Robust and Data-Driven Markov Decision Processes

Wolfram Wiesemann
Imperial College London

Markov decision process

Tuple $(\mathcal{S}, \mathcal{A}, q, p, r, \lambda)$ where

- $\mathcal{S} = \{1,\ldots,S\}$ is the (finite) state space;
- $\mathcal{A} = \{1,\ldots,A\}$ is the (finite) action space;
- $q = (q_1, \ldots, q_S) \in \Delta(\mathcal{S})$ is the initial state distribution;
- $p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel with elements $p(s' \,|\, s, a)$;
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are the expected one-step rewards;
- $\lambda \in (0,1)$ is the discount factor.

Markov Decision Processes

Markov decision process

Tuple $(\mathscr{S}, \mathscr{A}, q, p, r, \lambda)$ where

- $\mathscr{S} = \{1, \ldots, S\}$ is the (finite) state space;
- $\mathscr{A} = \{1, \ldots, A\}$ is the (finite) action space;
- $q = (q_1, \ldots, q_S) \in \Delta(\mathscr{S})$ is the initial state distribution;
- $p : \mathscr{S} \times \mathscr{A} \to \Delta(\mathscr{S})$ is the transition kernel with elements $p(s' \mid s, a)$;
- $r : \mathscr{S} \times$
- $\lambda \in (0, 1$

Objective

find policy π that maximizes the expected total discounted rewards:

$$\underset{\pi \in \Pi}{\text{maximize}} \quad \mathbb{E}_p \left[\sum_{t=1}^{\infty} \lambda^{t-1} \cdot r\big(s_t, \pi_t[s_t]\big) \right]$$

MDPs: Key Results

- **Stationary deterministic policies are optimal:**

$$\pi = \{\pi_t\}_{t=1}^{\infty}$$

$$\pi_t : (\mathcal{S} \times \mathcal{A})^{t-1} \times \mathcal{S} \to \Delta(\mathcal{A})$$

history-dependent, randomized

$\supset$

$$\pi : \mathcal{S} \to \Delta(\mathcal{A})$$

stationary, randomized

$\supset$

$$\pi : \mathcal{S} \to \mathcal{A}$$

stationary, deterministic

MDPs: Key Results

- **Stationary deterministic policies are optimal.**
- **Discounted rewards of a fixed policy**

- **Stationary deterministic policies are optimal.**
- **Discounted rewards of a fixed policy**

$$v^{\pi}(s) = \mathbb{E}_p\left[\sum_{t=1}^{\infty} \lambda^{t-1} \cdot r\left(s_t, \pi[s_t]\right) \,\middle|\, s_1 = s\right]$$

- **Stationary deterministic policies** are **optimal**.
- **Discounted rewards** of a **fixed policy** satisfy **linear equations**:

$$v^{\pi}(s) = r(s, \pi[s]) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s, \pi[s]) \cdot v^{\pi}(s')$$

- **Stationary deterministic policies are optimal.**
- **Discounted rewards of a fixed policy satisfy linear equations.**
- **Discounted rewards of an optimal policy satisfy *non*linear equations:**

$$v^\star(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s, a) \cdot v^\star(s') \right\}$$

- **Stationary deterministic policies are optimal.**
- **Discounted rewards of a fixed policy satisfy linear equations.**
- **Discounted rewards of an optimal policy satisfy *non*linear equations:**

$$v^\star(s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s'\in\mathcal{S}} p(s'\,|\,s,a)\cdot v^\star(s') \right\}$$

- $v^\star$**-greedy policy is optimal:**

$$\pi^\star(s) \in \arg\max_{a\in\mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s'\in\mathcal{S}} p(s'\,|\,s,a)\cdot v^\star(s') \right\}$$

MDPs: Solution Methods

- **Value iteration:**

$$v^\star(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s, a) \cdot v^\star(s') \right\}$$

- **Value iteration:**

$$\cancel{v^\star(s)} = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s, a) \cdot \cancel{v^\star(s')} \right\}$$

$$v^{i+1}(s) \longleftarrow v^i(s')$$

Starting from any $v^0 \in \mathbb{R}^S$, converges at linear rate to $v^\star$.

- **Value iteration.**
- **(Modified) Policy iteration:**

$$\pi^\star(s) \in \arg\max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s,a) \cdot v^\star(s') \right\}$$

- **Value iteration.**
- **(Modified) Policy iteration:**

$$\cancel{\pi^\star(s)} \in \arg\max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s, a) \cdot \cancel{v^\star(s')} \right\}$$

$$\pi^{i+1}(s) \longleftarrow v^{\pi^i}(s')$$

Policy improvement $\rightleftarrows$ **Policy evaluation**

Under suitable conditions, converges at superlinear rate to $v^\star$.
Converges to (ϵ-)optimal policy in finitely many iterations.

- **Value iteration**.
- **(Modified) Policy iteration**.
- **Linear programming**:

$$\begin{array}{lll} \underset{v \in \mathbb{R}^S}{\text{minimize}} & \displaystyle\sum_{s \in \mathcal{S}} q(s) \cdot v(s) & \\ \text{subject to} & v(s) = \displaystyle\max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right\} & \forall s \in \mathcal{S} \end{array}$$

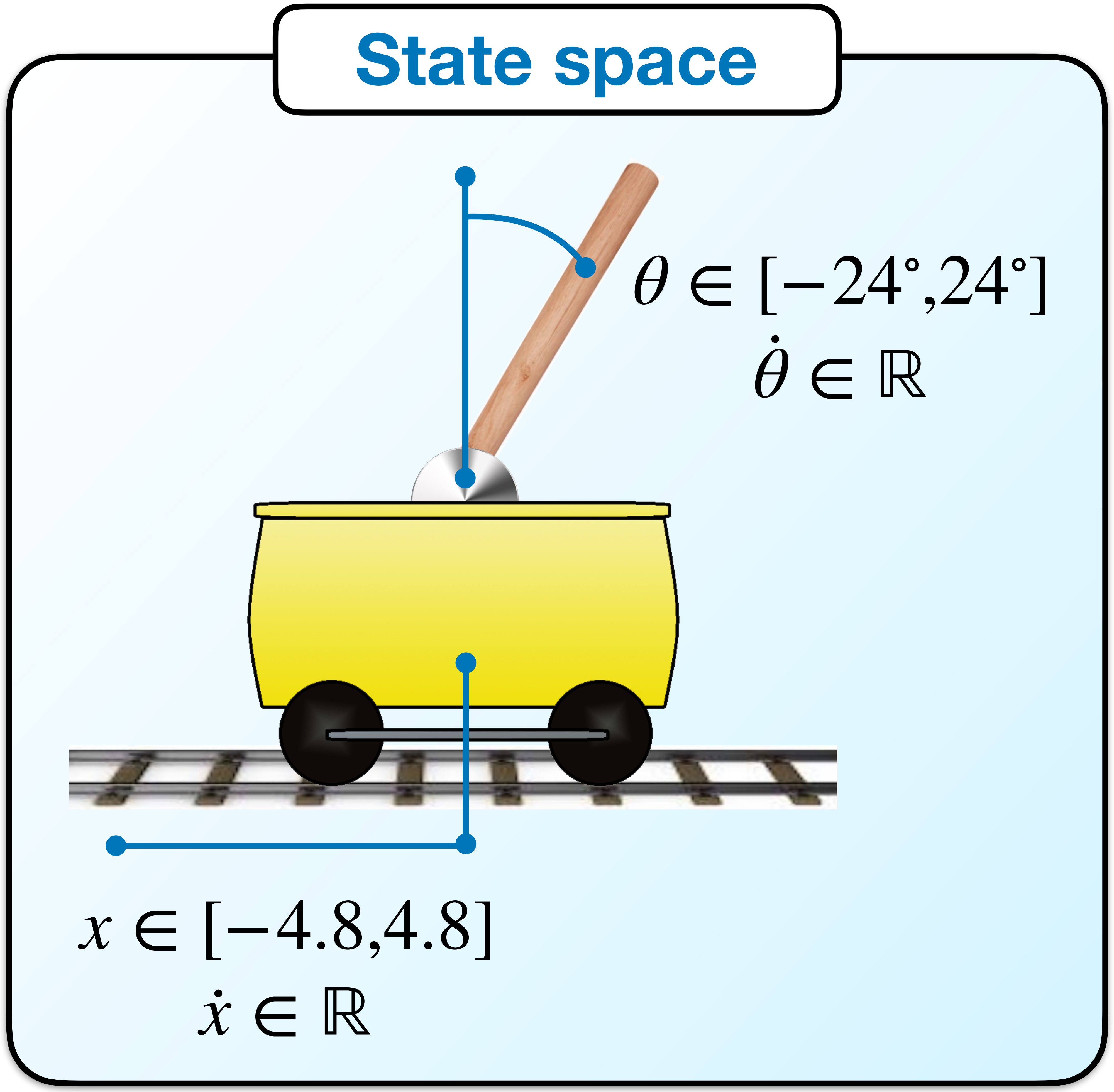
- **Value iteration.**
- **(Modified) Policy iteration.**
- **Linear programming:**

$$
\begin{array}{lll}
\underset{v \in \mathbb{R}^S}{\text{minimize}} & \displaystyle\sum_{s \in \mathcal{S}} q(s) \cdot v(s) & \\
\text{subject to} & v(s) \geq \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \displaystyle\sum_{s' \in \mathcal{S}} p(s' \,|\, s, a) \cdot v(s') \right\} & \forall s \in \mathcal{S}
\end{array}
$$

- **Value iteration.**
- **(Modified) Policy iteration.**
- **Linear programming:**

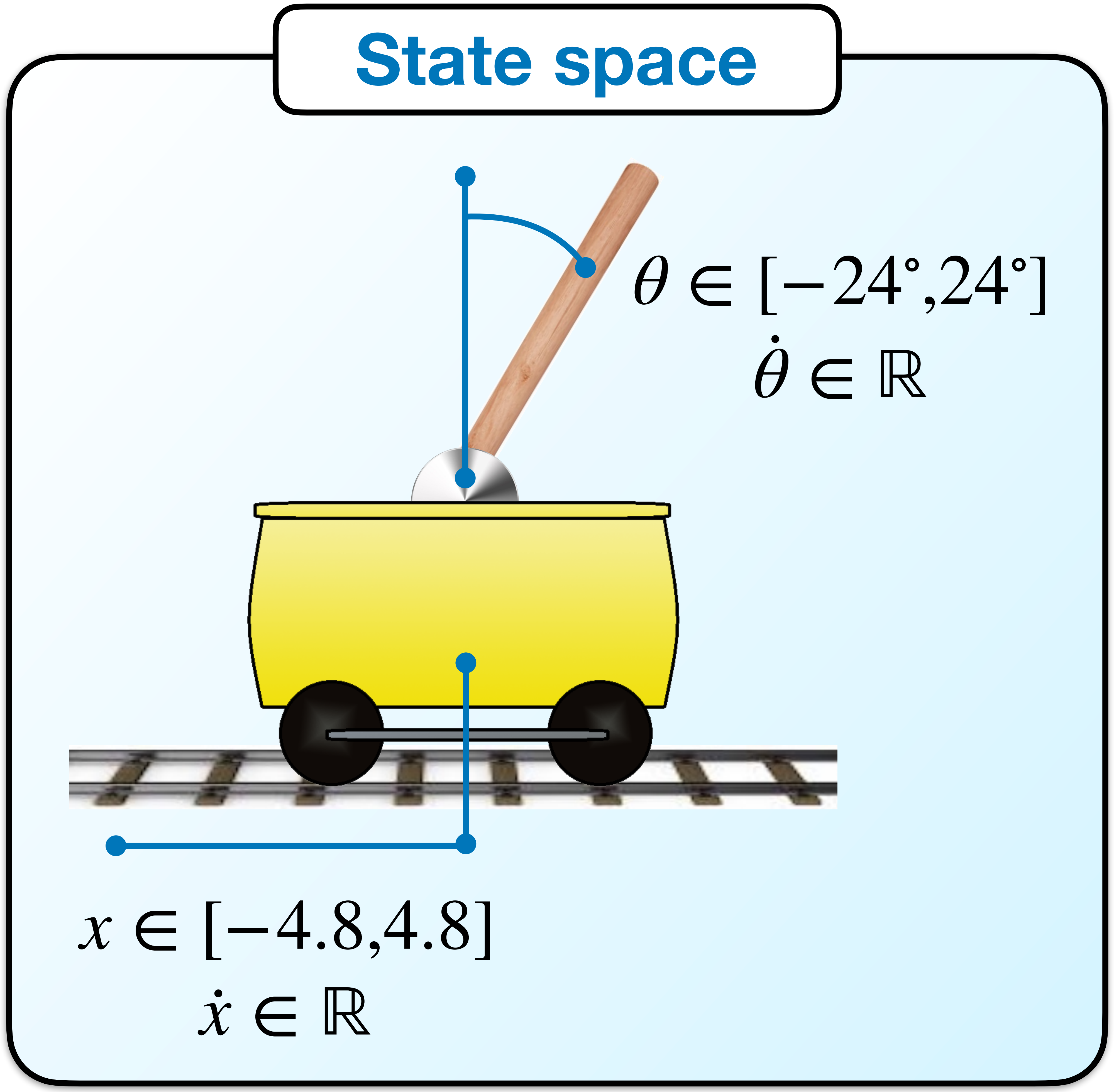
$$\begin{array}{lll} \underset{v \in \mathbb{R}^S}{\text{minimize}} & \displaystyle\sum_{s \in \mathcal{S}} q(s) \cdot v(s) & \\ \text{subject to} & v(s) \geq r(s,a) + \lambda \displaystyle\sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') & \forall s \in \mathcal{S}, \forall a \in \mathcal{A} \end{array}$$

