A Survey of Safe Reinforcement Learning Methods in Robotics

Yuen Xie

College of Computer Science and Electronic Engineering, Hunan University, Changsha, China

Abstract. Reinforcement learning has achieved remarkable results in complex decision-making tasks such as robot control, but its deployment in real-world environments still faces safety challenges. Safe reinforcement learning aims to ensure that the behavior of the agent does not violate key safety constraints while exploring the environment and optimizing long-term rewards. This review paper aims to explore safe reinforcement learning methods and introduces three main types of safe reinforcement learning methods: control theory-based methods, formal method-based methods, and constrained optimization-based methods. After summarizing the advantages and disadvantages of these methods, this paper further analyzes their performance on the mainstream safe reinforcement learning framework Omnisafe. By citing existing comparative experimental results, the performance of different methods in multiple tasks is evaluated. It is found that the constrained optimization method can better improve task performance while ensuring safety due to its strong adaptability and scalability. Finally, this paper summarizes the current research status of safe reinforcement learning and points out the challenges that still exist and future development directions.

1 Introduction

In reinforcement learning (RL), the agent learns how to maximize rewards by interacting with the environment. The goal is usually to optimize the strategy to achieve the maximum cumulative reward [1]. However, reward optimization and safety goals are often not completely consistent. The reward mechanism of reinforcement learning may prompt the agent to adopt certain high-return but potentially dangerous strategies, such as performing high-risk actions or causing drastic changes in the environment, while safety requirements emphasize that the agent must comply with specific safety constraints in the decision-making process, such as physical limitations, avoiding collisions, or controlling variables beyond the tolerance range. Therefore, between reward optimization and safety, the agent needs to find a reasonable balance to ensure that the task can be completed efficiently while avoiding safety risks.

In the field of robotics, the value of safe reinforcement learning is mainly reflected in improving the safety and reliability of the system and making reinforcement learning

xieyuen@hnu.edu.cn

algorithms more realistic and deployable. The exploration process of traditional reinforcement learning may cause robots to perform dangerous operations, while safe reinforcement learning can effectively meet safety requirements while ensuring task performance through methods such as constraint optimization, barrier functions, and formal verification, so that robots can still comply with safety rules when making autonomous decisions in complex environments. In addition, the research on safe reinforcement learning also provides important support for the development of robotics technology, promoting the construction of more intelligent and adaptive systems that can coexist safely with humans and further expand the application boundaries of robots in the real world.

A review of this direction can systematically analyze the conflict between reward and safety, and organize the understanding and solutions to this problem in the existing literature. By summarizing the research results and remaining challenges of existing safe reinforcement learning methods, it can provide ideas for researchers in related fields to promote the direction of further improvement.

At present, in order to ensure the safety of reinforcement learning, many safe reinforcement learning methods have been proposed [2]. The research mainly focuses on how to effectively deal with the trade-off between reward optimization and safety goals. According to different theoretical foundations and technical methods, the existing safe reinforcement learning methods can be roughly divided into three categories: methods based on control theory, methods based on formal methods, and methods based on constrained optimization [3].

Koller proposed a predictive control method that combines LSTM learning model with nominal model to solve the problem of model uncertainty and control constraints in nonlinear systems [4]. Chow et al. proposed a safe reinforcement learning method based on Lyapunov function. Under the CMDP model, the global safety guarantee in the strategy training process is achieved by constructing Lyapunov function [5,6]. Fulton and Platzer proposed a "provably safe" learning method combining formal verification and reinforcement learning. By embedding formal verification and runtime monitoring in the learning process, the learning agent is ensured to be safe when the model is met [7]. Li and Belta proposed a safe reinforcement learning framework that combines temporal logic, control Lyapunov function and control barrier function, which improves the exploration efficiency and safety in complex tasks [8]. The constrained policy optimization (CPO) method proposed by Achiam et al. uses constraint optimization theory. This is the first general policy search method for constrained reinforcement learning. It can approximately satisfy the constraints in each iteration. This method is suitable for high-dimensional control tasks and has verified its ability to achieve effective policy optimization while ensuring safety constraints in simulated robot motion [9]. In the latest research, Gu et al. proposed the PCRPO method [3], which is a soft switching policy optimization method based on gradient operation theory. By introducing the primal-dual update mechanism in the policy optimization process, safety constraints are satisfied in a more stable and efficient way, while improving the convergence and task performance of reinforcement learning.

