IBM Research AI

Tsinghua University



Safe Policy Optimization with Local Generalized Linear Function Approximations

Akifumi Wachi
IBM Research AI

Yunyue Wei
Tsinghua University

Yanan Sui
Tsinghua University

NeurIPS, 2021

Background

- Safety is an essential requirement for applying reinforcement learning (RL) in real applications.
- In safety-critical systems where even a single mistake could lead to a catastrophic failure, conventional algorithms cannot be applied.
- To guarantee safety during training, safe exploration problems have been actively studied.

$$\text{maximize:} \quad V^{\pi}_{\mathcal{M}}(\boldsymbol{s}_t) = \mathbb{E}\left[\sum_{\tau=0}^{H} \gamma^{\tau} r(\boldsymbol{s}_{t+\tau}) \mid \pi\right] \qquad \text{subject to:} \quad g(\boldsymbol{s}_t) \geq h$$

Maximization of the cumulative reward
(Typical RL objective)

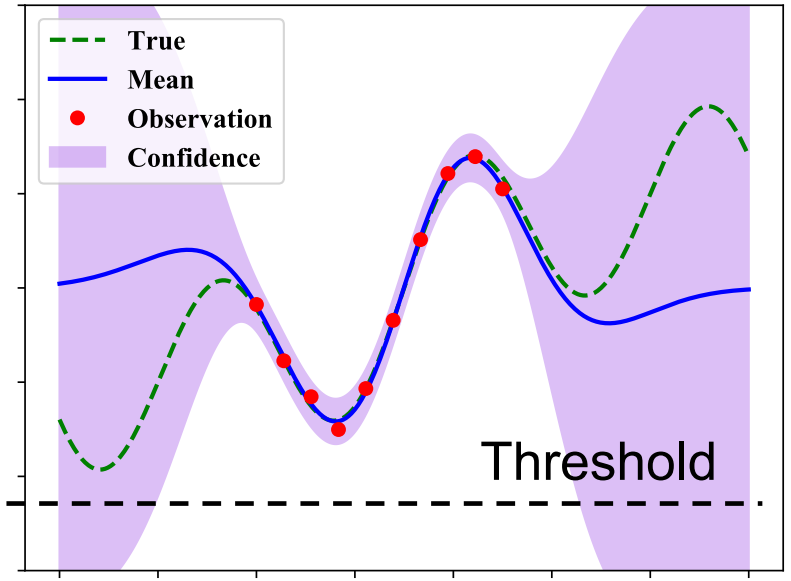
Safety constraint
(g is unknown a priori)

Previous Work: GP-based Safe Exploration

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

GP-based method basically:

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



☺ Theoretical guarantee on safety and near-optimality

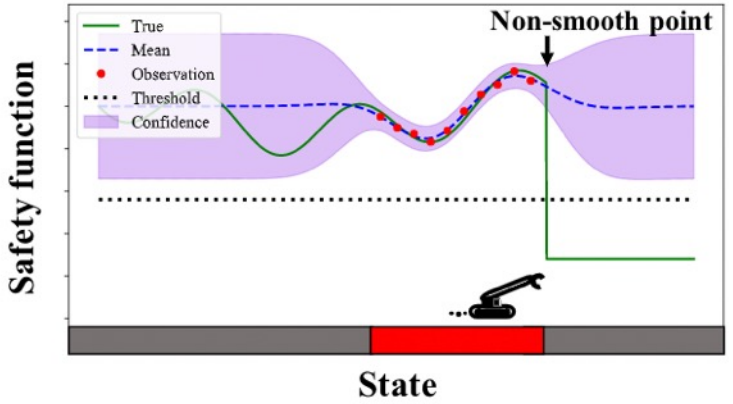
☹ Strong assumptions (i.e., regularity, Lipschitz continuity)

☹ Inconsistency between theoretical results and computational cost

Previous Work: GP-based Safe Exploration

☹ Strong assumptions (i.e., regularity, Lipschitz continuity)

- If degree of safety drastically changes, GP-based safe exploration will fail.
- Examples of non-smooth safety
  - Cliff, rock, presence and absence of pedestrians



