CAPO: A Unified Policy Gradient Approach for Reward and Cost Optimization in Safe Reinforcement Learning (Student Abstract)

Liu Xiaotao[1], **Prashant Mohit**[1], **Arvind Easwaran**[1]
[1]Nanyang Technological University, College of Computing and Data Science
XLIU057@e.ntu.edu.sg, MOHIT010@e.ntu.edu.sg, arvinde@ntu.edu.sg

Abstract

In safe reinforcement learning (SRL), there exists an inherent conflict between maximizing reward and minimizing cost. We propose a novel approach that effectively resolve the conflict between maximizing reward and minimizing cost in joint optimization.When the cost exceeds the threshold, we perform cost-reducing updates. Otherwise, we compute policy gradients that maximize expected rewards, while using second-order Taylor approximation to evaluate whether these reward-maximizing gradients would violate the cost constraint. If constraint violation is detected, we adjust the gradient direction to maintain safety compliance; otherwise, we execute standard reward-increasing policy updates. This approach helps ensure that reward-seeking updates do not inadvertently increase costs, thereby reducing the likelihood of constraint violations. Empirical tests show our framework successfully manages reward-cost trade-offs through reward augmentation and cost shaping, improving both performance and safety without switching optimization strategies. Results demonstrate that concurrent treatment of both objectives in one policy gradient update is viable for improving safe reinforcement learning methods.

Introduction

Reinforcement learning (RL) provides a powerful framework for sequential decision-making, enabling agents to learn optimal policies through trial-and-error interactions with environments to maximize cumulative rewards. However, real-world applications often require agents to satisfy safety constraints alongside performance objectives, leading to Safe Reinforcement Learning (SRL). In SRL, agents must optimize rewards while ensuring cost functions remain within acceptable thresholds.

Policy gradient (PG) methods have become the dominant approach in RL, with algorithms like TRPO and PPO achieving state-of-the-art performance. These methods optimize policies by computing gradients of expected returns with respect to policy parameters, providing a flexible framework for both standard RL and safe learning scenarios.

Constraint Rectified Policy Optimization (CRPO) (Xu et al. 2021) represents a key advancement in SRL, alternating between reward maximization and cost minimization phases based on constraint satisfaction. However, CRPO suffers from fundamental limitations: reward maximization and cost minimization often exhibit inherent conflicts rather than orthogonal objectives. This correlation can cause inefficient oscillation between safety and performance optimization, where alternating updates destabilize previously satisfied constraints.

We propose CAPO, a novel approach that uses second-order Taylor approximation of the cost function to detect and resolve conflicts between reward maximization and cost minimization within single policy updates, eliminating the need for optimization strategy switching while maintaining safety guarantees.

Methods

We propose CAPO to address the limitations of CRPO in handling conflicts between reward maximization and cost reduction in the case that the cost constraint is satisfied.

Intuition

CRPO cannot recognize when the reward gradient increases the expected cost. CAPO uses a second-order Taylor approximation to detect this conflict. If

$$g_c^\top g_r + \frac{1}{2} g_r^\top H g_r > 0, \tag{1}$$

where $g_r = \nabla J_r(\pi_\theta)$, $g_c = \nabla J_c(\pi_\theta)$, and H is the Hessian of the cost, then the update direction will undesirably increase the cost.

Conflict Handling

To correct this, we rescale the update direction by

$$g_r' = k g_r, \quad k = -2\frac{g_c^\top g_r}{g_r^\top H g_r}. \tag{2}$$

Substituting g_r' into the left-hand side of Equation (1), its value becomes 0, meaning that the cost will not increase under the second-order approximation.

Features

As shown in Algorithm 1 CAPO is agnostic to the choice of policy gradient method and only modifies the update direction. It avoids inefficient oscillations caused by CRPO while ensuring safe and efficient learning.

