AlwaysSafe: Reinforcement Learning without Safety Constraint Violations during Training[1]


Thiago D. Simão*


Nils Jansen†


Matthijs Spaan*

*Delft University of Technology

†Radboud University, Nijmegen

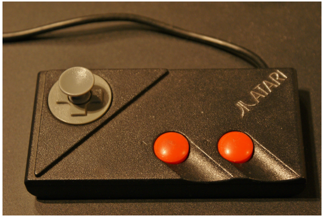
Planning and Reinforcement Learning Workshop

Aug 2021

[1] Extended abstract of a paper published at AAMAS-21

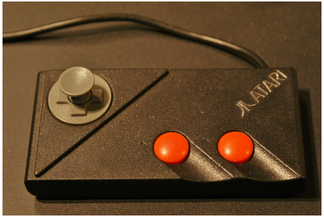
Gap between Research and Real-world

Simulations

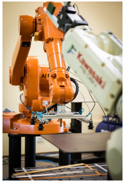


Gap between Research and Real-world

Simulations



Real-world tasks



Challenges for RL

Challenges to bring reinforcement learning from research to real-world applications[2]:

- Safety constraints
- Off-line training
- Limited interactions with the environment
- Partially observable tasks
- Explanability
- ⋮

[2] G. Dulac-Arnold et al. "Challenges of real-world reinforcement learning: definitions, benchmarks and analysis". In: *Machine Learning* (2021)

Challenges for RL

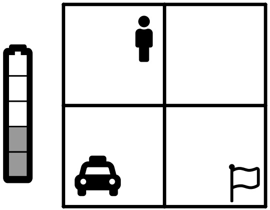
Challenges to bring reinforcement learning from research to real-world applications[2]:

- Safety constraints
- Off-line training
- Limited interactions with the environment
- Partially observable tasks
- Explanability

⋮

[2] G. Dulac-Arnold et al. “Challenges of real-world reinforcement learning: definitions, benchmarks and analysis”. In: *Machine Learning* (2021)

Electric Taxi



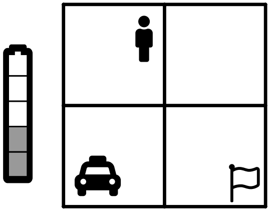
ⓟ : passenger info

ⓣ : taxi location

ⓑ : battery

◇ⓡ : passenger delivered

Electric Taxi



ⓟ : passenger info

ⓣ : taxi location

ⓑ : battery

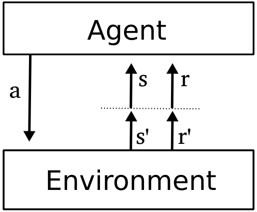
◇r : passenger delivered

◇c : out of power

Typical RL

MDP[3]: $\mathcal{M} = \langle \mathbb{S}, \mathbb{A}, P, R, \mu \rangle$

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



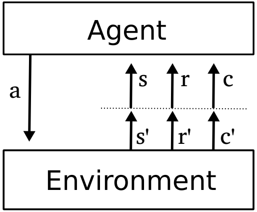
[3] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1st. John Wiley & Sons, Inc., 1994

Constrained RL

CMDP[4]: $\mathcal{M} = \langle \mathbb{S}, \mathbb{A}, P, R, \mu, C, \hat{c} \rangle$

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

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



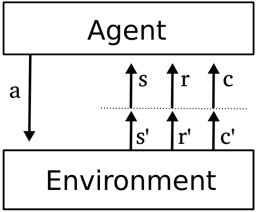
[4] E. Altman. *Constrained Markov Decision Processes*. Vol. 7. CRC Press, 1999

Constrained RL

CMDP[4]: $\mathcal{M} = \langle \mathbb{S}, \mathbb{A}, P, R, \mu, C, \hat{c} \rangle$

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

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



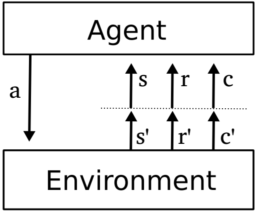
[4] E. Altman. *Constrained Markov Decision Processes*. Vol. 7. CRC Press, 1999

Constrained RL

CMDP[4]: $\mathcal{M} = \langle \mathbb{S}, \mathbb{A}, P, R, \mu, C, \hat{c} \rangle$

