Towards incorporating safety in reinforcement learning

Maryam Kamgarpour
École Polytechnique Fédérale de Lausanne, Switzerland

Spring School on Control & Reinforcement Learning
CWI Amesterdam, The Netherlands

Mar 20,2025

EPFL

sycamore lab
SYSTEMS CONTROL AND MULTIAGENT OPTIMIZATION RESEARCH

Stochastic control framework

Stochastic control system

- state x_{t+1} is a sample from $P(\cdot \mid x_t, u_t)$

Problem: design controller $u_t = \pi(x_t)$ to

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_P \sum_t c(x_t, u_t) \\ \text{s.t.} \quad & x_{t+1} \sim P(\cdot \mid x_t, u_t) \end{aligned}$$

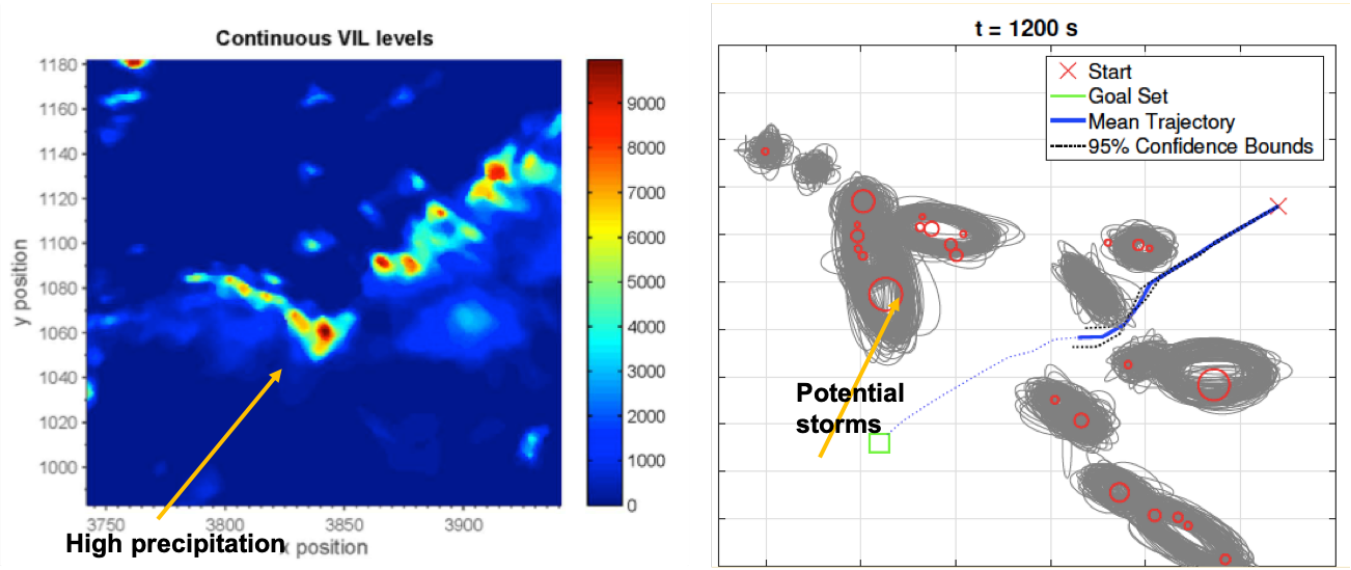
Dynamic programming (DP) [Bellman 1952]

$P(\cdot \mid x_t, u_t)$, objective $\rightarrow$ $\boxed{\text{DP}}$ $\rightarrow$ optimality conditions for π

Dynamic programming progress and limitations

Progress over the past decades

- addressing more general objectives
- multiagent formulation
- computational tractability



DP to design aircraft trajectory which maximizes probability of safety

Limitations: tractability, incorporating real-time data

Reinforcement learning (RL) approach

~~Given $x_{t+1} \sim P(\cdot \mid x_t, u_t)$~~

- Optimize π by interacting with the system

RL successes: chess, Go, Starcraft, protein design ...

Key challenge: guarantees for safe control systems

This talk

Towards incorporating performance guarantees in reinforcement learning for safe control systems

Safety in learning and control

- Constrained RL approaches
  - Lagrangian formulations [Bharadhwaj et al 2021], [Efroni et al. 2020], [Ding et al. 2021], ...
  - Constrained policy optimization [Achiam, et al. 2017], [Tsung-Yen et al. 2022], [Xu et al. 2021], ...
  - Model-based approaches [Zheng et al. 2020], [Turchetta et al. 2016], [Vaswani et al. 2022], [As et al. 2022], ...

- Control community approaches
  - Learning-based model predictive control [Hewig et al. 2019], [Coulson et al. 2019], [Zanon et al. 2020], [Berberich et al. 2021], [Maddalena et al. 2021], ...
  - Safely training neural net controllers [Zhao et al. 2020], [Xiao et al. 2021], ...
  - Formal methods [Alshiekh et al. 2017], [Fulton et al. 2019], [Hasanbeig et al. 2020], ...
  - Certificate functions, e.g. Lyapunov or control barrier functions [Chow et al. 2018], [Dutta et al. 2018], [Taylor et al. 2019], [Perkins et al. 2002], [Ma et al. 2022], [Emam et al. 2022], [Cohen et al. 2023], [Dowson et al. 2023], , ...
  - Gaussian processes [Akametalu et al. 2014], [Wachi et al. 2018], ...

Table of Contents

How objectives are incorporated in reinforcement learning

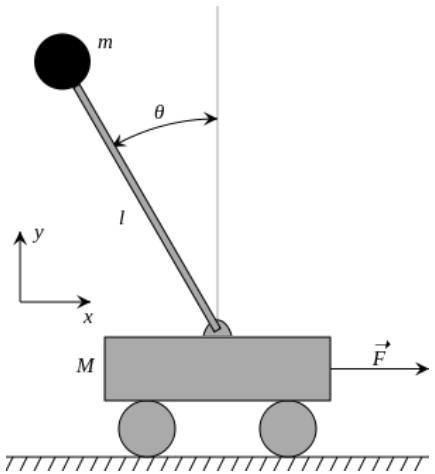
- Finite horizon: $\mathbb{E}_{P,\pi}\left[\sum_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right]$
  - can encode probability of trajectory staying inside a safe set: $\mathcal{P}(x_t \in S, t = 0, \ldots, T)$
- Discounted : $\mathbb{E}_{P,\pi}\left[\sum_{t=0}^{\infty} \gamma^t c(x_t, u_t)\right], \quad 0 < \gamma < 1$
  - cannot ensure infinite horizon safety/stability/reachability
- Average cost : $\limsup_{T\to\infty} \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{\infty} \frac{1}{T} c(x_t, u_t)\right]$
  - can capture infinite horizon properties above
  - learning would require assumptions

The first two are most commonly used but can they ensure any infinite horizon property such as stability or safety of the system?

Stability of discrete-time dynamical system

Discrete-time dynamical system:

- $x_{t+1} = f(x_t)$, $x_t \in \mathbb{R}^n$, $x_0 \in \mathbb{R}^n$
- Equilibrium $x_e \in \mathbb{R}^n$: $x_e = f(x_e)$



The equilibrium x_e is

- stable (in the sense of Lyapunov):
  $\forall \epsilon > 0,\ \exists \delta > 0$ s.t $\|x_0 - x_e\| \leq \delta \implies \|x_t - x_e\| \leq \epsilon,\ \forall t$
- locally asymptotically stable: it is stable and $\exists$ neighborhood of x_e, $U \subset \mathbb{R}^n$ s.t. $x_0 \in U \implies x_t \to x_e$

Let's see through an example why a discounted cost cannot ensure stability

Linear quadratic control optimal solution

- The infinite horizon discounted LQR problem

$$\min_{u_0,u_1,\ldots} \quad \sum_{t=0}^{\infty} \gamma^t (x_t^T Q x_t + u_t^T R u_t)$$
$$\text{s. t.} \quad x_{t+1} = Ax_t + Bu_t + w_t$$

  - $x_t \in \mathbb{R}^n, u_t \in \mathbb{R}^m$, $w_t \in \mathcal{N}(0, \sigma^2)$
  - $Q \in \mathcal{S}_+^n, R \in \mathcal{S}_{++}^m$, $0 < \gamma < 1$

- Motivation
  - many systems are approximated by linear models
  - rich theory exists for control of linear systems [Linear system theory, Callier, Desoer, 1982]
  - linearity gives insight into more complex nonlinear systems and systems with continuous state and actions

Linear quadratic control optimal solution

- The infinite horizon discounted LQR problem

$$\min_{\pi} \quad \sum_{t=0}^{\infty} \gamma^t (x_t^T Q x_t + u_t^T R u_t)$$
$$\text{s.t.} \quad x_{t+1} = A x_t + B u_t + w_t$$

