Towards Decentralized Reinforcement Learning Architectures for Social Dilemmas*

Nicolas Anastassacos
The Alan Turing Institute and University College London
London, United Kingdom
nanastassacos@turing.ac.uk

Mirco Musolesi
The Alan Turing Institute and University College London
London, United Kingdom
m.musolesi@ucl.ac.uk

ABSTRACT

Multi-agent reinforcement learning has received significant interest in recent years notably due to the advancements made in deep reinforcement learning which have allowed for the developments of new architectures and learning algorithms. In this extended abstract we present our initial efforts towards the development of decentralized architectures for multi-agent systems in order to understand and model societies. More specifically, using social dilemmas as the training ground, we present a novel learning architecture, Learning through Probing (LTP), where agents utilize a probing mechanism to incorporate how their opponent's behavior changes when an agent takes an action.

KEYWORDS

Multi-agent Systems; Reinforcement Learning; Cooperation.

ACM Reference Format:
Nicolas Anastassacos and Mirco Musolesi. 2019. Towards Decentralized Reinforcement Learning Architectures for Social Dilemmas. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), Montreal, Canada, May 13–17, 2019,* IFAAMAS, 2 pages.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) has garnered a significant amount of interest in recent years also due to the advancements in deep RL which has allowed for extensive study on agent behaviors. There has been emphasis on designing cooperative agents for decades [2, 9] yet extending this success to multi-agent environments has proven difficult as the Markov property is not satisfied since agent behaviors are continuously changing [8] and the use of experience replay does little to inhibit unstable learning in presence of multiple learners. Indeed, there are still challenges to be tackled in order to enable broader applications, e.g., in automated decision-making such as self-driving cars, personalized assistants, and the eventuality of artificial agents operating in society. A central aspect of this evolution lies in understanding the competitive and collaborative nature of environments and the emergence of such behaviors [1, 6].

Social dilemmas have been a staple when studying the emergence of cooperative and competitive strategies [3, 4, 7]. They reveal interesting tensions between the desires of an individual and what is best for the group but both game-theory and reinforcement learning approaches have struggled to tackle these types of games due to the added complexity of predicting how behaving in an environment will influence the learning and future behaviour of an opponent. Though recent approaches in MARL involve incorporating knowledge of opponent behaviour, they are concerned with optimizing against an opponent's known behaviour instead of their potential future behaviours [5, 10]. In contrast, our approach focuses directly on understanding the consequence of actions on opponent's behavior and incorporates that knowledge directly into agent learning via an adjusted reward function.

The Prisoner's Dilemma (PD) is a simple game that serves as the basis for research on social dilemmas. The premise of the game is that two partners in crime are imprisoned separately and each are offered leniency if they provide evidence against the other. Each player can choose between two actions: cooperation or defection. The dominant strategy is to defect, however, if both players take this action then they arrive at a Nash Equilibrium that is *socially deficient*. Originally, the PD is a one round game, but the IPD is a *sequential* PD often studied to understand the effects of previous outcomes and the emergence of cooperative behaviors.

2 LEARNING THROUGH PROBING

We propose a training mechanism, *Learning through Probing*, which allows agents to gather experiences that have been adjusted to reflect behavioral changes in a sequence of events over a period of time via an adjusted reward signal and, therefore, enables them to learn cooperative strategies. We identify two distinct phases, the *probing phase* and the *playing phase* and two components, the *prober* and the *player*. During the *probing phase* the probes explore the environment (and consequently, their opponent's current strategy). Each agent can probe the opponent agent in order to gather information about how their opponent's strategy changes after an update and adjust any collected experiences. We use a defined time horizon T to determine the number of updates to the opponent's strategy to consider. Experiences are then grouped according to the chosen actions of the agent and the opponent. To explore the outcome of taking an action a_t on an opponent's behaviour (which corresponds to a state s_t and leads to a reward r_{t+1}) the probe then updates on the subset of experiences and continues to play versus its opponent. After taking a one-step update based on these initial experiences, the *probes* play against each other according to their learned policies and repeat the process for T updates. The final trajectory $\tau = (s_0, a_0, r_1, s_1, ..., s_{T-1}, a_{T-1}, r_T, s_T)$ is stored and will be used to train the player component.

In the second phase, the agent trains on the adjusted experiences only that now account for changes to the opponent's behaviour over time. Secondly, with the addition of the probing phase, the

*An extended in-progress paper on the topics discussed in this paper can be found at: https://arxiv.org/abs/1809.10007.

Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019), N. Agmon, M. E. Taylor, E. Elkind, M. Veloso (eds.), May 13–17, 2019, Montreal, Canada.