This paper will summarize the principles, advantages and disadvantages, and application scenarios of three major types of safe reinforcement learning methods to help readers fully understand the latest developments in the field of safe reinforcement learning. Next, this paper will focus on analyzing the performance of these safe reinforcement learning methods on the mainstream framework Omnisafe. Omnisafe is a platform for accelerating safe reinforcement learning research, which provides implementation and evaluation tools for a variety of reinforcement learning algorithms [10]. Through

experimental comparisons on the Omnisafe platform, this paper will evaluate the performance of different methods in multiple tasks and compare how they balance task performance and efficiency while ensuring safety. Finally, this paper will summarize the current research status in the field of safe reinforcement learning, discuss the challenges and limitations of existing methods, and propose future research directions.

2 Overview of Mainstream Methods

In recent years, research in the field of safe reinforcement learning (Safe RL) has mainly focused on ensuring that the intelligent agent does not violate safety constraints while learning the optimal strategy. Mainstream methods can be roughly divided into three categories: safe RL based on control theory, safe RL based on formal methods, and safe RL based on constrained optimization:

Safe reinforcement learning methods based on control theory ensure safety while learning the optimal strategy by utilizing principles such as Lyapunov functions and model predictive control. Safe reinforcement learning based on formal methods strictly ensures that the learned strategy meets safety specifications through formal verification techniques such as temporal logic, but such methods require external knowledge to define safe state and action spaces. Control theory methods use techniques such as model predictive control and Lyapunov functions to ensure safety during the learning process, but usually require accurate system models, which limits their universality. Formal methods ensure that the reinforcement learning process meets predetermined safety specifications by using tools such as temporal logic, but their deployment requires external knowledge to define the safety space, which may be limited in practical applications. Safe reinforcement learning methods based on constrained optimization introduce constraints to optimize strategies while meeting safety standards. They are currently a relatively mature and widely used type of method. There are two main branches of safe reinforcement learning methods based on constrained optimization: primal-dual methods and primal optimization methods [11]. Primitive dual methods such as CPO [9] and PPO-Lagrangian [12, 13] ensure safety during the optimization process by introducing dual variables and line search mechanisms, while primitive methods such as CRPO [14] directly optimize in the primitive strategy space without dual variables, making it easier to implement and less dependent on initialization. Table 1 shows the representative methods, characteristics, advantages and disadvantages of the three mainstream methods.

Table 1 Representative methods, characteristics, advantages and disadvantages of the three mainstream methods

Method category	Representative methods	Main features	Advantages	Disadvantages
Control-Based Safe RL	Lyapunov-based RL[5][6]	Using Lyapunov function to constrain the exploration process and ensure the safety of the learning process.	In theory, it can provide strict safety guarantees.	It requires a system model and is difficult to generalize to general RL problems.

Formal Methods-Based Safe RL	Temporal Logic Verification[7]	Using temporal logic verification to ensure that the strategy meets safety constraints.	Strictly ensures safety and has a solid theoretical foundation.	External knowledge is required to define the safety state and action space, which is difficult to deploy in real applications.
Constrained Optimization-Based Safe RL				
→ Primal-Dual Methods	CPO[9]	Using TRPO[15] for constrained optimization, use line search to ensure constraint satisfaction.	Accurately controls constraint satisfaction	The amount of calculation involved in the process is large and the optimization process is complicated.
	PPO-Lagrangian[12][13]	Using PPO combined with Lagrangian multipliers to dynamically adjust weights.	Training stability is good and it is suitable for large-scale problems.	It requires fine-tuning of hyperparameters (Lagrangian multipliers) and is prone to suboptimal solutions.
	PCPO[16]	Combines CPO and PPO to optimize the dual variable update strategy.	It has both the security of CPO and the stability of PPO.	It still requires dual variable adjustment, which is complex to optimize.
	CUP[17]	Using an adaptive dual update strategy to improve training stability.	It has fewer hyperparameter adjustments and is more efficient.	It still relies on dual variables and cannot completely eliminate gradient conflicts.
→Primal Methods CRPO[14]	CRPO[14]	Focuses on optimizing rewards and only turns to safety optimization when safety constraints are seriously violated.	High training efficiency and simple implementation.	It is easy to oscillate near the safety constraint boundary, affecting the optimization stability.

	PCRPO[3]	Using Projected Gradient Descent and introduce slack variables to handle gradient conflicts.	Successfully avoid CRPO oscillation, making training smoother, and eliminate the need for dual variables.	The slack variable parameters need to be carefully designed.

