Safe Reinforcement Learning in Constrained Markov Decision Processes

Akifumi Wachi
IBM Research AI

Yanan Sui
Tsinghua University

ICML | 2020
Thirty-seventh International
Conference on Machine Learning

Overview

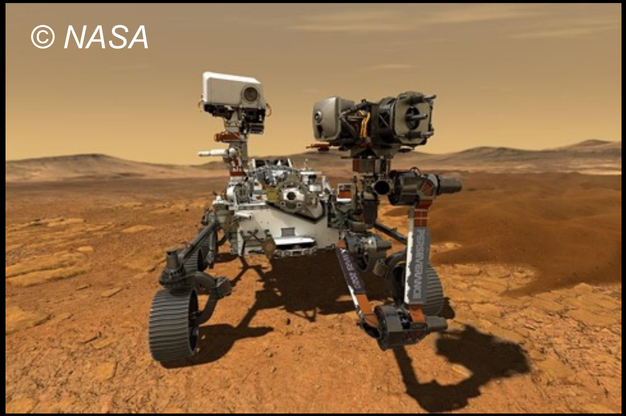
Details are in page 11 – 18.

Background

There is increasing need for automated exploration of the *unknown* environment

- *Unknown where is safe/unsafe*
- *Unknown where is scientifically worthwhile to visit*

An agent needs to maximize the cumulative reward while guaranteeing safety.



Problem Statement

We consider a safety-constrained Markov Decision Processes (MDPs).

$$\mathcal{M} = \langle S, A, f, r, g, \gamma \rangle$$

S : finite state space

A : finite action space

$f(\cdot,\cdot)$: deterministic transition

$\boldsymbol{r}(\cdot)$: reward function

$g(\cdot)$: safety function

γ : discount factor

Problem Formulation

$$\max \quad \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^{\tau} r(s_{t+\tau})\right]$$

$$\text{subject to} \quad g(s_{t+\tau}) \geq h, \quad \forall \tau = [0, \infty]$$

Problem Statement

- Both reward function r and safety function g are unknown a priori.
- It is intractable to solve this problem without further assumptions.

$$\max \quad \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^{\tau} r(s_{t+\tau})\right]$$

$$\text{subject to} \quad g(s_{t+\tau}) \geq h, \quad \forall \tau = [0, \infty]$$