Algorithm 1: Constraint-Aware Policy Optimization (CAPO)

Input: Policy parameters θ, cost limit $C_{\max}$, learning rate α, max epochs M
Output: Optimized policy parameters θ

Initialize θ, cost limit $C_{\max}$
for epoch $= 1$ to M **do**
  Collect trajectories, estimate gradients g_r, g_c
  if $J_c(\pi_\theta) > C_{\max}$ **then**
    update $\leftarrow -g_c$
  else if no conflict **then**
    update $\leftarrow g_r$
  else
    $\beta \leftarrow \min\{-2(g_c^\top g_r)/(g_r^\top H g_r), 1\}$
    update $\leftarrow \beta g_r$
  end if
  $\theta \leftarrow \theta + \alpha \cdot$ update
end for

Experiments and Discussion

Environment Settings

We evaluated CAPO in the Safety Gym environment PointGoal1-v0 , which are widely used benchmarks for safe reinforcement learning(Ray, Achiam, and Amodei 2019). In this environment, a dot-like agent that can be moved freely on the 2D-plane needs to reach a series of goals to get rewards and avoid entering the specified unsafe area to avoid cost increase.We compared our approach with current benchmark CPO(Achiam et al. 2017). Our implementation is based on the OmniSafe framework(Ji et al. 2024).

Result

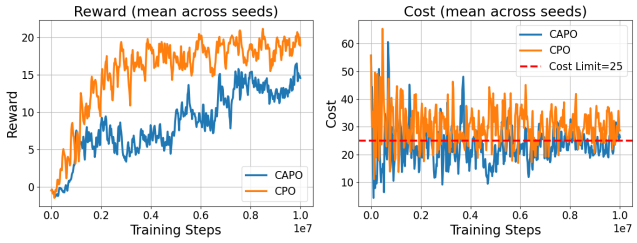

Figure 1: Training performance on PointGoal1-v0. Left: episode reward averaged across seeds. Right: episode cost averaged across seeds, with the red dashed line indicating the cost limit (25). CAPO achieves lower cost while maintaining competitive reward compared to CPO.

Overall, CAPO violated the safety threshold less frequently than CPO, with CPO consistently exceeding the constraint of 25 even after training stabilized. CPO achieved higher reward values compared to CAPO throughout the training process.

Discussion

Our goal was to maximize reward while ensuring cost remained below the threshold. Although CPO achieved higher reward, its persistent constraint violations indicate that it may not satisfy safety requirements in real-world deployment. In contrast, CAPO demonstrates a safer trade-off by prioritizing constraint satisfaction. The lower reward of CAPO arises from the inherent conflict between reducing cost and increasing reward: enforcing the cost constraint restricts the agent's feasible action space, which in turn limits the achievable optimal reward.

The significance of this work lies in advancing gradient-based constrained reinforcement learning methods. When the parameter update direction for maximizing reward conflicts with the direction for minimizing cost, our approach provides a principled mechanism for handling this trade-off. Unlike traditional methods that may blindly follow the reward gradient, CAPO explicitly considers the geometric relationship between reward and cost gradients, enabling more informed decision-making during optimization.

For future work, we aim to improve the conflict handling mechanism to achieve better balance between reward optimization and constraint satisfaction. Specifically, we hope to develop adaptive strategies that do not overly restrict the reward improvement rate in early training stages, while ensuring stricter adherence to cost constraints in the final policy. Additionally, exploring more sophisticated gradient projection techniques and adaptive penalty methods could lead to even safer and more efficient policy optimization algorithms.

Conclusion

We presented CAPO, a novel policy gradient approach for safe reinforcement learning that unifies reward maximization and cost minimization within a single optimization framework. By using second-order Taylor approximation to detect and resolve conflicts between reward and cost gradients, CAPO effectively reduces constraint violations while maintaining competitive reward performance. Experimental results demonstrate that CAPO outperforms CPO in safety compliance, making it a promising approach for practical safe reinforcement learning applications.

Acknowledgements

I would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project.

References

Achiam, J.; Held, D.; Tamar, A.; and Abbeel, P. 2017. Constrained policy optimization. In *International Conference on Machine Learning*, 22–31. PMLR.

Ji, J.; et al. 2024. OmniSafe: An Infrastructure for Accelerating Safe Reinforcement Learning Research. *arXiv preprint arXiv:2305.09304*.

Ray, A.; Achiam, J.; and Amodei, D. 2019. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*.

Xu, M.; Wu, Y.; Zheng, Q.; Shi, Y.; Yang, Z.; and Wang, Z. 2021. CRPO: Conservative Reward Penalization Operator for Safe Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 34, 17283–17294.