Attention-Based Recurrency for Multi-Agent Reinforcement Learning under State Uncertainty

Extended Abstract

Thomy Phan
LMU Munich
thomy.phan@ifi.lmu.de

Fabian Ritz
LMU Munich

Jonas Nüßlein
LMU Munich

Michael Kölle
LMU Munich

Thomas Gabor
LMU Munich

Claudia Linnhoff-Popien
LMU Munich

ABSTRACT

State uncertainty poses a major challenge for decentralized coordination. However, state uncertainty is largely neglected in multi-agent reinforcement learning research due to a strong focus on state-based *centralized training for decentralized execution (CTDE)* and benchmarks that lack sufficient stochasticity like *StarCraft Multi-Agent Challenge (SMAC)*. In this work, we propose *Attention-based Embeddings of Recurrence In multi-Agent Learning (AERIAL)* to approximate value functions under agent-wise state uncertainty. AERIAL uses a learned representation of multi-agent recurrence, considering more accurate information about decentralized agent decisions than state-based CTDE. We then introduce *MessySMAC*, a modified version of SMAC with stochastic observations and higher variance in initial states, to provide a more general and configurable benchmark. We evaluate AERIAL in a variety of MessySMAC maps, and compare the results with state-based CTDE.

KEYWORDS

Dec-POMDP; State Uncertainty; Multi-Agent Learning; Recurrence

ACM Reference Format:
Thomy Phan, Fabian Ritz, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. 2023. Attention-Based Recurrency for Multi-Agent Reinforcement Learning under State Uncertainty: 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

Multi-agent reinforcement learning (MARL) is a popular approach to solving general Dec-POMDPs with remarkable progress in recent years [16, 17]. State-of-the-art MARL is based on *centralized training for decentralized execution (CTDE)*, where training takes place in a laboratory or a simulator with access to global information [2, 4]. For example, *state-based CTDE* exploits true state information to learn a centralized value function in order to derive coordinated policies for decentralized decision making [9, 10, 13, 16, 18]. Due to its effectiveness in the *StarCraft Multi-Agent Challenge (SMAC)* as the current de facto standard for MARL evaluation, state-based CTDE has become very popular and is widely considered an adequate approach to general Dec-POMDPs, leading to many increasingly complex algorithms [5, 6].

However, merely relying on state-based CTDE and SMAC can be a pitfall in practice as state uncertainty is largely neglected. Since the real-world is generally messy and only observable through noisy sensors, state uncertainty is an important aspect of general Dec-POMDPs to be considered though [3, 5, 7]:

From an *algorithm perspective*, purely state-based value functions are insufficient to evaluate and adapt multi-agent behavior, since all agents make decisions on a completely different basis, i.e., individual histories of noisy observations and actions. True Dec-POMDP value functions consider more accurate closed-loop information about decentralized agent decisions though [8]. Furthermore, the optimal state-based value function represents an upper-bound of the true optimal Dec-POMDP value function thus state-based CTDE can result in overly optimistic behavior in general Dec-POMDPs [5].

From a *benchmark perspective*, SMAC has very limited state uncertainty due to deterministic observations and low variance in initial states [1]. Therefore, SMAC scenarios only represent simplified special cases rather than general Dec-POMDP challenges, being insufficient for evaluating generality of MARL [5].

2 METHODS

2.1 Attention-Based Embeddings of Recurrence

We propose *Attention-based Embeddings of Recurrence In multi-Agent Learning (AERIAL)* to approximate true optimal Dec-POMDP value functions according to [8]. Our setup uses a *factorization operator* Ψ like QMIX or QPLEX according to [11, 13, 14, 16]. All agents process their local observation-action histories $\tau_{t,i}$ via RNNs.

To consider more accurate closed-loop information about decentralized agent decisions, we exploit all *individual recurrences* by replacing the true state s_t in CTDE with the *joint memory representation* $\mathbf{h_t} = \langle h_{t,i} \rangle_{i \in \mathcal{D}}$ of all agents' RNNs. Since the individual recurrences encoded by memory representations $h_{t,i} \in \mathbf{h_t}$ are not conditionally independent in general, we additionally process $\mathbf{h_t}$ with a transformer to automatically consider the latent dependencies of all memory representations $h_{t,i} \in \mathbf{h_t}$ through self-attention [15]. The resulting approach, called `AERIAL`, is depicted in Fig. 1.

2.2 SMAC with State Uncertainty

MessySMAC is a modified version of SMAC with *observation stochasticity*, where the observation values are negated with a probability of $\phi \in [0, 1)$, and *initialization stochasticity*, where K random

This extended abstract is a short version of [12].