  - $x_t \in \mathbb{R}^n, u_t \in \mathbb{R}^m$, $w_t \in \mathcal{N}(0, \sigma^2)$
  - $Q \in \mathcal{S}_+^n, R \in \mathcal{S}_{++}^m$, $0 < \gamma < 1$

- Optimal control policy is linear $u_t = K x_t$ where
  - optimal cost: $x_0^T P x_0 + \sigma^2 \frac{\gamma}{1-\gamma} \mathsf{Tr}(P)$, $P \in \mathcal{S}_{++}^m$
  - P is a solution to the Riccati equation:
    $P = \gamma A^T P A - \gamma^2 A^T P B (R + \gamma B^T P B)^{-1} B^T P A + Q$
  - optimal controller: $K = -(R + B^T P B)^{-1} B^T P A$

[Bertsekas, Dynamic programming and optimal control, Chapter 4]

Minimizing a discounted cost cannot ensure stability

$$\min_{u_0,u_1,\dots} \quad \sum_{t=0}^{\infty} \gamma^t (x_t^T Q x_t + u_t^T R u_t)$$
$$\text{s.t.} \quad x_{t+1} = Ax_t + Bu_t + w_t$$

- Let $A = 2, B = 1, R = 1, Q = 1$: scalar linear system
- For $\gamma \in (0, 1/3)$, the controller is not stabilizing
  - example: $\gamma = 1/4 \rightarrow A + BK = 2.3904$
  - $x_{k+1} = (A + BK)x_k$ divergent geometric series
- Intuitively, the discount forgets about long term
- It is a problem with formulation of the cost, not any approach to solve the problem

[Postoyan et al. 2017] conditions for stability

Cost formulation for stability

One can derive conditions based on

- Controllability of the system (holds in example above)
- Detectability of the cost (holds in above)
- Sufficiently large discounted factor (doesn't hold above)

to ensure the optimal controller from

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_P \left[\sum_t \gamma^t c(x_t, u_t) \right] \\ \text{s.t.} \quad & x_{t+1} \sim P(\cdot \mid x_t, u_t) \end{aligned}$$

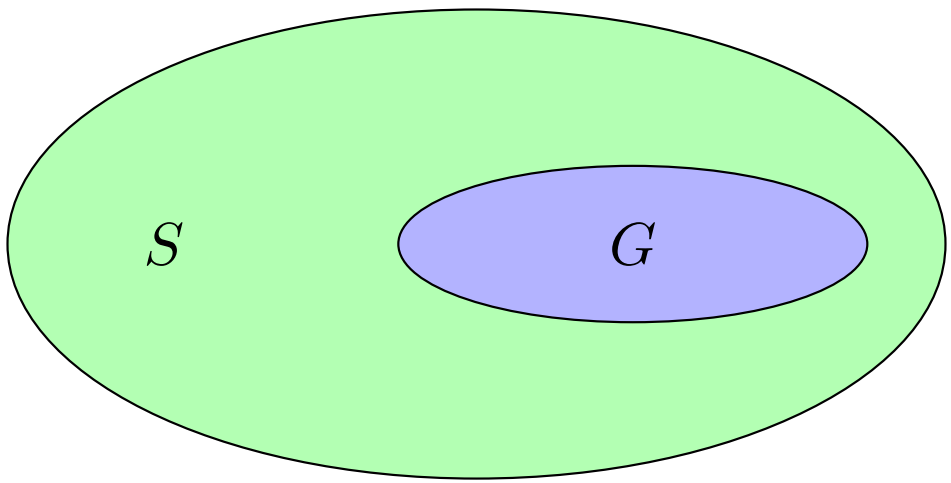
is stabilizing

- Catch: the conditions requires knowing the system dynamics, but the dynamics are not assumed known in an RL setting

What about other properties, such as safety or reachability?

Infinite horizon safety and reachability specifications

- Probabilistic safety: $\mathbb{P}_\pi(S) = \mathbb{P}(x_t \in S,\ \forall t)$
- Probabilistic reachability: $\mathbb{P}_\pi(G) = \mathbb{P}(\exists t \text{ s.t. } x_t \in G)$
- Safety and reachability are dual concepts
  - $S := X \setminus G$ ($X \setminus G = G^c$: complement of set G): $\mathbb{P}_\pi(S) = 1 - \mathbb{P}_\pi(G)$



Safety cannot be cast as a discounted reward

- Alur et a. Theorem 1: given a reachability specification, for any discounted reward R there exists a transition kernel such that the optimal policy corresponding to R will not correspond to the optimal policy for $P_\pi(G)$

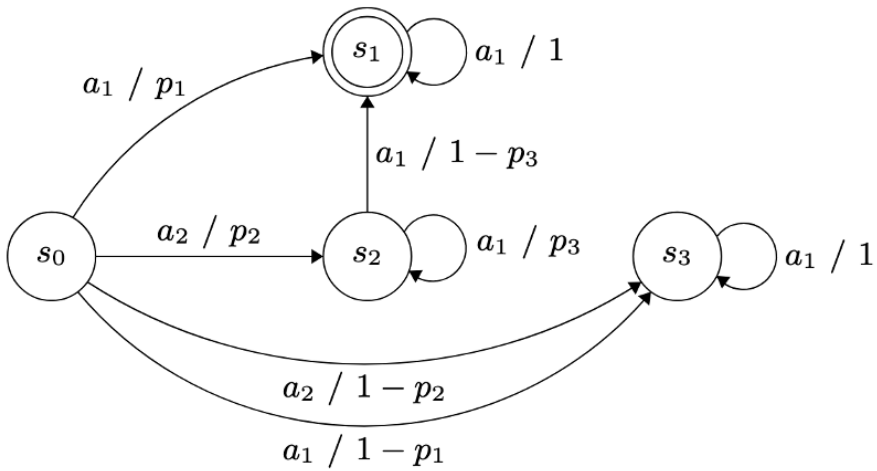


Fig. 1: Counterexample for reducing reach specification to a discounted RM.

Figure from Alur et. al, 2022

See [Figure 1, Topcu & Littman, 2017] for additional example and discussion

Takeaway message

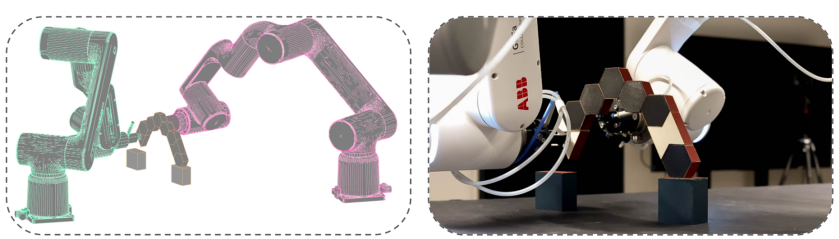
No guarantees for infinite horizon performance with discounted or finite horizon cost in RL

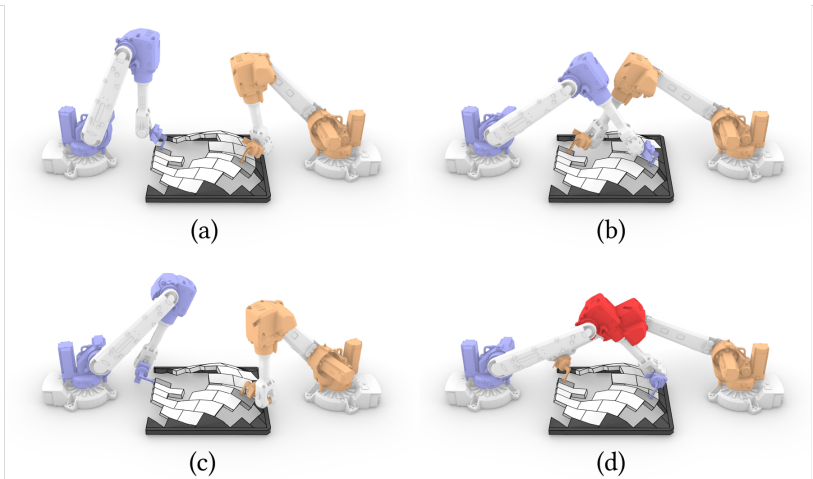
- Can we have any guarantees for stability of an inverted pendulum with a controller that optimizes a discounted cost?
- Would an average cost help?
  - Average cost can capture safety specifications [Theorem 2, from Alur et al. 2022]
  - But reinforcement learning with average cost is more challenging

