Curriculum Induction for Safe Reinforcement Learning

Marvin Sextro[1] and Jonas Loos[1]

[1]Technical University of Berlin

1. Introduction

1.1 Key Ideas[2]

- A teacher trains a student to solve a task
- The teacher keeps the student safe during training
- For this, the teacher is given a set of pre-defined interventions and learns to apply them optimally
  $\rightarrow$ *curriculum policy*
- Interventions are pairs of trigger states and transitions guiding the student back into a safe state

1.2 Our Approach

- We compare the students trained by the Optimized curriculum policy from the paper [2] to students trained with our own curriculum policies

2. Background

2.1 Constrained Markov Decision Process[2]

- The student is a RL agent trained in a CMDP:

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \mathcal{D} \rangle$$

- $\mathcal{S}, \mathcal{A}$: State and action space
- $\mathcal{P}(s'|s,a)$: Transition kernel
- $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$: Reward function
- $\mathcal{D}$: Set of unsafe terminal states

2.2 Curriculum Induction for Safe RL[2]

- In CISR, the teacher gets a set $\mathcal{I}$ of interventions $\{\langle \mathcal{D}_i, \mathcal{T}_i \rangle\}_{i=1}^{K}$ as input, which consist of *trigger states* $\mathcal{D}_i \subset \mathcal{S}$ and reset distributions $\mathcal{T}_i : \mathcal{S} \rightarrow \Delta_{\mathcal{S}\setminus\mathcal{D}_i}$
- **Curriculum**: Sequence of CMDPs $\mathcal{M}_{i_1}, ..., \mathcal{M}_{i_{N_s}}$, where during the n^{th} curriculum step, the student interacts with the CMDP $\mathcal{M}_{i_n}$ induced by an intervention $i_n \in \mathcal{I}$
- **Curriculum Policy**: A curriculum policy $\pi^T : \mathcal{H} \rightarrow \mathcal{I}$ maps the teacher's observation history of statistics $\phi(\pi_1), ..., \phi(\pi_{n-1}) \in \mathcal{H}$ about the student's policy to an intervention at the start of the n^{th} curriculum step
- For curriculum policies independent of the student's policy (e.g. SR, HR, Back or Incremental), this can be simplified to a mapping $\pi^T : [N_s] \rightarrow \mathcal{I}$

