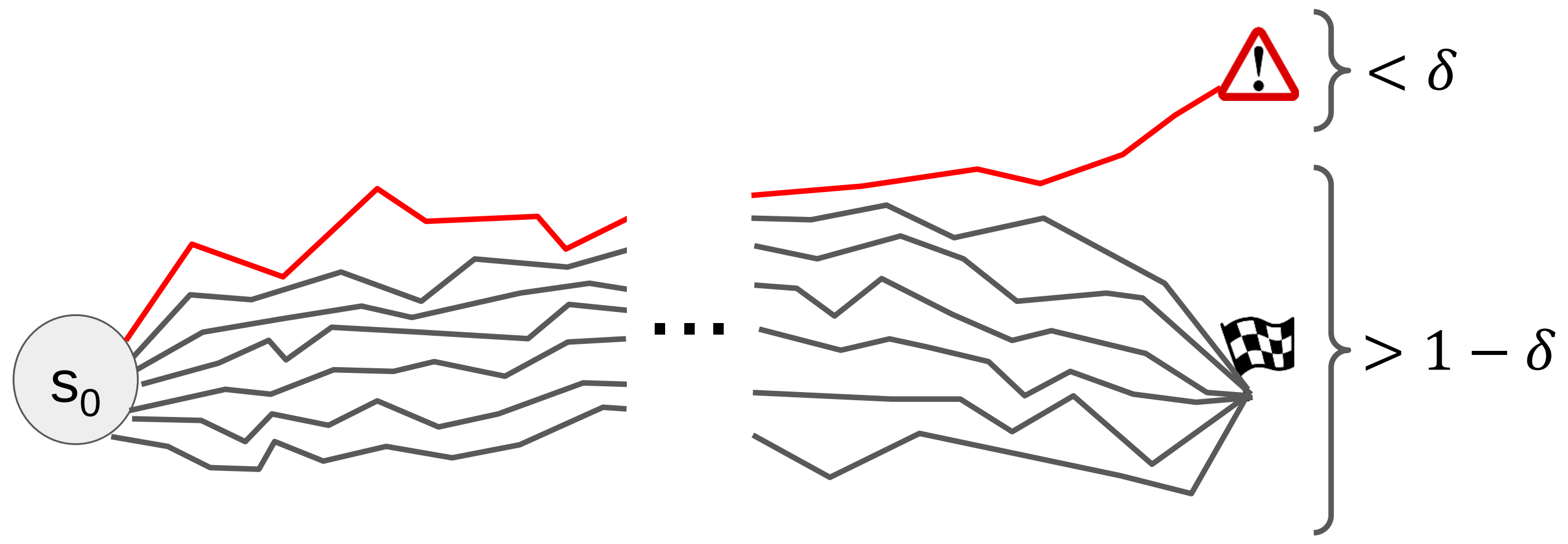
Provably Safe PAC-MDP Exploration Using Analogies

Melrose Roderick, Vaishnavh Nagarajan, J. Zico Kolter
Carnegie Mellon University

Applying Reinforcement Learning (RL) in many real-world problems requires strict safety guarantees.

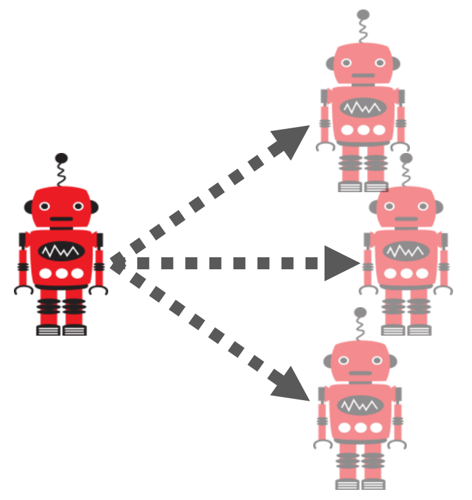
In safety-critical domains a single mistake can cause significant harm.

We must ensure the agent never reaches an unsafe state during the entire training trajectory.

