Constrained Variational Policy Optimization for Safe Reinforcement Learning

Zuxin Liu[1], Zhepeng Cen[1], Vladislav Isenbaev[2], Wei Liu[2], Zhiwei Steven Wu[1], Bo Li[3], Ding Zhao[1]

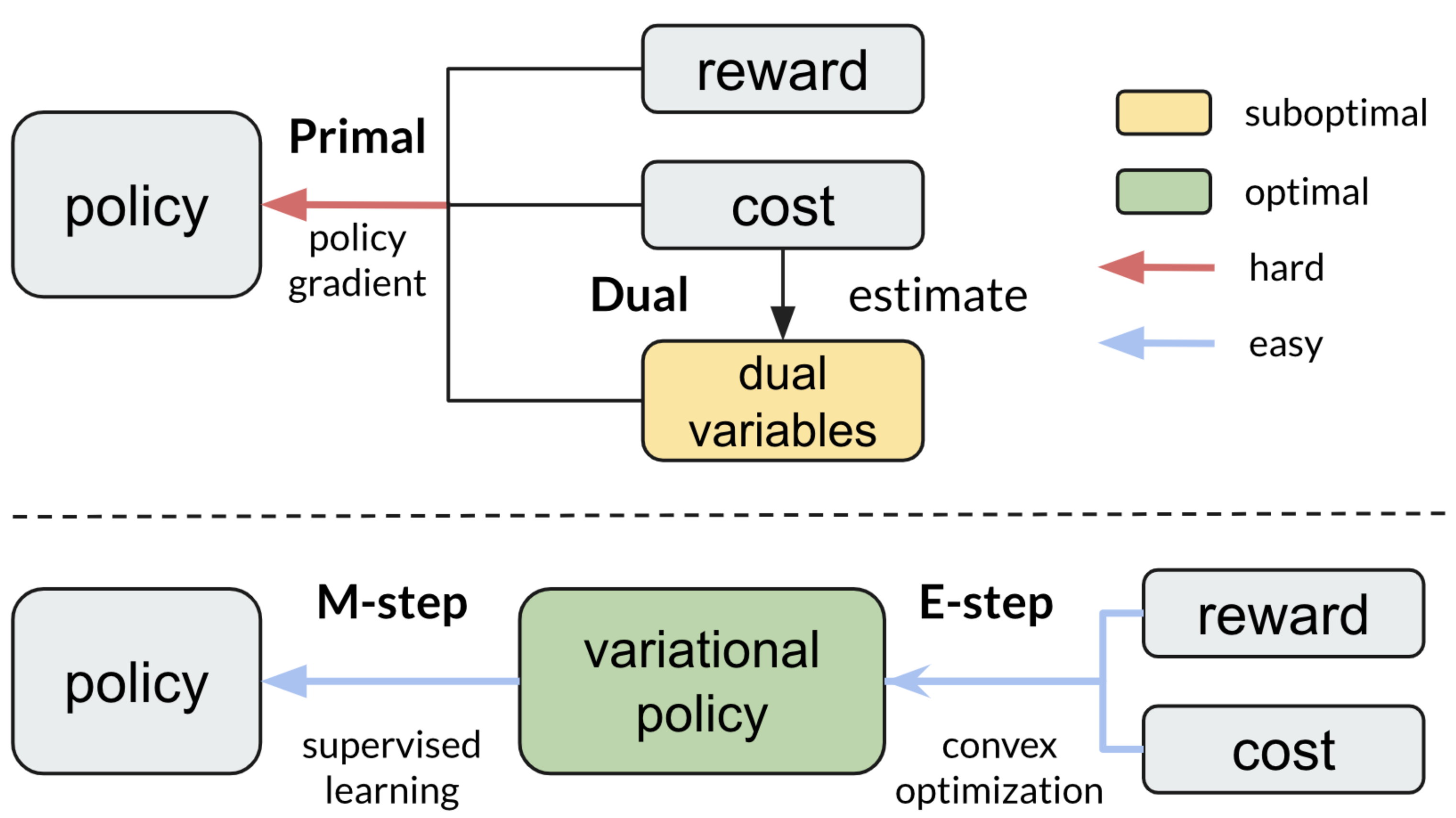
[1]*Carnegie Mellon University*, [2]*Nuro, Inc.*, [3]*UIUC*

Introduction & Background

Safe reinforcement learning (RL) aims to learn policies that satisfy certain constraints before deploying to safety-critical applications:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_t \gamma^t r_t\right], \text{s.t.}\, \mathbb{E}\left[\sum_t \gamma^t c_t\right] \le \epsilon$$

- Previous primal-dual style approaches suffer from instability issue and lack optimality guarantees.
- We solve the safe RL problem from the probabilistic inference perspective and propose an EM-style method **CVPO** (constrained variational policy optimization) with 3 advantages: (1) Sample efficient (2) stable performance (3) With optimality guarantees.



Method: CVPO

The benefits of viewing safe RL as inference:

- There is no inaccurate *dual variable optimization* and difficult *policy improvement*.
- Introducing a variational distribution and solving constrained optimization by EM algorithm.

