Balancing Rational and Other-Regarding Preferences in Cooperative-Competitive Environments

Extended Abstract

Dmitry Ivanov*
JetBrains Research; HSE University
Russian Federation, Saint-Petersburg
diivanov@hse.ru

Vladimir Egorov*
JetBrains Research; HSE University
Russian Federation, Saint-Petersburg
vsegorov@edu.hse.ru

Aleksei Shpilman
JetBrains Research; HSE University
Russian Federation, Saint-Petersburg
alexey@shpilman.com

ABSTRACT

Recent reinforcement learning studies extensively explore the interplay between cooperative and competitive behaviour in mixed environments. Unlike cooperative environments where agents strive towards a common goal, mixed environments are notorious for the conflicts of selfish and social interests. As a consequence, purely rational agents often struggle to maintain cooperation. A prevalent approach to induce cooperative behaviour is to assign additional rewards based on other agents' well-being. However, this approach suffers from the issue of multi-agent credit assignment, which can hinder performance. This issue is efficiently alleviated in cooperative setting with such state-of-the-art algorithms as QMIX and COMA. Still, when applied to mixed environments, these algorithms may result in unfair allocation of rewards. We propose BAROCCO, an extension of these algorithms capable to balance individual and social incentives. The mechanism behind BAROCCO is to train two distinct but interwoven components that jointly affect agents' decisions. We experimentally confirm the advantages of BAROCCO.

KEYWORDS

Multi-Agent Reinforcement Learning; Cooperation; Fairness

ACM Reference Format:
Dmitry Ivanov, Vladimir Egorov, and Aleksei Shpilman. 2021. Balancing Rational and Other-Regarding Preferences in Cooperative-Competitive Environments: Extended Abstract. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021*, IFAAMAS, 3 pages.

1 INTRODUCTION

Human cooperation is considered an evolutionary puzzle in the economic literature [3, 6, 9, 16, 20]. Despite the predictions of the rational choice theory to act selfishly [23], people of different age, gender, culture, and socioeconomic status engage into cooperation in a multitude of economic situations [2, 4, 5, 7, 11, 17]. A possible mechanism to resolve the paradox implies that the agents take social preferences into account during decision making [8, 9].

The questions of emergence and maintenance of cooperation are mirrored in the Multi-Agent Reinforcement Learning (MARL) literature [10, 18, 19, 21, 24, 25]. Numerous works have demonstrated that purely rational agents are unable to maintain cooperation and perform worse than those guided by social incentives [14, 15, 19, 27]. Despite this, training fully social agents can be undesirable when fairness is a concern. Such agents may prefer sacrificing own payoffs for the common good to equal reward distribution. A possible compromise that trades-off fairness and group performance is to balance selfish and social preferences of the agents.

The simplest way to achieve such balance is to train agents on a mixture of social [19, 28] and selfish rewards, which we refer to as Cooperative Reward Shaping (CRS):

$$r_i^{crs} = (1-\lambda) r_i + \lambda SW(\mathbf{r}) \tag{1}$$

where r_i is the native reward of i-th agent, $\mathbf{r}$ is a vector of rewards of all agents, SW denotes social welfare function like sum or minimum, λ is prosociality coefficient, and r_i^{crs} is mixed reward. However, CRS implies decentralized training and does not address such crucial issues of MARL as non-stationarity and credit assignment [1, 12, 13]. On the other hand, these issues are addressed in the techniques from fully cooperative MARL like QMIX [21] or COMA [10] that were shown to outperform decentralized training in such complex environments as StarCraft 2 [26]. Still, these techniques are only concerned with performance and ignore fairness.

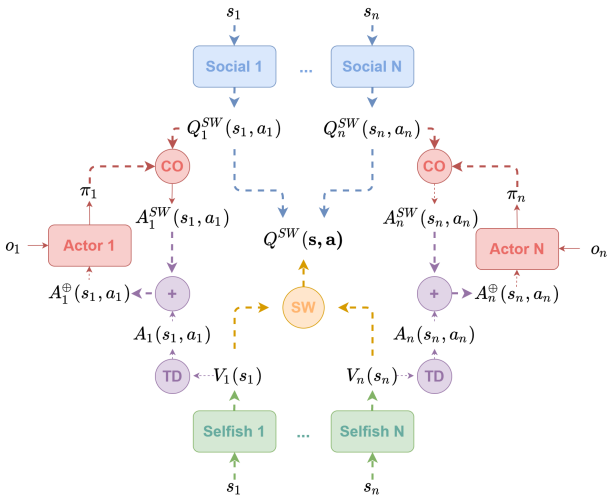
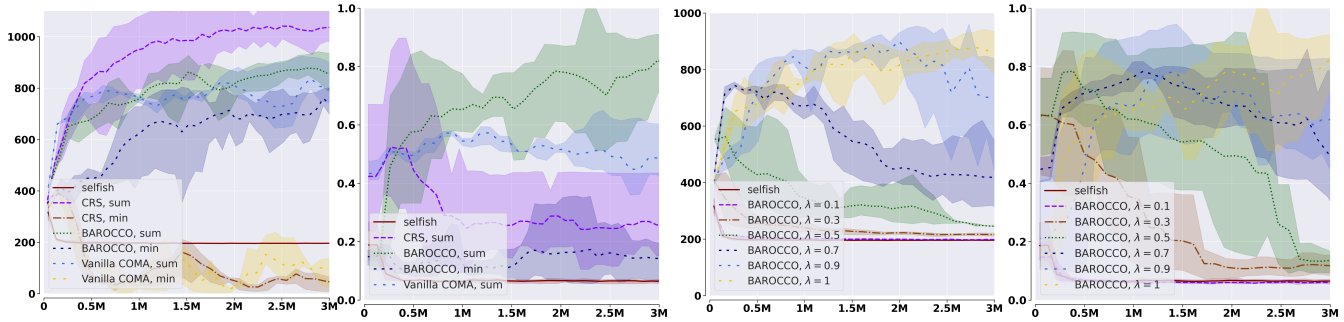


