Constrained Policy Optimization via Bayesian World Models

Yarden As, Ilnura Usmanova, Sebastian Curi, Andreas Krause

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

AUTOMATIC CONTROL LABORATORY

Learning & Adaptive Systems

Reinforcement learning agents demonstrate high potential in solving complex tasks.

How can we make them *safe*?

Constrained Markov decision processes

Formulating safety in reinforcement learning

Goal: $\max_{\pi \in \Pi} J(\pi, p^\star)$

$J = \mathbb{E}_{\pi, p^\star}\left[\sum_t r_t\right]$

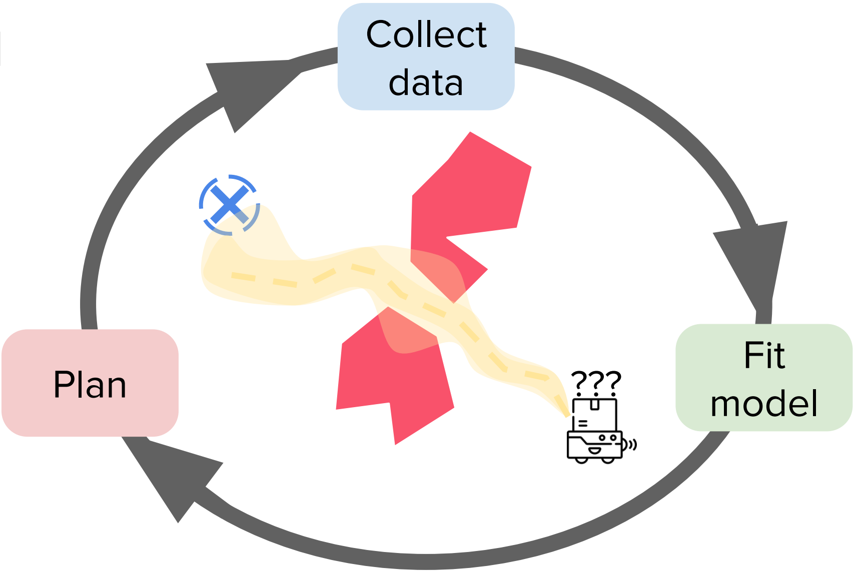
s.t.

$J_c(\pi, p^\star) \leq d$

$J_c = \mathbb{E}_{\pi, p^\star}\left[\sum_t c_t\right]$

Where $p^\star$ *are the true unknown dynamics*

Model-based RL training loop



Altman (1999). Constrained Markov Decision Processes

How do we use optimism and pessimism?

Idea: use the dynamics to be optimistic for the rewards and pessimism for the costs.

Goal: $\max_{\pi \in \Pi} J(\pi, p^\star)$

s.t.

$$J_c(\pi, p^\star) \leq d$$

Where $p^\star$ *are the* *true unknown* *dynamics*

$$\max_{\pi \in \Pi} \max_{p \in \mathcal{P}} J(\pi, p)$$

s.t.

