Gifting in Multi-Agent Reinforcement Learning (Student Abstract)

Andrei Lupu, Doina Precup
Mila / McGill University
3480 University St., Montreal, QC H3A 2A7
andrei.lupu@mail.mcgill.ca, dprecup@cs.mcgill.ca

Abstract

This work performs a first study on multi-agent reinforcement learning with deliberate reward passing between agents. We empirically demonstrate that such mechanics can greatly improve the learning progression in a resource appropriation setting and provide a preliminary discussion of the complex effects of gifting on the learning dynamics.

Introduction

Multi-agent reinforcement learning (MARL) is a rapidly growing field, with scenarios incorporating competitive, collaborative or mixed dynamics (Nguyen, Nguyen, and Nahavandi 2018). However, due to agents learning and changing their behaviour throughout the training, MARL is intrinsically non-Markovian and thus violates one of the core assumptions behind reinforcement learning methods. For that reason, MARL comes with its unique set of challenges, and in some settings can be subject to feedback loops that drive the system away from an optimal collective behaviour.

In order to prevent even further complications, previous MARL studies have thus adhered to a simplification inherited from classical RL – that is to treat the reward function as an exclusive property of the environment. Although this function can be non-stationary, it is then assumed that it evolves only in response to actions performed upon the environment and that the reward is entirely independent from the internal state of the agents present within it. Whereas this assumption is necessary and natural in classical single-agent RL, it is an artificial restriction in multi-agent settings and results in agents learning to interact only insofar as it helps them maximizing environmental returns. It is also very limiting as it greatly restricts the type of interactions allowed between agents. For instance, it makes it impossible to simulate negotiations, gifts and trade deals, which are at the core of many interactions within our societies.

Thus, two questions arise naturally. The first is how to stabilize MARL systems and prevent divergence during learning. The second, more broad, is how do current reinforcement learning algorithms behave when used in more general multi-agent settings, where rewards are no longer an exclusive property of the environment. In this work, we show that these two seemingly unrelated questions are linked. Specifically, we demonstrate that simply enabling agents to deliberately *gift* rewards to each other can lead to a drastically different learning progression that almost entirely avoids pitfalls observed in a previous study (Perolat et al. 2017).

To our knowledge, this work is the first to investigate the effects of deliberate reward gifting in MARL settings.

Background and Related Work

As a setting for the study of mutually rewarding multi-agent systems. we choose the problem of common pool resource (CPR) appropriation, where agents must compete to exploit a shared environment, while being careful not to deplete the resource pool through greedy approaches. This problem is ubiquitous in modern societies, encompassing scenarios such as access to fresh water or fisheries.

CPR appropriation has been studied as a MARL task by Perolat et al. In their work, they introduce the *Harvest* environment, where agents gather resources called apples, but can also tag each other by firing a beam. A tagged agent is then temporarily removed from the environment, thus limiting the number of apples it can collect. The key is that the apples regenerate at a rate dependent upon the remaining number of nearby apples, and so agents that are too greedy will permanently deplete the resource pool.

The agents in that work do ultimately converge to a collectively sustainable behaviour by tagging each other in order to reduce the effective population to the carrying capacity of the environment. However, they only do so after traversing a period of greed marked by a quick depletion of the resource pool. This period, called "*tragedy of the commons*" could have potentially disastrous consequences if those algorithms were to be deployed in a real resource exploitation scenario, and so presents a first test case for the benefit of reward gifting in MARL.

Our Approach

We modify the capabilities of the agents in the Harvest environment by extending their action space $\mathcal{A}$ with a reward

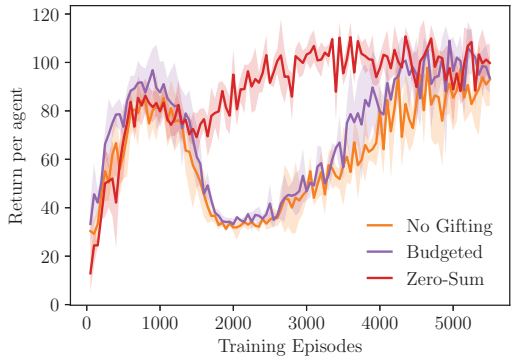

Figure 1: Return per agent for the Harvest environment with different gifting mechanics. Each curve is the average of 3 runs with different hyper-parameters.

