Is an Exponentially Growing Action Space Really That Bad? Validating a Core Assumption for using Multi-Agent RL

Extended Abstract

Ruan de Kock
University of Cape Town
Cape Town, South Africa
dkcrua001@myuct.ac.za

Arnu Pretorius
InstaDeep
Cape Town, South Africa
a.pretorius@instadeep.com

Jonathan Shock
University of Cape Town
Cape Town, South Africa Institut
National de la Recherche Scientifique
Québec City, Canada
jonathan.shock@uct.ac.za

ABSTRACT

One of the core challenges frequently cited in the multi-agent reinforcement learning (MARL) literature motivating the framing of a sequential decision-making problem as a multi-agent problem, instead of a centralised single-agent problem, is the exponential growth in the action space with the number of agents. The assumption that this is always a challenge suggests that this exponentially larger action space poses two specific problems compared with centralised approaches: (1) overwhelming memory requirements and (2) low sample efficiency due to the large optimisation space. Although a core tenet within the MARL community, few works have concretely tested this assumption empirically within a controlled setting to give some indication of its severity in practice.

In this work, we compare fully centralised learning with fully decentralised learning. Using a novel N-agent array game akin to the canonical Climbing matrix game, we re-establish a well-known result; that fully centralised learning is able to find the globally optimal solution while decentralised learning fails. We further demonstrate that these trends hold for more modern MARL benchmarks that run on hardware accelerators and leverage the computational efficiency gains of the JAX framework.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent systems**.

KEYWORDS

Multi-Agent Reinforcement Learning, Multi-Agent Systems

ACM Reference Format:
Ruan de Kock, Arnu Pretorius, and Jonathan Shock. 2025. Is an Exponentially Growing Action Space Really That Bad? Validating a Core Assumption for using Multi-Agent RL: 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

When considering a multi-agent systems task, a common first approach is to model such a problem as a centralised single-agent task. Single-agent RL has been well studied, provides numerous methods for a wide range of settings and, in some cases, has convergence guarantees. The question of how sensible this approach is has existed for over two decades in terms of limitations on learnt behaviours and computational complexity [1, 2], yet in the authors experience, practitioners still consider centralised controllers to be a competitive alternative to truly decentralised MARL methods. We posit that these question may have arisen due to (1) computational advances and (2) a lack of a clear empirical evidence in the literature comparing these two approaches.

To this end, we perform an empirical investigation to test whether the centralised approach is indeed sensible. We begin by showing that in canonical settings, that were standard in early MARL research, fully centralised learning can achieve superior performance using modern hardware acceleration, even as the action space and the number of agents are scaled. Therefore, in the context of simpler problems and modern hardware, fully centralised learning can be considered a strong approach. However, when transitioning to more recent higher-dimensional research environments, we show that computational memory and hardware simply can not keep up when trying a fully centralised approach. In the context of multi-agent systems, we hope that these findings provide useful ways to think about which approach is likely sensible for a given scenario.

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).

2 EXPERIMENTS

2.1 Benchmarking environments

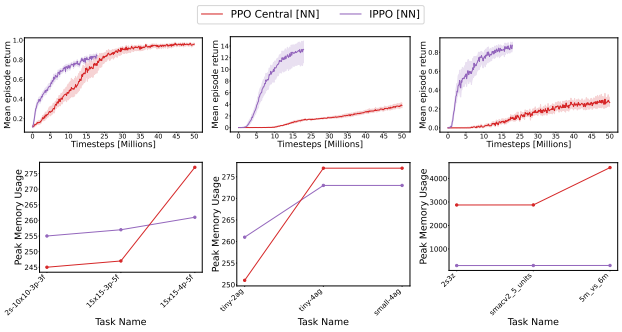
Climbing game. The Climbing game [4] is a matrix game with a pay-off matrix

$$\mathbf{r}_c = \begin{bmatrix} 11 & -30 & 0 \\ -30 & 7 & 6 \\ 0 & 0 & 5 \end{bmatrix} \tag{1}$$

and has a shadowed equilibrium [7] leading independent agents to favour sub-optimal local optima.

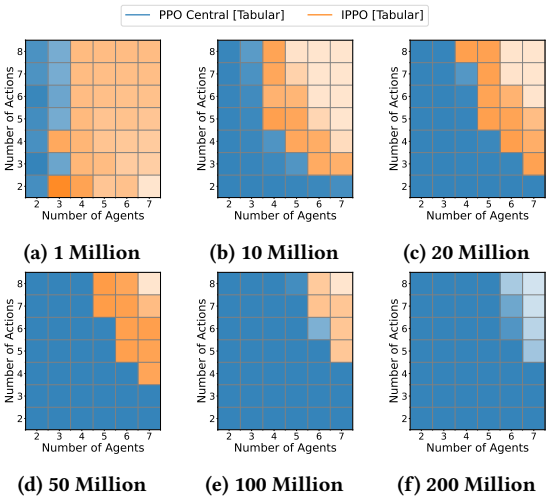
N-player array games. We construct two *N-player* array games which can be adapted to large numbers of agents and actions: (1) the *shadowed equilibrium* game which maintains the shadowed equilibrium properties of the Climbing game and (2) the *needle-in-a-haystack* game which tests for effective exploration. Both allow for arbitrary numbers of agents N and agent actions u.

Modern MARL benchmark environments. We consider the sparse reward Robotic Warehouse (RWARE), [9], the Level-based



(a) LBF. (b) RWARE. (c) SMAX.