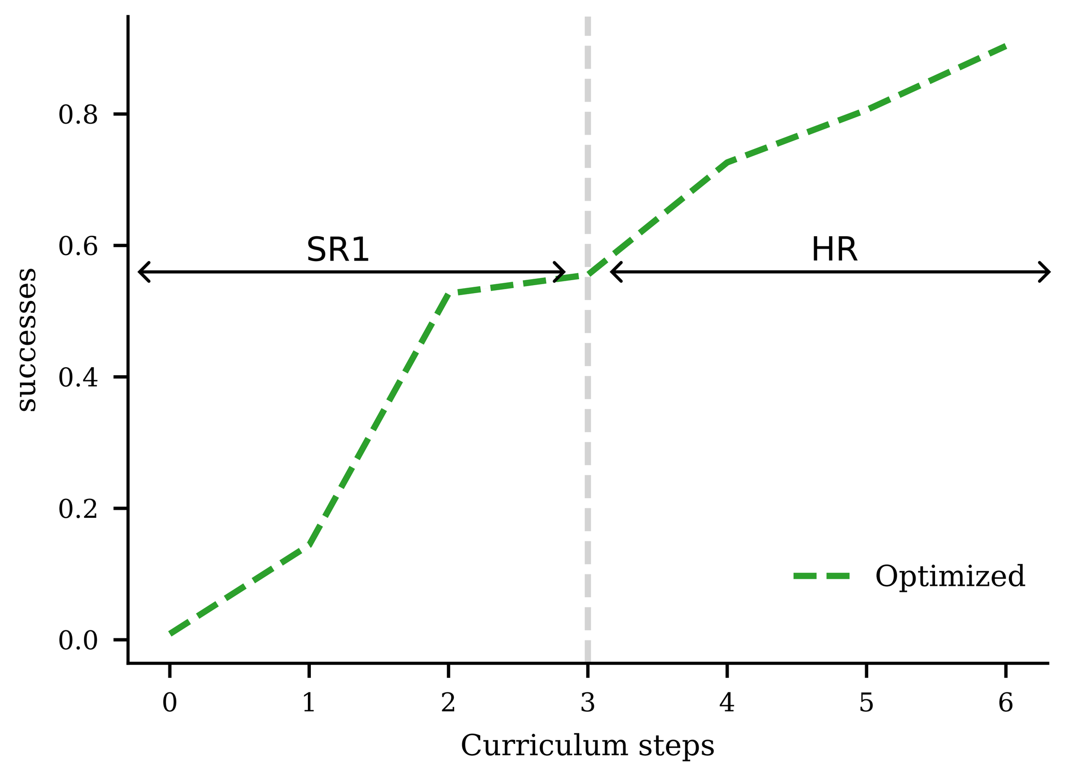

Figure 1: The Optimized curriculum policy switching interventions from Soft Reset 1 (SR1 moves the agent one step back) to Hard Reset (HR resets the agent back to the start).

3. Experiments

3.1 Curriculum Policies

3.1.1 Back

- The Back$_x$ curriculum policy always resets the agent by a constant number of x steps (we tested $x \in [1, 9]$)

3.1.2 Incremental

- The Incremental curriculum policy gradually changes from exploration to exploitation
- We define Incremental$_x$ to reset the agent by $\lceil \frac{1}{2^x} \cdot n \rceil$ steps during the n^{th} curriculum step
- The parameter x can be adjusted for environments of different size or complexity (we tested $x \in [0, 4]$)

3.2 Environments

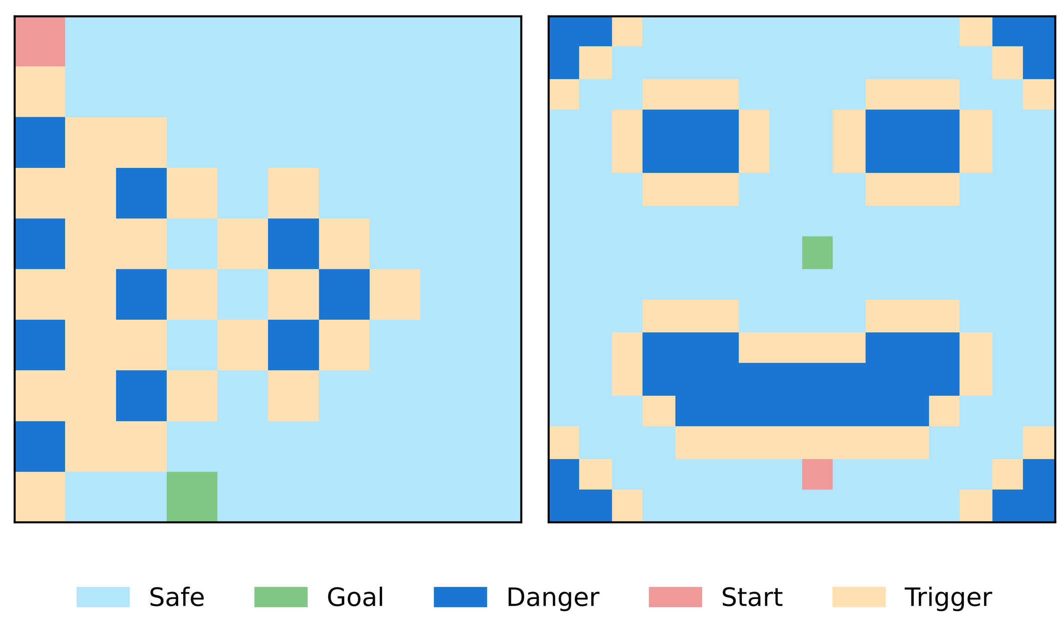

Figure 2: The Frozen Lake environment used in the paper [2] on the left (size 10x10) and our Frozen Smiley environment on the right (size 16x16). Interventions are triggered at distance = 1 from holes.

4. Results

- For all policies with teacher interventions the agent was kept safe during training
- Both the Back and the Incremental curriculum policy perform better than the Optimized one
- For Back, with increasing environment size and longer paths, it is beneficial to increase reset steps
- For Incremental, increasing the reset steps more slowly to allow for longer exploration is advantageous in larger environments

