Safety and Reliability in Reinforcement Learning

Thiago D. Simão

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

10 Apr 2025

Gap between research and real world

Simulations

Real-world tasks

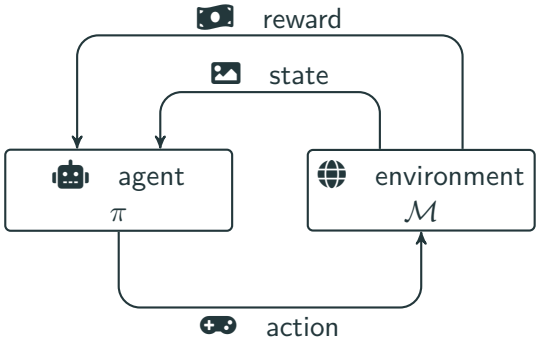
Reinforcement Learning (RL)

Markov decision process (MDP)[1]:
$\mathcal{M} = \langle S, A, \mathcal{T}, \mathcal{R}, \gamma \rangle$

- $\mathcal{T}: S \times A \to \Delta(S)$
- $\mathcal{R}: S \times A \to \mathbb{R}$
- γ: discount factor
- $\pi: S \to \Delta(A)$

Return

$$\arg\max_{\pi} \mathbb{E}_{\pi}\left[\sum_{t=1} \gamma^t R_t\right]$$



The RL agent has no knowledge about the problem, so it must learn from experiences

[1] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1st. John Wiley & Sons, Inc., 1994

Challenges for RL

Challenges to bring RL from research to real-world applications[2]:

- Safety constraints
- Off-line training
- Limited interactions with the environment
- Partially observable tasks
- Explainability
- ...

[2] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester. "Challenges of real-world reinforcement learning: definitions, benchmarks and analysis". In: *Mach. Learn.* 110.9 (2021), pp. 2419–2468

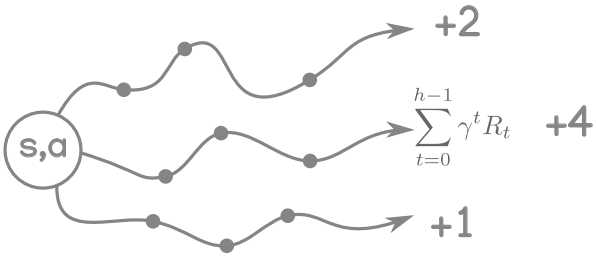
Safe Reinforcement Learning

"Safe Reinforcement Learning can be defined as the process of learning policies that **maximize the expectation of the return** in problems in which it is important to **ensure reasonable system performance** and/or **respect safety constraints** during the learning and/or deployment processes."[3]

[3] J. García and F. Fernández. "A Comprehensive Survey on Safe Reinforcement Learning". In: *Journal of Machine Learning Research* 16 (2015), pp. 1437–1480

- Ensure reasonable system performance.
- Respect safety constraints

Distribution of Returns



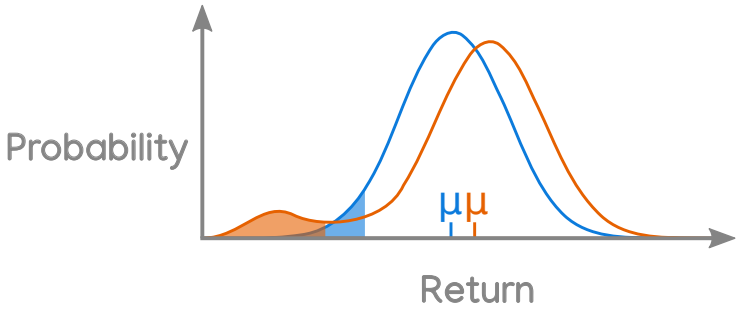
Trajectories and returns[4]

Sources of aleatoric uncertainty

- transition function
- initial states distribution
- stochastic policy

[4] R. Munos. *Distributional Reinforcement Learning*. Horizon Maths `https://vimeo.com/304849090`. 2018

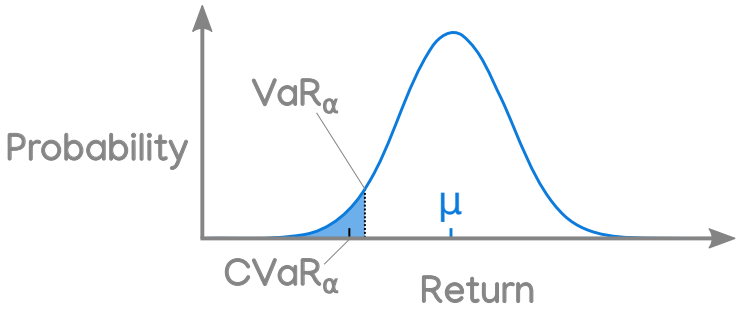
Average vulnerability ⚠



Routing

A direct route through the city center might have the lowest expected duration, but it can sporadically cause significant delays due to traffic.

Conditional value at risk (CVaR)



$\mathrm{CVaR}_\alpha(Z)$ can be interpreted as the expected value of the α-portion of the left tail of the distribution of Z.

α provides control over the risk level.

Handling aleatoric uncertainty

The default criterion only considers the mean value.

