GOV-REK: Governed Reward Engineering Kernels for Designing Robust Multi-Agent Reinforcement Learning Systems

Extended Abstract

Ashish Rana
Institute for Enterprise Systems
University of Mannheim, Germany
ashish.rana@students.uni-mannheim.de

Michael Oesterle
Institute for Enterprise Systems
University of Mannheim, Germany
michael.oesterle@uni-mannheim.de

Jannik Brinkmann
Institute for Enterprise Systems
University of Mannheim, Germany
jannik.brinkmann@uni-mannheim.de

ABSTRACT

For multi-agent reinforcement learning (MARL) systems, the problem task often involves massive problem-specific reward engineering effort. This effort is usually not directly transferable to other problems; worse, this problem is further exacerbated for sparse reward scenarios. We propose **GOV**erned **R**eward **E**ngineering **K**ernels (GOV-REK), which dynamically assign reward distributions to agents in MARLs during the learning stage. We also introduce governance kernels, which exploit the underlying structure in either state or joint action space for assigning meaningful agent rewards. We demonstrate, using a Hyperband-like problem-agnostic algorithm, that this approach successfully learns to solve different MARL problems by iteratively exploring multiple reward models.

KEYWORDS

Cooperative Multi-Agent Systems; Sparse Reinforcement Learning; Robust Multi-Agent Systems; Reward Shaping

ACM Reference Format:
Ashish Rana, Michael Oesterle, and Jannik Brinkmann. 2024. GOV-REK: Governed Reward Engineering Kernels for Designing Robust Multi-Agent Reinforcement Learning Systems: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

The interactions formulated in MARLs are intricate to learn at larger scales, and this problem is further aggravated for sparse problem scenarios [10, 13, 19, 25]. Previously, many approaches have explored designing reward model problems in single-agent and multi-agent settings, but these efforts are problem-specific and often not generalizable to other MARLS [4–6, 15, 20, 27]. Past approaches have also improved sample efficiency by using novelties like attention [28], curiosity [1, 7], and experience sharing [8], but they have not directly influenced agent motivations. Therefore, building effective and robust reward models for agents in MARL tasks in an automated and problem-agnostic manner to improve baselines is still a challenging problem [30].

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

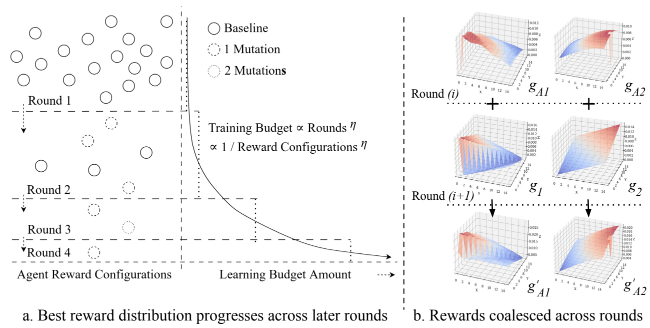

Figure 1: The reward model exploration and superposition mechanisms with increasing training timestep budgets.

Previously, reward shaping approaches that incorporate novel mechanisms, like learning ethical human behavior demonstrations [27], multi-objective reward shaping (MORS) [6], additional rewards for sub-goal completion [17], and context-sensitive rewards for agents [4], have shown further improvements. However, reward shaping is often susceptible to falling under continuous positive reward cycle traps, especially for sparse environments. For finding optimal policies, a multitude of systems like Autonomic Electronic Institutions (AEI) [3], Normative Multi-Agent Systems (NorMAS) [22], and Governed Multi-Agent Systems (GMAS) [24] have demonstrated their efficacy, where agents are provided with governing information for learning [9, 16, 26]. Our approach also proposes an intermediary governance layer between agents and environment, which directly incentivizes agents with additional rewards selected in an automated and problem-agnostic manner to improve the baseline MARL algorithms. Further, we define *governance kernels* for each agent, which are the reward distribution signals that generate similar additional rewards for similar states or joint actions depending on the MARL problem. Similar to problem-agnostic hyperparameter optimization algorithms like Successive Halving (SH) and Hyperband [18], the GOV-REK framework finds suitable reward models for agents by iteratively searching over different governance kernel configurations [2, 29]. Figure 1 demonstrates the execution of multiple SH rounds alongside the superposition of sample governance kernel configurations across these rounds.