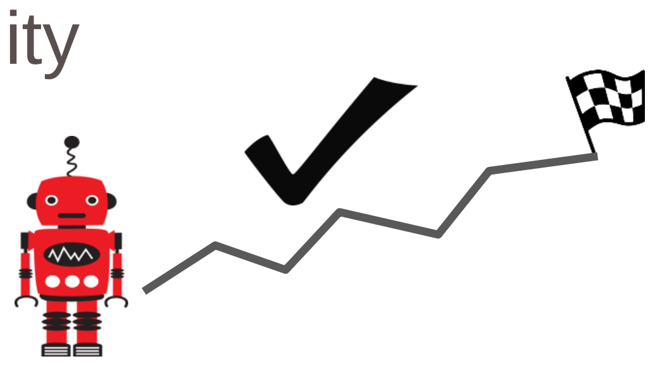


Our safe RL method is uniquely able to address 3 goals simultaneously:

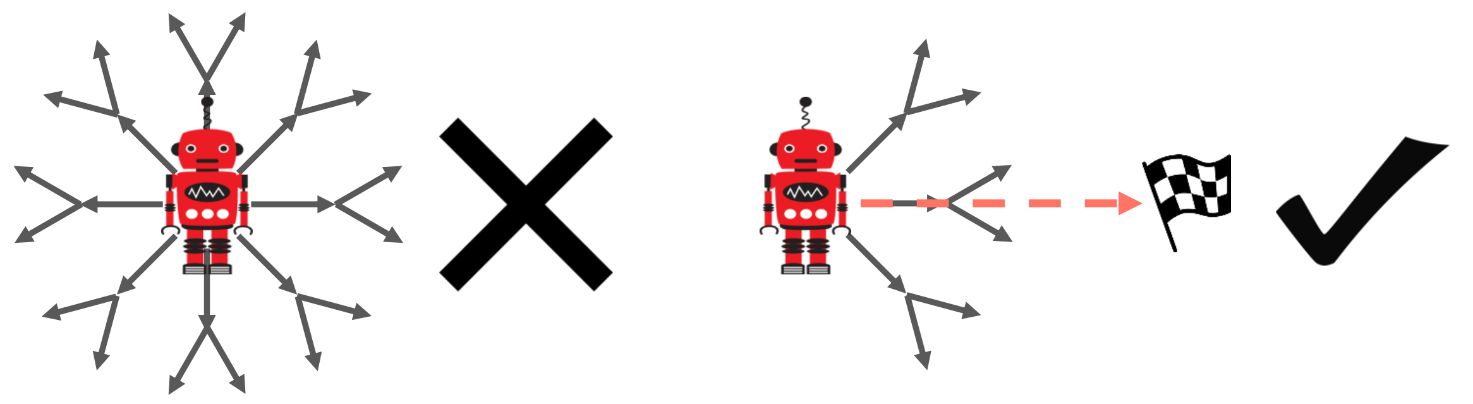
1. Unknown stochastic dynamics



2. PAC-MDP optimality

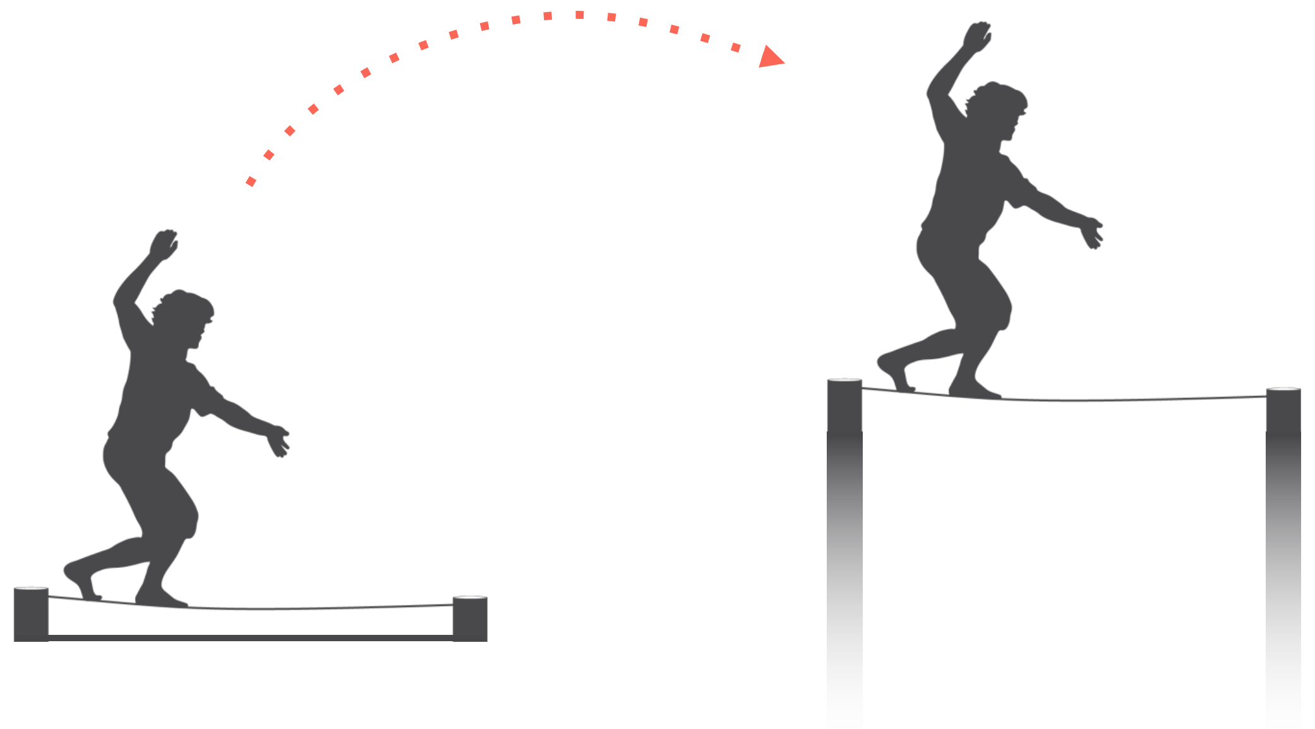


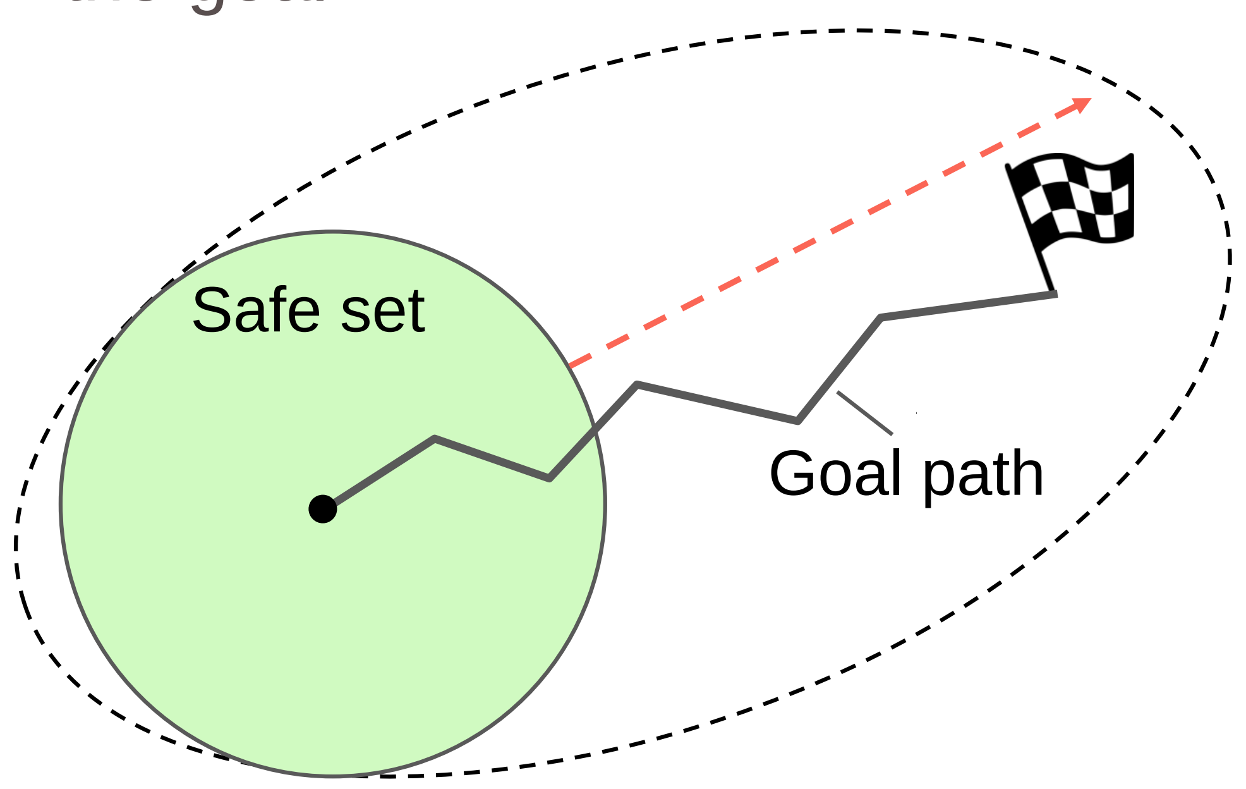
3. Guided exploration



Analogous Safe-state Exploration (ASE) explores safe state-actions to determine the safety of analogous state-actions.

1. ASE safely learns the dynamics of analogous states.
2. Then expands the safe set toward the goal.