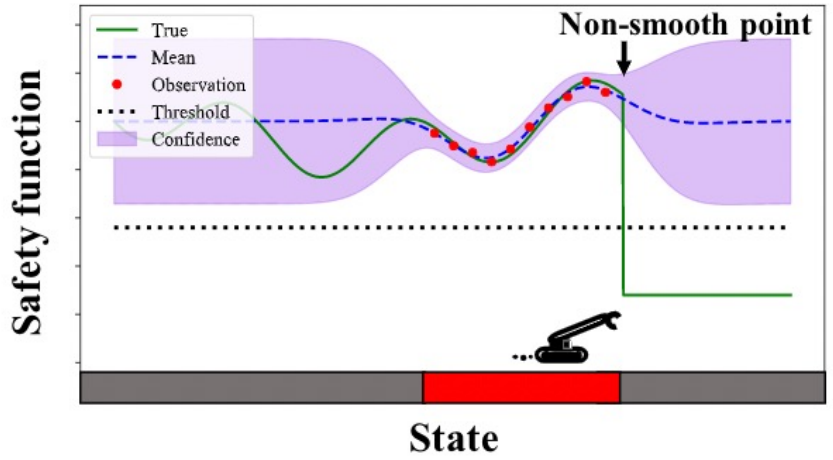
☹ Inconsistency between theoretical results and computational cost of GPs

- Previous work has proved
  - Completeness of the predicted safe region (e.g., Turchetta et al., 2016)
  - Optimality of the acquired policy (e.g., Wachi and Sui, 2020)
- Both theoretical results are achieved if the number of samples n is sufficiently large.
- Computational cost of GP is known to be large.
  - Naïve GP: $O(n^3)$

Previous Work: GP-based Safe Exploration

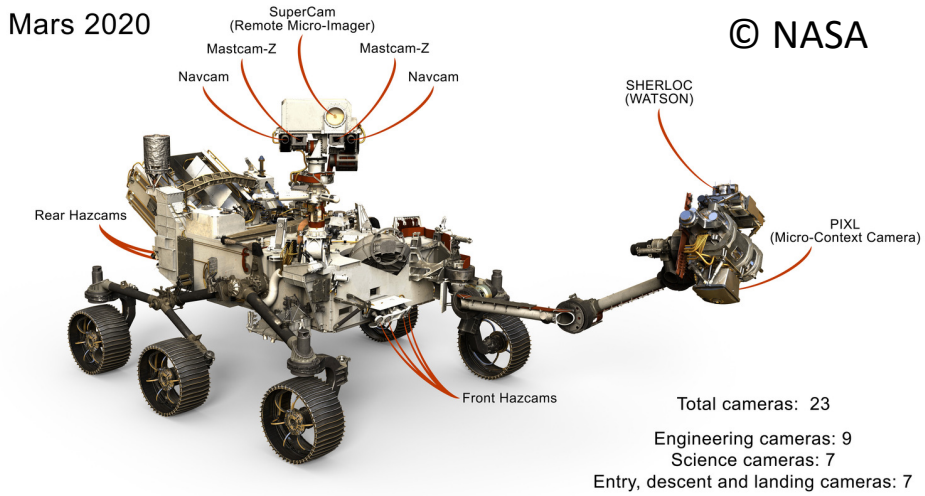
What is the fundamental problem?

1. It is assumed that an agent can observe only the current state.
2. No hint is provided for inferring safety of the neighboring states
3. It is necessary to assume the function structure (i.e., regularity, Lipschitz continuity)



Robots are equipped with sensors

- For example, Mars rover *Perseverance* has 16 cameras.
- It is reasonable to assume that an agent can observe "feature vectors" for inferring the degree of safety



Problem Statement

Safety-constrained Markov decision processes (MDPs) incorporating feature

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, f, H, r, g, \phi, \psi \rangle$$

S : finite state space A : finite action space $f(\cdot,\cdot)$: deterministic transition

