Safety-Prioritizing Curricula for Constrained Reinforcement Learning

Cevahir Koprulu[1], Thiago D. Simão[2], Nils Jansen[3], Ufuk Topcu[1]

[1] TEXAS The University of Texas at Austin [2] TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY [3] RUHR UNIVERSITÄT BOCHUM RUB

Problem Setting

Objective of curriculum learning

Automatically generate a sequence of tasks/contexts to accelerate learning.

Gap

Existing CL approaches overlook constraints!

Contextual Constrained MDP

$$\underbrace{\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{X}}_{\text{Context Space}}, \underbrace{\mathrm{M}}_{\text{From contexts to Constrained MDPs}}, D, \gamma \rangle$$

Context Space

$\mathbf{x} \in \mathcal{X}$

From contexts to Constrained MDPs

$$M(x) = \langle \mathcal{S}, \mathcal{A}, p_{\mathbf{x}}, p_{0,\mathbf{x}}, r_{\mathbf{x}}, c_{\mathbf{x}} \rangle$$

Optimal Policy

Given: Target context distribution φ

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

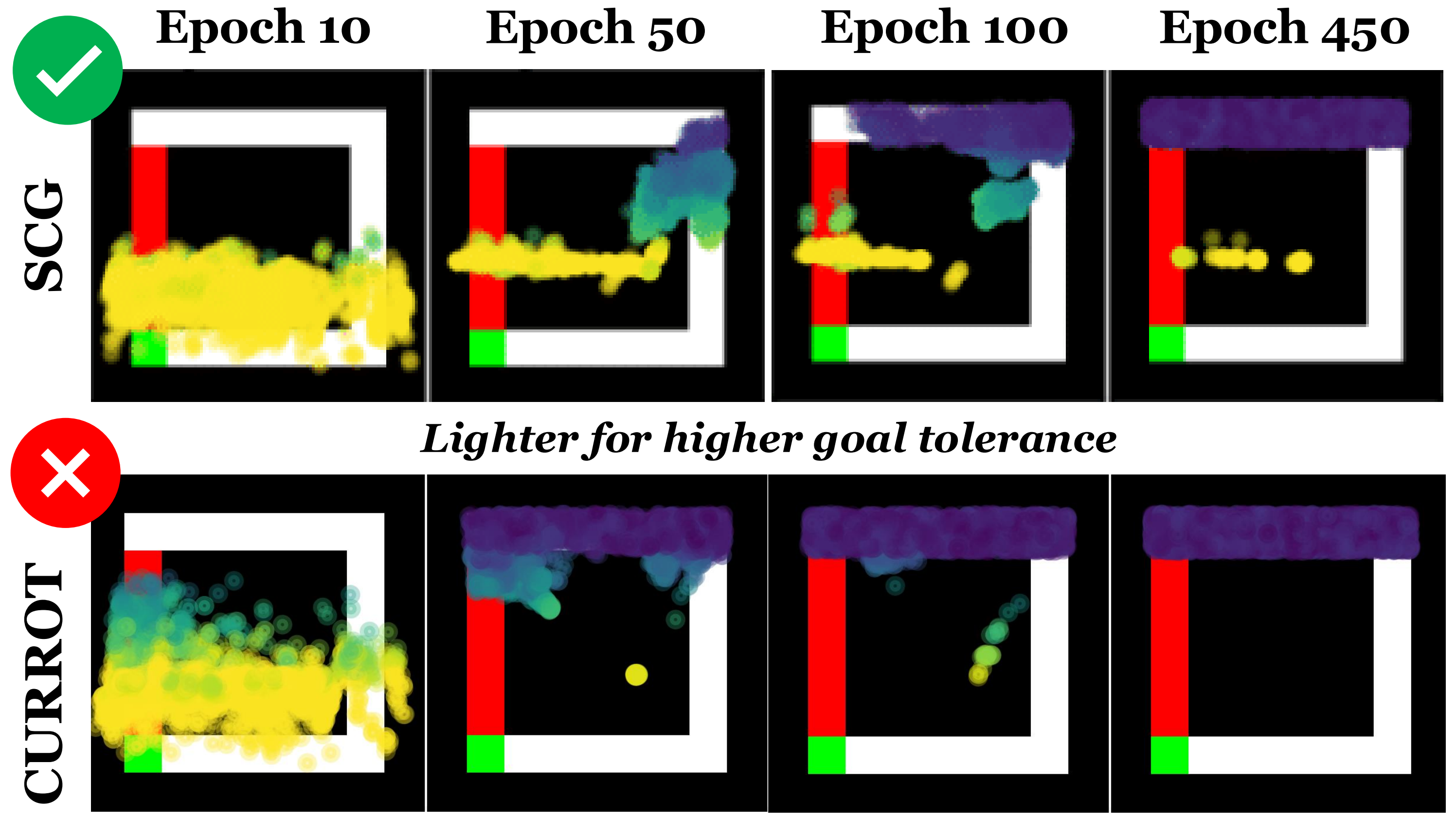
Constraint Violation Regret

$$\textstyle\sum_{l\in[L]} \max\{\underbrace{\mathbb{E}_{\mathbf{x}\sim\rho_l}\left[V_c^{\pi_l}(\mathbf{x})\right]}_{\text{Expected cost of policy } \pi_l \text{ in context distribution } \rho_l \text{ at iteration } l} - D, 0\}$$

Expected cost of policy π_l in context distribution ρ_l at iteration l

Safe vs Unsafe Curricula

Naïve attempts lead to constraint violations early on during training!

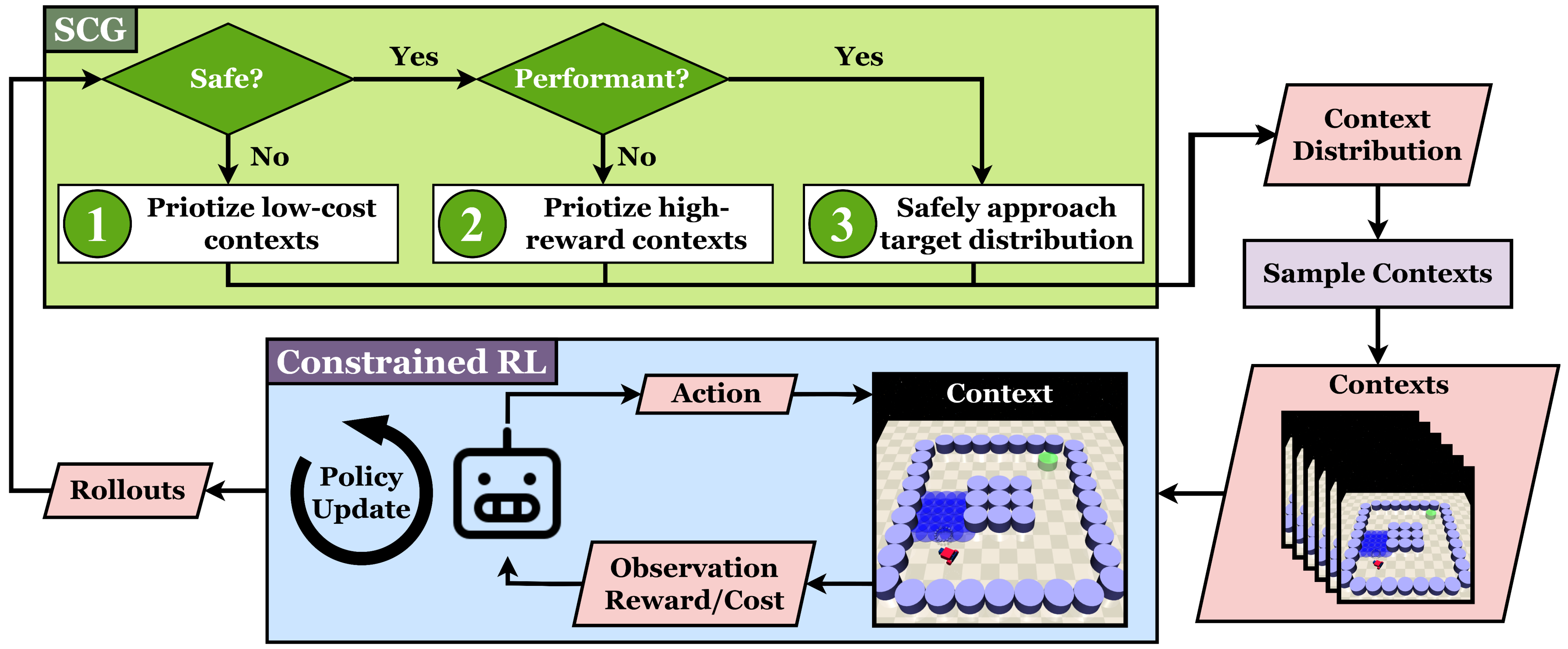


Lighter for higher goal tolerance

Klink, P., Yang, H., D'Eramo, C., Peters, J., & Pajarinen, J. (2022). Curriculum reinforcement learning via constrained optimal transport. In ICML.

