Safe Exploration in Reinforcement Learning: A Generalized Formulation and Algorithms

Akifumi Wachi[1], Wataru Hashimoto[2], Xun Shen[2], Kazumune Hashimoto[2]

[1]LINE Corp [2]Osaka University Correspondence: akifumi.wachi@linecorp.com

NEURAL INFORMATION PROCESSING SYSTEMS

Paper

Safe Reinforcement Learning (Safe RL)

Safe RL = reinforcement learning incorporating safety issues

Safe RL is typically formulated as a policy optimization problem under safety constraint(s)

$$\text{CMDP} \quad \mathcal{M} = \langle \mathcal{S}, \mathcal{A}, H, \mathcal{P}, r, g, s_1 \rangle$$

$\mathcal{S}$: State space $\mathcal{A}$: Action space H : Horizon $\mathcal{P}$: State transition probability

r : Reward function g : Safety function s_1 : Initial state

$$\max \quad \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, \pi, \rho \right] \quad \text{s.t.} \quad \text{Safety constraint}$$

Conventional RL objective

Two issues in safe RL research

1. There are various safety constraint formulations, and their relationships have not been sufficiently discussed
2. Applicability of resulting algorithms are low or unclear

Generalized Safe Exploration (GSE) Problem

Battery	Obstacle	Speed limit
Cumulative safety cost	State constraint	Instantaneous constraint
$\Pr\left[\sum_{h=1}^{H} \gamma_g^h g(s_h, a_h) \le \xi_1 \,\middle|\, \mathcal{P}, \pi \right] = 1$	$\mathbb{E}\left[\sum_{h=1}^{H} \gamma_g^h \mathbb{I}(s_h \in S_{\text{unsafe}}) \,\middle|\, \mathcal{P}, \pi \right] \le \xi_2$	$\Pr\left[g(s_h, a_h) \le \xi_3 \mid \mathcal{P}, \pi \right] = 1, \quad \forall h \in [1, H]$

Constraint in GSE problem

$$\Pr\left[g(s_h, a_h) \le b_h \mid \mathcal{P}, \pi \right] = 1, \quad \forall h \in [1, H]$$

Theorem 1 (informal)

Three common safe RL problems can be transformed into the GSE problem

Advantage 1

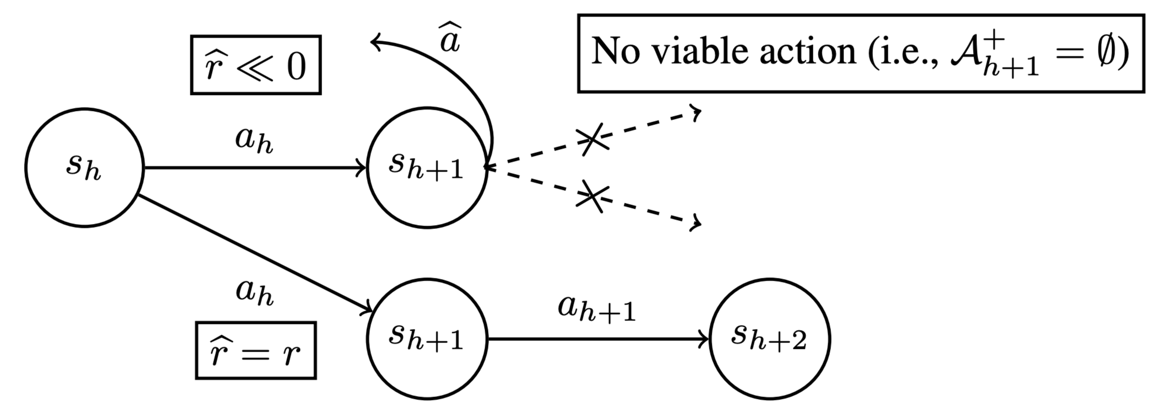
An algorithm for the GSE problem can also solve many safe RL problems

Advantage 2

The constraint of the GSE problem is instantaneous, which is easy to handle both theoretically and empirically

Meta Algorithm for Safe Exploration (MASE)

For solving the GSE problem, we propose MASE



Step 1: Construct an uncertainty quantifier Γ

$$|\, g(s, a) - \mu(s, a) \,| \le \Gamma(s, a), \quad \forall (s, a)$$

Step 2: Compute a set of (conservative) safe actions

$$\mathcal{A}_h^+ := \{\, a \in \mathcal{A} \mid \min\{1, \mu(s_h, a) + \Gamma(s_h, a)\} \le b_h \,\}$$

Step 3: Choose the next action from $\mathcal{A}_h^+$ and transit to s_{h+1}

Step 4: If $\mathcal{A}_{h+1}^+$ is empty, execute "emergency stop actions" and then transit to initial safe state. As a sacrifice, an agent receives a large penalty for (s_h, a_h) .

$$\widehat{r}(s_h, a_h) = \begin{cases} -c / \min_{a \in \mathcal{A}} \Gamma(s_{h+1}, a) & \text{if } \mathcal{A}_{h+1}^+ = \emptyset \\ r(s_h, a_h) & \text{otherwise,} \end{cases}$$

(c : A sufficiently large positive scaler)