H : horizon r : reward function g : safety function

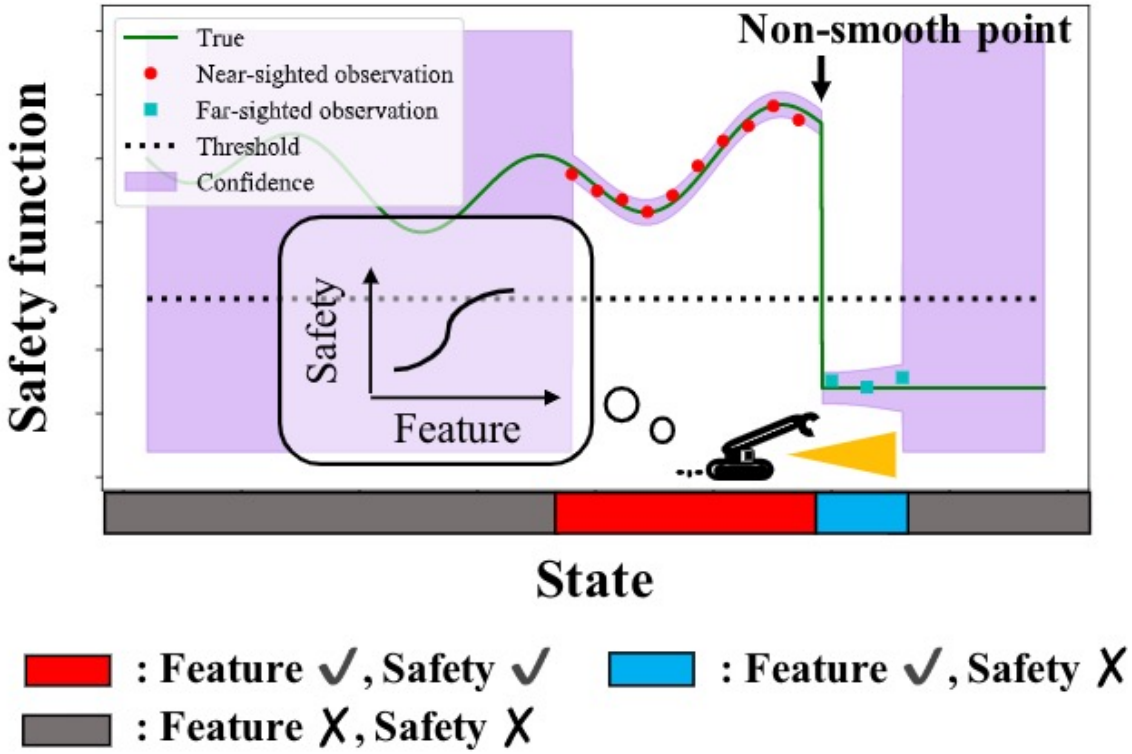
ϕ : feature mapping function ψ : function returning a set of states within sensor range

Reward function r, safety function g, and feature mapping function ϕ are unknown a priori

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

Problem Statement

- We introduce two notions of observations.
  - 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



- Agent basically needs to
  1. learn the relationship between feature and reward/safety functions via near-sighted observations
  2. infer the reward and safety values using feature vectors obtained by far-sighted observation

Our Contributions

- Formulate a new safe exploration problem incorporating locally-available feature
- Propose Safe Policy Optimization with Local Feature (SPO-LF) algorithm
- Prove that, with high probability, the SPO-LF will
  - achieve a near-optimal policy
  - satisfy safety constraints at every time step
- Evaluate our algorithm in two experiments.

	SPO-LF	**GP-based methods**	**Advanced deep RL methods**
Safety guarantee	✓	✓	✗
Near-optimality guarantee	✓	✓	✗
Scalability	Medium	Low	High
Sample complexity	Small	Medium	Large
Notable algorithms	-	SNO-MDP (Wachi and Sui, 2020)	CPO (Achiam et al., 2017)

SPO-LF Algorithm

- We are concerned about generalized linear models (GLMs)
- Confidence intervals of reward and safety functions are summarized in the right table.