3 Mainstream Frameworks and Evaluation Standards

In the research system of safe reinforcement learning, the construction of the infrastructure framework and the setting of evaluation standards together constitute the dual-wheel drive of technological evolution. As a representative platform in this field, OmniSafe Benchmark realizes the systematic integration of algorithm development and verification processes through modular architecture design. This framework not only provides a standardized safe reinforcement learning algorithm implementation library, but also innovatively develops an extensible constraint injection interface, allowing researchers to quickly configure multi-dimensional safety parameters such as joint torque limits and motion speed thresholds, laying a solid foundation for horizontal comparison of algorithm performance [10].

In terms of test environment construction, the research results cited in this article use two typical control tasks of the OmniSafe platform, Hopper and Ant, as benchmark verification scenarios. The former focuses on the dynamic balance ability of the algorithm under high-speed motion by modeling the dynamics of single-legged jumping; the latter verifies the safe motion strategy of complex mechanical systems in confined spaces by relying on the multi-degree-of-freedom control requirements of quadruped robots. These environments simulate key challenges such as inertial delay and sudden changes in ground friction in real scenarios through precise physical engine simulation, providing a high-fidelity experimental field for reliability testing of safety constraint mechanisms.

The scientific design of the evaluation system directly determines the orientation of algorithm optimization. OmniSafe adopts comprehensive evaluation indicators: the first dimension is safety compliance, which quantifies the degree of compliance of the algorithm with the preset constraint boundaries by real-time monitoring of key parameters such as joint angular velocity and end-effector displacement; the second is reward acquisition efficiency, which compares the cumulative reward convergence values of different algorithms in the same training cycle under the premise of meeting the safety threshold; the third dimension focuses on the training dynamics characteristics, and analyzes the learning efficiency of the algorithm through the policy entropy change curve and the loss function convergence rate; finally, the safety cost function is introduced to dynamically weighted punish constraint violations. This indicator is of special importance in high-risk scenarios such as autonomous driving. This multi-scale evaluation mechanism can not only fully reflect the performance characteristics of the algorithm, but also build a quantifiable technical bridge for safe reinforcement learning from theoretical verification to engineering application.

4 The Current Level Achieved in This Direction

This paper cites the results of researchers at PCRPO comparing several representative primal-dual optimization-based secure reinforcement learning methods on the popular benchmark Omnisafe. Figure 1 shows the performance of each method.

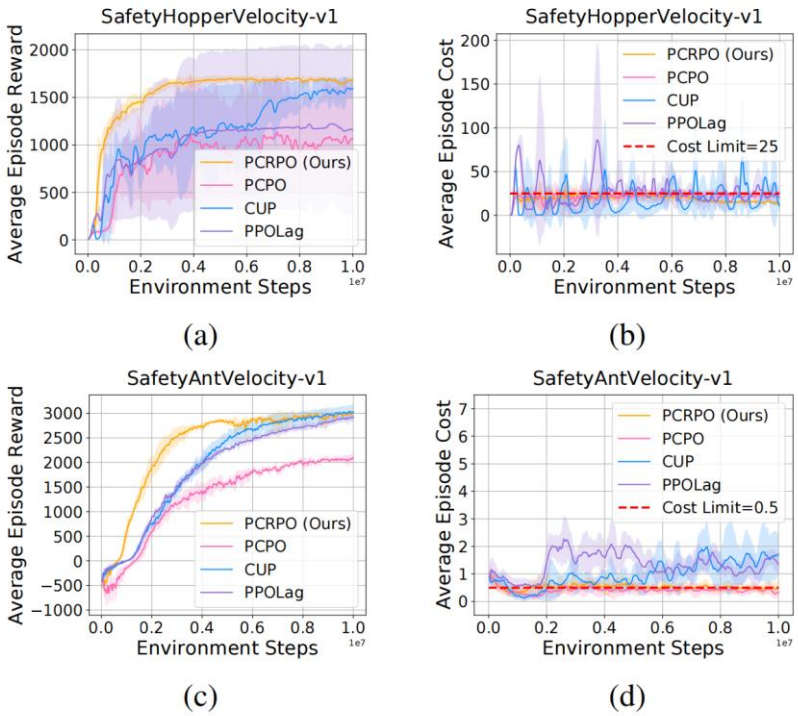

Figure 1 Comparison of PCRPO, PCPO, CUP, PPOLag baselines on SafetyHopperVelocity-v1 and SafetyAntVelocity-v1 tasks [3]

