Paper Critique

Anirud N, CE21B014

Course: DA7400, Fall 2024, IITM
Paper: Constrained Update Projection Approach to Safe Policy Optimization
Date: 27 September 2024

1 Problem Statement

Most of the current RL algorithms assume that the agent is free to explore any behavior. But this is not always possible in real-world scenarios. There is a need to consider safety and constraints. Recent approaches involve the use of some convex approximations to non-convex problems, which leads to sources of errors. A few implementations also involve the use of a first or second-order Taylor expansion, but this still lacks a theory to show the errors due to these approximations. These current approaches also involve an inverse of the Fischer information matrix - which is computationally costly,

2 Key Contributions

This paper introduces the CUP algorithm (Constrained update projection), with a theoretical safety guarantee. The CUP algorithm is a two-step process: (i) policy improvement that might temporarily violate constraints and (ii) a projection step that brings the policy back into a safe region. CUP is adaptable to high-dimensional problems without relying on convex approximations of the objective or constraints. The paper shows theoretically the bounds of the difference between two arbitrary policies. The paper shows a theoretical bound with respect to the Generalised advantage estimator, which reduces variance and has some decent level of bias acceptable. The paper unifies CPO -CPO is a special case of the bounds proven in the paper. The paper shows the theoretical aspect for using GAE to extend surrogate functions with respect to GAE. The paper shows the effectiveness of CUP on high-dimensional tasks.

3 Proposed Algorithm/Framework

Generalised Advantage Estimator:

$$\hat{A}_t^{\texttt{GAE}(\gamma,\lambda)}(s_t, a_t) = \sum_{\ell=0}^{\infty} (\gamma\lambda)^\ell \delta_{t+\ell}^V, \tag{1}$$

where $\delta_t^V = r_{t+1} + \gamma V(s_{t+1}) - V(s_t)$ is TD error, and $V(\cdot)$ is an estimator of value function. Extending this to the auxiliary cost function implies:

$$\hat{A}_{C,t}^{\texttt{GAE}(\gamma,\lambda)}(s_t, a_t) = \sum_{\ell=0}^{\infty} (\gamma\lambda)^\ell \delta_{t+\ell}^C, \tag{2}$$

where $\delta_t^C = r_{t+1} + \gamma C(s_{t+1}) - C(s_t)$ is TD error, and $C(\cdot)$ is an estimator of cost function c. The paper shows a theoretical bound on the error between the estimated returns of two different policies, expressed as a linear combination of the TD error between the policies and the discounted distributional difference between them.

3.1 Algortithm

Performance Improvement The objective is a minimization-maximization problem, maximizing the returns while minimizing the cost.

$$\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \left\{ \underset{\substack{s \sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot) \\ a \sim \pi_{\boldsymbol{\theta}_k}(\cdot|s)}}{E} \left[\frac{\pi_{\boldsymbol{\theta}}(a|s)}{\pi_{\boldsymbol{\theta}_k}(a|s)} A^{\mathtt{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}}(s,a) \right] - \alpha_k \sqrt{E_{s \sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k},)[s] \right]} \right\}. \quad (3)$$

Projection This involves projection of policy found after improvement onto the safe constraint set.

$$\pi_{\boldsymbol{\theta}_{k+1}} = \arg\min_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} D\left(\pi_{\boldsymbol{\theta}}, \pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}\right), \text{ s.t. } C_{\pi_{\boldsymbol{\theta}_k}}(, \beta_k) \leq b, \quad (4)$$

$$C_{\pi_{\boldsymbol{\theta}_k}}(, \beta) = J^c(\pi_{\boldsymbol{\theta}_k}) + \frac{1}{1-\tilde{\gamma}} E_{s \sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot), a \sim (\cdot|s)} \left[A^{\mathtt{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}}, C}(s,a) \right] + \beta \sqrt{E_{s \sim d^{\lambda}_{\pi_{\boldsymbol{\theta}_k}}(\cdot)} \left[\mathrm{KL}(\pi_{\boldsymbol{\theta}_k},)[s] \right]}.$$

3.2 Implementation

They used Empirical KL divergence :

$$\hat{D}_{\mathrm{KL}}(\pi_{\theta}, \pi'_{\theta}) = \frac{1}{T} \sum_{t=1}^{T} \mathrm{KL}((a_t|s_t), (a_t|s_t)).$$

Performance update is done as follows:

$$\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}} = \arg\max_{\pi_{\boldsymbol{\theta}} \in \Pi_{\theta}} \left\{ \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_k}(a_t|s_t)} \hat{A}_t - \alpha_k \sqrt{\hat{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})} \right\}, \quad (5)$$

$\hat{A}_t$ is an estimator of $A^{\mathtt{GAE}(\gamma,\lambda)}_{\pi_{\boldsymbol{\theta}_k}}(s,a)$. The projection step in equation 4 is solved by using a primal-dual approach.

$$(\pi_{\boldsymbol{\theta}_{k+1}}, \nu_{k+1}) = \arg\min_{\pi_{\boldsymbol{\theta}} \in \Pi_{\boldsymbol{\theta}}} \max_{\nu \geq 0} \left\{ \hat{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}_{k+\frac{1}{2}}}, \pi_{\boldsymbol{\theta}}) + \nu \hat{C}(, \pi_{\boldsymbol{\theta}_k}) \right\}$$

where $\hat{C}(, \pi_{\boldsymbol{\theta}_k}) = \hat{J}^C + \frac{1}{1-\tilde{\gamma}} \cdot \frac{1}{T} \sum_{t=1}^{T} \frac{\pi_{\boldsymbol{\theta}}(a_t|s_t)}{\pi_{\boldsymbol{\theta}_k}(a_t|s_t)} \hat{A}^C_t + \beta_k \sqrt{\hat{D}_{\mathrm{KL}}(\pi_{\boldsymbol{\theta}_k}, \pi_{\boldsymbol{\theta}})} - b$, $\hat{J}^C$and $\hat{A}^C_t$ are estimators for cost-return and cost-advantage.

4 Conclusions and Results

Experiments show that even a smaller level of hyperparameter tuning is enough and the algorithm performs well. CUP is robust to hyper-parameter tuning and cost limit variations, and it consistently learns stable policies despite diverse task difficulties. It was observed that CUP shows faster convergence and stable constraint returns in lesser steps. CUP ensures safety constraints, even in difficult tasks like Swimmer-v3 and HumanoidCircle, where other algorithms struggled. The paper shows the superior performance of CUP over algorithms like CPO, PCPO that approximates a non convex problem to a convex problem.