We demonstrate the efficacy of this dynamic reward-based inductive bias to explore topologically similar state or joint-action spaces which incentivizes better cooperation amongst agents in MARLs.

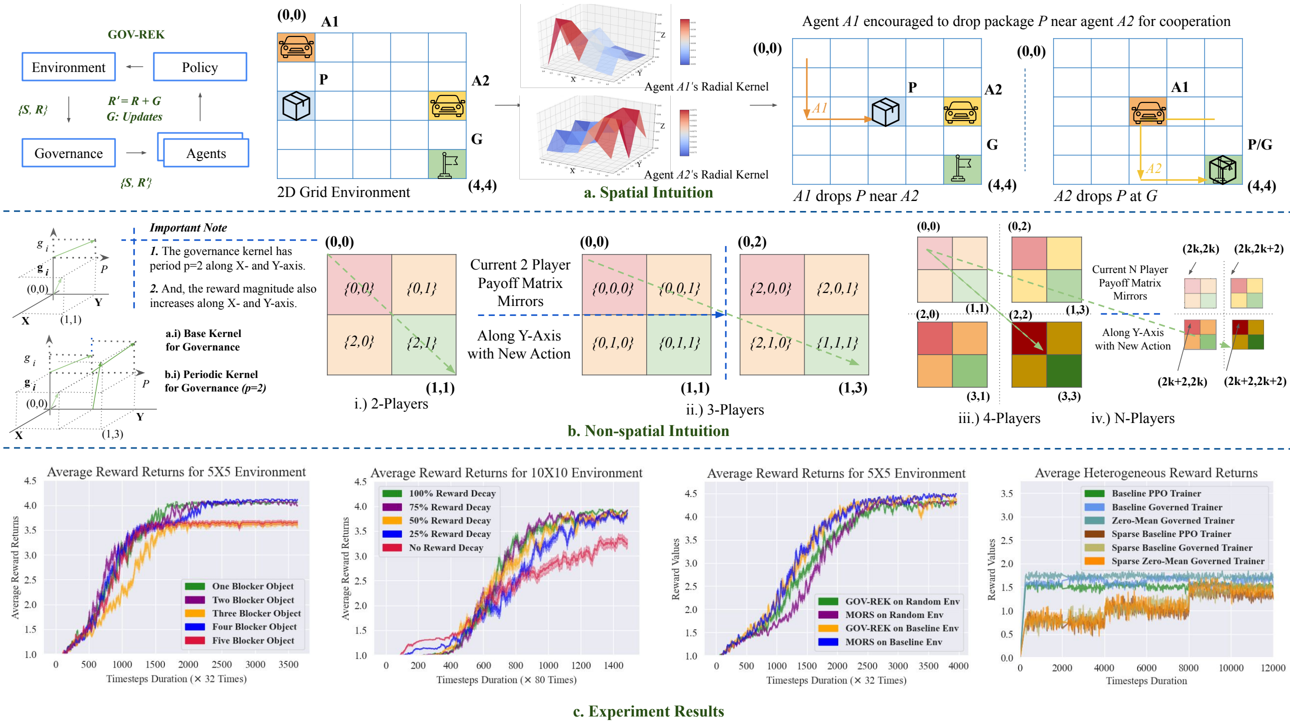

Figure 2: *a.* The additional sample radial governance kernels in the package delivery cooperative MARL problem are represented, where the kernels encourage higher local exploration to promote cooperation. *b.* The topological trend in the flattened joint-action space for the non-spatial social dilemma problem is exploited to define increasing periodic governance kernels that incentivize cooperation. *c.* The experiment demonstrates approach efficacy in four aspects: i.) robustness against the blocker objects, ii.) scalability with reward decays usage for the already explored states, iii.) better performance against manually defined MORS rewards, iv.) extensibility with higher average reward accumulation for the social dilemma problem.