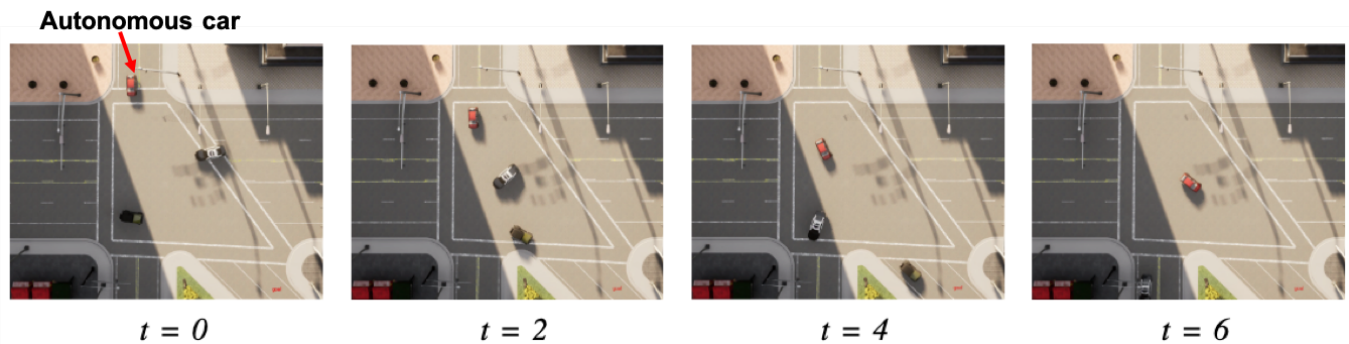
For the rest of the talk, we will focus on safety specifications, but in finite horizon

Table of Contents

Tasks we are addressing with RL





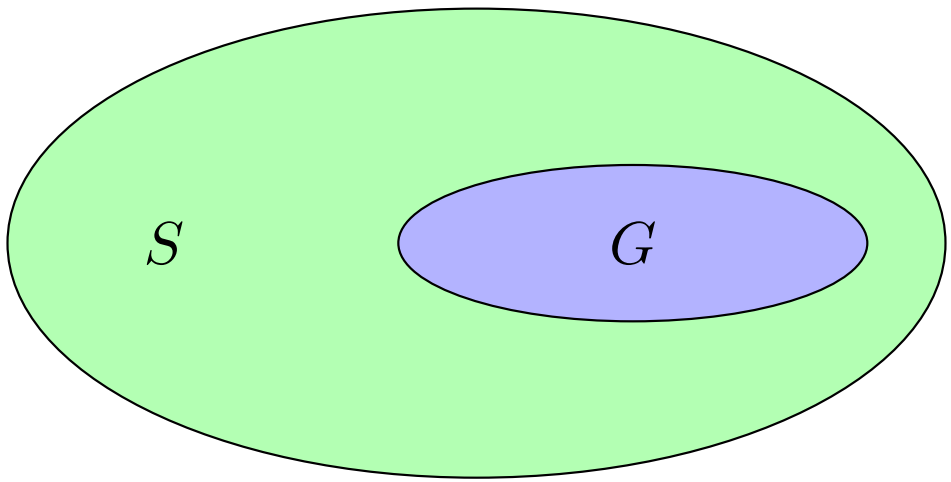


- finite horizon
- safety constraints

Safety and reachability

Consider a stochastic control system with finite horizon T:

- Safety (Invariance): given S, $\forall t \in [T]\ x_t \in S$
- Reachability: given G, $\exists t \in [T]$, s.t. $x_t \in G$
- Reach-avoid: given S, G, $\exists t \in [T]$, s.t. $x_t \in G$, $x_\tau \in S \setminus G$, $\forall \tau \in [t-1]$
  - includes safety and reachability properties above



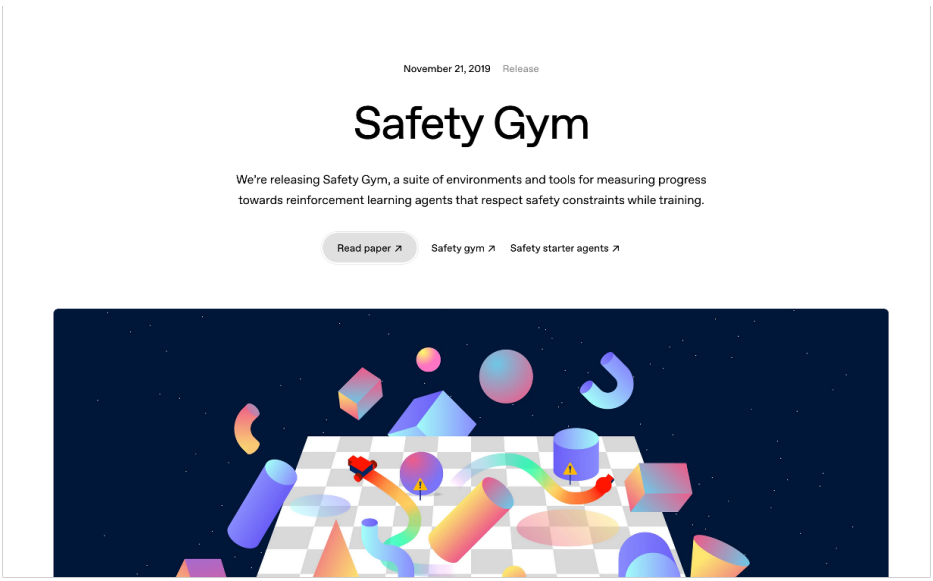
How do past works in RL incorporate safety constraints?

- A high negative reward for violation of constraints: $w\mathbb{1}_{x\notin S}(.)$
- Formulating a cumulative constraint threshold:

$$\min_{\pi} \quad \mathbb{E}_{P,\pi} \sum_{t} \gamma^t c(x_t, u_t)$$

$$\text{s.t.} \quad \mathbb{E}_{P,\pi} \sum_{t} \gamma^t \mathbb{1}_{x\notin S}(.) \leq 0.$$

