Near-Optimal Sample Complexity Bounds for Constrained MDPs

Sharan Vaswani*, Lin F. Yang*, Csaba Szepesvari



https://arxiv.org/abs/2206.06270

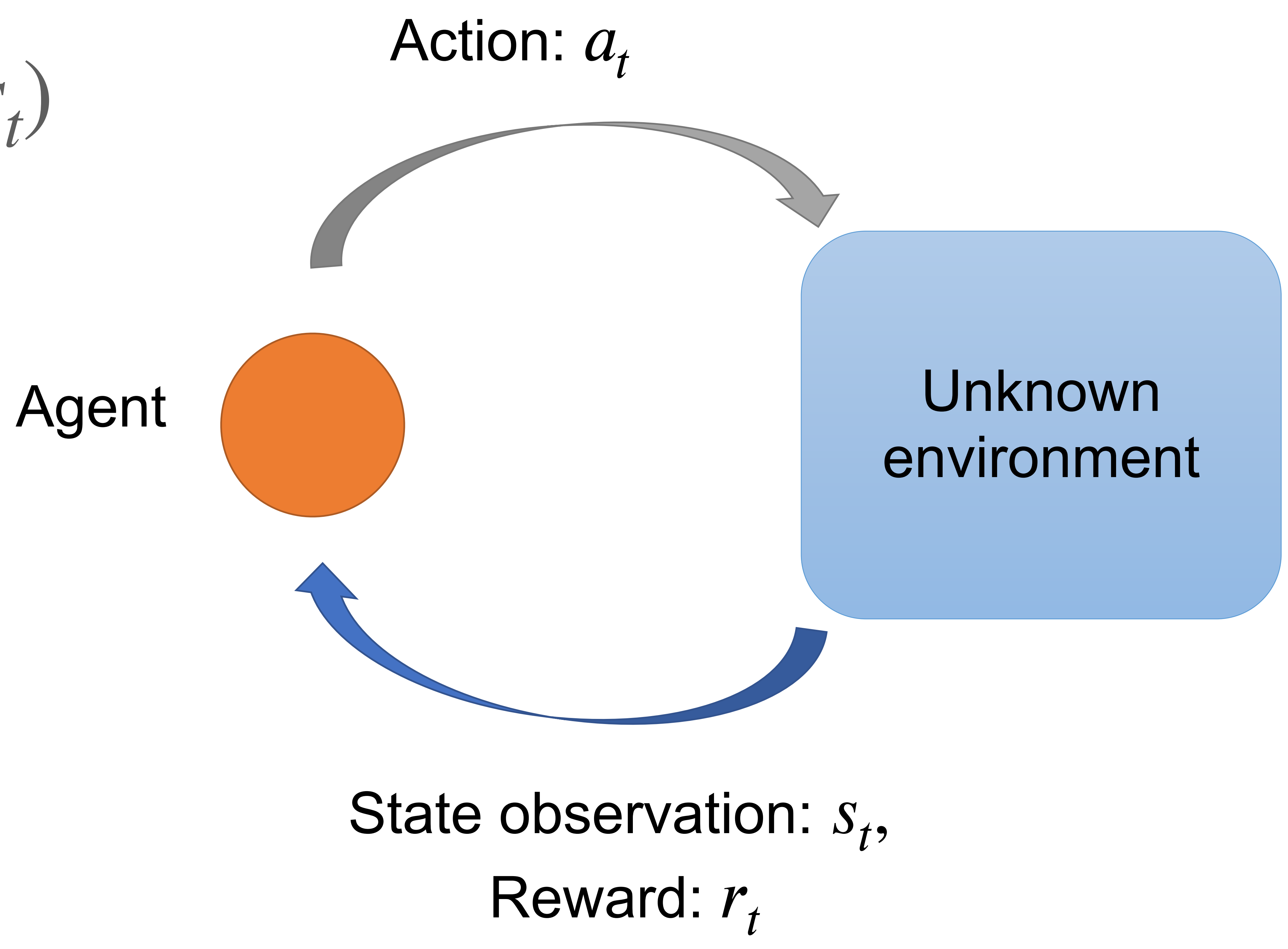
Reinforcement Learning

Learn to interact with an unknown environment through trial and error

Goal: Find a policy π to maximize the value (cumulative infinite-horizon discounted reward)

Value: $V_r^\pi := \mathbb{E}[r_0 + \gamma r_1 + \gamma^2 r_2 + \ldots] \ ; a_t = \pi(s_t)$

