Learning Fairness in Multi-Agent Systems

Jiechuan Jiang
Peking University
jiechuan.jiang@pku.edu.cn

Zongqing Lu
Peking University
zongqing.lu@pku.edu.cn

Overview

Problem:
In a multi-agent system, agents learn to share limited common resources (e.g. CPU and memory) to achieve **efficiency** and **fairness** simultaneously.

Solution:
- Propose **fair-efficient reward** to learn both efficiency and fairness.
- Design a **hierarchy** for easing the learning difficulty.
- Learning in a decentralized way coordinated by **average consensus**.

Motivation

Fairness is essential for human society, and also helps multi-agent systems become both efficient and stable.

However, learning efficiency and fairness simultaneously is a **complex**, **multi-objective**, **joint-policy** optimization.

Previous works are limited:

- The mainstream multi-agent RL methods do not consider the fairness.
- Existing work on fair division mainly focuses on **static settings**.
- Handcrafted RL methods designed for specific resource allocation applications require **domain-specific knowledge**.
- Methods for social dilemma might help, but they cannot guarantee the fairness.

Method

Fair-Efficient Network, FEN

The environmental reward r_i is only related to occupied resources of agent i.
The **coefficient of variation** of agents' utilities $u_t^i = \frac{1}{t}\sum_{j=0}^{t} r_j^i$ measures fairness.

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\frac{(u^i-\bar{u})^2}{\bar{u}^2}},$$ where $\bar{u}$ is agents' average utility.

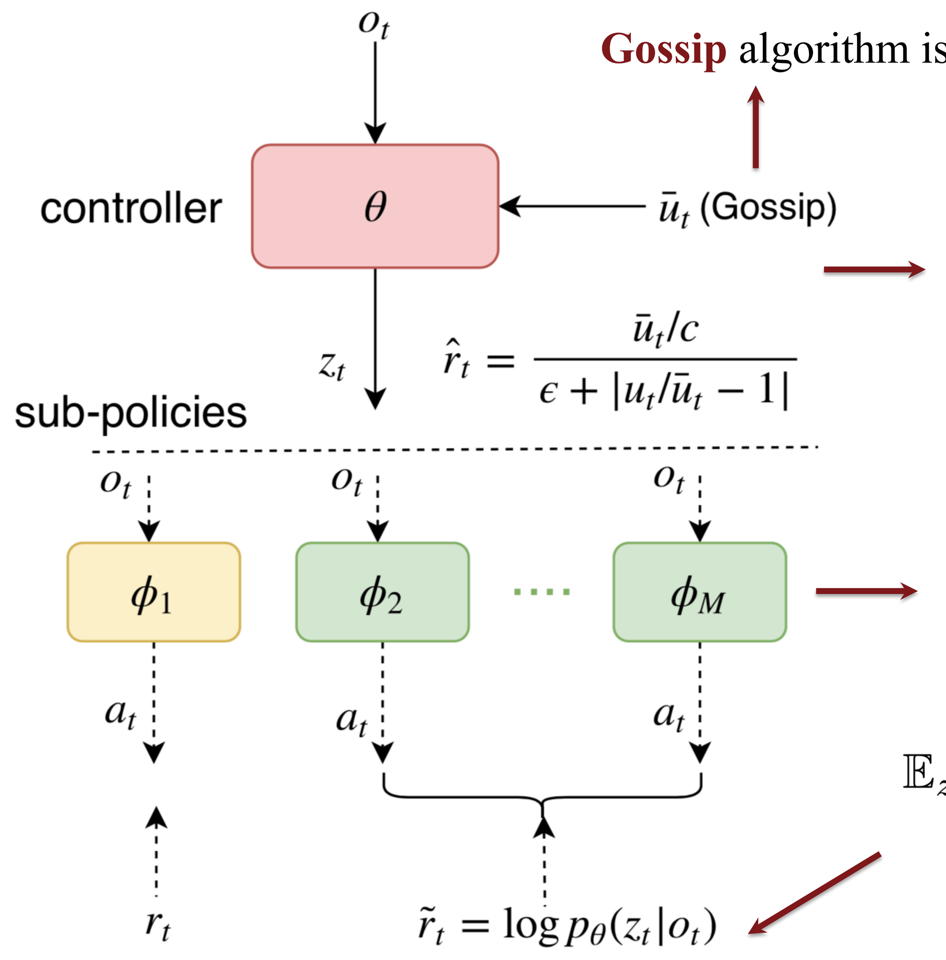
We propose fair-efficient reward for each agent $$\hat{r}_t^i = \frac{\bar{u}_t/c}{\epsilon + \left|u_t^i/\bar{u}_t - 1\right|}$$

$\bar{u}_t/c$ is the resource **utilization**, encouraging efficiency.
$\left|u_t^i/\bar{u}_t - 1\right|$ punishes the agent's utility **deviation** from the average.

