Paper:

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

Dmitry Ivanov*, Vladimir Egorov*, Aleksei Shpilman

AAMAS 2021

Mixed MARL

Centralized Training with Decentralized Execution

- Algorithms like MADDPG [1]
- Addresses non-stationarity
- Reduce variance of PG

Cooperative Reward Shaping (CRS) [2]

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

- Each agent i maximizes mixture of selfish and social rewards
- Selfish reward
- Prosociality coefficient, between 0 and 1
- Social Welfare - sum or min of rewards of all agents

- Does not address credit assignment problem

References

[1] Lowe, Ryan, et al. "Multi-agent actor-critic for mixed cooperative-competitive environments." - 2017
[2] Peysakhovich, Alexander, et. al. "Prosocial learning agents solve generalized stag hunts better than selfish ones." - 2017
[3] Schulman, John, et al. "Proximal policy optimization algorithms." - 2017
[4] Rashid, Tabish, et al. "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning." - 2018.
[5] Foerster, Jakob, et al. "Counterfactual multi-agent policy gradients." - 2018
[6] Hughes, Edward, et al. "Inequity aversion improves cooperation in intertemporal social dilemmas." - 2018

Our Method: BAROCCO

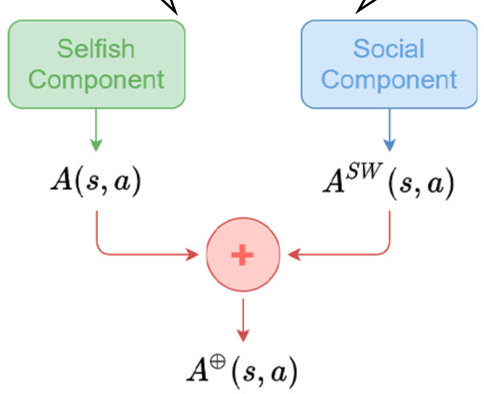
$$V_i^{\oplus}(s) = \mathbb{E}_{\pi_i} \sum_t \gamma^t \left((1-\lambda) r_{i_t} + \lambda SW(\mathbf{r}_t)\right)$$

$$= (1-\lambda)\mathbb{E}_{\pi_i} \sum_t \gamma^t r_{i_t} + \lambda \mathbb{E}_{\pi_i} \sum_t \gamma^t SW(\mathbf{r}_t)$$

$$= (1-\lambda) V_i(s) + \lambda V^{SW}(s)$$

Trained via MADDPG on selfish rewards

Trained via COMA on Social Values



Maximized via PPO [3]

Combining **selfish** and **social** incentives yields **tradeoff** between **group performance** and **fairness** in multi-agent systems

Cooperative MARL

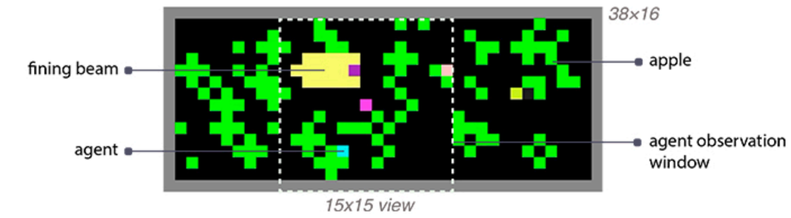
Centralized Training with Decentralized Execution

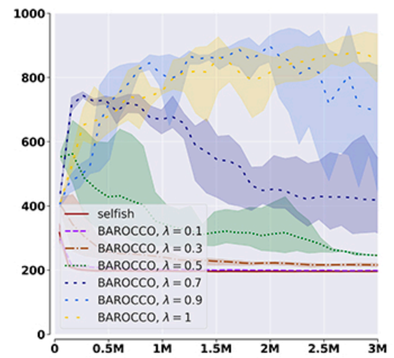
- Algorithms like QMIX [4] and COMA [5]
- Addresses non-stationarity and credit assignment
- Addresses growth of state and action spaces

- Inapplicable to mixed environments

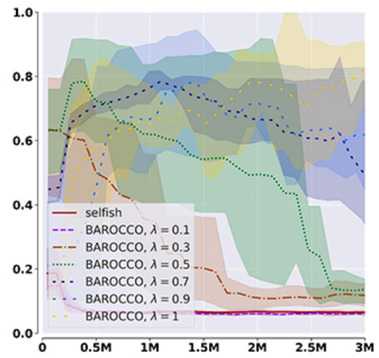
Results

Harvest Environment [6]





Group performance (Apples)



Fairness (Gini index)