Learning Optimal "Pigovian Tax" in Sequential Social Dilemmas

Extended Abstract

Yun Hua*
East China Normal University
Shanghai, China
yunhua@stu.ecnu.edu.cn

Shang Gao*
East China Normal University
Shanghai, China
shanggao@stu.ecnu.edu.cn

Wenhao Li
The Chinese University of Hong Kong, Shenzhen
Shenzhen, China
liwenhao@cuhk.edu.cn

Bo Jin
Tongji University
Shanghai, China
bjin@tongji.edu.cn

Xiangfeng Wang†
East China Normal University
Shanghai, China
xfwang@cs.ecnu.edu.cn

Hongyuan Zha
The Chinese University of Hong Kong, Shenzhen& Shenzhen Institute of AI and Robotics for Society
Shenzhen, China
zhahy@cuhk.edu.cn

ABSTRACT

In multi-agent reinforcement learning (MARL), each agent acts to maximize its individual accumulated rewards. Nevertheless, individual accumulated rewards could not fully reflect how others perceive them, resulting in selfish behaviors that undermine global performance, which brings the social dilemmas. This paper adapt the famous externality theory in economic area to analyze social dilemmas in MARL, and propose the method called **L**earning **O**ptimal **P**igovian **T**ax (LOPT) to internalize the externalities in MARL. Furthermore, a reward shaping mechanism based on the approximated optimal "Pigovian Tax" is applied to reduce the social cost of each agent and tries to alleviate the social dilemmas. Compared with existing state-of-the-art methods, the proposed LOPT leads to higher collective social welfare in both the Escape Room and the Cleanup environments, which shows the superiority of our method in solving social dilemmas.

KEYWORDS

Multi-Agent Reinforcement Learning; Sequential Social Dilemmas; Reward Shaping; Externality

ACM Reference Format:
Yun Hua, Shang Gao, Wenhao Li, Bo Jin, Xiangfeng Wang, Hongyuan Zha. 2023. Learning Optimal "Pigovian Tax" in Sequential Social Dilemmas: Extended Abstract. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), London, United Kingdom, May 29 – June 2, 2023*, IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement Learning [21] has achieved wide success in various tasks [8, 10, 15, 27] and has been successfully expanded into the multi-agent area, especially in fully-cooperative games [13, 24, 25].

However, most recent centralized learning [4, 17, 18, 20] and decentralized learning methods [2, 19, 22] is either not suitable for self-interested agents or have difficulty in dealing with coordination among agents.. In many real-world environments with mixed motives, such as those within exclusionary and subtractive common-pool resources [11, 12, 16], selfish agents may fall into social dilemmas because of the temptation to evade any cost, which harms social welfare.

The concept of the social dilemma originates from economics and describes the situations in which individual rationality leads to collective irrationality [9]. In multi-agent reinforcement learning, it is specified as a conflict between agents' self-interest based on their local rewards and social welfare [11]. *Externality theory* is proposed to deal with social dilemmas in economics [23], which present whenever the well-being of a consumer or the production possibilities of a firm are directly affected by the actions of another agent [14]. Therefore, it may become a practical tool to measure self-interested agents' influence on social welfare.

In this paper, we introduce the externality to analysis social dilemma in MARL. Furthermore, motivated by "Pigovian Tax", which is one of the most popular solutions [3] in non-market economics [1, 3] to deal with externalities. We build a typical reward shaping mechanism to promote social welfare.

Our proposed method is called **L**earning **O**ptimal **P**igovian **T**ax (**LOPT**), where a centralized agent, called **Tax planner**, is built to learn the Pigovian tax/allowance based on the global reward. In learning process, Tax planner aims to maximizes the long-term global reward, which is equivalent to approximating the optimal Pigovian tax. Based on the learned tax/allowance rates, a reward shaping with a distinctive structure, *Optimal Pigovian Tax Reward Shaping*, is established. As a result, such a reward shaping structure visualizes each agent's social cost and alleviates the social dilemmas.

2 METHOD

Externality in MARL: In economics, an externality occurs whenever the activities of one economic actor affect the activities of another in ways that are not reflected in market transactions [14]. In this paper, we expand the definition of externality to the multi-agent reinforcement learning area:

* These authors have contributed equally to this work; † Corresponding author (xfwang@cs.ecnu.edu.cn).

Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023), A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom.

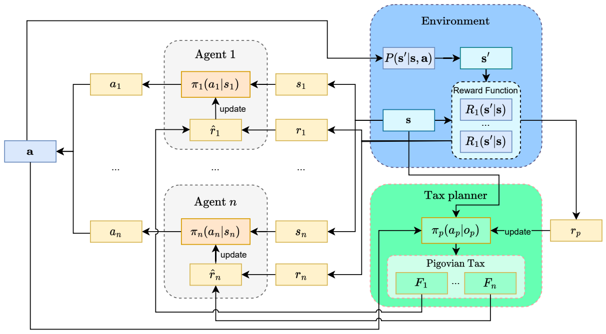


Figure 1: The Architecture of the LOPT.

Definition 1. *An* ***externality*** *occurs whenever the actions of an agent affect others in ways that are not reflected in local rewards.*

We consider a decentralized multi-agent reinforcement learning scenario with a N-player partially observable general-sum Markov game on a finite set of states $\mathcal{S}$. In each timestep, agents receive their d-dimensional views from the observation function $O : s \times \{1, ..., N\} \rightarrow \mathbb{R}^d$ based on the current state $s \in \mathcal{S}$. Then, agents select action $\{a_i\}_{i=1}^N \in \{\pi_i(a|o_i)\}_{i=1}^N$ from the set of actions $\{\mathcal{A}\}_{i=1}^N$, which transfers to the next states s' according to the transition function $P(s|\{a_i\}_{i=1}^N)$. Based on Defination. 1, the externality of agent i can be defined as:

$$E^i\left(s, \mathbf{a}_{-i}{}^*, a_i\right) = Q^*\left(s, \mathbf{a}^*\right) - Q\left(s, \mathbf{a}_{-i}{}^*, a_i\right), \tag{1}$$

where $Q^*\left(s, \mathbf{a}^*\right)$ is the optimal joint state action value and $Q\left(s, \mathbf{a}_{-i}{}^*, a_i\right)$ is the joint state action value with agent i's current action and other agents' optimal actions. For internalizing the externality and solving the social dilemma. The optimal Pigovian tax based reward shaping is written as follows:

$$F_i\left(s, \mathbf{a}_{-i}{}^*, a_i\right) = Q^*\left(s, \mathbf{a}^*\right) - Q\left(s, \mathbf{a}_{-i}{}^*, a_i\right). \tag{2}$$

It can further be reshaped as follows:

$$F_*^i\left(s^t, \mathbf{a}_{-i}^t{}^*, a_i^t\right) = \sum_{j=0}^N r_j\left(s^t, \mathbf{a}^{t*}\right) - \sum_{j=0}^N r^j\left(s^t, \mathbf{a}_{-i}^t{}^*, a_i^t\right). \tag{3}$$

Learning **O**ptimal **P**igovian **T**ax (**LOPT**) method is proposed to learn the optimal Pigovian tax based reward shaping. In LOPT, we design the Pigovian tax reward shaping within percentage tax/allowance formulation as:

$$\begin{aligned} F_{\theta,\delta}^i\left(s^t, \mathbf{a}_{-i}^t{}^*, a_i^t\right) =& -\theta_i(s^t, \mathbf{a}^t) r_i\left(s^t, \mathbf{a}_{-i}^t{}^*, a_i^t\right) + \\ & \delta_i(s^t, \mathbf{a}^t) \textstyle\sum_{j=0}^N \theta_j(s^t, \mathbf{a}^t) r_j\left(s^t, \mathbf{a}_{-i}^t{}^*, a_i^t\right). \end{aligned} \tag{4}$$

where $\boldsymbol{\theta}$ is the tax rates on all agents, θ_i is the specific tax rate for agent i, while $\boldsymbol{\delta}$ is the allowance rates on all agents, δ_i is the specific allowance rate for agent i. They are treated as functions based on the current joint state and action. Learning the optimal Pigovian tax reward shaping needs to learn $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ so as to let all $F_{\theta,\delta}^i\left(s^t, \mathbf{a}_{-i}^t{}^*, a_i^t\right)$ equal to the $F_*^i\left(s^t, \mathbf{a}_{-i}^t{}^*, a_i^t\right)$.