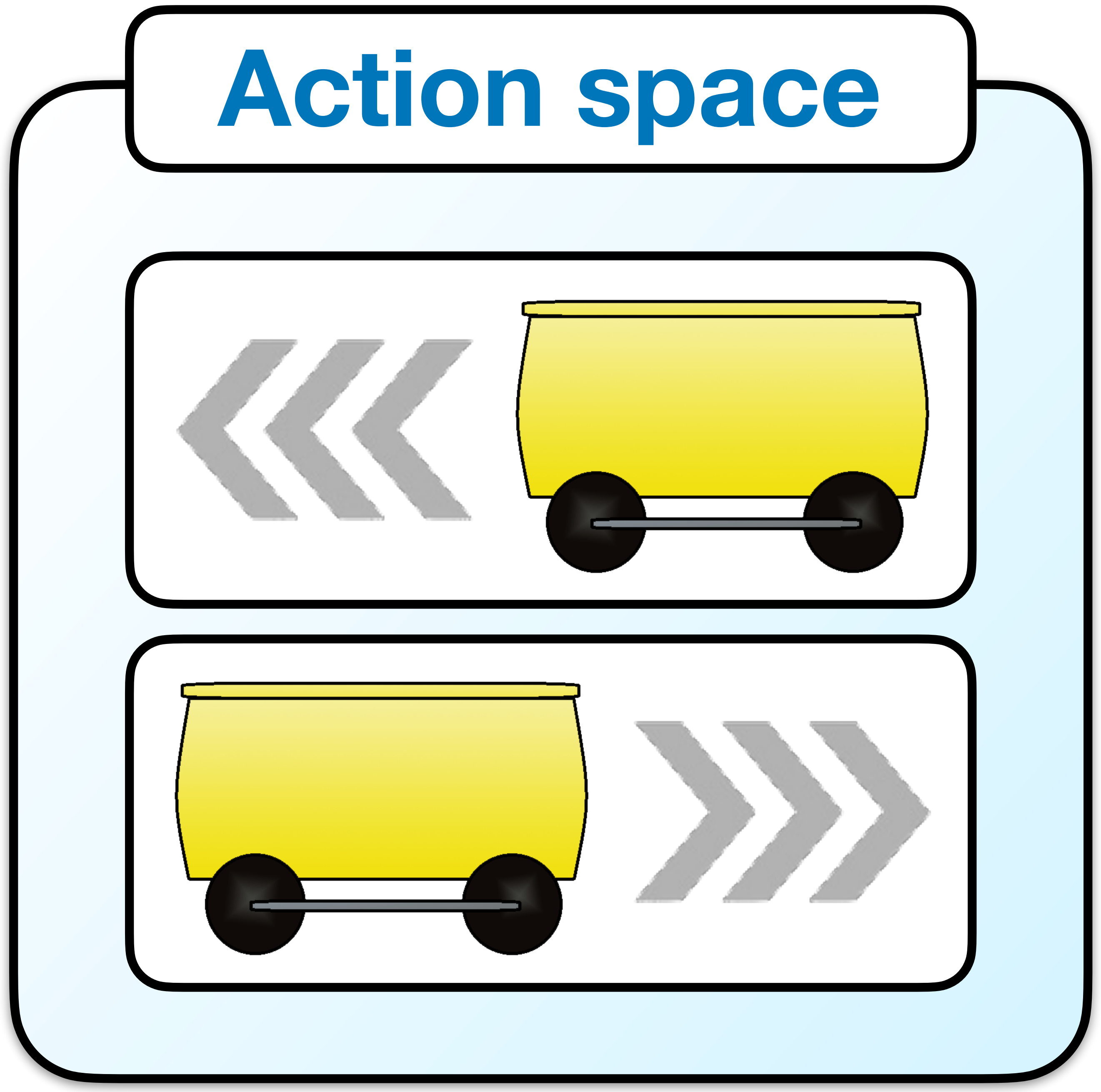
Optimally solved in polynomial time with standard solvers.



Barto et al. (1983), *Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems*.

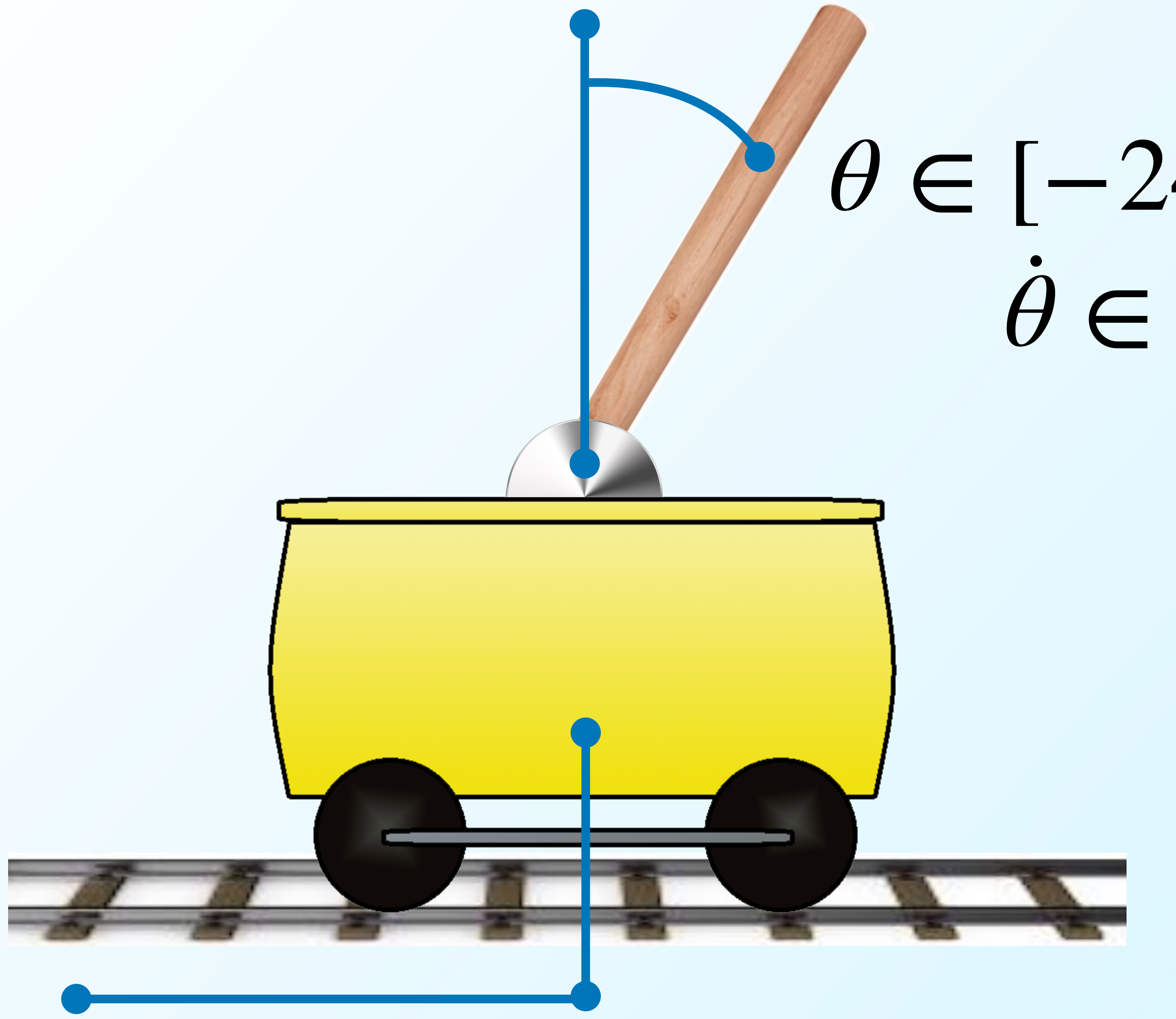
Cart Pole Example





Barto et al. (1983), *Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems*.

