Safe Entropic Agents under Team Constraints

Extended Abstract

Ayhan Alp Aydeniz*
Oregon State University
Corvallis (OR), USA
aydeniza@oregonstate.edu

Enrico Marchesini*
Massachusetts Institute of Technology
Cambridge (MA), USA
emarche@mit.edu

Robert Loftin
University of Sheffield
Sheffield, UK
r.loftin@sheffield.ac.uk

Christopher Amato
Northeastern University
Boston (MA), USA
c.amato@northeastern.edu

Kagan Tumer
Oregon State University
Corvallis (OR), USA
kagan.tumer@oregonstate.edu

ABSTRACT

Safety is a critical concern in multiagent reinforcement learning (MARL), yet typical safety-aware methods constrain agent behaviors, limiting exploration—essential for discovering effective cooperation. Existing approaches mainly enforce individual constraints, overlooking potential benefits of joint (*team*) constraints. We analyze team constraints theoretically and practically, introducing *entropic exploration for constrained MARL* (E2C). E2C maximizes observation entropy to encourage exploration while ensuring safety at the individual and team levels. Experiments across diverse domains demonstrate that E2C matches or outperforms common baselines in task performance while reducing unsafe behaviors by up to 50%.

KEYWORDS

Multiagent reinforcement learning; safety; entropy maximization

ACM Reference Format:
Ayhan Alp Aydeniz*, Enrico Marchesini*, Robert Loftin, Christopher Amato, and Kagan Tumer. 2025. Safe Entropic Agents under Team Constraints: 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, 4 pages.

1 INTRODUCTION

Training many agents for real-world applications demands cooperative behaviors, balancing task performance and safety objectives. However, existing multi-agent reinforcement learning (MARL) methods enforce individual agent constraints to ensure safety but overlook its inherently team-based nature (*e.g.,* inter-agent collisions affecting overall success) [1, 6, 19, 23, 25]. Moreover, constraints limit exploration—critical for discovering cooperative behaviors—leading to suboptimal policies [7, 12–14]. This paper analyzes the impact of team constraints theoretically and introduce *entropic exploration for constrained MARL* (E2C) to enhance exploration through observation entropy maximization (OEM) [2, 22] while maintaining constraints. Our results in standard benchmarks show that E2C matches or outperforms existing baselines while maintaining significantly safer behaviors, particularly in complex coordination tasks where prior constrained methods fail.

Preliminaries and Related Work. Our cooperative multiagent tasks are commonly modeled as decentralized Markov decision processes (Dec-MDPs) [17], and focus on the popular paradigm of *centralized training with decentralized execution* (CTDE) [10, 11, 18, 26]. In particular, policy gradient-based methods such as MAPPO [26] have demonstrated strong performance in cooperative settings under the CTDE paradigm. Given MAPPO's robustness, we build E2C on top of it to incorporate safety constraints. In MARL, constraints introduce three competing objectives: individual agent objectives, cooperative task performance, and safety compliance. Existing constrained MARL methods primarily extend single-agent approaches by enforcing independent agent constraints while overlooking the cooperative nature of multiagent safety. Some works refine cost estimation and credit assignment [8, 9], but they fail to address how constraints fundamentally impact exploration. We analyze team-level safety constraints theoretically [24] and evaluate their empirical benefits in constrained MAPPO, using OEM to address the impact on exploration.

2 TEAM-BASED TRUST REGION BOUNDS

In this section, we extend the cost improvement bounds derived by the works [6, 24] for trust region MARL with individual constraints to the team settings. We follow the same assumptions of such previous works and extend their stateful lower bound on the cost improvement to team constraints. [1] When cooperative agents use joint (team) constraints, we define a set of cost functions $C := \{c_j\}_{j \in m}$ (the team has m cost functions). These functions take the form $c_j : \mathcal{S} \times \mathcal{U} \to \{0, 1\}$ with cost-limiting values $\boldsymbol{l} := \{l_j\}_{j \in m}$. After performing the joint action in the environment, the agents receive joint costs $c_j(s_t, \boldsymbol{u}_t)\ \forall j = 1, \dots, m$. On top of maximizing the expected discounted return, the agents now also try to satisfy a joint constrained objective for which optimal policies maximize the return for feasible policies (*i.e.,* the ones satisfying the constraints):

$$J_j(\pi) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c_j(s_t, \boldsymbol{u}_t) \right] \leq l_j,\ \forall j = 1, \dots, m, \qquad (1)$$

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).

* Equal contribution.

[1] When their state assumption does not hold, authors assume that agents using recurrent networks as decentralized policies can overcome partial observability.

