Safe Reinforcement Learning for Trustworthy AI: Theory, Algorithms, and Applications

Honghao Wei

[1] School of EECS, Washington State University
honghao.wei@wsu.edu

Abstract

Safe reinforcement learning (RL) has emerged as a key paradigm for deploying AI in high stakes domains such as autonomous driving, robotics, healthcare, and recommender systems. By embedding constraints into the learning process, safe RL enables agents to optimize performance while satisfying critical requirements, including collision avoidance, resource limits, and system reliability. Such guarantees are indispensable for real world AI, where failures can cause physical harm, economic loss, or loss of trust. At the same time, demand for trustworthy AI continues to grow as machine learning is increasingly deployed in human centered applications. This makes it essential to design RL algorithms that are not only efficient but also reliable, robust, and aligned with societal needs.

Introduction

I will survey recent progress on the design of safe and efficient RL algorithms with theoretical guarantees, focusing on both online and offline settings. I will begin by outlining the fundamental differences between standard RL and safe RL, highlighting unique challenges such as the absence of an optimality Bellman equation, which necessitates stochastic policies, and the impracticality of assuming full dataset coverage in offline settings. These structural gaps underscore the need for new algorithms that provide both efficiency and rigorous safety guarantees.

For advancing **online safe RL**, we developed the first model-free, simulator-free algorithm with sublinear regret and zero constraint violations. The algorithm, *Triple-Q* (Wei, Liu, and Ying 2022), maintains Q-functions for reward and cost along with a virtual queue that tracks violations, combining optimism for rewards with pessimism for constraints. Triple-Q achieves efficient, violation-free learning and has been extended to non-stationary environments and to best policy identification with optimal-order guarantees. In parallel, we also established the first order-optimal model-based solution under adversarial settings, providing new insights into safety under hostile conditions. These results show that efficiency and strict safety can be achieved simultaneously, a key step toward real-world deployment.

For **offline safe RL**, where learning is based on pre-collected data. We introduced a primal–dual LP-based algorithm that improves data efficiency while relaxing unrealistic full-coverage assumptions. We also developed *Weighted Safe Actor-Critic (WSAC)* (Wei et al. 2024), which guarantees safe policy improvement under function approximation even with partial coverage. WSAC outperforms reference policies while maintaining safety, achieves the optimal statistical rate of $1/\sqrt{N}$, and is robust across a wide range of hyperparameters. Together, these results provide a pathway to safe decision-making in environments where new interactions are costly or risky, such as healthcare and industrial systems.

This talk also investigates **emerging directions** that connect safe RL theory to practice. These include *Rectified Policy Optimization (RePO)* for safe large language models (Peng et al. 2025), switching-policy mechanisms for handling unknown thresholds, and warm-start strategies that accelerate safe RL while reducing unsafe exploration. We also apply safe RL to physical robotic systems, where hardware reliability and safety constraints are critical. These directions not only broaden the applicability of safe RL, but also demonstrate how theoretical guarantees can guide the design of algorithms for real-world safety-critical systems.

This talk surveys the algorithmic foundations and recent advances in safe RL, showing how rigorous theory can be translated into scalable, trustworthy AI.

References

Peng, X.; Guo, H.; jiawei zhang; Zou, D.; Shao, Z.; Wei, H.; and Liu, X. 2025. Enhancing Safety in Reinforcement Learning with Human Feedback via Rectified Policy Optimization. In *Advances Neural Information Processing Systems (NeurIPS)*.

Wei, H.; Liu, X.; and Ying, L. 2022. A Provably-Efficient Model-Free Algorithm for Infinite-Horizon Average-Reward Constrained Markov Decision Processes. In *AAAI Conf. Artificial Intelligence*.

Wei, H.; Peng, X.; Ghosh, A.; and Liu, X. 2024. Adversarially Trained Weighted Actor-Critic for Safe Offline Reinforcement Learning. *nips*, 37: 52806–52835.