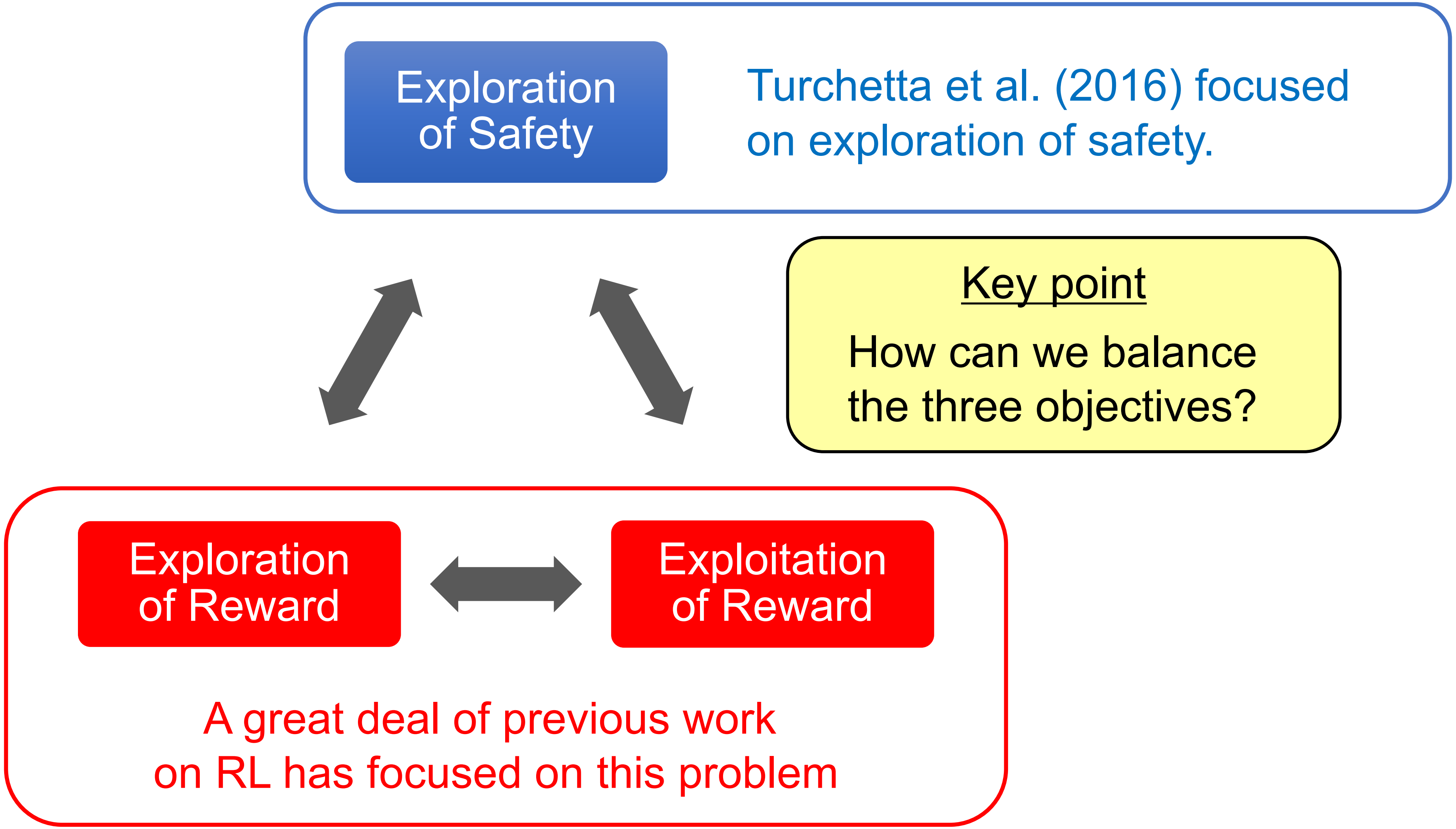
⇩

- We adapt two assumptions from Sui et al. (2015) and Turchetta et al. (2016).
  - **Assumption 1**. Agent starts in a set of safe states, which is known to be safe.
  - **Assumption 2**. Reward and safety functions exhibit regularity.
    - ➢ We model them using Gaussian Processes.

Sui et al., A. Safe exploration for optimization with Gaussian processes. In ICML, 2015.

Turchetta et al., A. Safe exploration in finite Markov decision processes with Gaussian processes. In NeurIPS, 2016.

Exploration and Exploitation



Our Main Contribution

Wachi et al. "Safe Exploration and Optimization of Constrained MDPs using Gaussian Processes." AAAI 2018.

- Safety: probabilistic guarantee
- Optimality: no guarantee

⇩

This paper: Safe Near-optimal MDP (SNO-MDP)

Wachi and Sui, “Safe Reinforcement Learning in Constrained Markov Decision Processes.” ICML 2020.

- Safety: probabilistic guarantee
- Optimality: probabilistic guarantee

Step-wise Approach

Wachi et al. (2018) tried to solve the three-way trade-off simultaneously.

⇩

This work takes a **step-wise approach**.

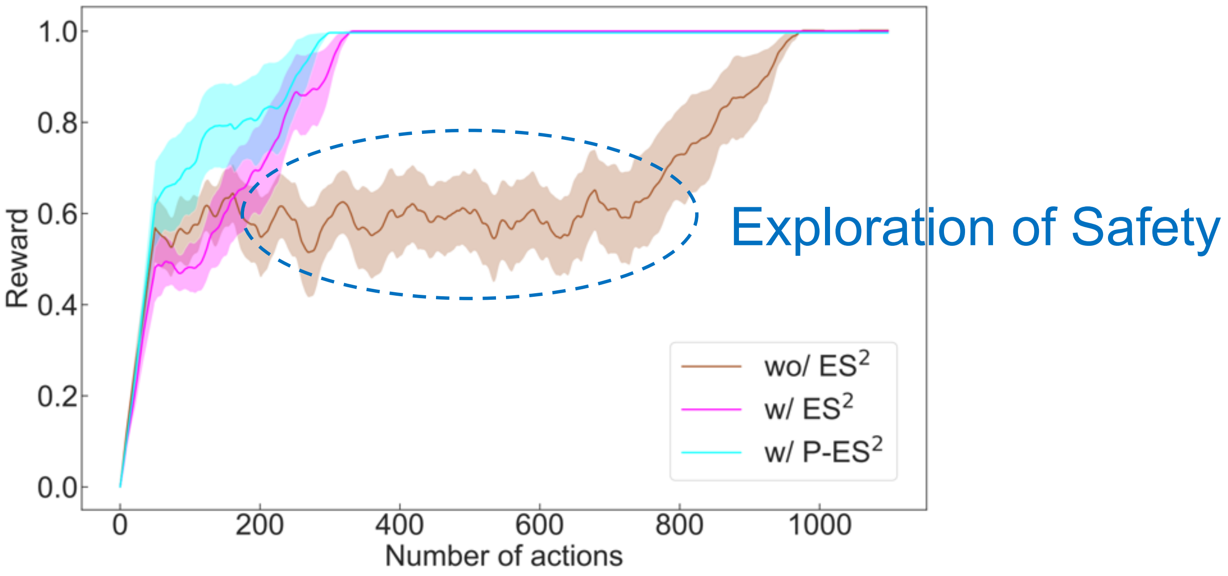
1. Exploration of safety.
2. Optimization of the cumulative reward in the certified safe region.