Pareto efficiency and **equal allocation** are guaranteed in infinite-horizon sequential decision-making.

Hierarchy

Gossip algorithm is adopted for distributed average consensus.



environmental reward
(for **efficiency**)

Every T timesteps, the controller chooses one of the sub-policy to interact with the environment, and receives fair-efficient reward.

Information-theoretic objective
(exploring **diverse** possible **fairness** behaviors)

$$I(Z;O) + H(A|O)$$

$$\mathbb{E}_{z\sim\pi_\theta, o\sim\pi_\phi}[\log p(z|o)] + \mathbb{E}_{a\sim\pi_\phi}[-p(a|o)\log p(a|o)]$$

(selection probability output by the controller) (entropy regulariztion)

The hierarchy reduces the difficulty of learning both efficiency and fairness.

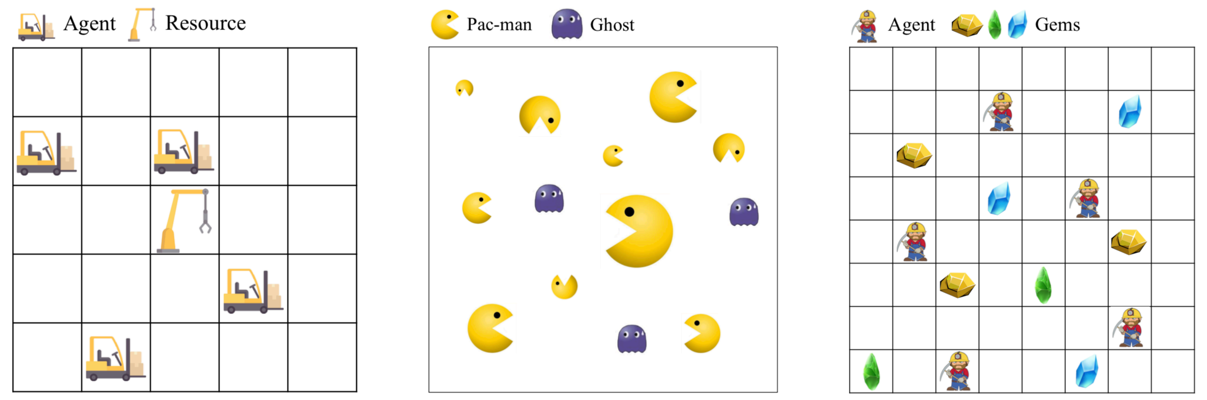
Controller:
- long-time horizon plan
- without direct interaction

Sub-policies:
- together decompose the complex objective
- each focuses its own easy objective

Algorithm 1 FEN training

```
1:  Initialize u_i, ū_i the controller θ and sub-policies φ
2:  for episode = 1, ..., M do
3:      The controller chooses one sub-policy φ_z
4:      for t = 1, ..., max-episode-length do
5:          The chosen sub-policy φ_z acts to the environment
            and gets the reward { r_t              if z = 1,
                                { log p_θ(z|o_t)   else
6:          if t%T = 0 then
7:              Update φ_z using PPO
8:              Update ū_i (with gossip algorithm)
9:              Calculate r̂^i = (ū_i/c) / (ε + |u^i/ū_i − 1|)
10:             The controller reselects one sub-policy
11:         end if
12:     end for
13:     Update θ using PPO
14: end for
```

Experiments



Job Scheduling   The Matthew Effect   Manufacturing Plant

Job Scheduling

	resource utilization	CV	min utility	max utility
Independent	96% ±11%	1.57 ±0.26	0	0.88 ±0.17
Min	47% ±8%	0.30 ±0.07	0.07 ±0.02	0.16 ±0.05
Avg	84% ±7%	0.75 ±0.13	0.05 ±0.03	0.46 ±0.17
Min+αAvg	63% ±5%	0.39 ±0.03	0.09 ±0.03	0.24 ±0.06
Inequity Aversion	72% ±9%	0.69 ±0.17	0.04 ±0.02	0.33 ±0.11
FEN	**90%** ±5%	**0.17** ±0.05	**0.18**±0.03	0.28 ±0.07
FEN w/o Hierarchy	57% ±13%	0.22 ±0.06	0.10 ±0.03	0.18 ±0.11