In the field of safety reinforcement learning, current mainstream methods have demonstrated significant safety assurance capabilities. In the OmniSafe framework, safety control is usually achieved by setting a threshold for the speed of the agent. When the speed exceeds the preset range, the system will trigger a penalty cost of 1 unit per step. In multiple comparative experiments, algorithms such as PCRPO, PCPO, CUP and PPOLag all demonstrated reliable safety performance, among which PCRPO performed particularly well. This algorithm not only ensured complete safety in the SafetyAntVelocity-v1 task, but also achieved comprehensive performance indicators that were superior to baseline methods such as CUP and PPOLag at the same safety level, demonstrating unique algorithmic advantages.

In terms of task reward acquisition capabilities, PCRPO also demonstrated excellent balancing capabilities. In the representative SafetyHopperVelocity-v1 test environment, this method significantly surpassed the comparative algorithms such as PCPO, CUP and PPOLag in terms of cumulative reward value under the premise of strictly following safety constraints. It is worth noting that some traditional safety reinforcement learning methods such as CUP and PPOLag often have significant reward decay when pursuing safety compliance, while PCRPO successfully solves this typical contradiction through an innovative strategy optimization mechanism.

From the analysis of algorithm convergence characteristics, PCRPO shows a significant learning efficiency advantage. Experimental data show that compared with PCPO and other benchmark methods, the algorithm can reach a stable state in a shorter training cycle, and the performance index corresponding to the final convergence position is higher. This shows that PCRPO can not only maintain a balance between reward and safety during training, but also accelerate the learning process, thereby improving training efficiency.

Overall, current safety reinforcement learning methods, especially in the OmniSafe benchmark, have reached a high level. Mainstream methods (such as PCRPO, PCPO, CUP, and PPOLag) can balance rewards and safety in multiple environments, ensuring that the agent does not violate safety constraints during the learning process. However, there are still some challenges, especially how to better balance the conflict between safety and reward, and how to improve the convergence speed and overall performance of the algorithm in more complex tasks. As a newly proposed method, PCRPO demonstrates its superiority over traditional methods, especially in terms of the trade-off between safety and reward.

5 Challenges and Prospects

In the field of secure reinforcement learning, the inherent conflict between reward mechanism and safety constraints is a key challenge that restricts algorithm performance. Although existing constrained optimization methods (such as the policy gradient constraint framework) can alleviate the opposition between the two through theoretical modeling, this contradiction still leads to gradient direction conflicts and policy oscillation in complex scenarios such as high-dimensional state space and nonlinear dynamic systems. In response to this challenge, soft switching gradient manipulation technology based on dynamic balance mechanism shows unique advantages. Advanced algorithms represented by PCRPO establish a reward-safety dual-objective adaptive adjustment module, calculate the gradient projection angle in real time during the policy update process, and dynamically adjust the optimization weight according to environmental feedback. This flexible processing mechanism not only reduces the decision-making risk in critical states, but also significantly improves the learning stability through the collaborative optimization of gradient space.

The single-dimensional limitations of the existing evaluation system are also worthy of attention. Current research generally uses discrete safety thresholds as core evaluation indicators. This simplified processing is difficult to fully reflect the multi-dimensional safety requirements in real scenarios. For example, in the field of robot control, in addition to the basic joint speed constraints, composite indicators such as actuator temperature, mechanical structure stress, and environmental interaction safety often constitute a complete safety assessment system. To this end, it is recommended to build a multi-level safety assessment framework: retain traditional constraint indicators at the basic layer, introduce dimensions such as equipment health monitoring and dynamic risk prediction at the application layer, and integrate comprehensive evaluation parameters such as environmental disturbance tolerance and emergency recovery capabilities at the system layer. This hierarchical and progressive evaluation system can not only maintain the rigor of theoretical research, but also enhance the adaptability to actual engineering needs.

In addition, the lack of authenticity of the algorithm verification environment has become an important bottleneck restricting the implementation of technology. Although mainstream test platforms (such as OmniSafe) have built standardized safety constraint scenarios, their simplified dynamic models and discrete state spaces are difficult to accurately simulate the continuity and uncertainty characteristics of the real world. In order to break through this limitation, it is recommended to integrate high-precision simulation technology of physical engines into the construction of data sets, and create a training environment with spatiotemporal continuity characteristics by introducing real traffic flow data, complex terrain modeling, and multi-agent dynamic interaction.

