# Cooperation in Multi-Agent Reinforcement Learning

**Karush Suri, Dian Gadjov, Lacra Pavel**
Department of Electrical & Computer Engineering, University of Toronto, Canada.
karush.suri@mail.utoronto.ca

## Abstract

Advancements in Multi-Agent Reinforcement Learning (MARL) are motivated by cooperation in agents arising from Game Theory (GT). Agents must collaborate in practical scenarios in order to achieve complex objectives and attain strategies which depict optimal behavior. The need for cooperation is further highlighted in the case of partially-observed settings wherein agents have restricted access to environment observations. We revisit cooperation in MARL from the viewpoint of GT and stochastic dynamics of environments. The contributions of our work are twofold. (1) We analyze and demonstrate the effectiveness of cooperative MARL in the case of complex and partially-observed tasks consisting of high-dimensional action spaces and stochastic dynamics. (2) We leverage the empirical demonstrations to construct a novel optimization objective which addresses the detrimental effects of spurious states across agents. Our large-scale experiments carried out on the StarCraft II benchmark depict the effectiveness of cooperative MARL and our novel objective for obtaining optimal strategies under stochastic dynamics.

## 1 Introduction

Reinforcement Learning (RL) has seen tremendous growth in applications such as arcade games [1], board games [2, 3], robot control tasks [4, 5] and lately, real-time games [6]. The rise of RL has led to an increasing interest in the study of multi-agent systems [7, 8], commonly known as Multi-Agent Reinforcement Learning (MARL). MARL provides significant benefits in comparison to contemporary single-agent methods [9]. The Multi-Agent framework allows the modelling of complex real-world systems which consist of dynamic and large-scale interactions between multiple agents [10]. Additionally, MARL enables the learning of diverse strategies which are essential for executing a range of different tasks by the same set of agents.

In the case of partially observable settings, MARL enables the learning of strategies from a GT perspective by utilizing cooperation across agents[11]. Agents collaborate with each other in a given environment to optimize the cumulative payoffs by means of a single utility function. Optimization of the joint utility function leads to optimal behavior [12, 13] in the long-horizon which is characterized by each agent executing its optimal strategy irrespective of other agents. Such a framework of learning strategies with collaborators and executing behaviors independently is often referred to as centralized training with decentralized control [14].

The regime of decentralized control is hindered by intrinsic stochasticity in the environment. Spurious states are a common phenomenon observed in the case of single-agent RL methods. In the case of model-based RL [15], agents build a model of the environment which learns the dynamics of the environment. Such a scheme is used as an effective planning tool in the case of long-horizon tasks [16]. In the case of model-free RL methods, environment stochasticity is addressed by utilizing robust utility functions [17, 18] and effective exploration strategies [19]. On the other hand, MARL does not account for spurious states across agents as a result of which the system remains unaware of drastic changes in the environment [20]. Thus, addressing the learning of stochastic dynamics in the case of multi-agent settings requires attention from a critical standpoint.

We revisit cooperation in MARL from the perspective of GT and stochastic dynamics in the agents' environment. Our work assesses and demonstrates collaborative schemes in MARL under partially-observed settings which pose ill-conditioned objectives for the multi-agent system. More specifically, our twofold contributions are the following-

- We analyze and demonstrate the effectiveness of cooperative MARL in the case of complex and partially-observed tasks consisting of high-dimensional action spaces and spurious dynamics.
- We leverage the empirical demonstrations to construct a novel optimization objective which addresses the detrimental effects of spurious states across agents.

Our large-scale experiments carried out on the StarCraft II benchmark depict the effectiveness of cooperative MARL and our novel objective for obtaining optimal strategies under stochastic dynamics.

## 2 Related Work

Growing advances in GT have given rise to efficient MARL algorithms and implementations in stochastic scenarios. Thsi section highlights some of the main contributions in learning of stochastic games which have paved the way for Multi-Agent learning.

### 2.1 Learning in Games

### 2.2 Multi-Agent Learning

Most multi-agent methods are based on the paradigm of centralized training and decentralized control [14, 21, 22] wherein agents learn to collaborate [23] and optimize their utility function [24]. methods continue to suffer from overestimation bias [25, 26].

Despite the recent success of RL [27, 28] MARL agents suffer from spurious state spaces and encounter sudden changes in trajectories. These anomalous transitions between consecutive states are often termed as surprise [17]. Quantitatively, surprise can be inferred as a measure of deviation [16, 19] among states encountered by the agent during its interaction with the environment. In the case of single-agent methods, surprise results in sample-inefficient learning [17]. This can be tackled by making use of rigorous exploration strategies [29, 30]. However, such solutions do not show evidence for multiple agents consisting of individual partial observations [31].

