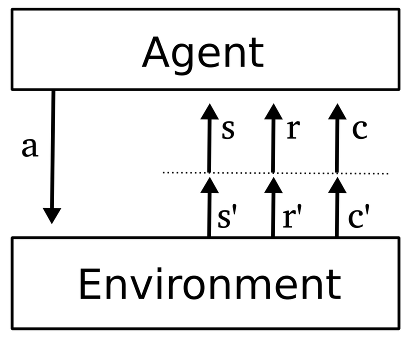
AlwaysSafe

Reinforcement Learning without Safety Constraint Violations during Training

Thiago D. Simão
Nils Jansen
Matthijs Spaan

- Constrained MDPs models safety requirements explicitly.
- How to learn without violating the safety constraints?

Constrained RL



$\mathcal{M} = \langle \mathbb{S}, \mathbb{A}, P, R, \mu, C, \hat{c} \rangle$

$$\max_{\pi} V_R^{\pi}(\mu) = \mathbb{E}_{\pi}\left[\sum_{t=1}^{H} r_t \mid \mu\right]$$

$$\text{s. t. } \underbrace{V_C^{\pi}(\mu) = \mathbb{E}_{\pi}\left[\sum_{t=1}^{H} c_t \mid \mu\right] \le \hat{c}}_{\text{Safety constraint}}$$

Cost-model-irrelevant Abstraction

$$\bar{\mathcal{M}}_{\phi} = \langle \bar{\mathbb{S}}, \mathbb{A}, \bar{P}, \bar{R}, \bar{\mu}, \bar{C}, \hat{c} \rangle$$

$$\phi : \mathbb{S} \to \bar{\mathbb{S}}$$

$$\mathcal{M} = \langle \mathbb{S}, \mathbb{A}, P, R, \mu, C, \hat{c} \rangle$$

ϕ preserves the expected cost:

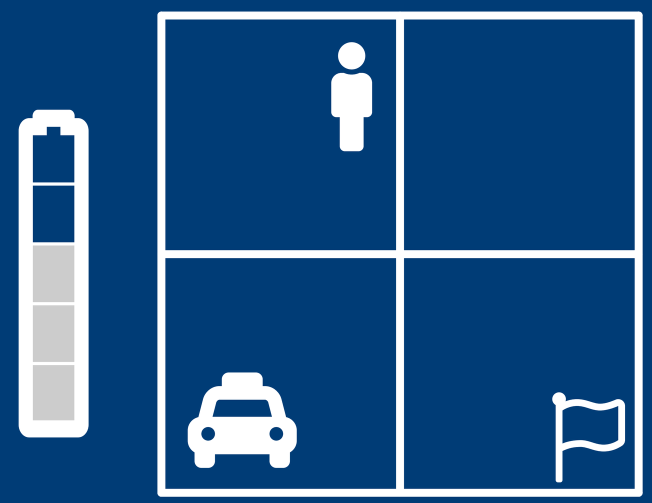
$$V_{\bar{C}}^{\pi, \bar{\mathcal{M}}_{\phi}}(\bar{\mu}) = V_C^{\pi, \mathcal{M}}(\mu)$$

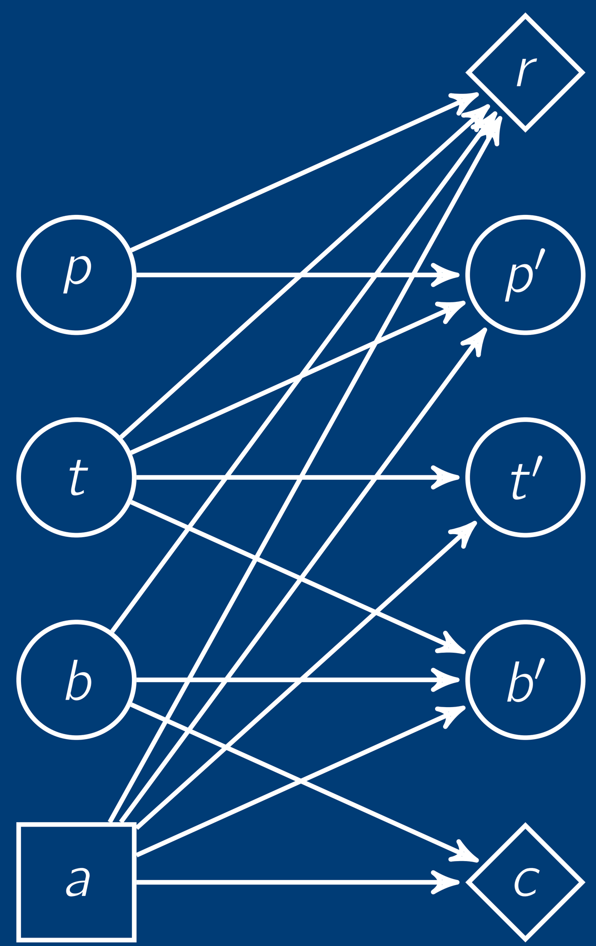
- The abstract policy π_A is safe but might be suboptimal.
- The ground policy π_G can reach optimality but has no safety guarantees.

