Policy Refinement with Human Feedback for Safe Reinforcement Learning

Ali Baheri

Rochester Institute of Technology
akbeme@rit.edu

Abstract

In this position paper, we discuss policy refinement in reinforcement learning (RL), focusing on safety-critical applications. We propose an integrated approach that combines Bayesian optimization, inverse RL, human feedback, and natural language processing to address challenges in policy refinement. We also examine the limitations of these methods and provide an outlook for the future of policy refinement in RL. Our aim is to contribute to the ongoing conversation and foster collaboration in this crucial area, driving the development of safe and responsible RL policies for real-world, safety-critical applications.

Introduction

RL has emerged as a powerful paradigm in artificial intelligence, empowering autonomous agents to learn optimal strategies for complex tasks through trial-and-error (Sutton and Barto 2018). The success of RL in diverse domains, including robotics, autonomous driving, aviation, healthcare, and gaming, underscores the importance of not only efficient but also safe policies, especially in safety-critical applications (Kober, Bagnell, and Peters 2013; Kiran et al. 2021; Razzaghi et al. 2022). In such contexts, policy refinement becomes crucial, enabling RL agents to *modify* their policies to satisfy safety constraints without compromising performance. As RL policies are increasingly deployed in safety-critical settings, guaranteeing safe and responsible behavior in unknown and uncertain environments is essential (Tambon et al. 2022). Effective policy refinement allows agents to adapt to changing conditions and align with human values, paving the way for widespread adoption in applications involving human lives and property. In this position paper, we will discuss the current landscape of policy refinement in RL, delving into a novel solution that integrates Bayesian optimization, inverse RL, and the incorporation of human feedback and natural language processing. We will also examine the limitations of these methods, highlighting their computational complexity, scalability challenges, and dependence on accurate environmental models. Moreover, we will provide an outlook for the future of policy refinement in RL, discussing anticipated advancements, integration within the broader RL research community, growing awareness and adoption in industry applications, and the ethical considerations and societal implications of refining RL policies. By offering a comprehensive overview of policy refinement in RL, we aim to contribute to the ongoing conversation and foster collaboration in this crucial area, driving the development of safe and responsible RL policies for real-world, safety-critical applications.

Background

Policy refinement is a crucial aspect of RL research, especially in safety-critical applications. The primary objective of policy refinement is to iteratively enhance an agent's policy, ensuring safe and optimal behavior while complying with environmental constraints and task requirements. This process generally involves examining the agent's current policy, pinpointing suboptimal or unsafe actions, and updating the policy to rectify these issues. In the context of safety-critical applications, safe RL has gained prominence as a vital research area (Garcıa and Fernández 2015; Baheri et al. 2020; Isele, Nakhaei, and Fujimura 2018; Baheri 2022). It concentrates on developing algorithms and techniques that guarantee RL agent safety during the learning process. Two notable techniques, counterexample-guided abstraction refinement (CEGAR) and counterexample-guided inductive synthesis (CEGIS), have emerged as promising approaches to policy refinement in RL, focusing on refining policies using counterexamples to ensure safety and optimality.

CEGAR, initially developed for the formal verification of finite-state systems, iteratively refines an abstract system model based on counterexamples discovered during verification (Clarke et al. 2000). Within the context of RL, CEGAR can be applied to policy refinement by generating an abstract representation of the agent's policy and iteratively refining this abstraction using counterexamples found during policy analysis (Jin et al. 2022). Similarly, CEGIS is an approach focused on synthesizing a correct-by-construction program or policy that satisfies a given specification (Solar-Lezama et al. 2006). In CEGIS, the process commences with an initial candidate policy or program, which is subsequently refined based on counterexamples encountered during the verification phase. CEGIS has been successfully implemented in various domains, such as program synthesis (Solar-Lezama 2008; Alur et al. 2013) and controller synthesis (Henzinger, Jhala, and Majumdar 2003; Ravanbakhsh and Sankaranarayanan 2016). By leveraging the strengths of

CEGAR, CEGIS, and safe RL strategies, policy refinement in RL can be rendered more efficient, ultimately resulting in safer and more reliable learning-enabled systems for safety-critical applications. These approaches demonstrate the importance of counterexample-based techniques in improving the safety and optimality of RL policies and highlight the potential for further advancements in the field of policy refinement.