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

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



[4] E. Altman. *Constrained Markov Decision Processes*. Vol. 7. CRC Press, 1999

Solving a CMDP

Occupancy measure of state and action: $x(s, a, t) = \mathbb{E}\left[s_t = s, a_t = a\right]$

$$\max_x \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) R(s, a) \quad \text{s. t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) C(s, a) \leq \hat{c}$$

$$x \text{ respects } T$$

Solving a CMDP

Occupancy measure of state and action: $x(s, a, t) = \mathbb{E}[s_t = s, a_t = a]$

$$\max_x \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) R(s, a) \quad \text{s. t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) C(s, a) \leq \hat{c}$$

$$x \text{ respects } T$$

Solving a CMDP

Occupancy measure of state and action: $x(s,a,t) = \mathbb{E}\left[s_t = s, a_t = a\right]$

$$\max_x \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) R(s,a) \quad \text{s. t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) C(s,a) \leq \hat{c}$$

$$x \text{ respects } T$$

Solving a CMDP

Occupancy measure of state and action: $x(s, a, t) = \mathbb{E}\left[s_t = s, a_t = a\right]$

$$\max_x \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) R(s, a) \quad \text{s. t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) C(s, a) \leq \hat{c}$$

x respects T

Solving a CMDP

Occupancy measure of state and action: $x(s,a,t) = \mathbb{E}\left[s_t = s, a_t = a\right]$

$$\max_x \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) R(s,a) \quad \text{s.t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) C(s,a) \leq \hat{c}$$

$$x \text{ respects } T$$

Optimism in Face of Uncertainty

`OptCMDP`[5] optimistically chooses a transition function within the uncertainty set, uses the lower bound of the reward function and the upper bound of the cost function.

$$\Sigma = \left[T' \,\middle|\, \| \hat{T}(\cdot \mid s, a) - T'(\cdot \mid s, a) \| \leq e_\delta^T(s, a) \right] \Rightarrow \mathbb{P}(T \in \Sigma) \geq 1 - \delta$$

$$\max_{x, T'} \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) \left(\hat{R}(s, a) + e_\delta^R(s, a) \right) \quad \text{s. t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) \left(\hat{C}(s, a) - e_\delta^C(s, a) \right) \leq \hat{c}$$

$$x \text{ respects } T'$$

$$T' \in \Sigma$$

[5] Y. Efroni et al. "Exploration-Exploitation in Constrained MDPs". In: ICML Workshop on Theoretical Foundations of Reinforcement Learning. 2020

Optimism in Face of Uncertainty

`OptCMDP`[5] optimistically chooses a transition function within the uncertainty set,
uses the lower bound of the reward function and the upper bound of the cost function.

$$\Sigma = \left[T' \,\middle|\, \| \hat{T}(\cdot \mid s, a) - T'(\cdot \mid s, a) \| \leq e^T_\delta(s, a) \right] \Rightarrow \mathbb{P}(T \in \Sigma) \geq 1 - \delta$$

$$\max_{x, T'} \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) \left(\hat{R}(s, a) + e^R_\delta(s, a) \right) \quad \text{s. t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) \left(\hat{C}(s, a) - e^C_\delta(s, a) \right) \leq \hat{c}$$

$$x \text{ respects } T'$$

$$T' \in \Sigma$$

[5] Y. Efroni et al. “Exploration-Exploitation in Constrained MDPs”. In: ICML Workshop on Theoretical Foundations of Reinforcement Learning. 2020

Optimism in Face of Uncertainty

`OptCMDP`[5] optimistically chooses a transition function within the uncertainty set,
uses the lower bound of the reward function and the upper bound of the cost function.

$$\Sigma = \left[T' \,\middle|\, \| \hat{T}(\cdot \mid s, a) - T'(\cdot \mid s, a) \| \leq e_\delta^T(s, a) \right] \Rightarrow \mathbb{P}(T \in \Sigma) \geq 1 - \delta$$

$$\max_{x, T'} \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) \left(\hat{R}(s, a) + e_\delta^R(s, a) \right) \quad \text{s. t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s, a, t) \left(\hat{C}(s, a) - e_\delta^C(s, a) \right) \leq \hat{c}$$

$$x \text{ respects } T'$$

$$T' \in \Sigma$$

[5] Y. Efroni et al. "Exploration-Exploitation in Constrained MDPs". In: ICML Workshop on Theoretical Foundations of Reinforcement Learning. 2020

