Quantifying Agent Interaction in Multi-agent Reinforcement Learning for Cost-efficient Generalization

Extended Abstract

Yuxin Chen
University of California, Berkeley
Berkeley, CA, USA
yuxinc@berkeley.edu

Chen Tang
The University of Texas at Austin
Austin, TX, USA
chen.tang@austin.utexas.edu

Ran Tian
University of California, Berkeley
Berkeley, CA, USA
rantian@berkeley.edu

Chenran Li
University of California, Berkeley
Berkeley, CA, USA
chenran_li@berkeley.edu

Jinning Li
University of California, Berkeley
Berkeley, CA, USA
jinning_li@berkeley.edu

Masayoshi Tomizuka
University of California, Berkeley
Berkeley, CA, USA
tomizuka@berkeley.edu

Wei Zhan
University of California, Berkeley
Berkeley, CA, USA
wzhan@berkeley.edu

ABSTRACT

Generalization in Multi-agent Reinforcement Learning (MARL) is challenging. Introducing a diverse set of co-play agents typically boosts the agent's generalization to unseen co-players. However, the extent to which an agent is influenced by co-players varies across scenarios and environments; thus, the improvement in generalization introduced by diversifying co-players also varies. In this work, we introduce *Level of Influence* (LoI), a novel metric measuring the interaction intensity among agents within a given scenario and environment. We show that LoI can effectively predict the disparities in the benefits of diversifying co-player distribution across scenarios, offering insights into optimizing training cost for varied situations. The code is available at: https://github.com/ThomasChen98/Level-of-Influence.

KEYWORDS

Multi-agent reinforcement learning, Learning agent-to-agent interactions, Multi-agent systems

ACM Reference Format:
Yuxin Chen, Chen Tang, Ran Tian, Chenran Li, Jinning Li, Masayoshi Tomizuka, and Wei Zhan. 2024. Quantifying Agent Interaction in Multi-agent Reinforcement Learning for Cost-efficient Generalization: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Developing agents that effectively interact with others remains a significant challenge [2, 5]. Model-free reinforcement learning (RL) has enabled agents to achieve or exceed human performance in various games and domains through self-play (SP) [6, 16, 17, 20, 22]. Yet, training exclusively with identical replicas limits their adaptability and robustness to new co-players with differing behaviors [3, 12, 15, 18]. Improving policy robustness through *diversifying co-player distribution* is a promising strategy [21]. This approach, effective in complex games, includes techniques such as population-based training (PP) [4, 9, 10], league-based training [20], fictitious self-play [7, 18], and agent hyperparameter diversification [8, 14]. Although these methods boost generalization, they often require more training resources and time as a trade-off.

Nevertheless, the key question remains: *whether diversifying co-player distribution during training justifies its high cost*, given that real-world applications often demand specialized RL policies for various scenarios [6, 13]. The value of this diversity hinges on the scenario's level of interaction. We introduce the *Level of Influence* (LoI), a metric quantifying how much an ego agent's reward changes with the behavior of other agents. Defined as the mutual information (MI) between the ego agent's reward and non-ego agents' policy choices, LoI helps in identifying when diversifying training partners can effectively improve an ego agent's generalization in various scenarios. Our findings demonstrate that the LoI metric is highly correlated with the benefits of increasing co-player diversity on the generalization of the ego agent within given scenarios. It indicates the potential of LoI in guiding training schedule optimization to achieve cost-effective generalization for MARL.

2 METHODOLOGY

In MARL, an agent's policy performance is gauged by its expected reward when interacting with various co-players, influenced by the reward structure or payoff matrix. We define *environments* as games with unique reward setups and *scenarios* as variations within these

environments, created by altering map features like size, shape, and obstacle placement.

To measure the intensity of interactions between agents in a scenario, we introduce a metric called the "Level of Influence" (LoI), drawing from the concept of causal influence [11]. Specifically, in this work, we consider a two-agent symmetric game with two agents named Alice and Bob, where Alice is controlled by our algorithm (the ego agent) with policy $\phi \in \Phi$ and Bob is another algorithm-driven agent or a human (the non-ego agent) with policy $\theta \in \Theta$. Alice and Bob choose their policies following the distribution $P_\varphi(\phi) = \mathbb{P}[\varphi = \phi]$ and $P_\vartheta(\theta) = \mathbb{P}[\vartheta = \theta]$ respectively. Let $r \in \mathbb{R}$ denote the total reward of Alice paired with Bob. Intuitively, we want the LoI to measure the degree to which Alice's reward distribution changes induced by Bob's policy choice, given Alice's own policy choice. Therefore, the LoI is defined as the conditional mutual information of Alice's reward and Bob's policy with respect to Alice's policy:

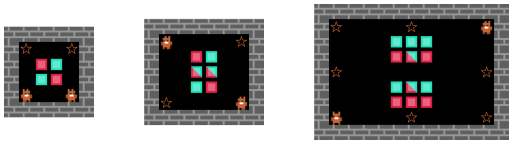
$$I(R;\vartheta|\varphi) = \sum_{\phi\in\Phi} P_\varphi(\phi) \sum_{\theta\in\Theta} P_\vartheta(\theta) D_{\mathrm{KL}}\left(P_{R|\vartheta=\theta,\varphi=\phi} \| P_{R|\varphi=\phi}\right), \quad (1)$$

where $P_{R|\vartheta=\theta,\varphi=\phi} = \mathbb{P}[R = r|\vartheta = \theta, \varphi = \phi]$ is the conditional reward distribution of Alice given Alice's policy $\varphi = \phi$ and Bob's policy $\vartheta = \theta$. Marginalizing ϑ results in $P_{R|\varphi=\phi}$.

When $I(R;\vartheta|\varphi) = 0$, Alice's reward is unaffected by Bob's policy choice, making diverse opponent training less beneficial. However, as this value increases, Bob's policy has a greater impact on Alice's reward. Thus, training Alice with a variety of Bob's policies enhances performance with new partners and justifies a larger training budget.

3 EXPERIMENTS

We adopt DeepMind *Melting Pot* environment [1] for evaluation. We choose the two-agent substrate named "* *in the Matrix*" [19]. We define four environments: *Chicken*, *Pure Coordination*, *Prisoners Dilemma*, and *Stag Hunt* [1] with distinct payoff matrices. For each environment, we modify the map size and resource/object layout to create three scenarios: small, medium, and large (Figure 1).



(a) Small **(b) Medium** **(c) Large**

Figure 1: Three * *in the Matrix* scenarios.

To validate LoI, we examine the effect of co-player diversity on generalization by training policies with different diversities and testing them against a fixed group of checkpoints from various stages of a single SP policy for evaluation, which we refered to as "*Fixed-Bobs*". We also train additional SP policies with different seeds and PP policies in groups of 3 and 5 populations, then test these against the Fixed-Bobs. Matches are played 10 times to compute average rewards, followed by calculating the average improvement for each training method across environments and scenarios. The results are reported in Table 1: The advantage of PP over SP (*i.e.*, average improvement) varies across different scenarios, and the correlations between scenario and reward increment vary across different environments.

Table 1: Average improvement on individual reward between SP, PP3, and PP5 under each scenario and environment.

	Chicken	*Pure Coordination*	*Prisoners Dilemma*	*Stag Hunt*
Small	1.4130	**1.7986**	7.0535	5.1652
Medium	3.8312	1.0248	**9.4688**	**8.0993**
Large	**4.9293**	0.9117	3.7931	5.2341

Table 2: LoI (and standard deviations, reported in parentheses) across three scenarios under four environments.

	Chicken	*Pure Coordination*	*Prisoners Dilemma*	*Stag Hunt*
Small	1.291 (0.14)	**1.117** (0.12)	1.377 (0.11)	1.397 (0.14)
Medium	1.364 (0.09)	1.071 (0.15)	**1.385** (0.11)	**1.431** (0.13)
Large	**1.438** (0.09)	0.976 (0.09)	1.180 (0.09)	1.424 (0.07)

Then we calculate the LoI introduced in Section 2 for each scenario and environment. Due to the unavailability of full policy spaces Φ and Θ, directly computing the true LoI is impractical. Instead, we approximate LoI by training 6 SP policies, randomly assigning 1 to represent Alice and 5 to Bob. From these, we select 4 late-stage checkpoints per Alice policy to capture variations in skill, and 9 checkpoints from all stages per Bob policy to sample his policy diversity. We model Alice and Bob's policy distributions uniformly across their checkpoints (with probabilities $P_\varphi = 1/4$ and $P_\vartheta = 1/9$), conducting six games per pair. The results are presented in Table 2: LoI exhibits varying trends across three specified scenarios in different environments.

Finally, we analyze the correlation between LoI and average improvement across three scenarios in each environment using the Pearson correlation coefficient r. The average coefficient across all four environments is $\bar{r} = 0.85$. This indicates a strong correlation between LoI and the benefits of a diverse co-player distribution for the ego agent's generalization within given scenarios.