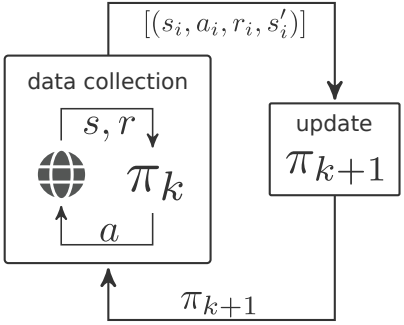
$$\max_{\pi} \mathbb{E}\left[\sum_{t=1} \gamma^t R_t\right].$$

We would like a more risk averse criterion, such as the CVaR:

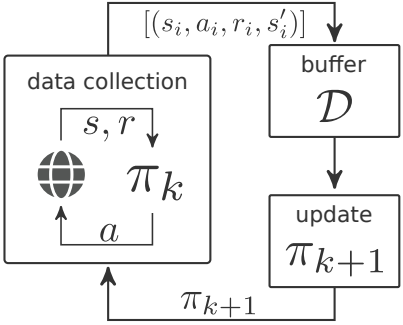
$$\max_{\pi} \mathrm{CVaR}_{\alpha}\left[\sum_{t=1} \gamma^t R_t\right],$$

$\mathrm{CVaR}_{\alpha}[Z] = \mathbb{E}[Z \mid Z \leq \mathrm{VaR}_{\alpha}(Z)]$ and $\alpha \in (0, 1]$.

Typical Reinforcement Learning: Online Approach



on-policy



off-policy

Opportunities

Direct interactions with the environment:

- expensive
- time consuming

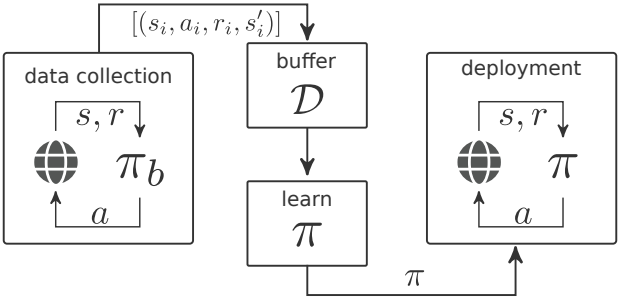
Many applications already have historical data available.

Offline RL

- Use historical data for RL.[5]
- Avoid direct interactions with the environment.

[5] S. Levine, A. Kumar, G. Tucker, and J. Fu. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems". In: *arXiv preprint arXiv:2005.01643* (2020)

Offline RL

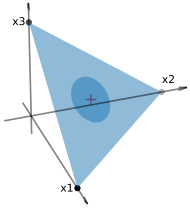

$[(s_i, a_i, r_i, s_i')]$
data collection
s, r
π_b
a
buffer
$\mathcal{D}$
learn
π
deployment
s, r
π
a
π

Handling epistemic uncertainty

▲ : $\Delta(S)$

$+$: $\hat{\mathcal{T}}(s' \mid s, a) = \frac{N(s,a,s')}{N(s,a)}$

● : $\Sigma = \left\{\mathcal{T}' \in \Delta(S) \mid \|\hat{\mathcal{T}}(\cdot \mid s, a) - \mathcal{T}'(\cdot \mid s, a)\| \leq \underbrace{e(s,a)}_{\approx \frac{1}{N(s,a)}}\right\}$



Σ is the set of probable transition functions and $\hat{\mathcal{T}}$ is the MLE of $\mathcal{T}$.

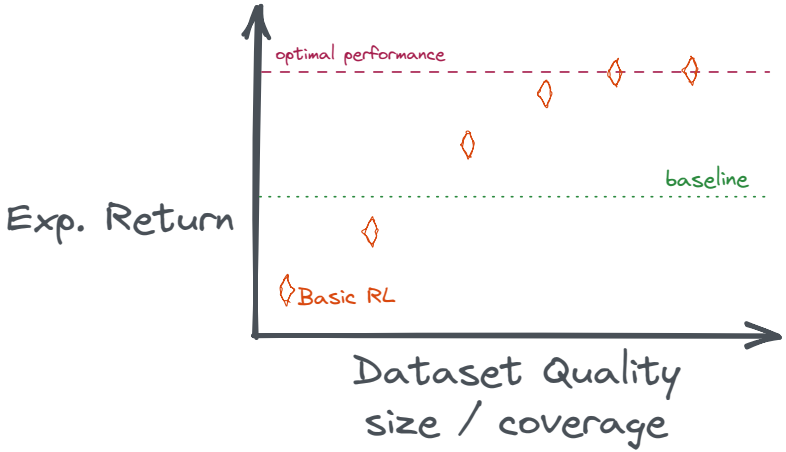
Worst-case criterion[6,7]

$$\max_{\pi} \min_{\mathcal{T}' \in \Sigma} \mathop{\mathbb{E}}_{\pi, \mathcal{T}'} \left[\sum_{t=0}^{\infty} \gamma^t R_t \right].$$

[6] A. Nilim and L. El Ghaoui. "Robust Control of Markov Decision Processes with Uncertain Transition Matrices". In: *OR* 53.5 (2005), pp. 780–798

[7] M. Suilen, T. D. Simão, D. Parker, and N. Jansen. "Robust Anytime Learning of Markov Decision Processes". In: *NeurIPS*. 2022, pp. 28790–28802

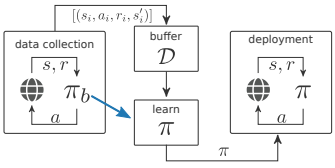
Reliability of Offline RL

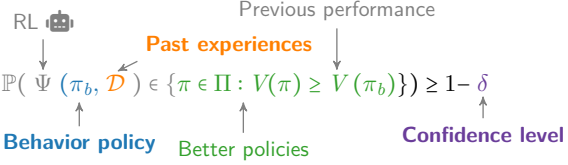


Safe Policy Improvement

The safety criterion

Safe Policy Improvement





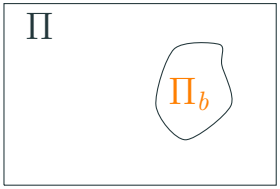
→ 🤖 Given the dataset $\mathcal{D}$ and the behavior policy π_b,

🤖 → **reliably** compute a policy π that outperforms π_b.