The effect of Hierarchy   High   Lowest   Highest

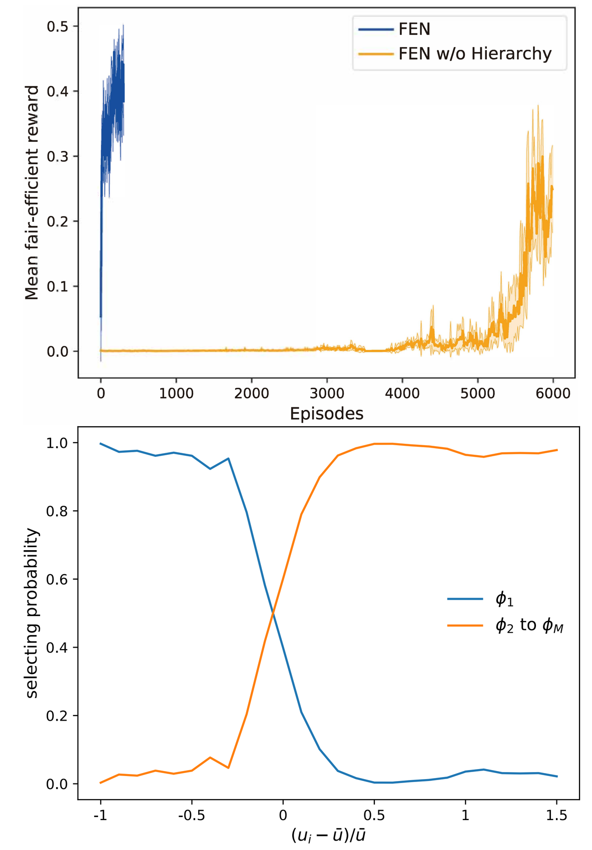
Manufacturing Plant

	resource utilization	CV	no. products
Independent	28% ±5%	0.38 ±0.08	19 ±3
Min	29% ±6%	0.26 ±0.01	20 ±4
Avg	13% ±3%	0.63 ±0.07	9 ±2
Min+αAvg	34% ±6%	0.28 ±0.01	23 ±4
Inequity Aversion	27% ±7%	0.29 ±0.07	19 ±5
FEN	**82%** ±5%	**0.10** ±0.03	**48** ±3
FEN w/o Hierarchy	22% ±3%	0.18 ±0.07	15 ±1

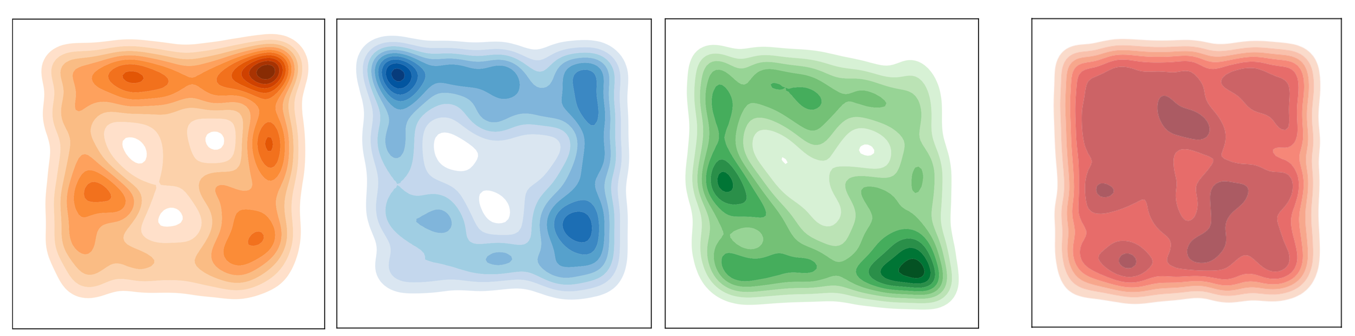
The highest products of FEN is the result of both efficiency and fairness.

Ablation

1. FEN learns much faster and converges to a higher fair-efficient reward than FEN w/o Hierarchy.
2. The controller is more likely to select ϕ_1 to occupy the resources when $u_i < \bar{u}$ and tends to select other sub-policies to maintain the fairness when $u_i > \bar{u}$.
3. Visualizations of position distribution verify the effect of the information-theoretic objective.



Three learned sub-policies   Random



*In the Matthew effect, we fix three ghosts at the center. The three learned sub-policies keep away from the three ghosts **for fairness** and their distributions are **different**, concentrated at different corners.