Local Advantage Networks for Cooperative Multi-Agent Reinforcement Learning

Extended Abstract

Raphaël Avalos
Vrije Universiteit Brussel
Brussels, Belgium
raphael.avalos@vub.be

Mathieu Reymond
Vrije Universiteit Brussel
Brussels, Belgium
mathieu.reymond@vub.be

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
ann.nowe@vub.be

Diederik M. Roijers
Vrije Universiteit Brussel (BE) &
HU Univ. of Appl. Sci. Utrecht (NL)
diederik.roijers@vub.be

ABSTRACT

Multi-agent reinforcement learning (MARL) enables us to create adaptive agents in challenging environments, even when the agents have limited observation. Modern MARL methods have focused on finding factorized value functions. While successful, the resulting methods have convoluted network structures. We take a radically different approach and build on the structure of independent Q-learners. Our algorithm LAN leverages a dueling architecture to represent decentralized policies as separate individual advantage functions w.r.t. a centralized critic that is cast aside after training. The critic works as a stabilizer that coordinates the learning and to formulate DQN targets. This enables LAN to keep the number of parameters of its centralized network independent in the number of agents, without imposing additional constraints like monotonic value functions. When evaluated on the SMAC, LAN shows SOTA performance overall and scores more than 80% wins in two super-hard maps where even QPLEX does not obtain almost any wins. Moreover, when the number of agents becomes large, LAN uses significantly fewer parameters than QPLEX or even QMIX. We thus show that LAN's structure forms a key improvement that helps MARL methods remain scalable.

KEYWORDS

Reinforcement Learning; Multi-Agent; Cooperation; Coordination

ACM Reference Format:

Raphaël Avalos, Mathieu Reymond, Ann Nowé, and Diederik M. Roijers. 2022. Local Advantage Networks for Cooperative Multi-Agent Reinforcement Learning: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) [1, 5, 12] introduces additional layers of complexity over single-agent RL: (a) the exponential growth of the joint action space in the number of agents; (b) the non-stationarity induced by the presence of multiple learners also known as the moving target problem [16]; (c) and the multi-agent credit assignment problem when the agents need to disambiguate their contribution in a team reward. In this paper, we consider the cooperative setting with partial observability modeled as a DecPOMDPs [8, 9].

Centralized Training with Decentralized Execution (CTDE) [4, 6, 9], has become a popular learning paradigm for MARL. The core idea behind CTDE is that even though decentralized execution is required the learning is allowed to be centralized. Specifically, during training, it is often possible to access the global state of the environment, as well as the observations and actions of all agents. This allows to break the partial observability, to mitigate the moving target problem, or the credit assignment problem depending on the algorithm. Most of the research in off-policy CTDE MARL for DecPOMDPS focuses on factorizing the joint Q-Value into local agent utilities [10, 13, 14, 18]. Our main contribution is a novel algorithm, called LAN, that learns the decentralized Advantages of best response policies through the centralized learning of a Q-value proxy with the size of its centralized network independent in the number of agents. Empirical evaluation on SMAC [11] shows that LAN performs on par with the current SOTA algorithms on the easy and hard-maps while outperforming on two super hard-maps with more than 80% wins. We thus conclude that LAN is highly promising, as it not only performs well but also scales better in the number of agents in terms of the number of parameters required. The full paper is available at https://arxiv.org/abs/2112.12458.

2 METHOD

Local Advantage Networks (LAN) is a novel value-based algorithm for collaborative partially observable MARL. LAN goes in the opposite direction of the current SOTA, which focuses on factorizing $Q^{\boldsymbol{\pi}}$, the Q-value of the joint policy $\boldsymbol{\pi}$. Instead, LAN learns for each agent the advantage of the best response policy to the other agents' policies.

As in [3], we derive for each agent a the single-agent POMDP induced by fixing the other agents policies. LAN's best response policies are expressed as local advantage functions, $A^{\pi_a}(\tau_a, u_a)$, of the induced POMDPs. The local advantages are solely conditioned on the agent's observation-action history. To coordinate their learning, LAN leverages full information about the states and the joint history at training time via a centralized value function $V^{\boldsymbol{\pi}}$. More

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

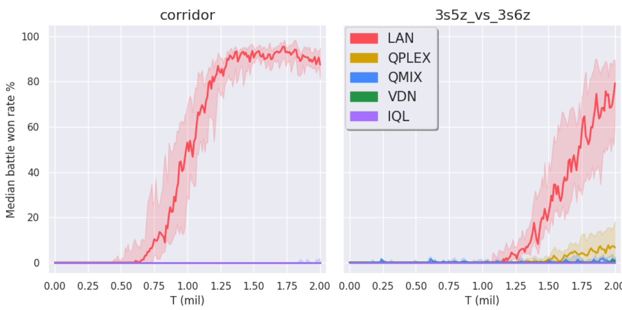

Figure 1: Averaged median test win rate on two SMAC super-hard maps.