Safe Policy Improvement With Baseline Bootstrapping (SPIBB)[8]

Min number of samples to stop bootstrapping s, a

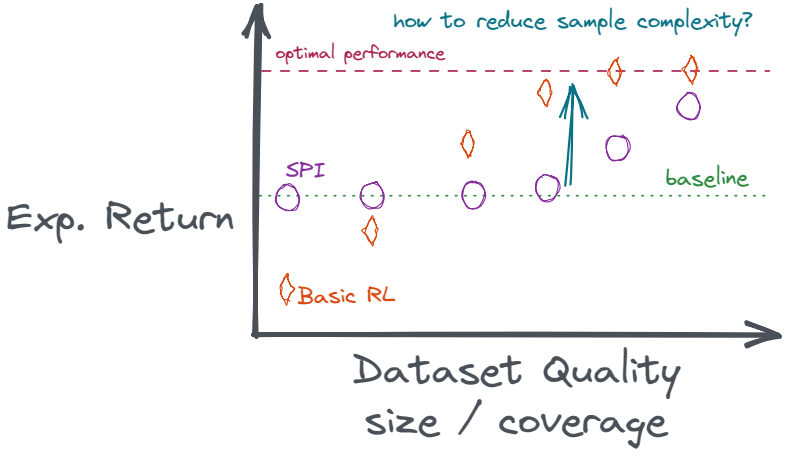
$$\Pi_b = \{\pi \in \Pi \,|\, \pi(a|s) = \pi_b(a|s) \text{ if } N_{\mathcal{D}}(s,a) < N_\wedge\}$$



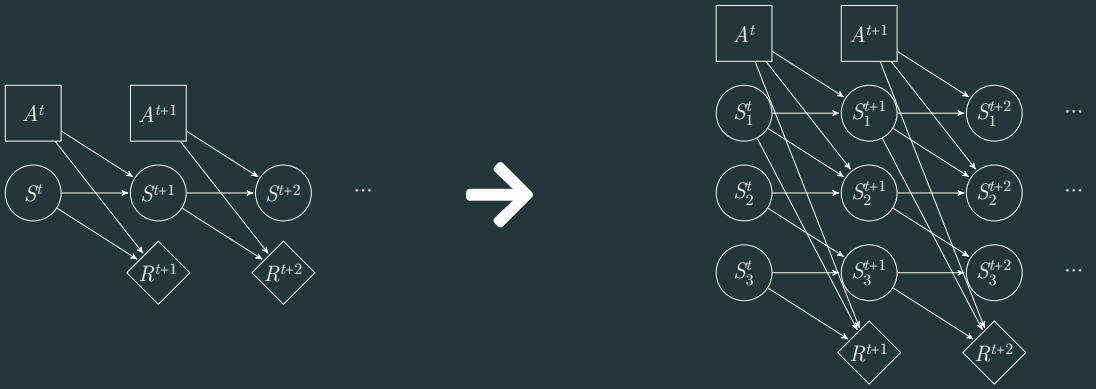
$$\Psi(\pi_b, \mathcal{D}) = \arg\max_{\pi \in \Pi_b} V(\pi, \hat{\mathcal{T}})$$

[8] R. Laroche, P. Trichelair, and R. Tachet des Combes. "Safe Policy Improvement with Baseline Bootstrapping". In: *ICML*. PMLR, 2019, pp. 3652–3661

Safe Policy Improvement

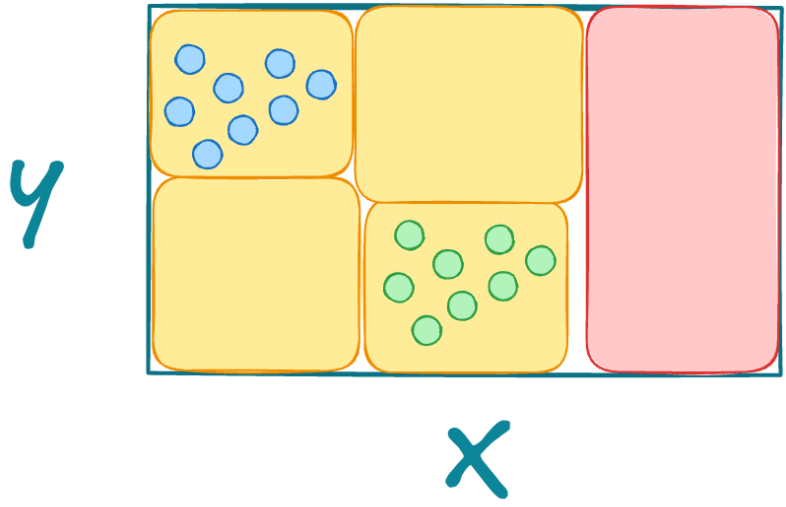


From flat to factored representation.