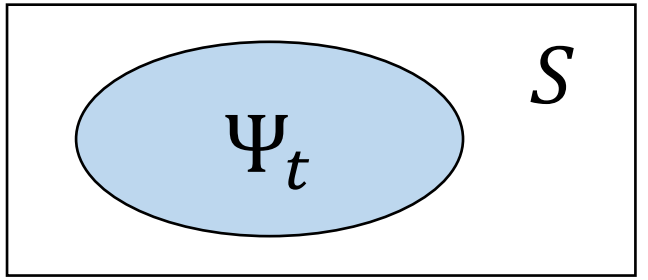
	$s \in \Psi_t$ (FEATURE AVAILABLE)	$s \notin \Psi_t$ (FEATURE UNAVAILABLE)
REWARD	$[\mu(\phi_s^\top \tilde{\theta}_r) \pm \beta_r \cdot \|\phi_s\|_{W_t^{-1}}]$	$[0, \mu(\|\tilde{\theta}_r\|) \pm \beta_r \cdot \lambda_{\max}(W_t^{-1})]$
SAFETY	$[\mu(\phi_s^\top \tilde{\theta}_g) \pm \beta_g \cdot \|\phi_s\|_{W_t^{-1}}]$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

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



Set of states for which an agent has observed feature vector

Unified Exploration

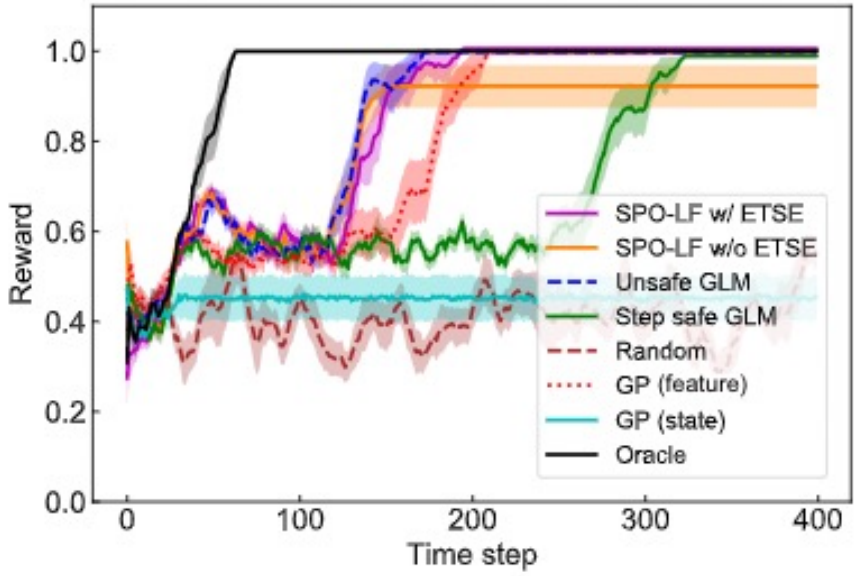
- Previous work based on GPs (Wachi and Sui, 2020) took a step-wise approach
  - Step-1: Exploration of safety
  - Step-2: Exploration and exploitation of reward
- An advantage of SPO-LF is that it is possible to explore reward and safety simultaneously
- If exploration and exploitation of reward are balanced, then exploration of safety is also conducted
- SPO-LF is more sample-efficient and simpler than GP-based methods

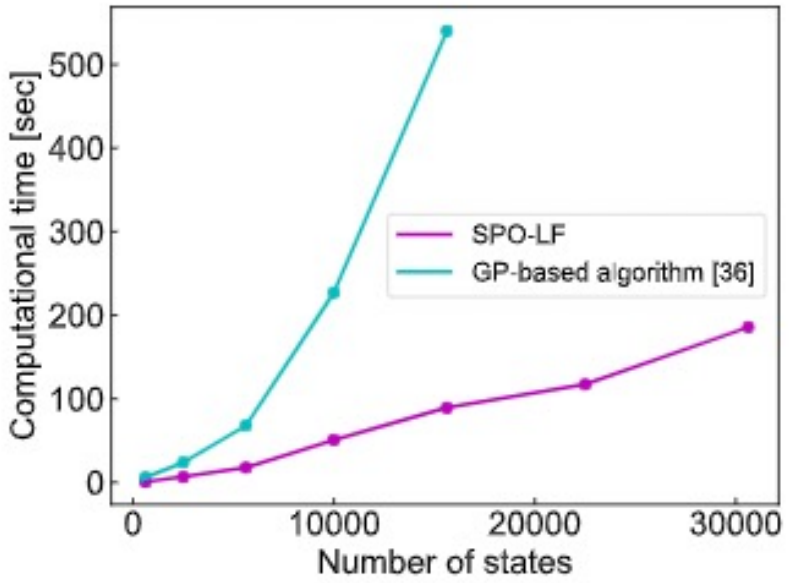
	$s \in \Psi_t$ (FEATURE AVAILABLE)	$s \notin \Psi_t$ (FEATURE UNAVAILABLE)
REWARD	$[\mu(\phi_s^\top \tilde{\theta}_r) \pm \beta_r \cdot \|\phi_s\|_{W_t^{-1}}]$	$[0, \mu(\|\tilde{\theta}_r\|) \pm \beta_r \lambda_{\max}(W_t^{-1})]$
SAFETY	$[\mu(\phi_s^\top \tilde{\theta}_g) \pm \beta_g \cdot \|\phi_s\|_{W_t^{-1}}]$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