Optimism in Face of Uncertainty

`OptCMDP`[5] optimistically chooses a transition function within the uncertainty set, uses the lower bound of the reward function and the upper bound of the cost function.

$$\Sigma = \left[T' \,\middle|\, \| \hat{T}(\cdot \mid s,a) - T'(\cdot \mid s,a) \| \leq e^T_\delta(s,a) \right] \Rightarrow \mathbb{P}(T \in \Sigma) \geq 1 - \delta$$

$$\max_{x, T'} \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{R}(s,a) + e^R_\delta(s,a) \right) \quad \text{s.t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{C}(s,a) - e^C_\delta(s,a) \right) \leq \hat{c}$$

$$x \text{ respects } T'$$

$$T' \in \Sigma$$

[5] Y. Efroni et al. "Exploration-Exploitation in Constrained MDPs". In: ICML Workshop on Theoretical Foundations of Reinforcement Learning. 2020

Optimism in Face of Uncertainty

`OptCMDP`[5] optimistically chooses a transition function within the uncertainty set,
uses the lower bound of the reward function and the upper bound of the cost function.

$$\Sigma = \left[T' \,\middle|\, \|\hat{T}(\cdot \mid s,a) - T'(\cdot \mid s,a)\| \le e_\delta^T(s,a) \right] \Rightarrow \mathbb{P}(T \in \Sigma) \ge 1 - \delta$$

$$\max_{x,T'} \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{R}(s,a) + e_\delta^R(s,a) \right) \quad \text{s. t.} \quad \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{C}(s,a) - e_\delta^C(s,a) \right) \le \hat{c}$$

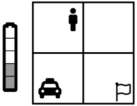
$$x \text{ respects } T'$$

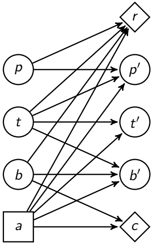
$$T' \in \Sigma$$

Bounded regret in terms of performance and safety but policy might be unsafe.

[5] Y. Efroni et al. "Exploration-Exploitation in Constrained MDPs". In: ICML Workshop on Theoretical Foundations of Reinforcement Learning. 2020

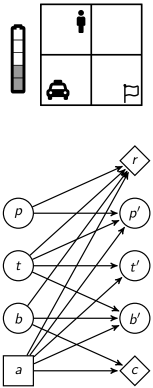
Not Everything is Relevant for Safety



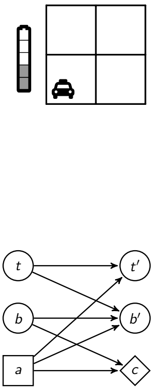


Factored MDP with cost function related to safety

Not Everything is Relevant for Safety



Factored MDP with cost function related to safety



Abstract factored MDP with safety dynamics

Cost-model-irrelevant Abstraction

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

$$\uparrow$$

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

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

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

ϕ preserves the expected cost of the policy.

Cost-model-irrelevant Abstraction

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

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

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

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

ϕ preserves the expected cost of the policy.

AbsOptCMDP

We can compute a safe policy in the abstract CMDP using a new variable z.

$$\max_{x,T',z} \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{R}(s,a) + e_{\delta}^{R}(s,a) \right) \quad \text{s. t.} \quad \sum_{\bar{s},a} \sum_{t=1}^{H} z(\bar{s},a,t) C(\bar{s},a) \le \hat{c}$$

$$x \text{ respects } T'$$

$$T' \in \Sigma$$

$$z \text{ respects } \bar{T}$$

$$z(\bar{s},a,t) = \sum_{s \in \phi^{-1}(\bar{s})} x(s,a,t) \quad \forall \bar{s},a,t$$

AbsOptCMDP

We can compute a safe policy in the abstract CMDP using a new variable z.