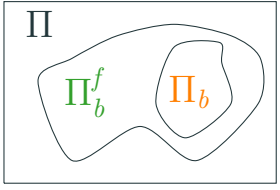
$$\mathcal{T}(s' \mid s, a) = \prod_{i=1}^{n} \mathcal{T}(s'[i] \mid s[\mathbf{Pa}_a(i)], a)$$

Exploiting the factored structure



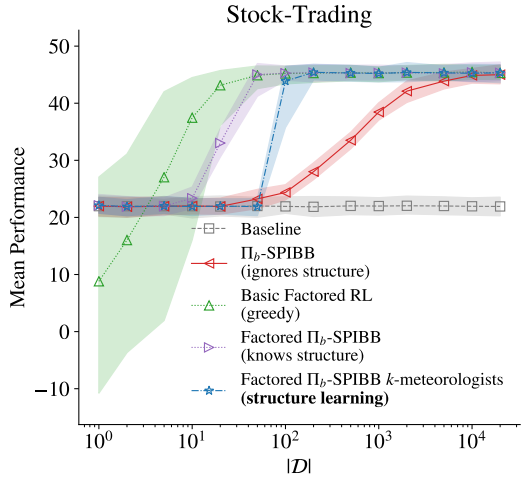
Factored Baseline Bootstrapping

$$\Pi_b^f = \left\{ \pi \,\middle|\, \pi(a|s) = \pi_b(a|s) \text{ if } \exists i : N_{\mathcal{D}}(s[\mathrm{Pa}_a(i)], a) < N_\wedge^i \right\}$$



$$\Psi(\pi_b, \mathcal{D}) = \arg\max_{\pi \in \Pi_b^f} V(\pi, \hat{T})$$

Improved Sample Complexity[9]



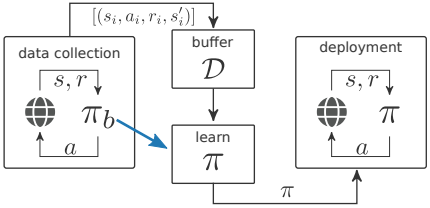
[9] T. D. Simão and M. T. J. Spaan. "Structure Learning for Safe Policy Improvement". In: *IJCAI*. ijcai.org, 2019, pp. 3453–3459

Overview

Safe Policy Improvement ensures a reasonable performance in offline RL with respect to the behavior policy.

further considerations

- Generalization[10]
- Unknown behavior policy[11]
- Partial-observability[12]
- Sample-complexity[13]
- Scalability[14]
- Multi-agents[15]



[10] T. D. Simão and M. T. J. Spaan. "Safe Policy Improvement with Baseline Bootstrapping in Factored Environment". In: *AAAI*. 2019, pp. 4967–4974

[11] T. D. Simão, R. Laroche, and R. Tachet des Combes. "Safe Policy Improvement with an Estimated Baseline Policy". In: *AAMAS*. 2020, pp. 1269–1277

[12] T. D. Simão, M. Suilen, and N. Jansen. "Safe Policy Improvement for POMDPs via Finite-State Controllers". In: *AAAI*. 2023, pp. 15109–15117

[13] P. Wienhöft, M. Suilen, T. D. Simão, C. Dubslaff, C. Baier, and N. Jansen. "More for Less: Safe Policy Improvement With Stronger Performance Guarantees". In: *IJCAI*. 2023, pp. 4406–4415

[14] A. Castellini, F. Bianchi, E. Zorzi, T. D. Simão, A. Farinelli, and M. T. J. Spaan. "Scalable Safe Policy Improvement via Monte Carlo Tree Search". In: *ICML*. 2023, pp. 3732–3756

[15] F. Bianchi, E. Zorzi, A. Castellini, T. D. Simão, M. T. J. Spaan, and A. Farinelli. "Scalable Safe Policy Improvement for Factored Multi-Agent MDPs". In: *ICML*. 2024

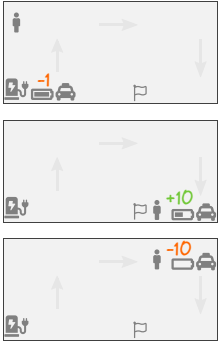
- Ensure reasonable system performance
- Respect safety constraints

Constrained Reinforcement Learning

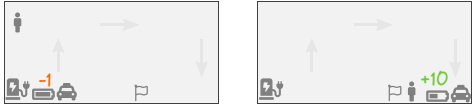
Learning through experience

Decoupling safety from the reward

- Reward might be inadequate to specify the desired behavior.
- We need mechanisms to express safe behaviors explicitly.