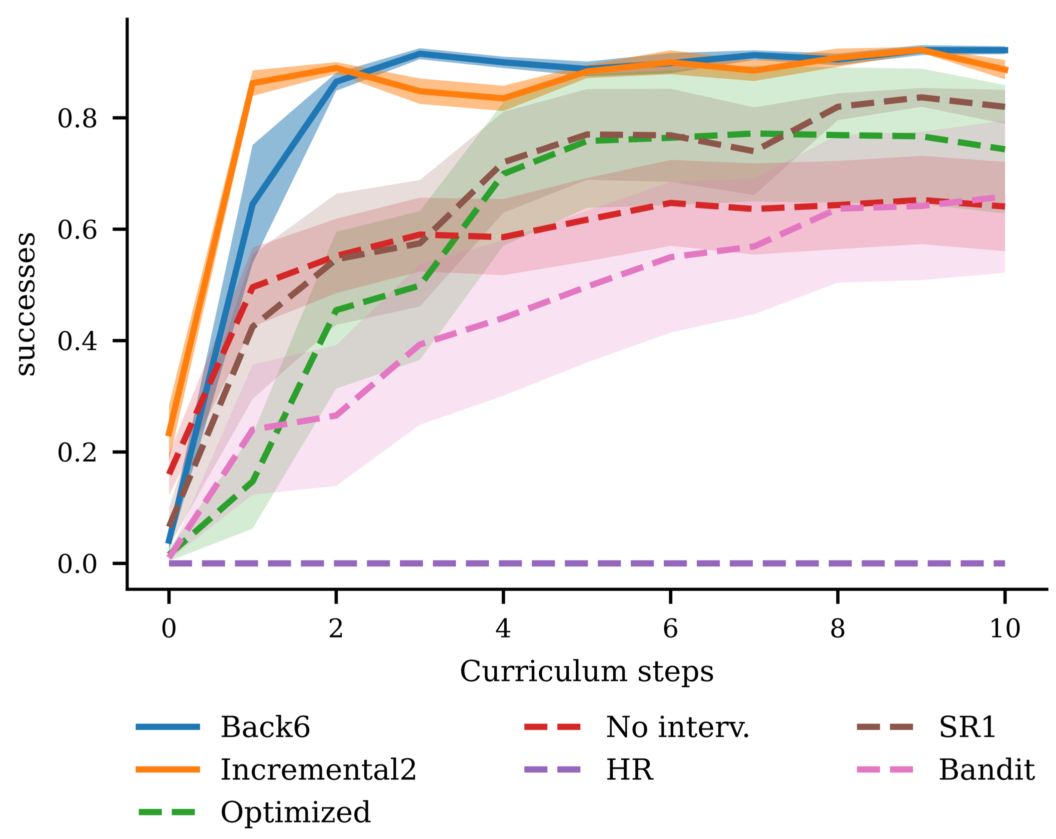

Figure 3: Success rates of different curriculum policies on the Frozen Lake environment. For our policies, the best found parameters x are used.