Objective: optimize the evidence lower bound (ELBO) in a feasible (constraint satisfied) policy set Π^{ϵ_1}.

$$\max \mathcal{J}(q,\theta) \triangleq \mathbb{E}_{\tau\sim q}[\sum(\gamma^t r_t - \alpha D_{KL}[q|\pi_\theta])] + \log p(\theta),$$
$$s.t.\, q \in \Pi^{\epsilon_1}$$

E-step: to find the optimal variational distribution q to

- Maximize the return of task reward;
- Satisfy the safety constraints meanwhile.

$$\max_q \mathbb{E}_{\tau\sim q}\left[\int q(a|s) Q_r^{\pi_{\theta_i}}(s,a)da\right],$$

$$s.t. \begin{cases} \mathbb{E}_{\tau\sim q}\left[\int q(a|s) Q_c^{\pi_{\theta_i}}(s,a)da\right] \le \epsilon_1 \\ \mathbb{E}_{\tau\sim q}\left[D_{KL}\left[q(a|s)|\pi_{\theta_i}\right]\right] \le \epsilon_2 \\ \int q(a|s)da = 1 \end{cases}$$

With Slater's condition, the above problem has closed-form solution,

$$q^*(a|s) = \frac{\pi_\theta(a|s)}{Z}\exp\left(\frac{Q_r^{\pi_\theta}(s,a) - \lambda^* Q_c^{\pi_\theta}(s,a)}{\eta^*}\right),$$

where λ^*, η^* can be solved by **convex optimization**.

M-step: To improve the ELBO w.r.t. θ by fitting π_θ to the optimal variational policy q^*.

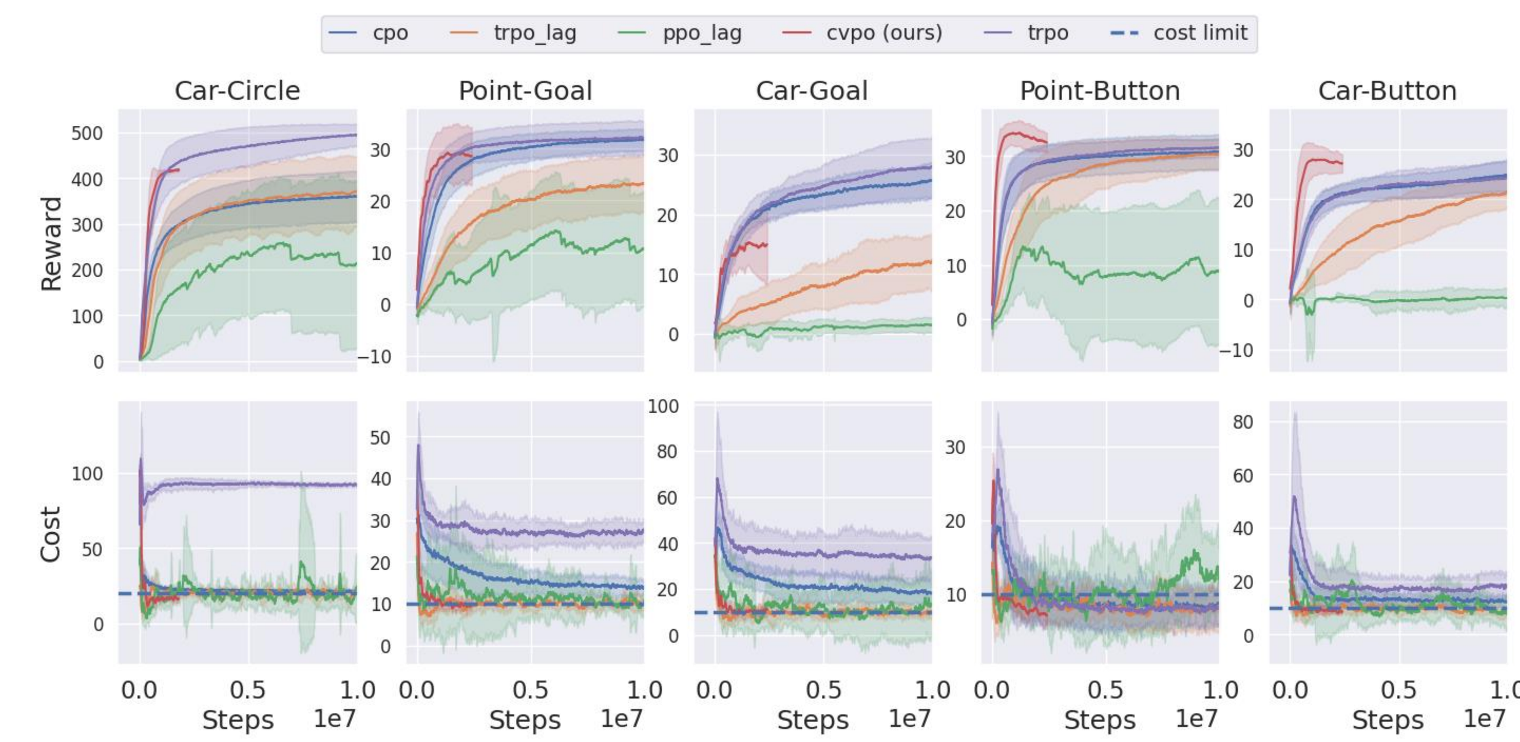
$$\max_\theta \mathbb{E}_{\tau\sim q}\left[\mathbb{E}_{q^*(\cdot|s)}[\log \pi_\theta(a|s)]\right],$$
$$s.t.\, \mathbb{E}_{\tau\sim q}\left[D_{KL}[\pi_{\theta_i}|\pi_\theta]\right] \le \epsilon$$

The M-step is a **supervised learning** problem that is easier to solve than policy gradient.

