Towards Safe Reinforcement Learning via Constraining Conditional Value at Risk

Chengyang Ying, Xinning Zhou, Dong Yan, Jun Zhu

Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua University; Contact: ycy21@mails.tsinghua.edu.cn

A Blessing in Disguise: The Prospects and Perils of Adversarial Machine Learning

Workshop at ICML 2021

Introduction

❑ **Deep Reinforcement Learning (DRL)** has achieved enormous success among a variety of tasks, ranging from playing Atari games and Go to manipulating complex robotics in the real world.

❑ However, standard reinforcement learning (RL) objective only focuses on cumulative reward without taking account of the risk of the policy, which may cause catastrophic results.

❑ The uncertainty of RL can be categorized into the inherent uncertainty and the parameter uncertainty (Garcıa & Fern´andez, 2015). Thus there are two main categories of safe RL. The first one is based on the modification of the optimality objective, and the second one is based on the modification of the exploration process

❑ **Limitations of current safe RL methods**:

- ❑ The current safe RL methods always consider minmax problems, which don't have general solutions and traditional methods usually require high computation complexity.
- ❑ The current safe RL methods always focus on the worst trajectories, which may cause over pessimistic behaviors.
- ❑ The direct usage of variance to penalize risk in current safe RL methods will not only eliminate the probability of particularly bad trajectories, but also particularly good ones, and thus causes a drop in the agents' performance.

Contributions

❑ We choose CVaR as the metric for quantifying the risk of policy and formalize safe RL as a constrained optimization problem in order to achieve policies with high cumulative reward and low risk.

❑ We acquire a gradient-based method called CVaR Proximal Policy Optimization (CPPO) to maximize the expected reward while keeping the risk within a reasonably low level by extending Proximal Policy Optimization method.

❑ We theoretically analyze the reduced cumulative rewards of policies against state observation disturbance and transition probability disturbance. Moreover, we analyze the connection between these two kinds of noises.

❑ We prove that our method exhibits stronger robustness under maliciously crafted perturbations than general Proximal Policy Optimization theoretically and experimentally.

Method

❑ **Problem Formulation**:

- Value at Risk (VaR) and Conditional Value at Risk (CVaR) are useful metrics for evaluating risk and their definitions are

$$\mathrm{VaR}_\alpha(Z) = \min\{z|F(z) \geq \alpha\}$$

$$\mathrm{CVaR}_\alpha(Z) = E\{z|z \geq \mathrm{VaR}_\alpha(z)\}$$

- By using the property of CVaR, we can loose the minmax problems of safe RL into asolvable optimization problem.
- Balance the standard RLobjective and safe RL objective, we can propose our constrained optimization problemas below:

$$\max_\theta J(\pi_\theta)$$
$$s.t. -\mathrm{CVaR}_\alpha(-D(\pi_\theta)) \geq \beta.$$

- By using the property of CVaR and Lagrangian relaxation method, we e need to solve the saddle pointof the function L(θ,ν,λ):

$$\max_{\lambda\geq 0}\min_{\theta,\nu} L(\theta,\nu,\lambda) \triangleq -J(\pi_\theta) + \lambda(-\nu + \frac{1}{1-\alpha}E[(-D(\pi_\theta)+\nu)^+] + \beta)$$

❑ **Properties of Our Objective:**

We assume the optimal policy of our objective is $\pi_c(\alpha,\beta)$, and we can give a lower bound of its cumulative reward:

$$J(\pi_c(\alpha,\beta)) \geq \frac{J(\pi_s) - \alpha M}{1-\alpha}$$

here M is the upper bound of the cumulative of every trajectory.

❑ **CPPO:**

By calculating the derivative of L(θ,ν,λ) to θ, ν and λ, we can extend Proximal Policy Optimization and propose our CVaR Proximal Policy Optimization (CPPO)

Algorithm 1 CVaR Proximal Policy Optimization(CPPO)

Require: confidence level α and reward tolerance β

Ensure: θ of parameterized policy π_θ(always be random policy), ϕ of parameterized value function V_ϕ.

for $k = 1, 2, ..., N_{iter}$ **do**

Generate N trajectories $\mathcal{D}_k = \{\xi_i\}_{i=1}^N$ by following the current policy π_θ.

Compute reward $\hat{R}_i^t$ of each state $s_{i,t}$ in each trajectory ξ_i and the cumulative reward $D(\xi_i)$.

Compute advantage estimates $\hat{A}_i^t$ of each state $s_{i,t}$ in each trajectory ξ_i.