Reward: dedicated to main task



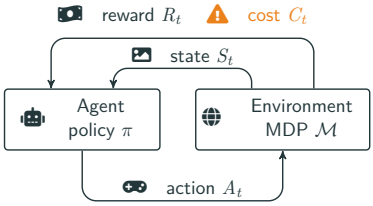
Cost: indicates unsafe interactions

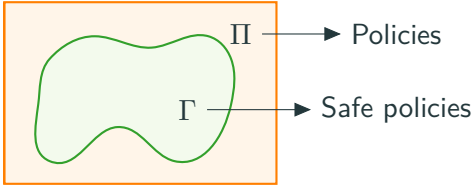
Constrained RL

Constrained MDP[16]: $\mathcal{M} = \langle S, A, \mathcal{T}, \mathcal{R}, \mathcal{C}, \hat{c} \rangle$

- $\mathcal{C} : S \times A \to \mathbb{R}$
- $\hat{c}$: cost bound

$$\max_{\pi} \mathbb{E}_{\pi}\left[\sum_t R_t\right]$$
$$\text{s.t. } \underbrace{\mathbb{E}_{\pi}\left[\sum_t C_t\right] \leq \hat{c}}_{\text{Safety constraint}}$$





[16] E. Altman. *Constrained Markov Decision Processes*. Vol. 7. CRC Press, 1999

How to learn while satisfying the safety constraints?

- Rely on an initial safe policy[17]
- Use prior knowledge about safety dynamics[18]

Alternatives

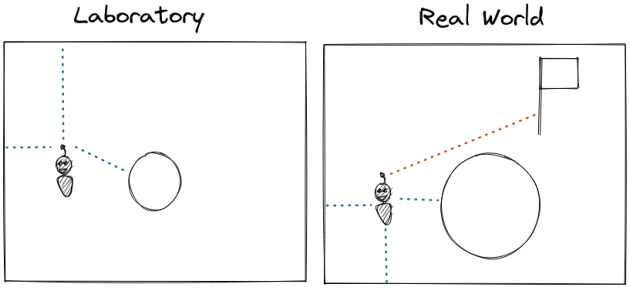
- Safe Transfer
- Curriculum Learning

[17] J. Achiam, D. Held, A. Tamar, and P. Abbeel. "Constrained Policy Optimization". In: *ICML*. 2017

[18] T. D. Simão, N. Jansen, and M. T. J. Spaan. "AlwaysSafe: Reinforcement Learning Without Safety Constraint Violations During Training". In: *AAMAS*. 2021, pp. 1226–1235

Safe Transfer

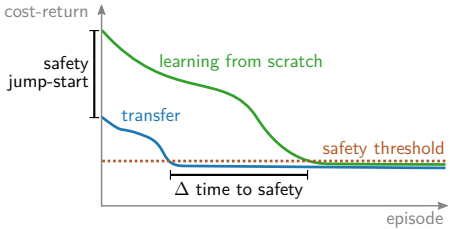
Transfer from lab to real world



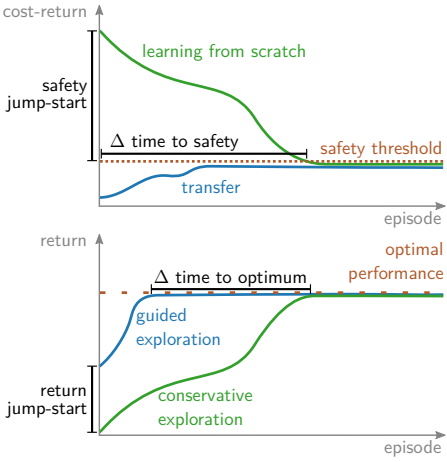
	source	target
costs	✓	✓
safety violation	✓	✗
rewards	✗	✓

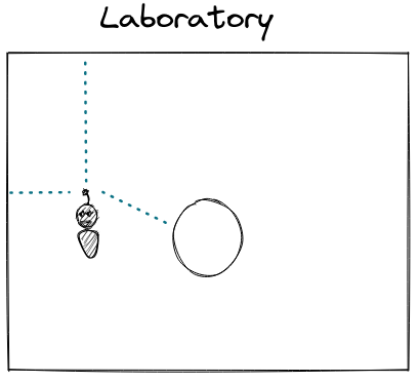
Safe-Transfer Metrics

Unsafe Transfer



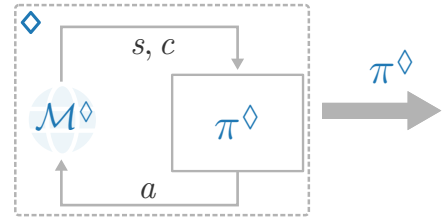
Safe Transfer





Part I: Learning in the Lab

Learning in the Lab



transfer

Source task ($\Diamond$) objective

Training $\pi^{\Diamond}$ to explore safely.

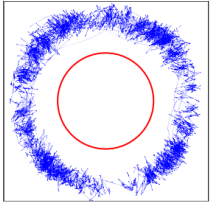
$$\max_{\pi} \mathop{\mathbb{E}}_{(s_t,a_t)\sim\rho_\pi} \left[\sum_t \overbrace{\mathcal{H}(\pi(\cdot \mid s_t))}^{\text{Policy Entropy}} + \overline{\alpha} \overbrace{r'(s_t, a_t)}^{\text{Novelty Bonus}} \right]$$

$$\text{s.t.} \quad \mathop{\mathbb{E}}_{(s_t,a_t)\sim\rho_\pi} \left[\sum_t c(s_t, a_t) \right] \leq d$$

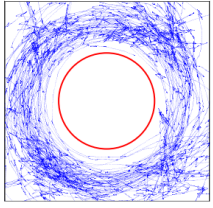
$$r'(s_t, a_t) = \mathbb{E}\Big[\underbrace{\delta(s_t, s_{t+1})}_{\text{distance between states}} \mid s_{t+1} \sim \mathcal{T}(\cdot \mid s_t, a_t) \Big]$$

