Safe Policy Optimization with Local Generalized Linear Function Approximations

Akifumi Wachi, Yunyue Wei, Yanan Sui

(akifumi.wachi@ibm.com, weiyy20@mails.tsinghua.edu.cn, ysui@tsinghua.edu.cn)

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	Safety constraint
maximize: $V^{\pi}_{\mathcal{M}}(s_t) = \mathbb{E}\left[\sum_{\tau=0}^{H} \gamma^{\tau} r(s_{t+\tau}) \mid \pi\right]$	subject to: $g(s_t) \geq h$

Previous work

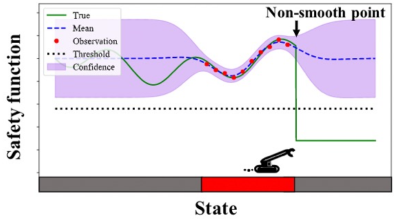
A mainstream of safe exploration research is based on Gaussian process (GP).

- 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)



If degree of safety drastically changes, GP-based safe exploration will fail

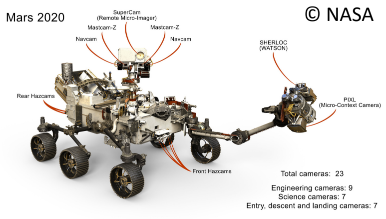
Fundamental problem of previous GP-based method.

1. Agent can observe only the current state.
2. No hint for inferring safety of the neighboring states.

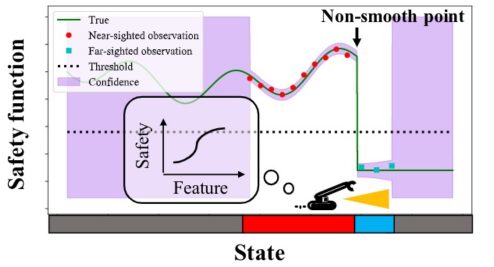
Problem Formulation

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.



: Feature ✓, Safety ✓ : Feature ✓, Safety ✗

: Feature ✗, Safety ✗

Near-sighted observation

- Reward, safety and feature vector are observed for the current state.

Far-sighted observation

- Only feature vectors are observed for visible states.

SPO-LF Algorithm

We are concerned about generalized linear models (GLMs)

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

	$s \in \Psi_t$ (FEATURE AVAILABLE)	$s \notin \Psi_t$ (FEATURE UNAVAILABLE)
REWARD	$[\mu(\phi_s^\top \tilde{\theta}_r) \pm \beta_r \cdot \|\phi_s\|_{W_t^{-1}}]$	$[0, \mu(\|\tilde{\theta}_r\|) \pm \beta_r \cdot \lambda_{\max}(W_t^{-1})]$
SAFETY	$[\mu(\phi_s^\top \tilde{\theta}_g) \pm \beta_g \cdot \|\phi_s\|_{W_t^{-1}}]$	$[0, \mu(\|\tilde{\theta}_g\|) + \beta_g \cdot \lambda_{\max}(W_t^{-1})]$

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, then 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**

Theory

Our paper provides two theorems.

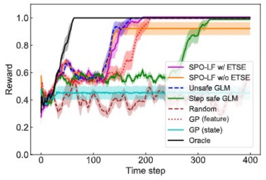
Theorem 1 (Near-optimality)
SPO-LF achieves near-optimal policy after a sufficiently large number of time step with a high probability

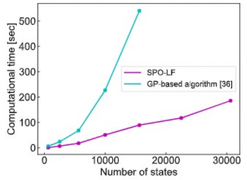
Theorem 2 (Safety)
SPO-LF satisfies the safety constraint for every time step with a high probability

Experiments

Gym-MiniGrid

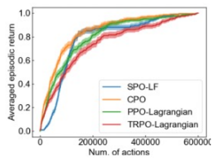
- SPO-LF achieves a near-optimal policy while satisfying safety constraints
- **SPO-LF performs better than baselines in terms of sample efficiency and scalability**

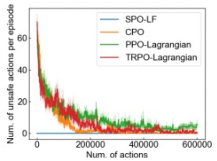




Safety-Gym

- 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.

OpenReview arXiv GitHub