$$\max_{x, T', z} \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{R}(s,a) + e_{\delta}^{R}(s,a) \right) \quad \text{s. t.} \quad \sum_{\bar{s},a} \sum_{t=1}^{H} z(\bar{s},a,t) C(\bar{s},a) \leq \hat{c}$$

$$x \text{ respects } T'$$

$$T' \in \Sigma$$

$$z \text{ respects } \bar{T}$$

$$z(\bar{s},a,t) = \sum_{s \in \phi^{-1}(\bar{s})} x(s,a,t) \quad \forall \bar{s}, a, t$$

AbsOptCMDP

We can compute a safe policy in the abstract CMDP using a new variable z.

$$\max_{x,T',z} \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{R}(s,a) + e_\delta^R(s,a) \right) \quad \text{s. t.} \quad \sum_{\bar{s},a} \sum_{t=1}^{H} z(\bar{s},a,t) C(\bar{s},a) \leq \hat{c}$$

$$x \text{ respects } T'$$

$$T' \in \Sigma$$

$$z \text{ respects } \bar{T}$$

$$z(\bar{s},a,t) = \sum_{s \in \phi^{-1}(\bar{s})} x(s,a,t) \quad \forall \bar{s},a,t$$

AbsOptCMDP

We can compute a safe policy in the abstract CMDP using a new variable z.

$$\max_{x, T', z} \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{R}(s,a) + e_{\delta}^{R}(s,a) \right) \quad \text{s. t.} \quad \sum_{\bar{s},a} \sum_{t=1}^{H} z(\bar{s},a,t) C(\bar{s},a) \leq \hat{c}$$

$$x \text{ respects } T'$$

$$T' \in \Sigma$$

$$z \text{ respects } \bar{T}$$

$$z(\bar{s},a,t) = \sum_{s \in \phi^{-1}(\bar{s})} x(s,a,t) \quad \forall \bar{s},a,t$$

AbsOptCMDP

We can compute a safe policy in the abstract CMDP using a new variable z.

$$\max_{x, T', z} \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{R}(s,a) + e_{\delta}^{R}(s,a) \right) \quad \text{s. t.} \quad \sum_{\bar{s},a} \sum_{t=1}^{H} z(\bar{s},a,t) C(\bar{s},a) \leq \hat{c}$$

$$x \text{ respects } T'$$

$$T' \in \Sigma$$

$$z \text{ respects } \bar{T}$$

$$z(\bar{s},a,t) = \sum_{s \in \phi^{-1}(\bar{s})} x(s,a,t) \quad \forall \bar{s}, a, t$$

- z induces an abstract policy π_A.

AbsOptCMDP

We can compute a safe policy in the abstract CMDP using a new variable z.

$$\max_{x, T', z} \sum_{s,a} \sum_{t=1}^{H} x(s,a,t) \left(\hat{R}(s,a) + e_\delta^R(s,a) \right) \quad \text{s. t.} \quad \sum_{\bar{s},a} \sum_{t=1}^{H} z(\bar{s},a,t) C(\bar{s},a) \leq \hat{c}$$

$$x \text{ respects } T'$$

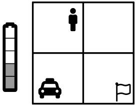
$$T' \in \Sigma$$

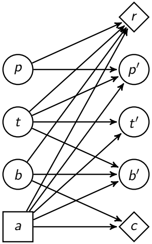
$$z \text{ respects } \bar{T}$$

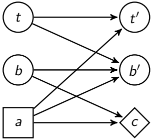
$$z(\bar{s},a,t) = \sum_{s \in \phi^{-1}(\bar{s})} x(s,a,t) \quad \forall \bar{s}, a, t$$

- z induces an abstract policy π_A.
- x induces a ground policy π_G.

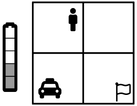
We May Need Everything to Compute an Optimal Policy

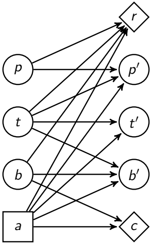


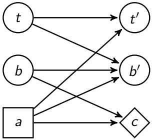




We May Need Everything to Compute an Optimal Policy

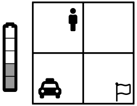


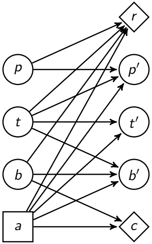


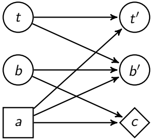


- The abstract policy π_A is safe

We May Need Everything to Compute an Optimal Policy

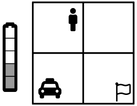


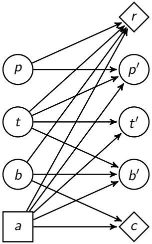


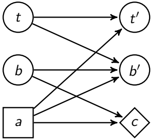


- The abstract policy π_A is safe but might be suboptimal.