Exploration Analysis

Single Obstacle

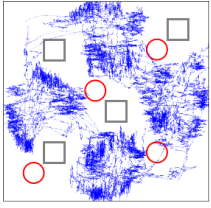


Entropy

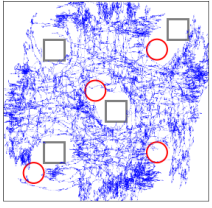


Entropy + Novelty

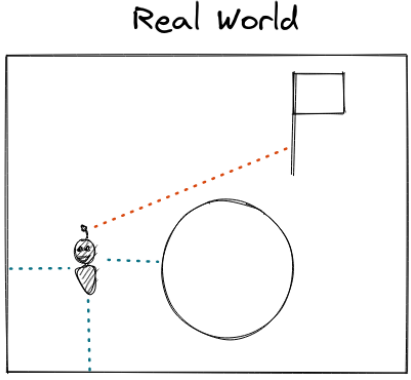
Multiple Obstables



Entropy



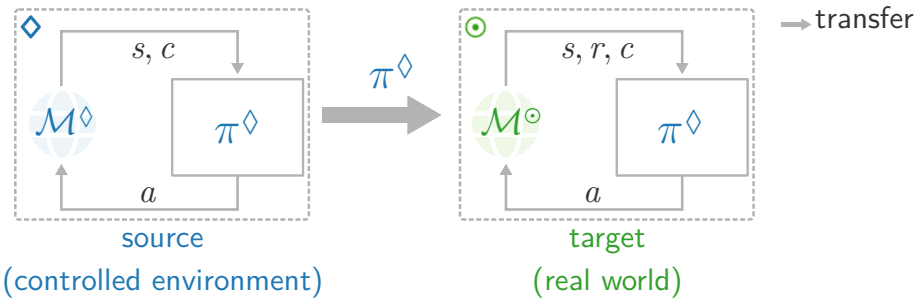
Entropy + Novelty



Part II: Learning in the Real Wolrd

Direct Transfer (Pre-training)

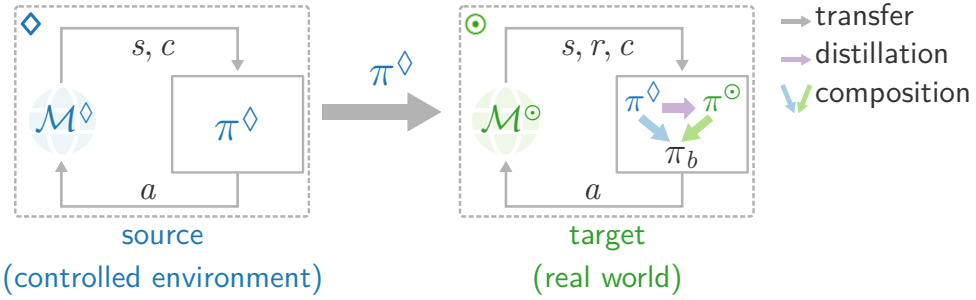
Adapt the trained policy ($\pi^{\diamond}$) to the target task ($\odot$).



Challenge: $\pi^{\diamond}$ might overfit to the source task and fail to solve the target task.

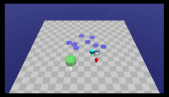
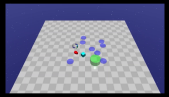
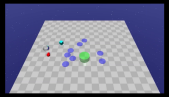
SaGui: Safe Guide[19]

Proposal: Train a student ($\pi^{\odot}$) in the target task ($\odot$) using the guide's advice ($\pi^{\diamond}$).



[19] Q. Yang, T. D. Simão, N. Jansen, S. H. Tindemans, and M. T. J. Spaan. "Reinforcement Learning by Guided Safe Exploration". In: *ECAI*. 2023, pp. 2858–2865