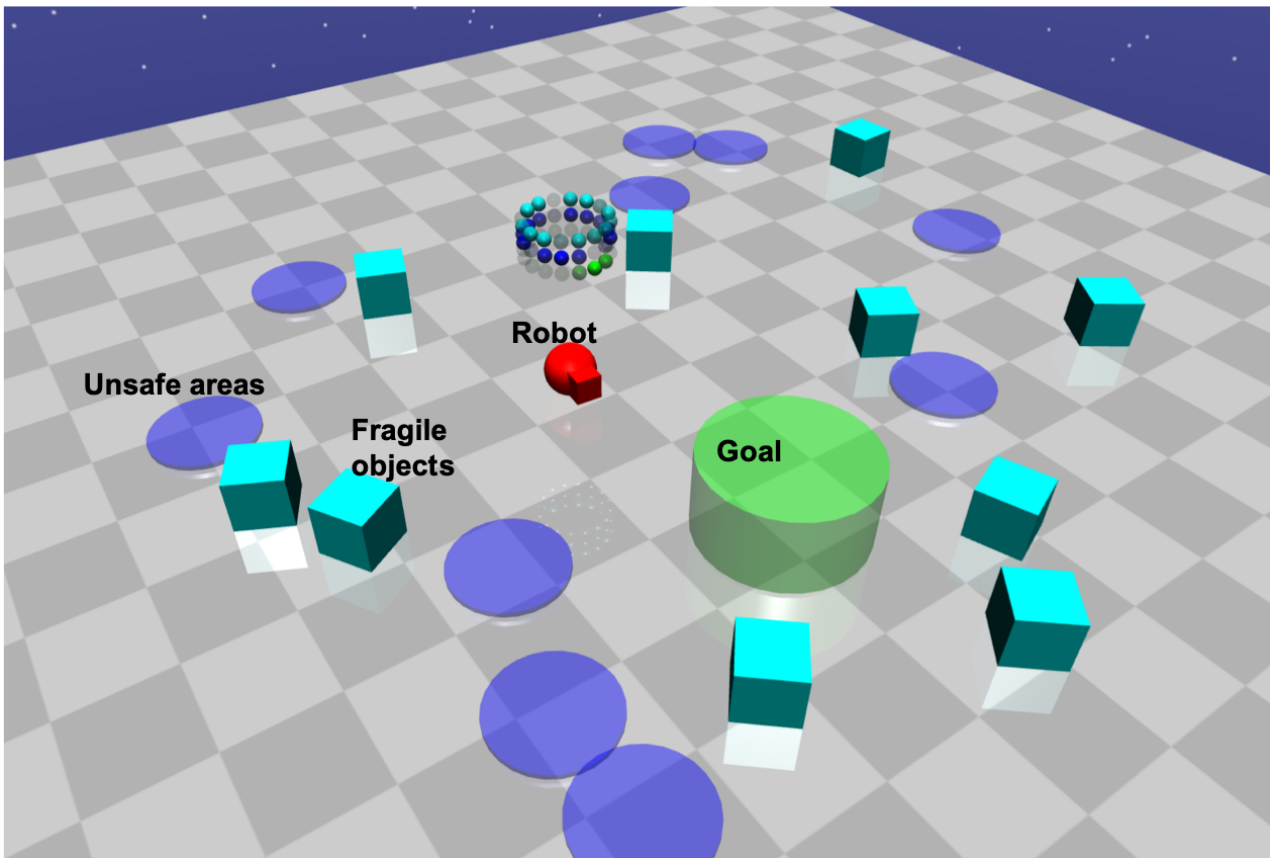
- no a priori guarantees on safety



From Open AI

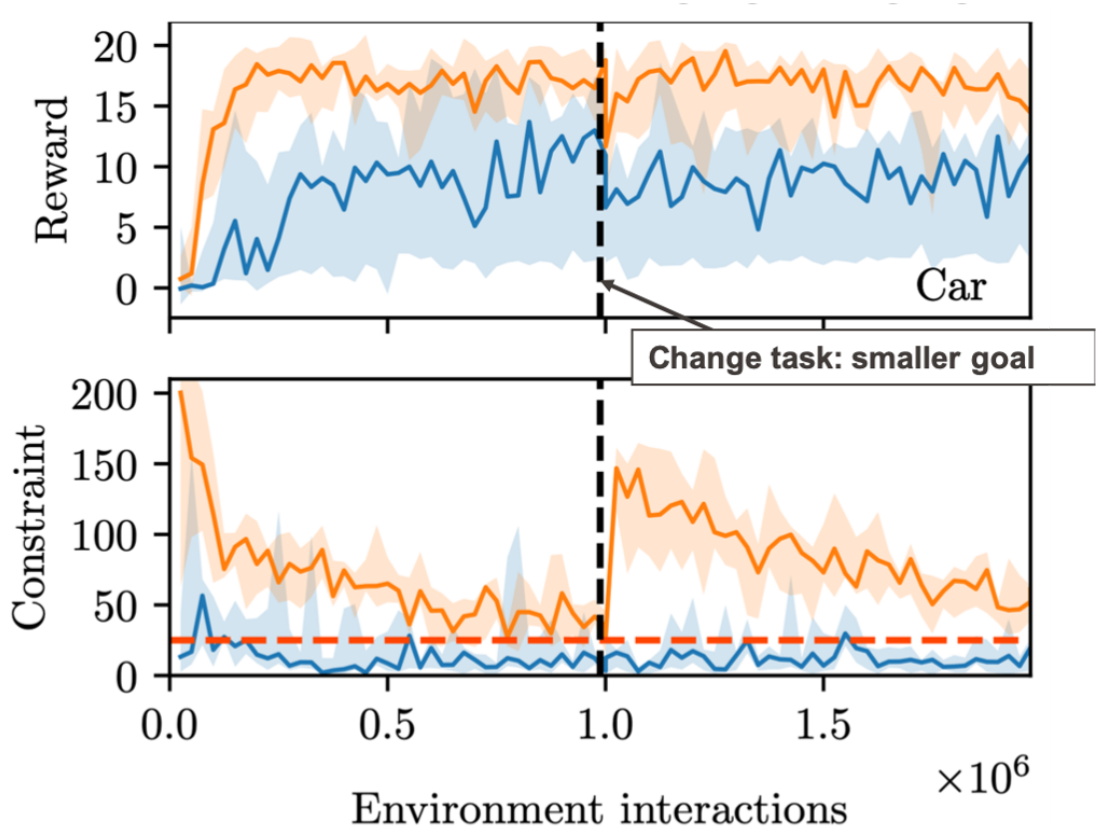
Example from SafetyGym

- Objective: reach the goal while avoiding obstacles
- Challenge: unknown dynamics and environment
- Approach: learn a neural network policy directly from images



From Open AI's Safety Gym

Performance constrained MDP approaches



Our approach in blue, Lagrangian approach in orange

Lagrangian: solves constrained RL, without safety during learning [As et al. 2022]

[Usmanova, As, **MK**, Krause, JMLR2024]

Constrained Markov decision process

Consider a finite horizon CMDP [Puterman, 1998]:

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right] \\ \text{s.t.} \quad & \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{T-1} c_s(x_t, u_t) + c_{s,T}(x_T)\right] \leq 0. \end{aligned}$$

Can we cast finite horizon safety/reachability constraint as a CMDP?

Probabilistic safety and reachability constraints through CMDP and solution through a policy gradient approach

Work with Ph.D. student: Tingting Ni

Reach-avoid constrained stochastic control

- Trajectory satisfying the reach-avoid condition:

$$\mathcal{C}_{S,G} := \{x_{0:T} \mid \exists t \in [T], x_t \in G, x_\tau \in S \setminus G, \forall \tau \in [t-1]\}.$$

Find π^* that solves:

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right] \\ \text{s.t.} \quad & \mathbb{P}_\pi(\mathcal{C}_{S,G}) \geq 1 - \delta, \end{aligned}$$

$\delta \in [0, 1)$: risk tolerance parameter

[Abate et al. 2008, Summers et al. 2010, Kamgarpour et al. 2013, Schmid et al. 2025]

Class of policies

History h_t at time t:

$$h_t := \{x_0, u_0, \ldots, x_{t-1}, u_{t-1}, x_t\}.$$

To select an action u_t, we consider a stochastic policy π_t:

- History-dependent: $u_t \sim \pi_t(\cdot \mid h_t)$
- Markovian: $u_t \sim \pi_t(\cdot \mid x_t)$
  - Stationary: $\pi_{t_1}(\cdot \mid x) = \pi_{t_2}(\cdot \mid x)$, $\forall t_1, t_2$

Optimal policies in a CMDP

Consider a CMDP

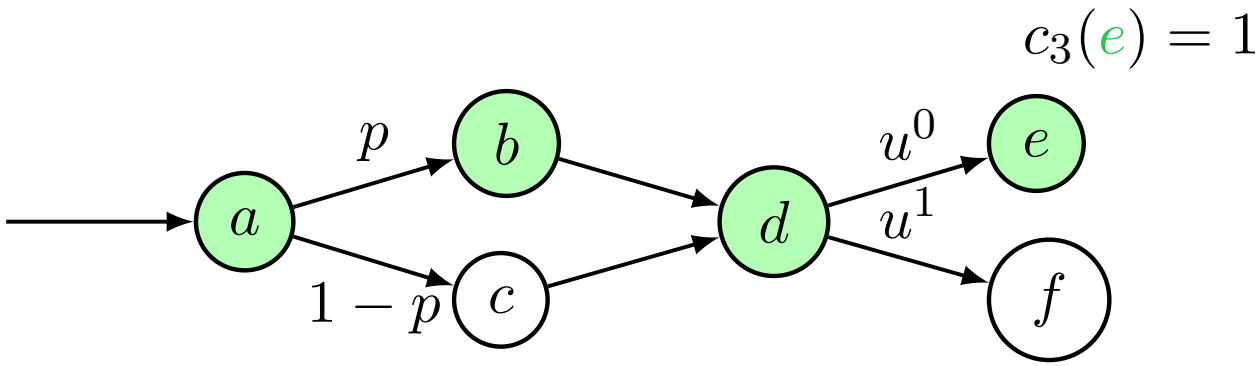
$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right] \\ \text{s.t.} \quad & \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{T-1} c_s(x_t, u_t) + c_{s,T}(x_T)\right] \leq 0. \end{aligned}$$

- Optimal policy is Markov [Puterman 1998]

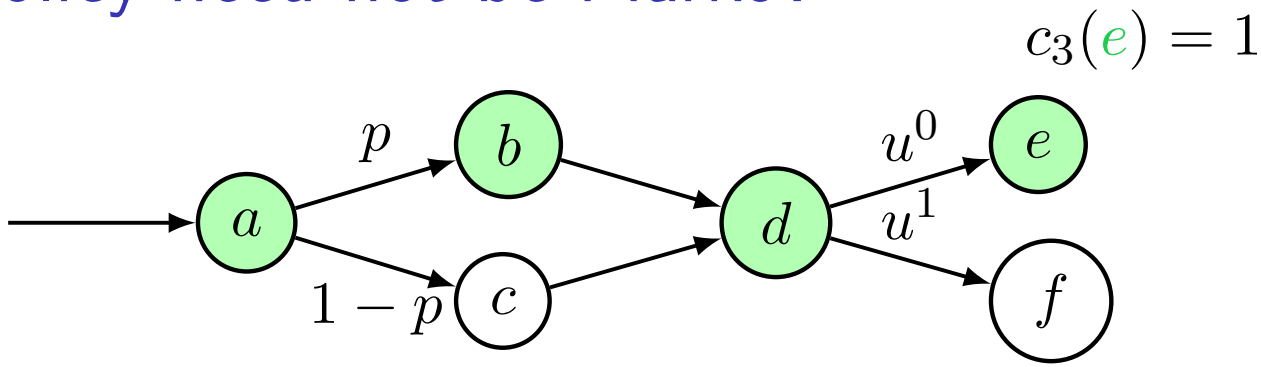
Optimal policies in reach-avoid constrained MDP

