Exploration and Communication for Partially Observable Collaborative Multi-Agent Reinforcement Learning

Doctoral Consortium

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

ABSTRACT

Multi-agent reinforcement learning (MARL) enables us to create adaptive agents in challenging environments, even when the agents have limited observation. The cooperative multi-agent setting introduces numerous challenges compared to the single-agent setting, such as the moving target problem and the curse of dimensionality with respect to the action space. It also aggravates the credit assignment problem, as the credit is not only spread across a sequence of actions, but also multiple agents. This setting also introduces new possibilities, such as task parallelization, specialization, and communication. The Centralized Training with Decentralized Execution paradigm has emerged as a popular strategy to mitigate some of the difficulties in MARL, while still ensuring that the policy of the agents is only conditioned on their local history. However, how to fully leverage this paradigm is still an open question. During the first year of my Ph.D., I developed a novel algorithm, Local Advantage Networks (LAN), that proposes an alternative direction to value factorization, that is more scalable, not limited in its representation and state-of-the-art. The next parts of my research will focus on multi-agent exploration and learning to communicate.

KEYWORDS

Reinforcement Learning; Multi-Agent; Cooperation; Exploration; Communication

ACM Reference Format:
Raphaël Avalos. 2022. Exploration and Communication for Partially Observable Collaborative Multi-Agent Reinforcement Learning: Doctoral Consortium. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 4 pages.

1 INTRODUCTION

Reinforcement learning (RL) [38] is the branch of machine learning dedicated to learning through trial-and-evaluation by interaction between an agent and an environment.

While single-agent RL has been highly successful, many real-word tasks – such as sensor networks [24], wildlife protection [46], and space debris cleaning [21] – require multiple agents to act autonomously. When these agents need to act on local observations, or the problem becomes too large to centralize due to the exponential growth of the joint action space in the number of agents, an explicitly multi-agent perspective is required. As such, *Multi-Agent Reinforcement Learning (MARL)* [8, 13, 34] introduces additional layers of complexity over single-agent RL.

In my research I focus on partially observable cooperative MARL where the agents aim to optimize a team reward. This setting, modeled as a Dec-POMDP, introduces two main challenges that do not exist in single-agent RL. 1) The *moving target problem* [40]: the presence of multiple learners in an environment makes it impossible for an agent to infer the conditional probability of future states. This invalidates most single-agent approaches, as the Markovian property no longer holds. 2) The *multi-agent credit assignment problem*: to learn a policy each agent needs to determine the actions that yield the maximum reward. While in single agent RL this problem is only temporal, as the reward can be sparse and delayed, the shared reward increases the complexity of this problem as the agents also need to determine their individual contribution.

Centralized Training with Decentralized Execution (CTDE) [12, 22, 28], 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, the observations and actions of all agents allowing to break partial observability, to mitigate the moving target problem and the credit assignment problem.

The subsequent parts are about the three main axes of my Ph.D. research. The first part focuses on Local Advantage Network (LAN) [2] a new CTDE algorithm for Dec-POMDPs. LAN offers an alternative to value factorization by learning a local sufficient representation of its incoming influences, and a best response to that. The second part considers exploration strategies for Dec-POMDP, as this area is mainly unexplored. The third part is about learning to communicate to improve the quality of the learned policies but also their robustness against unexpected events.

2 A NOVEL ALGORITHM

Modern MARL methods have hitherto focused on finding factorized value functions into individual utilities conditioned on local observation-action history that can be used for decentralized execution [32, 37, 44]. This approach is appealing as it transforms the multi-agent problem into a single-agent one while still being able to extract decentralized policies. However, to mitigate the curse of dimensionality those algorithms limit the type of function learnable or require convoluted network structures, which might limit their applicability in complex environments, as well as their scalability. Much recent work builds on the structure of QMIX [32] and offers gradual improvement. However, we believe that another approach is needed to widen the possibilities of cooperative deep MARL.

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.

