Safe Reinforcement Learning via Curriculum Induction

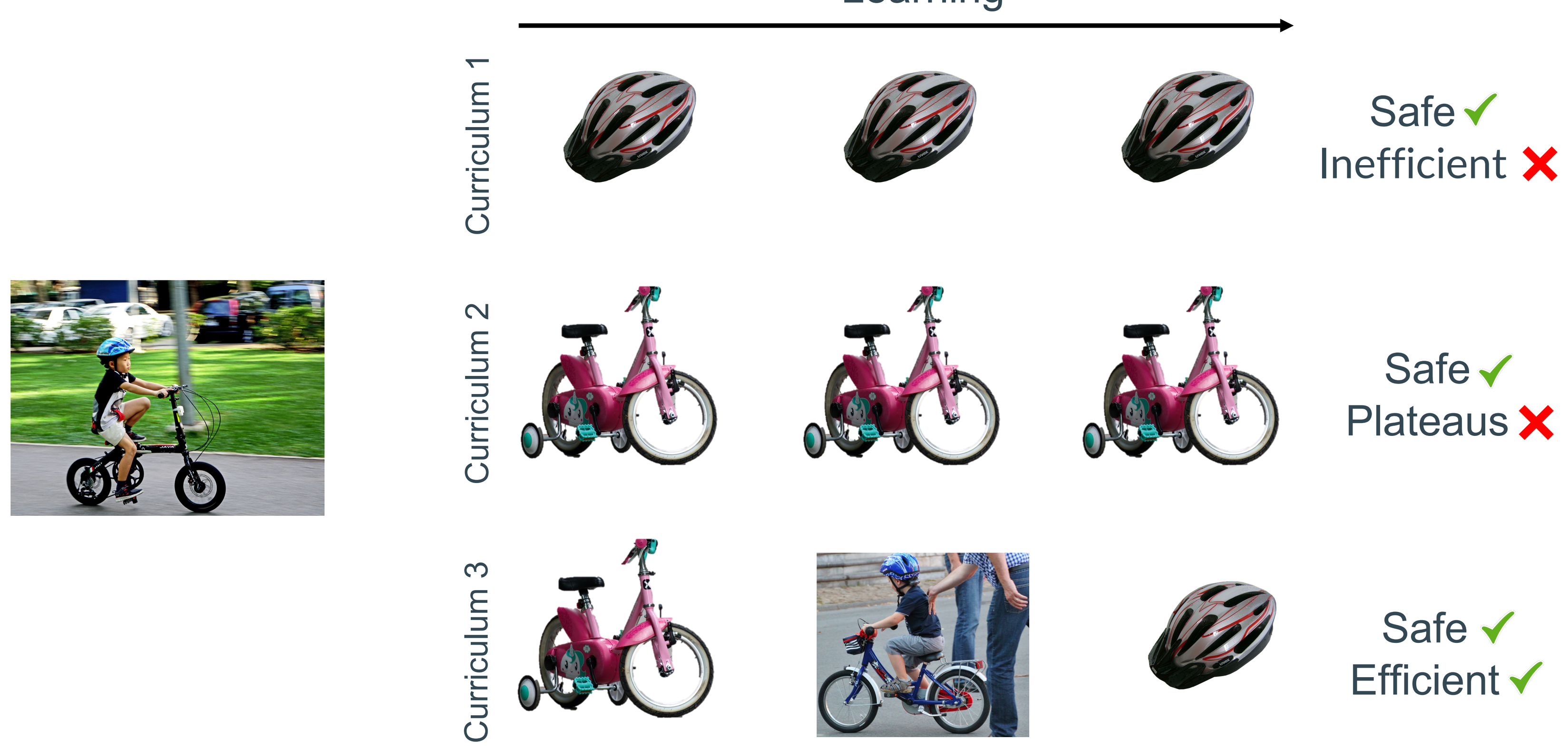
Matteo Turchetta, Andrey Kolobov, Shital Shah, Andreas Krause, Alekh Agarwal