Consider a MDP with a horizon $T = 3$:

- $S = \{a, b, d, e\}$
- Constraint: $\mathbb{P}(x_{0:3} \in S) \geq 1 - \delta$
- Cost function: $c_T(x) = \mathbf{1}_{\{e\}}(x)$



Optimal policy need not be Markov



Safety: $\mathbb{P}(x_{0:3} \in S) = \mathbb{P}(a, b, d, e)$

Cost: $\mathbb{E}_{P,\pi}[c_3(x_3)] = \mathbb{P}(x_3 = e)$

$= \mathbb{P}(a, b, d, e) + \mathbb{P}(a, c, d, e)$

- Markovian policy
  - Safety: $p\pi(u^0 \mid d)$
  - Cost: $p\pi(u^0 \mid d) + (1-p)\pi(u^0 \mid d) = \pi(u^0 \mid d)$
- Non-Markovian policy
  - Safety: $p\pi(u^0 \mid a, b, d)$
  - Cost: $p\pi(u^0 \mid a, b, d) + (1-p)\pi(u^0 \mid a, c, d)$

Optimal policy need not be Markov

- Markovian policy

$$\begin{aligned} \min_{\pi} \quad & \pi(u^0 \mid d) \\ \text{s.t.} \quad & p\pi(u^0 \mid d) \geq 1 - \delta. \end{aligned}$$

  - Optimal cost: $(1-\delta)/p$

- Non-Markovian policy

$$\begin{aligned} \min_{\pi} \quad & p\pi(u^0 \mid a, b, d) + (1-p)\pi(u^0 \mid a, c, d) \\ \text{s.t.} \quad & p\pi(u^0 \mid a, b, d) \geq 1 - \delta \end{aligned}$$

  - Optimal cost: $1 - \delta$

- optimal non-Markovian has lower cost than Markovian policy

Optimal policy need not be Markov

- Markovian policy

$$\begin{aligned} \min_{\pi} \quad & \pi(u^0 \mid d) \\ \text{s.t.} \quad & p\pi(u^0 \mid d) \geq 1 - \delta. \end{aligned}$$

  - Optimal cost: $(1-\delta)/p$

- Non-Markovian policy

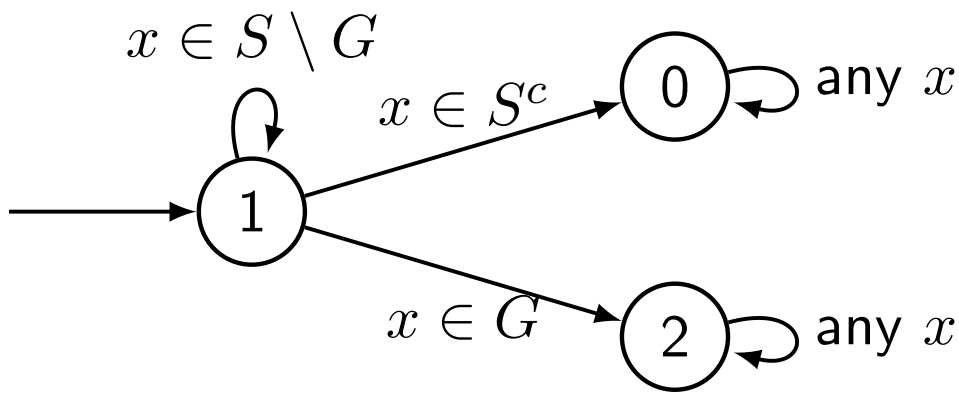
$$\begin{aligned} \min_{\pi} \quad & p\pi(u^0 \mid a, b, d) + (1-p)\pi(u^0 \mid a, c, d) \\ \text{s.t.} \quad & p\pi(u^0 \mid a, b, d) \geq 1 - \delta \end{aligned}$$

  - Optimal cost: $1-\delta$

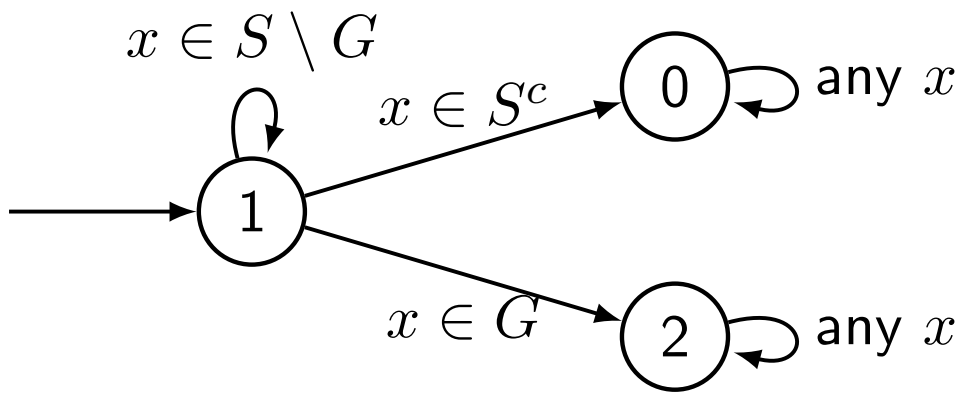
- optimal non-Markovian has lower cost than Markovian policy

Casting the reach-avoid constrained MDP as a CMDP

- Introduce additional states $Q = \{0, 1, 2\}$
  - reaching S^c, being in $S \setminus G$, and reaching G
- New state-space $X \times Q$
- Transition probability on $X \times Q$
  - $x_{t+1} \sim P(\cdot \mid x_t, u_t)$
  - q_{t+1} is decided by x_t, q_t:

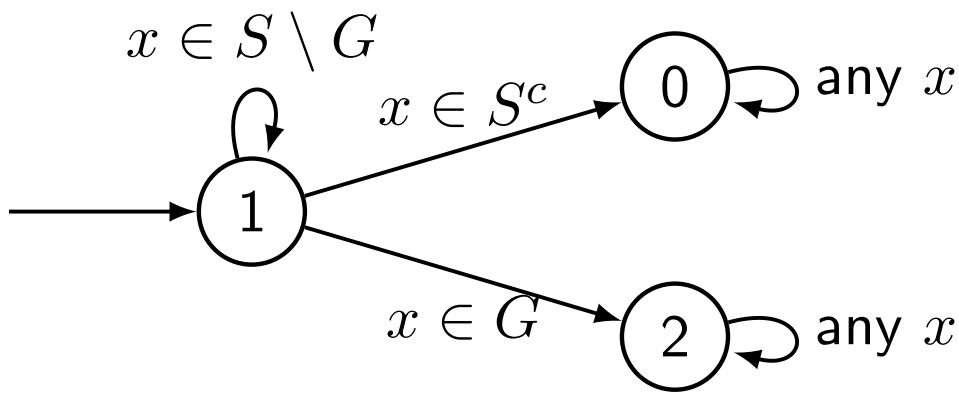


Transition dynamic for q



- Given x, q, the transition dynamics for q are:
  - Given $x_{1:T}$, we can generate $q_{1:T}$
  - $x_{1:T} \in \mathcal{C}_{S,G}$ if and only if $q_T = 2$
  - $\mathbb{E}_{P,\pi}\left[c_{s,T}(x_T, q_T)\right] = \mathbb{P}_\pi(\mathcal{C}_{S,G})$.
  - Cost: $c((x,q),u) = c(x,u), c_T(x,q) = c_T(x)$ for $q \in Q$
  - Constraint: $\mathbb{E}_{P,\pi}\left[c_{s,T}(x_T, q_T)\right] = \mathbb{E}_{P,\pi}\left[\mathbf{1}_{\{2\}}(q_T)\right]$

Transition dynamic for q



- Given x, q, the transition dynamics for q are:
  - Given $x_{1:T}$, we can generate $q_{1:T}$
  - $x_{1:T} \in \mathcal{C}_{S,G}$ if and only if $q_T = 2$
  - $\mathbb{E}_{P,\pi}\left[c_{s,T}(x_T, q_T)\right] = \mathbb{P}_\pi(\mathcal{C}_{S,G})$.
  - Cost: $c((x,q),u) = c(x,u), c_T(x,q) = c_T(x)$ for $q \in Q$
  - Constraint: $\mathbb{E}_{P,\pi}\left[c_{s,T}(x_T, q_T)\right] = \mathbb{E}_{P,\pi}\left[\mathbf{1}_{\{2\}}(q_T)\right]$

