Fast Adaptation to External Agents via Meta Imitation Counterfactual Regret Advantage

Extended Abstract

Mingyue Zhang
Key Lab of High Confidence Software Technologies (MoE),
Peking University
Beijing, China
mingyuezhang@pku.edu.cn

Zhi Jin
Key Lab of High Confidence Software Technologies (MoE),
Peking University
Beijing, China
zhijin@pku.edu.cn

Yang Xu
University of Electronic Science and Technology of China
Chengdu, China
xuyang@uestc.edu.cn

Zehan Shen
Nanjing University
Nanjing, China
shenzehan1995@smail.nju.edu.cn

Kun Liu
Key Lab of High Confidence Software Technologies (MoE),
Peking University
Beijing, China
kunl@pku.edu.cn

Keyu Pan
University of Electronic Science and Technology of China
Chengdu, China
201822080224@std.uestc.edu.cn

ABSTRACT

This paper focuses on the *multi-agent credit assignment* problem. We propose a novel multi-agent reinforcement learning algorithm called *meta imitation counterfactual regret advantage* (MICRA) and a three-phase framework for training, adaptation, and execution of MICRA. The key features are: (1) a *counterfactual regret advantage* is proposed to optimize the target agents' policy; (2) a meta-imitator is designed to infer the external agents' policies. Results show that MICRA outperforms state-of-the-art algorithms.

KEYWORDS

Multi-agent Reinforcement Learning; Meta Learning; Imitation Learning; Counterfactual Regret Minimization

ACM Reference Format:
Mingyue Zhang, Zhi Jin, Yang Xu, Zehan Shen, Kun Liu, and Keyu Pan. 2021. Fast Adaptation to External Agents via Meta Imitation Counterfactual Regret Advantage. In *Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), Online, May 3–7, 2021,* IFAAMAS, 3 pages.

1 INTRODUCTION

Many multi-agent reinforcement learning (MARL) problems are naturally modeled as mixed cooperative-competitive multi-agent systems [8, 12, 15]. Such a system usually involves two teams of agents, i.e., the *target agents* that are learning-based and therefore controllable agents, and the *external agents* that are not controllable by the learning algorithm. *Multi-agent credit assignment* is important in this mixed setting, that is, how to deduce the target agent's contribution to the team from a global reward. The key challenge in solving the problem is twofold, the *confounding* global reward [3, 5], and the *non-stationarity* of the external agents [1, 6].

For addressing the challenge, we design a *meta-imitation counterfactual regret advantage* (MICRA) algorithm, and propose a three-phase framework to support the training, adaptation, and execution of MICRA. The main features of our proposal are: (1) The framework introduces the training-adaptation paradigm in meta-learning, i.e., *online adaptation*, into the training paradigm in MARL, i.e., *offline training* and *online execution.* That is for using the meta-learning to avoid overfitting to certain policies of the external agents when training the policies of the target agents. (2) MICRA adopts the centralized critic to estimate the *counterfactual regret advantage* (CRA) for optimizing the target agent's policy. Here, we propose the *meta imitation learning* (MI) by combining the imitation learning with the meta learning to enable the algorithm being able to model the non-stationary policies of the external agents. In this way, fast adaptation to the changing policies is possible in online execution as the learning algorithm has already taken the changing external agents into account.

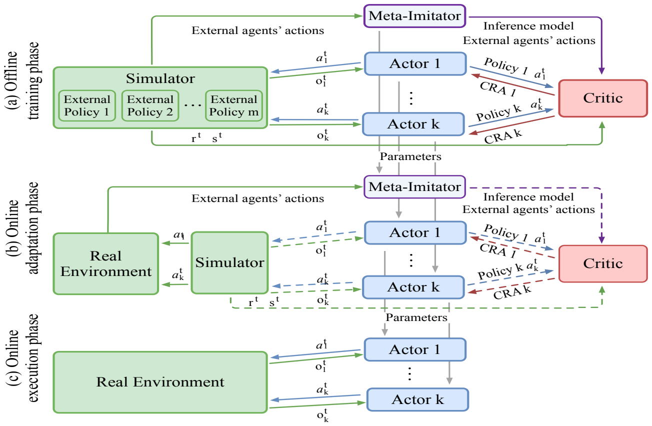


Figure 1: The Proposed Framework

Proc. of the 20th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2021), U. Endriss, A. Nowé, F. Dignum, A. Lomuscio (eds.), May 3–7, 2021, Online.