Intuitions. *Suppose an agent can sufficiently expand the safe region. Then, the agent only has to optimize the cumulative reward in the certified safe region.*

Early Stopping of Exploration of Safety

- Pure step-wise approach (brown line) has an issue.
  - Spend much time for the exploration of safety.

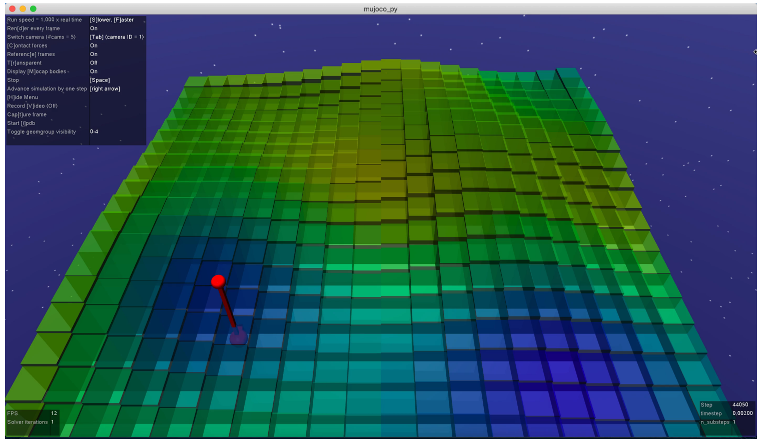
- We additionally proposed early stopping of exploration of safety (ES^2).
  - ES^2 maintains the theoretical guarantees on near-optimality.
  - $P\text{-}ES^2$ empirically performs better than ES^2.



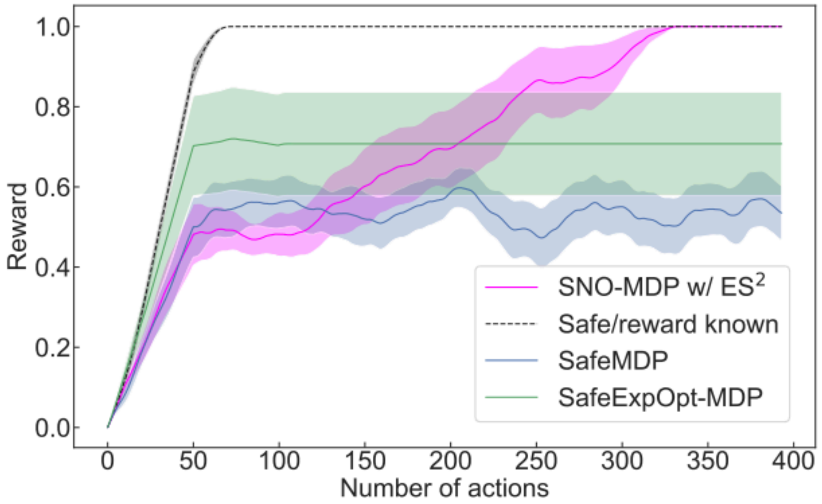
Experiments

- Developed a new simulation environment, called GP-Safety-Gym, which is based on Open AI SafetyGym (Ray et al., 2019).

- Achieved better empirical performance than other baselines.
  - SafeMDP (Turchetta et al., 2016)
  - SafeExpOpt-MDP (Wachi et al., 2018)



Reward (high: yellow, low: blue)
Safety: height



Details

Overview of SNO-MDP

Step 1: Exploration of Safety (Turchetta et al., 2016)

loop

- Step 1-1: Update of GPs based on the observations
- Step 1-2: Conventional terminal condition

This terminal condition is related to the “completeness” of the safe region.

Step 2: Exploration and Exploitation of Reward

Overview of SNO-MDP with ES^2

Step 1: Exploration of Safety

loop

Step 1-1: Update of GPs based on the observations

Step 1-2: Conventional terminal condition

Step 1-3: ES^2 algorithm

If ES^2 algorithm identifies that "exploration of safety" can be stopped, then break the loop.

Step 2: Exploration and Exploitation of Reward

Step 1: Exploration of Safety

To expand the safe region, we use the same scheme as in Turchetta et al. (2016).