High level algorithm:

- Compute a optimistic policy, $\overline{\pi}$, using MBIE [1].
- Do targeted exploration in the safe set to learn the safety of state-actions along .
- If some state-actions turn out to be unsafe, recompute $\overline{\pi}$ and repeat above steps.
- Once all state-actions along $\overline{\pi}$ are in the safe set, execute the optimistic policy

Algorithm 1 Analogous Safe-state Exploration

```
Initialize: $\hat{Z}_{\text{safe}} \leftarrow Z_0$
Compute confidence intervals, then $\overline{\pi}_{\text{goal}}, \overline{Z}_{\text{goal}}$, and $Z_{\text{explore}}$.
Compute $\overline{\pi}_{\text{explore}}, \overline{\pi}_{\text{switch}}$ using value iteration.
for $t = 1, 2, 3, \dots$ do
    $a_t \leftarrow \begin{cases} \overline{\pi}_{\text{goal}}(s_t) & \textbf{if } \overline{Z}_{\text{goal}} \subset \hat{Z}_{\text{safe}} \\ \overline{\pi}_{\text{explore}}(s_t) & \textbf{otherwise} \end{cases}$
    Take action $a_t$ and observe next state $s_{t+1}$.
    if $n(s_t, a_t) < m$ then
        Recompute confidence intervals and expand $\hat{Z}_{\text{safe}}$, then recompute policies.
```