TUDelft Radboud University

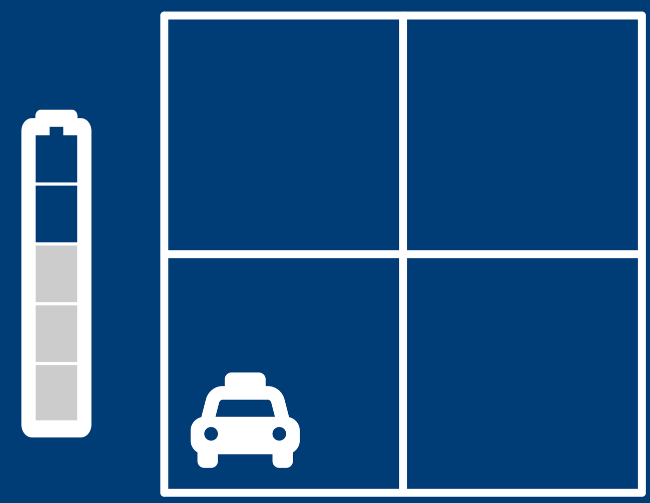
Not everything is relevant for safety

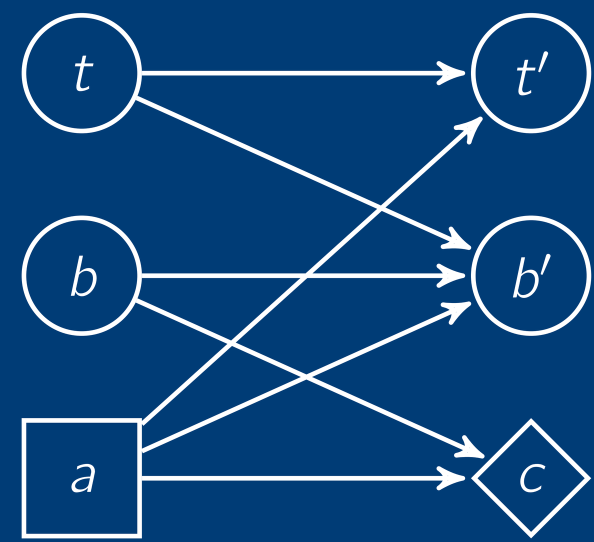
To prevent a taxi from running out of fuel it is not necessary to know the position of the passenger.





Factored MDP
with cost function related to safety

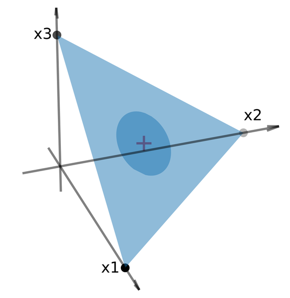




Abstraction of the safety dynamics

Find more at:
https://tdsimao.github.io/publications/Simao2021alwayssafe/

Uncertainty set

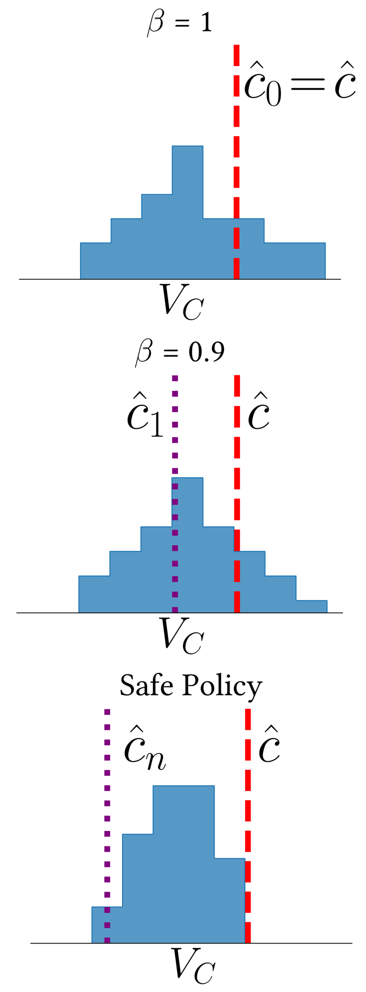


$$\Sigma = \Big\{ T \in \mathcal{P}(\mathbb{S}) \mid \|\hat{T}(\cdot \mid s, a) - T(\cdot \mid s, a)\| \le \underbrace{e(s, a)}_{\approx \frac{1}{N(s,a)}} \Big\}$$

Σ contains the true transition function with high probability.

Conservative policy

Tight safety constraint until ground policy is safe in all probable CMDPs (Σ).



Results