ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Policy-Based Reinforcement Learning in the Generalized Rock-Paper-Scissors Game

Mali Imre Gergely and Gabriela Czibula *

Babeş-Bolyai University, Department of Computer Science, Romania

Abstract. The *Rock-Paper-Scissors* game is a popular zero-sum game of cyclic nature, with a mixed-strategy Nash-equilibrium that has been the subject of a large number of studies and is of particular interest for economy, sociology and artificial intelligence. While there are numerous studies exploring evolutionary dynamics and learning, the overwhelming majority of these consider the game in its classical form, and two important axes with potential relevance remain unexplored. First, studies with policy-based reinforcement algorithms are lacking, and second, few existing investigations attempted to study such cyclic games with more than two players. The present work aims to address both of these matters.

1 Introduction

Multi-Agent Reinforcement-Learning (MARL) has shown great potential in many real-world applications in various domains such as robotics, medicine, agriculture, economy, etc. [1, 2]. Due to the fact that these systems consists of autonomous agents with learned policies, it is of crucial importance to understand and explain the incentives and the nature of interactions between agents. To this end, game-theory is highly relevant to MARL as it provides robust mathematical tools to analyse, interpret and predict agent behavior such as cooperation, conflict, and coordination [3]. Many existing studies consider social dilemmas such as the Prisoner's Dilemma, the Snowdrift game, etc. [4, 5]. These models however mostly experiment with games where there exists some pure-strategy Nash-equilibra. The *Rock-Paper-Scissors* (RPS) game is a classical game-theoretic model with a mixed-strategy Nash-equilibrium, which has been subject to vast academic scrutiny, since it proves to be an intuitive model of species competition in ecological systems and price cycling in economy [6, 7]. However, multi-agent systems modeling such situations generally need to account for interactions among potentially more than two agents, possibly involving multiple actions. We argue that a generalised RPS game can capture a larger class of scenarios that the classical 2-player 3-action version cannot account for.

Within the *reinforcement learning* (RL) domain, policy-based methods have become increasingly popular, the introduction of Proximal Policy Optimisation (PPO) being a major breakthrough for the field [8]. This method searches directly through the policy space and addresses the issue of large policy updates by means of a clipped objective function. Unlike value-based methods that most

*This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI â UEFISCDI, project number PN-III-P4-ID-PCE-2020-0800, within PNCDI III.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

existing investigations use in game-theoretic scenarios, PPO is a naturally good fit for learning mixed strategies.

The RPS game has been studied abundantly. A version with configurable number of actions of the game has been incorporated into the Petting-Zoo library [9], however it still only accounts for two agents. Other studies, such as [10] use RPS as an example benchmark. Most of the literature however considers evolutionary dynamics in the RPS [11, 12, 13]. In [12] the effects of population flow are investigated on the community of agents, while Jia et al. [11] considers feedback from the environment and finds evolutionarily stable strategies.

The aim of this paper is to study the RPS game in its general form that can account for multiple players ($N \geq 2$) and multiple actions ($M \geq 3$). While the generalization of the game across the number of actions is well known, to our best knowledge no existing work generalized across the number of players as well. This works aims to address this, and we call the resulting game the *N-Player M-action Rock-Paper-Scissors* game (NMRPS). The work of Lanctot et al. [14] comes closest to our approach. They proposed a benchmark to evaluate RL agents against RPS playing bots. The proposed benchmark however is based on the 2-Player iterated RPS, with population dynamics. We argue that a simultaneous game of N players is equally relevant. Nonetheless, our metrics are inspired by this work and thus comparable to it. We borrow the concept of exploitability and measure performance according to it when applicable. Our second contribution is studying the behavior of RL agents driven by PPO in the iterated NMRPS game, guided by the following research questions: **Q1**: *To what extent can PPO exploit fixed strategies and dominate them?* **Q2**: *What emergent behavior can be observed on multiple PPO-agents co-evolving in this environment, and how do these behaviors depend on the number of agents, actions and histories?*