Reach-avoid constrained MDP Optimal solution

$$\begin{cases} \min\limits_{\pi} & \mathbb{E}_{P,\pi}\left[\sum\limits_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right] \\ \text{s. t.} & \mathbb{P}_{\pi}(\mathcal{C}_{S,G}) \geq 1 - \delta. \end{cases} \iff \begin{cases} \min\limits_{\pi} & \mathbb{E}_{P,\pi}\left[\sum\limits_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right] \\ \text{s. t.} & \mathbb{E}_{P,\pi}\left[c_{s,T}(x_T, q_T)\right] \geq 1 - \delta. \end{cases}$$

Theorem

- *The above two problems are equivalent.*
- *Sufficient to consider Markovian policies on product space*

$$\{\pi_t(\cdot \mid x_t, q_t)\}_{t \in [T-1]}$$

- $\pi_t^*(\cdot \mid x_t, q_t) \mapsto \pi_t^*(\cdot \mid h_t)$

Take-away message

- Can formulate safety and reachability constraints on the trajectory as a constrained MDP over an extended state-space
- Safety is guaranteed by design
- Note: any specification given by an automaton can be cast as a reachability problem on an extended state-space, and thus, also formulated as a CMDP

Table of Contents

Constrained reinforcement learning approach

Find $\pi(\cdot \mid x, q)$ such that it solves the problem below:

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right] \\ \text{s.t.} \quad & \mathbb{E}_{P,\pi}\left[c_{s,T}(x_T, q_T)\right] \geq 1 - \delta. \end{aligned}$$

- Data: trajectories from the system for a given policy
- Ensure safety of the policies during learning

Constrained reinforcement learning

~~Given $x_{t+1} \sim P(\cdot \mid x_t, u_t)$~~, parametrize policy: $u_t \sim \pi_t^{\theta_t}(\cdot \mid x_t, q_t)$

$$\min_{\theta := \{\theta_t\}_{t \in [T-1]}} \quad J(\pi^\theta) := \mathbb{E}_{P,\pi^\theta}\left[\sum_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right]$$

$$\text{s.t.} \quad C(\pi^\theta) := 1 - \delta - \mathbb{E}_{P,\pi^\theta}\left[c_{s,T}(x_T, q_T)\right] \leq 0$$

Data: system trajectory

Safe learning: Design an algorithm such that $\pi^{\theta(k)}$ satisfies constraints and converges to the optimal policy

Safe learning as constrained optimization

Policy parametrization: $\theta \in \mathbb{R}^d$,

- Linear: $\pi^\theta(x, q) = \theta^T(x, q) + \omega$
- Gaussian: $\pi^\theta(u \mid x, q) = \mathcal{N}(\phi^\theta(x, q), \Sigma)$
- Softmax: $\pi^\theta(u \mid x, q) \propto \exp(\phi^\theta(x, q))$
- ...

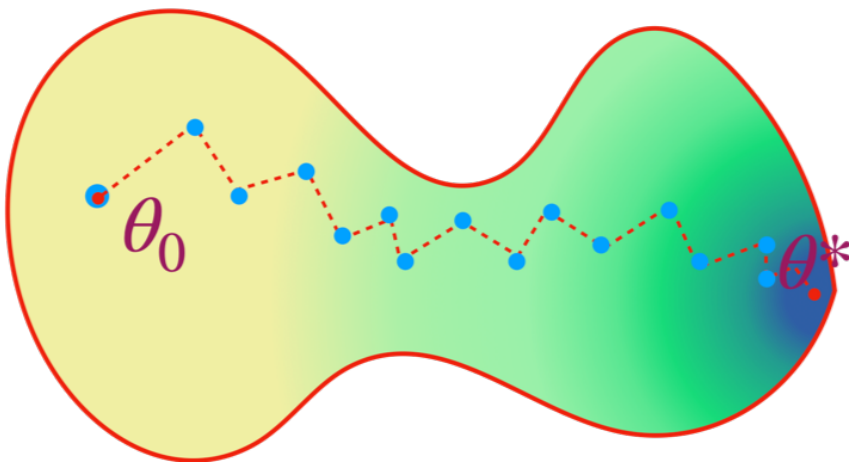
$$\begin{cases} \min\limits_{\pi} \ \mathbb{E}_{P,\pi}\left[\sum\limits_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right] \\ \text{s. t.} \ \ 1 - \delta - \mathbb{E}_{P,\pi}\left[c_{s,T}(x_T, q_T)\right] \leq 0 \end{cases} \Longrightarrow \begin{cases} \min\limits_{\theta} & J(\theta) \\ \text{s. t.} & C(\theta) \leq 0 \end{cases}$$

Safe learning as blackbox constrained optimization

~~Given $x_{t+1} \sim P(\cdot \mid x_t, u_t)$~~ $\implies$ $J(\cdot), C(\cdot)$ unknown

$$\begin{aligned} \min_{\theta} \quad & J(\theta) \\ \text{s.t.} \quad & C(\theta) \leq 0 \end{aligned}$$

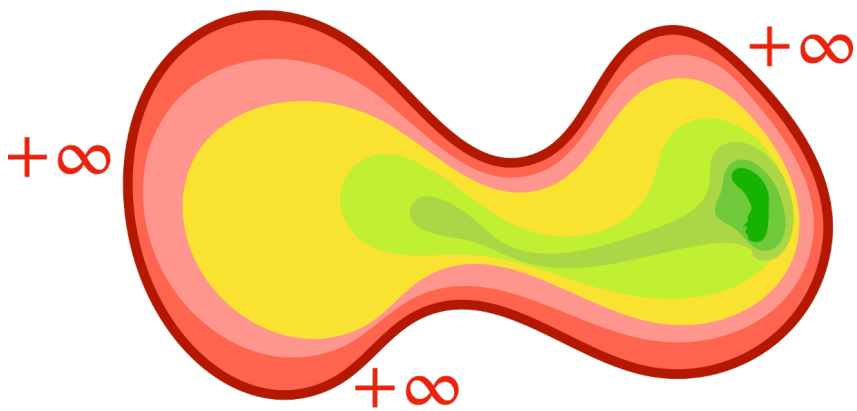
Safe learning: design $\{\theta(k)\}_k$ such that $C(\theta(k)) \leq 0$ and $\pi^{\theta(k)} \to \pi^*$



Challenges: $J(\cdot)$, $C(\cdot)$ non-convex and unknown

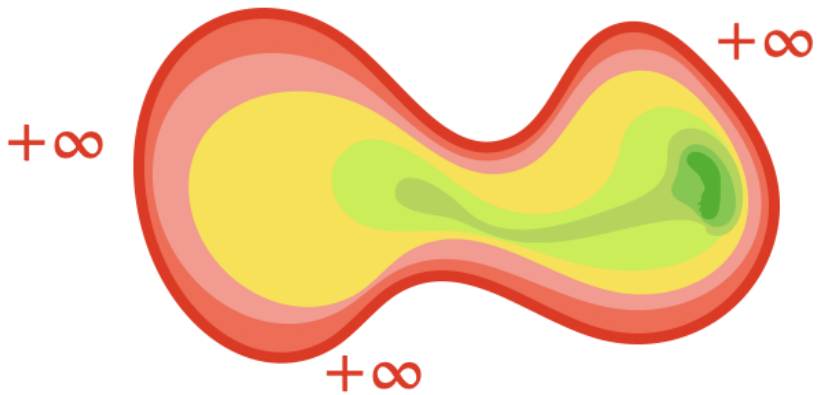
Overview of the proposed approach

- Design a barrier to stay inside the feasible set
- Estimate gradients to find a descent direction
- Take a carefully chosen step in the descent direction



Log barrier of the constrained optimization

- Log barrier of the constraint: $-\log\big(-C(\theta)\big)$



- Unconstrained optimization $\tilde{J}(\theta) = J(\theta) - \eta \log\big(-C(\theta)\big)$
  - $\eta \to 0$: approximate solution $\to$ true solution

Log barrier policy gradient approach

Algorithm: $\theta(k+1) = \theta(k) - \gamma(k)\nabla_\theta \tilde{J}(\theta(k))$

1. Would it converge to optimal policy parameters?
2. How to construct a good estimate of log barrier gradient?
3. How to choose $\gamma(k)$ for safety and convergence?