6 Conclusion

This paper reviews the research progress of safe reinforcement learning in the field of robotics, focusing on three main methods: methods based on control theory, methods based on formal methods, and methods based on constrained optimization. Subsequently, this paper analyzes the experimental performance of these methods under the Omnisafe framework and evaluates their ability to balance safety and performance in different tasks. Through this review, we find that the application of safe reinforcement learning in the field of robotics still faces many challenges, such as how to improve learning efficiency while ensuring safety, how to ensure the stability of algorithms in high-dimensional complex environments, and how to design more general methods to meet the needs of different tasks. Future research can start from combining deep learning with safe reinforcement learning, improving the interpretability of algorithms, and exploring more efficient constrained optimization methods to promote the implementation of safe reinforcement learning in practical applications. It is hoped that this study can provide a systematic reference for relevant researchers and provide theoretical and practical support for the development of safe reinforcement learning in robots.

References

1. Kaelbling, L. P., Littman, M. L., & Moore, A. W.: 'Reinforcement learning: A survey.' Journal of Artificial Intelligence Research, 1996, 4, 237-285
2. Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., & Knoll, A.: 'A review of safe reinforcement learning: Methods, theory and applications.' arXiv preprint arXiv:2205.10330. 2022
3. Gu, S., Sel, B., Ding, Y., Wang, L., Lin, Q., Jin, M., & Knoll, A.: 'Balance reward and safety optimization for safe reinforcement learning: A perspective of gradient manipulation.' In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 19, pp. 21099-21106). 2024
4. Koller, T., Berkenkamp, F., Turchetta, M., & Krause, A.: 'Learning-based model predictive control for safe exploration.' In 2018 IEEE conference on decision and control (CDC) (pp. 6059-6066). IEEE. 2018
5. Chow, Y., Nachum, O., Duenez-Guzman, E., & Ghavamzadeh, M.: 'A lyapunov-based approach to safe reinforcement learning.' Advances in neural information processing systems, 31. 2018
6. Chow, Y., Nachum, O., Faust, A., Duenez-Guzman, E., & Ghavamzadeh, M.: 'Lyapunov-based safe policy optimization for continuous control.' arXiv preprint arXiv:1901.10031. 2019.
7. Li, X., & Belta, C.: 'Temporal logic guided safe reinforcement learning using control barrier functions.' arXiv preprint arXiv:1903.09885. 2019
8. Fulton, N., & Platzer, A.: 'Safe reinforcement learning via formal methods: Toward safe control through proof and learning.' In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1). 2018
9. Achiam, J., Held, D., Tamar, A., & Abbeel, P.: 'Constrained policy optimization.' In International conference on machine learning (pp. 22-31). PMLR. 2017
10. Ji, J., Zhou, J., Zhang, B., Dai, J., Pan, X., Sun, R., ... & Yang, Y.: 'Omnisafe: An infrastructure for accelerating safe reinforcement learning research.' Journal of Machine Learning Research, 25(285), 1-6. 2024

11. Boyd, S. P. and L. Vandenberghe.: 'Convex optimization.' Cambridge university press. 2004
12. Zhou, Z., Huang, M., Pan, F., He, J., Ao, X., Tu, D., & He, Q.: 'Gradient-adaptive pareto optimization for constrained reinforcement learning.' In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 9, pp. 11443-11451). 2023
13. Calian, D.A. et al.: 'Balancing Constraints and Rewards with Meta-Gradient D4PG.' In ICLR. 2020
14. Xu, T., Liang, Y., & Lan, G.: 'Crpo: A new approach for safe reinforcement learning with convergence guarantee.' In International Conference on Machine Learning (pp. 11480-11491). PMLR. 2021
15. Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P.: 'Trust region policy optimization.' In International conference on machine learning, 1889 – 1897. PMLR. 2015
16. Yang, T. Y., Rosca, J., Narasimhan, K., & Ramadge, P. J.: 'Projection-based constrained policy optimization.' arXiv preprint arXiv:2010.03152. 2020
17. Yang, L., Ji, J., Dai, J., Zhang, L., Zhou, B., Li, P., & Pan, G.: 'Constrained update projection approach to safe policy optimization.' Advances in Neural Information Processing Systems, 35, 9111-9124. 2022