Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Constrained Policy Optimisation]
Date: [11-09-2024]

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 problem addressed in the paper is the optimization of policies in reinforcement learning under constraints, a setting often modeled by Constrained Markov Decision Processes (CMDPs). The challenge is to develop a policy search algorithm that guarantees constraint satisfaction throughout the training process while optimizing for high-dimensional control tasks.

This problem can be formulated as follows:

$$\text{Maximize:} \quad J(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})\right]$$

subject to:

$$J_{C_i}(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t C_i(s_t, a_t, s_{t+1})\right] \leq d_i, \quad \forall i = 1, \ldots, m$$

Where π is the policy to be optimized, $R(s_t, a_t, s_{t+1})$ is the reward function, $C_i(s_t, a_t, s_{t+1})$ are auxiliary cost functions representing the constraints, d_i are the limits on the expected costs, $\gamma \in [0, 1)$ is the discount factor. The goal is to find a policy π that maximizes the expected reward $J(\pi)$ while ensuring that all constraint costs $J_{C_i}(\pi)$ remain below their limits d_i.

2 Key contributions of the paper

The key contributions of the paper are as follows:

- **Constrained Policy Optimization (CPO) Algorithm:** The paper introduces Constrained Policy Optimization (CPO), a novel policy search algorithm designed for Constrained Markov Decision Processes (CMDPs). The algorithm guarantees near-constraint satisfaction throughout the entire training process while optimizing the policy's performance.
- **Theoretical Performance Bound:** The authors provide a new theoretical result that bounds the difference in expected returns between two policies, based on an average divergence. This result tightens existing policy search bounds and helps ensure both policy improvement and constraint satisfaction. The performance bound is given by:

$$J(\pi') - J(\pi) \geq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\pi, a \sim \pi'}\left[A_\pi(s, a) - \frac{2\gamma \epsilon_{\pi'}}{1-\gamma} D_{TV}(\pi' \| \pi)\right]$$

where $J(\pi)$ is the performance of policy π, $A_\pi(s,a)$ is the advantage function, and D_{TV} is the total variation divergence between policies.

- **Empirical Validation:** The paper demonstrates the effectiveness of CPO on several high-dimensional robotic locomotion tasks, where neural network policies are trained to satisfy safety constraints. CPO is shown to outperform primal-dual optimization methods and other baseline algorithms in terms of constraint satisfaction and policy performance. Empirically, the paper evaluates CPO on tasks like "Circle" and "Gather", showing the performance and constraint values:

$$\text{Maximize: } \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t R(s_t,a_t)\right] \quad \text{subject to} \quad \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t C(s_t,a_t)\right] \leq d$$

3 Proposed algorithm/framework

Algorithm 1 Constrained Policy Optimization

Input: Initial policy $\pi_0 \in \Pi_\theta$, tolerance α

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

1: Sample a set of trajectories $D = \{\tau\} \sim \pi_k = \pi(\theta_k)$
2: Form sample estimates $\hat{g}, \hat{b}, \hat{H}, \hat{c}$ with D
3: **if** approximate CPO is feasible **then**
4: Solve dual problem (12) for λ_k^*, ν_k^*
5: Compute policy proposal θ^*
6: **else**
7: Compute recovery policy proposal θ^*
8: **end if**
9: Obtain θ_{k+1} by backtracking line search to enforce satisfaction of sample estimates of constraints

end for

4 How the proposed algorithm addressed the problem

- Surrogate Objective and Constraints: CPO optimizes a surrogate objective:

$$L(\pi_k) = \mathbb{E}_{s\sim d_{\pi_k}, a\sim\pi_k}\left[A_{\pi_k}(s,a)\right]$$

and maintains constraints:

$$J_{C_i}(\pi_k) + \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_{\pi_k}, a\sim\pi}\left[A_{\pi_k C_i}(s,a)\right] \leq d_i$$

- Trust Region: Policy updates are constrained by KL-divergence:

$$D_{KL}(\pi_{k+1}\|\pi_k) \leq \delta$$

ensuring stable improvements.

- Constraint Satisfaction: CPO guarantees constraint satisfaction during each iteration, directly incorporating constraints into the optimization.

- Performance Bound: The algorithm guarantees bounded performance degradation:

$$J(\pi_{k+1}) - J(\pi_k) \geq -\frac{\sqrt{2\delta\gamma\epsilon_{\pi_{k+1}}}}{(1-\gamma)^2}$$