To derive the cost improvement bound for team constraints, we define the corresponding joint cost value functions. For the j^{th} cost function, we define the j^{th} (stateful) value functions as follows:

$$Q_j^\pi(s,\boldsymbol{u}) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t c_j(s_t,\boldsymbol{u}_t)|s_0=s,\boldsymbol{u}_0=\boldsymbol{u}\right],$$
$$V_j^\pi(s) := \mathbb{E}_{\boldsymbol{u}\sim\pi}\left[Q_{j,\pi}(s,\boldsymbol{u})\right],\ A_j^\pi(s,\boldsymbol{u}) = Q_{j,\pi}(s,\boldsymbol{u}) - V_{j,\pi}(s). \tag{2}$$

In trust region-based methods, Equation 1 is difficult to optimize directly when considering a joint policy π and some other policy $\bar{\pi}^i$ of agent i. Hence, we define the surrogate objective for team constraints following the individual constraint case of Gu et al. [6].

Lemma1. *Let π be a joint policy, and $\bar{\pi}^i$ be some other policy of agent i. Then, for any of the joint costs of index $j = 1,\dots,m$, we define the surrogate cost objective as follows:*

$$L_j^\pi(\bar{\pi}^i) = \mathbb{E}_{\boldsymbol{u}^{-i}\sim\pi^{-i},u^i\sim\bar{\pi}^i}\left[A_j^\pi(s,\boldsymbol{u})\right],$$

where π^{-i} indicates the policy of all the agents except i.

Finally, we extend Lemma 4.3 of Gu et al. [6] to the case of team constraints, deriving a lower bound on how the expected joint costs change when the agents update their policies.

Lemma2. *Let π and $\bar{\pi}$ be joint policies. Let $i \in \mathcal{N}$ be an agent, and $j = 1,\dots,m$ be one of the joint cost indexes. The following holds:*

$$J_j(\bar{\pi}) \le J_j(\pi) + L_j^\pi(\bar{\pi}^i) + \nu_j \sum_{h=1}^{|\mathcal{N}|} D_{KL}^{max}(\pi^h,\bar{\pi}^h),$$
$$\text{where } \nu_j = \frac{4\gamma\max_{s,\boldsymbol{u}}|A_j^\pi(s,\boldsymbol{u})|}{(1-\gamma)^2}.$$

In practice, trust region algorithms ensuring Lemma 3 (or its equivalent version for the individual constraints proposed by Gu et al. [6]) are replaced by approximations relying on neural networks and tractable clipping operators that can scale to large state and action spaces [6, 20, 21], on top of which we build E2C. Proofs of the lemmas are discussed in the supplementary [4].

3 E2C

In the safe RL literature, the Lagrangian method [16] is commonly used to transform the constrained problem into an equivalent unconstrained one $\forall i \in \mathcal{N}$, using a dual variable as follows:

$$\mathcal{L}^\pi(\boldsymbol{\lambda}) = J_r^\pi - \mathcal{L}_C^\pi(\boldsymbol{\lambda}),$$
$$\mathcal{L}_C^\pi(\boldsymbol{\lambda}) = \begin{cases}\lambda_j^i\left(J_j^i(\pi) - l_j^i\right)\ \forall j=1,\dots,m^i & \text{individual}\\ \lambda_j\left(J_j(\pi)-l_j\right)\ \forall j=1,\dots,m & \text{team}\end{cases}, \tag{3}$$

where $\boldsymbol{\lambda}$ are the so-called *Lagrangian multipliers* and act as a penalty in the optimization objective of each agent. The goal is thus to solve the resulting *max min* problem: $\max_\pi \min_{\boldsymbol{\lambda}\ge 0}\mathcal{L}^\pi(\boldsymbol{\lambda})$. A typical solution to that is to iteratively take gradient ascent steps in π and descent in $\boldsymbol{\lambda}$. We build E2C on top of the Lagrangian MAPPO—a strong baseline across a variety of scenarios [6, 26]. The resultant E2C-MAPPO algorithms address the challenges of using constraints in multiagent systems by using *entropy enhanced agents* as in [5]. Following the MAPPO baseline, we learn a centralized advantage estimator $A_\phi(s,\boldsymbol{u})$ parametrized by ϕ, while each agent $i \in \mathcal{N}$ learns a policy π_{θ_i} parametrized by θ_i. Policies' parameters are updated using the following clipped objective:

$$\max_{\theta_i}\min_{\boldsymbol{\lambda}}\ \mathbb{E}_{\pi_{\theta_i}}\Big[\min\Big(q(\theta_i,\theta_i')A_\phi(s,\boldsymbol{u}),\text{clip}\left(q(\theta_i,\theta_i'),1-\epsilon,1+\epsilon\right)$$
$$A_\phi(s,\boldsymbol{u})\Big) + q(\theta_i,\theta_i')\mathcal{L}_C^{\pi_{\theta_i}}(\boldsymbol{\lambda})\Big], \tag{4}$$

where the centralized advantage measures the overall effect of selecting a joint action, $\mathcal{L}_C^{\pi_{\theta_i}}$ depends on the nature of constraints (*i.e.,* individual or team as in Equation 3), and $q(\theta_i,\theta_i') = \frac{\pi_{\theta_i}(u_i|h_i)}{\pi_{\theta_i'}(u_i|h_i)}$.

