Entropy Seeking Constrained Multiagent Reinforcement Learning

Extended Abstract

Ayhan Alp Aydeniz*
Collaborative Robotics and Intelligent Systems Institute
Oregon State University

Enrico Marchesini*
Laboratory for Information Decision Systems
Massachusetts Institute of Technology

Christopher Amato
Khoury College of Computer Sciences
Northeastern University

Kagan Tumer
Collaborative Robotics and Intelligent Systems Institute
Oregon State University

ABSTRACT

Multiagent Reinforcement Learning (MARL) has been successfully applied to domains requiring close coordination among many agents. However, real-world tasks require safety specifications that are not generally considered by MARL algorithms. In this work, we introduce an Entropy Seeking Constrained (ESC) approach aiming to learn safe cooperative policies for multiagent systems. Unlike previous methods, ESC considers safety specifications while maximizing state-visitation entropy, addressing the exploration issues of constrained-based solutions.

KEYWORDS

Multiagent Reinforcement Learning; Safety; Exploration

ACM Reference Format:
Ayhan Alp Aydeniz*, Enrico Marchesini*, Christopher Amato, and Kagan Tumer. 2024. Entropy Seeking Constrained Multiagent Reinforcement Learning: 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

Multiagent Reinforcement Learning (MARL) has achieved impressive results in many realistic cooperative domains such as remote exploration missions [7, 11, 22]. Such approaches generally do not handle safety specifications, which are crucial for real-world tasks [15, 18, 19, 30]. In contrast, constrained optimization has been widely used in single-agent RL to foster safety [24, 29]. However, constraints limit exploration, which is critical to discovering effective cooperative behaviors among multiple agents [14, 25, 27].

We introduce *Entropy Seeking Constrained (ESC)*-MARL, fostering efficient exploration while satisfying constraints for multiagent systems. To this end, we derive an *exploration-driven reward* that increases the diversity of visited states, leading to *state entropy maximization* [2, 4, 26]. Entropy maximization encourages agents to actively explore the scenario when constraints are satisfied, addressing the typical trade-off between safety and performance [3].

Our evaluation in the multi-rover exploration task [1], shows the efficacy of ESC-MARL at optimizing two crucial objectives; *defined constraint* (*i.e.*, collision avoidance), and team *coordination*, outperforming the underlying constrained MARL baseline.

*Equal contribution; aydeniza@oregonstate.edu, emarche@mit.edu.

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).

2 PRELIMINARIES

The multiagent domain that we consider is a fully cooperative tasks. In the literature, these problems are commonly represented as Dec-POMDPs [21]. In particular, the *Centralized Training with Decentralized Execution (CTDE)* learning paradigm has been recently investigated as a way to centralize information at learning time while maintaining decentralized execution in Dec-POMDPs [16, 17, 23, 31]. Hence, it overcomes typical real-world issues related to sharing information (*e.g.*, limited communication bandwidth) and centralization. In addition, real multiagent systems require satisfying desired safety specifications, that are usually modeled as constraints. However, previous methods disregard the negative impact of constraints on exploration [9, 12]. This leaves a significant margin for developing novel safe MARL algorithms such as ESC-MARL.

2.1 Entropy Seeking Constrained Agents

Considering MAPPO as the baseline algorithm [31], ESC-MARL learns a centralized advantage estimator $A_\phi(\boldsymbol{h}, \boldsymbol{u})$ over the joint history $\boldsymbol{h}$ and actions $\boldsymbol{u}$ parametrized by ϕ, and a policy π_{θ_i} parametrized by θ_i for each agent $i \in \mathcal{N}$. Policies' parameters are updated as:

$$\max_{\theta_i} \min_{\boldsymbol{\lambda}} \mathbb{E}_{\pi_{\theta_i}} \Big[\min \Big(q(\theta_i, \theta_i') A_\phi(\boldsymbol{h}, \boldsymbol{u}), \\ \text{clip}\left(q(\theta_i, \theta_i'), 1-\epsilon, 1+\epsilon\right) A_\phi(\boldsymbol{h}, \boldsymbol{u}) \Big) \\ + q(\theta_i, \theta_i') \mathcal{L}_C(\boldsymbol{\lambda}) \Big] \tag{1}$$

