Paper Critique

Anirud N, CE21B014

Course: DA7400, Fall 2024, IITM
Paper: Projection-based constrained policy optimization
Date: 18 September 2024
Make sure your critique addresses 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 Problem Statement

In real-world applications such as self-driving cars and aerial vehicles, safety must be prioritized. Agents exploring these environments face additional costs, meaning they cannot be given complete freedom during exploration. This paper addresses the problem of learning control policies that optimize reward functions while ensuring that safety constraints are satisfied. Current solutions include: 1) forming a constrained policy optimization problem solved via conditional gradient descent, and 2) using a hyperparameter-weighted copy of the constraints. The former can lead to infeasibility if the current policy violates constraints, while the latter demands extensive hyperparameter tuning, making it computationally expensive.

2 Key Contributions

The paper introduces a projection-based constrained policy optimization approach, where policy updates occur in two stages:

- First, maximize the reward using a trust region optimization method without enforcing constraints. This may produce an intermediate policy that violates constraints.
- The second stage checks for constraint violations. If violations occur, the policy is projected onto the constraint set, selecting the policy closest to the intermediate one.
- The paper provides theoretical analysis, offering lower bounds on reward improvement and upper bounds on constraint violations.
- The proposed method is compared on four control tasks.

3 Proposed Algorithm/Framework

Step 1: Reward Improvement Optimize the reward function by maximizing the advantage function $A_R^{\pi^k}(s,a)$, subject to a KL divergence constraint, ensuring the updated policy stays within a δ-neighborhood of the previous policy.

$$\pi^{k+\frac{1}{2}} = argmax_\pi \, \mathbb{E}_{\substack{s\sim d^{\pi^k} \\ a\sim\pi}} \left[A_R^{\pi^k}(s,a) \right]$$
$$\text{s.t.} \, \mathbb{E}_{s\sim d^{\pi^k}} \left[\mathrm{KL}(\pi \,||\, \pi^k)[s] \right] \leq \delta. \tag{1}$$

This constraint transforms the optimization into a Trust Region Policy Optimization (TRPO) problem.

Step 2: Projection Step

$$\begin{aligned} \pi^{k+1} = argmin_{\pi} \quad & D(\pi, \pi^{k+\frac{1}{2}}) \\ \text{s.t.} \quad & J^C(\pi^k) + \mathbb{E}_{\substack{s\sim d^{\pi^k} \\ a\sim\pi}} \left[A_C^{\pi^k}(s,a) \right] \leq h. \end{aligned} \tag{2}$$

Here, D measures the distance between the intermediate policy and the next policy iteration. The paper explores two distance metrics: L^2 norm and KL divergence, with KL divergence projecting within probability distributions.

3.1 Implementation

Linearize the objective function in the reward improvement step. At iteration k, the linearized problem is solved subject to a second-order approximation of the KL divergence:

$$\begin{aligned} \theta^{k+\frac{1}{2}} = argmax_{\theta} \quad & \mathbf{g}^T(\theta - \theta^k) \\ \text{s.t.} \quad & \frac{1}{2}(\theta - \theta^k)^T \mathbf{H}(\theta - \theta^k) \leq \delta. \end{aligned} \tag{3}$$

If the projection is defined on a parameter space use L^2 norm. if it is defined on probability space- use KL divergence. This is approximated using a second-order expansion. Linearise the cost constraint at π^k

$$\begin{aligned} \theta^{k+1} = argmin_{\theta} \quad & \frac{1}{2}(\theta - \theta^{k+\frac{1}{2}})^T \mathrm{L}(\theta - \theta^{k+\frac{1}{2}}) \\ \text{s.t.} \quad & a^T(\theta - \theta^k) + b \leq 0, \end{aligned} \tag{4}$$

where $\mathrm{L} = I$ for L^2 norm projection, and $L = H$ for KL divergence projection.

4 Conclusions and Results

PCPO improves the reward function while having the fastest ability to satisfy the constraints for all the tasks analysed in the paper. The paper also claims it is the only model to learn constraint-bound policies across all tasks. CPO violates the constraint more, and PDO is very conservative. But this algorithm is balanced. FPO needs a lot of effort for tuning hyperparameters. PCPO is a more robust learning and constraint-satisfying policy than the other algorithms.
The paper claims that with a relatively small step of policy update, the constraint violation and the reward degradation in the worst-case scenario are tolerable. The paper also identifies that their model convergence of PCPO is deeply affected by the Fischer information matrix that is used during the training. Numerically, PCPO achieve 3.5 times fewer constraint violation and 15% higher reward than the current algorithms.