Log barrier policy gradient approach

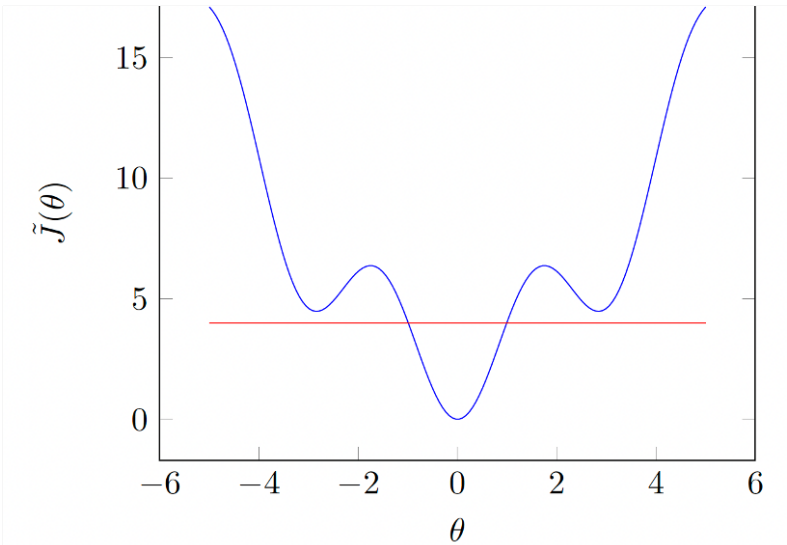
Algorithm: $\theta(k+1) = \theta(k) - \gamma(k)\nabla_\theta \tilde{J}(\theta(k))$

1. Would it converge to optimal policy parameters?
2. How to construct a good estimate of log barrier gradient?
3. How to choose $\gamma(k)$ for safety and convergence?

1. Stationary points of $\tilde{J}$ are nearly optimal

- transition kernel and policy parameterization $\implies$ gradient dominance:

$$J(\theta) - J(\theta^*) \leq \eta + \frac{1}{\nu}\|\nabla_\theta \tilde{J}(\theta)\|_2,\ \nu > 0$$



- What policy parameterization ensure $\nu > 0$?

1. Stationary points of $\tilde{J}$ are nearly optimal

Fisher non-degenerate policies: $\exists \mu_F > 0$ such that $\forall \pi^\theta$ and t,

$$\mathbb{E}[\nabla_{\theta_t} \log \pi_t^\theta(u \mid x, q)(\nabla_{\theta_t} \log \pi_t^\theta(u \mid x, q))^T] \geq \mu_F \mathbf{I},$$

- Gaussian and Cauchy parameterization
- Softmax classes of policies fail (intuition, no uniform lower bound as policy gets deterministic)

$\nearrow$ randomness of the policy $\implies$ $\mu_F \nearrow$

1. Stationary points of $\tilde{J}$ are nearly optimal

Fisher non-degenerate policies: $\exists \mu_F > 0$ such that $\forall \pi^\theta$ and t,

$$\mathbb{E}[\nabla_{\theta_t} \log \pi_t^\theta(u \mid x, q)(\nabla_{\theta_t} \log \pi_t^\theta(u \mid x, q))^T] \geq \mu_F \mathbf{I},$$

- Gaussian and Cauchy parameterization
- Softmax classes of policies fail (intuition, no uniform lower bound as policy gets deterministic)

$$\nearrow \text{ randomness of the policy } \implies \mu_F \nearrow$$

1. Stationary points of $\tilde{J}$ are nearly optimal

Lemma

Consider M_g-smooth Fisher non-degenerate policies. For any θ,

$$J(\theta) - J(\pi^*) \leq \eta + \frac{T^{\frac{1}{2}} M_g}{\mu_F} \left\| \nabla_\theta \tilde{J}(\theta) \right\| + \sqrt{\varepsilon_{bias}} T \left(1 - \frac{\eta}{C(\theta)} \right).$$

- "gradient-dominance"-like property
- $\nearrow$ richness of policy parameterization $\longrightarrow \varepsilon_{bias} \implies$
  - in finite action: $\varepsilon_{bias} = 0$ for softmax
  - neural softmax: $\nearrow$ depth $\implies \varepsilon_{bias} \searrow$

[Ni, **MK**, AISTATS 2025]

2. Constructing high confidence gradient estimator

$$\nabla_\theta \tilde{J}(\theta) = \nabla_\theta J(\theta) - \eta \frac{\nabla_\theta C(\theta)}{C(\theta)}$$

- Sample average to estimate $\mathbb{E}_{P,\pi}\left[c_{s,T}(x_T, q_T)\right]$ and thus, $C(\cdot)$

- Policy gradient theorem to estimate $\nabla_\theta J(\cdot), \nabla_\theta C(\cdot)$

2. Policy gradient theorem

- Consider the cost-to-go V_t^θ:

$$\begin{aligned} & V_t^\theta((x,q),u) \\ =& \mathbb{E}_{P,\pi^\theta}\left[\sum_{\tau=t}^{T-1} c(x_\tau,u_\tau) + c_T(x_T) \mid (x_t,q_t)=(x,q), u_t = u\right]. \end{aligned}$$

- For any θ, we have [Sutton et al. 1999],[Klein et al. 2024]

$$\nabla_\theta J(\theta) = \sum_{t\in[T-1]} \mathbb{E}_{P,\pi^\theta}\left[\nabla_\theta \log \pi_t^{\theta_t}(u \mid x,q) V_t^\theta((x,q),u)\right].$$

2. Policy gradient for $\nabla_\theta J(\cdot), \nabla_\theta C(\cdot)$

Apply π^θ to obtain n trajectories

Estimate gradients based on policy gradient theorem

$$\hat{\nabla}_\theta J(\theta) := \frac{1}{n} \sum_{i\in[n]} \sum_{t\in[T-1]} \nabla_\theta \log \pi_t^{\theta_t}(u_t^i \mid x_t^i, q_t^i) \Big(\sum_{m=t}^{T-1} c(x_m^i, u_m^i) + c_T(x_T^i)\Big),$$

Similarly, compute $\hat{\nabla}_\theta C(\theta)$.

2. Constructing high confidence gradient estimator

$$\hat{\nabla}_\theta \tilde{J}(\theta) = \hat{\nabla}_\theta J(\theta) - \eta \frac{\hat{\nabla}_\theta C(\theta)}{\hat{C}(\theta)}$$

- Sample average estimates of $C(\cdot)$:

- Policy gradient theorem: $\hat{\nabla}_\theta J(\cdot), \hat{\nabla}_\theta C(\cdot)$

- $n \geq \frac{\eta^2 \ln(\alpha^{-1})}{\epsilon^2 (C(\theta))^4} \implies$

$$\mathbb{P}(\|\hat{\nabla}_\theta \tilde{J}(\theta) - \nabla_\theta \tilde{J}(\theta)\| \leq \epsilon) \geq 1 - \alpha$$

3. Ensuring safety of iterates with high probability

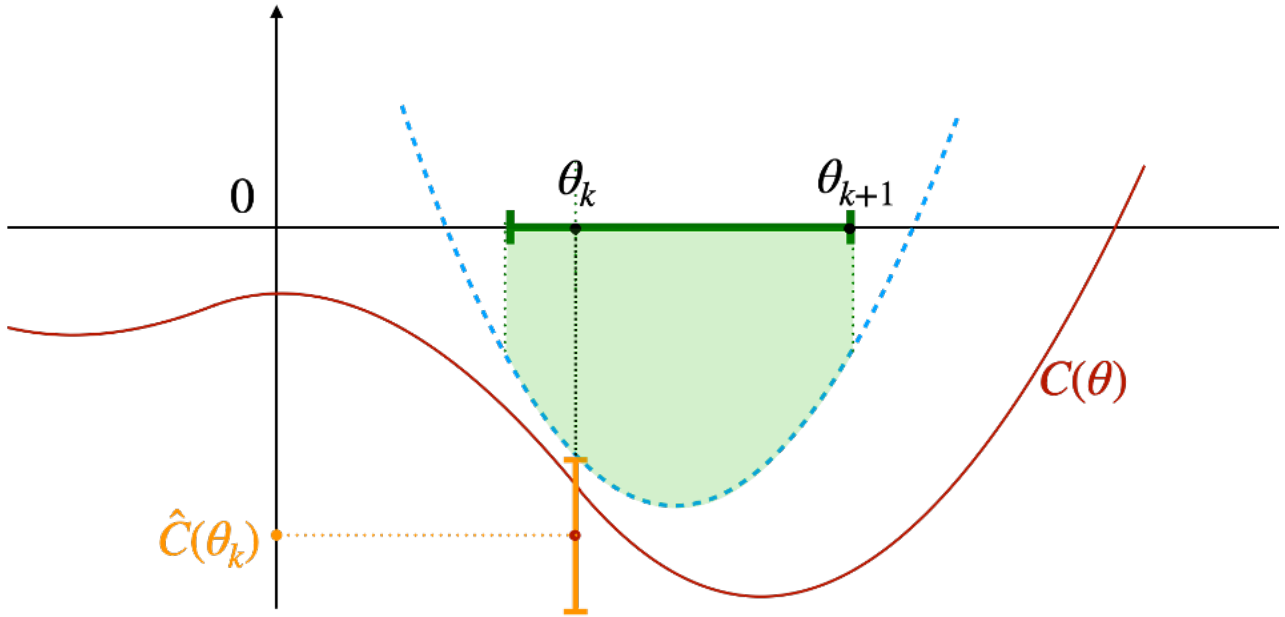
$$\theta(k+1) = \theta(k) - \boxed{\gamma(k)}\hat{\nabla}_\theta \tilde{J}(\theta(k))$$