Inspired by influence-based abstraction [29], we developed a novel algorithm called *Local Advantage Networks (LAN)* [2]. Instead of learning a factorization of the centralized Q-value, LAN learns the advantage of the best response policy to the other agents' policies for every agent. These local advantages, which are solely conditioned on the agent observation-action history, are sufficient to build a decentralized policy. In this sense, the architecture of LAN resembles independent Q-learners more than other CTDE approaches such as QMIX or QPLEX. A key element of our solution is to derive a proxy of the local Q-value that leverages CTDE to stabilize the learning of the local advantages. For each agent, the Q-value proxy is composed of the sum of the local advantage with the centralized value of the joint policy. Compared to the local Q-value, LAN's proxy is able to provide better updates by breaking the partial observability while mitigating the moving target problem by integrating the changes of the other agents' policies faster. As LAN learns the local advantage function for each agent it naturally reduces the multi-agent credit assignment problem as well. LAN is also highly scalable as the centralized value network reuses the hidden states of the local advantages to represent the joint observation-action history and the number of parameters of the centralized value does not depend on the number of agents.

In future work, I plan to explore how classic single-agent extensions of DQN [25] can be adapted to LAN such as Prioritized Experience Replay [33] or Hindsight Experience Replay [1], and also how LAN can leverage environment structures through Graph Neural Networks [5, 20].

3 EXPLORING EFFICIENTLY

Exploring efficiently is a key element to learning good policies in complex environments [3, 16]. In MARL the difficulty increases compared to single-agent, as the agents need to balance local and global exploration in a coordinated manner [23, 47]. Indeed, in many real-world environments agents depend on each other to succeed.

One important area of RL exploration consists in measuring the different types of uncertainty in an attempt to reduce the epistemic uncertainty [30, 42] while being aware of the zones with high aleatoric uncertainty [26, 27]. In MARL, the presence of multiple learners increases the difficulty to measure those uncertainties as the different outcomes could be a consequence of the exploration of the other agents or of their policy evolving. I plan to explore how CTDE, by breaking the partial observability, accessing the other agents' policies, and knowing when the other agents are exploring, might allow to mitigate this problem.

The other main area of RL exploration focuses on designing intrinsic rewards when the environment reward is not informative enough or too sparse. In single RL, the intrinsic rewards are usually derived from novelty scores [7], prediction errors [31], or information gain [14]. In a multi-agent setting, those elements are harder to measure, due to the increased dimension of the problem and the moving target problem. They also induce a trade-off between the local and global scope of those measures [6]. However, the presence of multiple learners also presents opportunities such as deriving intrinsic rewards from the influence agents have on each other [17]. I plan to extend this idea to Q-learning algorithms [45] as it is currently limited to policy-based algorithms.

As three key elements for collaborative MARL exploration are adaptivity, commitment, and diversity [10], I plan to build on an algorithm that will leverage successor feature representations [4] and Thompson sampling [39]. LAN will be used as a building block for its adaptivity, while the successor features and Thompson sampling will ensure the commitment and diversity aspect.

4 LEARNING TO COMMUNICATE

Communication is an essential tool to achieve deep coordination as it allows to mitigate partial observability and the stochasticity of the world. Communication can exist in many forms, it can be indirect, such as when an agent performs actions in an environment to send a message, or direct, through a communication channel allowing agents to broadcast or send targeted messages depending on the channel.

While in the tabular setting communication is typically limited to observation sharing or sending discrete messages, deep neural networks allow for continuous communication to be learned in an end-to-end fashion [11].

Many recent papers focus on persistent and perfect communication between all the agents [9, 15, 35, 36]. This setting, while appealing, goes against the principle of autonomous agents, at least in its implementation, as the algorithms can be seen as centralized learning and execution with alternative neural network architectures. Furthermore, persistent and perfect communication is usually not available in the real world and it creates single points of failure if the communication goes down or an agent stops to communicate, as the rest of the agents would not be able to act. While some work has already been done in this setting [18, 19, 43, 48], mainly by introducing a communication budget, it is still unexplored.

We will investigate how to learn to communicate under realistic conditions such as a limited bandwidth channel. We will leverage attention neural networks [41] to handle the variation of incoming messages, while ensuring scalability.

5 CONCLUSION

My research focuses on collaborative Multi-Agent Reinforcement Learning and is built around three main axes. The first one is the development of LAN, a new CTDE algorithm for Dec-POMDP that learns best responses policies instead of value factorization. I aim to explore how LAN can be further enhanced by leveraging different types of replay buffer and network architecture. The second one is designing new exploration strategies. To this end, I will focus on how CTDE can be exploited in uncertainty measurement, but also on how to extend societal influence to Q-learning algorithms, and on combining successor features and Thompson sampling in a novel way. The third one is learning how to communicate to increase the quality and robustness of the policies. To match the autonomy requirement in multi-agent systems and to be widely applicable, I will focus on a limited bandwidth setting.