1) Probabilistic Safety Guarantee

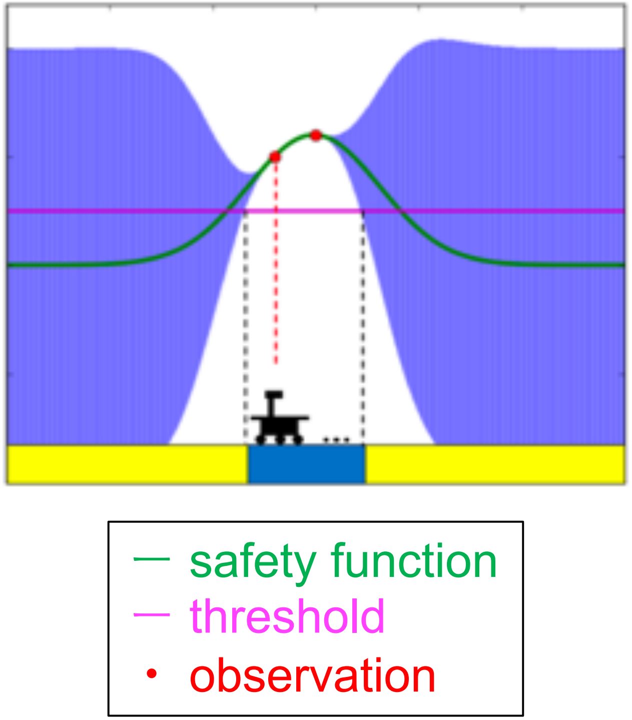
Original safety constraint: $g(s) \geq h$

$$\Downarrow$$

If a state s satisfies the following condition, safety is guaranteed with high probability.

$$\boxed{l(s)} \geq h$$

Lower bound of g inferred by GP.



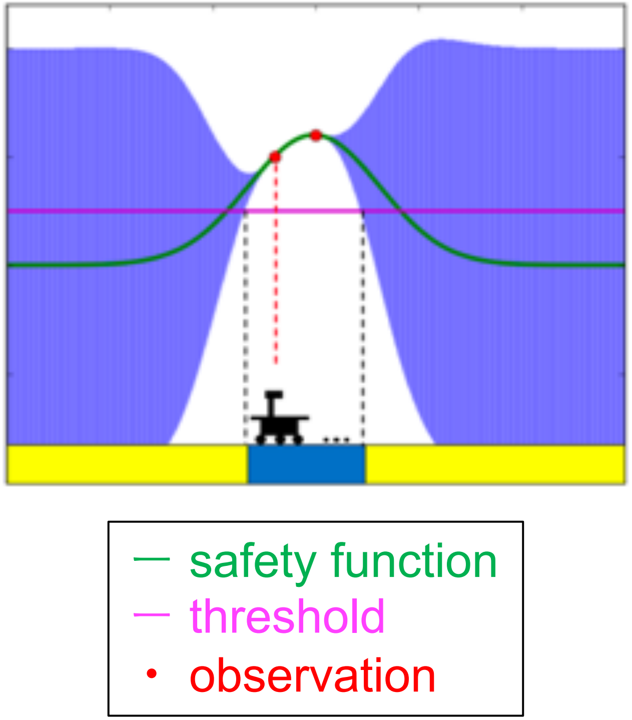
Step 1: Exploration of Safety

2) Expansion of Safe Region

- The *efficiency* of expanding the safe region is measured by the width of the safety function's confidence interval.

$$w(s) = u(s) - l(s)$$

- Sample the next state with the maximum w within the safe space.



3) Conventional Terminal Condition of Exploration of Safety

- The previous work (Sui et al., 2015; Turchetta et al., 2016) terminated the exploration if the following equation holds for all states in safe space.

$$\max w(s) \leq \epsilon_g$$

Step 2: Optimization of Reward

- All the agent has to do is optimize the cumulative reward in the certified safe region.
- Leverage algorithms for optimizing unconstrained MDPs.
- A simple approach is to follow *the optimism in the face of uncertainty* principle.

Probabilistic upper confidence bound of reward

$$J(s_t) = \max_{\boxed{s_{t+1} \in \mathcal{X}^-}} \left[\boxed{U_t(s_{t+1})} + \gamma J(s_{t+1}) \right]$$

Next state must be in pessimistically identified safe space

ES^2 algorithm

- Consider a new MDP, where reward function is defined as in the figure below.
  - Reward is set to be the **lower bound in the currently identified safe region.**
  - **Otherwise, set to be the upper bound.**

- This reward settings encourage the agent to explore outsides the currently identified safe region.

- Suppose the set of next states that the agent will visit based on the optimal policy is **a *subset* of the currently identified safe region**

⇩

- we can stop exploring the safety function.

Conclusion

- We have proposed **SNO-MDP**, a stepwise approach for exploring and optimizing a safety-constrained MDP.
- Theoretically, we proved a bound of the sample complexity to achieve **near-optimal policy while guaranteeing safety, with high probability**.
- We also proposed the **ES^2 algorithm** for improving the efficiency in obtaining rewards.
- We developed **GP-SAFETY-GYM** to test the effectiveness of SNO-MDP.
- Our proposed algorithm, SNO-MDP overperforms other baselines.

Thank you!