Reducing Variance Caused by Communication in Decentralized Multi-agent Deep Reinforcement Learning

Extended Abstract

Changxi Zhu
Utrecht University
Utrecht, Netherlands
c.zhu@uu.nl

Mehdi Dastani
Utrecht University
Utrecht, Netherlands
m.m.dastani@uu.nl

Shihan Wang
Utrecht University
Utrecht, Netherlands
s.wang2@uu.nl

ABSTRACT

In decentralized multi-agent deep reinforcement learning (MADRL), communication can help agents to gain a better understanding of the environment to better coordinate their behaviors. Nevertheless, communication may involve uncertainty, which potentially introduces variance to the learning of decentralized agents. In this extended abstract, we report on our research that focuses on a specific decentralized MADRL setting with communication and a theoretical analysis to study the variance caused by communication in policy gradients. We argue for modular techniques to reduce the variance in policy gradients during training. We show a pseudo algorithm to illustrate the integration of the modular techniques into existing decentralized MADRL with communication methods.

KEYWORDS

Multi-agent Deep Reinforcement Learning, Communication, Variance Reduction

ACM Reference Format:
Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2025. Reducing Variance Caused by Communication in Decentralized Multi-agent Deep Reinforcement Learning: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Multi-agent Deep Reinforcement Learning (MADRL) has been widely used to develop cooperative behaviors of agents in partially observable environments [3, 10, 14]. MADRL agents can communicate various types of information, including observations, intentions, and experiences, to mitigate the limitations in agent observability and enhance the coordination of their behaviors [3, 15, 16]. In recent years, there has been growing research interest in MADRL that focuses on communication via a vector or range of values as encoded messages, rather than directly sharing agents' private and massive local information, known as MADRL with learning communication (Comm-MADRL) [16].

Practical considerations such as security and privacy require that agents act independently and keep control of their individual information during execution [11]. Among various MADRL settings, **D**ecentralized **C**ommunicating **C**ritics and **D**ecentralized **A**ctors (DCCDA) setting [4, 7], which is based on actor-critic methods, enables communication among critics during training while disabling communication among actors (policies) during both training and execution. In DCCDA, agents can enjoy the benefit of communication for better coordination during training, while they can operate fully decentralized (without communication) during the execution. Despite the promising applications of DCCDA, communicated messages are often initiated in a stochastic manner [1, 5], adding uncertainty from the receiver's perspective. This can result in high variance in the policy gradient estimator of receiver agents, leading to low sample efficiency and performance degradation.

In this work, we argue for the need of a theoretical analysis regarding the variance in policy gradients within Comm-MADRL under the DCCDA setting. Variance analysis is a vigorous method that allows us to investigate the variability and dispersion of policy learning. Previous research has focused on variance analysis in policy gradients without communication [8, 9], and thus not measuring variance caused by communication. Specifically, Lyu et al. [8] claim that the variance in policy gradients using a centralized critic (without communication) can be equal to or higher than the variance in policy gradients using decentralized critics (without communication). However, it remains unclear how communication (in the DCCDA setting) affects the variance in policy gradients. This extended abstract reports on our variance analysis [17] which has been used to prove that under both idealistic communication settings (where agents communicate sound & complete information) and non-idealistic communication settings (where sound & complete information is corrupted with noise), policy gradients under DCCDA have equal or higher variance than under the setting using a centralized critic.

Our variance analysis has motivated us to propose a novel message-dependent *baseline* technique to reduce the variance caused by communication, which is inherently different from other baseline techniques considering states and actions [2, 6, 12, 13]. To improve the learning of critics, we also propose a regularization technique to align non-communicating actors and communicating critics. The proposed baseline and regularization techniques can be applied to any Comm-MADRL method under DCCDA. We then propose a pseudo algorithm to show how to extend existing MADRL methods under the DCCDA setting with our proposed technique.