State space



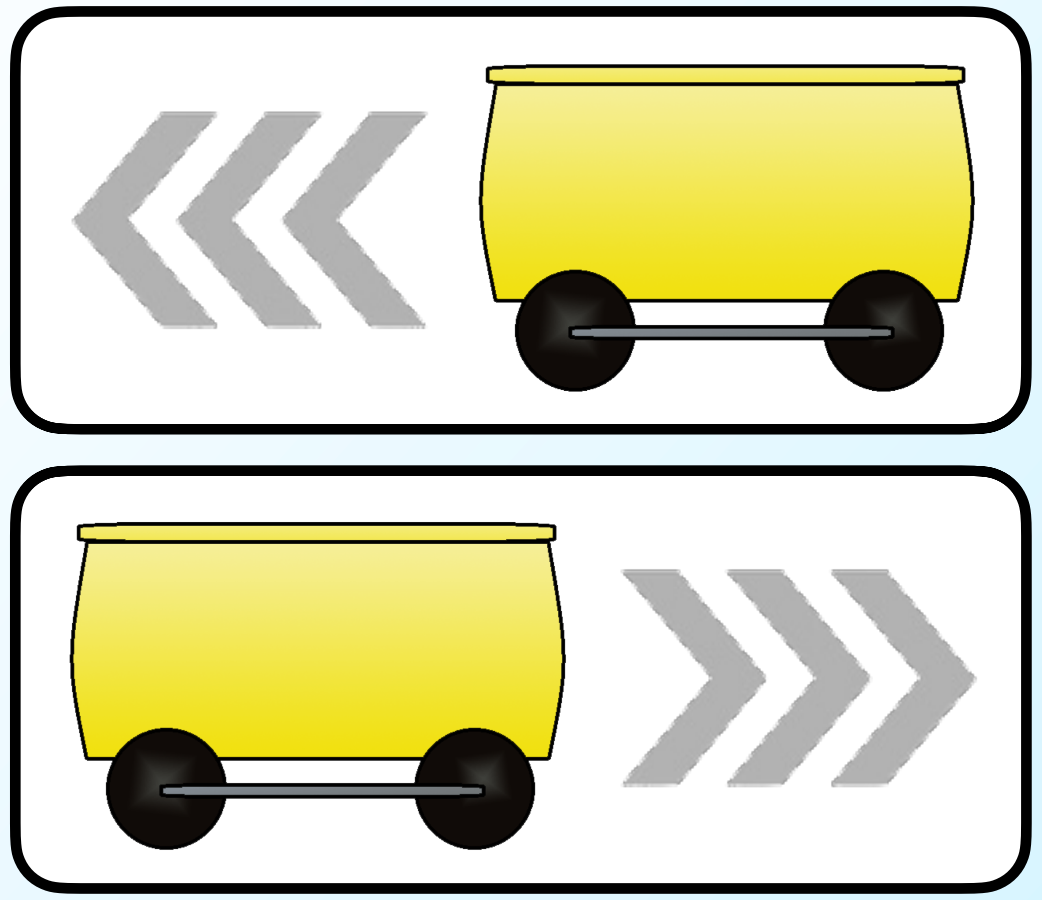
$\theta \in [-24°, 24°]$

$\dot{\theta} \in \mathbb{R}$

$x \in [-4.8, 4.8]$

$\dot{x} \in \mathbb{R}$

Action space

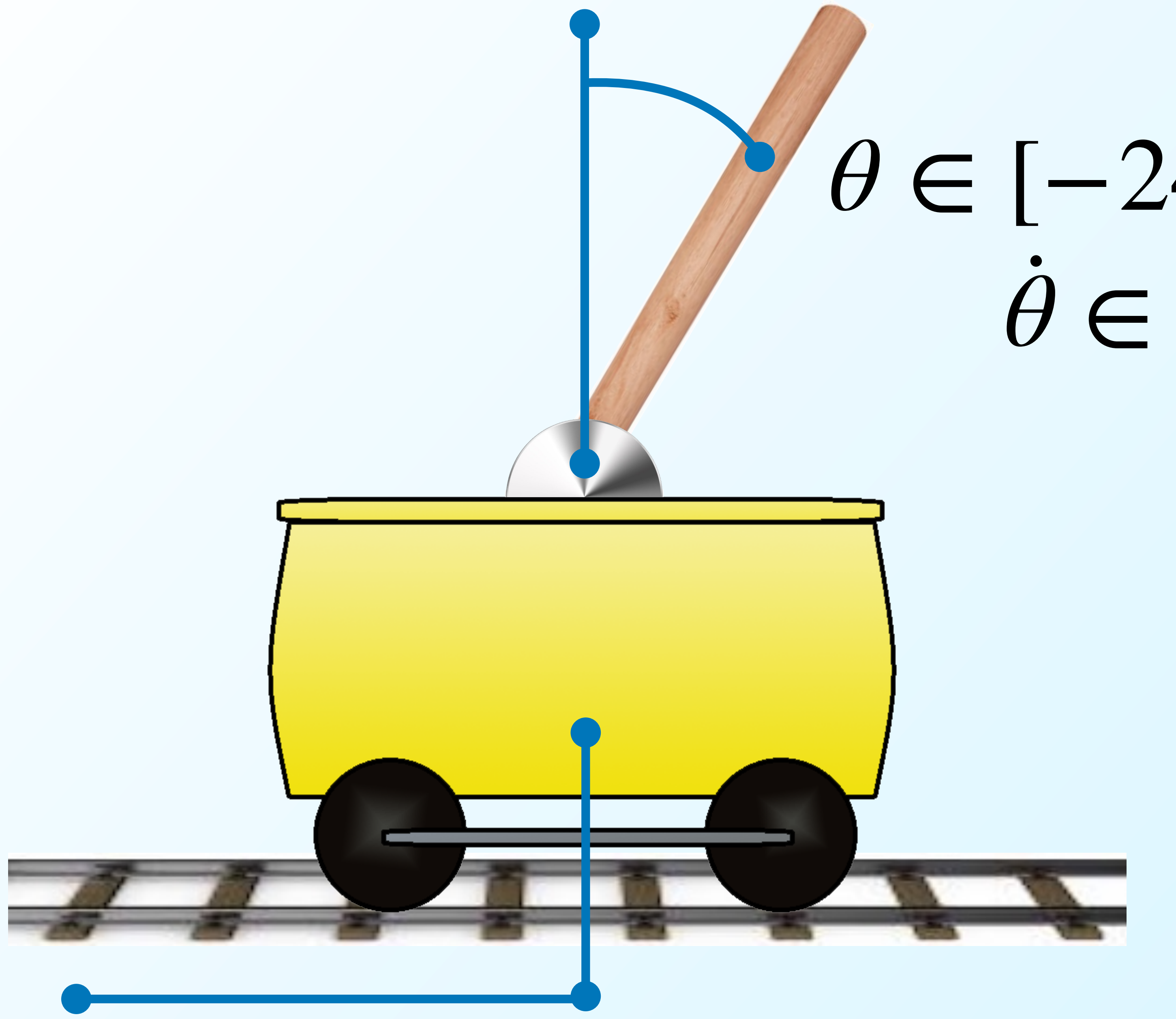


Initial state

$$x, \dot{x}, \theta, \dot{\theta} \sim \mathcal{U}[-0.05, 0.05]$$

Barto et al. (1983), *Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems*.

State space



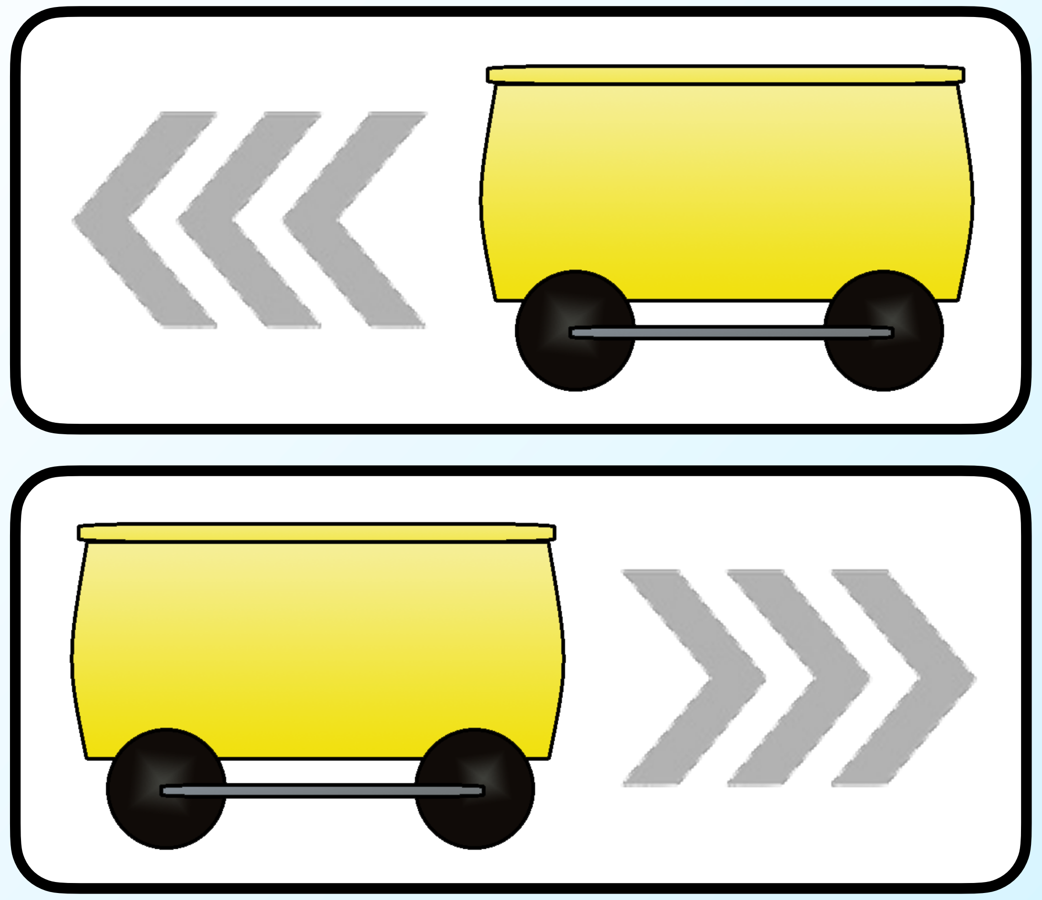
$\theta \in [-24°, 24°]$

$\dot{\theta} \in \mathbb{R}$

$x \in [-4.8, 4.8]$

$\dot{x} \in \mathbb{R}$

Action space



Transitions

- deterministic via laws of mechanics
- terminate if

$x \notin [-2.4, 2.4]$

or $\theta \notin [-12°, 12°]$

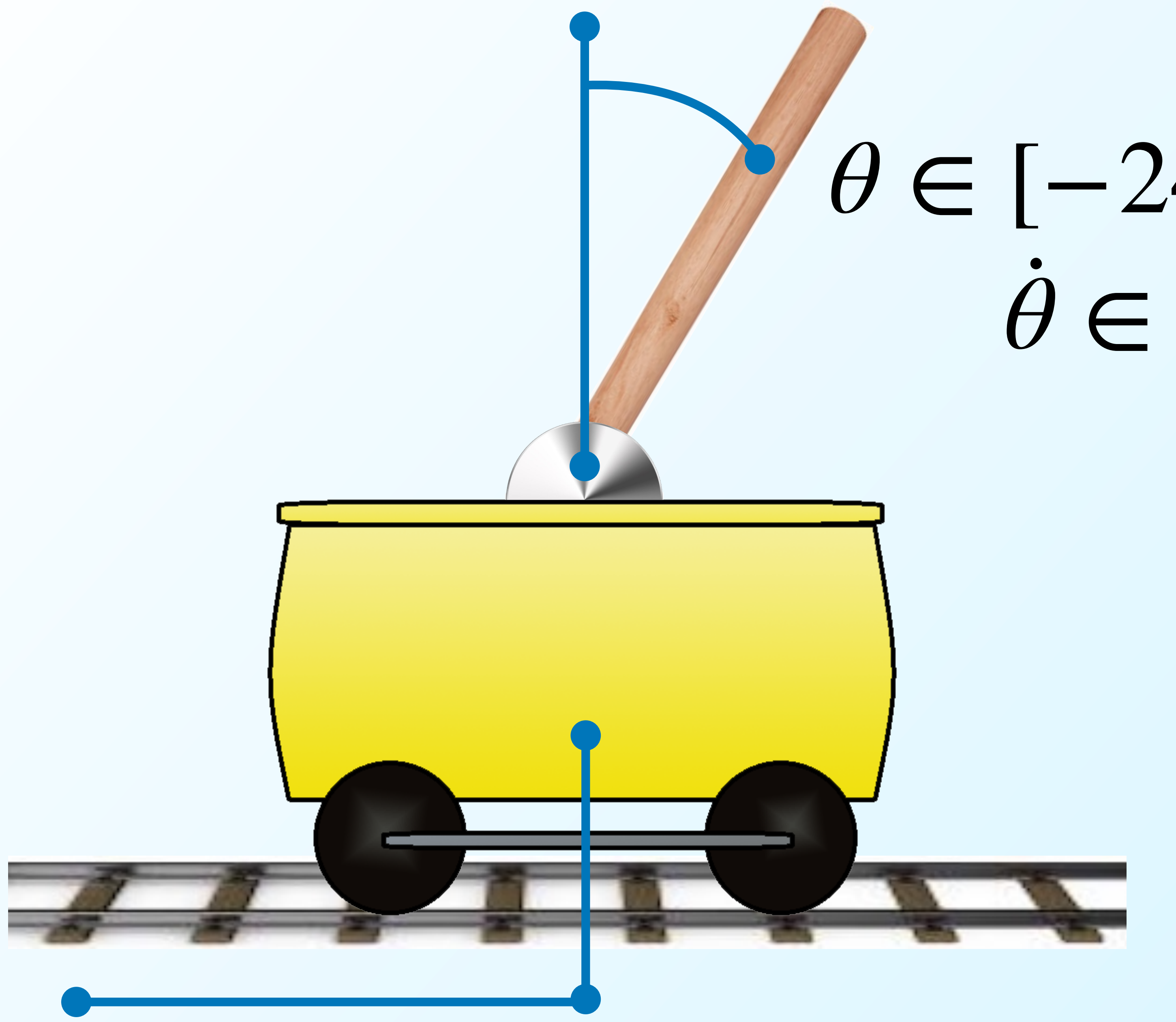
Initial state

$x, \dot{x}, \theta, \dot{\theta} \sim \mathcal{U}[-0.05, 0.05]$

Barto et al. (1983), *Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems*.