Update parameters respectively:

$$\eta \leftarrow \eta - lr_\eta\left(-\lambda + \frac{\lambda}{N(1-\alpha)}\sum_{i=1}^N \mathbf{1}\{\eta \geq D(\xi_i)\}\right)$$

$$\theta \leftarrow \theta + lr_\theta \frac{1}{NT}\sum_{i=1}^N\sum_{t=0}^T \nabla_\theta \min\left(\frac{\pi_\theta(a_i^t|s_i^t)}{\pi_{\theta_k}(a_i^t|s_i^t)}\hat{A}_i^t, g(\epsilon,\hat{A}_i^t)\right)$$
$$- lr_\theta \frac{1}{N}\sum_{i=1}^N(\nabla_\theta \log P_\theta(\xi_i))\frac{\lambda}{1-\alpha}(-D(\xi_i)+\eta)$$
$$\mathbf{1}\{\eta \geq D(\xi_i)\}$$

$$\lambda \leftarrow \lambda + lr_\lambda\left(-\eta + \frac{\sum_{i=1}^N(-D(\xi_i)+\eta)^+}{N(1-\alpha)} + \beta\right)$$

$$\phi \leftarrow \phi + lr_\phi\left(\frac{1}{NT}\sum_{i=1}^N\sum_{t=0}^T 2(V_\phi(s_{i,t}) - \hat{R}_i^t)\nabla_\phi V_\phi(s_{i,t})\right)$$

end for

Theoretical Analysis

❑ **We calculate the reduced reward against state observation disturbance and give an upper bound of it.**

Theorem 3 *For any policy π and any adversary ν, we have:*

$$J_\mathcal{M}(\pi) - J_\mathcal{M}(\tilde{\pi}_\nu) = \frac{\gamma}{1-\gamma}E_{s\sim d_\mathcal{M}^{\tilde{\pi}_\nu}}E_{a\sim\pi(\cdot|\nu(s))}\left(1 - \frac{\pi(a|s)}{\pi(a|\nu(s))}\right)E_{s'\sim P(\cdot|s,a)}V_{\mathcal{M},\pi}(s')$$
$$+\frac{1}{1-\gamma}E_{s\sim d_\mathcal{M}^{\tilde{\pi}_\nu}}E_{a\sim\pi(\cdot|\nu(s))}\left(1 - \frac{\pi(a|s)}{\pi(a|\nu(s))}\right)R(s,a)$$

Furthermore, we can give a upper bound of it:

$$|J_\mathcal{M}(\pi) - J_\mathcal{M}(\tilde{\pi}_\nu)| \leq \frac{\gamma}{1-\gamma}\max_s D_{TV}(\pi(\cdot|s),\pi(\cdot|\nu(s)))\left(\max_s V_{\mathcal{M},\pi}(s) - \min_s V_{\mathcal{M},\pi}(s)\right)$$
$$+\frac{2}{1-\gamma}\max_s D_{TV}(\pi(\cdot|s),\pi(\cdot|\nu(s))\max_{s,a}|R(s,a)|$$

❑ **We calculate the reduced reward against transition probability disturbance and give an upper bound of it.**

Theorem 4 *For any policy π and any disturbed environment $\hat{\mathcal{M}} = (\mathcal{S},\mathcal{A},\hat{\mathcal{P}},\mathcal{R})$, we have:*

$$J_\mathcal{M}(\pi) - J_{\hat{\mathcal{M}}}(\pi) = \frac{\gamma}{1-\gamma}E_{s\sim d_\mathcal{M}^\pi}E_{a\sim\pi(\cdot|s)}E_{s'\sim\hat{P}(\cdot|s,a)}\left(1 - \frac{P(s'|s,a)}{\hat{P}(s'|s,a)}\right)V_{\mathcal{M},\pi}(s')$$

Furthermore, we can give a upper bound of it:

$$J_\mathcal{M}(\pi) - J_{\hat{\mathcal{M}}}(\pi) = \frac{2\gamma}{1-\gamma}\max_{s,a} D_{TV}(P(\cdot|s,a),\hat{P}(\cdot|s,a))\left(\max_s V_{\mathcal{M},\pi}(s) - \min_s V_{\mathcal{M},\pi}(s)\right)$$

Experimental Findings

❑ We choose MuJoCo as our environment and use VPG, TRPO, PPO as our baselines.

❑ The left figure is the performance of agents trained by diffrent algorithms in the training stage. The middle and the right figures show the robustness of trained agents under state observation disturbance and transition probability disturbance respectively.