specifically, LAN derives $\tilde{Q}_a^{\boldsymbol{\pi}}$ a proxy of the local Q-value Q^{π_a} for each agent a. The proxy is constructed by summing the local advantage A^{π_a} with the centralized value of the joint policy $V^{\boldsymbol{\pi}}$. While $\tilde{Q}_a^{\boldsymbol{\pi}}$ is not a real Q-value and it is conditioned on the full state and the joint history $\boldsymbol{\tau}$ it can still be used to extract decentralized policies as the maximizing actions only depend on the agent's history τ_a.

$$\tilde{Q}_a^{\boldsymbol{\pi}}(s, \boldsymbol{\tau}, u_a) = V^{\boldsymbol{\pi}}(s, \boldsymbol{\tau}) + A^{\pi_a}(\tau_a, u_a) \tag{1}$$

LAN uses Double DQN [7, 17] to learn $\tilde{Q}_a^{\boldsymbol{\pi}}$ for all agents $a \in A$ simultaneously. This allows LAN to learn both the centralized value $V^{\boldsymbol{\pi}}$ and the local advantages A^{π_a} at the same time by optimizing a unique loss, resulting in an efficient learning scheme. The local advantages use a GRU [2] to build a representation of the local history. To ensure scalability, the centralized Value network reuses the GRUs' hidden states of the local advantages to build a representation of the joint history. This results in the number of parameters of the centralized value being independent of the number of agents.

In addition to synchronizing the learning, the centralized nature of $V^{\boldsymbol{\pi}}$ allows reducing the partial observability while mitigating the moving target problem and the multi-agent credit assignment problem. Combined all those elements allow building more accurate DQN targets for the local advantages resulting in better policies.

Compared to LAN, IQL [15] applies DQN on the local Q-values Q^{π_a} simultaneously for all agents ignoring the moving target problem, which leads to sub-optimal policies.

Two key differences with a factorized Q-function are: (1) that LAN does not learn the Q-value of the joint policy but the value (V), and (2) that LAN can represent all decentralized policies, while the type of functions that can be represented by VDN [14] and QMIX [10] is limited.

3 EXPERIMENTS

To benchmark LAN we use SMAC[1] [11], a set of environments that runs in the popular video game StarCraft II. We compare LAN to IQL [15], VDN [14], QMIX [10] and QPLEX [18].

Figure 1 shows the median test win rate, averaged on at least 6 different seeds, on the two super-hard maps corridor and 3s5z_vs_3s6z. In corridor, 6 agents of type 'zealot' fight a team of 24 enemies of type 'zerlings'. The difficulty of this map comes from the fact that if the zerlings surround the zealots they will defeat our agents. In the other map, 3s5z_vs_3s6z, the agents are heterogeneous and the enemy team has an extra unit. While the baselines did not achieve any wins, except QPLEX in 3s5z_vs_3s6z with less than 15%, LAN is able to reach a final performance of at least 80% in both maps.

For those two maps, LAN learns a baiting strategy: one agent lures part of the enemies to a remote location giving a decisive advantage to the rest of the agents. The reason for LAN's performance in the two super-hard maps is the ability for an agent to learn to sacrifice itself for the survival of its team. We believe that this behavior is easier to discover with LAN than with the mixing algorithms for two reasons: (1) the shared Value network allows dead agents to benefit directly from the rewards scored by the other agents after their death; (2) LAN, by focusing on learning best response policies instead of factorizing a joint Q-value, directly learns for each agent the policy that maximizes the team return.

In 10 of the 12 remaining maps, LAN obtains the same performance as the best performing SOTA algorithm. In the last 2 remaining maps, LAN is 20% below QPLEX which equals VDN's performance in one map and is 20% above QMIX and 40% above VDN in the other. Finally, in the complete benchmark, the average performance over the 14 maps, LAN outperforms QPLEX by 10% and ranks first.

4 CONCLUSION

We proposed Local Advantage Networks (LAN); a novel value-based MARL algorithm for Dec-POMDPs. LAN leverages the CTDE approach by building, for each agent, a proxy of the local Q-value composed of the local advantage and the joint value. The centralized learning allows to overcome the partial observability during training, and learning both networks in parallel, by applying DQN to the Q-value proxy, synchronizes all value functions to the ever-changing policies. This results in more accurate DQN targets and mitigates the moving target problem. To ensure scalability, LAN's joint value efficiently summarizes the hidden states produced by the GRUs of the local advantages to represent the joint history, keeping its number of parameters is independent of the number of agents.

