Robotics: Science and Systems (RSS) 2024

Jean-Baptiste Bouvier
Kartik Nagpal
Negar Mehr

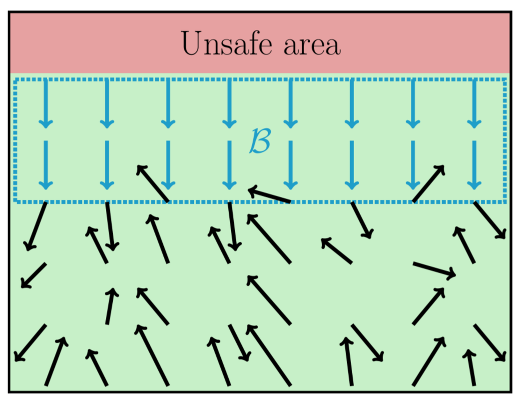
POLICEd RL: Learning Closed-Loop Robot Control Policies with Provable Satisfaction of Hard Constraints

Introduction

We propose **POLICEd RL**, a novel RL algorithm to **guarantee** satisfaction of an affine constraints in closed-loop with a black-box environment.

Key insights:

- make the learned policy affine around the unsafe area,
- use this affine region as a repulsive buffer to keep trajectories safe.

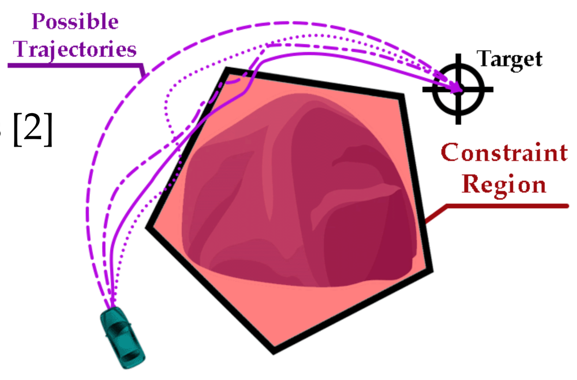


Enforcing constraints in RL

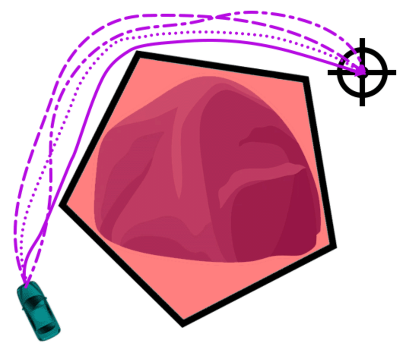
Soft Constraint
Constraint violations are penalized but no guarantees of respect.

Typical safe RL:
- reward shaping
- Constrained Markov Decision Processes [2]

no safety guarantees



Hard Constraint (Ours)
Adheres to constraints by construction, and guarantees no violations.