Figure 3: ***Central controllers offer competitive performance on the small-scale LBF environment suite, but suffer from low sample efficiency and high memory requirements on larger scale suites like RWARE and SMAX.***



(a) 1 Million (b) 10 Million (c) 20 Million

(d) 50 Million (e) 100 Million (f) 200 Million

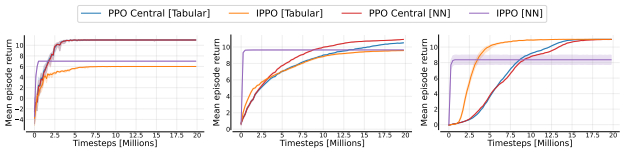
Figure 1: As the training time budget increases, central controllers outperform independent learners and converge to the true optimal solution.

Foraging (LBF), [3] and The StarCraft Multi-agent Challenge in JAX (SMAX) environments1 [10].

2.2 Experiment outline

We first test policies on the Climbing matrix game and both N-player array games. Finally we evaluate policies on JAX-based modern MARL benchmarks and report the per-task peak memory usage and episode returns aggregated over tasks for each environment suite.

In all cases we consider proximal policy optimisation (PPO) [11] as the centralised algorithm and its independent learners (IL) extension, IPPO [6]. All algorithm baselines were implemented using the Mava JAX-based research library [5] and follow the evaluation protocol outlined in [8].



(a) Climbing game. (b) Shadowed equilibrium game. (c) Needle-in-a-haystack game.

Figure 2: ***Central controllers are able to overcome sub-optimal solutions in both the climbing game and shadowed equilibrium game while independent learners offer comparable exploration of the joint action space in the needle-in-a-haystack game.***

2.3 Results

We note from Figure 1 that, given ample environment transitions, centralised controllers indeed converge to better returns than their IL counterparts. Similarly in Figure 2 we see, as expected, that for the Climbing game, the centralised controllers escape the non-global optimum. This extends to the N-agent case as illustrated in Figure 2. Interestingly, tabular IL agents do not overfit as easily as their neural network counterparts and can explore effectively in higher dimensions. This could be due to their comparably lower parameter counts. From Figure 3, we note that for simple tasks with few agents and agent actions, like those in the LBF environment suite, centralised controllers offer competitive performance. However, for more complex tasks with larger agent action spaces like in the RWARE and SMAX suites, this quickly breaks down. This is particularly illustrated in the large memory requirements for centralised learning agents in SMAX due to the much larger action spaces per agent. Overall, we find that centralised controllers are indeed only effective in small-scale settings.

3 CONCLUSION

We empirically assessed the trade-offs of fully centralised versus fully decentralised learning, specifically testing the assumption that an exponentially growing action space indeed poses a significant challenge, even on modern hardware. We first considered simpler array-game settings, showing that as the action space grows, fully centralised learning is still able to efficiently find the optimal solution, while decentralised learning fails. However, as we consider more complex, and higher-dimensional settings, this trend no longer continues. Specifically, to use more compute and train for longer becomes infeasible to achieve similar successes as in simpler settings.

Although this finding, that in complex environments, the exponentially growing action space does indeed pose a problem, may be obvious, especially in the theoretical limit, we hope that having concrete empirical evidence supporting its theoretical conclusions even when using modern hardware and software gives a simple framework for deciding on whether a given multi-agent system problem should be modelled using centralised or decentralised frameworks.

ACKNOWLEDGMENTS

We thank the Google TPU Research Cloud (TRC) for generously providing the cloud TPUs required for conducting our experiments.

REFERENCES

[1] Craig Boutilier. 1996. Planning, learning and coordination in multiagent decision processes. In *TARK*, Vol. 96. Citeseer, 195–210.

[2] Lucian Busoniu, Robert Babuska, 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 (2008), 156–172.

[3] Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. 2020. Shared Experience Actor-Critic for Multi-Agent Reinforcement Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[4] Caroline Claus and Craig Boutilier. 1998. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI* 1998, 746-752 (1998), 2.

[5] Ruan de Kock, Omayma Mahjoub, Sasha Abramowitz, Wiem Khlifi, Callum Rhys Tilbury, Claude Formanek, Andries Smit, and Arnu Pretorius. 2021. Mava: a research library for distributed multi-agent reinforcement learning in JAX. *arXiv preprint arXiv:2107.01460* (2021).

[6] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. 2020. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533* (2020).

[7] Nancy Fulda and Dan Ventura. 2007. Predicting and Preventing Coordination Problems in Cooperative Q-learning Systems.. In *IJCAI*, Vol. 2007. Citeseer, 780–785.

[8] Rihab Gorsane, Omayma Mahjoub, Ruan John de Kock, Roland Dubb, Siddarth Singh, and Arnu Pretorius. 2022. Towards a standardised performance evaluation protocol for cooperative marl. *Advances in Neural Information Processing Systems* 35 (2022), 5510–5521.

[9] Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. 2021. Benchmarking Multi-Agent Deep Reinforcement Learning Algorithms in Cooperative Tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS)*. http://arxiv.org/abs/2006.07869

[10] Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, Saptarashmi Bandyopadhyay, Mikayel Samvelyan, Minqi Jiang, Robert Tjarko Lange, Shimon Whiteson, Bruno Lacerda, Nick Hawes, Tim Rocktaschel, Chris Lu, and Jakob Nicolaus Foerster. 2023. JaxMARL: Multi-Agent RL Environments in JAX. *arXiv preprint arXiv:2311.10090* (2023).

[11] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).