ETH zürich
Microsoft Research

NEURAL INFORMATION PROCESSING SYSTEMS

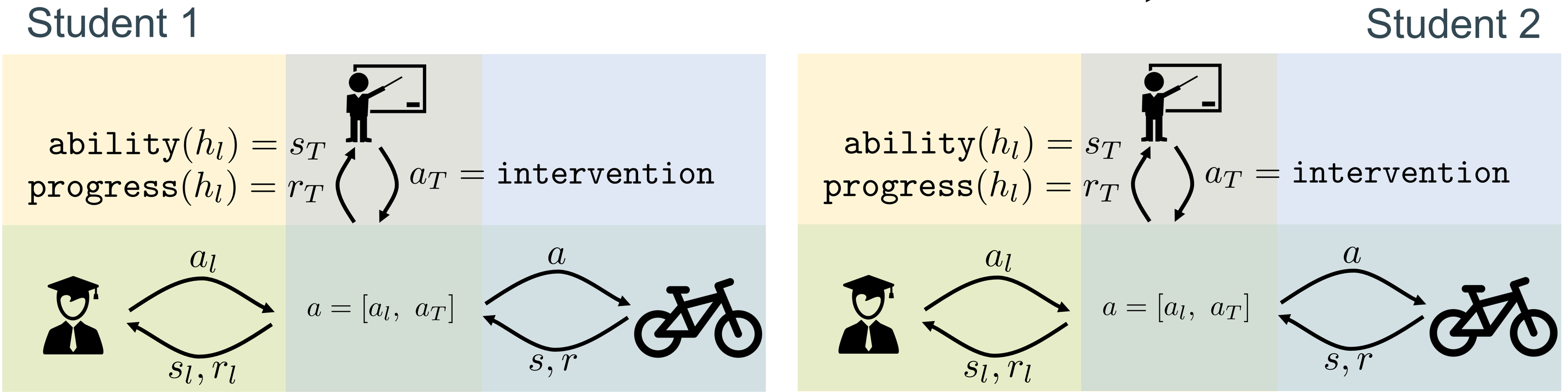
Intuition

In safety-critical environment, humans learn efficiently from well-structured, safety-centered curricula.



Learning Teaching Policies

Q: How can we automatically design a curriculum for safe and efficient learning?
A: By optimizing a teaching policy educating a sequence of students.

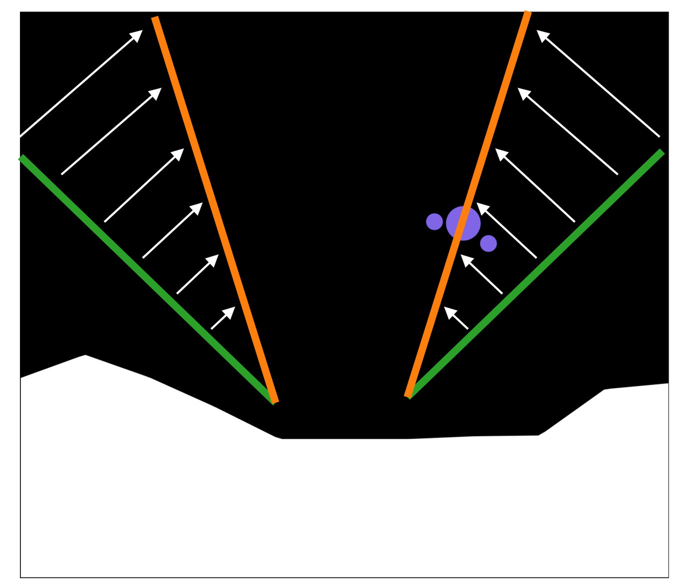


Related Work and Contributions

	No baseline policy	Online Training	Non-smooth environment	Safe Training	Non-smooth dynamics	Weak teacher	Optimized teaching policy
CISR (ours)	✓	✓	✓	✓	✓	✓	✓
Le et al. '19		✗					
Wachi et al'20			✗				
Berkenkamp et al '17					✗		
Achiam et al. '17				✗			
Chow et al. '18	✗						
LfD						✗	
Inverse RL						✗	
Curriculum learning (most)							✗