Proposed Solution

Formal methods offer a rigorous framework for verifying and validating the safety and correctness of RL policies. However, formal methods frequently face computational complexity and scalability challenges, especially in high-dimensional and continuous state spaces. To address these challenges, our proposed solution integrates the strengths of Bayesian optimization (BO) (Frazier 2018), inverse RL (Ng, Russell et al. 2000), human feedback, and natural language processing (NLP) (Chowdhary and Chowdhary 2020) in the policy refinement process for safety-critical applications. In essence, our proposed approach employs BO and inverse RL to identify and eliminate failure trajectories from the policy while preserving its performance. BO efficiently searches for failure trajectories in the state-action space, while IRL modifies the reward function to exclude these trajectories. This integrated approach aims to deliver a more efficient policy refinement process, providing significant improvements over existing techniques. The key components of our approach to RL policy refinement include:

Leveraging Bayesian Optimization: Integrating BO into the policy refinement process allows for the identification of promising regions for policy improvement while accounting for uncertainties. This strategy enables a more efficient search and helps to overcome the computational complexity and scalability issues often encountered in formal methods.

Inverse RL for Reward Function Estimation: Incorporating inverse RL into the policy refinement process facilitates the estimation of the underlying reward function that guides expert actions. This information can then be used to improve the policy, allowing the refined policy to better align with expert preferences and safety considerations, leading to more responsible and acceptable behavior in safety-critical applications.

Incorporating Human Feedback: Integrating human feedback into the policy refinement process allows the refined policy to better align with human values and safety considerations, resulting in more responsible and acceptable behavior in safety-critical applications. This approach also offers a more intuitive and accessible way for domain experts to participate in the policy refinement process, bridging the gap between domain knowledge and algorithmic refinement.

Utilizing Natural Language Processing: Employing NLP techniques enables parsing and interpreting human feedback, transforming it into actionable information that can be integrated into the policy refinement process. By leveraging NLP, our approach can better understand and incorporate expert guidance, leading to more effective policy improvements and ultimately safer RL policies in safety-critical applications.

Limitations and Potential Strategies

While the proposed approach combining BO, inverse RL, human feedback, and NLP offers several advantages, it also has some limitations. The computational complexity of the combined approach, dependence on the quality of human feedback, and the accuracy of NLP interpretation are among the challenges. To address these issues, we plan investigate more efficient algorithms, develop guidelines for clearer feedback, and utilize advanced NLP models specifically trained for the domain. The performance of the approach might also be sensitive to initial policy and reward function estimates, and it may rely on certain assumptions about the environment's dynamics. To tackle these challenges, we can explore adaptive strategies to escape local optima, investigate methods that can adapt to non-stationary environments, and develop rigorous validation methods such as worst-case analysis or formal verification techniques. By addressing these limitations, we can further enhance the safety of the proposed policy refinement approach in RL.

Outlook and Conclusions

As policy refinement continues to gain traction in the field of RL, several developments can be anticipated. Firstly, advancements in policy refinement techniques are expected, including novel methods and improvements to existing techniques, addressing current limitations and enabling more scalable policy refinement processes. Secondly, the integration of policy refinement into the broader RL research community is anticipated, with increased collaboration and knowledge sharing between researchers in policy refinement and other RL subfields leading to synergistic advancements and the development of more holistic solutions. Thirdly, as the importance of safety in RL applications becomes more widely recognized, industry adoption of policy refinement techniques is expected to grow, ensuring safe and responsible operation of RL systems in safety-critical domains. Lastly, ethical concerns and societal implications will arise as policy refinement techniques become more advanced, necessitating thoughtful discussions and guidelines to ensure that policy refinement aligns with human values and respects societal norms.