2 COMPARISON IN MADRL SETTINGS

To position our focused DCCDA setting within MADRL, we illustrate various settings, including Centralized Training and Decentralized Execution (CTDE) and Decentralized Training and Decentralized Execution (DTDE), with and without communication, across

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

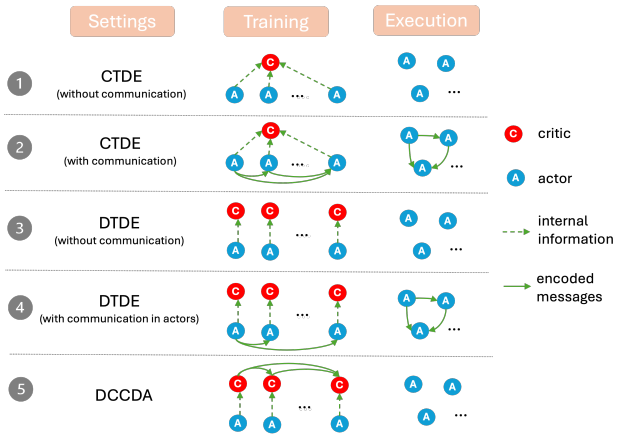

Figure 1: The training and execution phases for CTDE (without communication), CTDE (with communication), DTDE (without communication), and DCCDA using actor-critic methods.

training and execution phases in Figure 1. Note that we specifically focus on actor-critic methods, which align with the DCCDA setting used in our work. As motivated in the introduction, Settings 2&4 allows agents to communicate during policy execution, which may not satisfy practical requirements of security and privacy. Also, Setting 3 is fundamentally different from other settings as communication is not utilized during the training phase. It should be noted that Setting 1 (CTDE without communication) utilizes global information in a centralized critic during training, which can be comparable with a situation where all information is communicated in the training phase. This extended abstract reports on our research that compares our DCCDA setting with Setting 1 in both theoretical analysis and experiments.

3 APPROACH

In our theoretical analysis [17], we consider an idealistic communication setting, where communication induces sound & complete information from all agents. We also consider a non-idealistic communication setting, where sound & complete information is corrupted with noise, e.g., due to the imperfection of message decoders. In both settings, we theoretically prove that the variance in DCCDA policy gradients (where communication is utilized in decentralized critics) is equal to or higher than CTDE policy gradients (where a centralized critic is utilized).

Our theoretical variance analysis motivates us to propose a message-dependent baseline technique $b_i(h_i, m_{-i})$ to consider individual agent' history h_i and the other agents' message m_{-i}. To achieve minimal variance in the presence of communication, we set the derivatives of the variance w.r.t. the baseline as 0. By integrating the optimal baseline to DCCDA policy gradients, we achieve that the variance with the baseline is reduced compared to the setting where the baseline is not used. We also notice that communication is utilized by critics while not by actors, which could cause misalignment between critics and actors. Specifically, the actors generate actions that do not consider communication, while the corresponding Q-values do consider communication. To improve the consistency between actors and critics and thereby improve the learning of critics, we propose a KL objective $\mathcal{L}_{KL}$ to push the policy distribution of actors close to the policy distribution suggested by the critics. The optimal message-dependent baseline and the KL objective jointly constitute our proposed techniques regarding the variance reduction in policy gradients and the learning of critics.

Algorithm 1 DCCDA methods using message-dependent baseline

1: Initialize the parameters for each agent's actor and critic. Initialize the communication model for each agent.
2: **for** each training iteration **do**
3: Initialize data buffer D_i for each agent i
4: Get initial observations $\boldsymbol{o}_0$ and set initial history $\boldsymbol{h}_0$
5: **for** $t = 0$ to *max_steps_per_episode* **do**
6: **for** each agent i **do**
7: Decide action a_t^i and generate messages m_t^i
8: Send messages to other agents
9: **end for**
10: Get new observations $\boldsymbol{o}_{t+1}$ and rewards r_t
11: Insert experience into D_i and update h_t^i for each agent
12: **end for**
13: **for** each agent i **do**
14: Sample training batches from buffer D_i
15: Compute KL objective $\mathcal{L}_{KL}$ using sampled experience
16: Compute baseline $b_i^*(h_i, m_{-i})$ using sampled experience
17: Update policy and communication model
18: **end for**
19: Update critic for each agent i
20: **end for**

Algorithm 1 illustrates how communication is integrated into the MADRL learning process and how our proposed message-dependent baseline and KL divergence term are used during the MADRL learning process. Importantly, the exact procedures of generating messages (line 7), communicating messages (line 8), and updating the communication model (line 17) are determined by a DCCDA method. This reflects the adaptability and flexibility of our technique, making our technique model-agnostic to existing Comm-MADRL methods under DCCDA.