ACKNOWLEDGMENTS

Raphaël Avalos was supported by the FWO (Research Foundation – Flanders) under the grant 11F5721N.

REFERENCES

[1] Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. 2017. Hindsight experience replay. In *Advances in Neural Information Processing Systems*, Vol. 2017-Decem. 5049–5059. http://arxiv.org/abs/1707.01495

[2] Raphael Avalos, Mathieu Reymond, Ann Nowé, and Diederik M. Roijers. 2022. Local Advantage Networks for Cooperative Multi-Agent Reinforcement Learning. *AAMAS '22: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (2022).

[3] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. 2018. Efficient exploration through Bayesian deep Q-networks. In *2018 Information Theory and Applications Workshop, ITA 2018*. https://doi.org/10.1109/ITA.2018.8503252

[4] Andre Barreto, Will Dabney, Remi Munos, Jonathan J. Hunt, Tom Schaul, Hado P. van Hasselt, and David Silver. 2017. Successor Features for Transfer in Reinforcement Learning. *Advances in Neural Information Processing Systems* 30 (2017).

[5] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. 2018. Relational inductive biases, deep learning, and graph networks. (6 2018). http://arxiv.org/abs/1806.01261

[6] Thomas Bolander and Mikkel Birkegaard Andersen. 2011. Epistemic planning for single-and multi-agent systems. In *Journal of Applied Non-Classical Logics*. https://doi.org/10.3166/JANCL.21.9-34

[7] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by Random Network Distillation. (10 2018). https://arxiv.org/abs/1810.12894

[8] 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

[9] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Michael Rabbat, and Joelle Pineau. 2019. TarMAC: Targeted multi-agent communication. In *36th International Conference on Machine Learning, ICML 2019*, Vol. 2019-June. 2776–2784. http://arxiv.org/abs/1810.11187

[10] Maria Dimakopoulou, Ian Osband, and Benjamin Van Roy. 2018. Scalable coordinated exploration in concurrent reinforcement learning. In *Advances in Neural Information Processing Systems*, Vol. 2018-Decem. 4219–4227. https://arxiv.org/abs/1805.08948

[11] 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*. 2145–2153. http://arxiv.org/abs/1605.06676

[12] 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

[13] 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

[14] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. 2016. VIME: Variational Information Maximizing Exploration. *Advances in Neural Information Processing Systems* 0 (5 2016), 1117–1125. http://arxiv.org/abs/1605.09674

[15] Shariq Iqbal and Fei Sha. 2019. Actor-attention-critic for multi-agent reinforcement learning. In *36th International Conference on Machine Learning, ICML 2019*, Vol. 2019-June. 5261–5270. http://arxiv.org/abs/1810.02912

[16] David Janz, Jiri Hron, Przemyslaw Mazur, Katja Hofmann, José Miguel Hernández-Lobato, and Sebastian Tschiatschek. 2019. Successor uncertainties: Exploration and uncertainty in temporal difference learning. In *Advances in Neural Information Processing Systems*, Vol. 32. http://arxiv.org/abs/1810.06530

[17] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro A. Ortega, D. J. Strouse, Joel Z. Leibo, and Nando de Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *36th International Conference on Machine Learning, ICML 2019*, Vol. 2019-June. 5372–5381. http://arxiv.org/abs/1810.08647

[18] Jiechuan Jiang and Zongqing Lu. 2018. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems*, Vol. 2018-Decem. 7254–7264. http://arxiv.org/abs/1805.07733

[19] Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, and Yung Yi. 2019. Learning to schedule communication in multi-agent reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019*. International Conference on Learning Representations, ICLR. http://arxiv.org/abs/1902.01554

[20] Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*. http://arxiv.org/abs/1609.02907

[21] Richard Klima, Daan Bloembergen, Rahul Savani, Karl Tuyls, Alexander Wittig, Andrei Sapera, and Dario Izzo. 2018. Space debris removal: Learning to cooperate and the price of anarchy. *Frontiers Robotics AI* 5, JUN (2018). https://doi.org/10.3389/FROBT.2018.00054/FULL

[22] 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

[23] Anuj Mahajan, Tabish Rashid, Mikayel Samvelyan, and Shimon Whiteson. 2019. MAVEN: Multi-Agent Variational Exploration. *Advances in Neural Information Processing Systems* 32 (10 2019). https://arxiv.org/abs/1910.07483v2

[24] Mihail Mihaylov, Karl Tuyls, and Ann Nowé. 2010. Decentralized Learning in Wireless Sensor Networks. In *Adaptive and Learning Agents*, Matthew Taylor and Karl Tuyls (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 60–73.