2 A theoretical model for the N-Player M-action RPS game

We introduce the *N-Player M-action RPS* (NMRPS) as a normal form game. The generalization across the number of actions is straightforward and is a well-known exercise in the community. To preserve the balance of the game, every action has to win over exactly as many actions as many it defeats (thus M must be odd). Further, consider the actions as the nodes of a directed graph, where the directed edge (a, b) means a beats b. This graph structure boils down to a *regular tournament*, since all actions have to be related. One straightforward way to construct such a graph is to label the actions as integers from 0 to M-1 and consider that the winner between two distinct actions $W(a, b) = max(a, b)$ when a and b are both odd or both even, and $W(a, b) = min(a, b)$ otherwise. Due to symmetry, min and max can be reversed. Further, generalizing across the number of players is similarly intuitive. Considering $N \geq 2$ players, the joint action profile of the game would be a vector of N numbers. The payoff of a player i is computed by adding the number of actions from the profile that i-s action beats and subtracting the ones that defeat it. Table 1 shows the flattened

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

payoff matrix of this generalized game for three players and three actions.

1/(2,3)	**S, S**	**S, P**	**S, R**	**P, S**	**P, P**	**P, R**	**R, S**	**R, P**	**R, R**
S	0,0,0	1,1,-2	-1,-1,2	1,-2,1	2,-1,-1	0,0,0	-1,2,-1	0,0,0	-2,1,1
P	-2,1,1	-1,2,-1	0,0,0	-1,-1,2	0,0,0	1,1,-2	0,0,0	1,-2,1	2,-1,-1
R	2,-1,-1	0,0,0	1,-2,1	0,0,0	-2,1,1	-1,2,-1	1,1,-2	-1,-1,2	0,0,0

Table 1: Payoff matrix with $N = 3$ and $M = 3$

It can be trivially seen that pure strategies are easily exploitable. A mixed-strategy Nash-equilibrium must make players indifferent to each other's strategies. If each player has a mixed strategy, then in a mixed strategy profile s, we can write the probability of player i choosing action j as p_{ij}. Further, let's denote an action profile as $a = (a_0, a_1, ..., a_{N-1})$ where a_i is the action chosen by player i. The payoff of player i given an action profile a is $P_i(a)$. With a slight abuse of notation, we can denote the utility of i doing action j in the presence of the other players' action profile a_{-i} as $P_i(j, a_{-i})$. Then, each player's utility for doing a certain fixed action j must be the same as the utility for all other actions, and this has to hold true for all players. The utility of a player i doing action j given a mixed strategy profile of opponents as s_{-i} is $U_i(j, s_{-i}) = \sum_{a_{-i}} P_i(j, a_{-i}) \cdot \mathbb{P}(a_{-i}), \quad \forall i \in \{0, \dots, N-1\}, \forall j \in \{0, ..., M-1\}$, where: $a_{-i} = (a_0, \dots, a_{i-1}, a_{i+1}, \dots, a_{N-1})$ and the probability of one action profile a_{-i} is defined as $\mathbb{P}(a_{-i}) = \prod_{k=0, k\neq i}^{N-1} p_{ka_k}$. Moreover, as every mixed strategy must be a valid distribution over actions, we must have for all players i that $\sum_{0 \leq j \leq M-1} p_{ij} = 1, \quad \forall i \in \{0, 1, ..., N-1\}$. Since for each action the number of defeated and defeating actions is the same, the unweighted sum of payoffs for each player $U_i(j, s_{-i}) = \sum_{a_{-i}} P_i(j, s_{-i})$ is zero, which implies that a solution of the system consists of those p_{ij}-s for which all $\mathbb{P}(a_{-i})$ values are equal. One can observe that $p_{ij} = \frac{1}{M}$ is a solution of the system for all actions j and all players i, yielding 0 reward in expectation. Thus we have demonstrated that in the extended game, uniform randomization is still a Nash-equilibrium, despite the variable number of agents and actions. Nonetheless, the proposed model exposes an interesting kind of strategic complexity. From the above rules it follows that independently of the opponents' strategies, any player that randomizes uniformly will receive 0 payoff in expectation. However, such strategy configurations are not equilibria, as playing non-uniformity still yields an incentive to exploitation from the opponents, which makes the game particularly interesting for larger N and M.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

3 Results and discussion

Having introduced a suitable game-theoretic test-bed for MARL, this section aims to present some experiments conducted on the proposed environment. We study the behavior of intelligent players in the iterated NMRPS, powered by PPO individually, against fixed strategies and against each other. Thus, the entire setting can be viewed as a decentralized MARL system. The environment state of each agent, fed to PPO is obtained by concatenating the numerical values of the last H actions made by each opponent into one vector. Each experiment runs 200,000 total timesteps, divided into episodes of 200 iterations of the game, over which average reward is computed to assess performance and convergence. We used the best-converging hyperparameters from N=2 experiments for all tests, as alterations didn't significantly affect results despite non-convergence for $N > 2$: $batch\_size = 200$, $\gamma = 0.99$, $\lambda = 0.95$ and a learning rate of $3 \cdot 10^{-4}$. We consider $H \in \{1, 3, 5, 10, 15\}$, $M \in \{3, 5, 7, 11\}$ and $N \in \{2, 3, 4, 5, 6\}$.