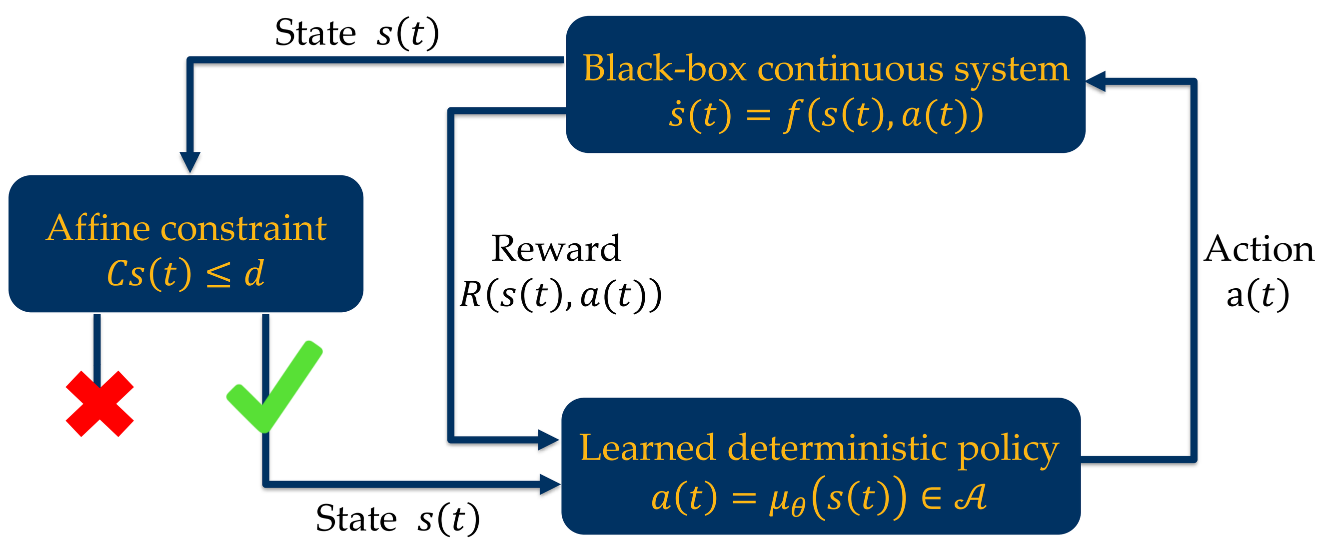


- HJB reachability
- Control Barrier Functions (CBFs)
- projection on safe set

all require *white-box dynamics*

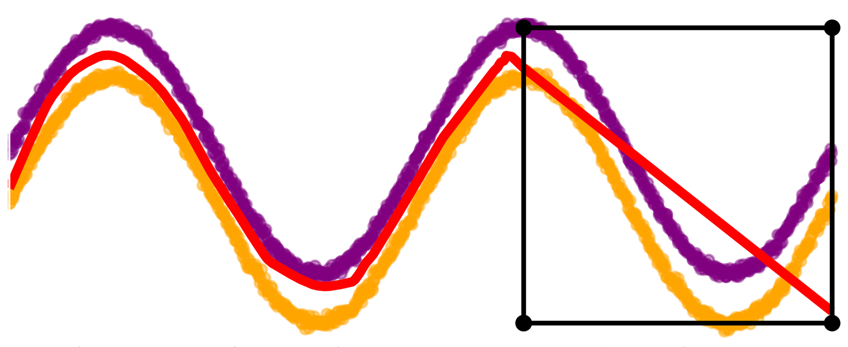
Learned CBF certificates and learned safety-critics provide *no safety guarantees*

Closed-loop constrained RL



The POLICE algorithm [2]

Bias modification to make a deep neural network affine in a user-provided region.



Classification of purple vs orange with red boundary, forced to be affine by POLICE [2] in **black** square.

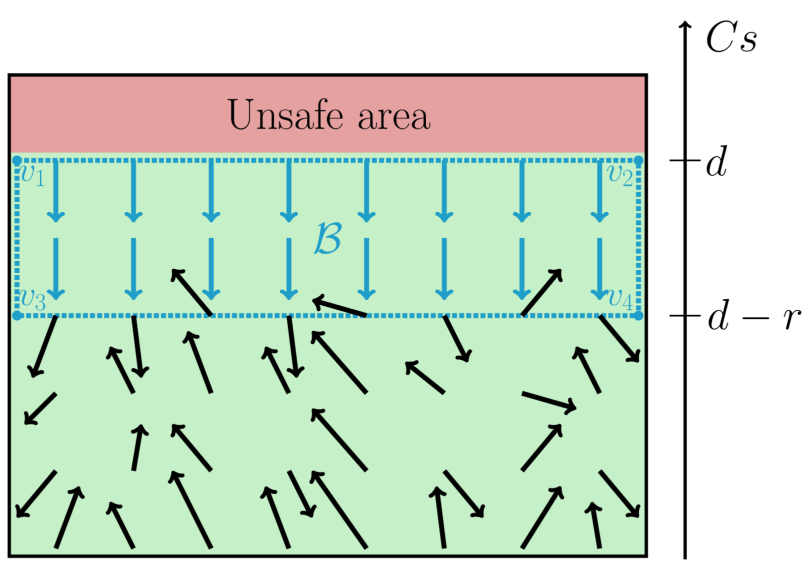
Our approach: *POLICEd RL*

Define a buffer $\mathcal{B} = \{s : Cs \in [d-r,\ d]\}$ of radius $r > 0$.

Use POLICE [2] to make policy μ_θ affine over buffer $\mathcal{B}$.

Estimate how far from affine are dynamics f with ε

$|Cf(s,a) - C(As + Ba + c)| \le \varepsilon$ for all $s \in \mathcal{B}$ and $a \in \mathcal{A}$.