Interventions

Original problem: CMDP constrained on the number of catastrophic events.
Introducing the teacher: Whenever the student approaches danger (determined by a set of **trigger states**), the teacher rescues it and resets it to a neighboring state (determined by a **conditional reset distribution**). To prevent the student from exploiting the teacher, we constrain the number of times the student can get its help. The strictness of the teacher is controlled by its **tolerance**.



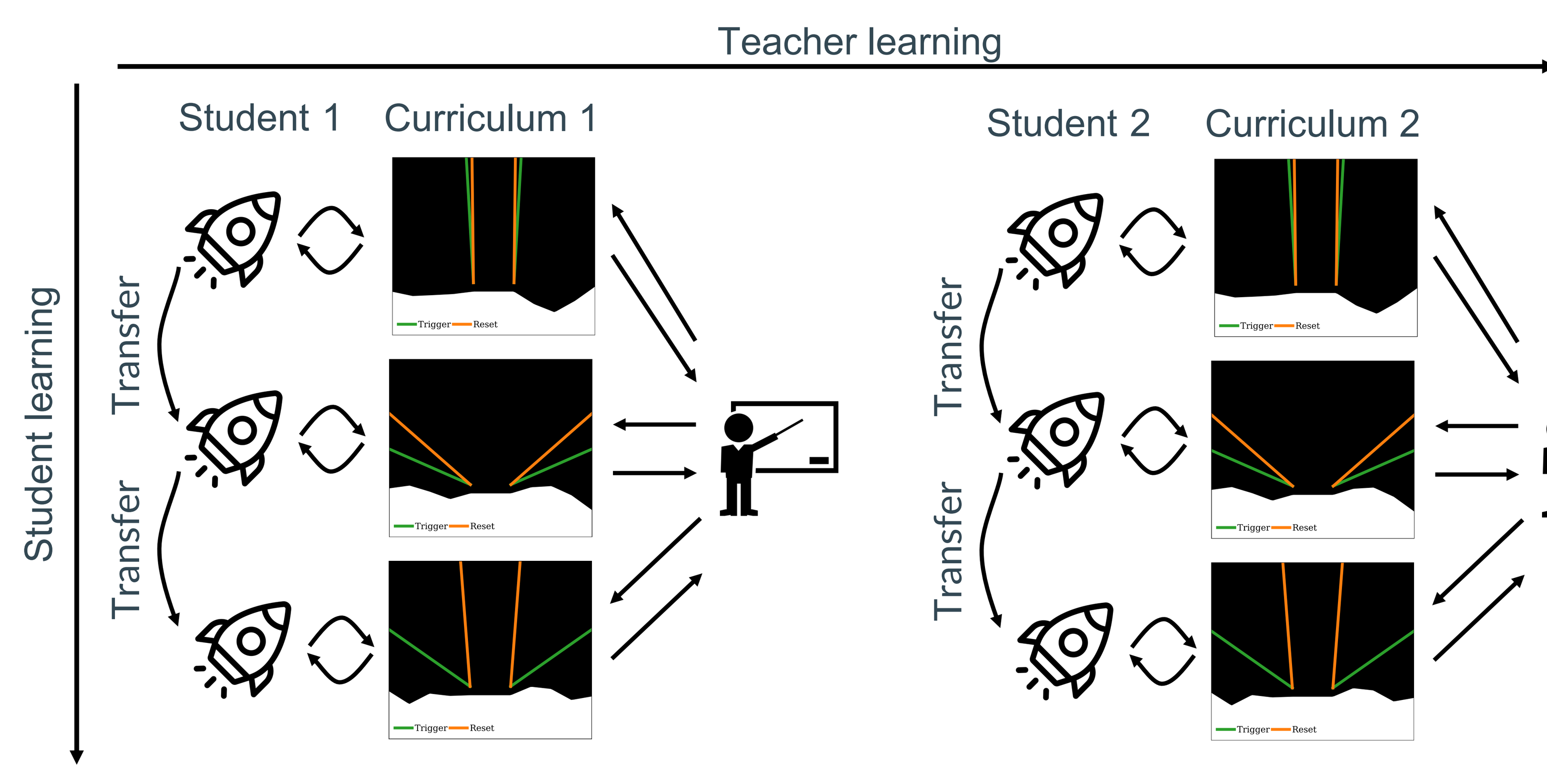
$$\pi^* = \operatorname{argmax} \ \mathbb{E}_{\rho_i^\pi} \sum_{t=0}^{T} r_i(s_t, a_t, s'_{t+1}),$$