We May Need Everything to Compute an Optimal Policy

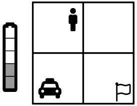


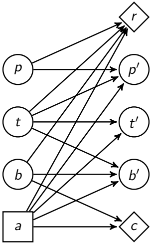


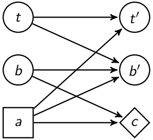


- The abstract policy π_A is safe but might be suboptimal.
- The ground policy π_G can reach optimality

We May Need Everything to Compute an Optimal Policy



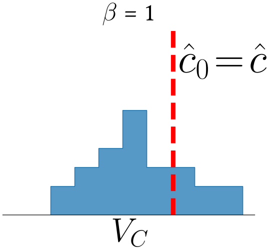




- The abstract policy π_A is safe but might be suboptimal.
- The ground policy π_G can reach optimality but has no safety guarantees.

AlwaysSafe π_α

Dynamically adjusting the safety constraint to ensure safety

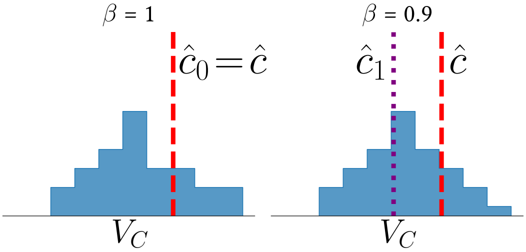


Search for policy that is safe in the whole uncertainty set.

1. $\hat{c}_t \leftarrow \beta_t \hat{c}$
2. compute π_α according to $\hat{c}_t$
3. $\beta_t \leftarrow \beta_{t-1} - \alpha \frac{\max\{\max_{T' \in \Sigma} V_C(\pi_\alpha) - \hat{c}, 0\}}{\hat{c}}$

AlwaysSafe π_α

Dynamically adjusting the safety constraint to ensure safety

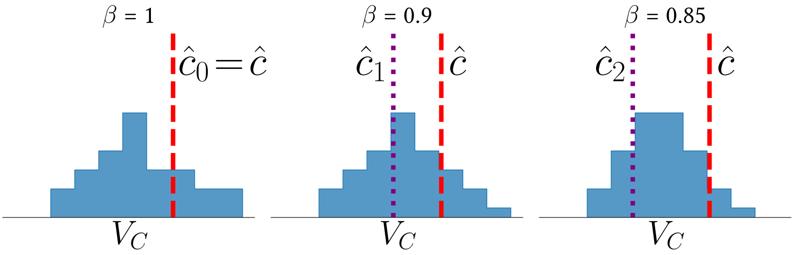


Search for policy that is safe in the whole uncertainty set.

1. $\hat{c}_t \leftarrow \beta_t \hat{c}$
2. compute π_α according to $\hat{c}_t$
3. $\beta_t \leftarrow \beta_{t-1} - \alpha \frac{\max\{\max_{T' \in \Sigma} V_C(\pi_\alpha) - \hat{c}, 0\}}{\hat{c}}$

AlwaysSafe π_α

Dynamically adjusting the safety constraint to ensure safety

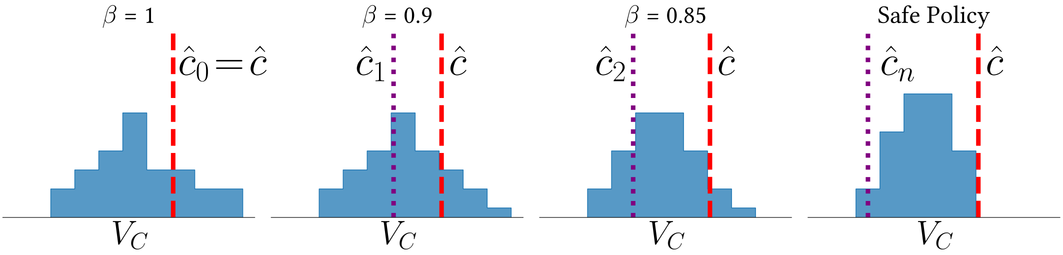


Search for policy that is safe in the whole uncertainty set.

