Learning Fair Policies in Decentralized Cooperative Multi-Agent Reinforcement Learning

Matthieu Zimmer*, Claire Glanois*, Umer Siddique and Paul Weng

Univ. of Michigan-Shanghai Jiao Tong Univ. Joint Institute

ICML - July 2021

UNIVERSITY OF MICHIGAN

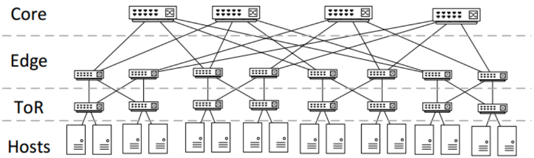
JOINT INSTITUTE

交大密西根学院

*Equal contribution

Motivations

Examples



Data center control

Traffic light control

Framework

Dec-POMDP $(\mathcal{S}, \mathcal{A} = (\mathcal{A}_i)_{i \in [N]}, (\mathcal{O}_i)_{i \in [N]}, P, (\Omega_i)_{i \in [N]}, \boldsymbol{r}, \gamma)$

- vectorial reward function $\boldsymbol{r} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^D$ where D is the number of users
- partial reward observability: an agent i observes $\boldsymbol{r}_{I_i} = (r_k)_{k \in I_i}$ where $I_i \subseteq [D]$

Propositions

Optimizing Social Welfare Functions

$$\max_{\boldsymbol{\theta}} \mathfrak{J}(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta}} \phi(\boldsymbol{J}_{I_1}(\theta_1), \ldots, \boldsymbol{J}_{I_N}(\theta_N))$$

where $\phi : \mathbb{R}^D \to \mathbb{R}$ is a social welfare function, such as:

- Generalized Gini social welfare function: $G_{\boldsymbol{w}}(\boldsymbol{u}) = \sum_{k \in [D]} w_k u_k^{\uparrow}$
- α-fairness: $\phi_\alpha(\boldsymbol{u}) = \sum_{k \in [D]} \frac{u_k^{1-\alpha}}{1-\alpha}$

Fairness properties

- Impartiality, Efficiency, Pigou-Dalton principle

Advantage sharing

- No need for a centralized critic
- Less communication needed

Theoretical Analysis

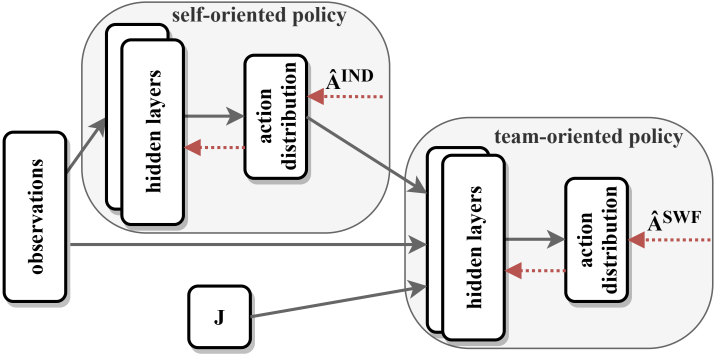
Theorem

Under standard assumptions, the SWF objective $\mathfrak{J}(\boldsymbol{\theta}^k)$ converges almost surely and with a sub-linear convergence rate within a radius of convergence $\tilde{\mathfrak{r}}$ of the optimal value $\mathfrak{J}^*$ where $\tilde{\mathfrak{r}}$ depend on the approximation errors of (a) estimating $\boldsymbol{J}$, (b) estimating $\boldsymbol{A}(\boldsymbol{o}, \boldsymbol{a})$, and (c) ignoring the effects of one agent's action over other agents.

- Corollary providing a high-probability bound on the number of iterations before convergence
- Reducing (b) by learning two critics per agents

Self-Oriented Team-Oriented (SOTO) Architecture

- Transfer learning with advice taking: the self-oriented policy advises the team-oriented policy
- Learning from two losses from two critics
- Progressively switch from the self-oriented policy to the team-oriented one



Conclusion

Fair optimization in multi-agent reinforcement learning

- Scalable (no centralized critic nor centralized policy)
- Evaluation on two scenarios and 5 domains
  - Centralized learning with decentralized execution
  - Fully decentralized
- Convergence proof

Future directions

- Learning the communications
- Relaxation of impartiality