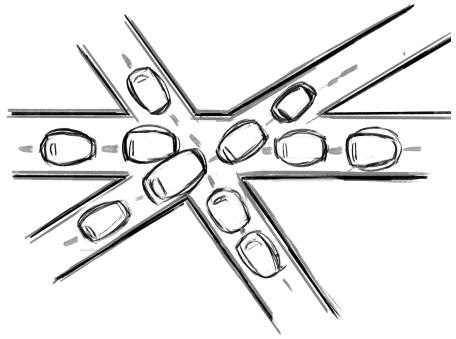
LOPT: Learning Optimal Pigovian Tax in Sequential Social Dilemmas

Yun Hua[1], Shang Gao[2], Wenhao Li[3] ,Haosheng Chen[2], **Xiangfeng Wang[2]**, Jun Luo[1] , Hongyuan Zha[4]

1. AnTai College of Economics and Management, Shanghai JiaoTong University
2. School of Computer Science and Technology, East China Normal University
3. School of Computer Science and Technology, Tongji University
4. Chinese University of Hong Kong, Shenzhen

LOPT: Learning Optimal Pigovian Tax in Sequential Social Dilemmas

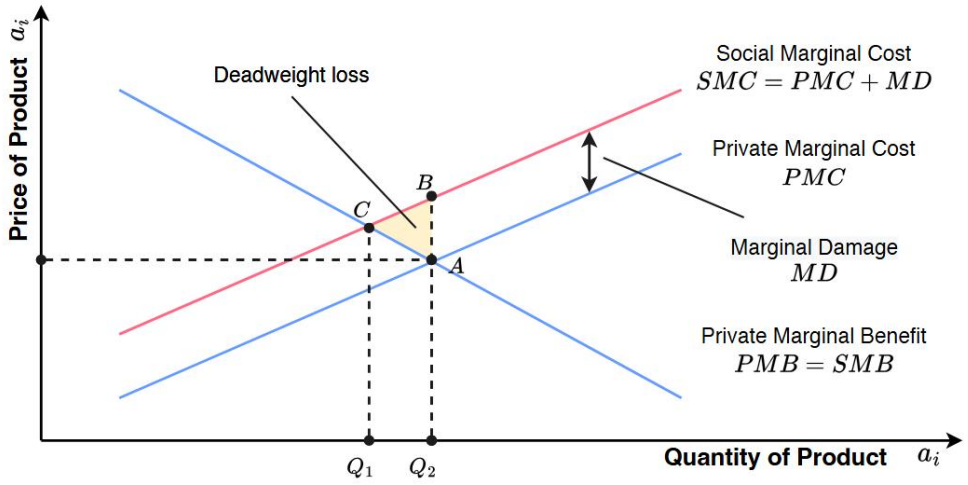
- Structural credit assignment in multi-agent reinforcement learning aims to characterize the influence of different agents' actions on the overall final outcome.
  - Current research on structural credit assignment in multi-agent reinforcement learning primarily focuses on fully cooperative tasks.
  - These methods have limitations in non-cooperative game scenarios, such as mixed multi-agent reinforcement learning settings.
  - In such scenarios, sequential social dilemmas can lead to difficulties in credit assignment. (Agents may harm each other's interests.)



Sequential Social Dilemma Example: Traffic Congestion

LOPT: Learning Optimal Pigovian Tax in Sequential Social Dilemmas

- This paper proposes a solution to this problem through Pigouvian reward shaping based on the externality principle.
- It introduces the economic concept of externalities to quantify credit assignment in mixed multi-agent reinforcement learning.
- It provides a mathematical definition and an effective quantification method for credit assignment in this context.



Definition 1. *An **externality** occurs whenever one economic actor's activities affect another's activities in ways that are not reflected in market transactions [25].*

Definition 2. *An **Externality** occurs whenever an agent's actions affect others in ways that are not reflected in individual local rewards.*

LOPT: Learning Optimal Pigovian Tax in Sequential Social Dilemmas

- This paper proposes a solution to this problem through Pigouvian reward shaping based on the externality principle.
- It introduces the economic concept of externalities to quantify credit assignment in mixed multi-agent reinforcement learning.
- This leads to the presentation of the optimal Pigouvian reward:

$$E^i\left(s, \boldsymbol{a}^{-i^*}, a^i\right) = Q^*\left(s, \boldsymbol{a}^*\right) - Q\left(s, \boldsymbol{a}^{-i^*}, a^i\right),$$

$$F^i\left(s, \boldsymbol{a}^{-i^*}, a^i\right) = Q^*\left(s, \boldsymbol{a}^*\right) - Q\left(s, \boldsymbol{a}^{-i^*}, a^i\right)$$