Cart Pole Example

State space



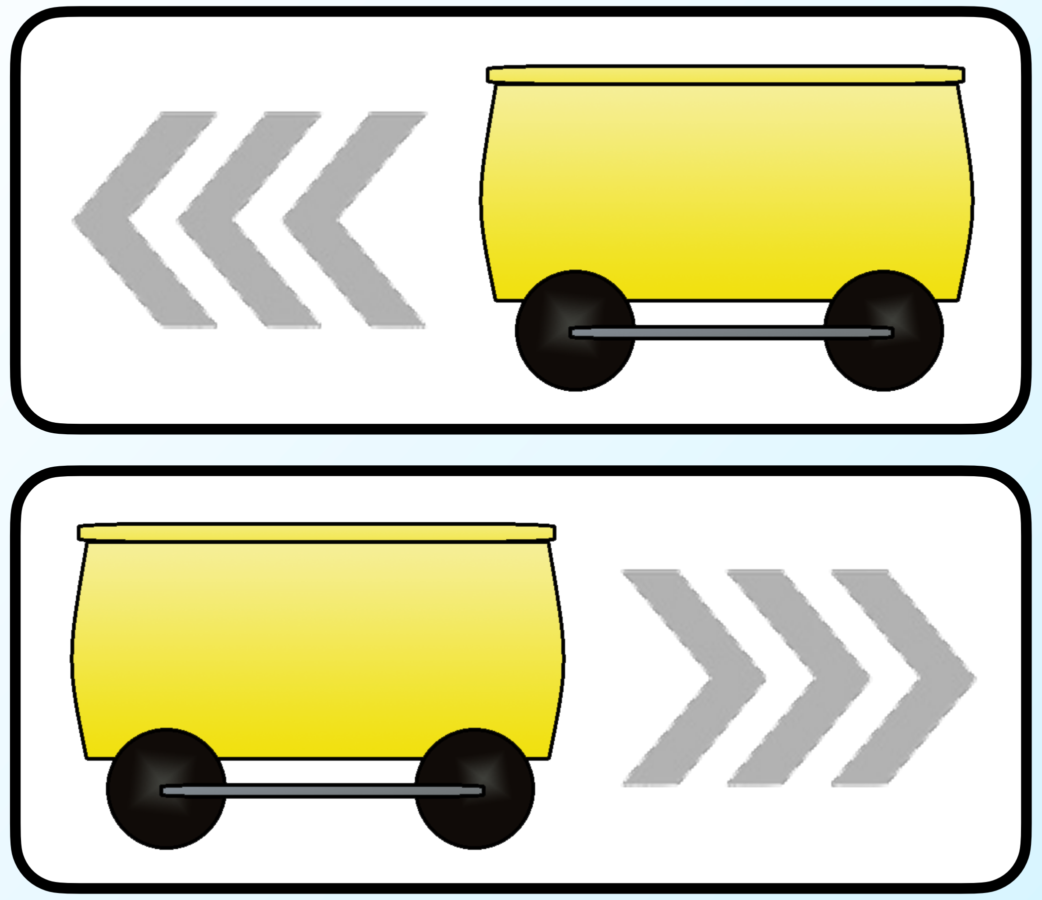
$\theta \in [-24°, 24°]$

$\dot{\theta} \in \mathbb{R}$

$x \in [-4.8, 4.8]$

$\dot{x} \in \mathbb{R}$

Action space



Transitions

- deterministic via laws of mechanics
- terminate if

$x \notin [-2.4, 2.4]$

or $\theta \notin [-12°, 12°]$

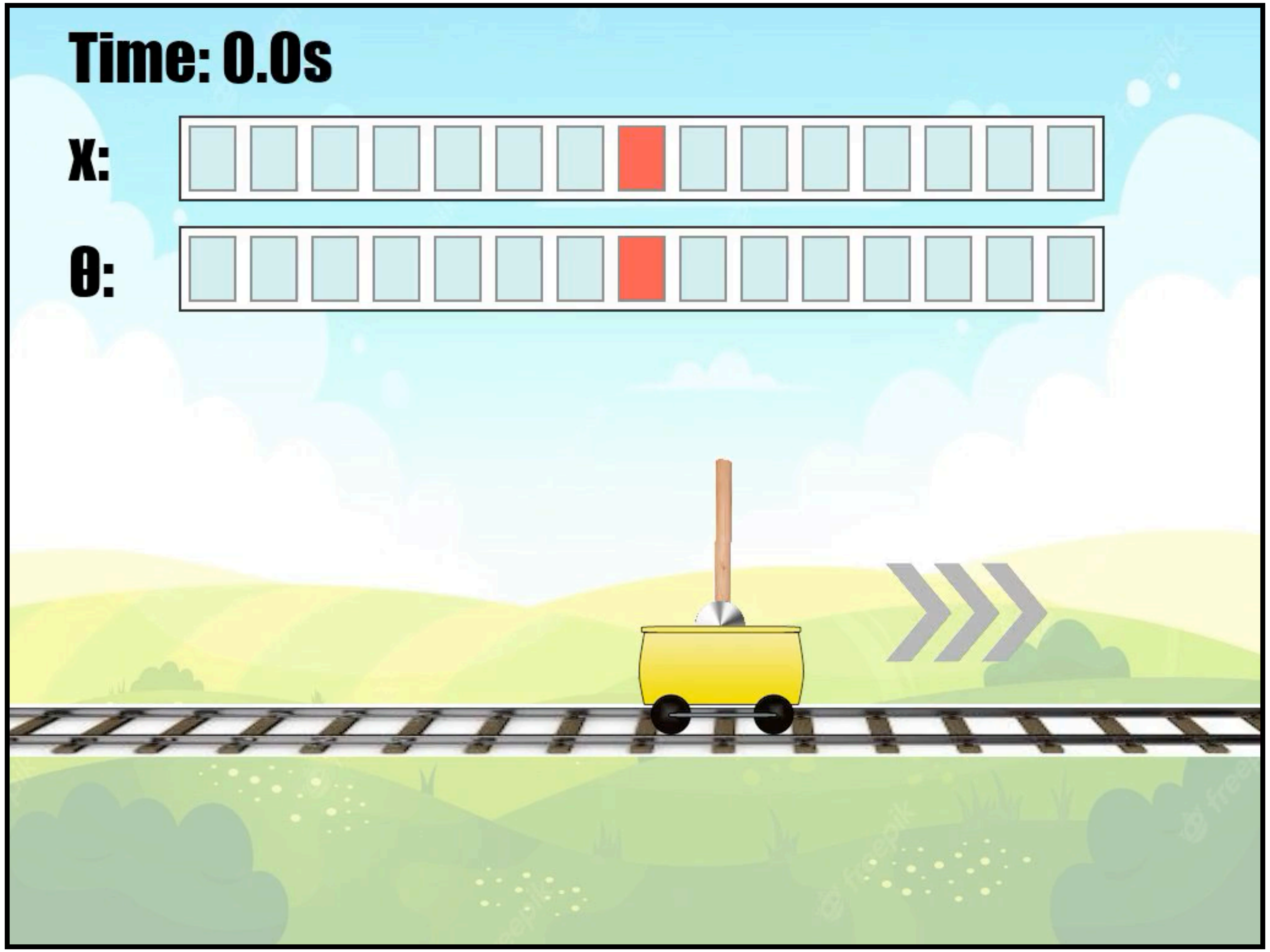
Initial state

$x, \dot{x}, \theta, \dot{\theta} \sim \mathcal{U}[-0.05, 0.05]$

Rewards

+1/non-terminated time step

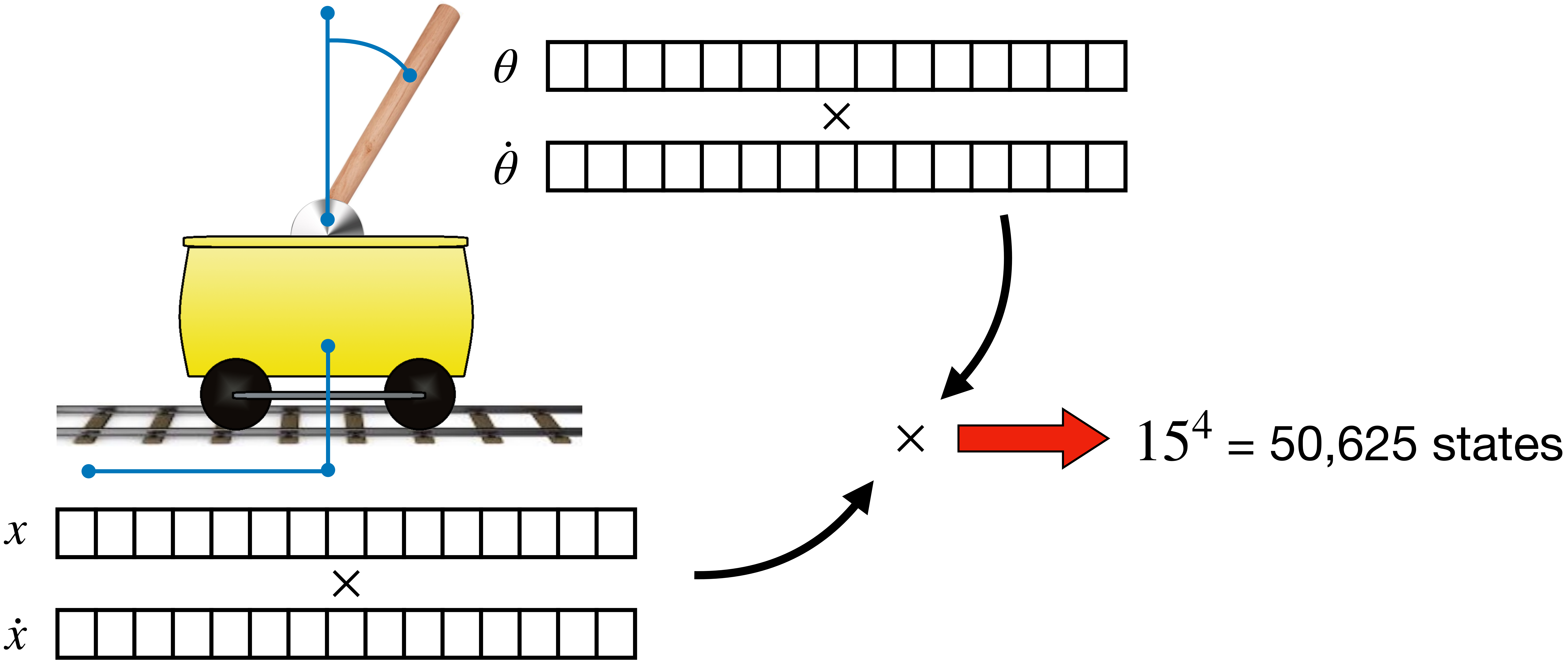
Barto et al. (1983), *Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems*.



Barto et al. (1983), *Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems*.

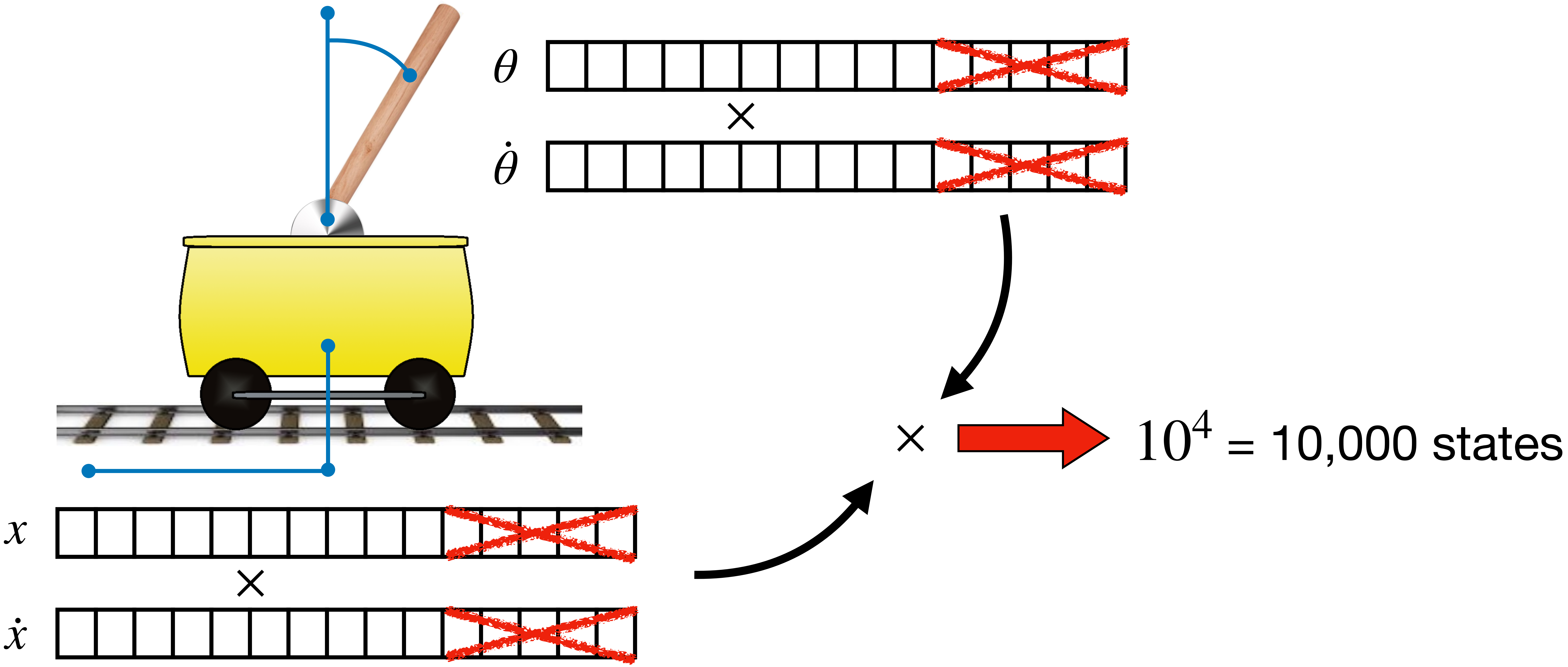
Two common sources of ambiguity:

- **Modelling errors:** 32.67 secs/run



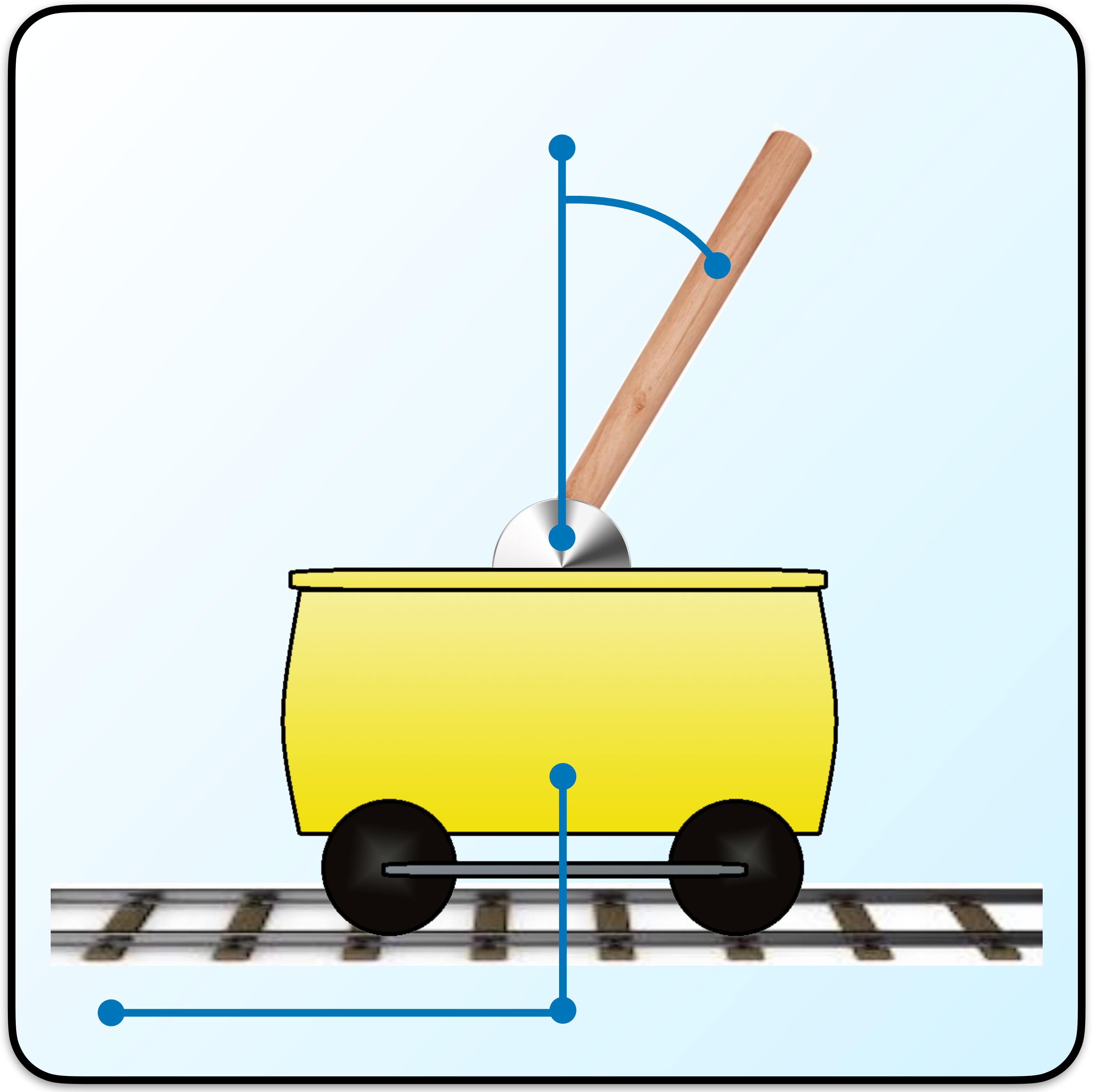
Two common sources of ambiguity:

- **Modelling errors**: 32.67 secs/run ➡ 2.45 secs/run



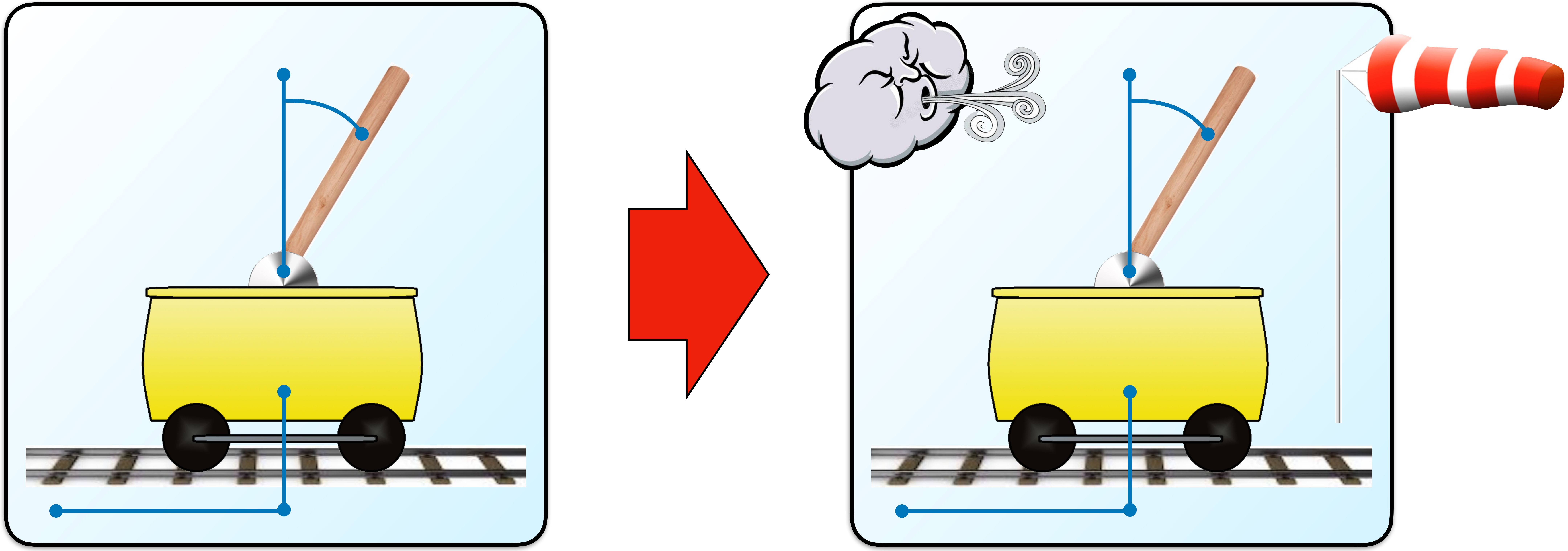
Two common sources of ambiguity:

- **Modelling errors**: 32.67 secs/run → 2.45 secs/run
- **Estimation errors**: 32.67 secs/run



Two common sources of ambiguity:

- **Modelling errors**: 32.67 secs/run → 2.45 secs/run
- **Estimation errors**: 32.67 secs/run → 4.68 secs/run



Two common sources of ambiguity:

- **Modelling errors**: 32.67 secs/run → 2.45 secs/run
- **Estimation errors**: 32.67 secs/run → 4.68 secs/run

Impact of ambiguity can be alleviated via robust optimization:

Robust MDP

$$\underset{\pi \in \Pi}{\text{maximize}} \quad \inf_{p \in \mathscr{P}} \mathbb{E}_p \left[\sum_{t=1}^{\infty} \lambda^{t-1} \cdot r\big(s_t, \pi[s_t]\big) \right]$$

ambiguity set

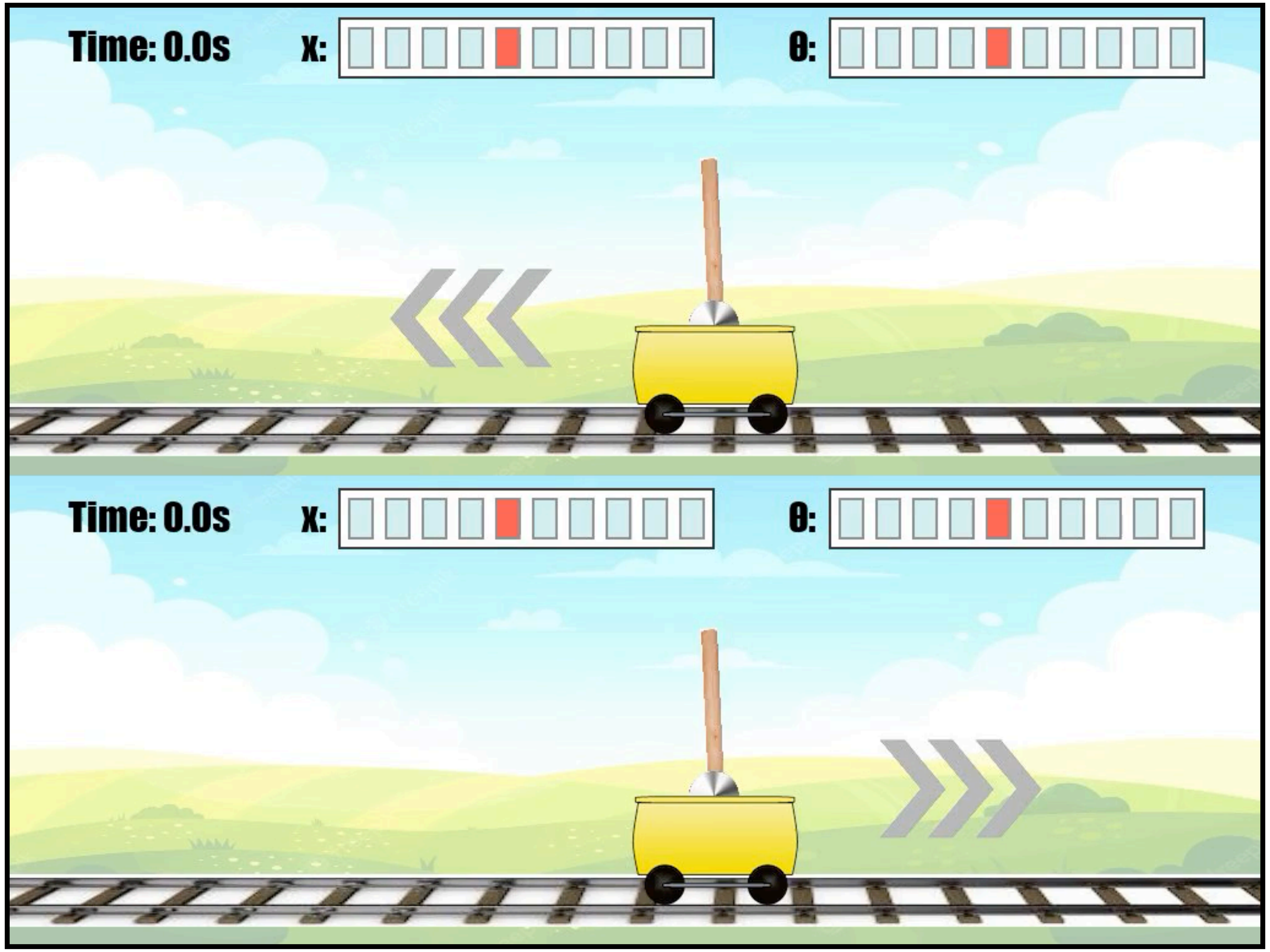
Robust MDPs admit interpretation as regularized MDPs!

Derman et al. (2023), *Twice Regularized Markov Decision Processes: The Equivalence between Robustness and Regularization*.