Figure 1: BAROCCO for Actor-Critic Framework. *CO*, *TD*, **and** *SW* **denote counterfactual baseline, temporal difference, and social welfare function, respectively.**

*Equal Contribution
The code for this paper can be found here

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.



(a) Apples, all algorithms (b) Gini, all algorithms (c) Apples, varying λ (d) Gini, varying λ

Figure 2: Experiments in Harvest. 'Apples' denotes total number of apples collected by all agents per episode. 'Gini' denotes gini coefficient that can be treated as a measure of unfairness.

2 BAROCCO

We propose a meta-algorithm that extends credit assignment techniques like COMA to mixed environments with capability to balance the incentives, which we refer to as BAROCCO, i.e. BAlancing Rational and Other-regarding preferences in Cooperative-COmpetitive environments. BAROCCO is based on the insight that instead of relying on a single model to balance the incentives via CRS, two distinct components can be trained concurrently and combined during decision making (Fig. 1). While the two approaches are mathematically equivalent, the latter approach allows us to train the social component via techniques that address credit assignment.

More specifically, for each agent we train selfish critic via a variant of MADDPG [18] and social critic via COMA [10]. Then, the actor is trained via proximal policy gradient [22] on a mixture of predictions of these two critics. The importance of each critic is controlled via predefined prosociality coefficient λ.

A crucial novelty of BAROCCO concerns the training of the social component. Instead of combining agents' rewards with *SW* function like in eq. 1, we directly combine agents' values. While in certain cases the two approaches are mathematically equivalent, the latter approach might be more suitable for mixed environments and supports a broader choice of *SW* functions. For example, an alternative to fairness through selfishness could be to train a fair centralized system to maximize minimum of agents' payoffs. We show that this can be viable if the system is trained with BAROCCO.

3 EXPERIMENTS AND DISCUSSION

We conduct our experiments in the Harvest environment [14], where five agents collect apples on a partially observable grid-like map. The regrowth rate of apples increases with the number of uncollected apples nearby. Therefore, the agents that harvest every apple in sight quickly exhaust its supplies. The optimal strategy for a group of agents is to balance harvesting and cultivating apples.

First, we evaluate the performance of BAROCCO, i.e. the total reward obtained by all agents. To this end, we compare it with several baselines: selfish decentralized agents, prosocial decentralized agents trained via CRS, prosocial centralized agents trained via COMA. We fix $\lambda = 1$ and *SW* function as sum for BAROCCO, CRS, and COMA. The results are presented in Figure 2a. 'BAROCCO, sum' performs slightly better than 'COMA, sum', meaning that our modifications of the training procedure can be beneficial. However, both underperform compared to 'CRS, sum', suggesting that additional complexity of centralized algorithms can hinder performance in some environments. This result contradicts the findings of the prior literature where centralization consistently improved performance [10, 21]. However, the algorithms suggested in this literature were not tested in complex mixed environments like Harvest before.

Second, we test whether BAROCCO can achieve both fairness and performance in a centralized training setup. To this end, we set *SW* as minimum, assuming that optimizing minimum of agents' payoffs might lead to fair reward distribution. From Figure 2b we can see that this is indeed the case: choosing *SW* as minimum yields more even distribution than *SW* sum. Unsurprisingly, this comes at some cost of performance (Fig. 2a). It is worth noting that performance of CRS and COMA plummets when sum is replaced with minimum as *SW* (Fig. 2a), which highlights flexibility of BAROCCO.

Third, we examine the effect of varying prosociality coefficient λ in BAROCCO. As expected, increasing λ improves performance (Fig. 2c) but can result in unfair reward allocation (Fig. 2d). We note that prosociality coefficient should be treated as a hyperparameter and appropriately tuned for a given environment to maximize performance while keeping fairness above an acceptable level.