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.

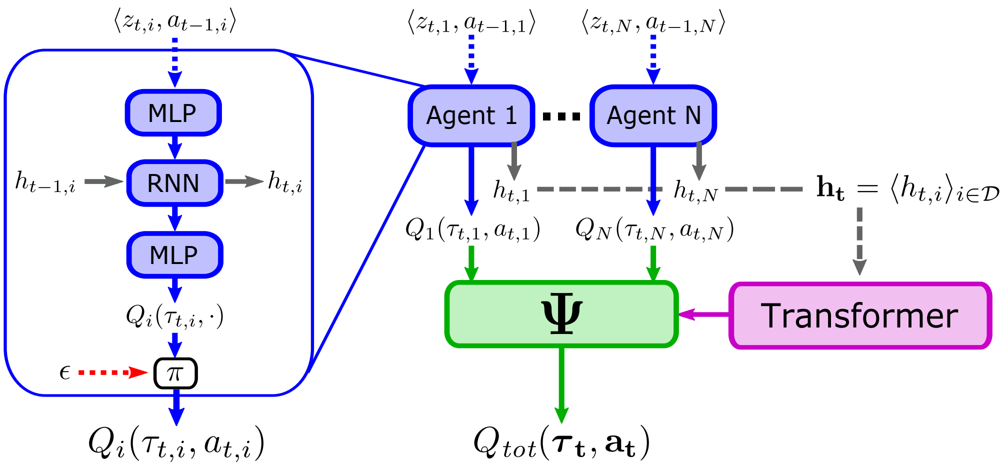


Figure 1: Illustration of the AERIAL setup. ***Left:*** **Recurrent agent network structure with memory representations $h_{t-1,i}$ and $h_{t,i}$.** ***Right:*** **Value function factorization via factorization operator Ψ using the joint memory representation $\mathrm{h_t} = \langle h_{t,i} \rangle_{i \in \mathcal{D}}$ of all agents' RNNs instead of true states s_t. All memory representations $h_{t,i}$ are detached from the computation graph (indicated by the dashed gray arrows) and passed through a simplified transformer before being used by Ψ for value function factorization.**

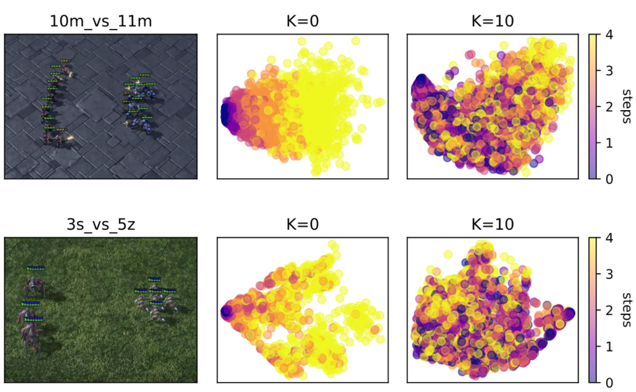


Figure 2: ***Left:*** **Screenshot of two SMAC maps.** ***Middle:*** **PCA visualization of the joint observations in original SMAC within the first 5 steps of 1,000 episodes using a random policy (with $K = 0$ initial random steps).** ***Right:*** **Analogous visualization for MessySMAC (with $K = 10$ initial random steps). For visual comparability, the observations are deterministic here.**

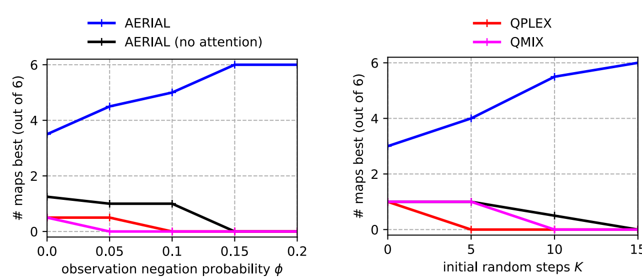
steps are initially performed before officially starting an episode. MessySMAC represents a more general Dec-POMDP challenge which enables systematic evaluation under various state uncertainty configurations according to ϕ and K.

Fig. 2 shows the PCA visualization of joint observations in two maps of SMAC ($K = 0$) and MessySMAC ($K = 10$) within the first 5 steps of 1,000 episodes using a random policy. While the observations of the initial state (dark purple) in original SMAC are very similar and can be easily distinguished from subsequent steps, the separability in MessySMAC is much harder due to significantly higher entropy, indicating higher state uncertainty.

