Towards Safe Reinforcement Learning via Constraining Conditional Value-at-Risk

Chengyang Ying, Xinning Zhou, Hang Su, Dong Yan, Ning Chen, Jun Zhu
Tsinghua University
Contact: ycy21@mails.tsinghua.edu.cn

Safe Reinforcement Learning

- Traditional reinforcement learning focuses on maximizing the cumulative return but ignore the risk.

- Safe reinforcement learning consider the uncertainty in the reinforcement learning, including transition uncertainty and observation uncertainty.

- Transition Uncertainty

$$\max_{\theta} \min_{\mathcal{P} \in \hat{\mathcal{P}}} J_{\mathrm{tr}}(\pi_\theta, \mathcal{P}) \triangleq \mathbb{E}\left[D(\pi_\theta) \triangleq \sum_{t=1}^{\infty} \gamma^t r_t \middle| \pi_\theta, \mathcal{P} \right]$$

- Observation Uncertainty

$$\max_{\theta} \min_{\nu \in \Gamma} J_{\mathrm{obs}}(\pi_\theta) \triangleq \mathbb{E}\left[D(\pi_\theta, \nu) \triangleq \sum_{t=1}^{\infty} \gamma^t r_t \right]$$

Value Function Range

- Value Function Range

Definition 2 (Value Function Range). *For MDP $\mathcal{M}$, the Value Function Range (VFR) of the policy π is*

$$\hat{V}_{\mathcal{M},\pi} \triangleq \max_s V_{\mathcal{M},\pi}(s) - \min_s V_{\mathcal{M},\pi}(s). \tag{6}$$

- We provide the connectation of transition disturbance and observation disturbance.

$$\left|J_{\mathcal{M}}(\pi) - J_{\hat{\mathcal{M}}}(\pi)\right| \leq \frac{2\gamma}{1-\gamma}\epsilon_{\mathcal{P}}\hat{V}_{\mathcal{M},\pi}$$

$$\left|J_{\mathcal{M}}(\pi) - J_{\mathcal{M}}(\hat{\pi}_{\nu})\right| \leq \frac{\gamma}{1-\gamma}\epsilon_{\pi}\hat{V}_{\mathcal{M},\pi} + \frac{2}{1-\gamma}\epsilon_{\pi}.$$

- We can control them both by controlling Value Function Range.

Theorem 3 (Proof in Appendix B.2). *For any $\alpha \in [0,1]$, we have*

$$-\mathrm{CVaR}_{\alpha}(-D(\pi)) \leq -\mathrm{CVaR}_{\alpha}(-V(s)). \tag{10}$$

- We introduce CVaR to loose the min in Value Function Range to avoid excessive pessimism.

CPPO

- We formulate our objective as a constrained optimization problem and use Lagrangian relaxation method to deform it as an unconstrained problem.

$$\min_{\theta,\eta} -J(\pi_\theta) \quad s.t. \ \frac{1}{1-\alpha}\mathbb{E}[(\eta - D(\pi_\theta))^+] - \eta \le -\beta.$$

$$\max_{\lambda \ge 0} \min_{\theta,\eta} L(\theta, \eta, \lambda)$$
$$\triangleq -J(\pi_\theta) + \lambda\left(\frac{1}{1-\alpha}\mathbb{E}[(\eta - D(\pi_\theta))^+] - \eta + \beta\right).$$

- Based on previous work [Chow and Ghavamzadeh, 2014], we can calculate the gradient of our objective and further propose CVaR Proximal Policy Optimization (CPPO).