where

$$q(\theta_i, \theta_i') = \frac{\pi_\theta(u_i|h_i)}{\pi_{\theta'}(u_i|h_i)}; \\ A_\phi(\boldsymbol{h}, \boldsymbol{u}) = r + \gamma V_\phi(\boldsymbol{h}') - V_\phi(\boldsymbol{h}) \tag{2}$$

with $\boldsymbol{h}'$ being the joint history after performing $\boldsymbol{u}$.

2.1.1 Team-level constraints. The role of $\mathcal{L}_C(\boldsymbol{\lambda})$ varies depending on the type of constraints we are considering. Here we discuss a single team-level constraint c, with a hard-coded threshold l. Following the Lagrangian formalization [20], we incorporate the constraint by learning a multiplier $\boldsymbol{\lambda}$, a centralized cost-advantage estimator $A_c(\boldsymbol{h}, \boldsymbol{u})$, and minimize the following:[1]

$$\min_{\lambda \geq 0} -\boldsymbol{\lambda}(A_c(\boldsymbol{h}, \boldsymbol{u}) - l) \tag{3}$$

The observation-visitation entropy component is then incorporated into the agent-specific reward as follows.

2.1.2 Observation Entropy Maximization. ESC-MAPPO considers rewards at two levels, *agent-level* (for agent-specific behaviors) and *team-level* (for overall team policy). In particular, the entropy-maximizing rewards are agent-specific and aim at decreasing the recurrence of similar observations [2]. Algorithm 1 describes how we compute such entropy-maximizing rewards. The core idea is that most states will likely occur once in continuous state spaces, which are typical of real-world applications. Hence, we first employ a *quantization* mechanism [2] to cluster similar observation vectors together. This allows us to distinguish observations to maximize the entropy of the distribution of the visited ones during an episode, enabling us to avoid premature entropy maximization.

Algorithm 1: Computes a sequence of local entropy seeking reward for agent i over a single episode.

Initialize state s^0, history $h_i^0 = \{o_i^0\}$.
for $t \in [0, h-1]$ **do**
  Retrieve action a_i^t, state s^{t+1} and observation o_i^{t+1}
  $count \leftarrow 1$
  for $o_i \in h_i^t$ **do**
    if $quantize(o_i^{t+1}) == quantize(o_i)$ **then**
      $count \leftarrow count + 1$
  $reward \leftarrow \frac{1}{count}$
  $h_i^{t+1} \leftarrow h_i^t \cup \{a_i^t, o_i^{t+1}\}$
  if $V(h_i^{t+1}) > 0$ **then**
    $reward \leftarrow reward * V(h_i^{t+1})$
  Yield *reward*

To summarize, the proposed framework enhances agents with the local entropy maximizing reward [2] to keep exploration active. Once constraints are satisfied and the optimization process "focuses" on maximizing the main task objective (the discounted team return), agents maximize the entropy of the *intra-episode* observation distribution. Specifically, they explore novel parts of the space to find policies with higher payoffs (*i.e.*, better behaviors).

3 EXPERIMENTS

We investigate the performance of our algorithm in a well-known multiagent coordination problem, multi-rover domain [1], considering 8 agents and couplings of 3 and 4. The coupling value indicates the number of rover that has to simultaneously visit a point of interest, in order to collect a positive reward. In this task, the safety requirement is collision avoidance; each collision triggers a positive cost and we aim at limiting their accumulation up to a hard-coded threshold highlighted as a dashed line in Figure 1.

Due to the importance of having statistically significant results, we report the Pareto frontier of average return versus cost at convergence over 10 runs per method in Figure 1. We refer to our approach as ESC-MARL (T), indicating we are using team-level constraints, and ESC-MARL when using a separate constraint per each agent. To address the partially observable nature of the tasks [10, 28], we incorporate a GRU [5] layer into the network and use weight sharing to speed up the training process [8].

