Dynamic preference allocation for multi-objective, multi-agent reinforcement learning

Asad Jeewa [1] Anban Pillay [1] Jonathan Shock [2] Benjamin Rosman [3]

[1]University of KwaZulu-Natal [2]University of Cape Town [2]University of Witwatersrand

How can agents coordinate their behaviour to achieve their individual and the collective objectives of the system?

Motivation and Objective

Goal: To train RL agents to balance conflicting objectives in multi-agent environments.

Motivation: Robust reinforcement learning algorithms are required to solve complex real-world problems that necessitate coordination among multiple agents as well as reasoning about their collective and individual benefits.

Problem: These environments are challenging due to:

- Divergent Rewards
- Temporally and spatially extended, interdependent objectives
- Severe non-stationarity
- Partial observability
- Sparse Rewards

Case Study: Clean Up

- Agents are rewarded for eating apples.
- Apples grow at a rate based on the cleanliness of a nearby river.
- The reward for cleaning the river is implicit.
- There is a tension between the short-term individual incentive and the long-term group interest.
- A turn-taking approach that benefits all agents would be an ideal solution.

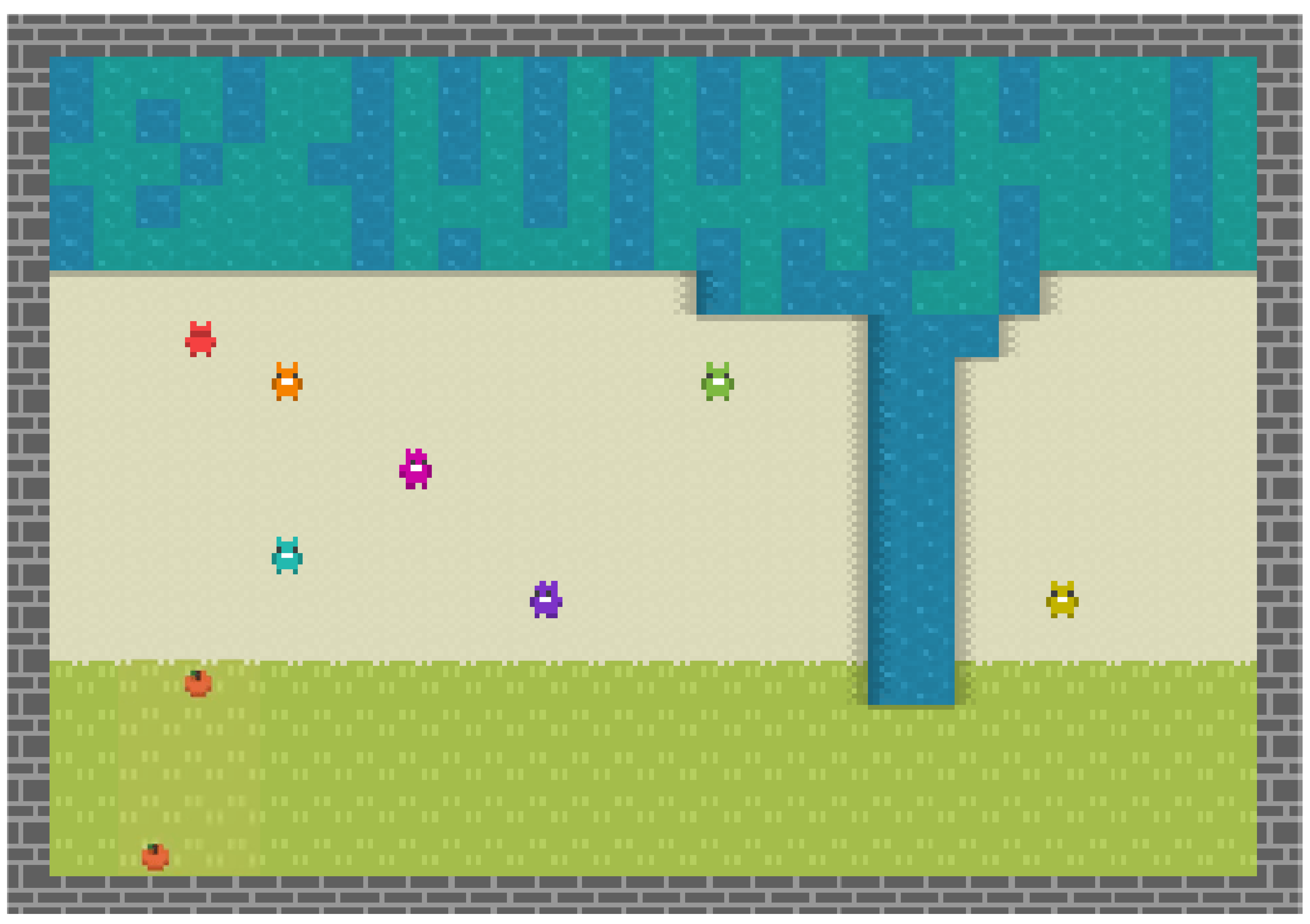
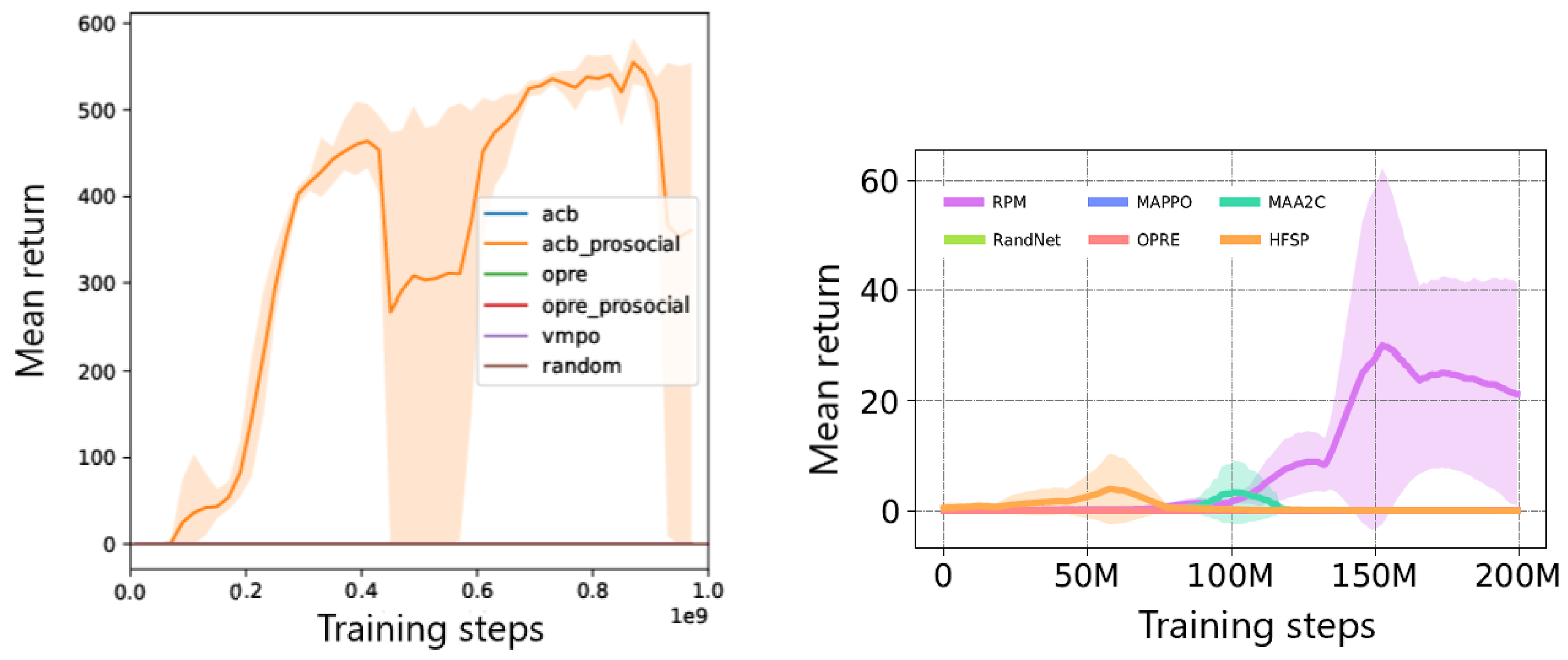


