ICML | 2020

Safe Reinforcement Learning in Constrained Markov Decision Processes

IBM Research AI Akifumi Wachi, Yanan Sui

Tsinghua University

Introduction

Conventional reinforcement learning (RL) literature has pursued efficiency and optimality of the cumulative reward.

When RL algorithms are applied to real systems, **safety** is an essential requirement.

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

- Both reward and safety are unknown *a priori.*
- Our objective is to maximize the cumulative reward while guarantee safety.

Method

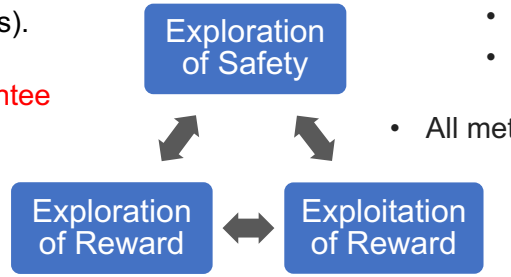
This problem requires an agent to balance 1) exploration of safety, 2) exploration of reward, and 3) exploitation reward.

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

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

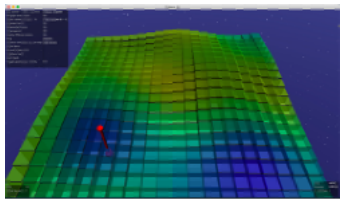
Our proposed algorithm Safe Near-Optimal MDP (SNO-MDP) provide **Theoretical Guarantees on both Safety and Near-Optimality.**

We also proposed **Early-Stopping of Exploration of Safety (ES^2)** algorithm for faster convergence.

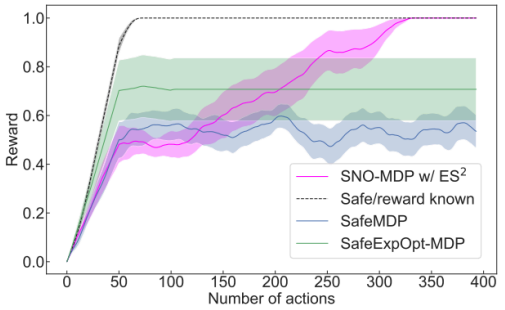


Experiment

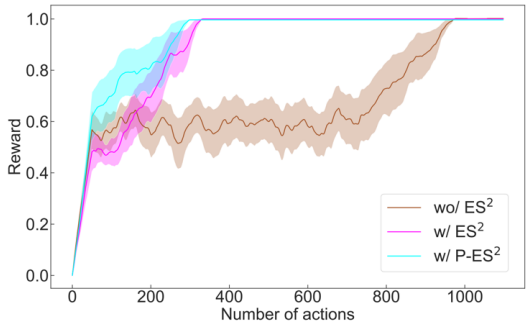
- Developed a new **simulation environment** called **GP-Safety-Gym** based on Open AI Safety Gym.
- **Achieved better empirical performance** than other baselines.
  - SafeMDP (Turchetta et al., 2016)
  - SafeExpOpt-MDP (Wachi et al., 2018)
- All methods, including the baselines, did not take any unsafe actions.



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



Performance comparison of SNO-MDP and other baselines



Effect of ES^2 algorithm
($P\text{-}ES^2$ is a practical version of ES^2)

Paper and Code at
www.yanansui.com