It cannot be obtained directly and requires approximation.

$$\hat{r}^i\left(s, \boldsymbol{a}\right) = r^i\left(s, \boldsymbol{a}\right) + F^i\left(s, \boldsymbol{a}^{-i^*}, a^i\right),$$

LOPT: Learning Optimal Pigovian Tax in Sequential Social Dilemmas

- We introduces the economic concept of externalities to quantify credit assignment in mixed multi-agent reinforcement learning.
- We develop a reward shaping framework based on the externality principle, termed "Pigouvian reward shaping" (a name inspired by the Pigouvian tax).

$$F^i_{\boldsymbol{\theta},\boldsymbol{\delta}}\left(s^t, \mathbf{a}^t\right) = F^i_{\boldsymbol{\theta},\boldsymbol{\delta}}\left(s^t, \mathbf{a}^t_{-i}{}^*, a^t_i\right).$$

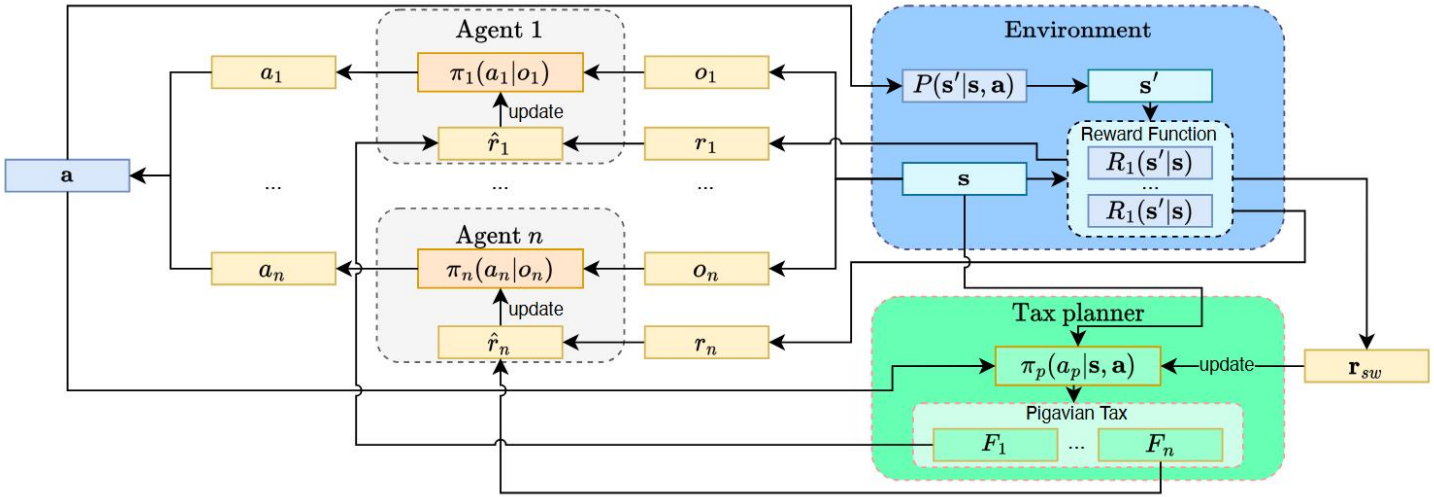


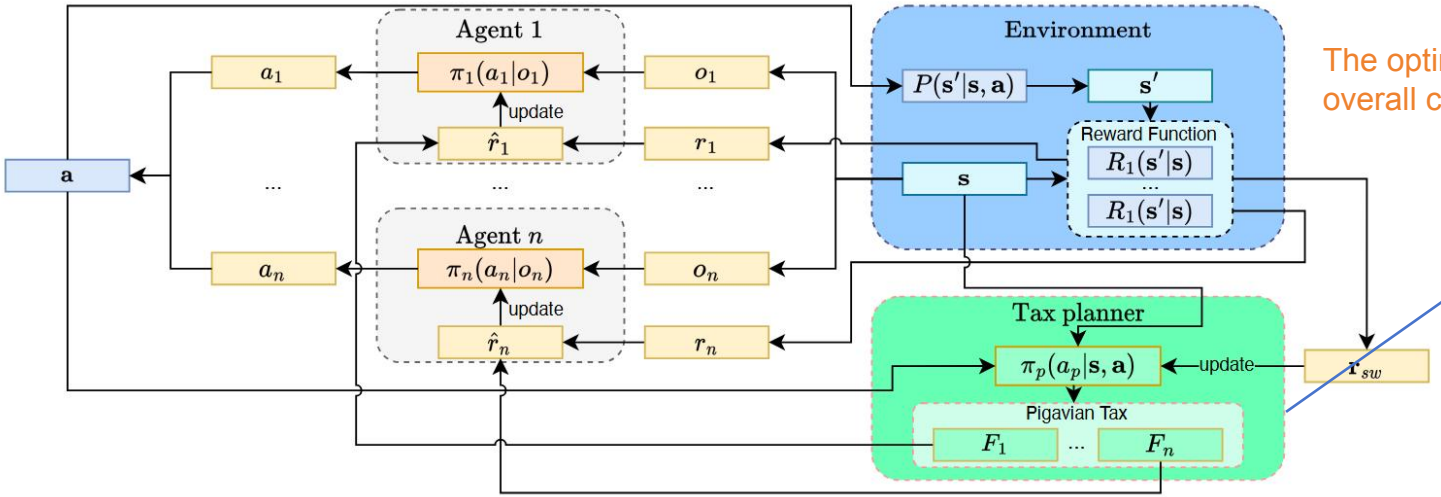
Figure 3: The Architecture of the **LOPT**. The centralized agent Tax planner allocate the Pigovian tax/allowance within a functional percentage formulation. Reward shaping is established based on the Pigovian tax/allowance to alleviate the social dilemmas.

LOPT: Learning Optimal Pigovian Tax in Sequential Social Dilemmas