We evaluated LAN on the challenging SMAC benchmark where we performed significantly better or on par compared to state-of-the-art methods, while its architecture is significantly more scalable in the number of agents. In the two most complex maps, LAN was able to learn a complex strategy where one agent would sacrifice itself for the survival of the team, therefore proving experimentally LAN's ability to mitigate the multi-agent credit assignment problem. We believe that the lean architecture of LAN for learning decentralized policies in a Dec-POMDP is key to learning efficiently in decentralized partially observable settings.

ACKNOWLEDGMENTS

Raphaël Avalos was supported by the FWO (grant 11F5721N). This research was supported by the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" program.

[1]We use version SC2.4.6.2.69232 and not SC2.4.10 - performances not comparable.

REFERENCES

[1] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. 2008. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews* 38, 2 (3 2008), 156–172. https://doi.org/10.1109/TSMCC.2007.913919

[2] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *EMNLP 2014 - 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference* (6 2014), 1724–1734. https://arxiv.org/abs/1406.1078v3

[3] Jakob Foerster, Nantas Nardell, Gregory Farquhar, Trtantafyllos Afouras, Philip H.S. Torr, Pushmeet Kohli, and Shimon Whiteson. 2017. Stabilising experience replay for deep multi-agent reinforcement learning. In *34th International Conference on Machine Learning, ICML 2017*, Vol. 3. International Machine Learning Society (IMLS), 1879–1888. http://arxiv.org/abs/1702.08887

[4] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*. 2974–2982. http://arxiv.org/abs/1705.08926

[5] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E. Taylor. 2019. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems 2019 33:6* 33, 6 (10 2019), 750–797. https://doi.org/10.1007/S10458-019-09421-1

[6] 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*, Vol. 2017-Decem. 6380–6391. http://arxiv.org/abs/1706.02275

[7] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2 2015), 529–533. https://doi.org/10.1038/nature14236

[8] Frans A Oliehoek and Christopher Amato. 2016. *A Concise Introduction to Decentralized POMDPs*. 1–116 pages. http://www.springer.com/us/book/http://link.springer.com/10.1007/978-3-319-28929-8http://www.springer.com/us/book/%0Ahttp://link.springer.com/10.1007/978-3-319-28929-8

[9] Frans A. Oliehoek, Matthijs T.J. Spaan, and Nikos Vlassis. 2008. Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research* 32 (10 2008), 289–353. https://doi.org/10.1613/jair.2447

[10] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *Proceedings of the35thInternational Conference on MachineLearning, Stockholm, Sweden, PMLR 80, 2018* (3 2018). https://arxiv.org/abs/1803.11485v2

[11] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS* 4 (2 2019), 2186–2188. https://arxiv.org/abs/1902.04043v5

[12] Yoav Shoham, Rob Powers, and Trond Grenager. 2007. If multi-agent learning is the answer, what is the question? *Artificial Intelligence* (2007). https://doi.org/10.1016/j.artint.2006.02.006

[13] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Hostallero, and Yung Yi. 2019. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *36th International Conference on Machine Learning, ICML 2019*, Vol. 2019-June. 10329–10346. http://arxiv.org/abs/1905.05408

[14] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. 2018. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems, AAMAS*, Vol. 3. 2085–2087. http://arxiv.org/abs/1706.05296

[15] Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. 2015. Multiagent Cooperation and Competition with Deep Reinforcement Learning. *PLoS ONE* 12, 4 (11 2015). http://arxiv.org/abs/1511.08779

[16] Karl Tuyls and Gerhard Weiss. 2012. Multiagent learning: Basics, challenges, and prospects. In *AI Magazine*. https://doi.org/10.1609/aimag.v33i3.2426

[17] Hado van Hasselt, Arthur Guez, and David Silver. 2015. Deep Reinforcement Learning with Double Q-learning. *30th AAAI Conference on Artificial Intelligence, AAAI 2016* (9 2015), 2094–2100. https://arxiv.org/abs/1509.06461v3

[18] Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. 2021. QPLEX: Duplex Dueling Multi-Agent Q-Learning. *International Conference on Learning Representations* (8 2021). https://arxiv.org/abs/2008.01062http://arxiv.org/abs/2008.01062https://openreview.net/forum?id=Rcmk0xxIQV