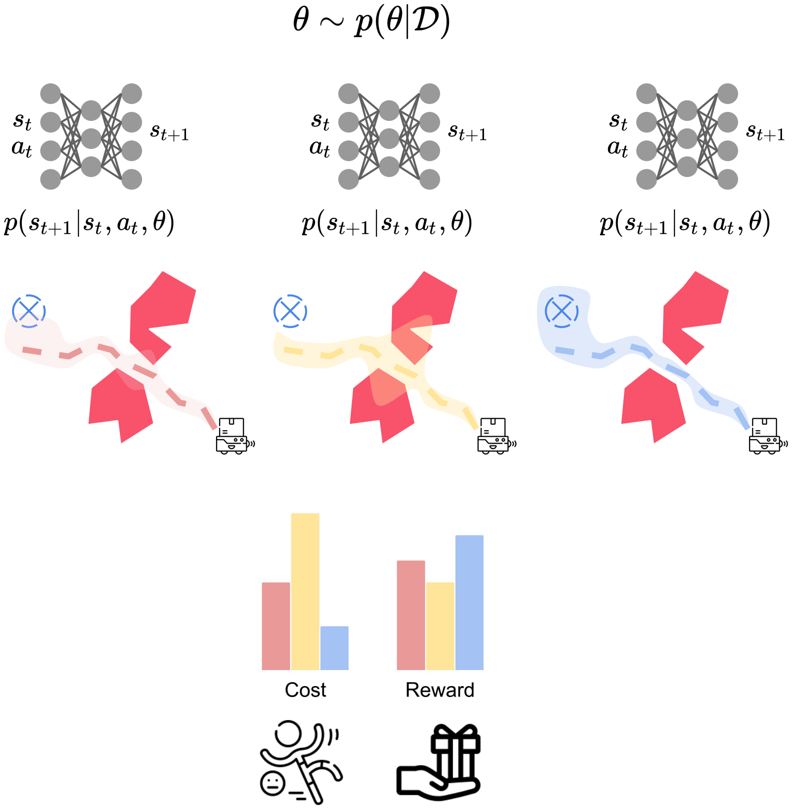
$$\max_{p \in \mathcal{P}} J_c(\pi, p) \leq d$$

Optimism

pessimism

But the inner maximization is intractable...

Upper bounds estimation via posterior sampling



1. **Sample model.**
2. **Sample trajectories.**
3. **Evaluate.** (via $J(\pi, p) \approx \mathbb{E}_{\pi,p}\left[\frac{1}{H}\sum_t^H \mathrm{V}(s_t)\right]$

   $J_c(\pi, p) \approx \mathbb{E}_{\pi,p}\left[\frac{1}{H}\sum_t^H \mathrm{V}_c(s_t)\right]$).
4. **Repeat steps 1-3.**
5. **Obtain optimistic/pessimistic evaluations.**

SWAG: *Maddox et al. (2019). A Simple Baseline for Bayesian Uncertainty in Deep Learning.* **RSSM:** *Hafner et al. (2019). Learning Latent Dynamics for Planning from Pixels. ICML (2019)*

LAgrangian Model-BaseD Agent (LAMBDA)

$$\max_{\pi \in \Pi} \max_{p \in \mathcal{P}} J(\pi, p)$$

s.t.

$$\max_{p \in \mathcal{P}} J_c(\pi, p) \leq d$$

Inner maximization

Via **posterior sampling**

$\theta \sim p(\theta|\mathcal{D})$

Solving the constrained problem

$$\max_{\pi \in \Pi} \min_{\lambda \geq 0} \left[J(\pi, p) - \lambda(J_c(\pi, p) - d)\right]$$