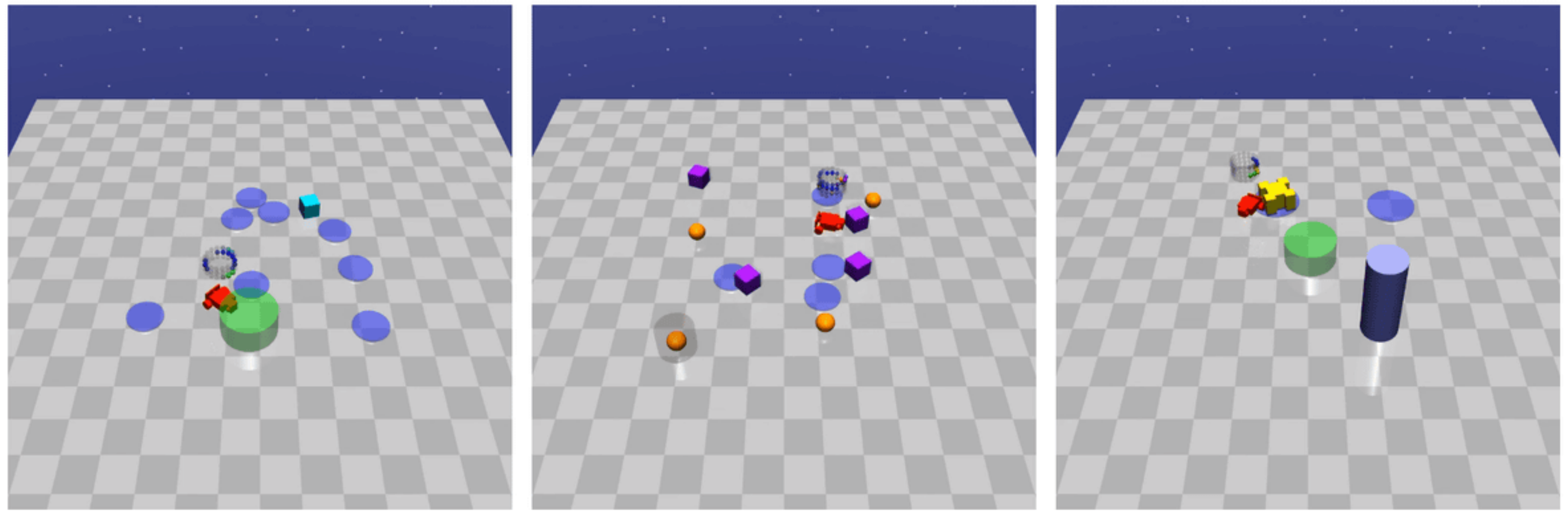
Discount factor



Constrained Reinforcement Learning

Maximize the reward value subject to a constraint

$$\max_{\pi} V_r^{\pi} \text{ s.t. } V_c^{\pi} \geq b$$



Goal: Move to a series of goal positions. **Button**: Press a series of goal buttons. **Push**: Move a box to a series of goal positions.