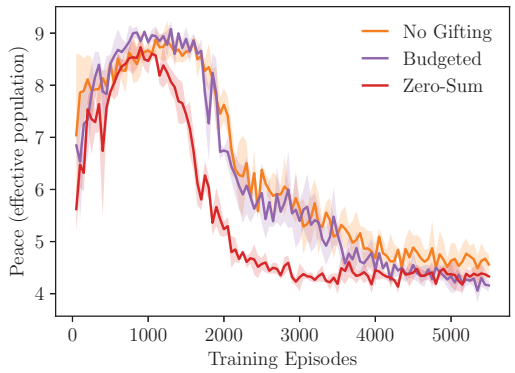

Figure 2: Peace metric for the Harvest environment with different gifting mechanics. Each curve is the average of 3 runs with different hyper-parameters.

gifting action. We experiment with two distinct gifting mechanics:

- **Budgeted gifting:** Agents have a fixed "budget" of apples they can gift to other agents, without observing any reward or penalty from doing so. Once the budget depleted, the gift action becomes equivalent to the null one.
- **Zero-sum gifting:** Agents can gift an apple as often as desired, but incur an immediate penalty of -1.

In both cases, the "gift" is equally split among all agents in a small region in front of the gifting agent, and simply added to their reward for that time step. If that region is devoid of other agents, the gift action is not processed.

Results and Discussion

We study the effect of the gifting mechanics in the Harvest environment, which we re-implement using the MiniGrid library (Chevalier-Boisvert, Willems, and Pal 2018).

We find that groups of agents enabled with budgeted gifting seem to entirely ignore that mechanic. Indeed, as can be seen from Fig. 1, their average return follows the same pattern through learning and is subject to the same tragedy of the commons as the baseline (non-gifting) agents. This could be anticipated since agents have no individual incentive for gifting, nor any process that allows them to represent the effect of that action on the reward perceived by other agents.

On the other side, Fig. 1 shows that groups enabled with zero-sum gifting exhibit a dramatically different learning progression, and almost entirely avoid the tragedy of the commons. This is highly unexpected and difficult to explain since (1) gifting in this variant is strongly penalized and (2) no net reward is added to the pool, leaving the incentive towards a greedy behaviour unchanged.

Furthermore, this drastic empirical change cannot be attributed exclusively to cooperation. Indeed, Fig. 2 shows that agents equipped with zero-sum gifting resort to tagging each other substantially earlier in the training process than standard or budgeted agents[1]. We use ε-greedy training and gifting only penalizes an agent if there are other agents present in its *gifting region*. Therefore, one could postulate that the effect is uniquely due to agents learning to tag each other earlier in order to avoid the negative reward when forced to explore the gift action. Although this might serve as a partial explanation, simultaneous consideration of both figures reveals that for behaviours with similar peace values, agents with zero-sum gifting can achieve very different returns from other types of agents. Thus, we conjecture that zero-sum gifting must affect not only their tagging rate, but also the way in which agents learn to collect resources.

Conclusion and Future Work

In this work we perform an initial study of multi-agent reinforcement learning with deliberate reward sharing (gifting) between agents. We demonstrate empirically that agents with a simple zero-sum gifting mechanic achieve a dramatically improved learning progression in the Harvest environment, as compared to standard agents or agents with a fixed gifting budget. We also provide a preliminary discussion on the dynamics behind the results obtained, and outline a direction for future investigation.

At the time of writing, we are running experiments aiming to better understand the effects of zero-sum gifting on the agents' behaviour.

Our further efforts will be directed towards extending this work to additional environments and gifting mechanics.

References

Chevalier-Boisvert, M.; Willems, L.; and Pal, S. 2018. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid.

Nguyen, T. T.; Nguyen, N. D.; and Nahavandi, S. 2018. Deep reinforcement learning for multi-agent systems: a review of challenges, solutions and applications. *arXiv preprint arXiv:1812.11794*.

Perolat, J.; Leibo, J. Z.; Zambaldi, V.; Beattie, C.; Tuyls, K.; and Graepel, T. 2017. A multi-agent reinforcement learning model of common-pool resource appropriation. In *Advances in Neural Information Processing Systems*, 3643–3652.

[1]Please see Perolat et al. (2017) for the Peace metric definition.