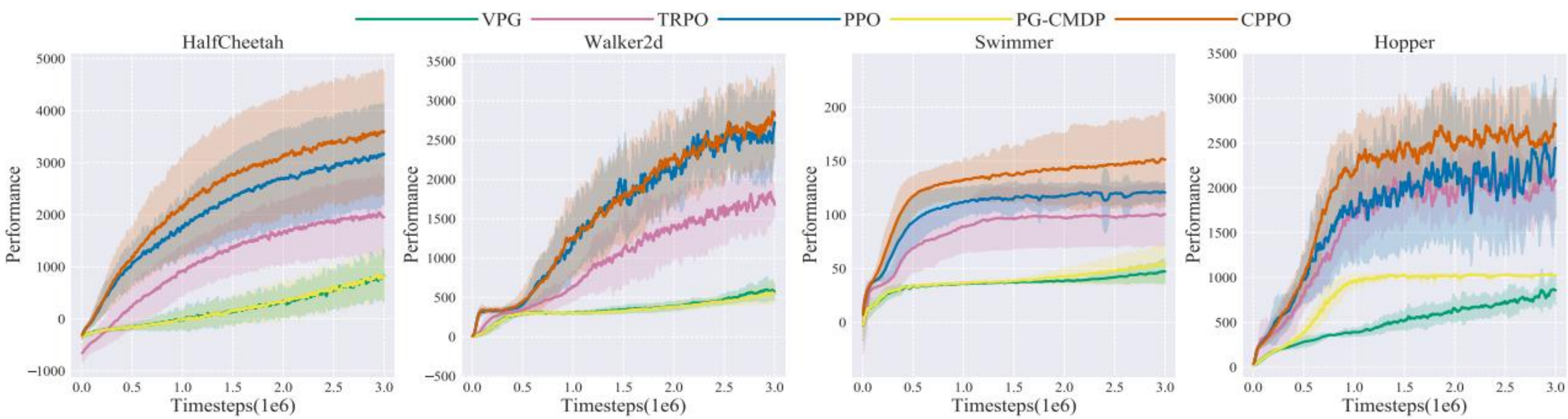
Algorithm 1 CVaR Proximal Policy Optimization (CPPO)

Require: confidence level α, learning rate $lr_\eta, lr_\theta, lr_\lambda, lr_\phi$
Ensure: parameterized policy π_θ and parameterized value function V_ϕ.

1: **for** $k = 1, 2, ..., N_{iter}$ **do**
2: Generate N trajectories with the current policy π_θ.
3: Compute advantage estimates $\hat{A}_i^t$ of each state $s_{i,t}$ in each trajectory ξ_i and the cumulative reward $D(\xi_i)$.
4: Update parameters $\eta, \theta, \lambda, \phi$ respectively with the calculated gradients.
5: Modify β as a function of current trajectories' return.
6: **end for**

