbook of Experimental Economics, eds JH Kagel, AE Roth, Princeton University Press, Princeton, NJ, USA.

[8] David G. Rand, Samuel Arbesman, and Nicholas A. Christakis. 2011. Dynamic social networks promote cooperation in experiments with humans. In *Proceedings of the National Academy of Sciences*. 19193–19198.

[9] Bettina Rockenbach and Manfred Milinski. 2006. The efficient interaction of indirect reciprocity and costly punishment. In *Proceedings of the National Academy of Sciences*. 718–723.

[10] T. W. Sandholm and R. H. Crites. 1996. Multiagent reinforcement learning in the iterated prisoner's dilemma. *Biosystems* 37, 1-2 (1996), 147–166.

[11] Matthijs van Veelen, Julian Garcia, David G. Rand, and Martin A. Nowak. 2012. Direct reciprocity in structured populations. In *Proceedings of the National Academy of Sciences*. 9929–9934.

[12] M. Wunder, M. Littman, and M. Babes. 2010. Classes of multiagent Q-learning dynamics with greedy exploration. In *Proceedings of the 27th International Conference on Machine Learning (ICML'10)*.