Learn safer with curriculum learning!

Safe Curriculum Generation: Prioritize safe tasks



$$\rho_{next} = \mathrm{argmin}_{\rho} \mathcal{W}_2(\rho, \varphi)$$

$$\text{s.t.} \quad \rho(x) > 0 \Rightarrow V_r^{\pi_k} > \zeta, \forall x \in \mathcal{X}$$
$$\rho(x) > 0 \Rightarrow V_c^{\pi_k} < \widetilde{D}, \forall x \in \mathcal{X}$$
$$\mathcal{W}_2(\rho, \rho_+) \leq \epsilon$$

Source distribution ρ_+ determined by each phase

Objective
Approach target φ

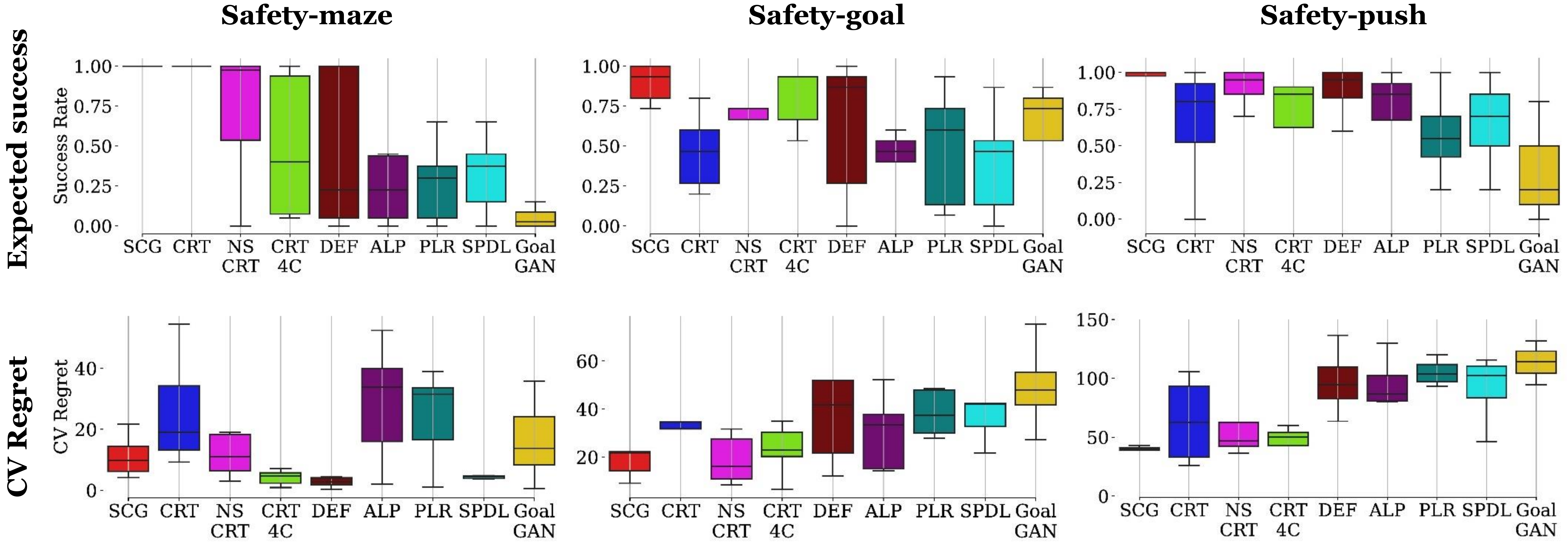
Constraints
Performance

Safety

Pace of learning

Results: Highest success rates with low constraint violation regret

In safety-maze, SCG achieves lower CV regret than CURROT, the only other method that yields 100% success. In safety-push and goal, SCG achieves the highest success rates, simultaneously reducing constraint violations.



CENTER FOR autonomy

Office of Naval Research Science & Technology

DEUCE. erc Data-Driven Verification and Learning under Uncertainty.