Step 5: Optimize a policy in the following **unconstrained MDP using a standard RL algorithm**

$$\widehat{\mathcal{M}} := \langle \mathcal{S}, \{\mathcal{A}, \widehat{a}\}, H, \mathcal{P}, \widehat{r}, s_1 \rangle \qquad (\widehat{a} : \text{Emergency stop action})$$

Theoretical Results

Thanks to the simplicity of the algorithmic flow, MASE **provides theoretical guarantees on safety and optimality**

Safety guarantee

Theorem 2 (informal)

By constructing proper uncertainty quantifier, MASE guarantees safety with a high probability

We present two variants of MASE: one based on generalized linear models (GLMs) and the other based on GP

Near-optimality guarantee

Theorem 3 (informal)

Under proper assumptions, the optimal policy in $\widehat{\mathcal{M}}$ is identical to that in $\mathcal{M}$

Assumption: Generalized Linear CMDP (GL-CMDP)

The true Q-function and safety function are subject to GLMs with a known same feature mapping function

Theorem 3' (informal)

Under the GL-CMDP assumption, the policy obtained by MASE is guaranteed to be near-optimal

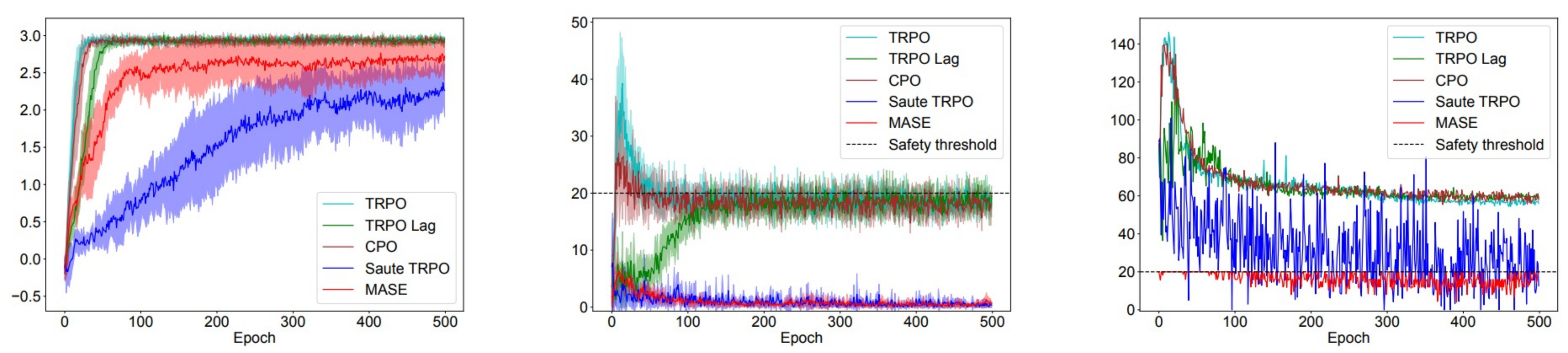
Experiments

Benchmark: Safety Gym

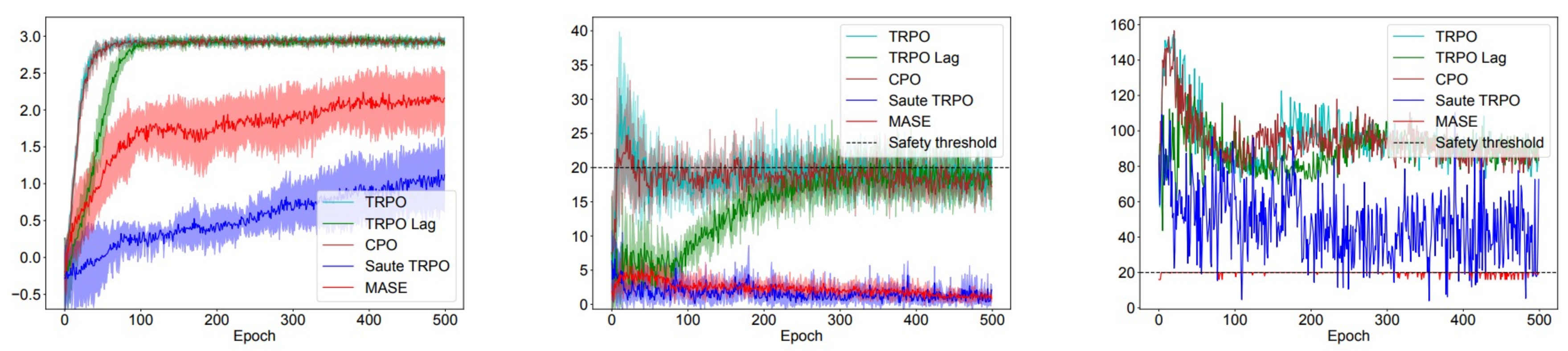
Baseline methods
TRPO, TRPO-Lagrangian, CPO, Saute RL

Metrics
- Average episode return (i.e., reward)
- Average episode safety
- Maximum episode safety

Results (Top: PointGoal1, Bottom : CarGoal1)
- MASE (proposed, red lines) learns a policy **without violating any safety constraint**
- Performance in terms of reward is slightly worse than baselines because exploration is prevented due to emergency stop actions



(a) Average episode return. (b) Average episode safety. (c) Maximum episode safety.



(d) Average episode return. (e) Average episode safety. (f) Maximum episode safety.