CRPO: A New Approach for Safe Reinforcement Learning with Convergence Guarantee

Tengyu Xu[1], Yingbin Liang[1], Guanghui Lang[2]

The Ohio State University[1]
Georgia Institute of Technology[2]

International Conference on Machine Learning (ICML) 2021

Safe Reinforcement Learning

- Agent receives both reward $r(s,a)$ and costs $c_i(s,a)$ $(i = 1, \cdots, m)$
- Goal of SRL:

$$\max_w \; J_0(w)$$
$$\text{s.t.} \quad J_i(w) \leq D_i \quad (i = 1, \cdots, m)$$

  - Objective function: $J_0(w) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, s_{t+1}) \Big| s_0 \sim \mu_0, \pi_w \Big]$
  - Cost function: $J_i(w) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t c_i(s_t, a_t, s_{t+1}) \Big| s_0 \sim \mu_0, \pi_w \Big]$
  - Constraints threshold: $D_i > 0$

Primal-Dual Approach

- Construct a Lagrangian function

$$\mathcal{L}(w, \lambda) := -J_0(w) + \sum_{i=1}^{m} \lambda_i (J_i(w) - D_i)$$

  - $\lambda = [\lambda_1, ..., \lambda_m]^\top$ is dual variable vector.

- Solve a minimax problem over Lagrangian function

$$\max_{\lambda \in \mathbb{R}_+^m} \min_{w} \mathcal{L}(w, \lambda)$$

  - Pro: guarantee converges to global optimal policy π^*
  - Con: Slow convergence rate, sensitive to hyper-parameter

- Motivation: propose an easy-to-implement SRL algorithm that has global optimality guarantee

Constraint-Rectified Policy Optimization (CRPO)

- CRPO update:
  - **Step 1 – Constraint Estimation:**
    - Estimate constraint function $\hat{J}_{i,t} \approx J_i(w_t)$ via policy evaluation
  - **Step 2 – Policy Optimization:**
    - If there exists $1 \leq i_t \leq m$ s.t. $\hat{J}_{i_t} \geq d_i + \eta \rightarrow$ minimize $J_{i_t}(\pi_{w_t})$
    - If exist multiple i_t, randomly choose one to minimize
    - If $\hat{J}_i \leq d_i + \eta$ for all $1 \leq i \leq m \rightarrow$ maximize $J_0(\pi_{w_t})$

- Key features:
  - Immediate response to constraint satisfaction/violation
  - No dual variable, easy to implement

Global Convergence of CRPO

Theorem (Tabular Setting)

With probability at least $1-\delta$, CRPO output satisfies

$$J_0(\pi^*) - \mathbb{E}[J_0(w_T)] \leq \Theta\left(\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^{1.5}\sqrt{T}}\right),$$

and for all $i = 1, \cdots, m$,

$$\mathbb{E}[J_i(w_T)] - D_i \leq \Theta\left(\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^{1.5}\sqrt{T}}\right).$$

- Both objective and cost converge at rate $\mathcal{O}(1/\sqrt{T})$
- This rate matches primal-dual approach (Ding et al. 2020)

Global Convergence of CRPO

Theorem (Function Approximation Setting)

With probability at least $1-\delta$, CRPO output satisfies

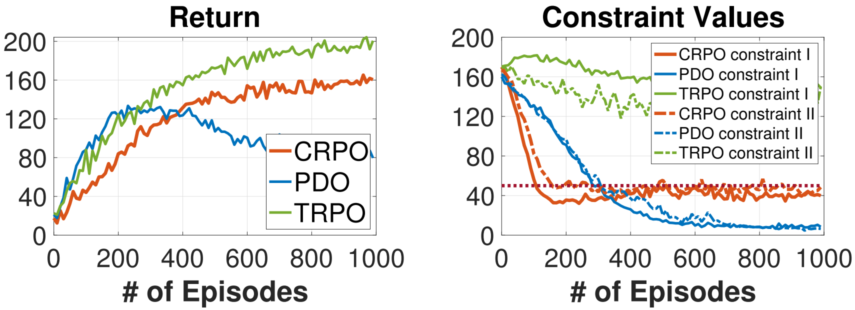
$$J_0(\pi^*) - \mathbb{E}[J_0(w_T)] \leq \Theta\left(\frac{1}{\sqrt{T}}\right) + \Theta\left(\varepsilon_{approx}\right),$$

and for all $i = 1, \cdots, m$,

$$\mathbb{E}[J_i(w_T)] - D_i \leq \Theta\left(\frac{1}{\sqrt{T}}\right) + \Theta\left(\varepsilon_{approx}\right).$$

- $\varepsilon_{\text{approx}}$ is introduced by function approximation
- This rate matches primal-dual approach (Ding et al. 2020)

Empirical Results



- Convergence
  - CRPO achieves much higher reward
- Constraint violation
  - CRPO drop below thresholds (and thus satisfy the constraints) much faster than that of PDO
  - CRPO tracks constraint thresholds almost exactly, which sufficiently explores boundary of feasible set to optimize reward
  - Primal-Dudal under-enforce constraints, and yields lower reward

- For more details about this work, please refere to CRPO: A New Approach for Safe Reinforcement Learning with Convergence Guarantee, T. Xu, Y. Liang, G. Lan, ICML 2021, https://https://arxiv.org/abs/2011.05869
- Feel free to contact me (xu.3260@osu.edu) for questions.

Thank You!