2 METHODS

Framework. As shown in figure 1, the proposed framework integrates the training paradigm in MARL, i.e., *centralized training and decentralized execution* (CTDE) [5, 8, 11], with the meta-learning process, i.e., *training-adaptation procedures* [4]. The features of the three phases in this framework are: (1) In *offline training* phase, MICRA (which consists of multiple independent *actors*, a *meta-imitator*, and a centralized *critic*) learns a *meta policy* over different external agents' policies; (2) In *online adaptation* phase, MICRA uses the meta policy to generate the real-taken policy in the real environment with the real external agents; (3) In *online execution* phase, each target agent takes actions independently by using the real-taken policy to complete the tasks in a collaborative manner without any centralized control.

Algorithm. A *counterfactual regret advantage* (CRA) is proposed based on COMA [5]. The main ideas are (1) A centralized critic evaluates a *regret* value for an agent with the assumption that other agents follow the current policies; (2) Multiple decentralized actors independently update their individual policies minimizing the regret value. The *immediate counterfactual regret advantage* is:

$$\begin{aligned}\mathcal{A}_{T,i,\pi^T}(s,\vec{a}) &= v_{\pi^T|s\mapsto a_i}(s) - v_{\pi^T}(s)\\ &= \sum_{\vec{a}_{\tau-i},\vec{a}_\epsilon} \pi^T_{\tau-i}(\vec{a}_{\tau-i}|s)\pi^T_\epsilon(\vec{a}_\epsilon|s)Q(s,[a_i,\vec{a}_{\tau-i},\vec{a}_\epsilon])\\ &- \sum_{\vec{a}_\tau,\vec{a}_\epsilon} \pi^T_\tau(\vec{a}_\tau|s)\pi^T_\epsilon(\vec{a}_\epsilon|s)Q(s,[\vec{a}_\tau,\vec{a}_\epsilon])\end{aligned} \tag{1}$$

where π^T denotes the policy at T learning episode, $\pi^T|s \mapsto a_i$ denotes that action a_i is always taken at state s, and the policy π is otherwise followed [7]; $\vec{a}_\tau$ denotes the joint actions of target agents; $\vec{a}_\epsilon$ denotes the joint actions of external agents; π_{-i} denotes the joint policy for all agents except agent i; $\pi_{\tau-i}$ denotes the joint policy for target agents except agent i. $\mathcal{A}_{T,i,\pi^T}$ is analogous to *immediate counterfactual regret* in *counterfactual regret minimization* (CFR) [7, 16, 18]. Eq.(1) is a general form of the advantage function; the advantage functions in previous works, e.g. [5, 13, 14], are in fact the special cases of Eq.(1).

A *discount cumulative* CRA is: $\mathcal{A}^\gamma_{T,i,\pi^T} = \gamma_c\mathcal{A}^\gamma_{T-1,i,\pi^{T-1}} + \mathcal{A}_{T,i,\pi^T}$ where $\gamma_c \in [0,1]$ is the discount rate. We further use the target Q-network to estimate discount cumulative CRA: $\mathcal{A}^\gamma_{T,i,\pi^T}(s,\vec{a}) \approx \gamma_c\Big(Q(s,\vec{a};\theta) - \sum_{a\in A_i}(\pi_i(a|o_i)Q(s,a,\vec{a}_{-i};\hat{\theta}))\Big) + \mathcal{A}_{T,i,\pi^T}(s,\vec{a})$. Then the CRA based policy gradient for agent i on trajectory data D is:

$$g_{\mathrm{cr},i} = \mathbb{E}_{s^t\sim D,\vec{a}^t\sim\pi}\Big[\sum_{t=0}^{H}\nabla_{\theta_i^a}\log(\pi_i(a_i^t|o_i^t;\theta_i^a))\mathcal{A}^\gamma_{i,\pi}(s^t,\vec{a}^t)\Big] \tag{2}$$

