Bi-Level Reinforcement Learning for Multi-Robot Systems

Doctoral Consortium

Arjun Prakash
Brown University
Providence, RI, USA
arjun_prakash@brown.edu

ABSTRACT

I aim to build safe and collaborative multi-robot systems. My research so far has focused on bi-level optimization. In the constrained zero-sum case, I solved the reach-avoid problem, where one robot must reach a target area defended by another. In the unconstrained general-sum case, I am currently working on improving actor-critic reinforcement learning (RL) algorithms with low-rank approximations of the inverse-Hessian vector product to capture the dependency between actor and critic. I hope to extend my research to heterogeneous teams of robots by augmenting RL with classic control algorithms through differentiable programming and through continual multi-agent RL to organically learn diverse policies. These directions aim to advance scalable, safe, and collaborative AI for dynamic real-world environments.

KEYWORDS

Reinforcement Learning, Robotics; Multi-Agent Systems; Bi-Level Optimization; Algorithmic Game Theory

ACM Reference Format:
Arjun Prakash. 2025. Bi-Level Reinforcement Learning for Multi-Robot Systems: Doctoral Consortium. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 2 pages.

1 EXTENDED ABSTRACT

As AI becomes more ubiquitous, my aim is to build collaborative, safe, and pro-social autonomous robots that can continuously adapt to changing environments. I believe this is possible by developing multi-agent reinforcement algorithms with game theoretic principles.

My first project, Differential Stackelberg Markov Games with Applications to Mobile Robots was core to my development and understanding of game theory, reinforcement learning (RL) and control theory. The objective was to learn a reinforcement learning policy to avoid collisions. We formulated the problem as a discrete-time zero-sum Markov game where one player, *the evader* had to reach a target area while another player, *the pursuer* tried to defend the target area [4]. Our objective was to learn a safe policy for the evader such that it would be robust to the pursuer's policy. We did this by modeling the game as a two-player zero-sum game with dependent constraints—specifically, the evader was constrained to remain outside the pursuer's striking distance—which we solved using RL. Specifically, we developed and implemented a nested policy gradient algorithm where the evader was constrained by the position of the pursuer. I implemented this algorithm to solve the game assuming both a discrete action space by applying an action mask, and a continuous action space by building a Lagrangian into the reward function. Finally, I deployed the algorithm on a pair of drones, overcoming the sim2real gap by adding random noise to the transitions. Building constraints into an RL agent's behavior was a step towards building safer and more reliable AI [5].

While the previous interaction was best modeled by a zero-sum Stackelberg game with dependent constraints, there are many win-win situations in the real-world that are better modeled as general-sum games. Two-player Stackelberg games in which the players have individual utility functions are also called bilevel optimization problems (BLOPs) [2], where the upper-level player's action is observed by the lower-level player. While computing a global equilibrium is not generally tractable [8], I have been investigating local solutions , which are guaranteed to exist assuming a strongly-convex lower-level utility function. Solving a BLOP involves following the hypergradient of the upper-level objective. One of the most popular ways to do this is to leverage the implicit function theorem (IFT), which reduces the problem to computing an inverse Hessian-vector-product (IHVP) [7]. While many turn to iterative methods like conjugate-gradient to compute the IHVP, I have observed that the Hessians of neural networks are usually ill-conditioned, leading to arbitrarily bad solutions. Indeed, using conjugate gradient to compute the IHVP can actually make the solution worse. However, we have found success with the Nyström method [6], which computes a low-rank approximation of the IHVP. With these results in hand, I have developed an improved Stackelberg actor-critic algorithm [9]. The Stackelberg view of actor-critic is that there is an inherent dependency between the actor and the critic, so training them independently, as most RL algorithms do, leaves some performance gains on the table. Our current paper, which uses the Nyström method to approximate the hypergradient and nests the critic as in my earlier work, has led to improved performance on continuous control tasks.

One natural progression of this line of work would be to investigate general-sum games between teams of agents in mobile-robotic applications, such as logistics and transportation. One specific problem of interest to me is the coordination and optimization of teams of vehicles, such as drones and trucks, for efficient delivery operations. This type of setting is challenging as it involves an interplay between individual and group-level objectives, where agents

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).

must learn coordinated strategies while adhering to safety and operational constraints. A promising approach to addressing these challenges is to augment RL policies with differentiable control principles, such as model predictive control or barrier functions. In this framework, RL agents could adaptively set parameters that force the behavior of the control agents into a desired solution. This hybrid approach leverages the flexibility of RL to optimize the behavior of the entire system while maintaining the robustness and stability offered by control-based methods, enabling joint optimization of safety, efficiency, and task performance in complex multi-agent systems.