$$\text{s.t.} \ \mathbb{E}_{\rho_i^\pi} \sum_{t=0}^{T} \mathbb{I}(\ ) \leq \kappa_i,$$

$$\mathbb{E}_{\rho_i^\pi} \sum_{t=0}^{T} \mathbb{I}(\ ) \leq \tau_i$$

Interaction

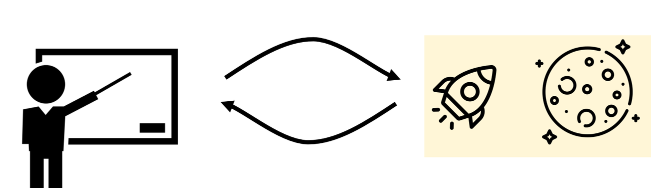
Two learning agents on different time scales: the students and the teachers. The students learn across a sequence (curriculum) of CMDPs proposed by the teacher. Thus, they are CMDP solvers with knowledge transfer mechanism. The teacher tries different teaching policies across students to identify an optimal one.



Teacher's POMDP

POMDP

$a_T = \texttt{intervention}$



$\texttt{ability}(h_l) = s_T$

$\texttt{progress}(h_l) = r_T$

Observation mapping, $\Pi \to \mathbb{R}^2$

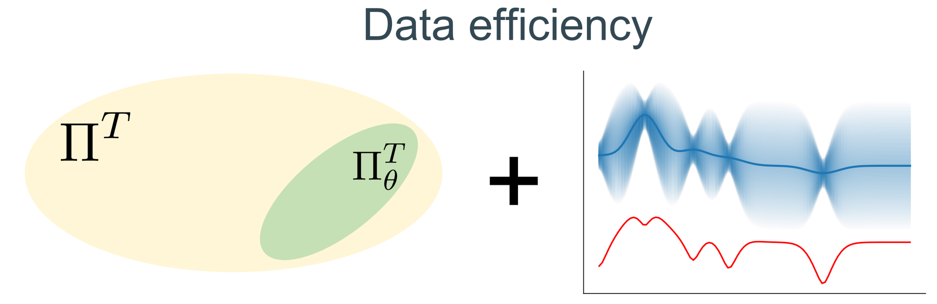
$$s_T^n = \mathbb{E}_{\rho_{i_n}^\pi}\left[\left(\sum_{t=0}^{T} r_{i_n}(s_t, a_t, s'_{t+1}), \ \sum_{t=0}^{T} \mathbb{I}(s_t \in \mathcal{D}_{i_n}) - \tau_{i_n}\right)\right]$$