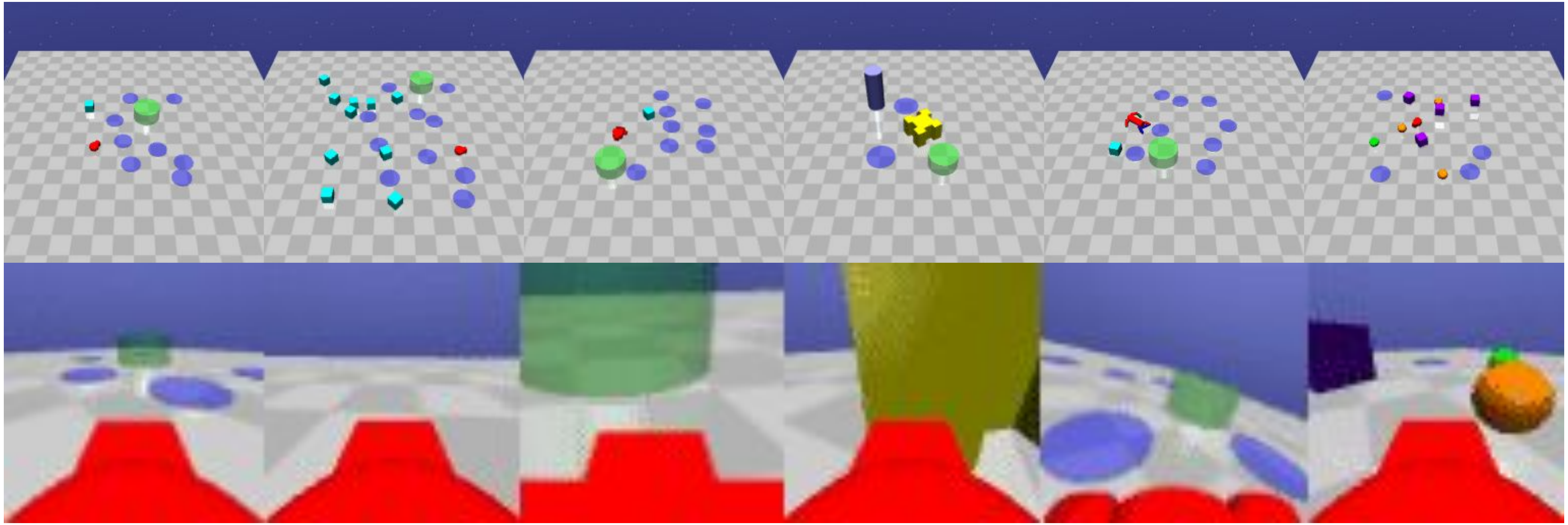
Performing **gradient ascent** on the **Lagrangian** and **gradient descent** on the **Lagrange multiplier**

Nocedal & Wright (2006). Numerical Optimization. Springer, 2nd Edition.

Benchmark with Safety-Gym

Robots, tasks and observations

PointGoal1 PointGoal2 CarGoal1 PointBox1 DoggoGoal1 PointButton1

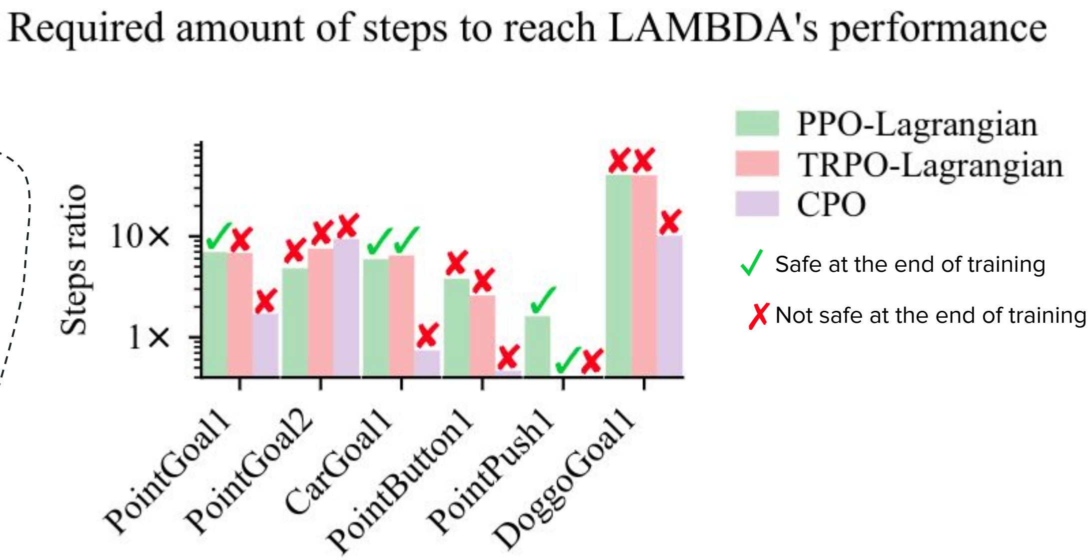


Experimental results

Higher is better

Model-based

Model-free



Between **x5-10** more sample efficient

Same amount of training

LAMBDA can *efficiently* learn to solve *complex* tasks which require *safe* behavior, from image observations.

Paper (pre-print, arXiv):

Code (github):

yarden.as@inf.ethz.ch