Zero-shot Adaptation to Partners in Mixed Cooperative-Competitive Tasks

Yuming Chen

Target: Develop a multi-agent reinforcement learning (MARL) algorithm to learn a policy that can cooperate in competition. Specifically, the policy can reach to a (relatively) **more efficient equilibrium** facing an unseen partner while maintaining a **high-level self-reward** in *mixed cooperative-competitive tasks*[12] (mixed task for short). Methodologically, we plan to discuss this problem from the perspective of general-sum games.

Introduction: Zero-shot adaptation to partners (ZSAP) is essential for agents in dynamic and open environments. It can be seen as the extension of ad hoc teamwork (AHT)[16], defined as the ability to exploit or coordinate with other partners with no prior knowledge. ZSAP has made progress in Markov games with special reward structures, including zero-sum reward (fully competitive) and potential reward (fully cooperative). However, ZSAP in general-sum games is not thoroughly discussed. Existing works related to general-sum games emphasize target and environment zero-shot ability. While theoretical efforts focus on matrix games[15] or assume the properties of partners[13, 5, 9].

We plan to develop an learning algorithm that satisfies our target with theory guarantee, and apply it to large scale environments, such as Hide-and-Seek[1], and real world swarms or even interaction with humans.

Related Works: Works related to ZSAP mostly discuss fully competitive tasks and cooperative tasks. For competitive tasks, including the GO[18] and Mahjong[10], the policies are trained from self-play with Policy Space Response Oracle (PSRO)[8] and evaluated against unseen opponents. In cooperative tasks, notably, [19] shows that agent trained with a variant of PSRO can coordinate with humans whose polices are not seen during training in overcooked[4]. Empirically, diversifying the meta-strategy of the other players in PSRO can enhance the ZSAP ability of agents in these tasks.

In mixed tasks, most policies are training under the paradigm of Centralized Training and Decentralized Executing (CTDE)[12], facing a fixed group of opponents[6] or jointly training all agents[17]. Such agents are not robust facing unseen partners as they have reached conventional equilibrium during training. Opponent Shaping[5] address this by assuming the learning dynamics of opponents. Following work[14] relaxes the assumption using meta-learning but can only apply to games with small scale. All the works do not directly address our target.

Sub-targets: To develop the desired algorithm, We plan to study the following sub-problems.
(1) Analyze the dynamics of general-sum games. Convergence in general-sum games is challenging[15]. We hope to prove the theoretical convergency of the algorithm, especially algorithm explicitly considering partners. Additionally, [2, 3] give us an inspiration to measure the equilibrium efficiency besides social welfare.
(2) Extend the above results to environments with large scale. from a model-free or a model-based approach. HATRPO[7] and HASAC[11] are good examples to guarantee monotonic improvement and convergency in large scale games. (Similar theoretical results are hopeful to be proved in individual-reward environments.)
(3) Apply the algorithm to real-world scenarios. Swarm, robotic arms, or interacting with humans. (If possible, we could try to cooperate with Intelligent Robotics Lab.)

References

[1] Bowen Baker, Ingmar Kanitscheider, Todor Markov, Yi Wu, Glenn Powell, Bob McGrew, and Igor Mordatch. Emergent tool use from multi-agent autocurricula. In *International Conference on Learning Representations*, 2020.

[2] David Balduzzi, Sebastien Racaniere, James Martens, Jakob Foerster, Karl Tuyls, and Thore Graepel. The mechanics of n-player differentiable games. In *International Conference on Machine Learning*, pages 354–363. PMLR, 2018.

[3] Ozan Candogan, Ishai Menache, Asuman Ozdaglar, and Pablo A Parrilo. Flows and decompositions of games: Harmonic and potential games. *Mathematics of Operations Research*, 36(3):474–503, 2011.

[4] Micah Carroll, Rohin Shah, Mark K Ho, Tom Griffiths, Sanjit Seshia, Pieter Abbeel, and Anca Dragan. On the utility of learning about humans for human-ai coordination. *Advances in neural information processing systems*, 32, 2019.

[5] Jakob N Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017.

[6] Qingxu Fu, Tenghai Qiu, Jianqiang Yi, Zhiqiang Pu, and Shiguang Wu. Concentration network for reinforcement learning of large-scale multi-agent systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9341–9349, 2022.

[7] Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. *CoRR*, abs/2109.11251, 2021.

[8] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[9] A Letcher, J Foerster, D Balduzzi, T Rocktaschel, and S Whiteson. Stable opponent shaping in differentiable games. In *2019 International Conference on Learning Representations*. OpenReview, 2019.

[10] Junjie Li, Sotetsu Koyamada, Qiwei Ye, Guoqing Liu, Chao Wang, Ruihan Yang, Li Zhao, Tao Qin, Tie-Yan Liu, and Hsiao-Wuen Hon. Suphx: Mastering mahjong with deep reinforcement learning. *arXiv preprint arXiv:2003.13590*, 2020.

[11] Jiarong Liu, Yifan Zhong, Siyi Hu, Haobo Fu, QIANG FU, Xiaojun Chang, and Yaodong Yang. Maximum entropy heterogeneous-agent reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

[12] Ryan Lowe, Yi I Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30, 2017.

[13] Christopher Lu, Timon Willi, Christian A Schroeder De Witt, and Jakob Foerster. Model-free opponent shaping. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14398–14411. PMLR, 17–23 Jul 2022.

[14] Christopher Lu, Timon Willi, Christian A Schroeder De Witt, and Jakob Foerster. Model-free opponent shaping. In *International Conference on Machine Learning*, pages 14398–14411. PMLR, 2022.

[15] Weichao Mao and Tamer Başar. Provably efficient reinforcement learning in decentralized general-sum markov games. *Dynamic Games and Applications*, 13(1):165–186, 2023.

[16] Reuth Mirsky, Ignacio Carlucho, Arrasy Rahman, Elliot Fosong, William Macke, Mohan Sridharan, Peter Stone, and Stefano V Albrecht. A survey of ad hoc teamwork research. In *European conference on multi-agent systems*, pages 275–293. Springer, 2022.

[17] Igor Mordatch and Pieter Abbeel. Emergence of grounded compositional language in multi-agent populations. *arXiv preprint arXiv:1703.04908*, 2017.

[18] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[19] DJ Strouse, Kevin McKee, Matt Botvinick, Edward Hughes, and Richard Everett. Collaborating with humans without human data. *Advances in Neural Information Processing Systems*, 34:14502–14515, 2021.