Another promising future direction is to investigate continual reinforcement learning for multi-agent systems. The main technique I propose to explore is continual backpropagation [3], which selectively resets the least important neurons to maintain plasticity in the network. This approach could yield particularly interesting results in the context of multi-agent systems, especially if it were to lead to the organic emergence of heterogeneous agents without pre-defining their roles or behaviors. As the agents continually learn and adapt their behavior to perform different tasks, their neural networks should develop distinct specializations, which can be beneficial for the overall performance of a team. For example, this natural differentiation could lead to more robust and adaptable systems, where agents collectively compensate for individual limitations or failures [1]. Furthermore, this approach could enable teams to develop both specialized expertise and the flexibility to adapt to new situations, addressing a fundamental tension in multi-agent system design.

These research directions present unique challenges that cannot be effectively addressed using traditional optimization methods alone due to the complexities introduced by the heterogeneous nature of the agents and the dynamic environments in which they operate. To tackle these challenges, I hope to break new ground in multi-robot research. The intractability of general-sum games means that scalable reinforcement learning will be key to capturing the intricate interactions and trade-offs between individual agent objectives and overall system performance. The potential impact of this research extends beyond specific applications, as the insights gained can contribute to the design of more efficient, robust, and adaptive multi-agent systems for various domains, such as manufacturing, healthcare, logistics and transportation. Furthermore, the frameworks and algorithms developed in this research can inform the design of human-robot collaborative systems, where humans and intelligent robots work together seamlessly to achieve shared objectives, and are core to building well aligned AI systems.

ACKNOWLEDGMENTS

Thank you to Amy Greenwald and Nora Ayanian for their supervision. This work was supported by the Office of Naval Research under grant number N000142412657 for Stackelberg Games: Applications and Solutions. Additional support was provided by Brown University through the OVPR Salomon Award, grant number GR100162.

REFERENCES

[1] Nora Ayanian. 2019. Dart: Diversity-enhanced autonomy in robot teams. *The International Journal of Robotics Research* 38, 12-13 (2019), 1329–1337.
[2] Stephan Dempe and Alain Zemkoho. 2020. Bilevel optimization. In *Springer optimization and its applications*. Vol. 161. Springer, Gewerbestrasse 11, 6330 Cham, Switzerland.
[3] Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. 2024. Loss of plasticity in deep continual learning. *Nature* 632, 8026 (Aug. 2024), 768–774. https://doi.org/10.1038/s41586-024-07711-7 Publisher: Nature Publishing Group.
[4] Jaime F Fisac, Mo Chen, Claire J Tomlin, and S Shankar Sastry. 2015. Reach-avoid problems with time-varying dynamics, targets and constraints. In *Proceedings of the 18th international conference on hybrid systems: computation and control*. Association for Computing Machinery, New York, NY, USA, 11–20.
[5] Denizalp Goktas, Arjun Prakash, and Amy Greenwald. 2024. Convex-Concave Zero-Sum Stochastic Stackelberg Games. *Advances in Neural Information Processing Systems* 36 (2024).
[6] Ryuichiro Hataya and Makoto Yamada. 2023. Nyström Method for Accurate and Scalable Implicit Differentiation. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 206)*, Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (Eds.). PMLR, 4643–4654. https://proceedings.mlr.press/v206/hataya23a.html
[7] Kaiyi Ji, Junjie Yang, and Yingbin Liang. 2021. Bilevel Optimization: Convergence Analysis and Enhanced Design. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 4882–4892. https://proceedings.mlr.press/v139/ji21c.html
[8] Quan Xiao and Tianyi Chen. 2025. Unlocking Global Optimality in Bilevel Optimization: A Pilot Study. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=2xvisNIfdw
[9] Liyuan Zheng, Tanner Fiez, Zane Alumbaugh, Benjamin Chasnov, and Lillian J Ratliff. 2022. Stackelberg actor-critic: Game-theoretic reinforcement learning algorithms. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 9217–9224.