Ambiguity: Modelling Errors

nominal

robust

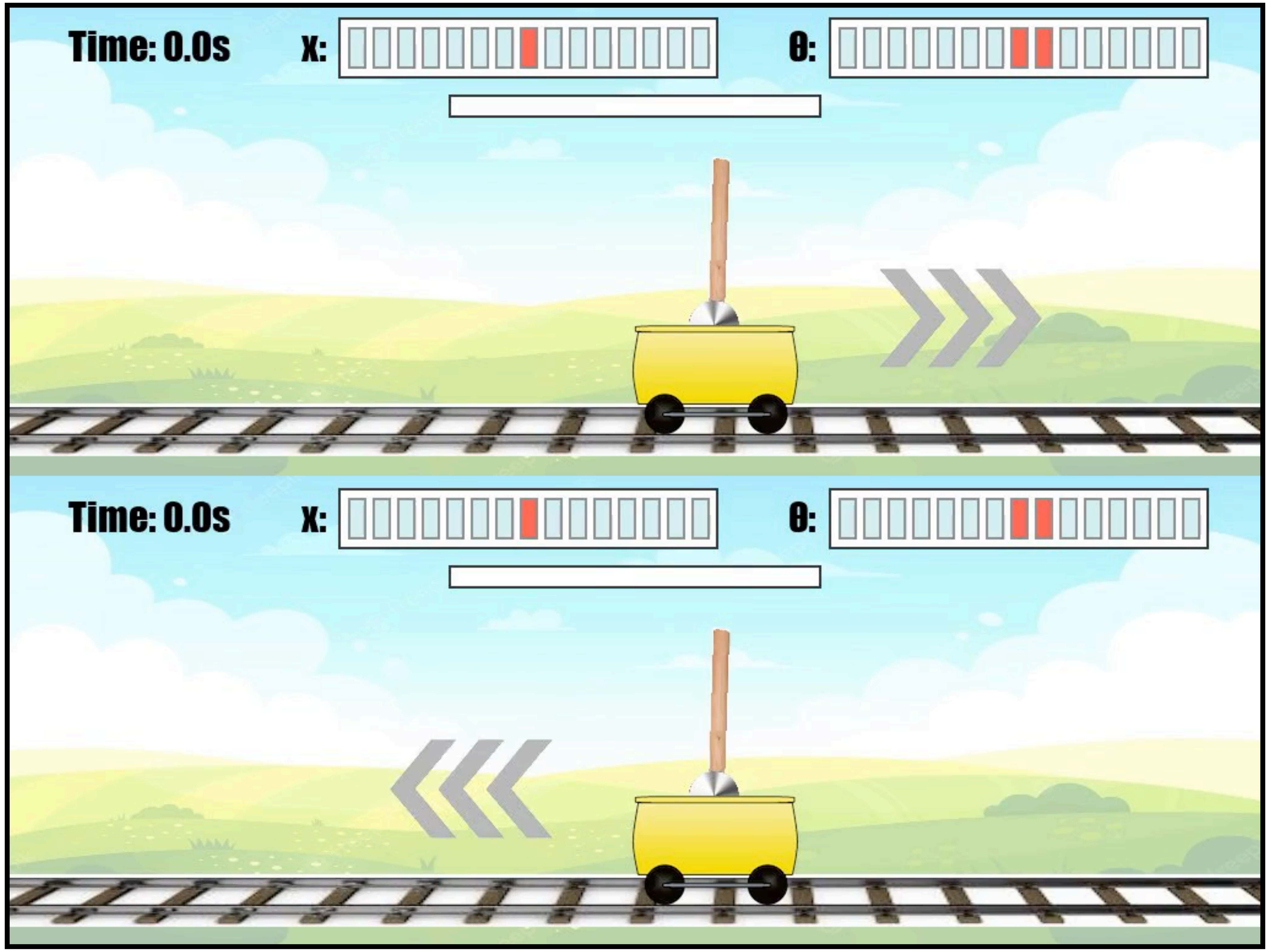


Modelling errors: 32.67 secs/run → 2.45 secs/run → 15.77 secs/run

Ambiguity: Estimation Errors

nominal

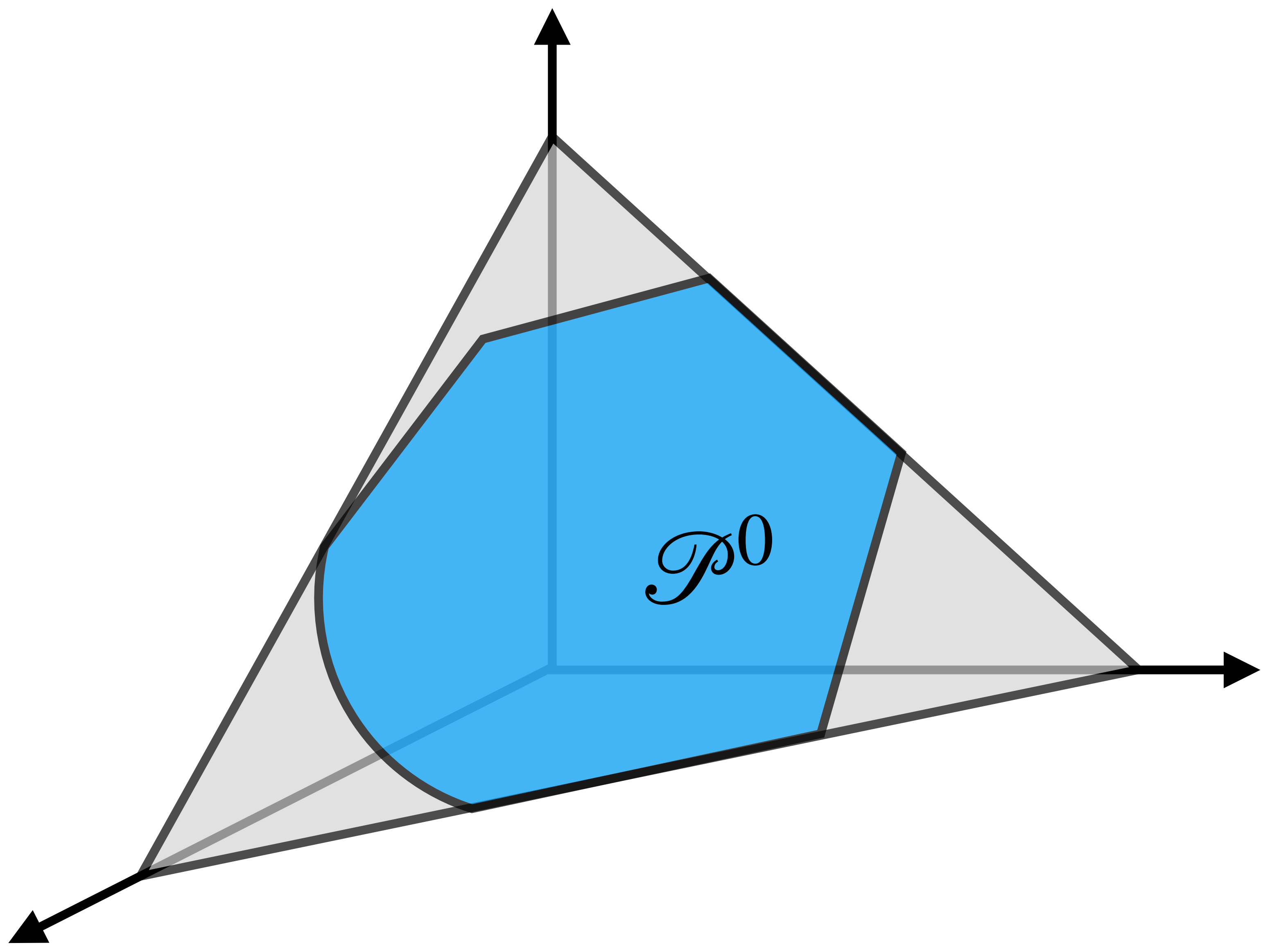
robust



Estimation errors: 32.67 secs/run → 4.68 secs/run → 15.76 secs/run

Ambiguity Sets

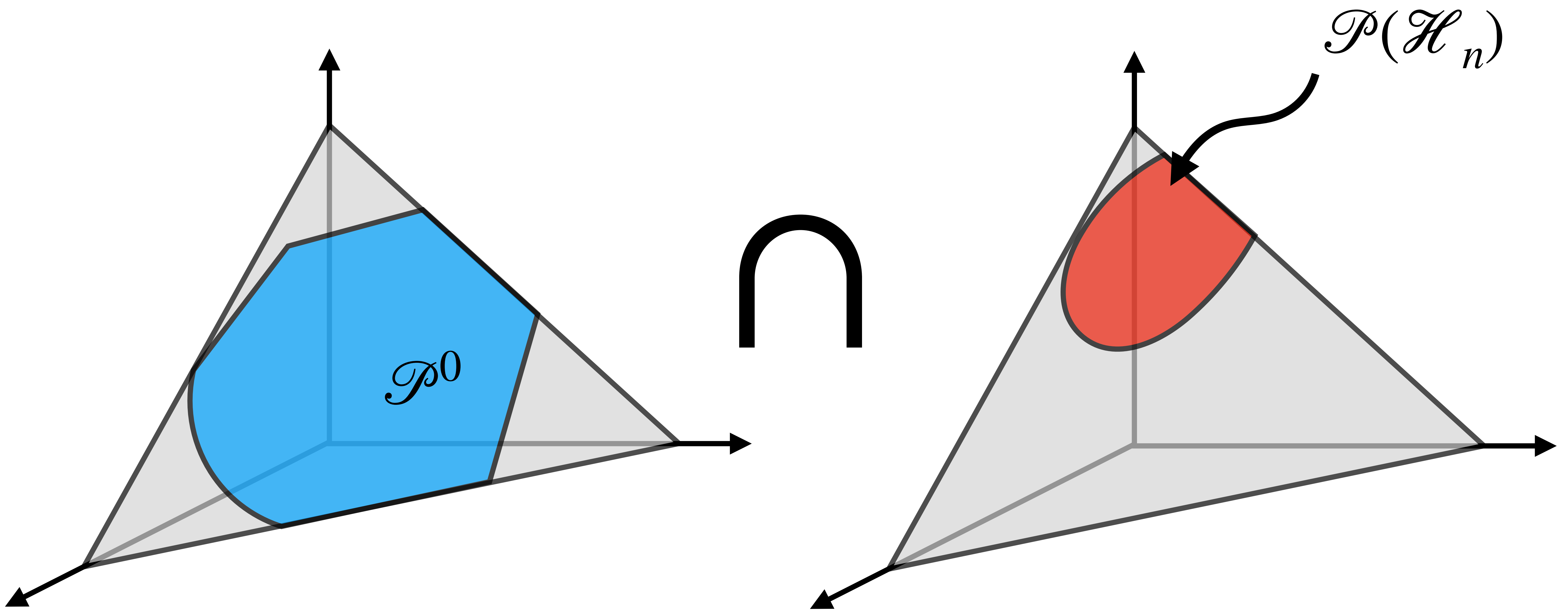
Structural ambiguity set



Wiesemann et al. (2013), *Robust Markov Decision Processes*.

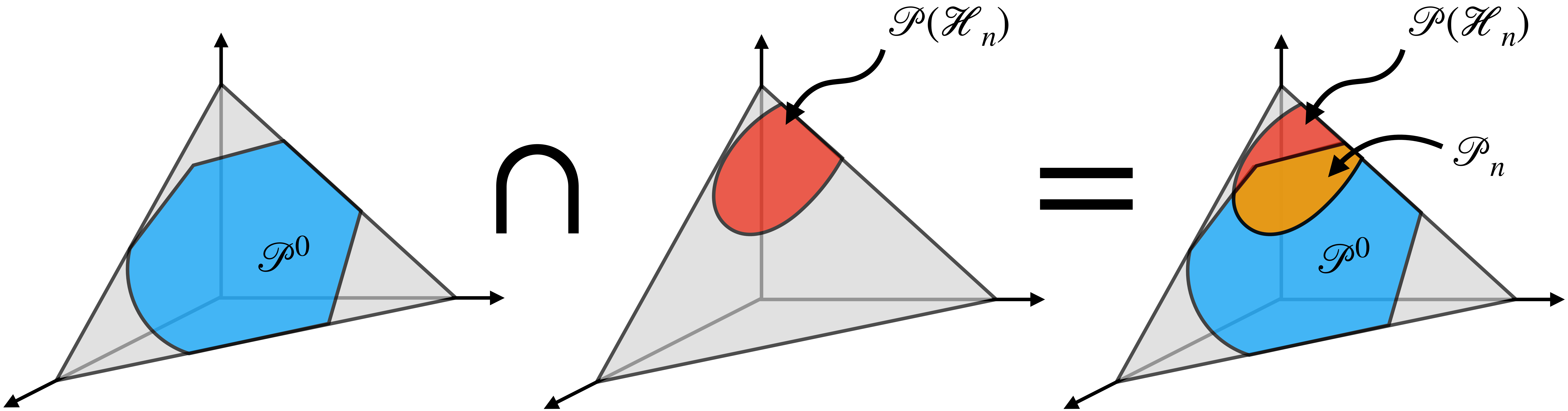
Ambiguity Sets

Structural ambiguity set **Historical sample**



Wiesemann et al. (2013), *Robust Markov Decision Processes*.