Additionally, we provide two intentionally suboptimal fixed strategies of the game to evaluate the performance of learning agents: (1) *Distribution* - chooses actions according to a distribution over actions; and (2) *Cycler* - chooses all actions in a round-robin fashion from a fixed array of actions.

Answering **Q1**, through our experiments we find that one *PPO* agent easily exploits non-uniform *Distribution* agents. By means of obtainable reward, playing against multiple distribution players is equivalent to playing against one with the mean distribution of all the rest. Exploitability of fixed players can be measured by comparing the obtained reward against the maximum achievable reward, and we use this as a performance metric against fixed players. For example, for $M = 3$ and distribution $(0.7, 0.2, 0.1)$ over the actions $(0, 1, 2)$, an opponent can exploit by playing 2 all the time, getting in expectation $0.7 - 0.2 + 0 = 0.5$, which in 200 rounds yields a maximum average reward of 100. Average reward was calculated as the running mean reward of the last 100 episodes. Experiments against fixed players have been repeated four times and averaged.

We find that playing against distribution players, PPO can achieve 98% of the maximum achievable reward when $H \leq 3$, and decreasing as H grows, getting down to as low ass 87% in average when $H = 15$. Increasing N decreases performance to a similar extent, however increasing M seems not to hurt it. The exploitability of cyclers mostly depends on the regularity of the cycles and on H. Here, increasing H yields better results (around 99% when H is at least as large than the cycle length and steadily decreasing). Increasing the number of agents does not have any effect(we consider all opponents have the same cycle length), however increasing M slightly improves performance since it increases the predictability of patterns in the history.

Addressing **Q2** is arguably the most compelling. Pitting multiple learning agents together, convergence to the Nash-equilibrium happens only in the simplest of cases. For $N = 2$, $M = 3$ and $H = 1$, the system successfully converges to the Nash equilibrium. For $N = 3$ and $M = 3$, the system converges with a smaller learning rate of 0.0001 to approximately uniform randomness after some

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

initial fluctuations. For both larger N or M, the non-stationarity of the system becomes prevalent and we observe divergence. Whenever all agents except one learn to uniformly randomize, each agent will get 0 payoff in expectation, including the last one, which is now thus indifferent to its own strategy. Then however, all the other agents are incentivised to exploit non-uniformity, making themselves exploitable. Systems comprised of more than three PPO agents diverge in such a cycling fashion, average rewards going up and down indefinitely. This phenomenon however effectively captures price or social cycling, empirically observed in humans [15]. Additionally, we find that this cycling behavior emerges even with $N = 2$, whenever $M > 5$. Longer history lengths and having more actions also seem to increase the amplitude of fluctuation in average reward. Figure 1 shows the average reward of PPO agents against each other as they evolve through time.