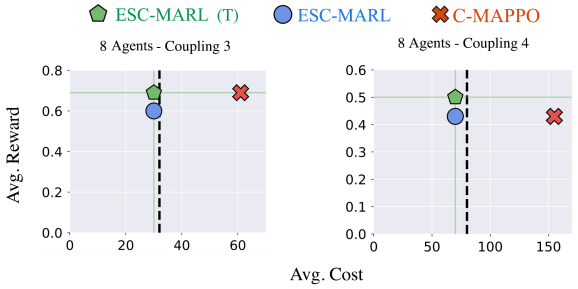

Figure 1: Average reward versus cost for the unconstrained MAPPO and (our) ESC-MARL and ESC-MARL (T) with 8 rovers and coupling factors of 3, and 4.

In general, ESC-MARL and ESC-MARL (T) have comparable task-objective performance (*i.e.*, average reward) to the unconstrained version, but, on average, they halve the cost. This translates into the rovers experiencing half of the number of collisions while navigating to visit the POIs. In addition, the team constraints of ESC-MARL (T) achieve higher performance than the individual constraints of ESC-MARL and the unconstrained MAPPO, especially when the number of rovers increases.

4 DISCUSSION AND FUTURE WORK

We address the problem of safe cooperative multiagent exploration, proposing a novel constrained MARL framework. Our framework leverages entropy maximization to enabling agents to have a much more effective search in the policy space, resulting in the discovery of policies that satisfy both the task objective and the constraints. We investigate the efficacy of the proposed method under various task complexities representing the dependency of each agent on its teammates. As this dependency increases, the performance of multiagent teams gets significantly affected as more complex cooperative behaviors are required. Results show that ESC-MARL achieves a comparable or better score on the team objective while learning safer policies that halve the number of collisions. Future work will consider extending our approach to different constraints (*e.g.*, probabilistic [6] and instantaneous) and research the effects of exploration-driven rewards in different multiagent contexts.

ACKNOWLEDGMENTS

This work was partially supported by the National Science Foundation with grant No. IIS-2112633 and the Air Force Office of Scientific Research with grant No. FA9550-19-1-0195.

[1]The centralized A_c does not scale well due to the cardinality of the joint action-observation history, but it leverages joint information to improve predictions [13].

REFERENCES

[1] Adrian K Agogino and Kagan Tumer. 2004. Unifying temporal and structural credit assignment problems. In *AAMAS*.

[2] Ayhan Alp Aydeniz, Robert Loftin, and Kagan Tumer. 2023. Novelty seeking multiagent evolutionary reinforcement learning. In *Proceedings of the Genetic and Evolutionary Computation Conference*. 402–410.

[3] Ayhan Alp Aydeniz, Enrico Marchesini, Robert Loftin, and Kagan Tumer. 2023. Entropy Maximization in High Dimensional Multiagent State Spaces. In *2023 International Symposium on Multi-Robot and Multi-Agent Systems (MRS)*. IEEE, 92–99.

[4] Ayhan Alp Aydeniz, Anna Nickelson, and Kagan Tumer. 2022. Entropy-based local fitnesses for evolutionary multiagent systems. In *GECCO*. 212–215.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[6] Stefan Bieniawski, Ilan Kroo, and David Wolpert. 2004. Discrete, continuous, and constrained optimization using collectives. In *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*. 4580.

[7] Mitchell K Colby and Kagan Tumer. 2012. Shaping fitness functions for coevolving cooperative multiagent systems.. In *AAMAS*, Vol. 1. Citeseer, 425–432.

[8] Jakob Foerster, Ioannis Alexandros Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to Communicate with Deep Multi-Agent Reinforcement Learning. In *NeurIPS*, Vol. 29.

[9] Shangding Gu, Jakub Grudzien Kuba, Munning Wen, Ruiqing Chen, Ziyan Wang, Zheng Tian, Jun Wang, Alois Knoll, and Yaodong Yang. 2021. Multi-Agent Constrained Policy Optimisation. In *arXiv*, Vol. abs/2110.02793.

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9 (1997).