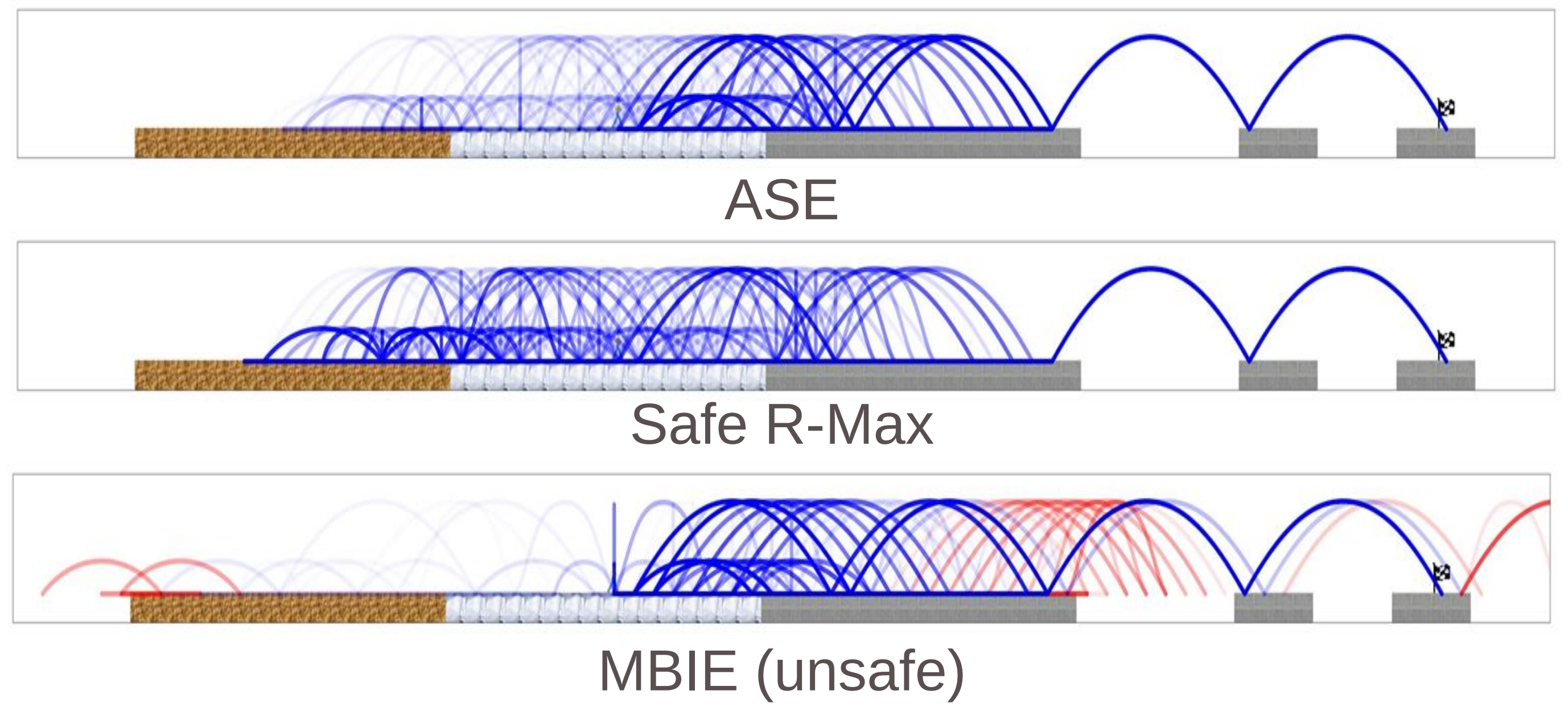
We prove ASE never (whp) reaches an unsafe state during training and finds an optimal policy (PAC-MDP).

Theorem. *For any $\epsilon, \delta \in (0, 1]$ and with probability at least $1 - \delta$, the agent never takes an unsafe state-action and makes a finite number of ϵ-sub-optimal steps bounded by:*

$$O(Hm|S|A|(1/\epsilon(1 - \gamma))\ln(1/\delta))$$

where $m \approx O((|S|\epsilon^2) + (1/\epsilon^2)\ln(|S||A|/\epsilon)$ and H is the communication time of the MDP.

In our experiments, ASE never enters an unsafe state and guides exploration towards the goal.