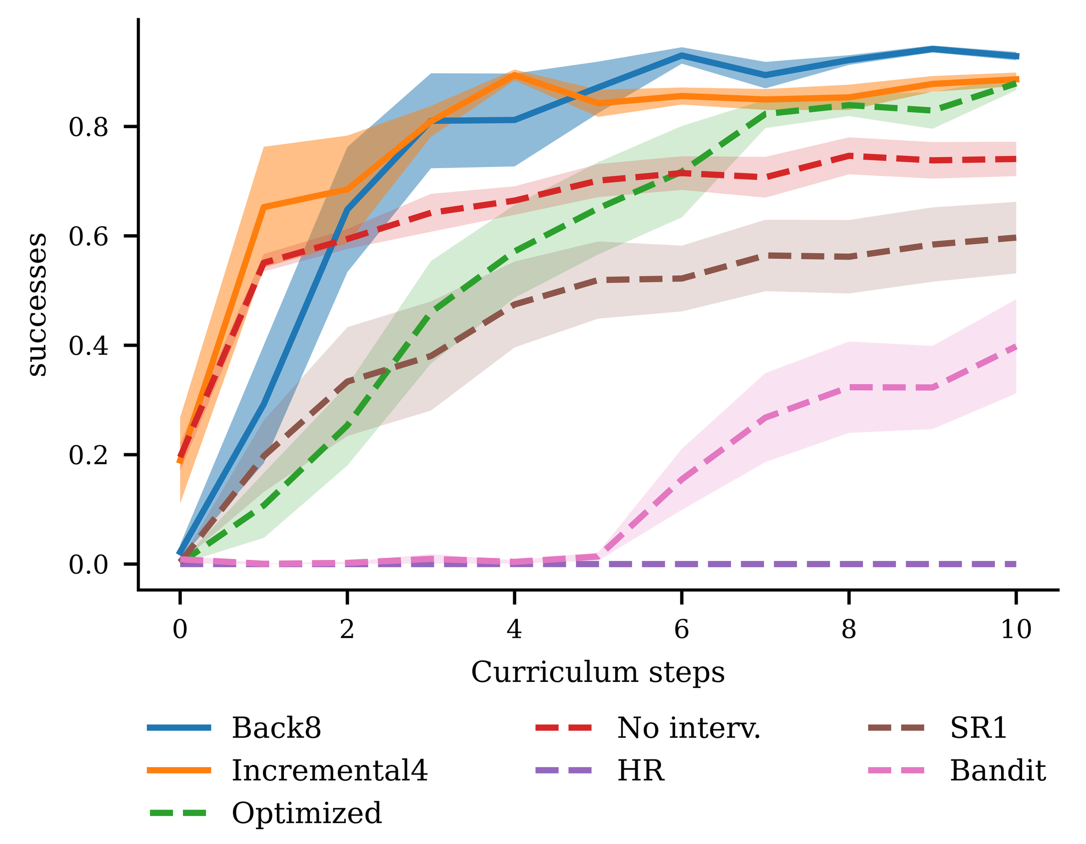

Figure 4: Success rates of different curriculum policies on the Frozen Smiley environment. For our policies, the best found parameters x are used.

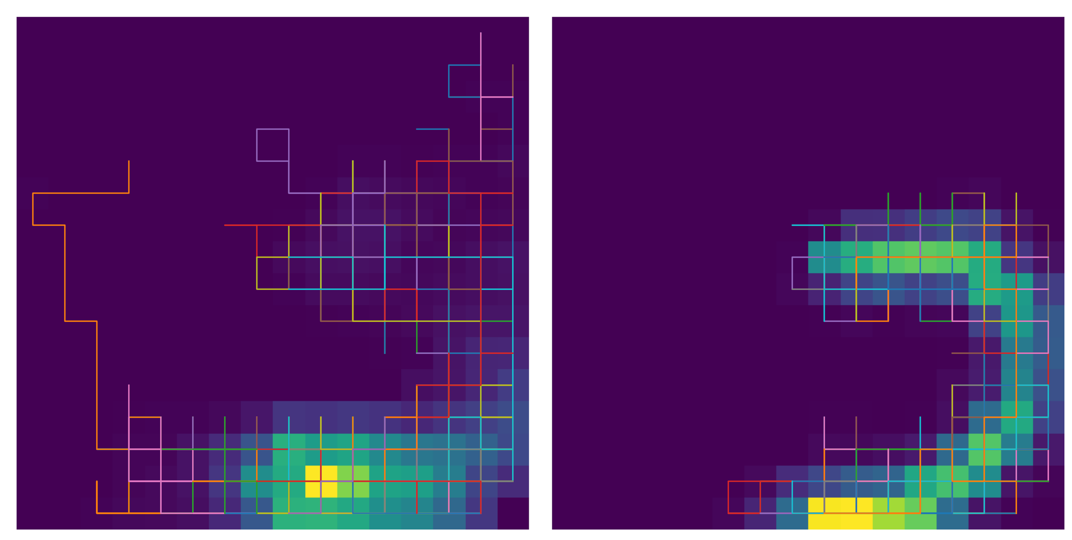

Figure 5: Exemplary trajectories for the Frozen Smiley environment with the Optimized policy. The lines represent the steps taken, while the background shows a heatmap of the student's positions. The trajectories show a progression from the first curriculum step (left) to a later step (right).

5. Conclusions

- For the Frozen environments, our curriculum policies outperform the Optimized one
- Larger environments require a longer exploration phase and more reset steps
- The original HR, SR and Bandit policies do not generalize well to larger environments
- Defining reset transitions which keep the student safe is easier than defining suitable trigger states
- This could become a problem when the state space is complex, dynamic or just partly observable

6. Outlook

- Apply the method to OpenAI's Safety Gym
- Increase the amount of available interventions for the Optimized curriculum policy
- Evaluate how well different curriculum policies generalize to dynamic or random environments

References

[1] OpenAI. Frozen Lake. URL: https://gymlibrary.ml/environments/toy_text/frozen_lake.

[2] Matteo Turchetta, Andrey Kolobov, S. Shah, Andreas Krause, and Alekh Agarwal. Safe Reinforcement Learning via Curriculum Induction. *ArXiv*, abs/2006.12136, 2020.