[11] Junyan Hu, Hanlin Niu, Joaquin Carrasco, Barry Lennox, and Farshad Arvin. 2020. Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning. *IEEE Transactions on Vehicular Technology* 69, 12 (2020), 14413–14423.

[12] Jiajing Ling, Arambam James Singh, Duc Thien Nguyen, and Akshat Kumar. 2022. Constrained Multiagent Reinforcement Learning for Large Agent Population. In *ECML PKDD*.

[13] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *NeurIPS*.

[14] Enrico Marchesini and Christopher Amato. 2023. Improving Deep Policy Gradients with Value Function Search. In *The Eleventh International Conference on Learning Representations*.

[15] Enrico Marchesini, Davide Corsi, and Alessandro Farinelli. 2022. Exploring Safer Behaviors for Deep Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 36, 7 (2022), 7701–7709. https://doi.org/10.1609/aaai.v36i7.20737

[16] Enrico Marchesini and Alessandro Farinelli. 2021. Centralizing State-Values in Dueling Networks for Multi-Robot Reinforcement Learning Mapless Navigation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 4583–4588. https://doi.org/10.1109/IROS51168.2021.9636349

[17] Enrico Marchesini and Alessandro Farinelli. 2022. Enhancing Deep Reinforcement Learning Approaches for Multi-Robot Navigation via Single-Robot Evolutionary Policy Search. In *ICRA*. 5525–5531. https://doi.org/10.1109/ICRA46639.2022.9812341

[18] Enrico Marchesini, Luca Marzari, Alessandro Farinelli, and Christopher Amato. 2023. Safe Deep Reinforcement Learning by Verifying Task-Level Properties. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*. 1466–1475.

[19] Luca Marzari, Enrico Marchesini, and Alessandro Farinelli. 2023. Online Safety Property Collection and Refinement for Safe Deep Reinforcement Learning in Mapless Navigation. In *International Conference on Robotics and Automation (ICRA)*.

[20] J. Nocedal and S. Wright. 2006. *Numerical Optimization* (2 ed.). Springer.

[21] Frans A Oliehoek and Christopher Amato. 2016. *A concise introduction to decentralized POMDPs*. Springer.

[22] Aida Rahmattalabi, Jen Jen Chung, Mitchell Colby, and Kagan Tumer. 2016. D++: Structural credit assignment in tightly coupled multiagent domains. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4424–4429.

[23] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *ICML*.

[24] Julien Roy, Roger Girgis, Joshua Romoff, Pierre-Luc Bacon, and Chris J Pal. 2022. Direct Behavior Specification via Constrained Reinforcement Learning. In *ICML*, Vol. 162. 18828–18843.

[25] Lukas Schäfer, Oliver Slumbers, Stephen McAleer, Yali Du, Stefano V. Albrecht, and David Mguni. 2023. Ensemble Value Functions for Efficient Exploration in Multi-Agent Reinforcement Learning. In *arXiv*, Vol. abs/2302.03439.

[26] Younggyo Seo, Lili Chen, Jinwoo Shin, Honglak Lee, Pieter Abbeel, and Kimin Lee. 2021. State entropy maximization with random encoders for efficient exploration. In *International Conference on Machine Learning*. PMLR, 9443–9454.

[27] Kagan Tumer, Zachary T Welch, and Adrian Agogino. 2008. Aligning social welfare and agent preferences to alleviate traffic congestion. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems-Volume 2*. Citeseer, 655–662.

[28] Paul J Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 10 (1990), 1550–1560.

[29] Qisong Yang and Matthijs T.J. Spaan. 2023. CEM: Constrained Entropy Maximization for Task-Agnostic Safe Exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[30] Connor Yates, Ayhan Alp Aydeniz, and Kagan Tumer. 2021. Reactive Multi-Fitness Learning for Robust Multiagent Teaming. In *2021 International Symposium on Multi-Robot and Multi-Agent Systems (MRS)*. IEEE, 92–100.

[31] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Wang, Alexandre M. Bayen, and Yi Wu. 2022. The Surprising Effectiveness of MAPPO in Cooperative, Multi-Agent Games. In *NeurIPS*.