$\gamma(k)$ should be

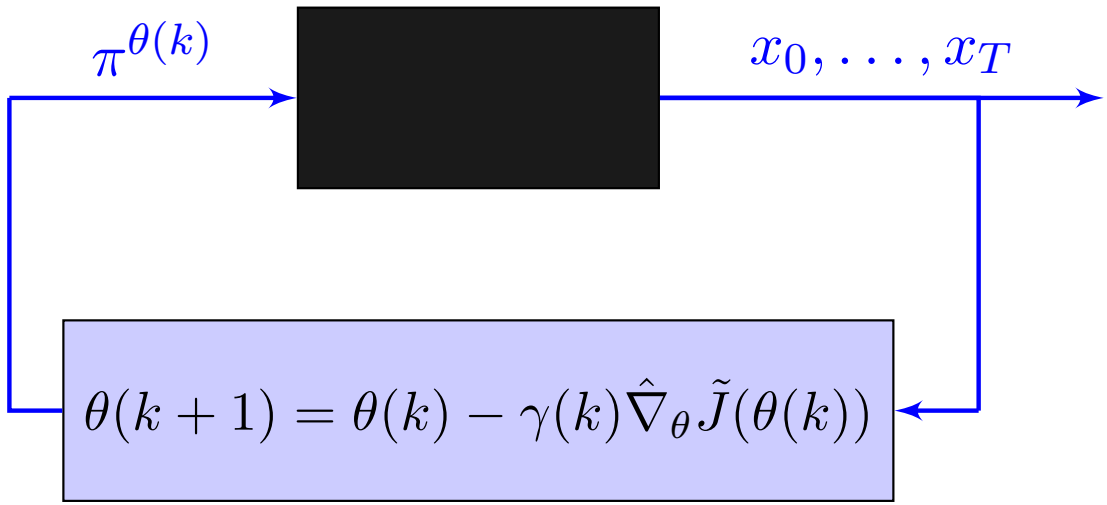
- sufficiently large to make progress
- sufficiently small to keep iterates safe

Approach

- Derive high probability local quadratic bounds on the objective and constraint



Theoretical guarantees for log barrier policy gradient



Theorem

With suitable choices of $\gamma(k), \eta$, we have

- *safety: policies $\pi^{\theta(k)}$ satisfy constraints with high probability*
- *convergence: $J(\theta(T)) < J(\pi^*) + \epsilon$, with $T = \tilde{O}(\epsilon^{-2})$, and $n = \tilde{O}(\epsilon^{-4})$ samples per trajectory*

[Usmanova, As, **MK***, Krause, JMLR 2024], [Ni,* **MK***, AISTATS 2025, HSCC 2025]*

Take-away message

- Can use the policy gradient approach to solve safety-constrained RL
- Provable safety (with high probability) and convergence even in continuous state and action spaces
- High sample complexity

Table of Contents

Summary

- The cost and constraints in reinforcement learning are important
  - discounted/finite horizon costs don't ensure stability or safety
- Safety and reachability constraints can be formulated through a constrained Markov decision process over an extended state-space
- Safe learning algorithms can ensure high probability safety during learning
  - increases trust in training the algorithm on real system

Ongoing work

Bridging the theory-practice gap

- Verification of the algorithm in realistic simulation
- Verification of the algorithm on hardware
- Improving sample efficiency of algorithms
- Generalization and robustness to mismatch

References

- Piunovskii, A. B. "Control of random sequences in problems with constraints.", Theory of Probability & Its Applications (1994).
- Sutton, Richard S., et al. "Policy gradient methods for reinforcement learning with function approximation.", NeurIPS (1999).
- Abate, Alessandro, et al. "Probabilistic reachability and safety for controlled discrete time stochastic hybrid systems.", Automatica (2008).
- Summers, Sean, et al. "A stochastic reach-avoid problem with random obstacles". HSCC (2011).
- Kamgarpour, Maryam, et al. "Control design for specifications on stochastic hybrid systems.", HSCC (2013).
- Littman, Michael L., et al. "Environment-independent task specifications via GLTL.", arXiv preprint arXiv:1704.04341 (2017).
- Liu, Yongshuai, et al. "IPO: Interior-point policy optimization under constraints.", AAAI (2020).
- Rui Yuan, et al. "A General Sample Complexity Analysis of Vanilla Policy Gradient.", AISTATS (2022).
- Archana, Bura, et al. "DOPE: Doubly optimistic and pessimistic exploration for safe reinforcement learning.", NeurIPS (2022).
- Alur, Rajeev, et al. "A framework for transforming specifications in reinforcement learning.", Principles of Systems Design: Essays Dedicated to Thomas A. Henzinger on the Occasion of His 60th Birthday. Cham: Springer Nature Switzerland (2022).
- Ding, Dongsheng, et al. "Last-iterate convergent policy gradient primal-dual methods for constrained MDPs.", NeurIPS (2023).
- Klein, Sara, et al. "Beyond stationarity: Convergence analysis of stochastic softmax policy gradient methods.", ICLR (2024).
- Usmanova, Ilnura, et al. "Log barriers for safe black-box optimization with application to safe reinforcement learning.", JMLR (2024).
- Ni, Tingting, et al. "Interior point constrained reinforcement learning with global convergence guarantees.", AISTATS (2025).
- Ni, Tingting, et al. "A learning-based approach to stochastic optimal control under reach-avoid constraint.", HSCC (2025).
- Schmid, Niklas, et al. "Computing optimal joint chance constrained control policies.", IEEE TAC (2025).

Acknowledgments

- This talk: PhD students Tingting Ni, Ilnura Usmanova, collaborators: Yarden As, Andreas Krause
- Funding : ERC, NSERC Canada, Swiss National Fund, NCCR Automation

Group members: R. Ouhamma, A. Schlaginhaufen, A. Maddux, K. Ren, G. Vallat, G. Salizzoni, T. Ni, S. Vaishampayan, P. Jordan

`https://www.epfl.ch/labs/sycamore/`

Costs in reinforcement learning - what else can go wrong

- Controllability: for any $x_0, x_1 \in \mathbb{R}^n$ there exist an input that can steer the state from x_0 to x_1 in finite time
- Observability: Given $x_0 \in \mathbb{R}^n$ can we determine the state trajectory from observing the cost $c(x_t, u_t)$ over a finite horizon
- reduce to algebraic tests for linear dynamical systems
- if we don't have either of them, we cannot have guarantees for stability of an RL approach, even with a non-discounted cost
  - easy to construct counterexamples with linear quadratic control

Constrained MDP

$$\begin{aligned} \min_{\pi} \quad & \mathbb{E}_{P,\pi}\left[\sum_{t=0}^{T-1} c(x_t, u_t) + c_T(x_T)\right] \\ \text{s.t.} \quad & \mathbb{E}_{P,\pi}\left[c_{s,T}(x_T, q_T)\right] \geq 1 - \delta \end{aligned}$$

Theorem

Under the assumptions:

- *Compact action space*
- $P(x'|x,u)$ *and* $c(x,u)$ *are continuous with respect to* u
- $\exists \pi$ *s.t.* $\mathbb{P}_\pi(\mathcal{C}_{S,G}) > 1 - \delta$

There exists an optimal Markov policy for the above [AB Piunovskii, 1994]

Extra notations

- Occupancy measure

$$\rho_t^\pi(\tilde{s}, a) = \mathbb{E}_{\tau \sim P_{\tilde{s}_0}^\pi}\left[\mathbf{1}_{(\tilde{s}_t, a_t)=(\tilde{s}, a)}\right]$$

$$= \begin{cases} \pi_0(a|s_0)\mathbf{1}_{\tilde{s}=\tilde{s}_0},\ t=0, \\ \pi_t(a|s)\int_{(\tilde{s}', a') \in \tilde{\text{s.t.}} \times \mathcal{A}} \rho_{t-1}^\pi(d\tilde{s}', da')P(\tilde{s}|d\tilde{s}', da'), \text{ otherwise.} \end{cases}$$

- Sample complexity
  $n = \mathcal{O}(\epsilon^{-4} \ln \frac{1}{\beta\epsilon})$ and $H = \mathcal{O}(\ln \frac{1}{\epsilon})$, and $T = \mathcal{O}(\epsilon^{-2})$ to ensure optimality and safe exploration with confidence $1 - \beta$