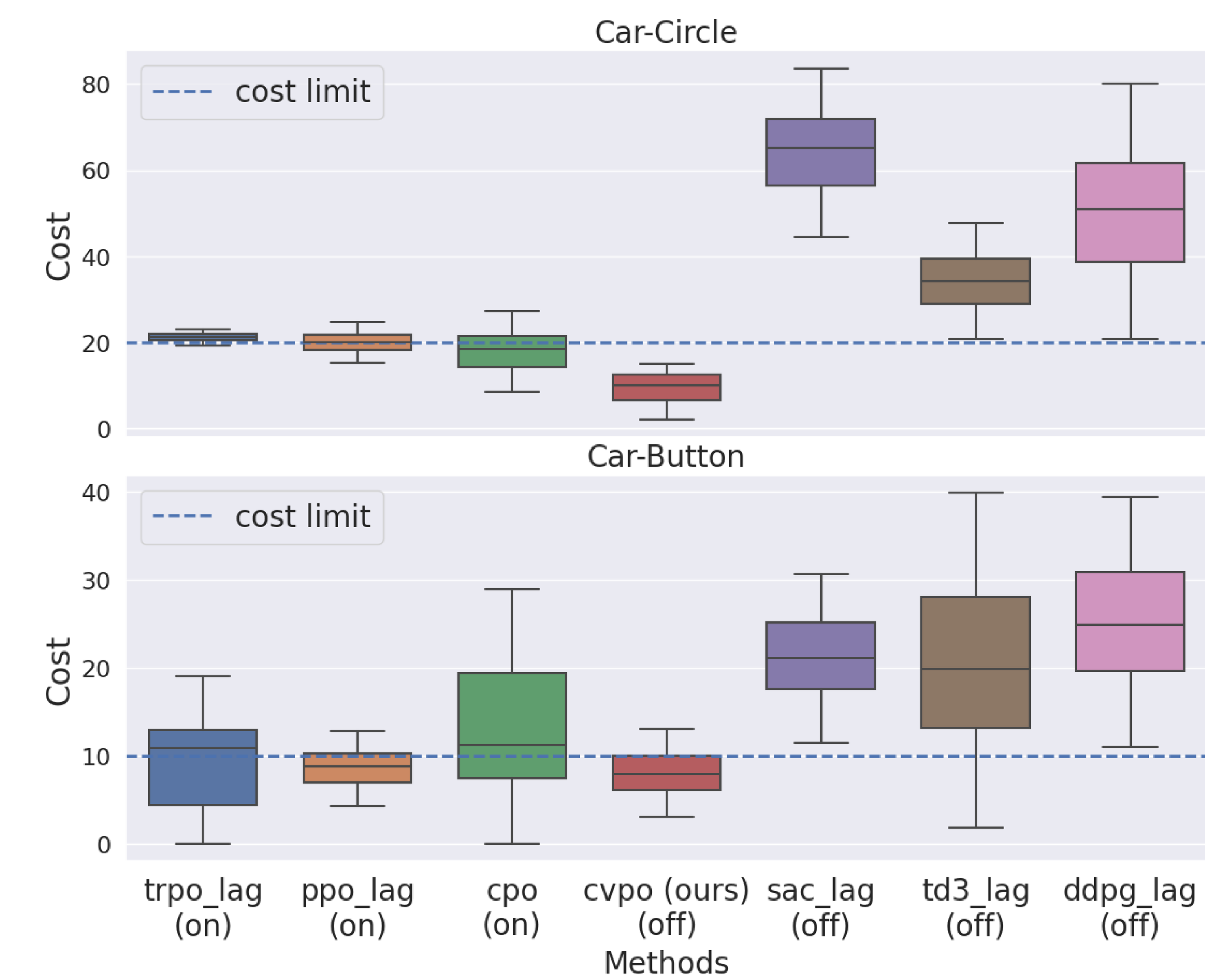
Theoretical guarantees:

- Monotonic improvement by EM algorithm.
- Bounded worst-case safety violation.
- Robust policy improvement.

Results & Conclusion



The training curves of different safe RL methods on safety-gym tasks



Boxplot of convergence cost. On/off denote on-policy/off-policy.

For safe RL problem, CVPO enjoys the advantages of

- High **sample-efficiency** from off-policy algorithm;
- **Stable** performance and constraint satisfaction;
- Theoretical **optimality** & **feasibility** guarantees.