2 APPROACH FORMULATION

We assume that the underlying learning algorithm is highly curious to select diverse actions, where all the relevant solution trajectories between the state-action transition pairs are explored, and define this as the *exploration expectation assumption*. Hence, we take an expectation with respect to the explored actions from the solution trajectories to define our reward models only as a function of state similarity, which is mathematically denoted by $E_a[R(s,a,s')] \longrightarrow R'(s,s')$, and we extend our results to joint-action spaces as well. This allows us to define governance kernels independent of agent transitions, and this is expressed by the relations $r'_i = r_i + g_{r,i}$ and $R' = R + G_r$ for agent-specific and agent-agnostic kernels respectively. In its generalized mathematical form, we express our governance kernels as $g_i(s_{a_i}) = \sigma^2\kappa(s_{a_i}, s'_{a_i}) + \xi$ in agent-specific and $G(s) = \sigma^2\kappa(s,s') + \xi$ in agent-agnostic non-parametric variations respectively, while staying compliant with a Potential Based Reward Shaping (PBRS) constraint for policy invariance [11, 12, 23]. Here, the kernel function is represented by κ, while σ upscales or downscales function values much similar to Gaussian kernels in Gaussian processes [14, 21], (s_{a_i}, s'_{a_i}) or (s,s') quantifies the magnitude variation between agent-specific or agent-agnostic states, and ξ represents the uniform noise in the kernel function. The GOV-REK framework uses repeated Hyperband executions and iteratively manipulates governance kernel configurations to find suitable agent reward models[1].

3 EXPERIMENTS

We evaluate the GOV-REK framework in CTCE and CTDE settings for a two-agent *sparse cooperative package delivery problem* and a sixteen-agent *heterogeneous social dilemma problem*. Figure 2 summarizes the experimental results quantifying robustness, scalability, performance and extensibility criteria with average expected reward returns for MARL tasks. Notably, we also observed a performance detriment at larger scales and asymmetric agent contribution problem in the CTCE setting experimentation.

4 CONCLUSION AND FUTURE WORK

We demonstrate that our proposed GOV-REK framework which defines simplistic reward models based on state or joint-action topological similarities helps agents to learn different MARL tasks effectively. Building upon this result, exploring a paradigm that trades between our rigid and simplistic reward exploration method against wholly fluid and complex state similarity learning methods is part of our future research effort [1, 28].

[1]The complete manuscript and experiment implementations are available at the repository: github.com/arana-initiatives/gov-rek-marls

REFERENCES

[1] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, et al. 2020. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038* (2020).

[2] James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of machine learning research* 13, 2 (2012).

[3] Eva Bou, Maite López-Sánchez, and Juan Antonio Rodríguez-Aguilar. 2006. Towards self-configuration in autonomic electronic institutions. In *International Workshop on Coordination, Organizations, Institutions, and Norms in Agent Systems*. Springer, 229–244.

[4] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. 2015. Reinforcement learning from demonstration through shaping. In *Twenty-fourth international joint conference on artificial intelligence*.

[5] Tim Brys, Anna Harutyunyan, Matthew E Taylor, and Ann Nowé. 2015. Policy Transfer using Reward Shaping.. In *AAMAS*. 181–188.

[6] Tim Brys, Anna Harutyunyan, Peter Vrancx, Matthew E Taylor, Daniel Kudenko, and Ann Nowé. 2014. Multi-objectivization of reinforcement learning problems by reward shaping. In *2014 international joint conference on neural networks (IJCNN)*. IEEE, 2315–2322.

[7] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894* (2018).

[8] Filippos Christianos, Lukas Schäfer, and Stefano Albrecht. 2020. Shared experience actor-critic for multi-agent reinforcement learning. *Advances in neural information processing systems* 33 (2020), 10707–10717.

[9] Rosaria Conte, Rino Falcone, and Giovanni Sartor. 1999. Agents and norms: How to fill the gap? *AI & L.* 7 (1999), 1.