[25] 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

[26] Thomas M. Moerland, Joost Broekens, and Catholijn M. Jonker. 2017. Efficient exploration with Double Uncertain Value Networks. (11 2017). http://arxiv.org/abs/1711.10789

[27] Nikolay Nikolov, Johannes Kirschner, Felix Berkenkamp, and Andreas Krause. 2019. Information-directed exploration for deep reinforcement learning. In *7th International Conference on Learning Representations, ICLR 2019*. http://arxiv.org/abs/1812.07544

[28] 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

[29] Frans A. Oliehoek, Stefan J. Witwicki, and Leslie P. Kaelbling. 2012. Influence-based abstraction for multiagent systems. In *Proceedings of the National Conference on Artificial Intelligence*.

[30] Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. 2016. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*. 4033–4041.

[31] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. 2017. Curiosity-driven exploration by self-supervised prediction. In *34th International Conference on Machine Learning, ICML 2017*, Vol. 6. 4261–4270. http://arxiv.org/abs/1705.05363

[32] 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

[33] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized experience replay. In *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*. http://arxiv.org/abs/1511.05952

[34] 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

[35] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. 2019. Learning when to communicate at scale in multiagent cooperative and competitive tasks. *7th International Conference on Learning Representations, ICLR 2019* (12 2019). http://arxiv.org/abs/1812.09755

[36] Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems*. 2252–2260. https://arxiv.org/abs/1605.07736

[37] 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

[38] R.S. Sutton and A.G. Barto. 1998. Reinforcement Learning: An Introduction. *IEEE Transactions on Neural Networks* (1998). https://doi.org/10.1109/tnn.1998.712192

[39] William R. Thompson. 1933. On the Likelihood that One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika* (1933). https://doi.org/10.2307/2332286

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

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, Vol. 2017-Decem. 5999–6009.

[42] Timothy Verstraeten, Eugenio Bargiacchi, Pieter J.K. Libin, Jan Helsen, Diederik M. Roijers, and Ann Nowé. 2020. Multi-Agent Thompson Sampling for Bandit Applications with Sparse Neighbourhood Structures. *Scientific Reports* 10, 1 (12 2020), 1–13. https://doi.org/10.1038/s41598-020-62939-3

[43] Rose E. Wang, Michael Everett, and Jonathan P. How. 2020. R-MADDPG for Partially Observable Environments and Limited Communication. (2 2020). http://arxiv.org/abs/2002.06684

[44] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado Van Hasselt, Marc Lanctot, and Nando De Frcitas. 2016. Dueling Network Architectures for Deep Reinforcement Learning. In *33rd International Conference on Machine Learning, ICML 2016*, Vol. 4. 2939–2947. https://arxiv.org/abs/1511.06581

[45] Christopher J. C. H. Watkins and Peter Dayan. 1992. Q-learning. *Machine Learning* (1992). https://doi.org/10.1007/bf00992698

[46] Lily Xu, Shahrzad Gholami, Sara Mc Carthy, Bistra Dilkina, Andrew Plumptre, Milind Tambe, Rohit Singh, Mustapha Nsubuga, Joshua Mabonga, Margaret Driciru, Fred Wanyama, Aggrey Rwetsiba, Tom Okello, and Eric Enyel. 2020. Stay ahead of poachers: Illegal wildlife poaching prediction and patrol planning under uncertainty with field test evaluations (Short Version). *Proceedings - International Conference on Data Engineering* 2020-April (4 2020), 1898–1901. https://doi.org/10.1109/ICDE48307.2020.00198

[47] Tianpei Yang, Hongyao Tang, Chenjia Bai, Jinyi Liu, Jianye Hao, Zhaopeng Meng, and Peng Liu. 2021. Exploration in Deep Reinforcement Learning: A Comprehensive Survey. (9 2021). http://arxiv.org/abs/2109.06668

[48] Chongjie Zhang and Victor Lesser. 2013. Coordinating multi-agent reinforcement learning with limited communication. In *12th International Conference on Autonomous Agents and Multiagent Systems 2013, AAMAS 2013*, Vol. 2. 1101–1108. www.ifaamas.org