performance safety

Reward

$r_T^n = \hat{V}(\pi_n) - \hat{V}(\pi_{n-1})$ Progress on proxy task

Data efficiency



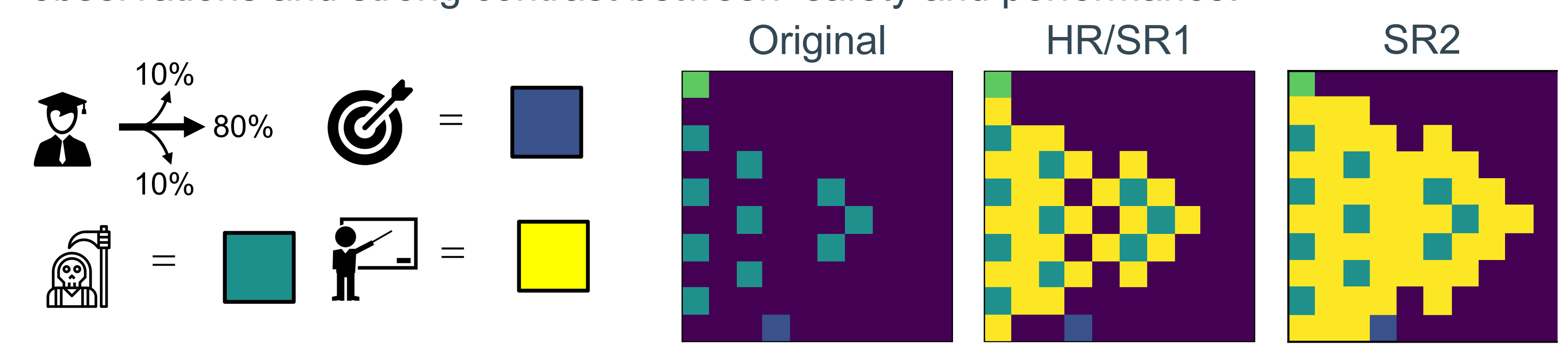
Bayesian optimization over parametrized policies

Theoretical results

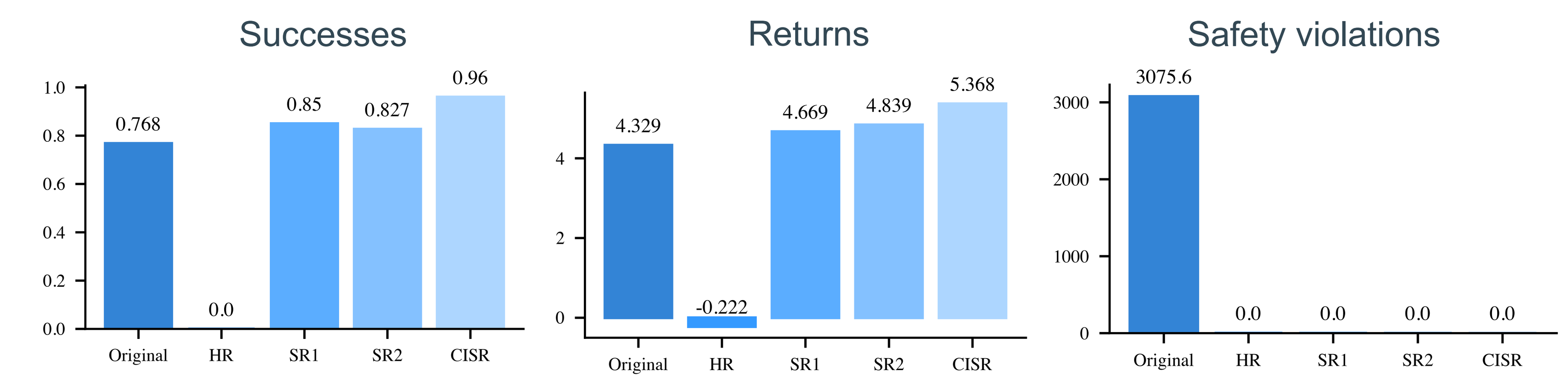
"Proposition"
If the teacher is sufficiently strict ($\tau_i + \kappa_i \leq \kappa$) and the trigger states of the teacher's intervention "blankets" the set of unsafe states ($\mathcal{D} \subseteq \mathcal{D}_i$), then, the interventions guarantee **safe learning** and **safe deployment**.

Frozen Lake Experiments

Challenging variant of the frozen lake environment with high dimensional observations and strong contrast between safety and performance.