3 EXPERIMENTS

To evaluate the robustness of `AERIAL` against various state uncertainty configurations in MessySMAC[1], we manipulate the observation negation probability ϕ and the number of initial random steps K as defined in Section 2.2. We compare the results with `QPLEX` and `QMIX` as the best performing state-of-the-art baselines in MessySMAC according to the findings of [12]. We present summarized plots, reporting the count of maps used in [12], where each approach performs best compared to the others.

The results w.r.t. observation and initialization stochasticity are shown in Fig. 3. `AERIAL` performs best in most maps, especially when $\phi \geq 15\%$ and $K \geq 10$. State-based CTDE approaches like QPLEX and QMIX are notably less effective when observation and initialization stochasticity increase.



(a) observation stochasticity **(b) initialization stochasticity**

Figure 3: The average number of maps best out of 6 for `AERIAL`, `AERIAL (no attention)`, and the best MessySMAC baselines for ϕ and K w.r.t. the maps used in [12] (20 runs per configuration). The legend at the top applies across all plots.

[1]Our code is available at https://github.com/thomyphan/messy_smac.

ACKNOWLEDGMENTS

This work was partially funded by the Bavarian Ministry for Economic Affairs, Regional Development and Energy as part of a project to support the thematic development of the Institute for Cognitive Systems.

REFERENCES

[1] Benjamin Ellis, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob N. Foerster, and Shimon Whiteson. 2022. SMACv2: An Improved Benchmark for Cooperative Multi-Agent Reinforcement Learning. https://arxiv.org/abs/2212.07489

[2] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. *Proceedings of the AAAI Conference on Artificial Intelligence* 32, 1 (Apr. 2018).

[3] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. 1998. Planning and Acting in Partially Observable Stochastic Domains. *Artificial intelligence* 101, 1-2 (1998), 99–134.

[4] 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*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[5] Xueguang Lyu, Andrea Baisero, Yuchen Xiao, and Christopher Amato. 2022. A Deeper Understanding of State-Based Critics in Multi-Agent Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 9 (Jun. 2022), 9396–9404. https://doi.org/10.1609/aaai.v36i9.21171

[6] Xueguang Lyu, Yuchen Xiao, Brett Daley, and Christopher Amato. 2021. Contrasting Centralized and Decentralized Critics in Multi-Agent Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and Multiagent Systems*. 844–852.

[7] Frans A Oliehoek and Christopher Amato. 2016. *A Concise Introduction to Decentralized POMDPs*. Vol. 1. Springer.

[8] Frans A Oliehoek, Matthijs TJ Spaan, and Nikos Vlassis. 2008. Optimal and Approximate Q-Value Functions for Decentralized POMDPs. *Journal of Artificial Intelligence Research* 32 (2008), 289–353.

[9] Thomy Phan, Lenz Belzner, Kyrill Schmid, Thomas Gabor, Fabian Ritz, Sebastian Feld, and Claudia Linnhoff-Popien. 2020. "A Distributed Policy Iteration Scheme for Cooperative Multi-Agent Policy Approximation". In *12th Adaptive and Learning Agents Workshop (ALA '20)*.

[10] Thomy Phan, Thomas Gabor, Andreas Sedlmeier, Fabian Ritz, Bernhard Kempter, Cornel Klein, Horst Sauer, Reiner Schmid, Jan Wieghardt, Marc Zeller, et al. 2020. Learning and Testing Resilience in Cooperative Multi-Agent Systems. In *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, 1055–1063.

[11] Thomy Phan, Fabian Ritz, Lenz Belzner, Philipp Altmann, Thomas Gabor, and Claudia Linnhoff-Popien. 2021. VAST: Value Function Factorization with Variable Agent Sub-Teams. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 24018–24032.

[12] Thomy Phan, Fabian Ritz, Jonas Nüßlein, Michael Kölle, Thomas Gabor, and Claudia Linnhoff-Popien. 2023. Attention-Based Recurrence for Multi-Agent Reinforcement Learning under State Uncertainty. arXiv:2301.01649 [cs.MA] https://arxiv.org/pdf/2301.01649.pdf

[13] 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 *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 4295–4304.

[14] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. 2019. QTRAN: Learning to Factorize with Transformation for Cooperative Multi-Agent Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 5887–5896.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc.

[16] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. In *International Conference on Learning Representations*.

[17] Muning Wen, Jakub Grudzien Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. 2022. Multi-Agent Reinforcement Learning is a Sequence Modeling Problem. *arXiv preprint arXiv:2205.14953* (2022).

[18] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. 2022. The Surprising Effectiveness of PPO in Cooperative Multi-Agent Games. In *36th Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.