Introduction

Safety is an essential requirement for applying reinforcement learning (RL) in real applications. To guarantee safety during training, safe exploration problems have been actively studied. Typical RL objective

\[
\max V(s) = \mathbb{E} \left[ \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 \right]
\]

Subject to \( g(s) \geq b \)

Safety constraint

Previous work

A mainstream of safe exploration research is based on Gaussian process (GP).

- Train GP-based model using observations
- Allow an agent to visit only the states that are conservatively identified as safe.

- Theoretical guarantee (safety and optimality)
- Computational cost
- Strong assumptions (i.e., regularity)

Problem Formulation

Robots are equipped with sensors.

- Mars rover Perseverance: >10 cameras.
- Reasonable to assume that agents observe “feature vectors” for inferring safety.

We formulate a problem as safety-constrained Markov decision processes incorporating feature.

SPO-LF Algorithm

We are concerned about generalized linear models (GLMs)

Confidence intervals of reward and safety functions are summarized in the table below.

| Reward | Safety |
|--------|--------|
| \( \mu(\hat{r}_t) \pm \delta \) | \( \mu(\hat{s}_t) \pm \delta \) |
| \( \hat{r}_t \pm \delta \) | \( \hat{s}_t \pm \delta \) |

How does SPO-LF deal with safety?

- Visit only “safe” states such that the lower bound of safety function satisfies the constraint

How does SPO-LF maximize the cumulative reward?

- Follow the “optimistic in the face of uncertainty” principle by leveraging upper bound of reward function

Advantage: Unified Exploration

- An advantage of SPO-LF is that it is possible to explore reward and safety simultaneously
- If exploration and exploitation of reward are balanced, exploration of safety is also conducted
- Previous work based on GPs (Wachi and Sui, 2020) took a step-wise approach
- SPO-LF is more sample-efficient and simpler than GP-based methods

Experiments

Gym-MiniGrid

- SPO-LF achieves a near-optimal policy while satisfying safety constraints
- SPO-LF performs better than baselines in terms of sample efficiency and scalability

Safety-Gym

- In terms of reward, SPO-LF achieved comparable performance compared with advanced deep RL methods (e.g., CPO)
- SPO-LF did not execute even a single unsafe action

Summary

- New formulation via CMDPs with local feature.
- Proposed the SPO-LF algorithm for safely optimizing a policy in an a priori unknown environment.
- Theoretical guarantee on optimality and safety.
- Experimental advantages with code available.

Theory

Our paper provides two theorems.

Theorem 1 (Near-optimality)

SPO-LF achieves near-optimal policy after a sufficiently large number of time step with a high probability

Theorem 2 (Safety)

SPO-LF satisfies the safety constraint for every time step with a high probability