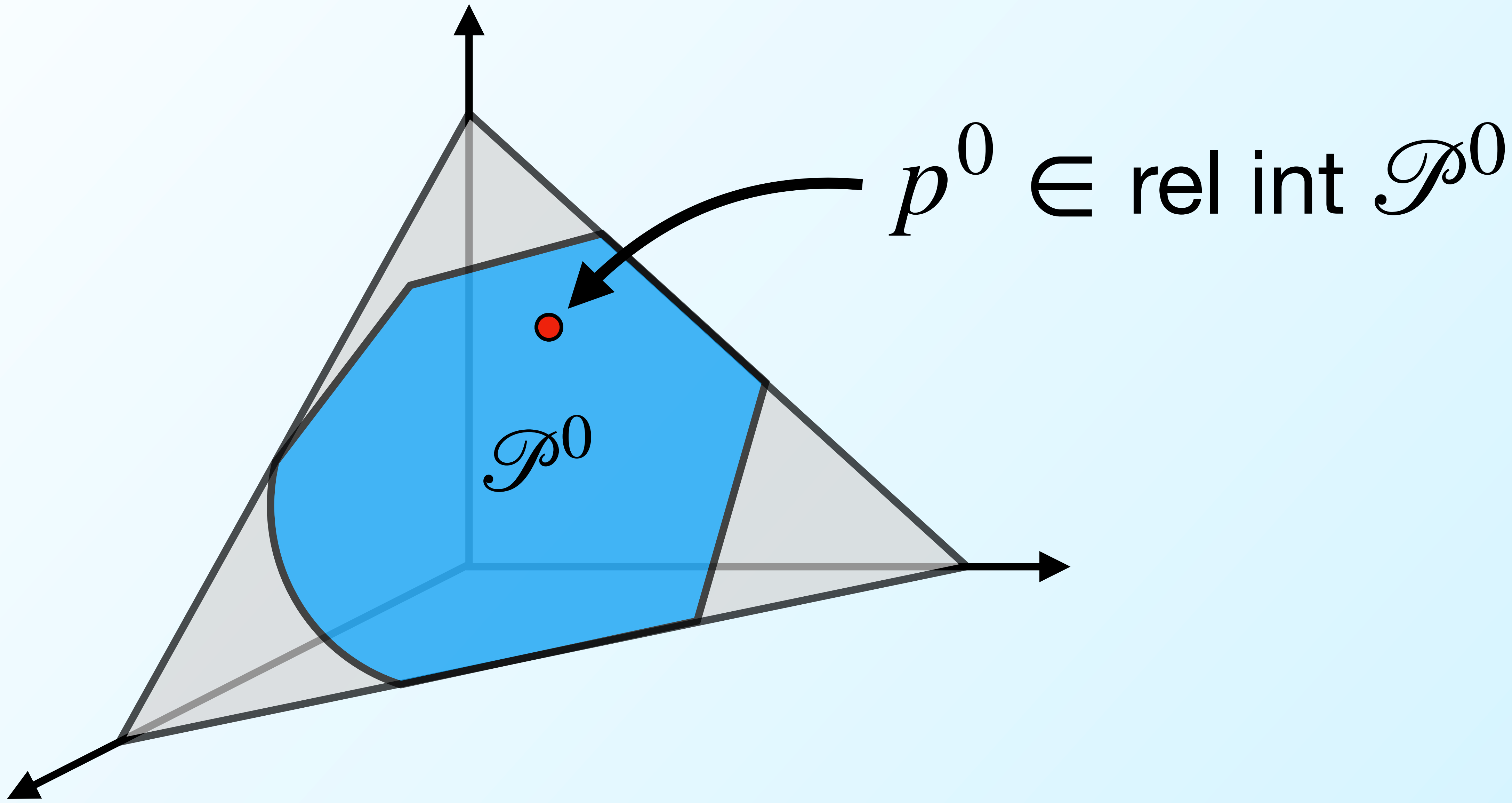
Structural ambiguity set **Historical sample** **Out-of-sample guarantee**



Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Structural ambiguity set

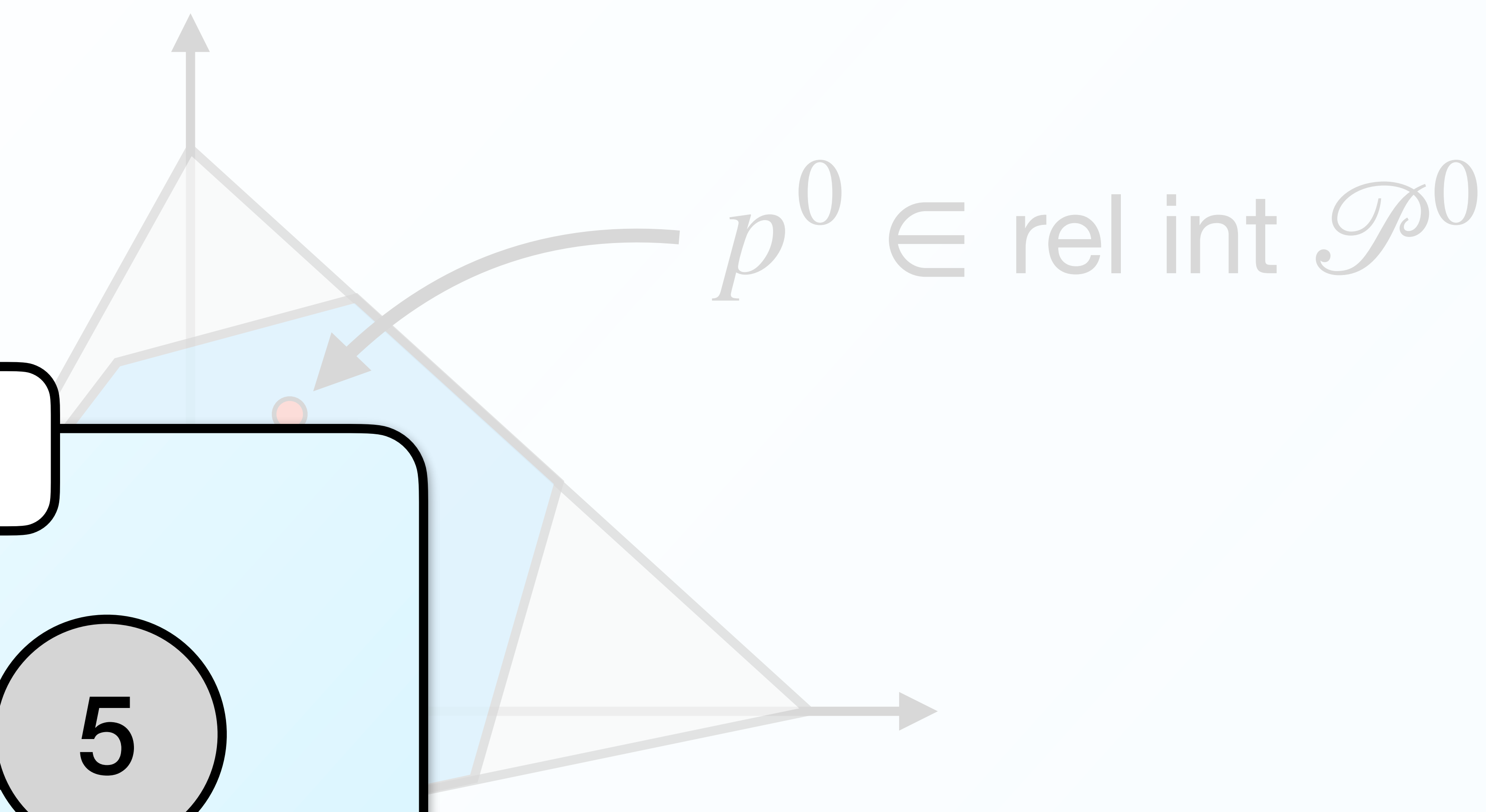
$$\mathscr{P}^0 \subseteq \{p : \mathscr{S} \times \mathscr{A} \to \Delta(\mathscr{S})\}$$



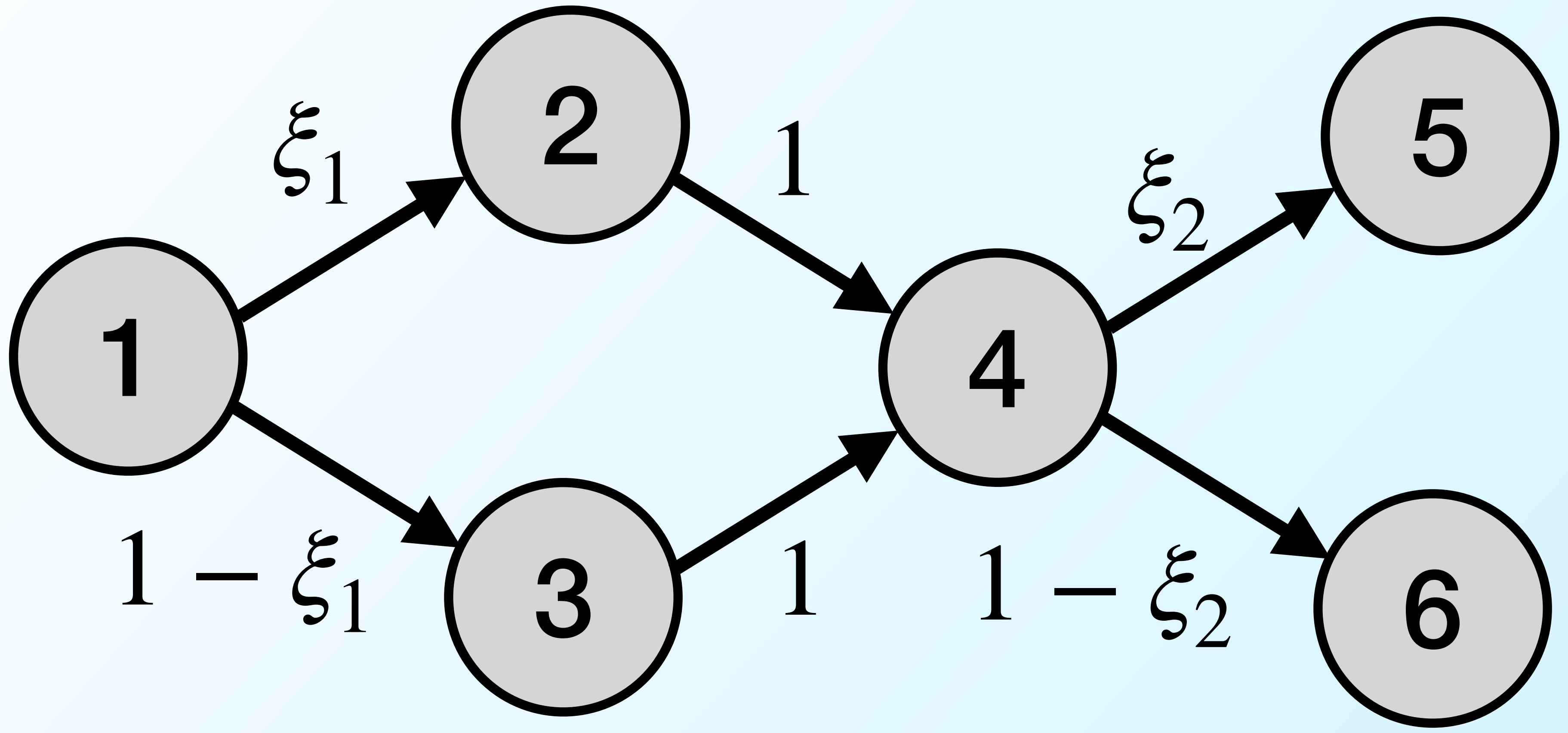
Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Structural ambiguity set

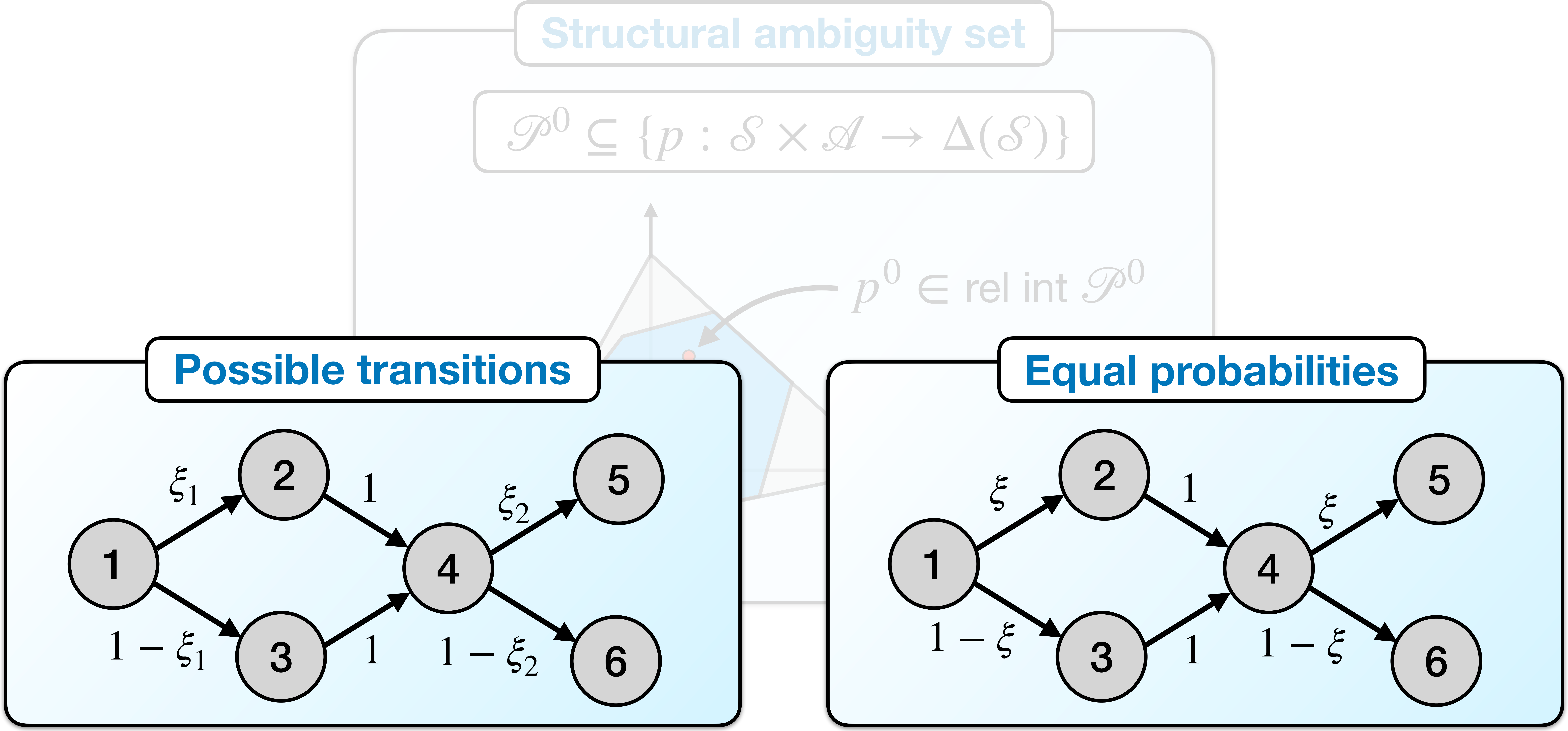
$$\mathscr{P}^0 \subseteq \{p : \mathscr{S} \times \mathscr{A} \to \Delta(\mathscr{S})\}$$