Experiments. We test how well does E2C-MAPPO solve standard cooperative tasks (with individual and team-based constraints) compared to a constrained (safe) baseline in two safe particle environment tasks [15], where the safety requirement is *collision avoidance*. Hence, when agents collide, they receive a positive cost value and they try to limit its accumulation under defined thresholds. Considering the twofold nature of E2C, we call E2C-MAPPO (T) the entropy maximizing algorithm using team constraints, and E2C-MAPPO the one with individual constraints. For a fair comparison, the threshold for each agent in the individual constraint case equals the team threshold divided by the number of agents (detailed in the following section). The results show in Fig. 1 show the the average return versus cost of 10 runs per method at convergence. Overall, E2C-based methods achieve higher performance than the baseline constrained algorithm, with the team version outperforming the others in one of the tasks.

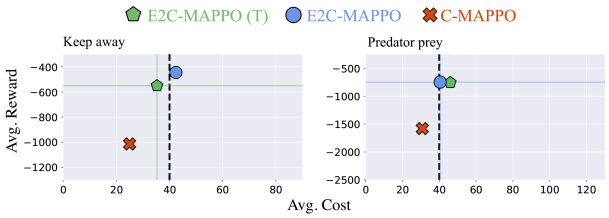

Figure 1: Average reward versus cost for the constrained C-MAPPO and (our) E2C-MAPPO (T) and E2C-MAPPO (team and individual constraints).

4 CONCLUSION

We address the challenges of team-based constrained MARL, where cooperation and safety are critical. E2C leverages observation entropy maximization [3] to enhance exploration while maintaining constraint satisfaction. By prioritizing team constraints and observation diversity, our approach mitigates excessive conservatism and fosters effective coordination. Experiments across diverse environments demonstrate E2C's ability to improve task performance while satisfying both individual and team constraints, outperforming conventional baselines. Future work can extend E2C to larger agent teams and real-world applications.

ACKNOWLEDGMENTS

This work was partially funded by the NSF award number 2044993 and by the Air Force Office of Scientific Research grant number FA9550-19-1-0195.

REFERENCES

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. 2017. Constrained policy optimization. In *International conference on machine learning*. PMLR, 22–31.

[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, Christopher Amato, and Kagan Tumer. 2024. Entropy Seeking Constrained Multiagent Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*. 2141–2143.

[4] Ayhan Alp Aydeniz, Enrico Marchesini, Robert Loftin, Christopher Amato, and Kagan Tumer. 2024. Safe Multiagent Coordination via Entropic Exploration. arXiv:2412.20361 [cs.MA] https://arxiv.org/abs/2412.20361

[5] 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.

[6] 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.

[7] Zhang-Wei Hong, Tzu-Yun Shann, Shih-Yang Su, Yi-Hsiang Chang, and Chun-Yi Lee. 2018. Diversity-Driven Exploration Strategy for Deep Reinforcement Learning. In *NeurIPS*.

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

[9] Chenyi Liu, Nan Geng, Vaneet Aggarwal, Tian Lan, Yuan Yang, and Mingwei Xu. 2021. CMIX: Deep Multi-agent Reinforcement Learning with Peak and Average Constraints. In *ECML-PKDD*.

[10] 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

[11] Enrico Marchesini and Alessandro Farinelli. 2022. Enhancing Deep Reinforcement Learning Approaches for Multi-Robot Navigation via Single-Robot Evolutionary Policy Search. In *2022 International Conference on Robotics and Automation (ICRA)*.

[12] 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.

[13] Luca Marzari, Changliu Liu, Priya L. Donti, and Enrico Marchesini. 2024. Improving Policy Optimization via ϵ-Retrain. In *arXiv*.

[14] Luca Marzari, Enrico Marchesini, and Alessandro Farinelli. 2023. Online Safety Property Collection and Refinement for Safe Deep Reinforcement Learning in Mapless Navigation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 7133–7139.

[15] Igor Mordatch and Pieter Abbeel. 2017. Emergence of Grounded Compositional Language in Multi-Agent Populations. *arXiv preprint arXiv:1703.04908* (2017).

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

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

[18] 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*.

[19] 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.

[20] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. 2015. Trust Region Policy Optimization. In *ICML*.

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. In *arXiv*.

[22] 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.

[23] Krishnan Srinivasan, Benjamin Eysenbach, Sehoon Ha, Jie Tan, and Chelsea Finn. 2020. Learning to be safe: Deep rl with a safety critic. *arXiv preprint arXiv:2010.14603* (2020).

[24] Mingfei Sun, Sam Devlin, Jacob Beck, Katja Hofmann, and Shimon Whiteson. 2023. Trust Region Bounds for Decentralized PPO Under Non-stationarity. In *AAMAS*.

[25] Brijen Thananjeyan, Ashwin Balakrishna, Suraj Nair, Michael Luo, Krishnan Srinivasan, Minho Hwang, Joseph E Gonzalez, Julian Ibarz, Chelsea Finn, and Ken Goldberg. 2021. Recovery rl: Safe reinforcement learning with learned recovery zones. *IEEE Robotics and Automation Letters* 6, 3 (2021), 4915–4922.

[26] 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*.