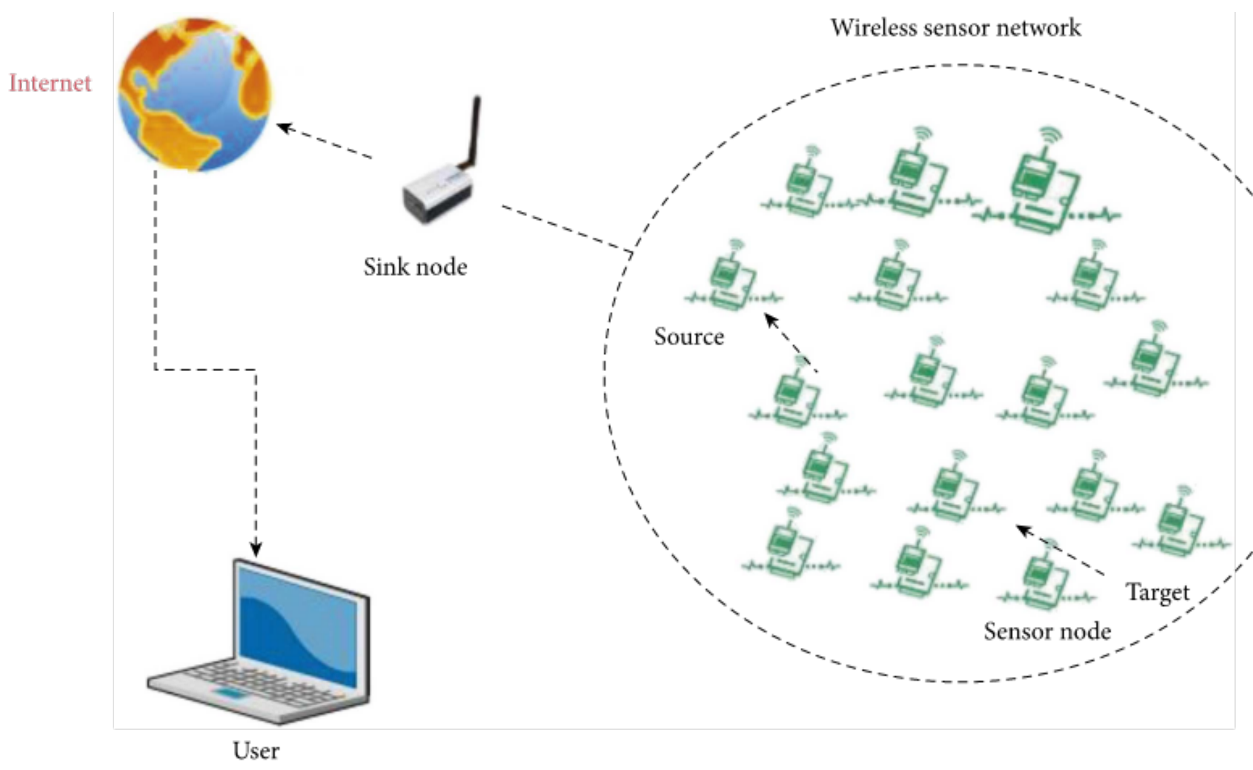
Safety Gym (OpenAI)

Example:

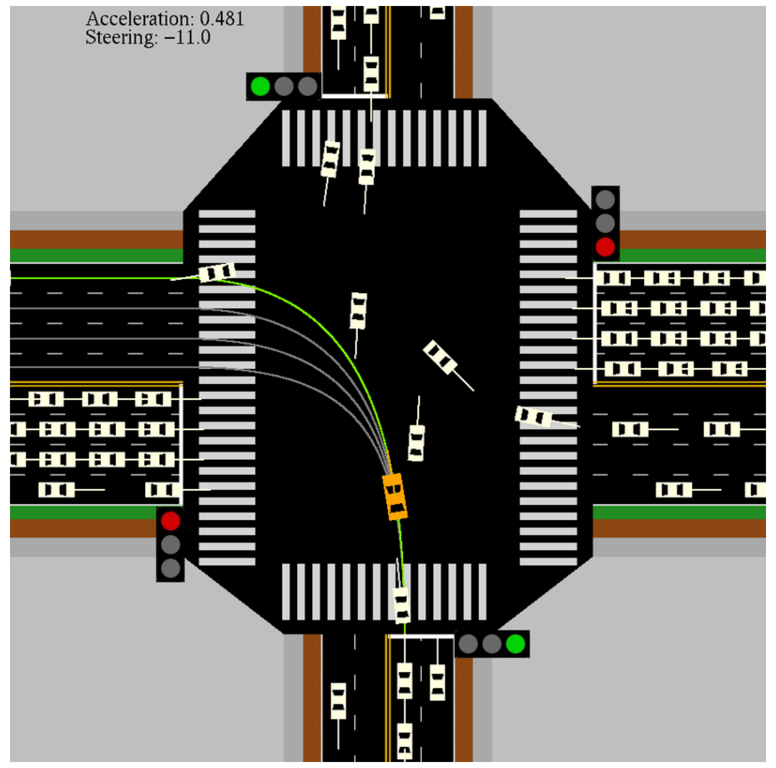
r : task reward

c : negative “energy used” (constraint reward)

Maximize task reward while keeping energy use below threshold.