- We introduces the economic concept of externalities to quantify credit assignment in mixed multi-agent reinforcement learning.
- We develop a reward shaping framework based on the externality principle, termed "Pigouvian reward shaping" (a name inspired by the Pigouvian tax).

$$F^i_{\boldsymbol{\theta},\boldsymbol{\delta}}\left(s^t, \mathbf{a}^t\right) = F^i_{\boldsymbol{\theta},\boldsymbol{\delta}}\left(s^t, \mathbf{a}^t_{-i}{}^*, a^t_i\right).$$

$$F^i_{\boldsymbol{\theta},\boldsymbol{\delta}}\left(s^t, \mathbf{a}^t_{-i}{}^*, a^t_i\right) = -\theta_i(s^t, \mathbf{a}^t) r_i\left(s^t, \mathbf{a}^t_{-i}{}^*, a^t_i\right) + \delta_i(s^t, \mathbf{a}^t) \sum_{j=0}^{N} \theta_j(s^t, \mathbf{a}^t) r_j\left(s^t, \mathbf{a}^t_{-i}{}^*, a^t_i\right).$$

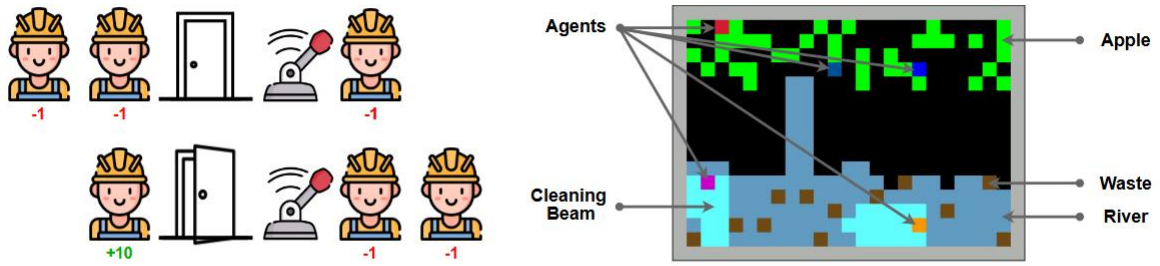


The optimization objective is to maximize the overall cumulative reward (i.e., social welfare).

Figure 3: The Architecture of the **LOPT**. The centralized agent Tax planner allocate the Pigovian tax/allowance within a functional percentage formulation. Reward shaping is established based on the Pigovian tax/allowance to alleviate the social dilemmas.