Final Policy
Point Goal with Dynamic Obstacles

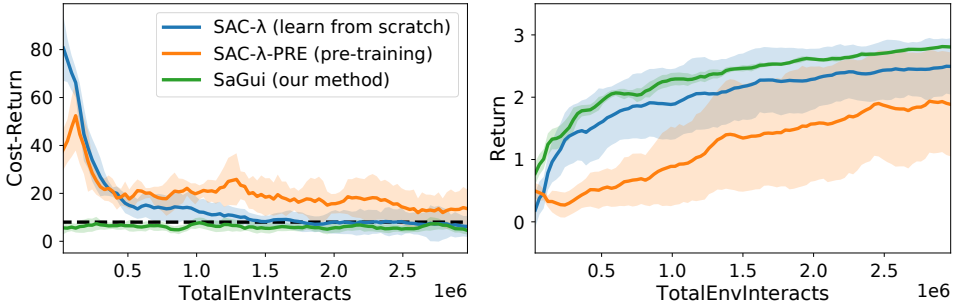




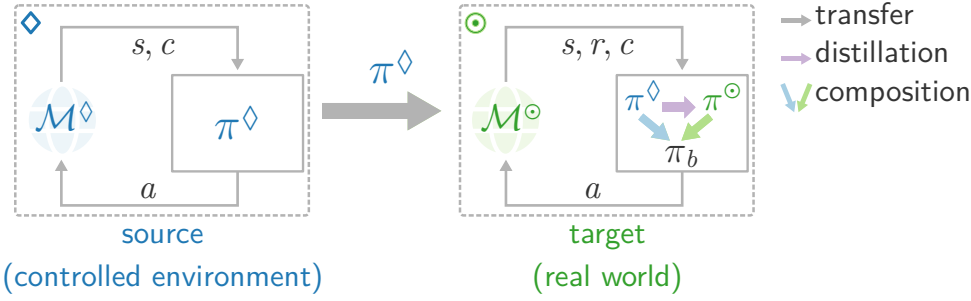
Learned
From Scratch

Pre-trained

Safe
Guide

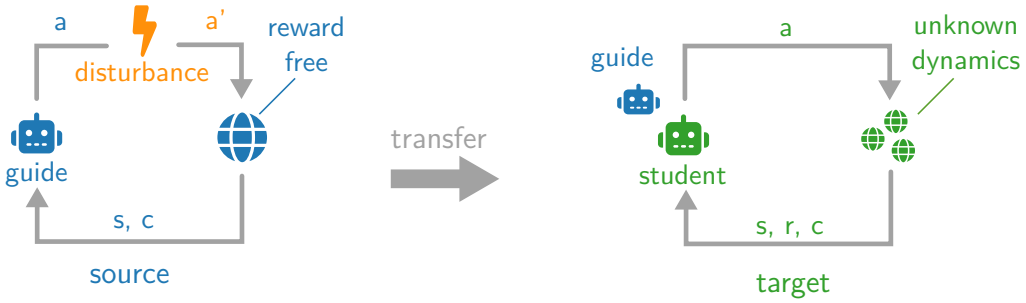
Empirical Analysis





What if the targe task is different from the source task?

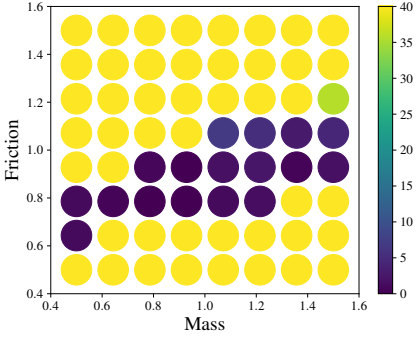
Training the guide to become robust[20]



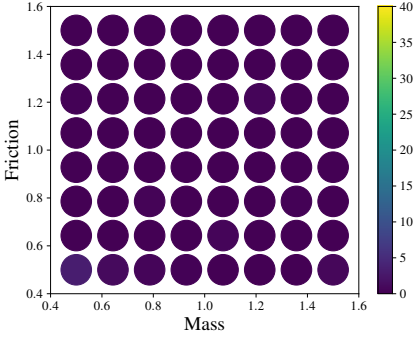
[20] M. Zubia, T. D. Simão, and N. Jansen. "Robust Transfer of Safety-Constrained Reinforcement Learning Agents". In: *ICLR*. 2025

Robustness heatmaps

- Guide trained with and without adversarial perturbations.
- Evaluate within target task's uncertainty set.



$\alpha = 0.$



$\alpha = 0.29.$

Expected cumulative cost

Curriculum Learning

Don't change the agent

Make the tasks safer

Curriculum Learning[21]

- Given a set of tasks and a target task,
- beginning with easy tasks, design a sequence of tasks in increasing difficulty
- to speed up the learning of the target task.

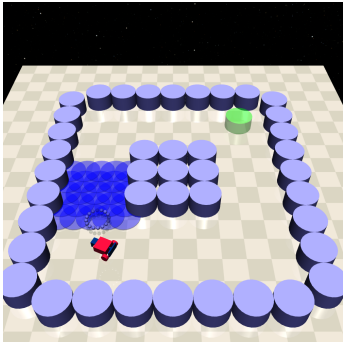
Can we design a sequence of tasks to improve safety during learning and speed up the learning?

[21] S. Narvekar, B. Peng, M. Leonetti, J. Sinapov, M. E. Taylor, and P. Stone. "Curriculum Learning for Reinforcement Learning Domains: A Framework and Survey". In: *J. Mach. Learn. Res.* 21 (2020), 181:1–181:50

Contextual Constrained RL

$$\mathcal{M} = \langle S, A, \mathcal{X}, \mathsf{M}, \hat{c} \rangle$$

- $\mathcal{X}$ is the context space
- M maps a context $\mathbf{x} \in \mathcal{X}$ to a constrained MDP $\mathsf{M}(\mathbf{x}) = \langle S, A, \mathcal{T}_{\mathbf{x}}, \mathcal{R}_{\mathbf{x}}, \mathcal{C}_{\mathbf{x}}, \hat{c} \rangle$
- Note: π is conditioned on the context
- φ is the target context distribution



Safety Maze

Objectives

Given a target distribution over the tasks φ, the agent should learn a policy π^*