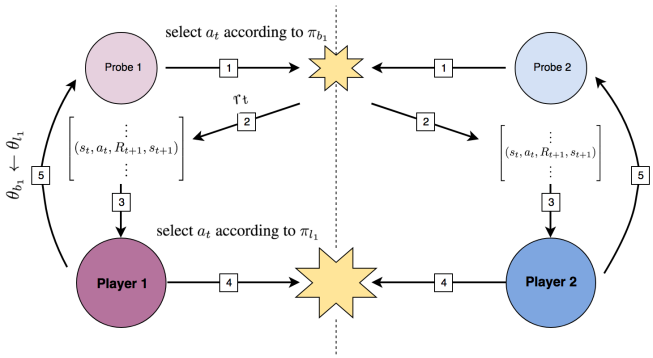

Figure 1: Learning through Probing architecture diagram involving two RL agents. 1) After exploring the environment, the *probe* component trains on subsets of experiences to learn consequences for actions. Actions are then selected according to a learned policy. 2) Experiences are collected into a replay buffer and adjusted. 3) The *player* component trains on the adjusted experiences. 4) The players are matched against each other after training. 5) In continuously adaptive games, probes could adopt learned player policies and adapt their strategies over time.

agents do not need to have information about the parameters of the opponent agent or need to track their strategy in advance.

Experimentally, we demonstrate that two RL agents trained with this approach learn to cooperate in the IPD and how this type of training mechanism results in a RL agent learning optimal policies when matched with other stationary and quasi-stationary strategies from Axelrod tournaments. Finally, we contrast this with current methodologies in multi-agent RL to highlight potential difficulties and we discuss how probing and using experiences through updates might help established methods achieve better performance in learning environments.

3 COOPERATIVE BEHAVIOUR IN STABLE SOCIETIES

The emphasis in current MARL is to stationarize the environment using techniques that give the agent more information about the dynamics at play in the environment. In the previous section, we use an adaptation to training procedure without changing the RL objective function. In this section we instead look to make changes to the training environment of Q-learning agents so they develop cooperative tendencies by inserting other agents into the environment that act as regulators. In open environments such as artificial societies, agents are likely to come into contact with unseen scenarios and their learning in these new environment is hard to predict which makes it difficult to come up with reasonable adjustments to algorithms. We are interested in understanding what happens when an agent is inserted in an unseen multi-agent environment with given dynamics, like a society with established social norms and how we can control for certain types of behavior without specifying changes to the agent's objective function. Evaluating these developmental aspects may provide key insights to understanding how types of behaviors are established in a society or how certain behaviors might provide the basis for stable social norms.

RL agents are trained with a vanilla Q-learning algorithm and learn to play against other RL agents. We evaluate how well these Q-learning agents perform against one another depending on the other agents that are present in the society. We start with an environment that features only two RL agents and systematically add agents that play a Tit-for-Tat strategy (TFT) one-by-one to the environment and observe the changes in Q-values. Other, less regulatory, agents are also added to see how cumulative reward changes in the society. The format was modeled as a random encounters where each agent was matched with a random opponent with equal probability. We can show that tailoring their overall experience without explicitly changing the cost function or reward function, they can learn to cooperate with other RL agents via regulating TFT agents. This is an interesting finding as it better represents how these agents would act in societal-like contexts and an understanding of what kind of regulatory techniques might be needed for reinforcement learning to be viable for facilitating multi-agent interactions.

ACKNOWLEDGMENTS

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

REFERENCES

[1] Robert Axelrod and William Donald Hamilton. 1981. The evolution of cooperation. *Science* 221, 4489 (1981), 1390–1396.
[2] Spiros Kapetanakis and Daniel Kudenko. 2002. Reinforcement learning of coordination in cooperative multi-agent systems. In *AAAI/IAAI*. 326–331.
[3] 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.
[4] Joel Z Leibo, Vinicius Zambaldi, Mac Lanctot, Janusz Marecki, and Thore Graepel. 2017. Multi-agent reinforcement learning in sequential social dilemmas. In *AAMAS*. 464–473.
[5] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *NIPS*. 6379–6390.
[6] Martin Nowak. 2006. Five rules for the evolution of cooperation. *Science* 314, 5805 (2006), 1560–1563.
[7] Elinor Ostrum, Roy Gardner, and James Walker. 1994. *Rules, games, and common-pool resources.* University of Michigan Press.
[8] Richard S Sutton and Andrew G Barto. 1998. *Reinforcement Learning: An Introduction.* MIT Press.
[9] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *ICML*, Vol. 10. 330–337.
[10] Gerald Tesauro. 2004. Extending Q-learning to general adaptive multi-agent systems. In *NIPS*. 871–878.