Sampling-Based Safe Reinforcement Learning

Luca Vignola, Bruce D. Lee, Manish Prajapat, Manuel Wendl, Melanie Zeilinger, Andreas Krause, Yarden As
ETH Zurich

Constrained Markov decision process

Finite horizon and unknown dynamics

$$x_{t+1} = f^\star(x_t, a_t) + w_t, \qquad w_t \sim \mathcal{N}(0, \sigma_w^2 I).$$

Maximize expected reward

$$\max_{\pi \in \Pi} J_r(\pi, f^\star) = \max_{\pi \in \Pi} \mathbb{E}_{\tau_\pi^{f^\star}} \left[\sum_{t=0}^{T-1} r(x_t, a_t) \right]$$

Safety at every episode

$$J_c(\pi_n, f^\star) = \mathbb{E}_{\tau_{\pi_n}^{f^\star}} \left[\sum_{t=0}^{T-1} c(x_t, a_t) \right] \leq d \qquad \forall n$$

The expectation is over the initial state and process noise under policy π and dynamics $f^\star$.

Learning loop with a Gaussian process model

Dynamics model

Assume RKHS dynamics and maintain a GP posterior with a calibrated confidence set $\mathcal{F}_n$ containing $f^\star$ with high probability.

At each episode n

1. **Select** a policy π_n that maximizes predicted reward subject to the safety and exploration constraints.
2. **Roll out** π_n to collect a trajectory.
3. **Update** the GP posterior and refine the dynamics samples before the next episode.

Safety across finitely many models

At episode n, the confidence set $\mathcal{F}_n$ contains all plausible dynamics.

Worst-case check

Maximize the policy's expected cost over the full confidence set.

$$\max_{f \in \mathcal{F}_n} J_c(\pi, f) \leq d$$

This search is generally intractable.

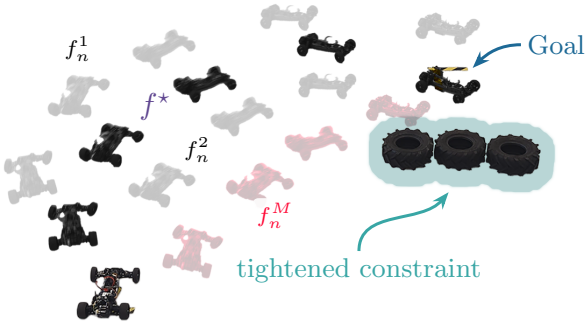
Finite checks in SBSRL

Require the policy to pass the safety check on every sampled model.

$$\max_{m \in [M]} J_c(\pi, f_n^m) \leq d - \Delta_\zeta$$

M is finite and Δ_ζ is a safety margin.

Illustration of the sampled safety constraint



Different model predictions for the same policy.

Infeasible policy

A sampled model predicts a violation (red). This policy lies outside the feasible set of SBSRL's constrained optimization.

Gaussian process models

With enough samples, at least one is close to the true dynamics with high probability.

Exploration as a constraint

Optimize over policies that pass the safety check. The quantity J_{s_n} is the expected cumulative model uncertainty along the predicted trajectory.

Exploration bonus

Add an uncertainty bonus to the reward.

$$\max_{\pi} \; J_r(\pi, \mu_n) + \lambda_n J_{s_n}(\pi, \mu_n)$$

The bonus weight λ_n can be hard to tune.

SBSRL exploration constraint

Require a minimum level of uncertainty.

$$\max_{\pi} \; J_r(\pi, \mu_n) \quad \text{s.t.} \quad J_{s_n}(\pi, \mu_n) \geq d_\sigma^n$$

Then maximize the expected reward.

Stopping rule Stop exploring when no safe policy meets the threshold.

Guarantees for Gaussian process models

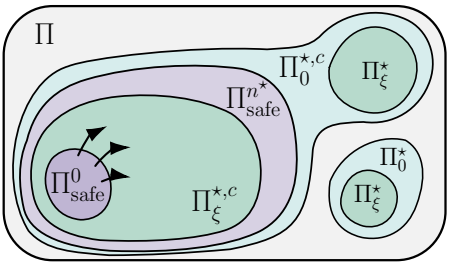
Under the paper's assumptions, both guarantees hold with probability at least $1 - 2\delta$.

Safety throughout training

$J_c(\pi_n, f^\star) \leq d$ for every training episode.

Near-optimality in finite time

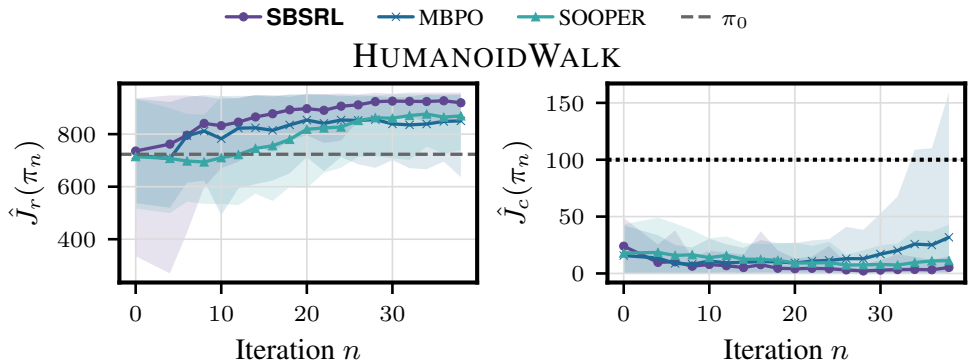
The returned policy is within ε of the best policy reachable from the initial safe set under a tightened budget.



At termination, the certified set covers the reachable component of the optimal tightened safe set.

Scalable implementation with neural network ensembles

Practical variant Check safety across models and explore using their disagreement.



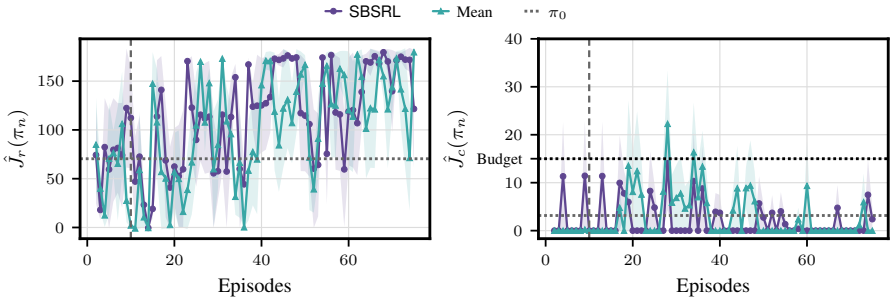
Reward (left) and cost (right). Mean and 95% empirical percentile interval over 25 runs.

Offline-to-online learning on an RC car



Hardware result

Higher reward than the offline policy and mean cost within budget.



Reward and collision cost. Mean ± standard error over three seeds.

Thank you!

Paper arXiv:2605.19469

Code github.com/lucavignola/safe-learning



Read the paper