For both reward and safety, efficiency of exploration depends on the same term

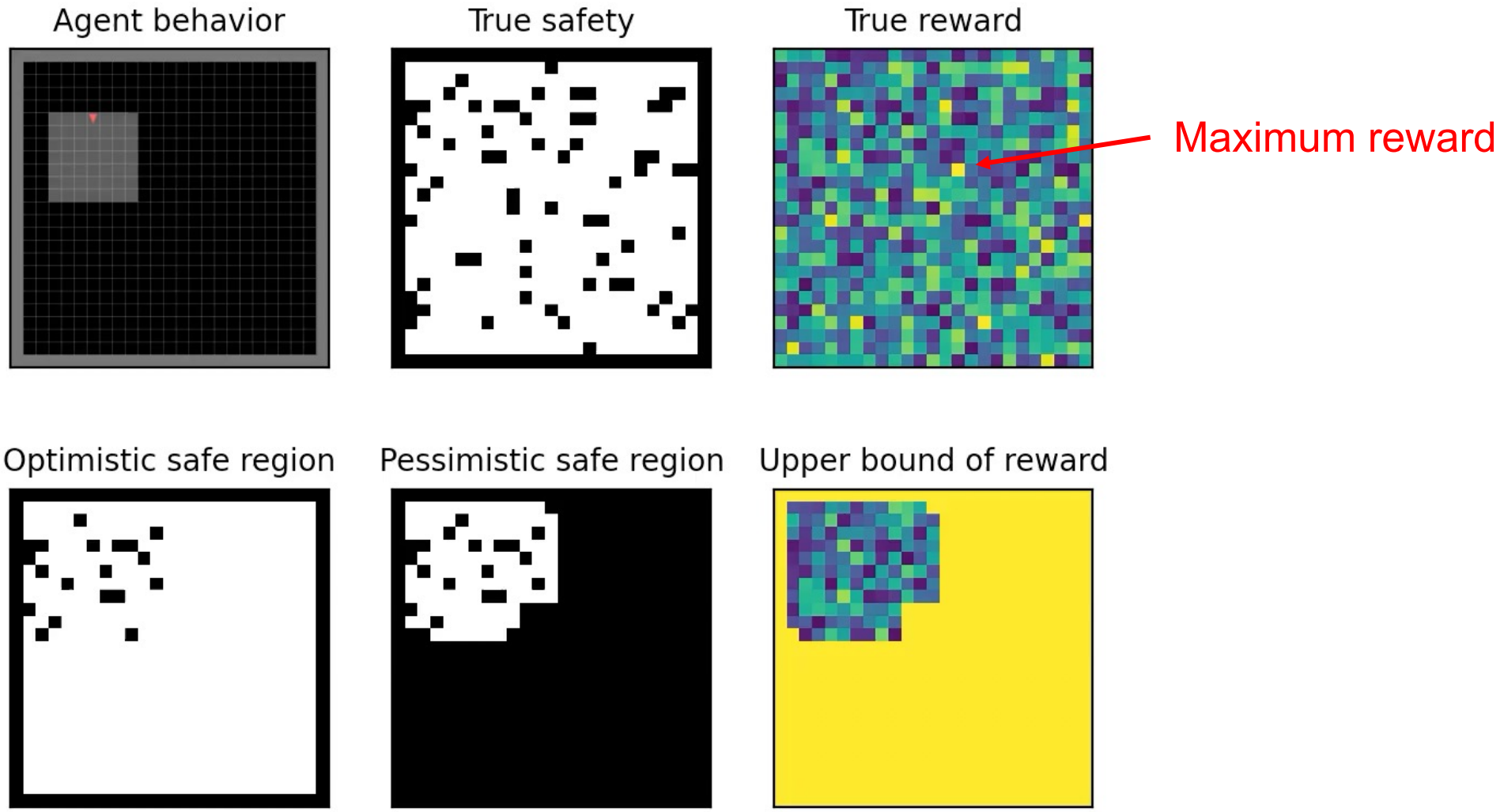
Experiments (Gym-MiniGrid)

