Specification-Guided Reinforcement Learning

Tanmay Ambadkar

Department of Computer Science and Engineering
The Pennsylvania State University
201 Old Main, University Park, PA, USA 16802
tsa5252@psu.edu

Abstract

While Reinforcement Learning (RL) has demonstrated remarkable success in solving complex sequential decision-making problems, its application in real-world, safety-critical systems is hindered by its reliance on carefully engineered reward functions. Designing effective rewards is notoriously challenging and can lead to unintended or unsafe behaviors, a phenomenon known as *reward hacking*. Specification-guided RL has emerged as a principled alternative, leveraging formal methods to directly encode high-level objectives, safety requirements, and behavioral constraints. However, the practical utility of this approach is often limited by coarse or under-specified logical formulas and the computational challenge of enforcing safety at scale. This thesis addresses these limitations by developing a unified framework for the automated refinement, scalable enforcement, and flexible adaptation of formal specifications in RL.

Introduction

Reinforcement Learning (RL) has emerged as a powerful framework for training agents to solve complex sequential decision-making problems in domains ranging from robotics to game playing. However, the practical impact of RL is often limited by its reliance on a carefully engineered scalar reward function, which serves as the principal mechanism for communicating task intent to the agent. Designing effective reward functions is notoriously challenging. Minor misalignments between the designer's intent and the formal specification can lead to undesirable or unsafe behaviors, a phenomenon widely recognized as *reward hacking* or *specification gaming*.

To address this challenge, specification-guided RL leverages formal languages, such as Linear Temporal Logic (LTL), to directly encode high-level objectives, safety requirements, and behavioral constraints. While this approach improves expressiveness, its effectiveness is often hampered by three critical challenges: coarse or under-specified logical formulas that provide insufficient learning guidance, the computational intractability of enforcing safety constraints in high-dimensional systems, and the difficulty of learning policies that can adapt to multiple, conflicting objectives. This thesis presents a unified framework to address these challenges, developing novel algorithms for the automated refinement of logical specifications, scalable enforcement of safety guarantees, and flexible adaptation to multi-objective preferences.

Automating The Refinement Of Reinforcement Learning Specifications

A core challenge in specification-guided reinforcement learning is that high-level specifications, while expressive, are often too coarse to effectively guide policy learning. This coarseness provides insufficient guidance for the agent, leading to poor sample efficiency and learning failures. Although logically correct, such specifications can miss environmental complexities, such as trap states or misaligned predicates, rendering the task intractable for the agent. Furthermore, the manual process of refining these specifications is labor-intensive and domain-specific, which undermines the promise of specification-guided methods to reduce engineering effort.

To address these limitations, I introduce **AUTOSPEC**, a framework for the automated refinement of logical specifications. When an agent fails to learn a satisfactory policy, AUTOSPEC identifies the failing subgoals in the specification's abstract graph and applies one of four sound refinement procedures: refining predicates (**SeqRefine**), adding waypoints (**AddRefine**), partitioning source regions (**PastRefine**), or finding alternative paths (**OrRefine**). This refinement process is exploration-guided, analyzing sampled trajectories from the failing policy to diagnose the cause of failure. The procedures then use this data to compute sound, geometric refinements. This process improves learnability while guaranteeing that the refined specification remains true to the user's original intent. The framework's value is confirmed through its theoretical soundness guarantees and empirical validation on large specification graphs.

Scalable Enforcement of Safety Specifications

While formal specifications can define *what* is safe, reliably *enforcing* these constraints during RL remains a major challenge (Bastani 2021). Existing methods present a difficult trade-off: symbolic techniques offer strong formal guarantees but fail to scale to the high-dimensional systems where deep RL excels, while scalable cost-based methods permit numerous violations during training. My work

resolves this trade-off with two complementary shielding frameworks that provide strong safety assurances in complex environments.

My first approach, **SPARKD**, enables scalable safety analysis by combining a learned, globally linear dynamics model with formal verification techniques. Using Koopman Operator theory, **SPARKD** learns a "lifted" representation where complex, nonlinear dynamics become linear, making them amenable to formal analysis. This structure allows **SPARKD** to use weakest precondition calculus to efficiently compute safe action sets, providing a shield that scales to environments where prior symbolic methods failed. This 'lift-and-linearize' approach avoids the curse of dimensionality inherent in traditional symbolic methods, which rely on explicitly partitioning the state space. At runtime, the formal analysis is compiled into a convex optimization, allowing the shield to find the closest safe action by solving a small and efficient Quadratic Program (QP).