## 3 Preliminaries

### 3.1 Stochastic Markov Games

### 3.2 Q-Learning

### 3.3 Multi-Agent Learning

## 4 Cooperation in Multi-Agent Learning

### 4.1 The Partial Observability Setting

### 4.2 Learning Model-Free Behaviors

## 5 Tackling Spurious Dynamics

## 6 Experiments

### 6.1 The StarCraft II Benchmark

### 6.2 Performance

### 6.3 Spurious Dynamics

## 7 Conclusion

## References

[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013.

[2] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016.

[3] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering atari, go, chess and shogi by planning with a learned model, 2019.

[4] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Manfred Otto Heess, Tom

Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *CoRR*, abs/1509.02971, 2015.

[5] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

[6] Oriol Vinyals, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani, Heinrich Küttler, John Agapiou, Julian Schrittwieser, John Quan, Stephen Gaffney, Stig Petersen, Karen Simonyan, Tom Schaul, Hado van Hasselt, David Silver, Timothy Lillicrap, Kevin Calderone, Paul Keet, Anthony Brunasso, David Lawrence, Anders Ekermo, Jacob Repp, and Rodney Tsing. Starcraft ii: A new challenge for reinforcement learning, 2017.

[7] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments, 2017.

[8] Oriol Vinyals, Igor Babuschkin, Wojciech Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David Choi, Richard Powell, Timo Ewalds, Petko Georgiev, Junhyuk Oh, Dan Horgan, Manuel Kroiss, Ivo Danihelka, Aja Huang, Laurent Sifre, Trevor Cai, John Agapiou, Max Jaderberg, and David Silver. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575, 11 2019.

[9] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2018.

[10] Gonçalo Neto. From single-agent to multi-agent reinforcement learning: Foundational concepts and methods. *Learning theory course*, 2005.

[11] Liviu Panait and Sean Luke. Cooperative multi-agent learning: The state of the art. *Autonomous agents and multi-agent systems*, 11(3):387–434, 2005.

[12] Ann Nowé, Peter Vrancx, and Yann-Michaël De Hauwere. Game theory and multi-agent reinforcement learning. In *Reinforcement Learning*, pages 441–470. Springer, 2012.

[13] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

[14] Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients, 2017.

[15] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, Afroz Mohiuddin, Ryan Sepassi, George Tucker, and Henryk Michalewski. Model-based reinforcement learning for atari, 2019.

[16] Glen Berseth, Daniel Geng, Coline Devin, Dinesh Jayaraman, Chelsea Finn, and Sergey Levine. Smirl: Surprise minimizing rl in entropic environments. 2019.

[17] Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning, 2017.

[18] Luis Macedo, Rainer Reisezein, and Amilcar Cardoso. Modeling forms of surprise in artificial agents: empirical and theoretical study of surprise functions. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 26, 2004.

[19] Jerry Zikun Chen. Reinforcement learning generalization with surprise minimization, 2020.

[20] Luis Macedo and Amilcar Cardoso. The role of surprise, curiosity and hunger on exploration of unknown environments populated with entities. In *2005 portuguese conference on artificial intelligence*, 2005.

[21] 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. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18, page 2085–2087, 2018.

[22] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML 2018: Proceedings of the Thirty-Fifth International Conference on Machine Learning*, 2018.

[23] Jianye Hao, Dongping Huang, Yi Cai, and Ho-Fung Leung. Reinforcement social learning of coordination in networked cooperative multiagent systems. In *AAAI workshop on multiagent interaction without prior coordination (MIPC 2014)*, 2014.

[24] Carlos Guestrin, Michail Lagoudakis, and Ronald Parr. Coordinated reinforcement learning. In *ICML*, volume 2, pages 227–234, 2002.

[25] Johannes Ackermann, Volker Gabler, Takayuki Osa, and Masashi Sugiyama. Reducing overestimation bias in multi-agent domains using double centralized critics. *arXiv preprint arXiv:1910.01465*, 2019.

[26] Xueguang Lyu and Christopher Amato. Likelihood quantile networks for coordinating multi-agent reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, 2020.

[27] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, 2016.

[28] Matteo Hessel, Joseph Modayil, Hado Van Hasselt, Tom Schaul, Georg Ostrovski, Will Dabney, Dan Horgan, Bilal Piot, Mohammad Azar, and David Silver. Rainbow: Combining improvements in deep reinforcement learning. *arXiv preprint arXiv:1710.02298*, 2017.

[29] Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

[30] Lisa Lee, Benjamin Eysenbach, Emilio Parisotto, Eric Xing, Sergey Levine, and Ruslan Salakhutdinov. Efficient exploration via state marginal matching. *arXiv preprint arXiv:1906.05274*, 2019.

[31] Wei Ren, Randal W Beard, and Ella M Atkins. A survey of consensus problems in multi-agent coordination. In *Proceedings of the 2005, American Control Conference, 2005.*, 2005.