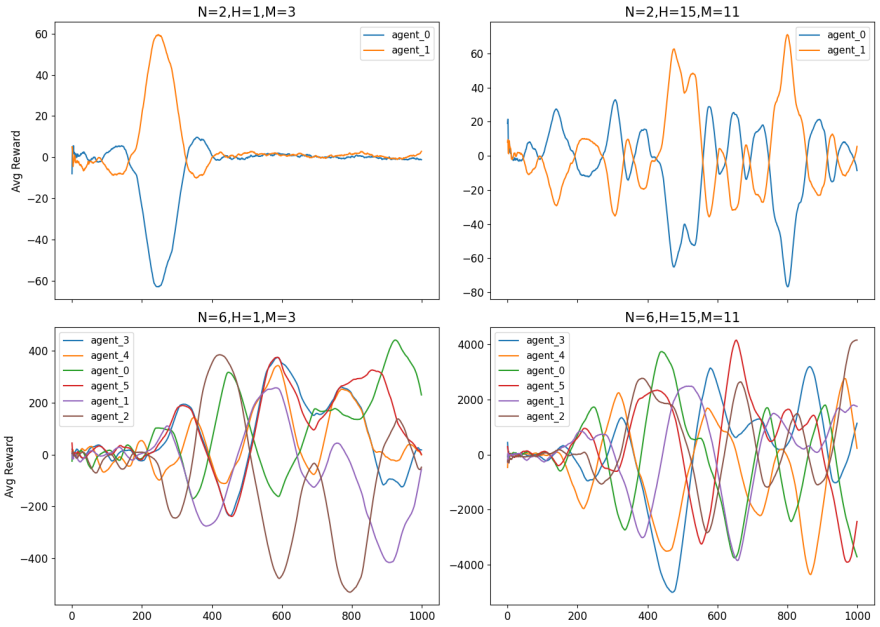

Fig. 1: Convergence and cycling of $N = 2$ and $N = 6$ PPO agents with different history lengths and numbers of actions

4 Conclusions and future work

In this work we have successfully generalized the well-known Rock-Paper-Scissors game into a model with the key property of having uniform randomization as Nash equilibrium. We showcased its relevance by experimenting with agents driven by deep RL methods, and explored some emergent behavior under this model. However, further investigations are necessary to properly understand the model itself, whenever N and M vary. One particularly interesting line of

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

future work would be to explore meta-learning against different fixed or adaptive opponents, in order to study how well learning agents can exploit different patterns. Better explanations and further investigations of the non-convergence of the system and the cycling behavior are also needed.

Understanding the behavior and incentives of individual agents in MARL systems remains a particularly difficult problem. The provided game model proves to be useful in this pursuit, and our conducted experiments uncover valuable insights into the intricacies of MARL.

References

[1] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

[2] Ammar Haydari and Yasin Yılmaz. Deep reinforcement learning for intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):11–32, 2020.

[3] Arnu Pretorius, Scott Cameron, Elan Van Biljon, Thomas Makkink, Shahil Mawjee, Jeremy du Plessis, Jonathan Shock, Alexandre Laterre, and Karim Beguir. A game-theoretic analysis of networked system control for common-pool resource management using multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 33:9983–9994, 2020.

[4] Jakob N Foerster, Richard Y Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. *arXiv preprint arXiv:1709.04326*, 2017.

[5] Alexander Peysakhovich and Adam Lerer. Towards ai that can solve social dilemmas. In *2018 AAAI Spring Symposium Series*, 2018.

[6] Hai-Jun Zhou. The rock–paper–scissors game. *Contemp. Phys.*, 57(2):151–163, 2016.

[7] Erik Brockbank and Edward Vul. Formalizing opponent modeling with the rock, paper, scissors game. *Games*, 12(3):70, 2021.

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

[9] J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, et al. Pettingzoo: Gym for multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:15032–15043, 2021.

[10] Vivek F Farias, Ciamac C Moallemi, Benjamin Van Roy, and Tsachy Weissman. Universal reinforcement learning. *IEEE Trans. Inf. Theory*, 56(5):2441–2454, 2010.

[11] Susu Jia, Zheng Kuang, Haihui Cheng, Xinzhu Meng, Tasawar Hayat, and Aatef Hobiny. Analysis and simulations of the rock-paper-scissors game with environmental feedback. In *Proceedings of the 3rd Asia-Pacific Conference on Image Processing, Electronics and Computers*, pages 547–551, 2022.

[12] Junpyo Park. Evolutionary dynamics in the rock-paper-scissors system by changing community paradigm with population flow. *Chaos, Solitons & Fractals*, 142:110424, 2021.

[13] Timothy N Cason, Daniel Friedman, and Ed Hopkins. Cycles and instability in a rock–paper–scissors population game: A continuous time experiment. *Review of Economic Studies*, 81(1):112–136, 2014.

[14] Marc Lanctot, John Schultz, Neil Burch, Max Olan Smith, Daniel Hennes, Thomas Anthony, and Julien Perolat. Population-based Evaluation in Repeated RPS as a Benchmark for Multiagent Reinforcement Learning. *arXiv preprint arXiv:2303.03196*, 2023.

[15] Zhijian Wang, Bin Xu, and Hai-Jun Zhou. Social cycling and conditional responses in the rock-paper-scissors game. *Scientific reports*, 4(1):1–7, 2014.