Building on this, **RAMPS** introduces a novel multi-step robust Control Barrier Function (CBF) formulation for providing strong, real-time safety guarantees. A key insight of this framework is that the CBF-based shield can operate with any learned linear dynamics model, from a simple regression to a complex Koopman operator. By explicitly accounting for model error and control delays through the robust multi-step CBF formulation over an adaptive horizon, **RAMPS** provides strong, minimally invasive shielding that significantly reduces safety violations. This robustness is achieved by formally incorporating a data-driven error bound (ϵ) into the CBF, creating a 'robust tightening term' that guarantees safety even when the learned linear model is imperfect. The multi-step, adaptive-horizon design is critical for handling real-world systems with high relative-degree constraints (i.e., control delays), where myopic, one-step CBFs would fail to prevent inevitable future violations.

My research now aims to bridge the gap between shielding methods and cost-based safe RL (Wachi and Sui 2020). I use the formal safety analysis techniques from **SPARKD** and **RAMPS** to automate the generation of robust and continuous cost specifications. This provides a more informative learning signal to guide cost-based algorithms, moving beyond simple binary penalties for violations.

Flexible Adaptation to Preference Specifications

Real-world tasks often require agents to balance multiple, conflicting objectives according to user preferences. While training a single, preference-conditioned policy is a flexible and efficient approach for Multi-Objective Reinforcement Learning (MORL) (Hayes et al. 2021), existing methods face significant challenges. They often suffer from **destructive gradient interference**, where conflicting objectives destabilize policy updates, and **representational mode collapse**, where the policy fails to produce diverse behaviors for different preferences. Furthermore, most approaches are restricted to convex preference weights, which limits their expressiveness and the portion of the Pareto front they can discover.

My third contribution, **D^{3}PO** (Decomposed, Diversity-Driven Policy Optimization), is a novel algorithm that trains a single, preference-conditioned policy to directly address these issues. **D^{3}PO** introduces two key innovations. First, a **decomposed optimization process** with late-stage weighting mitigates gradient interference by preserving raw, per-objective advantages before applying preferences. Second, a **scaled diversity regularizer** explicitly encourages the policy to produce distinct behaviors for distinct preferences, formally preventing mode collapse. The framework is supported by strong theoretical guarantees: we formally prove that our **Late-Stage Weighting (LSW)** is provably more robust against signal distortion than naive approaches, and that our **Scaled Diversity Regularizer** guarantees the policy *cannot* exhibit mode collapse. We also prove that the entire algorithm converges to a stationary point, matching the state-of-the-art guarantee for standard PPO.

D^{3}PO demonstrates state-of-the-art performance when compared to more complex multi-policy architectures. This architecture is the first to support non-convex and negative preference weights, enabling the discovery of more comprehensive Pareto fronts. My future work will extend **D^{3}PO** in two key directions. First, I will fully explore its unique support for negative preferences, an unexplored area in preference-conditioned RL that allows for the direct penalization of objectives. Second, I will extend **D^{3}PO**'s decomposed architecture to support non-linear preference functions, integrating its stability and diversity benefits into another major MORL paradigm.

Conclusion

Collectively, my thesis contributions form a unified framework that addresses the full lifecycle of using formal specifications in reinforcement learning: from authoring and refinement to enforcement and adaptation. By automating the repair of coarse specifications, enabling the scalable enforcement of safety guarantees, and providing flexible adaptation to complex preferences, my work aims to fundamentally advance the reliability and usability of RL in complex, safety-critical environments. The overarching goal is to lower the barrier to applying RL in domains where alignment with high-level objectives is paramount. By reducing the dependence on hand-engineered reward and cost functions, this research helps move reinforcement learning from academic demonstrations to practical, trustworthy tools for autonomous decision-making in the real world.

References

Bastani, O. 2021. Safe reinforcement learning with nonlinear dynamics via model predictive shielding. In *2021 American control conference (ACC)*, 3488–3494. IEEE.

Hayes, C. F.; Rădulescu, R.; Bargiacchi, E.; Källström, J.; Macfarlane, M.; Reymond, M.; Verstraeten, T.; Zintgraf, L. M.; Dazeley, R.; Heintz, F.; et al. 2021. A practical guide to multi-objective reinforcement learning and planning. *arXiv preprint arXiv:2103.09568*.

Wachi, A.; and Sui, Y. 2020. Safe reinforcement learning in constrained markov decision processes. In *International Conference on Machine Learning*, 9797–9806. PMLR.