(Malik et al, 2020)

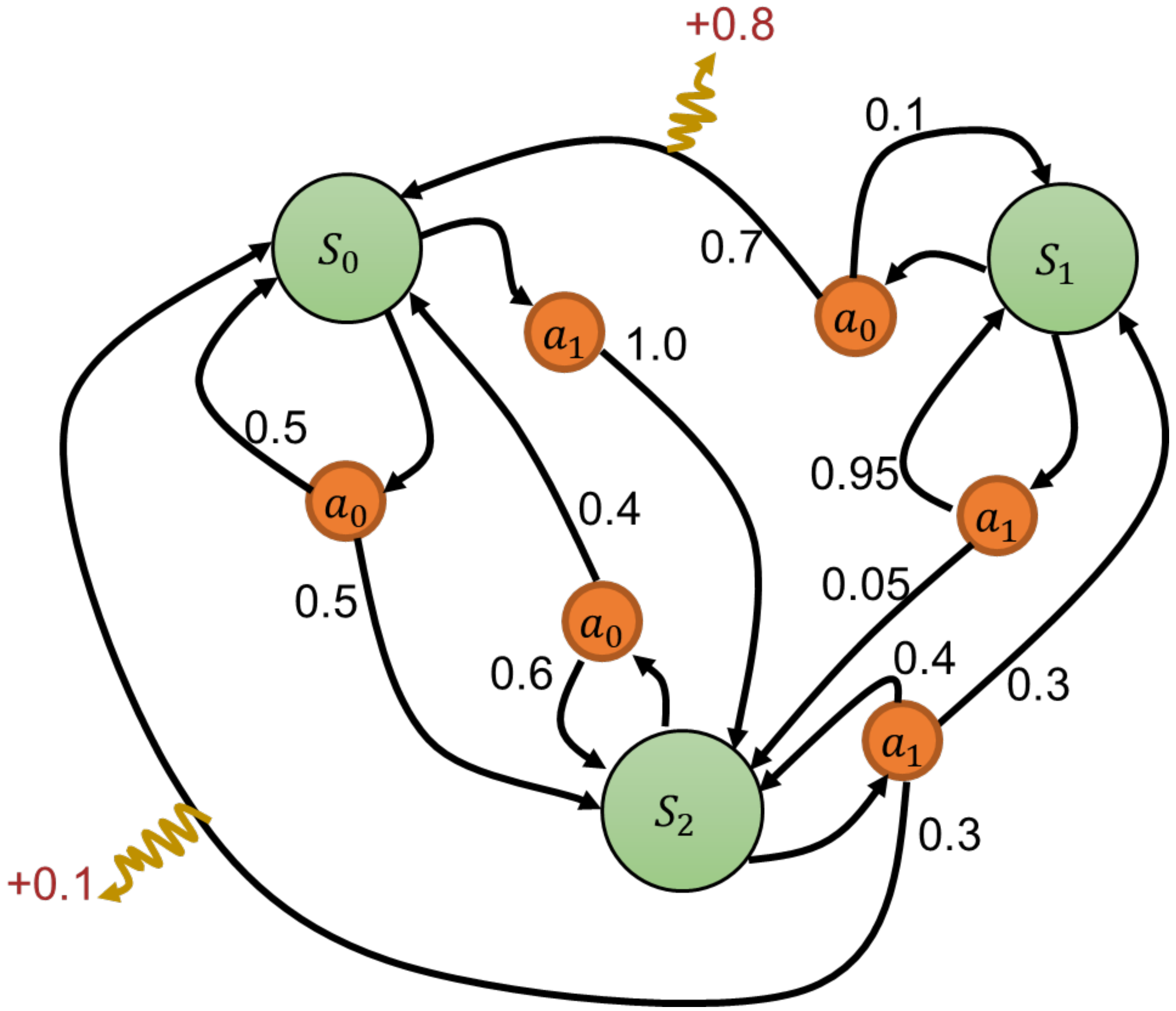


(Ma et al, 2021)

Constrained Markov Decision Processes

- States: $\mathcal{S}$; Actions: $\mathcal{A}$
- Rewards: $r_{s,a} \in [0,1]$
- State transitions: $P(s' \mid s, a)$
- Constraint rewards: $c_{s,a} \in [0,1]$
- Constraint threshold: b

- Initial state distribution: ρ
- Discount factor: $\gamma \in (0,1)$
- Policy: $\pi : \mathcal{S} \to \mathcal{A}$



