Paper Critique

Shuvrajeet Das, DA24D402

Course: DA7400, Fall 2024, IITM
Paper: [Reward Constrained Policy Optimization]
Date: [13-09-2024]

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

In this work, they present an approach for constrained policy optimization, called 'Reward Constrained Policy Optimization' (RCPO), which uses an alternative penalty signal to guide the policy towards a constraint-satisfying one. They proved the convergence of our approach and provide empirical evidence of its ability to train constraint satisfying policies.

Maximize the expected cumulative reward:

$$\max_{\pi \in \Pi} J_R^\pi = \mathbb{E}_{s\sim\mu}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$$

subject to the constraint on cumulative costs:

$$J_C^\pi = \mathbb{E}_{s\sim\mu}\left[C(s)\right] \leq \alpha$$

2 Key contributions of the paper

- The paper introduces a novel algorithm called *Reward Constrained Policy Optimization (RCPO)* that handles both discounted sum constraints and mean value constraints, which generalizes beyond standard reinforcement learning approaches.
- RCPO incorporates constraints as a penalty signal into the reward function to guide the policy towards constraint satisfaction.
- The paper proves that RCPO converges almost surely to a constraint-satisfying solution under mild assumptions.
- Empirical results demonstrate that RCPO converges faster and more stably than traditional constraint optimization methods, with experiments conducted on a toy domain and six robotics domains.
- The approach is reward-agnostic, meaning it is invariant to scaling of the underlying reward signal, and does not require prior knowledge to perform effectively.

3 Proposed algorithm/framework

Algorithm 1 Template for an RCPO implementation

```
1: **Input:** penalty $c(\cdot)$, constraint $C(\cdot)$, threshold $\alpha$, learning rates $\eta_1(k) < \eta_2(k) < \eta_3(k)$
2: Initialize actor parameters $\theta = \theta_0$, critic parameters $\nu = \nu_0$, Lagrange multipliers and $\lambda = 0$
3: **for** $k = 0, 1, \ldots$ **do**
4:     Initialize state $s_0 \sim \mu$
5:     **for** $t = 0, 1, \ldots, T-1$ **do**
6:         Sample action $a_t \sim \pi$, observe next state $s_{t+1}$, reward $r_t$ and penalties $c_t$
7:         $\hat{R}_t = r_t - \lambda c_t + \gamma V(\lambda, s_{t+1}, \nu_t)$
8:         Critic update: $\nu_{k+1} = \nu_k - \eta_3(k) \left[\frac{\partial \hat{R}_t - \hat{V}(\lambda, s_t; \nu)}{\partial \nu}\right]^2$
9:         Actor update: $\theta_{k+1} = \theta_k + \Gamma_\theta \left[\theta_k + \eta_2(k) \nabla_\theta V(\hat{V}, s)\right]$
10:        Lagrange multiplier update: $\lambda_{k+1} = \lambda_k + \eta_1(k)[\lambda_k + \eta_1(k)(J_C^\pi - \alpha)]$
11:    **end for**
12: **end for**
13: **return** policy parameters $\theta$
```

4 How the proposed algorithm addressed the problem

The proposed algorithm, *Reward Constrained Policy Optimization (RCPO)*, addresses the problem by reformulating the constrained optimization problem as a Lagrangian:

$$\min_{\lambda \geq 0} \max_{\theta} L(\lambda, \theta) = \min_{\lambda \geq 0} \max_{\theta} \left[J_R^\pi(\theta) - \lambda \left(J_C^\pi(\theta) - \alpha\right)\right]$$

The algorithm then uses a multi-timescale optimization approach:

- On the faster timescale, it updates the policy parameters θ using policy gradient methods to maximize the Lagrangian.
- On the slower timescale, it updates the Lagrange multiplier λ to ensure that the constraint is satisfied over time.

The RCPO algorithm also incorporates a guiding penalty signal $C_\gamma(s)$ in the value function, defined as:

$$V_C^\pi(s) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) \mid s_0 = s\right]$$

This penalty helps guide the policy towards constraint satisfaction while still optimizing for reward.