4 CONCLUSION

In our study, we introduce the *Level of Influence* (LoI) metric, a measure that quantifies the interaction intensity between agents across varied scenarios in multi-agent reinforcement learning. Our findings demonstrate that policies trained with larger population sizes exhibit improved performance when paired with unseen co-players in highly interactive scenarios. Our proposed metric can effectively predict the potential generalization improvement by having a more diverse set of co-player distribution during training.

ACKNOWLEDGMENTS

We thank Jiaxun Cui, Arrasy Rahman, and Micah Carroll for their helpful discussions and support on this work.

REFERENCES

[1] John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Duéñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, et al. 2022. Melting Pot 2.0. *arXiv preprint arXiv:2211.13746* (2022).

[2] Nolan Bard, Jakob N Foerster, Sarath Chandar, Neil Burch, Marc Lanctot, H Francis Song, Emilio Parisotto, Vincent Dumoulin, Subhodeep Moitra, Edward Hughes, et al. 2020. The hanabi challenge: A new frontier for ai research. *Artificial Intelligence* 280 (2020), 103216.

[3] Kalesha Bullard, Franziska Meier, Douwe Kiela, Joelle Pineau, and Jakob Foerster. 2020. Exploring zero-shot emergent communication in embodied multi-agent populations. *arXiv preprint arXiv:2010.15896* (2020).

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

[5] Allan Dafoe, Edward Hughes, Yoram Bachrach, Tantum Collins, Kevin R McKee, Joel Z Leibo, Kate Larson, and Thore Graepel. 2020. Open problems in cooperative AI. *arXiv preprint arXiv:2012.08630* (2020).

[6] Florian Fuchs, Yunlong Song, Elia Kaufmann, Davide Scaramuzza, and Peter Dürr. 2021. Super-human performance in gran turismo sport using deep reinforcement learning. *IEEE Robotics and Automation Letters* 6, 3 (2021), 4257–4264.

[7] Johannes Heinrich, Marc Lanctot, and David Silver. 2015. Fictitious self-play in extensive-form games. In *International conference on machine learning*. PMLR, 805–813.

[8] Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. "other-play" for zero-shot coordination. In *International Conference on Machine Learning*. PMLR, 4399–4410.

[9] Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. 2019. Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science* 364, 6443 (2019), 859–865.

[10] Max Jaderberg, Valentin Dalibard, Simon Osindero, Wojciech M Czarnecki, Jeff Donahue, Ali Razavi, Oriol Vinyals, Tim Green, Iain Dunning, Karen Simonyan, et al. 2017. Population based training of neural networks. *arXiv preprint arXiv:1711.09846* (2017).

[11] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*. PMLR, 3040–3049.

[12] Ryan Lowe, Abhinav Gupta, Jakob Foerster, Douwe Kiela, and Joelle Pineau. 2020. On the interaction between supervision and self-play in emergent communication. *arXiv preprint arXiv:2002.01093* (2020).

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

[14] Kevin R McKee, Ian Gemp, Brian McWilliams, Edgar A Duéñez-Guzmán, Edward Hughes, and Joel Z Leibo. 2020. Social diversity and social preferences in mixed-motive reinforcement learning. *arXiv preprint arXiv:2002.02325* (2020).

[15] Kevin R McKee, Joel Z Leibo, Charlie Beattie, and Richard Everett. 2022. Quantifying the effects of environment and population diversity in multi-agent reinforcement learning. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 21.

[16] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2018. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* 362, 6419 (2018), 1140–1144.

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

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

[19] Alexander Sasha Vezhnevets, Yuhuai Wu, Remi Leblond, and Joel Z Leibo. 2019. Options as responses: Grounding behavioural hierarchies in multi-agent RL. *arXiv preprint arXiv:1906.01470* (2019).

[20] 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.

[21] Tom Zahavy, Vivek Veeriah, Shaobo Hou, Kevin Waugh, Matthew Lai, Edouard Leurent, Nenad Tomasev, Lisa Schut, Demis Hassabis, and Satinder Singh. 2023. Diversifying AI: Towards Creative Chess with AlphaZero. *arXiv preprint arXiv:2308.09175* (2023).

[22] Daochen Zha, Jingru Xie, Wenye Ma, Sheng Zhang, Xiangru Lian, Xia Hu, and Ji Liu. 2021. DouZero: Mastering DouDizhu with Self-Play Deep Reinforcement Learning. arXiv:2106.06135 [cs.AI]