[10] Christoph Dann and Emma Brunskill. 2015. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems* 28 (2015).

[11] Sam Devlin and Daniel Kudenko. 2011. Theoretical considerations of potential-based reward shaping for multi-agent systems. In *The 10th International Conference on Autonomous Agents and Multiagent Systems*. ACM, 225–232.

[12] Sam Michael Devlin and Daniel Kudenko. 2012. Dynamic potential-based reward shaping. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems*. IFAAMAS, 433–440.

[13] Wei Du and Shifei Ding. 2021. A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications. *Artificial Intelligence Review* 54 (2021), 3215–3238.

[14] David Duvenaud. 2014. *Automatic model construction with Gaussian processes*. Ph.D. Dissertation. University of Cambridge.

[15] Mahmoud Elbarbari, Kyriakos Efthymiadis, Bram Vanderborght, and Ann Nowé. 2021. Ltlf-based reward shaping for reinforcement learning. In *Adaptive and Learning Agents Workshop*, Vol. 2021.

[16] Marc Esteva, Juan-Antonio Rodriguez-Aguilar, Carles Sierra, Pere Garcia, and Josep L Arcos. 2001. On the formal specification of electronic institutions. In *Agent Mediated Electronic Commerce: The European AgentLink Perspective*. Springer, 126–147.

[17] Anna Harutyunyan, Sam Devlin, Peter Vrancx, and Ann Nowé. 2015. Expressing arbitrary reward functions as potential-based advice. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[18] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. *Automated machine learning: methods, systems, challenges*. Springer Nature.

[19] Nan Jiang and Alekh Agarwal. 2018. Open problem: The dependence of sample complexity lower bounds on planning horizon. In *Conference On Learning Theory*. PMLR, 3395–3398.

[20] Hyeoksoo Lee, Jiwoo Hong, and Jongpil Jeong. 2022. MARL-Based Dual Reward Model on Segmented Actions for Multiple Mobile Robots in Automated Warehouse Environment. *Applied Sciences* 12, 9 (2022), 4703.

[21] Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press.

[22] Emery A. Neufeld. 2022. Reinforcement Learning Guided by Provable Normative Compliance. In *Proceedings of the 14th International Conference on Agents and Artificial Intelligence, ICAART 2022, Volume 3, Online Streaming, February 3-5, 2022*, Ana Paula Rocha, Luc Steels, and H. Jaap van den Herik (Eds.). SCITEPRESS, 444–453. https://doi.org/10.5220/0010835600003116

[23] Andrew Y Ng, Daishi Harada, and Stuart Russell. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, Vol. 99. 278–287.

[24] Michael Oesterle, Christian Bartelt, Stefan Lüdtke, and Heiner Stuckenschmidt. 2022. Self-learning governance of black-box multi-agent systems. In *International Workshop on Coordination, Organizations, Institutions, Norms, and Ethics for Governance of Multi-Agent Systems*. Springer, 73–91.

[25] Karl Tuyls and Gerhard Weiss. 2012. Multiagent learning: Basics, challenges, and prospects. *Ai Magazine* 33, 3 (2012), 41–41.

[26] Danny Weyns, Sven A Brueckner, and Yves Demazeau. 2008. *Engineering Environment-Mediated Multi-Agent Systems: International Workshop, EEMMAS 2007, Dresden, Germany, October 5, 2007, Selected Revised and Invited Papers*. Vol. 5049. Springer.

[27] Yueh-Hua Wu and Shou-De Lin. 2018. A low-cost ethics shaping approach for designing reinforcement learning agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[28] Baicen Xiao, Bhaskar Ramasubramanian, and Radha Poovendran. 2022. Agent-Temporal Attention for Reward Redistribution in Episodic Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2201.04612* (2022).

[29] Steven R Young, Derek C Rose, Thomas P Karnowski, Seung-Hwan Lim, and Robert M Patton. 2015. Optimizing deep learning hyper-parameters through an evolutionary algorithm. In *Proceedings of the workshop on machine learning in high-performance computing environments*. 1–5.

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