Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Constrained Update Projection Approach to Safe Policy Optimization]
Date: [26-09-2024]

Make sure your critique Address the following points:
1. The problem the paper is trying to address
2. Key contributions of the paper
3. Proposed algorithm/framework
4. How the proposed algorithm addressed the described problem
Note: Be concise with your explanations. Unnecessary verbosity will be penalized. Please don't exceed 2 pages.

---

1 The problem the paper is trying to address

The paper addresses the problem of safe policy optimization in reinforcement learning (RL). Specifically, it aims to develop a method where an agent not only maximizes the expected reward but also satisfies safety constraints during the learning process. Traditional RL methods often fail to meet safety constraints, which is undesirable in real-world applications like robotics or autonomous systems.

$$\begin{aligned} \max_{\pi_\theta} \quad & J(\pi_\theta) \\ \text{s.t.} \quad & J_c(\pi_\theta) \leq b \end{aligned}$$

- $J(\pi_\theta)$ is the expected return (reward),
- $J_c(\pi_\theta)$ is the cost function representing safety constraints,
- b is the upper bound for safety constraints.

2 Key contributions of the paper

- Proposed the Constrained Update Projection (CUP) algorithm for safe policy optimization in reinforcement learning.
- Developed surrogate functions with respect to performance bounds, improving empirical performance.
- Unified existing theoretical bounds to provide a better understanding and interpretability of existing algorithms.
- Implemented a non-convex optimization solution via first-order optimizers, avoiding strong convex approximations.
- Validated CUP against multiple safe reinforcement learning baselines, demonstrating its effectiveness in reward maximization and constraint satisfaction.

3 Proposed algorithm/framework

Algorithm 1 Constrained Update Projection (CUP)

Initialize: Policy parameters θ_0, cost parameters ν_0, step-size η

Hyper-parameters: Discount rate γ, trajectory horizon T, regularization coefficients α_k, β_k

for $k = 0, 1, 2, \ldots$ **do**

Collect batch data $\mathcal{D}_k = \{(s_t, a_t, r_{t+1}, c_{t+1})\}$ using current policy π_{θ_k}

Estimate returns and advantage functions:

$$J(\pi_{\theta_k}), J_c(\pi_{\theta_k}), A^{GAE}_{\pi_{\theta_k}}, A^{GAE,C}_{\pi_{\theta_k}}$$

Step 1: Performance Improvement

Update the policy by solving the following:

$$\pi_{\theta_{k+1/2}} = \arg\max_{\pi_\theta \in \Pi_\theta} \left[\mathbb{E}\left(A^{GAE}_{\pi_{\theta_k}} \right) - \alpha_k \sqrt{\mathbb{E}\left[D_{KL}(\pi_{\theta_k}, \pi_\theta) \right]} \right]$$

Step 2: Projection onto Safe Set

Project the improved policy back onto the safe set by solving:

$$\pi_{\theta_{k+1}} = \arg\min_{\pi_\theta \in \Pi_\theta} D(\pi_\theta, \pi_{\theta_{k+1/2}}), \quad \text{s.t.} \quad J_c(\pi_\theta) \leq b$$

Update θ_{k+1}, ν_{k+1} via primal-dual optimization.

end for

4 How the proposed algorithm addressed the problem

Step 1: Performance Improvement

$$\pi_{\theta_{k+1/2}} = \arg\max_{\pi_\theta \in \Pi_\theta} \left[\mathbb{E}_{s \sim d_{\lambda \pi_{\theta_k}}(s), a \sim \pi_{\theta_k}(a|s)} \left(\frac{\pi_\theta(a|s)}{\pi_{\theta_k}(a|s)} A^{GAE}_{\pi_{\theta_k}}(s,a) \right) - \alpha_k \sqrt{\mathbb{E}_{s \sim d_{\lambda \pi_{\theta_k}}} \left[D_{KL}(\pi_{\theta_k}, \pi_\theta) \right]} \right]$$

Step 2: Projection

$$\pi_{\theta_{k+1}} = \arg\min_{\pi_\theta \in \Pi_\theta} D(\pi_\theta, \pi_{\theta_{k+1/2}}), \quad \text{s.t.} \quad C_{\pi_{\theta_k}}(\pi_\theta, \beta_k) \leq b$$