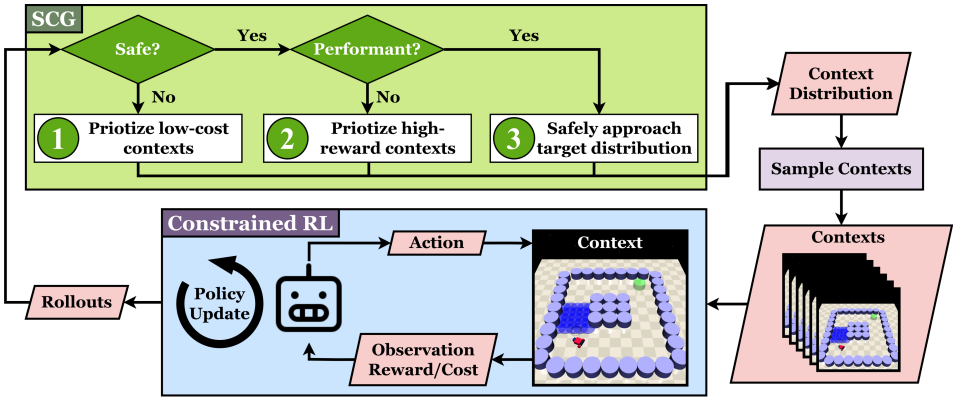
$$\pi^* \in \arg\max_{\pi} \mathbb{E}_{\varphi}[V_r^{\pi}(\mathbf{x})] \quad \text{s.t. } \mathbb{E}_{\varphi}[V_c^{\pi}(\mathbf{x})] \leq \hat{c}$$

The goal is to generate a sequence of context distributions $\{\varrho_l\}_{l=1}^{L}$ that allow an constrained RL agent to *sample-efficiently* learn π^* with minimal constraint violation regret:

$$\text{Reg}^{tr}(L, \{\varrho_l\}_{l=1}^{L}, \hat{c}) = \sum_{l=1}^{L} [\mathbb{E}_{\varrho_l}[V_c^{\pi_l}(\mathbf{x})] - \hat{c}]_+$$

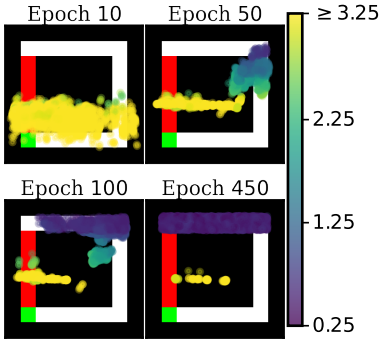
$[x]_+ = \max\{x, 0\}$

Safe Curriculum Generation[22]

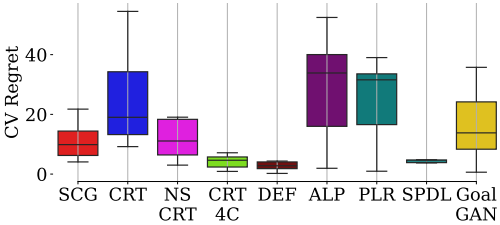


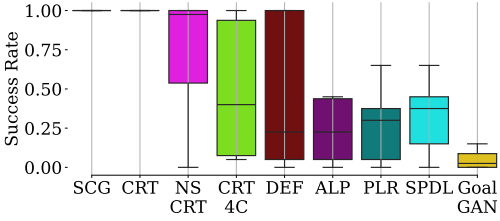
[22] C. Koprulu, T. D. Simão, N. Jansen, and U. Topcu. "Safety-Prioritizing Curricula for Constrained Reinforcement Learning". In: *ICLR*. 2025

Evaluation in the Safety Maze



Curriculum Progression





Overview

Constrained RL can

- model safety requirements explicitly, and
- automatically find a trade-off between safety and performance.

Transferring a safe exploration policy can

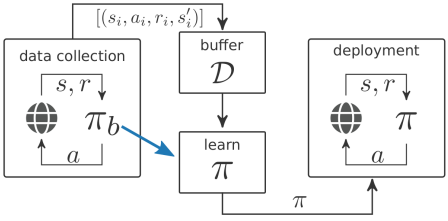
- provide safety on the target task, and
- make the exploration of the target task more effective.

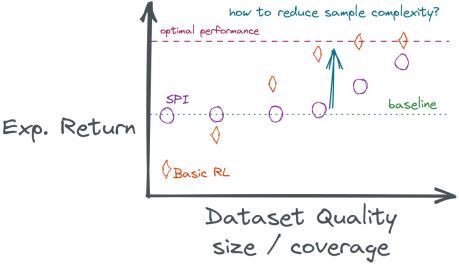
Safety-Prioritizing Curricula can

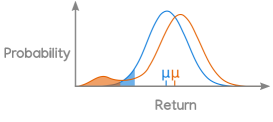
- improve learning on the target task distribution, and
- reduce the safety constraints during learning.

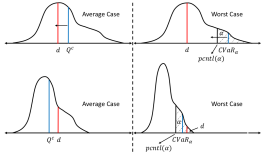
Safe RL

- Ensure reasonable system performance
  - Safe Policy Improvement
- Respect safety constraints
  - Safe Tranfer
  - Safe Curriculum Generation



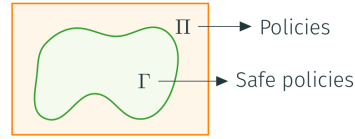


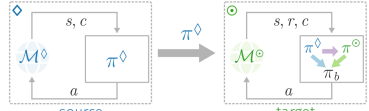


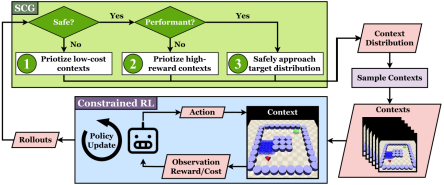


Reward: dedicated to main task

Cost: indicates unsafe interactions







Thank you!

Thiago D. Simão
t.simao@tue.nl
https://tdsimao.github.io