In this position paper, we have discussed the importance of policy refinement in RL, highlighting the proposed solutions, their limitations, and potential strategies to address these challenges. We have also provided an outlook for the future of policy refinement, touching upon anticipated advancements, integration within the broader RL research community, growing awareness and adoption in industry applications, and the ethical considerations and societal implications of refining RL policies. We call upon the research community to address the challenges and explore the potential of policy refinement in RL. Continued research in this area is essential for ensuring the safe and responsible deployment of RL policies in safety-critical applications, ultimately leading to more reliable RL systems that can be trusted to act in the best interests of both humans and the environment.

References

Alur, R.; Bodik, R.; Juniwal, G.; Martin, M. M.; Raghothaman, M.; Seshia, S. A.; Singh, R.; Solar-Lezama, A.; Torlak, E.; and Udupa, A. 2013. *Syntax-guided synthesis*. IEEE.

Baheri, A. 2022. Safe reinforcement learning with mixture density network, with application to autonomous driving. *Results in Control and Optimization*, 6: 100095.

Baheri, A.; Nageshrao, S.; Tseng, H. E.; Kolmanovsky, I.; Girard, A.; and Filev, D. 2020. Deep reinforcement learning with enhanced safety for autonomous highway driving. In *2020 IEEE Intelligent Vehicles Symposium (IV)*, 1550–1555. IEEE.

Chowdhary, K.; and Chowdhary, K. 2020. Natural language processing. *Fundamentals of artificial intelligence*, 603–649.

Clarke, E.; Grumberg, O.; Jha, S.; Lu, Y.; and Veith, H. 2000. Counterexample-guided abstraction refinement. In *Computer Aided Verification: 12th International Conference, CAV 2000, Chicago, IL, USA, July 15-19, 2000. Proceedings 12*, 154–169. Springer.

Frazier, P. I. 2018. A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.

Garcıa, J.; and Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1): 1437–1480.

Henzinger, T. A.; Jhala, R.; and Majumdar, R. 2003. Counterexample-guided control. In *Automata, Languages and Programming: 30th International Colloquium, ICALP 2003 Eindhoven, The Netherlands, June 30–July 4, 2003 Proceedings*, 886–902. Springer.

Isele, D.; Nakhaei, A.; and Fujimura, K. 2018. Safe reinforcement learning on autonomous vehicles. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 1–6. IEEE.

Jin, P.; Tian, J.; Zhi, D.; Wen, X.; and Zhang, M. 2022. Trainify: a CEGAR-driven training and verification framework for safe deep reinforcement learning. In *Computer Aided Verification: 34th International Conference, CAV 2022, Haifa, Israel, August 7–10, 2022, Proceedings, Part I*, 193–218. Springer.

Kiran, B. R.; Sobh, I.; Talpaert, V.; Mannion, P.; Al Sallab, A. A.; Yogamani, S.; and Pérez, P. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6): 4909–4926.

Kober, J.; Bagnell, J. A.; and Peters, J. 2013. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11): 1238–1274.

Ng, A. Y.; Russell, S.; et al. 2000. Algorithms for inverse reinforcement learning. In *ICML*, volume 1, 2.

Ravanbakhsh, H.; and Sankaranarayanan, S. 2016. Robust controller synthesis of switched systems using counterexample guided framework. In *Proceedings of the 13th International Conference on Embedded Software*, 1–10.

Razzaghi, P.; Tabrizian, A.; Guo, W.; Chen, S.; Taye, A.; Thompson, E.; Bregeon, A.; Baheri, A.; and Wei, P. 2022. A Survey on Reinforcement Learning in Aviation Applications. *arXiv preprint arXiv:2211.02147*.

Solar-Lezama, A. 2008. *Program synthesis by sketching*. University of California, Berkeley.

Solar-Lezama, A.; Tancau, L.; Bodík, R.; Seshia, S. A.; and Saraswat, V. A. 2006. Combinatorial sketching for finite programs. In *Proceedings of the 12th International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS 2006, San Jose, CA, USA, October 21-25, 2006*, 404–415.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Tambon, F.; Laberge, G.; An, L.; Nikanjam, A.; Mindom, P. S. N.; Pequignot, Y.; Khomh, F.; Antoniol, G.; Merlo, E.; and Laviolette, F. 2022. How to certify machine learning based safety-critical systems? A systematic literature review. *Automated Software Engineering*, 29(2): 38.