From Figure 1, The LOPT uses a centralized tax planner for learning Pigovian tax-based reward shaping functions. It can be described as a centralized reinforcement learning agent: $\left\langle \mathcal{S}_p, O_p, \mathcal{A}_p, R_p \right\rangle$, where $\mathcal{S}_p$ is the global state space for the tax planner, and O_p is the observation function to get observation o_p from its global state, $\mathcal{A}_p$ is its action space, and $\mathcal{R}_p$ is the reward function for it. Typically, the observation in timestep t, $o_p^t = \langle s^t, \mathbf{a}^t \rangle$ includes these general agents' joint state and action in the same timestep, while the action in timestep t includes the tax rates and allowance rates for all general agents $a_p^t = \langle \boldsymbol{\theta}^t, \boldsymbol{\delta}^t \rangle$. In the training process, we use the approximated state action function $Q_p(o_p, a_p)$ to replace the cumulative global reward (Social Welfare), the gradient loss for the tax planner is:

$$\mathbb{E}_{\pi_p^{\phi_p}}\left[\nabla_{\pi_p^{\phi_p}} \log \pi_p\left(a_p^t \mid o_p^t\right) Q^{p,\pi_p^{\phi_p}}\left(o_p^t, a_p^t\right)\right] + \eta f\left(\pi_p^{\phi_p}\right), \tag{5}$$

where $f(\pi_p) = \left|\sum_{t=0}^T \sum_{i=0}^T F_{\theta,\delta}^i\left(o^t, \mathbf{a}_{-i}^t{}^*, a_i^t\right)\right|$, which is the entropy to maintain the balance on tax and allowance. In light of the learning process of the tax planner, other general agents are trained within the approximated optimal Pigovian tax reward shaping as follows:

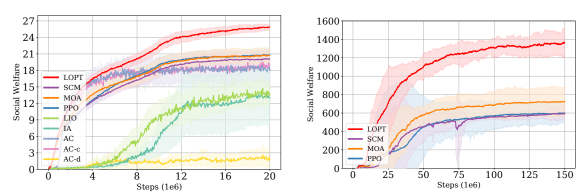
$$\mathcal{L}\left(\phi_i\right) = \mathbb{E}_{\pi_i^{\phi_i}}\left[\nabla_{\pi_i^{\phi_i}} \log \pi^i\left(a_i \mid s\right) \hat{Q}^{i,\pi_i^{\phi_i}}(s, \mathbf{a})\right], \tag{6}$$

where function $\hat{Q}^{i,\pi_i^{\phi_i}}(s, \mathbf{a})$ is defined as:

$$\hat{Q}^{i,\pi_{\phi_i}^i}(s, \mathbf{a}) = r_i(s, \mathbf{a}) + F^i\left(s, \mathbf{a}^{-i^*}, a_i\right) + \gamma \max_{\mathbf{a}'} \hat{Q}^{i,\pi_i^{\phi_i}}(s', \mathbf{a}'). \tag{7}$$

3 EXPERIMENT

To benchmark LOPT we use Cleanup [6], a set of environments with social dilemmas. We compare LOPT with **LIO** [26], **IA** [6], **MOA** [7], **SCM** [5], and common reinforcement learning algorithms in previous works [5–7, 26]. Figure. 2 LOPT can reach better social welfare, especially in more complex Cleanup($N = 5$) scenarios.



(a) Learning Curves for $N = 2$, 10×10 Map **(b) Learning Curves for $N = 5$, 18×25 Map**

Figure 2: Results on Cleanup Environment.

4 CONCLUSION

In this paper, the externality theory is first introduced to analysis social dilemmas in MARL. Based on it, **L**earning **O**ptimal **P**igovian **T**ax method is proposed to deal with social dilemmas. In LOPT, the tax planner learns each agent's tax/allowance allocation policy. Pigovian tax reward shaping internalizes each agent's externality to encourage them to promote social welfare. Experiments have shown the superiority of the proposed mechanism for alleviating social dilemmas in MARL. In the future, we aim to build a decentralized Pigovian tax/allowance mechanism to learn the reward shaping to internalize agents' externality with lower computation complexity.

5 ACKONWLEDGEMENT

This work was supported in part by National Key Research and Development Program of China (No. 2020AAA0107400), Postdoctoral Science Foundation of China (No. 2022M723039), NSFC (No. 12071145), STCSM (No. 22QB1402100), Shenzhen Science and Technology Program (No. JCYJ20210324120011032), Shanghai Trusted Industry Internet Software Collaborative Innovation Center and a grant from Shenzhen Institute of Artificial Intelligence and Robotics for Society.