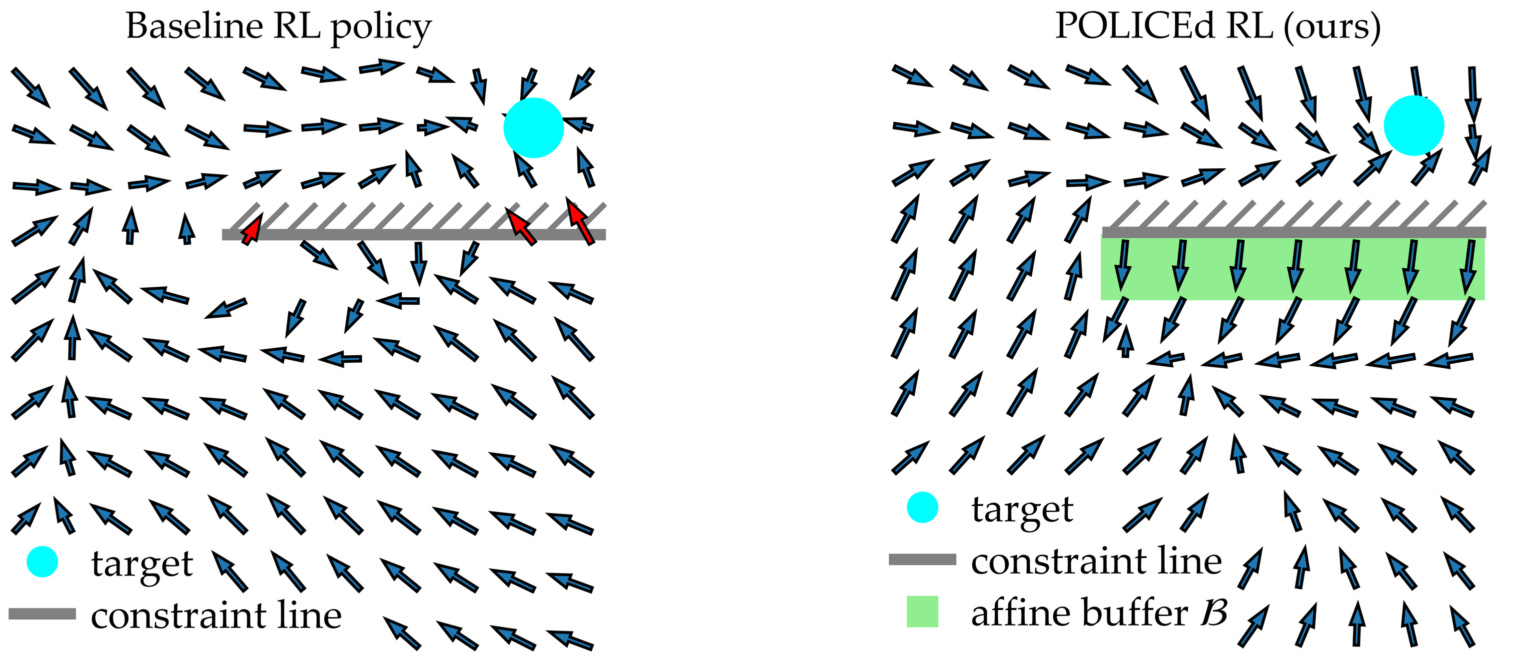
Theorem: If μ_θ is affine over $\mathcal{B}$ and for some affine measure ε, repulsion condition $Cf(v, \mu_\theta(v)) \le -2\varepsilon$ holds at all vertices v of $\mathcal{B}$, then $Cs(t) < d$ for all $t \ge 0$.

Algorithm

1. Calculate buffer radius r
2. Determine buffer $\mathcal{B}$ and its vertices
3. Sample transitions (s, a, s') with $s \in \mathcal{B}$ and estimate ε with least-square approximation
4. Train μ_θ until repulsion condition $Cf(s, \mu_\theta(s)) \le -2\varepsilon$ holds on the vertices of $\mathcal{B}$