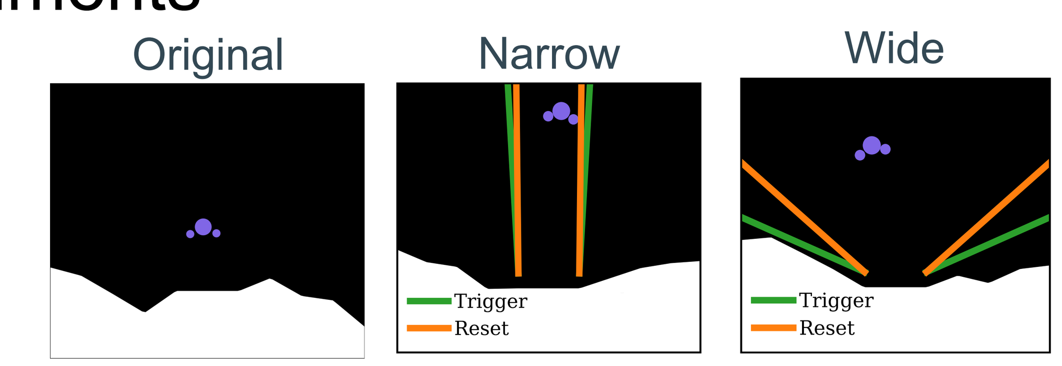


HR (reset to starting state): too hard to find the goal.
SR (reset to previous state): finds the goal but too different from original.

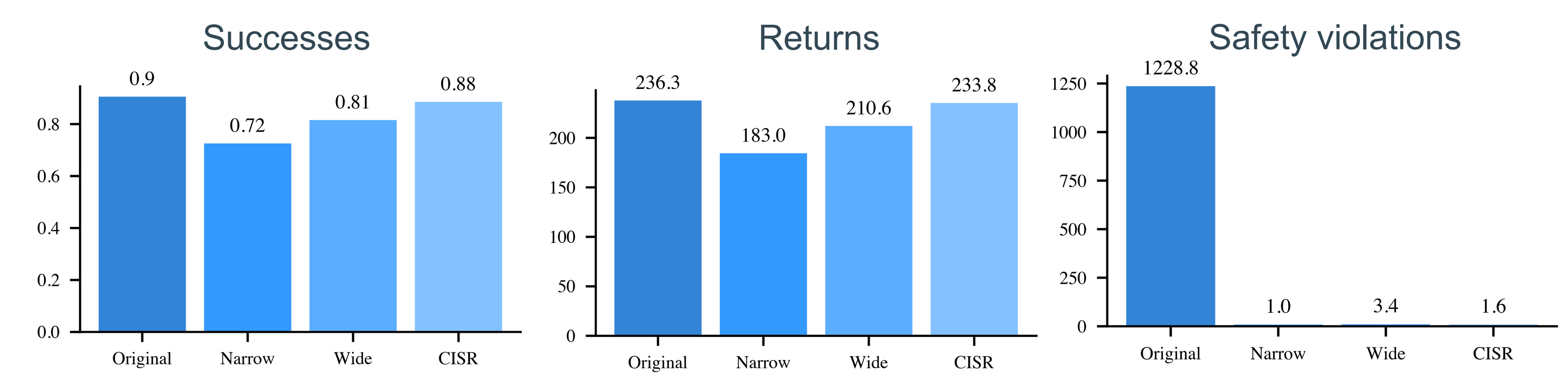


Lunar lander experiments

- Continuous environment.
- Cannot sense distance from the ground.
- Landing surface changes each episode.



Narrow: easy but plateaus.
Wide: hard to experience normal landing, therefore it is slow.



Conclusions

- CISR can learn teaching policies for **faster** and **safe** training of students.
- CISR requires **fewer assumptions** than most method in the literature.
- CISR was effectively applied in two challenging safety-critical environments.

Paper
SCAN ME

Code
SCAN ME