REFERENCES

[1] G Christopher Archibald. 1959. Welfare economics, ethics, and essentialism. *Economica* 26, 104 (1959), 316–327.

[2] Tamer Başar and Geert Jan Olsder. 1998. *Dynamic noncooperative game theory.* SIAM.

[3] Mark Blaug. 2011. Welfare economics. In *A Handbook of Cultural Economics, Second Edition*. Edward Elgar Publishing.

[4] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *AAAI*.

[5] HC Heemskerk. 2020. *Social curiosity in deep multi-agent reinforcement learning.* Master's thesis.

[6] Edward Hughes, Joel Z Leibo, Matthew Phillips, Karl Tuyls, Edgar Dueñez-Guzman, Antonio García Castañeda, Iain Dunning, Tina Zhu, Kevin McKee, Raphael Koster, et al. 2018. Inequity aversion improves cooperation in intertemporal social dilemmas. *NeurIPS* (2018).

[7] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A Ortega, DJ Strouse, Joel Z Leibo, and Nando de Freitas. 2018. Intrinsic social motivation via causal influence in multi-agent RL. (2018).

[8] Jens Kober, J Andrew Bagnell, and Jan Peters. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research* 32, 11 (2013), 1238–1274.

[9] Peter Kollock. 1998. Social dilemmas: The anatomy of cooperation. *Annual review of sociology* (1998), 183–214.

[10] Guillaume Lample and Devendra Singh Chaplot. 2017. Playing FPS games with deep reinforcement learning. In *AAAI*.

[11] JZ Leibo, VF Zambaldi, M Lanctot, J Marecki, and T Graepel. 2017. Multi-agent Reinforcement Learning in Sequential Social Dilemmas. In *AAMAS*.

[12] Adam Lerer and Alexander Peysakhovich. 2017. Maintaining cooperation in complex social dilemmas using deep reinforcement learning. *arXiv preprint arXiv:1707.01068* (2017).

[13] Wenhao Li, Xiangfeng Wang, Bo Jin, Dijun Luo, and Hongyuan Zha. 2022. Structured cooperative reinforcement learning with time-varying composite action space. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2022), 8618–8634.

[14] Andreu Mas-Colell, Michael Dennis Whinston, Jerry R Green, et al. 1995. *Microeconomic theory*. Vol. 1. Oxford university press New York.

[15] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529–533.

[16] Anatol Rapoport. 1974. Prisoner's dilemma—recollections and observations. In *Game Theory as a Theory of a Conflict Resolution*. 17–34.

[17] Tabish Rashid, Gregory Farquhar, Bei Peng, and Shimon Whiteson. 2020. Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *NeurIPS* (2020).

[18] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*.

[19] Jun Sun, Gang Wang, Georgios B Giannakis, Qinmin Yang, and Zaiyue Yang. 2020. Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In *AISTATS*.

[20] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2018. Value-Decomposition Networks For Cooperative Multi-Agent Learning Based On Team Reward. In *AAMAS*.

[21] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[22] Ming Tan. 1993. Multi-agent reinforcement learning: Independent versus Cooperative Agents. In *ICML*.

[23] Hanne van der Iest, Jacob Dijkstra, and Frans N Stokman. 2011. Not 'Just the two of us': Third party externalities of social dilemmas. *Rationality and Society* 23, 3 (2011), 347–370.

[24] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[25] Hua Wei, Nan Xu, Huichu Zhang, Guanjie Zheng, Xinshi Zang, Chacha Chen, Weinan Zhang, Yanmin Zhu, Kai Xu, and Zhenhui Li. 2019. Colight: Learning network-level cooperation for traffic signal control. In *CIKM*.

[26] Jiachen Yang, Ang Li, Mehrdad Farajtabar, Peter Sunehag, Edward Hughes, and Hongyuan Zha. 2020. Learning to incentivize other learning agents. *NeurIPS* (2020).

[27] Meixin Zhu, Xuesong Wang, and Yinhai Wang. 2018. Human-like autonomous car-following model with deep reinforcement learning. *Transportation Research Part C: Emerging Technologies* 97 (2018), 348–368.