By following the line of *agent modeling* [2, 9], we propose a *meta imitation learning* (MI) to learn an inference model $\delta_i(o_i;\theta_{i,j}) : O_i \rightarrow \Delta(A_i)$ based on MAML [4]. It is used to predict the action taken by external agent i. External agents' joint policy π_ϵ is computed with $\pi_\epsilon(\vec{a}_\epsilon|s) = \sum_{i\in\epsilon}\delta_i(a_i|o_i)$. The loss function of $\delta_i(\cdot)$ is:

$$L^{im}_{\mathcal{H}_i}(\delta_i(\cdot;\theta_i)) = -\mathbb{E}_{(o^t,a^t)\sim\mathcal{H}_i}\Big[\sum_{k=1}^{|A_i|} I(a^{(k)},a^t)\log\delta_i(o^t;\theta_i)\Big] \tag{3}$$

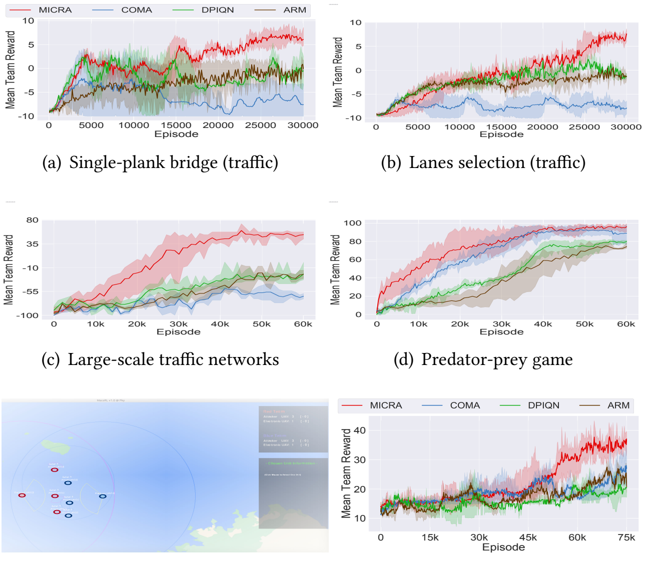
where $\mathcal{H}_i$ is the history behavior set of external agent i, $I(\cdot)$ is the ground truth indicator function. $\delta_i(\cdot)$ is performed via a multi-layer perceptron (MLP), of which the output layer is softmax. The objective of MI is:

$$\begin{aligned}&\min_{\theta_i}\sum_{\mathcal{T}_i\sim p(\mathcal{T})} L^{im}_{\mathcal{H}_i}(\delta(\cdot;\theta_i'))\\ &\text{s.t. } \theta_i' = \theta_i - \alpha_{\mathrm{adp}}\nabla_{\theta_i}L^{im}_{\mathcal{H}_i}(\delta(\cdot;\theta_i))\end{aligned} \tag{4}$$

where $p(\mathcal{T})$ is the distribution of all external agents' policies. θ_i is the meta parameters which will be used as initial parameters in online adaptation phase.

3 EVALUATION

We conduct several experiments, i.e. two standard MARL benchmarking tasks (*traffic control* and *predator-prey game*) in a grid environment [10, 17] and a practical application (electronic countermeasure based real-time strategy game). Three baseline MARL algorithms (COMA [5], DPIQN [6], and ARM [7]) are chosen for the experimental comparison. Figure 2 gives the results which show that our algorithm outperforms three baseline algorithms.