4 CONCLUSIONS

In this paper, we report on our research investigating the variance of policy gradients caused by communication in decentralized MADRL. Specifically, we focus on the Decentralized Communicating Critics and Decentralized Actors (DCCDA) setting, where communication is allowed only among critics, while actors do not communicate during training and execution. Our variance analysis suggests that DCCDA policy gradients have a higher or equal variance than the policy gradients under CTDE. We further propose a message-dependent baseline technique for variance reduction in policy gradients and a KL objective to improve the learning of critics. In the future, we would like to investigate variance reduction techniques under various communication settings.

REFERENCES

[1] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems 29 (NIPS)*. 2137–2145.

[2] Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. 2004. Variance Reduction Techniques for Gradient Estimates in Reinforcement Learning. *J. Mach. Learn. Res.* 5 (2004), 1471–1530.

[3] Sven Gronauer and Klaus Diepold. 2022. Multi-agent deep reinforcement learning: a survey. *Artif. Intell. Rev.* 55, 2 (2022), 895–943. https://doi.org/10.1007/S10462-021-09996-W

[4] Shariq Iqbal and Fei Sha. 2019. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 2961–2970.

[5] Jiechuan Jiang and Zongqing Lu. 2018. Learning Attentional Communication for Multi-Agent Cooperation. In *Advances in Neural Information Processing Systems 31 (NIPS)*. 7265–7275.

[6] Jakub Grudzien Kuba, Muning Wen, Linghui Meng, Shangding Gu, Haifeng Zhang, David Mguni, Jun Wang, and Yaodong Yang. 2021. Settling the Variance of Multi-Agent Policy Gradients. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (Eds.). 13458–13470.

[7] Yong Liu, Weixun Wang, Yujing Hu, Jianye Hao, Xingguo Chen, and Yang Gao. 2020. Multi-Agent Game Abstraction via Graph Attention Neural Network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*. 7211–7218.

[8] Xueguang Lyu, Andrea Baisero, Yuchen Xiao, Brett Daley, and Christopher Amato. 2023. On Centralized Critics in Multi-Agent Reinforcement Learning. *J. Artif. Intell. Res.* 77 (2023), 295–354. https://doi.org/10.1613/JAIR.1.14386

[9] Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. 2021. Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning. In *AAMAS '21: 20th International Conference on Autonomous Agents and Multiagent Systems, Virtual Event, United Kingdom, May 3-7, 2021*, Frank Dignum, Alessio Lomuscio, Ulle Endriss, and Ann Nowé (Eds.). 844–852.

[10] Afshin Oroojlooy and Davood Hajinezhad. 2023. A review of cooperative multi-agent deep reinforcement learning. *Appl. Intell.* 53, 11 (2023), 13677–13722. https://doi.org/10.1007/S10489-022-04105-Y

[11] Omar Sami Oubbati, Mohammed Atiquzzaman, Hyotaek Lim, Abderrezak Rachedi, and Abderrahmane Lakas. 2022. Synchronizing UAV Teams for Timely Data Collection and Energy Transfer by Deep Reinforcement Learning. *IEEE Trans. Veh. Technol.* 71, 6 (2022), 6682–6697. https://doi.org/10.1109/TVT.2022.3165227

[12] Lex Weaver and Nigel Tao. 2001. The Optimal Reward Baseline for Gradient-Based Reinforcement Learning. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, Jack S. Breese and Daphne Koller (Eds.). Morgan Kaufmann, 538–545.

[13] Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M. Bayen, Sham M. Kakade, Igor Mordatch, and Pieter Abbeel. 2018. Variance Reduction for Policy Gradient with Action-Dependent Factorized Baselines. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

[14] Yaodong Yang and Jun Wang. 2020. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective. *CoRR* abs/2011.00583 (2020). arXiv:2011.00583

[15] Mohamed Salah Zaïem and Etienne Bennequin. 2019. Learning to Communicate in Multi-Agent Reinforcement Learning : A Review. *CoRR* abs/1911.05438 (2019). arXiv:1911.05438

[16] Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2024. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems* 38, 4 (2024). https://doi.org/10.1007/s10458-023-09633-6

[17] Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2025. Reducing Variance Caused by Communication in Decentralized Multi-agent Deep Reinforcement Learning. arXiv:2502.06261 https://arxiv.org/abs/2502.06261