LOPT: Learning Optimal Pigovian Tax in Sequential Social Dilemmas

- For the experimental evaluation, two scenarios were adopted: "Escape Room" and "CleanUp".
  - The Escape Roomscenario represents a simple N-player prisoner's dilemma. It requires cooperation from at least M agents out of N to successfully escape the room.
  - The CleanUpscenario simulates a social dilemma in a more complex task. Agents gain individual rewards by collecting apples. However, if no agent performs the unrewarded cleaning work, the apple supply will be depleted, leading to a decrease in collective return for all.

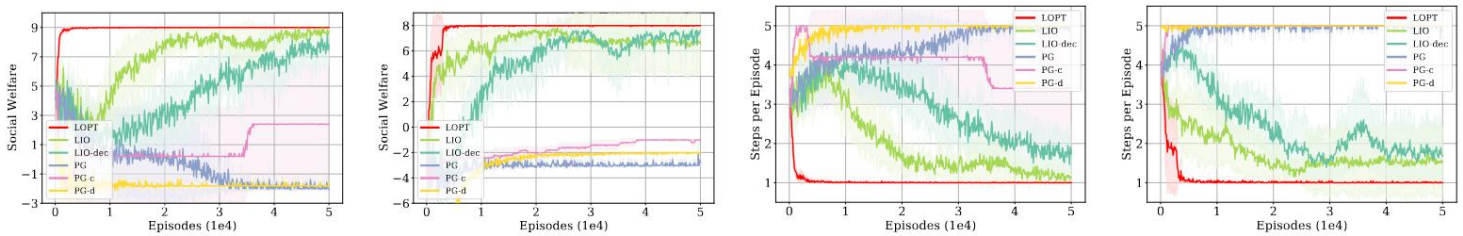


(a) Escape Room ($N = 3, M = 2$) (b) Cleanup($N = 5$)

Figure 4: Environment Examples

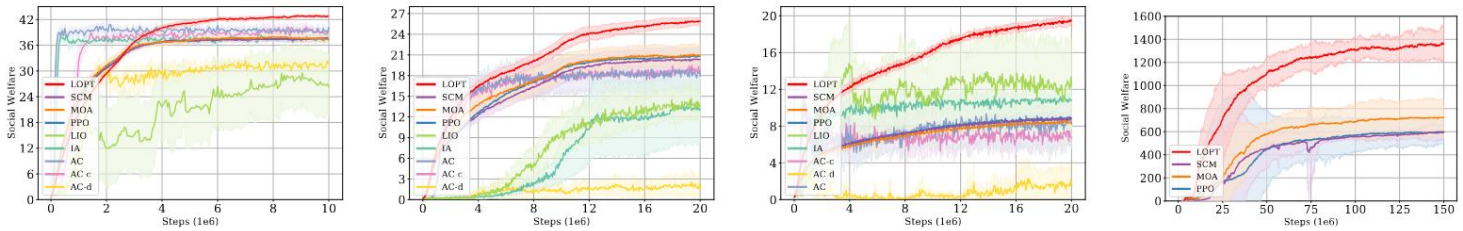
LOPT: Learning Optimal Pigovian Tax in Sequential Social Dilemmas

- Compared to the baseline methods, the reward shaping learned by LOPTRS significantly improves the total long-term cumulative reward in social dilemma scenarios.



(a) $N = 2$ (b) $N = 3$ (c) $N = 2$ (d) $N = 3$

Figure 5: Results on Escape Room Environment. (5(a), 5(b)) shows the learning curves of the proposed **LOPT**; which converges to the optimum and successfully solves the Escape Room social dilemmas. (5(c), 5(d)) shows that the proposed **LOPT** is able to end the episode in a single 1 step without any betrayal.



(a) $N = 2, (7 \times 7)$ (b) $N = 2, (10 \times 10)$ (c) $N = 2, (10 \times 10)$ Fixed (d) $N = 5, (18 \times 25)$

Figure 6: Results on Cleanup Environment. (6(a), 6(b)) shows the learning curves for the proposed **LOPT** in Cleanup($N = 2$); (6(c)) shows the learning curves for the proposed **LOPT** in Cleanup($N = 2$) with the fixed-orientated assumption. (6(d)) scales to a more complex environment with $N = 5$ agents.

Thank You