(a) Single-plank bridge (traffic)

(b) Lanes selection (traffic)

(c) Large-scale traffic networks

(d) Predator-prey game

(e) The environment of ECM-RTS game

(f) Learning curves in ECM-RTS game

Figure 2: Offline training: the learning curves on different tasks (red line is ours).

4 FUTURE WORK

How to deal with the dynamics of the system environment and the self-adaptation of the system is an important challenge in design complex cyber-physical systems. The proposed framework takes into account both concerns and is potentially evolved into a reference architecture. Modeling real *cyber-physical system* applications to verify its adaptability will be within our future work.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Hangyu Mao for comments on an early version of the paper. This work is supported by the National Natural Science Foundation of China under Grant No. 61620106007.

REFERENCES

[1] Maruan Al-Shedivat, Trapit Bansal, Yura Burda, Ilya Sutskever, Igor Mordatch, and Pieter Abbeel. 2018. Continuous Adaptation via Meta-Learning in Nonstationary and Competitive Environments. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

[2] Stefano V Albrecht and Peter Stone. 2018. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence* 258 (2018), 66–95.

[3] Yali Du, Lei Han, Meng Fang, Ji Liu, Tianhong Dai, and Dacheng Tao. 2019. LIIR: Learning Individual Intrinsic Reward in Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems, NeurIPS*. 4405–4416.

[4] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *ICML*.

[5] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[6] Zhangwei Hong, Shihyang Su, Tzuyun Shann, Yihsiang Chang, and Chunyi Lee. 2018. A Deep Policy Inference Q-Network for Multi-Agent Systems. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*. 1388–1396.

[7] Peter H. Jin, Kurt Keutzer, and Sergey Levine. 2018. Regret Minimization for Partially Observable Deep Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning,ICML (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 2347–2356.

[8] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*. 6379–6390.

[9] Hangyu Mao, Zhengchao Zhang, Zhen Xiao, and Zhibo Gong. 2019. Modelling the dynamic joint policy of teammates with attention multi-agent ddpg. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. 1108–1116.

[10] Hangyu Mao, Zhengchao Zhang, Zhen Xiao, Zhibo Gong, and Yan Ni. 2020. Learning multi-agent communication with double attentional deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 34, 1 (2020), 32.

[11] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 4292–4301.

[12] Mikayel Samvelyan, Tabish Rashid, Christian Schröder de Witt, et al. 2019. The StarCraft Multi-Agent Challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems, AAMAS*. 2186–2188.

[13] John Schulman, Philipp Moritz, Sergey Levine, Michael I Jordan, and Pieter Abbeel. 2015. High-Dimensional Continuous Control Using Generalized Advantage Estimation. *arXiv: Learning* (2015).

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

[15] Zheng Tian, Shihao Zou, Ian Davies, Tim Warr, Lisheng Wu, Haitham Bou-Ammar, and Jun Wang. 2020. Learning to Communicate Implicitly by Actions. In *2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference*. AAAI Press, 7261–7268.

[16] Runsheng Yu, Zhenyu Shi, Xinrun Wang, Rundong Wang, Buhong Liu, Xinwen Hou, Hanjiang Lai, and Bo An. 2019. Inducing Cooperation via Team Regret Minimization based Multi-Agent Deep Reinforcement Learning. *CoRR* abs/1911.07712 (2019). arXiv:1911.07712

[17] Lianmin Zheng, Jiacheng Yang, Han Cai, Ming Zhou, Weinan Zhang, Jun Wang, and Yong Yu. 2018. MAgent: A Many-Agent Reinforcement Learning Platform for Artificial Collective Intelligence. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*. AAAI Press, 8222–8223.

[18] Martin Zinkevich, Michael Johanson, Michael Bowling, and Carmelo Piccione. 2007. Regret Minimization in Games with Incomplete Information. In *Proceedings of the 20th International Conference on Neural Information Processing Systems* (Vancouver, British Columbia, Canada) *(NIPS'07)*. Curran Associates Inc., Red Hook, NY, USA, 1729–1736.