1. $\hat{c}_t \leftarrow \beta_t \hat{c}$
2. compute π_α according to $\hat{c}_t$
3. $\beta_t \leftarrow \beta_{t-1} - \alpha \frac{\max\{\max_{T' \in \Sigma} V_C(\pi_\alpha) - \hat{c}, 0\}}{\hat{c}}$

AlwaysSafe π_α

Dynamically adjusting the safety constraint to ensure safety

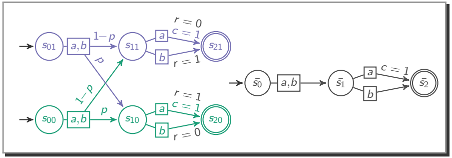


Search for policy that is safe in the whole uncertainty set.

1. $\hat{c}_t \leftarrow \beta_t \hat{c}$
2. compute π_α according to $\hat{c}_t$
3. $\beta_t \leftarrow \beta_{t-1} - \alpha \frac{\max\{\max_{T' \in \Sigma} V_C(\pi_\alpha) - \hat{c}, 0\}}{\hat{c}}$

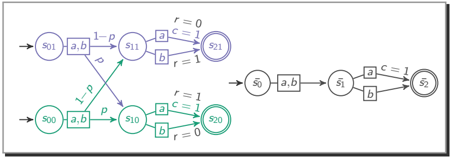
Empirical Results

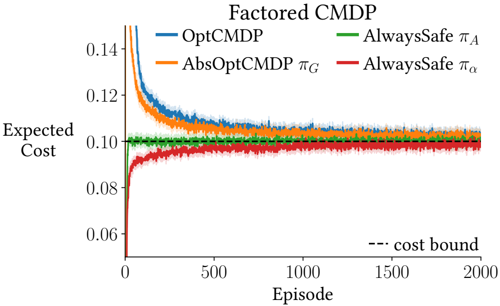
$p = 0.9$ and $\hat{c} = 0.1$



Empirical Results

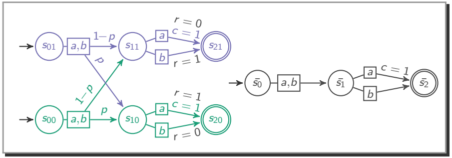
$p = 0.9$ and $\hat{c} = 0.1$

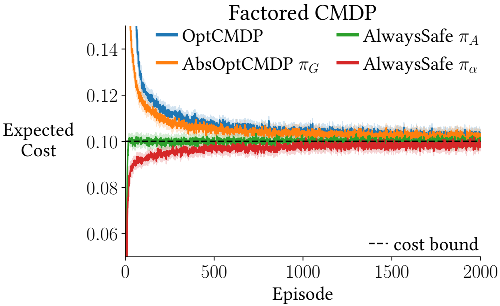


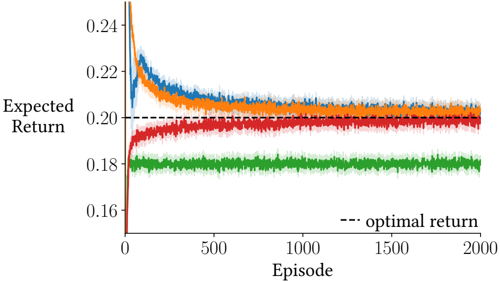


Empirical Results

$p = 0.9$ and $\hat{c} = 0.1$







Take Home Message

Constrained RL

- models safety requirements explicitly and
- avoids reward engineering/hacking.

Take Home Message

Constrained RL

- models safety requirements explicitly and
- avoids reward engineering/hacking.

The algorithm proposed

- is always safe during the learning process (with high probability),
- seamlessly switches from a conservative policy to a greedy policy and
- can explore optimistically.

Take Home Message

Constrained RL

- models safety requirements explicitly and
- avoids reward engineering/hacking.

The algorithm proposed

- is always safe during the learning process (with high probability),
- seamlessly switches from a conservative policy to a greedy policy and
- can explore optimistically.

Thank you!