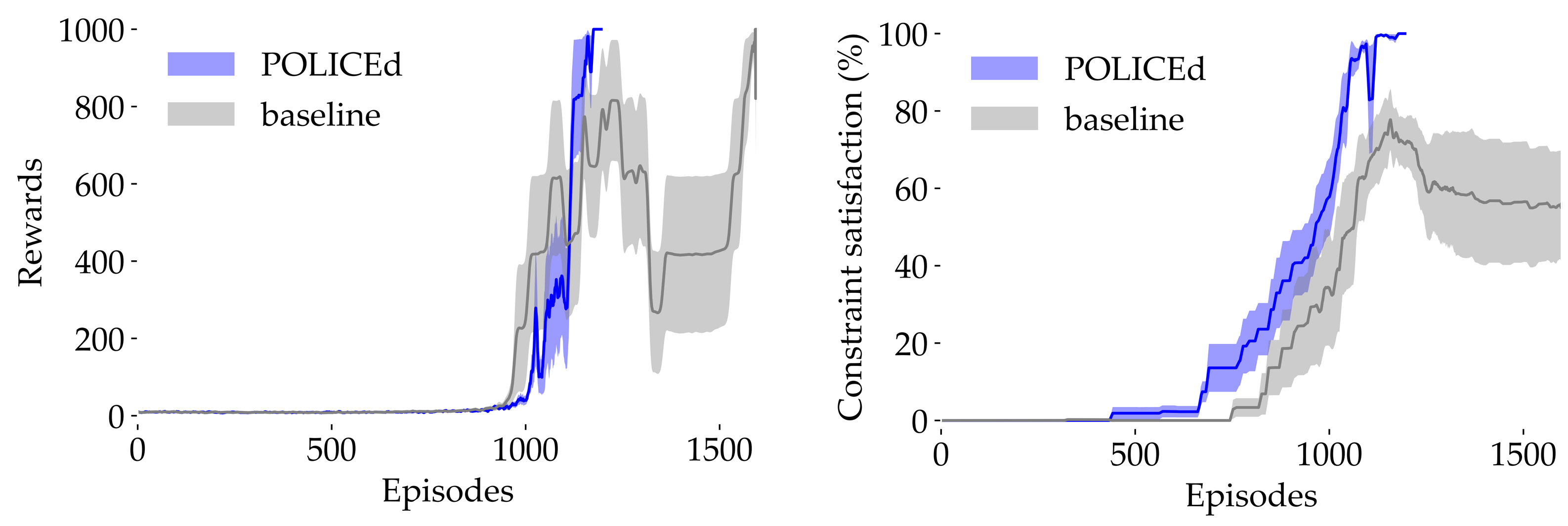
Guarantees $Cs(t) < d$ if $Cs(0) < d$.

2D illustrative example



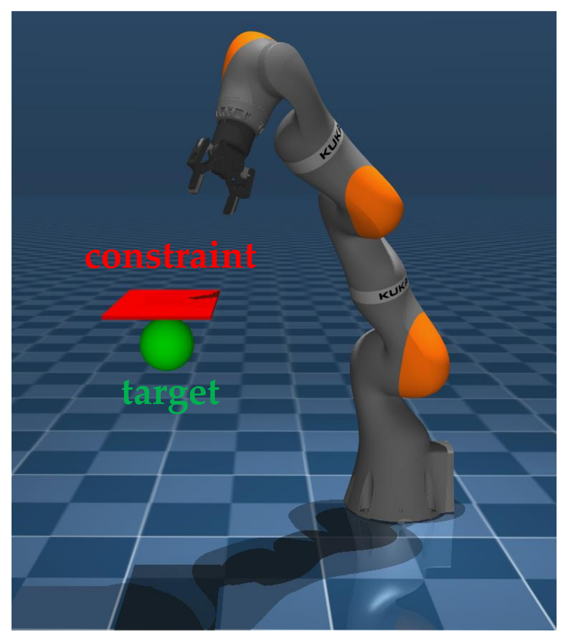
POLICEd RL learns to reach the target without any constraint violation thanks to its affine buffer.

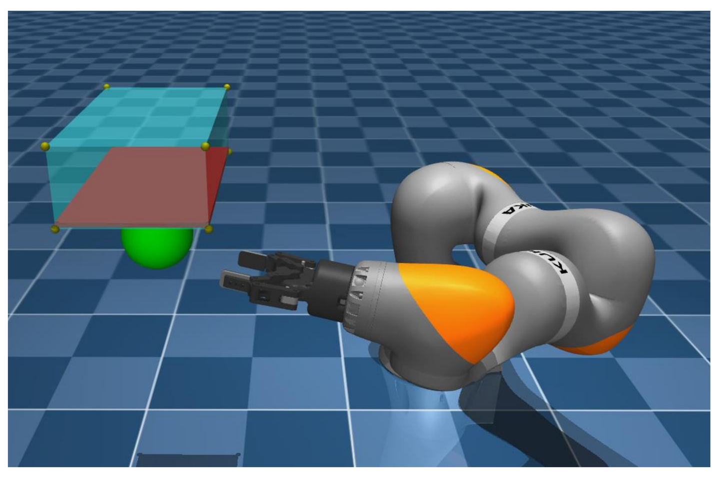
Stabilizing the MuJoCo inverted pendulum