- We first evaluate our SPO-LF in Gym-MiniGrid
- SPO-LF achieves a near-optimal policy while satisfying safety constraints
- SPO-LF performs better than baselines in terms of
  - Sample efficiency
  - Scalability
- Source-code: https://github.com/akifumi-wachi-4/spolf



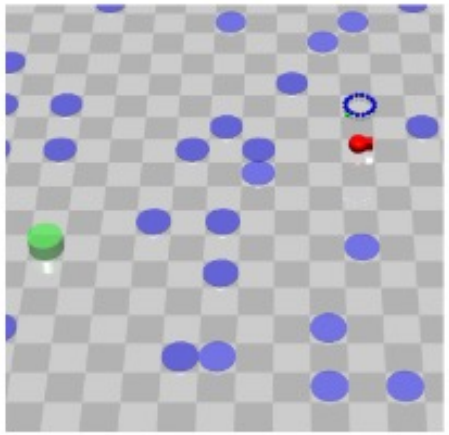


How Does the SPO-LF Agent Behave?

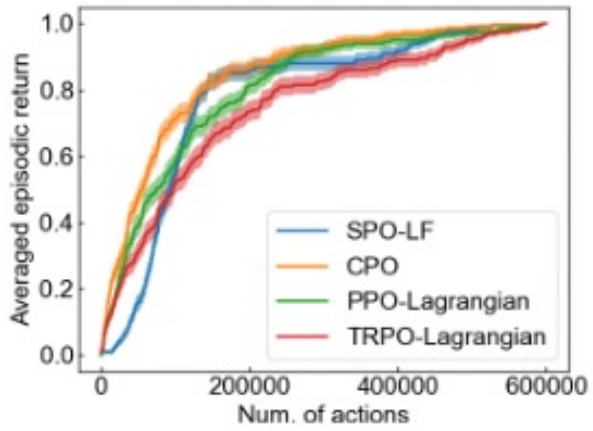


Experiments (Safety-Gym)

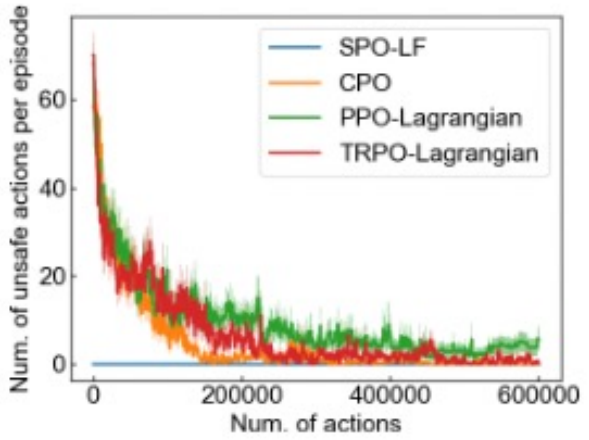
- We then evaluate our SPO-LF in Safety-Gym
- In terms of reward, SPO-LF achieved comparable performance compared with advanced deep RL methods (e.g., CPO)
- SPO-LF did NOT execute even a single unsafe action



(a) Environment



(b) Reward



(c) Number of unsafe actions

Conclusion

- Formulated **safety-constrained MDPs incorporating local features**
- Proposed **SPO-LF based on GLMs** that simultaneously balances three-way tradeoff (exploration of reward and safety, exploitation of reward)
- Theoretically, we proved a bound of the sample complexity to achieve **near-optimal policy while guaranteeing safety, with high probability**
- Showed that our SPO-LF
  1. achieved better efficiency and scalability than previous safe exploration methods with theoretical guarantees
  2. behaved more safely than existing advanced deep RL methods with constraints.

Thank you!

Akifumi Wachi

IBM Research AI

Yunyue Wei

Tsinghua University

Yanan Sui

Tsinghua University