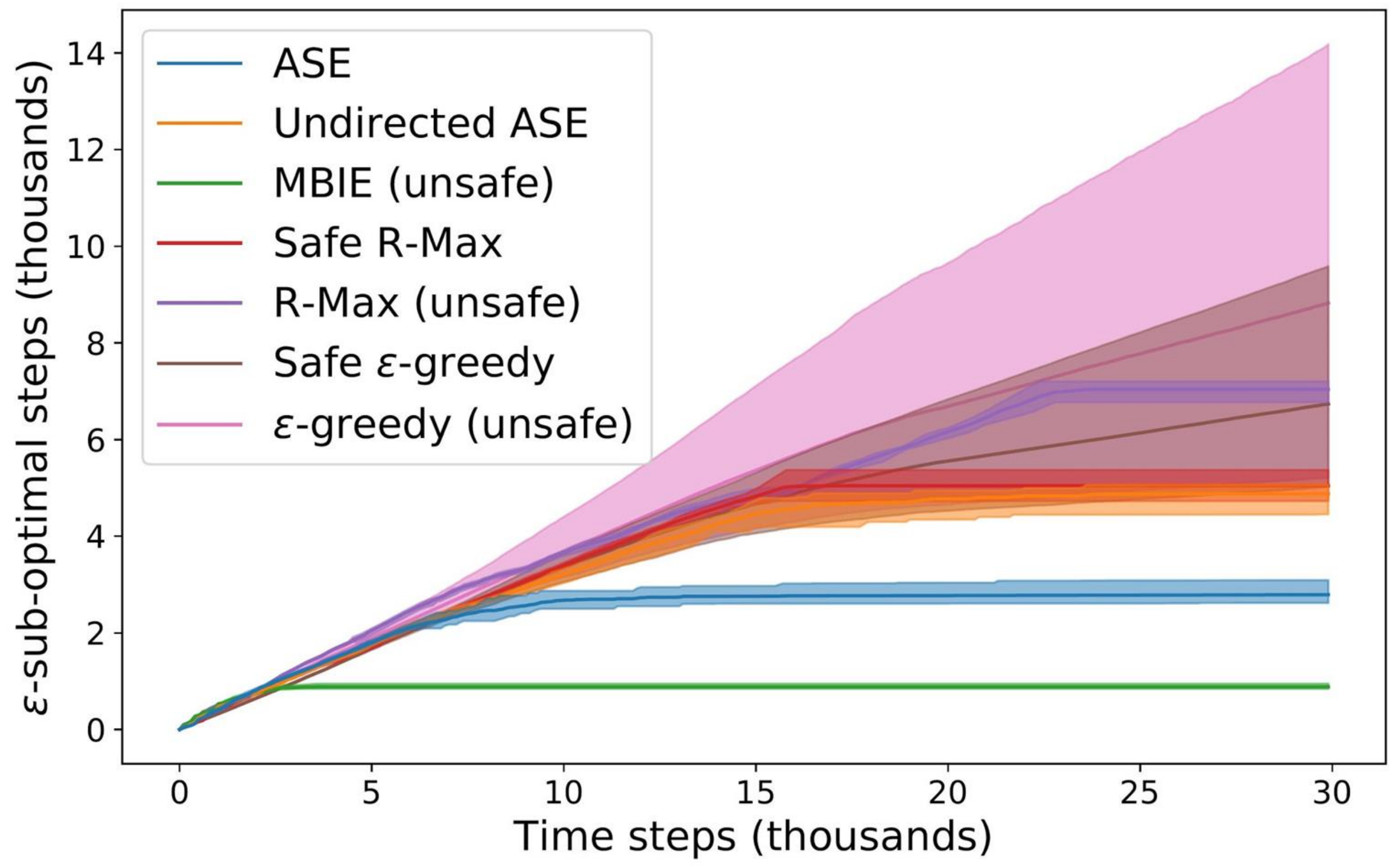
ASE

Safe R-Max

MBIE (unsafe)

All trajectories of different agents on the Discrete Platformer domain. Unsafe trajectories are drawn in red.

Empirically, ASE makes far fewer sub-optimal actions than other safe algorithms.



Number of ϵ-sub-optimal steps taken by each agent throughout training.

References

[1] Alexander L Strehl and Michael L Littman (2008). “An analysis of model-based interval estimation for markov decision processes.” Journal of Computer and System Sciences, 1309–1331.

[2] Matteo Turchetta, Felix Berkenkamp, and Andreas Krause (2016). “Safe exploration in finite markov decision processes with gaussian processes.” In Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016.

[3] Teodor Mihai Moldovan and Pieter Abbeel (2012). “Safe exploration in markov decision processes.” In Proceedings of the 29th International Conference on Machine Learning, ICML 2012.