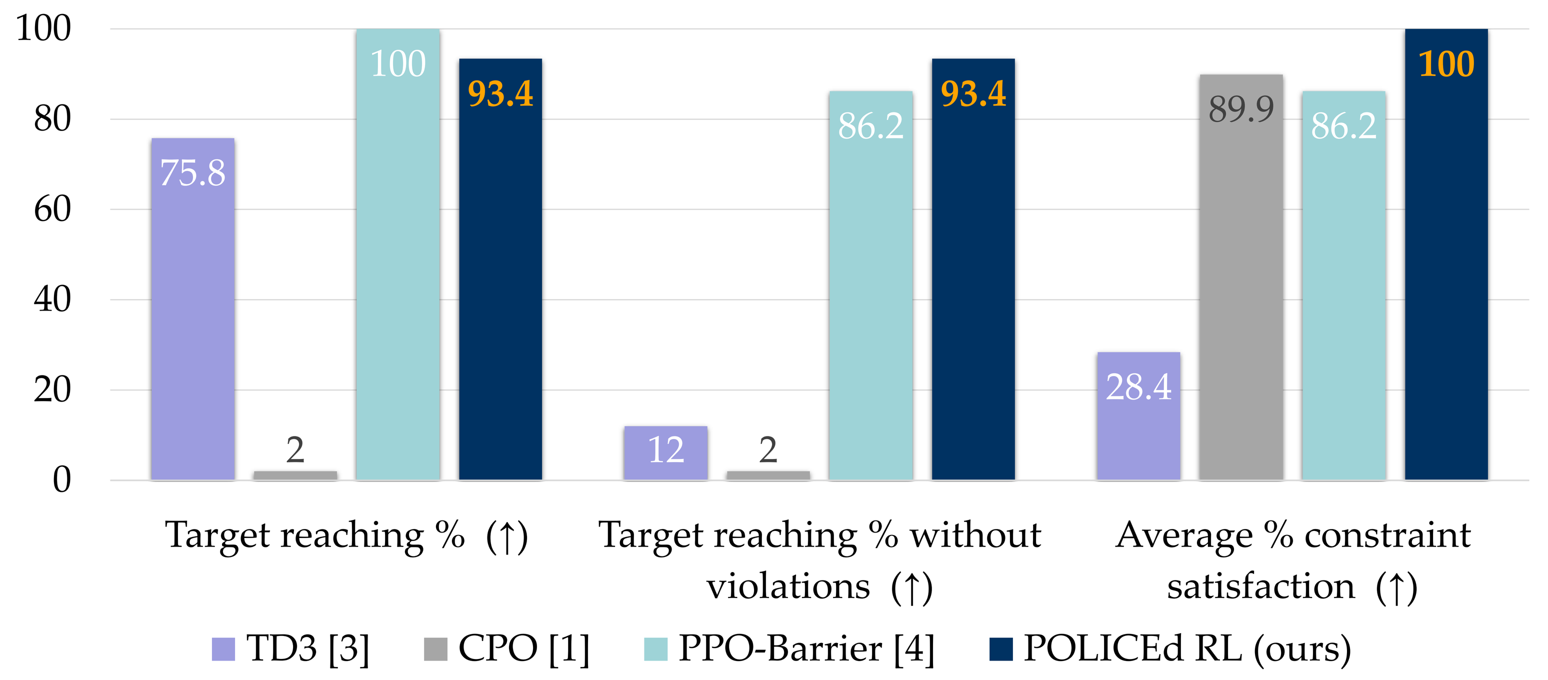
Objective: maintain $\theta \in [-\theta_{max},\ \theta_{max}]$
Constraint: $\dot{\theta} < 0$ near θ_{max} to prevent falling past θ_{max}

Reach-avoid with KUKA arm





POLICEd RL uses a buffer to push the KUKA arm away from the constraint



Conclusion

- POLICEd RL provably enforces an affine constraint
- Only requires a black-box model of the environment
- Tractable safety verification at the buffer vertices

References

Full text available on ArXiv at https://arxiv.org/pdf/2403.13297

[1] Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. "Constrained policy optimization." In *International Conference on Machine Learning*, pages 22 - 31, 2017.

[2] Randall Balestriero and Yann LeCun. "POLICE: Provably optimal linear constraint enforcement for deep neural networks." In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1 - 5, 2023.

[3] Scott Fujimoto, Herke Hoof, and David Meger. "Addressing function approximation error in actor-critic methods." In *International Conference on Machine Learning*, pages 1587 – 1596, 2018

[4] Yujie Yang, Yuxuan Jiang, Yichen Liu, Jianyu Chen, and Shengbo Eben Li. "Model-free safe reinforcement learning through neural barrier certificate." *IEEE Robotics and Automation Letters*, pages 1295 - 1302, 2023.