Safe Reinforcement Learning Using Advantage-Based Intervention

Nolan Wagener, Byron Boots, Ching-An Cheng

Georgia Tech Robotics & Intelligent Machines

Paul G. Allen School W

Microsoft Research

Motivation

- Reinforcement learning in real world requires **safety** during both training and deployment.
- Safe RL approaches are typically unsafe during training or need external safety mechanism forever.
- We desire:
  - Safety during training
  - Safety at deployment
  - High reward at deployment

Some Notation

- Reward function $r(s, a)$ is non-negative.
- Safety cost function $c(s, a) = \mathbf{1}\{s \in \text{unsafe set}\}$.
- Value function for reward $V^{\pi}(s)$ and cost $\bar{V}^{\pi}(s)$.

- Objective: Maximize return from s_0 while keeping cost below some threshold.

$$\max_{\pi} V^{\pi}(s_0) \quad \text{subject to} \quad \overline{V}^{\pi}(s_0) \leq \delta$$

Prior methods

Constrained RL

Solve a constrained optimization problem. To ensure safety, optimize a penalty for safety violation along with the RL policy.

+ Directly solves for problem of interest*
+ Good performance and safety at deployment*

- Not safe during training
- Optimization may be unstable

Intervention

Wrap a safety layer around RL policy so agent doesn't take unsafe actions. Then, run an RL algorithm on the induced unconstrained MDP.

+ Safe during training
+ We solve an unconstrained problem

- No guarantees on performance or safety of RL policy at deployment

Our approach (SAILR) provides the best of both paradigms.

*With some assumptions on the optimal policy

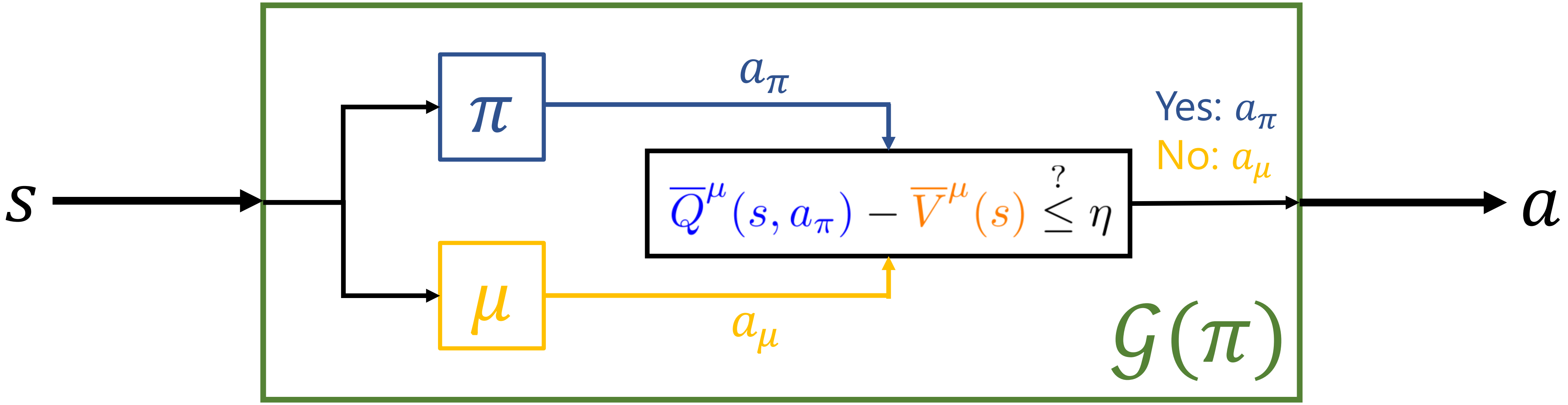
Our approach: SAILR (Intervention)

We assume access to a **baseline policy** μ that is safe starting from the initial state.

Intervention rule $\mathcal{G}$ defined by baseline μ and advantage threshold η.

Given **RL policy** π, construct **shielded policy** $\mathcal{G}(\pi)$ for exploration.