This work contributes to the broader discussion of what constitutes cooperation. Most MARL papers that study mixed environments focus on efficiency, but we argue that this metric can be too limiting. Agents that act towards a single common goal are more reminiscent of a swarm system than a group of distinct individuals that could mutually benefit from cooperation. We explore ways to incorporate the notion of fairness into such systems, either by preserving some individuality of the agents or by modifying the centralized objective. We hope that our work sparks further discussion regarding other desirable qualities of multi-agent systems and the means to achieve these qualities.

ACKNOWLEDGMENTS

This research was supported in part through computational resources of HPC facilities at HSE University. Support from the Basic Research Program of the National Research University Higher School of Economics is gratefully acknowledged.

REFERENCES

[1] Adrian K Agogino and Kagan Tumer. 2004. Unifying temporal and structural credit assignment problems. In *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems-Volume 2*. IEEE Computer Society, 980–987.

[2] Michael S Alvard. 2004. The ultimatum game, fairness, and cooperation among big game hunters. *Foundations of human sociality* (2004), 413–435.

[3] R Axelrod and WD Hamilton. 1981. The evolution of cooperation. *Science* 211, 4489 (1981), 1390–1396. https://doi.org/10.1126/science.7466396 arXiv:https://science.sciencemag.org/content/211/4489/1390.full.pdf

[4] Joyce F Benenson, Joanna Pascoe, and Nicola Radmore. 2007. Children's altruistic behavior in the dictator game. *Evolution and Human Behavior* 28, 3 (2007), 168–175.

[5] Yongxiang Chen, Liqi Zhu, and Zhe Chen. 2013. Family income affects children's altruistic behavior in the dictator game. *PloS one* 8, 11 (2013).

[6] Andrew M Colman. 2006. The puzzle of cooperation. *Nature* 440, 7085 (2006), 744–745.

[7] Rachel Croson and Nancy Buchan. 1999. Gender and culture: International experimental evidence from trust games. *American Economic Review* 89, 2 (1999), 386–391.

[8] Ernst Fehr and Urs Fischbacher. 2002. Why social preferences matter–the impact of non-selfish motives on competition, cooperation and incentives. *The economic journal* 112, 478 (2002), C1–C33.

[9] Ernst Fehr and Klaus M Schmidt. 1999. A theory of fairness, competition, and cooperation. *The quarterly journal of economics* 114, 3 (1999), 817–868.

[10] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Thirty-second AAAI conference on artificial intelligence*.

[11] Joseph Henrich, Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr, Herbert Gintis, Richard McElreath, et al. 2001. Cooperation, reciprocity and punishment in fifteen small-scale societies. *American Economic Review* 91, 2 (2001), 73–78.

[12] Pablo Hernandez-Leal, Michael Kaisers, Tim Baarslag, and Enrique Munoz de Cote. 2017. A survey of learning in multiagent environments: Dealing with non-stationarity. *arXiv preprint arXiv:1707.09183* (2017).

[13] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor. 2019. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 33, 6 (2019), 750–797.

[14] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. In *Advances in neural information processing systems*. 3326–3336.

[15] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, Dj Strouse, Joel Z Leibo, and Nando De Freitas. 2019. Social Influence as Intrinsic Motivation for Multi-Agent Deep Reinforcement Learning. In *International Conference on Machine Learning*. 3040–3049.

[16] Dominic DP Johnson, Pavel Stopka, and Stephen Knights. 2003. The puzzle of human cooperation. *Nature* 421, 6926 (2003), 911–912.

[17] Sara Elisa Kettner and Israel Waichman. 2016. Old age and prosocial behavior: Social preferences or experimental confounds? *Journal of Economic Psychology* 53 (2016), 118–130.

[18] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in neural information processing systems*. 6379–6390.

[19] Alexander Peysakhovich and Adam Lerer. 2018. Prosocial learning agents solve generalized stag hunts better than selfish ones. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2043–2044.

[20] David G Rand and Martin A Nowak. 2013. Human cooperation. *Trends in cognitive sciences* 17, 8 (2013), 413–425.

[21] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. *arXiv preprint arXiv:1803.11485* (2018).

[22] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[23] John Scott. 2000. Rational choice theory. *Understanding contemporary society: Theories of the present* 129 (2000), 671–85.

[24] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296* (2017).

[25] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*. 330–337.

[26] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, et al. 2017. Starcraft ii: A new challenge for reinforcement learning. *arXiv preprint arXiv:1708.04782* (2017).

[27] Jane X Wang, Edward Hughes, Chrisantha Fernando, Wojciech M Czarnecki, Edgar A Duéñez-Guzmán, and Joel Z Leibo. 2019. Evolving intrinsic motivations for altruistic behavior. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 683–692.

[28] Weixun Wang, Jianye Hao, Yixi Wang, and Matthew Taylor. 2019. Achieving cooperation through deep multiagent reinforcement learning in sequential prisoner's dilemmas. In *Proceedings of the First International Conference on Distributed Artificial Intelligence*. 1–7.