Constrained Markov Decision Processes

$$\max_{\pi} V_r^{\pi}(\rho) \text{ s.t. } V_c^{\pi}(\rho) \geq b$$

Initial state distribution

Constrained Markov Decision Processes

$$\max_{\pi} V_r^{\pi}(\rho) \text{ s.t. } V_c^{\pi}(\rho) \geq b$$

Initial state distribution

Optimal policy & value: π^*, $V_r^*(\rho)$

ϵ-optimal policy π: $V_r^{\pi}(\rho) \geq V_r^*(\rho) - \epsilon$

Compared to MDPs,

Optimal policy might need to randomize

Optimal policy changes with ρ

Constrained Markov Decision Processes

$$\max_{\pi} V_r^{\pi}(\rho) \text{ s.t. } V_c^{\pi}(\rho) \geq b$$

Initial state distribution

Optimal policy & value: π^*, $V_r^*(\rho)$

ϵ-optimal policy π: $V_r^{\pi}(\rho) \geq V_r^*(\rho) - \epsilon$

Compared to MDPs,

Optimal policy might need to randomize

Optimal policy changes with ρ

Feasibility Assumption: $\zeta := \max_{\pi} V_c^{\pi}(\rho) - b > 0$

Slater constant

Sample complexity of planning

Generative model

The agent can obtain samples from $P(\cdot \mid s, a)$ for every (s, a)

$r_{s,a}$, $c_{s,a}$ is known at all (s, a) pairs, but P is unknown

Sample complexity of planning

Generative model

The agent can obtain samples from $P(\cdot \mid s, a)$ for every (s, a)

$r_{s,a}$, $c_{s,a}$ is known at all (s, a) pairs, but P is unknown

Q: How many samples are needed from the generative model to output policy $\hat{\pi}$ such that:

1. **Relaxed Feasibility:** $V_r^{\hat{\pi}}(\rho) \geq V_r^{*}(\rho) - \epsilon$ **and** $V_c^{\hat{\pi}}(\rho) \geq b - \epsilon$

2. **Strict Feasibility:** $V_r^{\hat{\pi}}(\rho) \geq V_r^{*}(\rho) - \epsilon$ **and** $V_c^{\hat{\pi}}(\rho) \geq b$

Sample complexity of planning - Existing Bounds

MDPs:

Effective Horizon = $\frac{1}{1-\gamma}$

Lower Bound: $\Omega(H^3 SA\, \epsilon^{-2})$ [Azar et al' 2013]

Upper Bound: $\tilde{O}(H^3 SA\, \epsilon^{-2})$ [Sidford et al. 2018, Agarwal et al. 2020,Li et al. 2021]

Sample complexity of planning - Existing Bounds

MDPs:

Effective Horizon = $\frac{1}{1-\gamma}$

Lower Bound: $\Omega(H^3 SA\, \epsilon^{-2})$ [Azar et al' 2013]

Upper Bound: $\tilde{O}(H^3 SA\, \epsilon^{-2})$ [Sidford et al. 2018, Agarwal et al. 2020,Li et al. 2021]

CMDPs:

Trivial Lower Bound: $\Omega(H^3 SA\, \epsilon^{-2})$ (since MDPs are a special case of CMDPs)

Upper Bound:

1. Relaxed Feasibility:
$\tilde{O}(H^3 S^2 A\, \epsilon^{-2})$ [HasanzadeZonuzy et al 2021]
$\tilde{O}(H^5 SA\, \epsilon^{-2})$ [Ding et al 2021]

2. Strict Feasibility: $\tilde{O}(H^6 SA\, \epsilon^{-2} \zeta^{-2})$ [Bai et al 2022]

Sample complexity of planning - Our results

1. Relaxed Feasibility

Upper Bound: Model-based algorithm that requires $\tilde{O}\left(H^3SA\epsilon^{-2}\right)$ samples

Sample complexity of planning - Our results

1. Relaxed Feasibility

Upper Bound: Model-based algorithm that requires $\tilde{O}\left(H^3SA\epsilon^{-2}\right)$ samples

2. Strict Feasibility

Lower Bound:

For any $\delta \in (0,1)$, $\epsilon \in [0,H]$, there exists a CMDP with Slater constant ζ such that any (ϵ, δ)-algorithm requires $\Omega(H^5SA\,\epsilon^{-2}\zeta^{-2})$ samples

Sample complexity of planning - Our results

1. Relaxed Feasibility

Upper Bound: Model-based algorithm that requires $\tilde{O}\left(H^3 SA\epsilon^{-2}\right)$ samples

2. Strict Feasibility

Lower Bound:

For any $\delta \in (0,1)$, $\epsilon \in [0,H]$, there exists a CMDP with Slater constant ζ such that any (ϵ, δ)-algorithm requires $\Omega(H^5 SA\, \epsilon^{-2}\zeta^{-2})$ samples

Upper Bound: Model-based algorithm that requires $\tilde{O}\left(H^5 SA\epsilon^{-2}\zeta^{-2}\right)$ samples