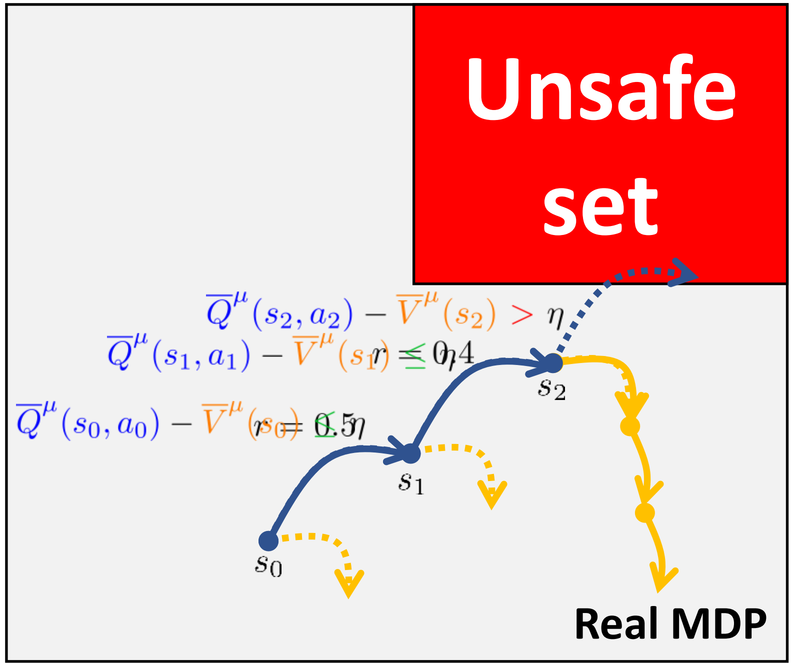
How we sample actions from $\mathcal{G}(\pi)$:

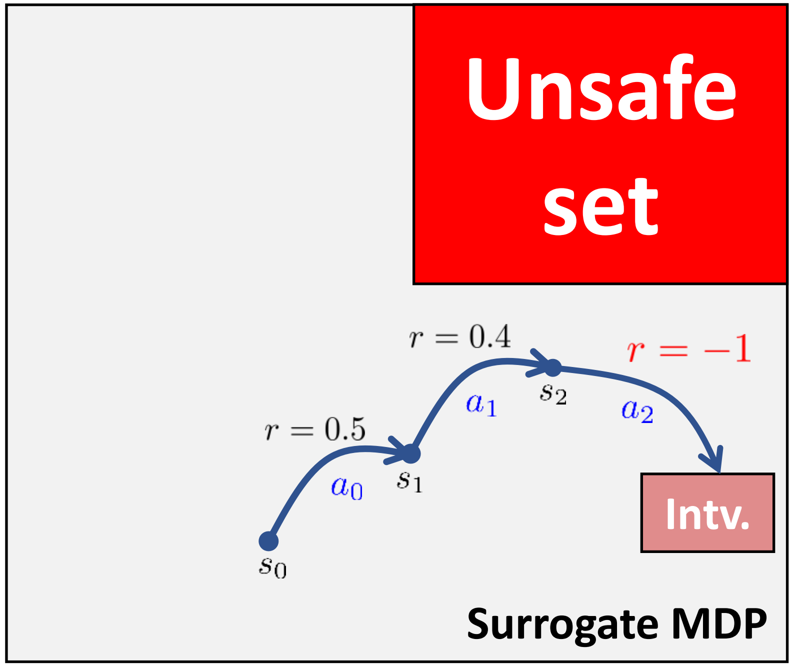


Our approach: SAILR (Algorithm)

$\mathcal{G}(\pi)$ runs in real MDP. π observes it's running in a **surrogate MDP**.

SAILR: Run unconstrained RL algorithm (e.g., PPO) in surrogate MDP.





Theoretical Guarantees

μ: Baseline policy
η: Threshold for intervention
$\mathcal{G}(\pi)$: Shielded policy for exploration

V^{π}: value function for reward
$\bar{V}^{\pi}$: value function for safety cost
π^*: Optimal policy for safety-constrained problem
$\hat{\pi}^*$: Optimized policy returned by SAILR

Safety During Training

$$\overline{V}^{\mathcal{G}(\pi)}(s_0) \leq \overline{V}^{\mu}(s_0) + \frac{\eta}{1-\gamma}$$

Shielded policy is roughly as safe as baseline policy

Safety at Deployment (Without Intervention!)

$$\overline{V}^{\hat{\pi}^*}(s_0) \leq \overline{V}^{\mu}(s_0) + \frac{\eta}{1-\gamma}$$