Figure 1. Clean Up Environment [1]

Current Approaches



(a) [1] (b) [2]

Figure 2. Training performance on Clean Up[1]

[1]Prosocial training uses the mean reward of all agents as a training signal. ACB is an actor-critic baseline built on top of A3C [1]. RPM is a novel MARL approach that randomly samples policies from a buffer [2]. Further details can be obtained from the papers.

- Numerous baselines fail to perform meaningful learning.
- Prosocial approaches have exhibited some success but such methods struggle with credit assignment [1].
- Agents should rather behave with a mixture of selflessness and selfishness [3].

Proposed Approach

We propose:

- Making the reward function explicit for all objectives.
- Vectorising the reward function to handle-trade-offs and incorporate preferences into the learning process [4].
- Training a universal policy that is parametrised by preferences [5].
- Introducing a high-level controller that dynamically allocates preferences to agents.
- Semi-sequential training for handling non-stationarity.

Future Work

- Dividing agents into teams
- Making the agents heterogeneous
- Modelling the behaviour of agents in the environment

References

[1] J. P. Agapiou, A. S. Vezhnevets, E. A. Duéñez-Guzmán, J. Matyas, Y. Mao, P. Sunehag, R. Köster, U. Madhushani, K. Kopparapu, R. Comanescu, D. J. Strouse, M. B. Johanson, S. Singh, J. Haas, I. Mordatch, D. Mobbs, and J. Z. Leibo, "Melting Pot 2.0," Jan. 2023, arXiv:2211.13746 [cs]. [Online]. Available: http://arxiv.org/abs/2211.13746

[2] W. Qiu, X. Ma, B. An, S. Obraztsova, Y. Shuicheng, and Z. Xu, "Rpm: Generalizable multi-agent policies for multi-agent reinforcement learning," in *The Eleventh International Conference on Learning Representations*, 2022.

[3] D. Radke, K. Larson, and T. Brecht, "The importance of credo in multiagent learning," in *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2023, pp. 2243–2252.

[4] C. F. Hayes, R. Rădulescu, E. Bargiacchi, J. Källström, M. Macfarlane, M. Reymond, T. Verstraeten, L. M. Zintgraf, R. Dazeley, F. Heintz, E. Howley, A. A. Irissappane, P. Mannion, A. Nowé, G. Ramos, M. Restelli, P. Vamplew, and D. M. Roijers, "A practical guide to multi-objective reinforcement learning and planning," *Autonomous Agents and Multi-Agent Systems*, vol. 36, no. 1, p. 26, Apr. 2022. [Online]. Available: https://doi.org/10.1007/s10458-022-09552-y

[5] T. Basaklar, S. Gumussoy, and U. Ogras, *PD-MORL: Preference-Driven Multi-Objective Reinforcement Learning Algorithm*, Aug. 2022.