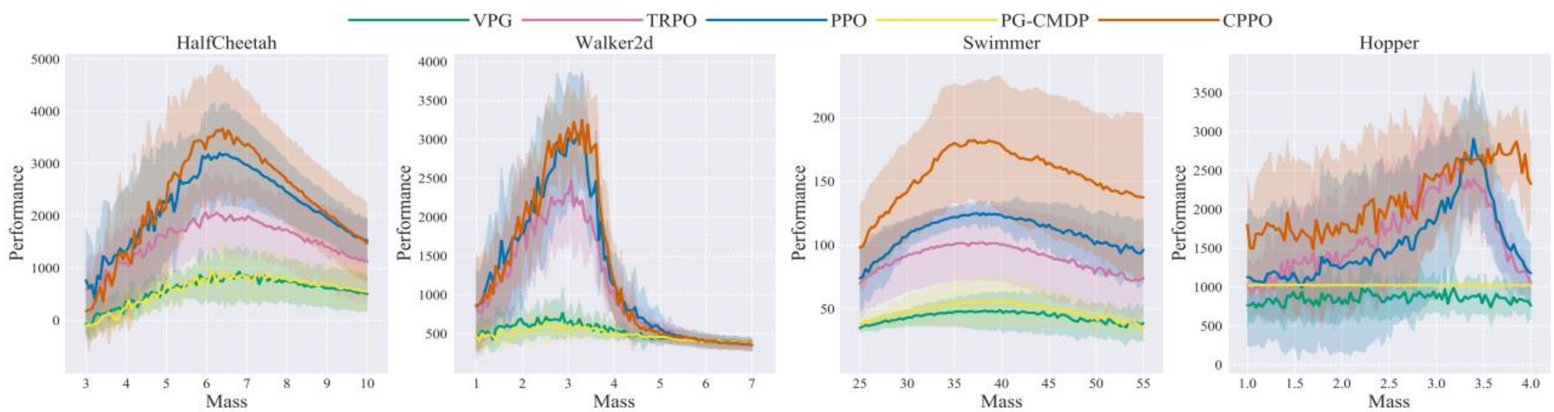
Evaluation on MuJoCo

Method	Ant-v3	HalfCheetah-v3	Walker2d-v3	Swimmer-v3	Hopper-v3
VPG	12.8± 0.0	896.9± 531.1	628.6± 229.4	48.3± 11.3	888.4± 209.5
TRPO	1625.4± 356.4	2073.8± 741.3	2005.6± 398.7	101.2± 29.3	2391.4± 455.3
PPO	3372.2± 301.4	3245.4± 947.3	2946.3± 944.3	122.0± 7.9	2726.0± 886.0
PG-CMDP	7.4 ± 3.6	928.7± 562.9	596.7± 219.9	55.4± 18.8	1039.2± 21.1
CPPO(ours)	**3514.7± 247.2**	**3680.5± 1121.3**	**3194.0± 648.2**	**182.5± 46.0**	**3144.6± 158.4**

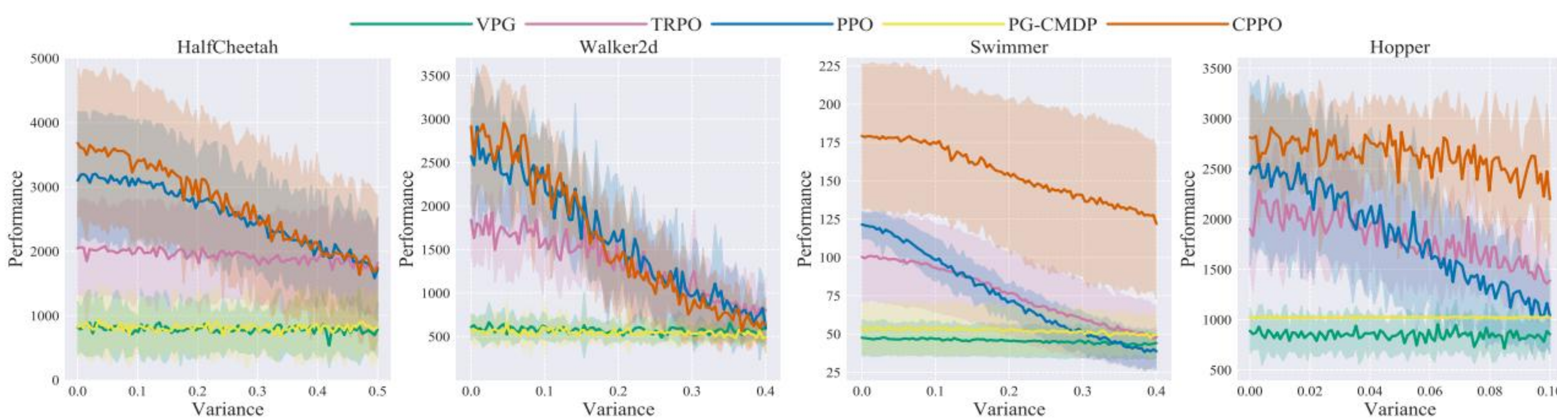


curve of the training performance

Evaluation on MuJoCo



curve of the testing performance under transition disturbance



curve of the testing performance under observation disturbance

Thanks