Optimized policy is roughly as safe as baseline policy

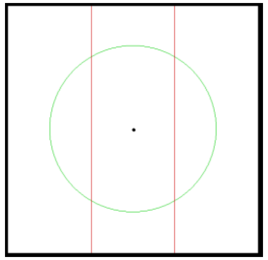
Return at Deployment (Without Intervention!)

$$V^{\pi^*}(s_0) - V^{\hat{\pi}^*}(s_0) \leq O\left(\frac{\text{Prob}(\pi^* \text{ is intervened by } \mathcal{G})}{1-\gamma}\right)$$

Suboptimality determined by how likely π^* would be intervened

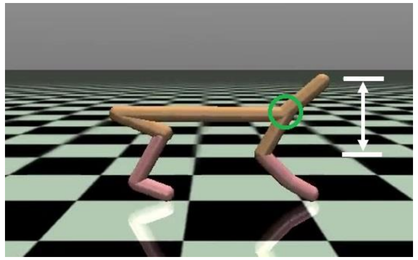
Experiments

Point Robot



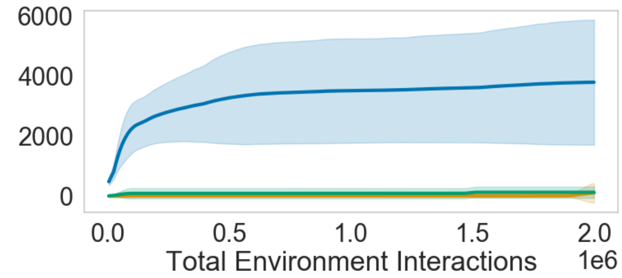
μ: Deceleration policy

Half-Cheetah

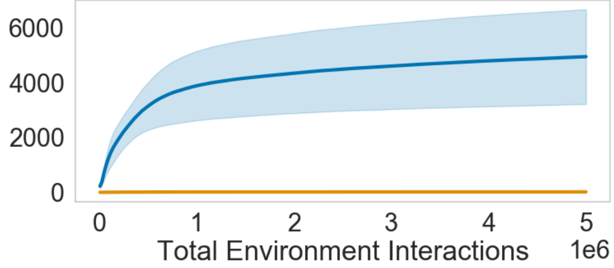


μ : Model predictive control

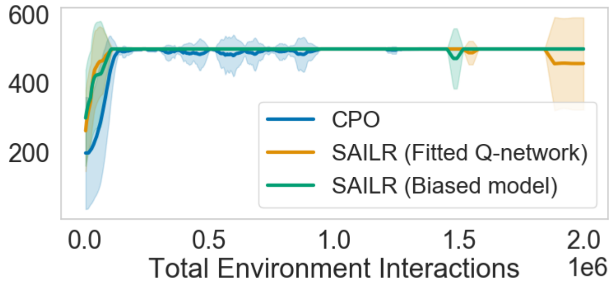
Safety violations during training



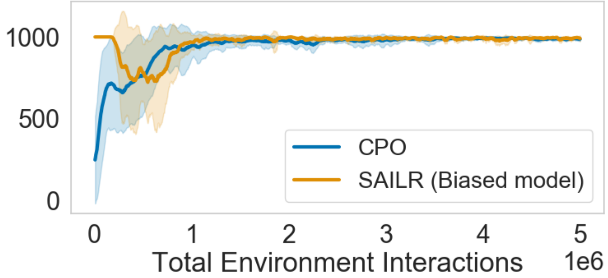
Far fewer safety violations during training



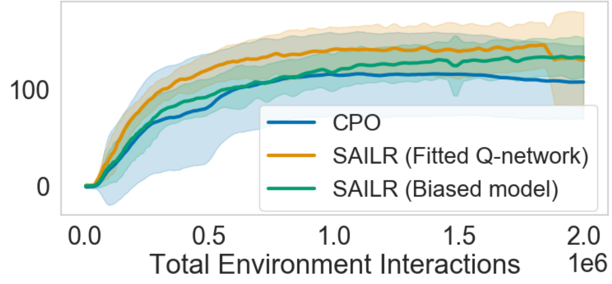
Episode length at deployment



Similar level of safety at deployment



Episode return at deployment



Similar returns at deployment