Possible transitions



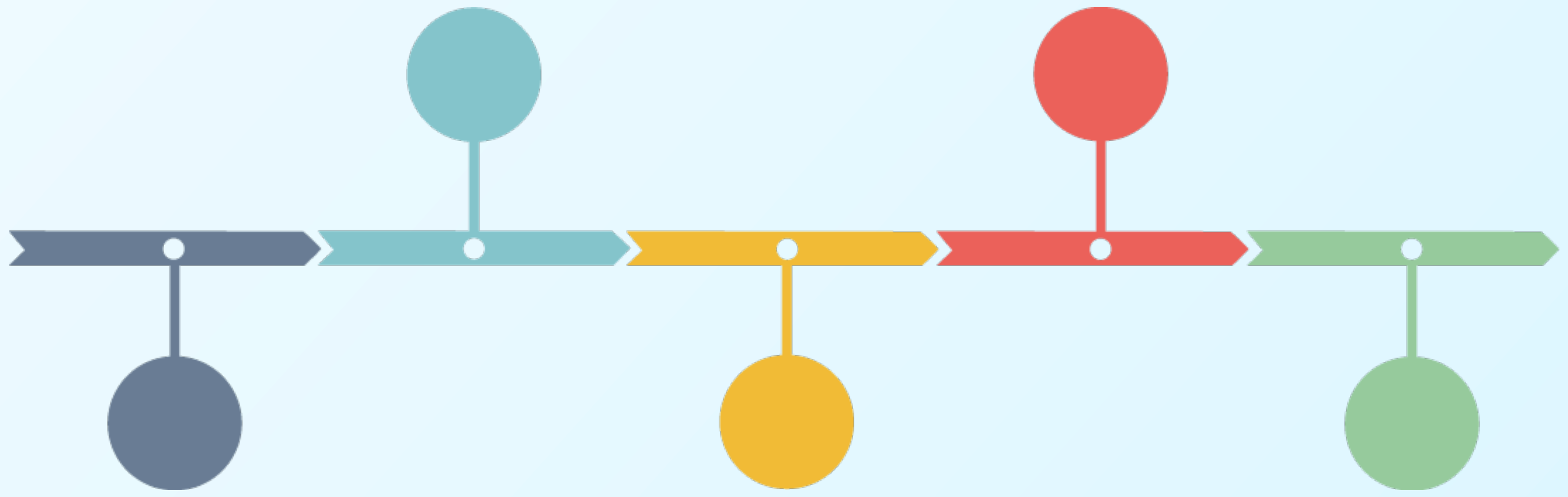
Wiesemann et al. (2013), *Robust Markov Decision Processes*.



Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Historical sample

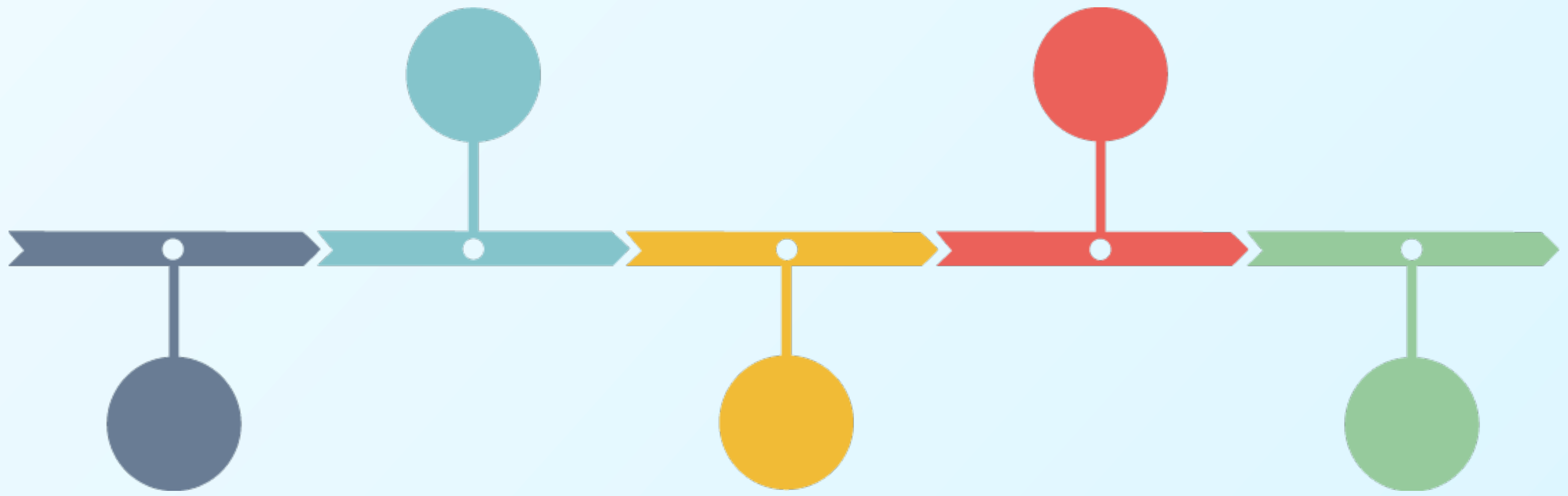
historical policy π^0
(stationary, randomized)



state-action history
$$\mathscr{H}_n = (s_1, a_1, \ldots, s_n, a_n) \in (\mathscr{S} \times \mathscr{A})^n$$

Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Historical sample



historical policy π^0
(stationary, randomized)

state-action history
$\mathscr{H}_n = (s_1, a_1, \ldots, s_n, a_n) \in (\mathscr{S} \times \mathscr{A})^n$

Likelihood, given history

$$\mathscr{L}_n(p) \;=\; q(s_1) \cdot \pi^0(a_n \mid s_n) \cdot \prod_{t=1}^{n-1} \left[\pi^0(a_t \mid s_t) \cdot p(s_{t+1} \mid s_t, a_t)\right]$$

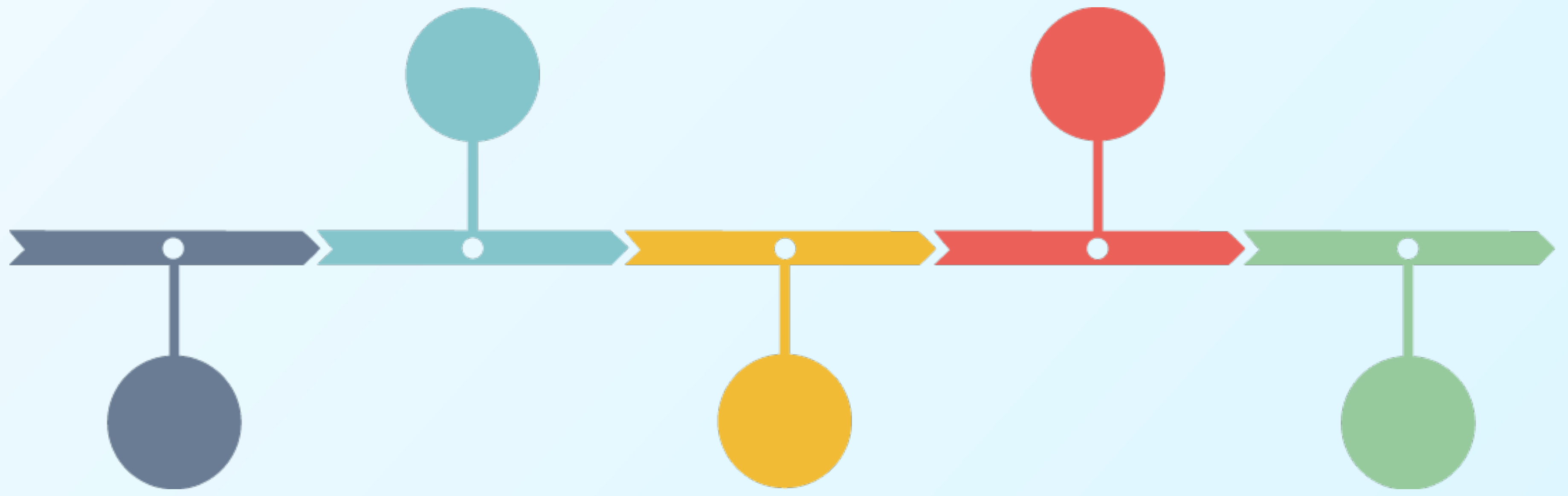
Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Ambiguity Sets

Historical sample

$$\mathscr{P}(\mathscr{H}_n) = \left\{p : \log \mathscr{L}_n(p) \geq \log \mathscr{L}_n(p^\star) - \delta\right\}$$

historical policy π^0
(stationary, randomized)



state-action history

$$\mathscr{H}_n = (s_1, a_1, \ldots, s_n, a_n) \in (\mathscr{S} \times \mathscr{A})^n$$

Likelihood, given history

$$\mathscr{L}_n(p) \;=\; q(s_1) \cdot \pi^0(a_n \,|\, s_n) \cdot \prod_{t=1}^{n-1} \left[\pi^0(a_t \,|\, s_t) \cdot p(s_{t+1} \,|\, s_t, a_t)\right]$$

Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Theorem

Assumption: Historical policy π^0 visits every $s \in \mathcal{S}$ infinitely often as $n \longrightarrow \infty$

Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Theorem

$\mathscr{P}_n = \mathscr{P}^0 \cap \mathscr{P}(\mathscr{H}_n)$ with $\delta = (1-\beta)$-quantile of χ^2-distribution with κ degrees of freedom

Assumption: Historical policy π^0 visits every $s \in \mathscr{S}$ infinitely often as $n \longrightarrow \infty$

Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Theorem

$\mathscr{P}_n = \mathscr{P}^0 \cap \mathscr{P}(\mathscr{H}_n)$ with $\delta = (1-\beta)$-quantile of χ^2-distribution with κ degrees of freedom

Assumption: Historical policy π^0 visits every $s \in \mathscr{S}$ infinitely often as $n \longrightarrow \infty$

1

$$\lim_{n \longrightarrow \infty} \mathbb{P}\left[p^0 \in \mathscr{P}_n\right] = 1 - \beta$$

Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Theorem

$\mathscr{P}_n = \mathscr{P}^0 \cap \mathscr{P}(\mathscr{H}_n)$ with $\delta = (1-\beta)$-quantile of χ^2-distribution with κ degrees of freedom

Assumption: Historical policy π^0 visits every (s, a) infinitely often as $n \longrightarrow \infty$

1

$$\lim_{n \longrightarrow \infty} \mathbb{P}\left[p^0 \in \mathscr{P}_n\right] = 1 - \beta$$

Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Theorem

$\mathscr{P}_n = \mathscr{P}^0 \cap \mathscr{P}(\mathscr{H}_n)$ with $\delta = (1-\beta)$-quantile of χ^2-distribution with κ degrees of freedom

Assumption: Historical policy π^0 visits every (s, a) infinitely often as $n \longrightarrow \infty$

1 $\lim_{n \longrightarrow \infty} \mathbb{P}\left[p^0 \in \mathscr{P}_n\right] = 1 - \beta$

2 $\operatorname{plim}_{n \longrightarrow \infty} \left[\sqrt{n} \cdot d^{\mathsf{H}}(\mathscr{P}_n, \{p^0\})\right] = 0$

Wiesemann et al. (2013), *Robust Markov Decision Processes*.

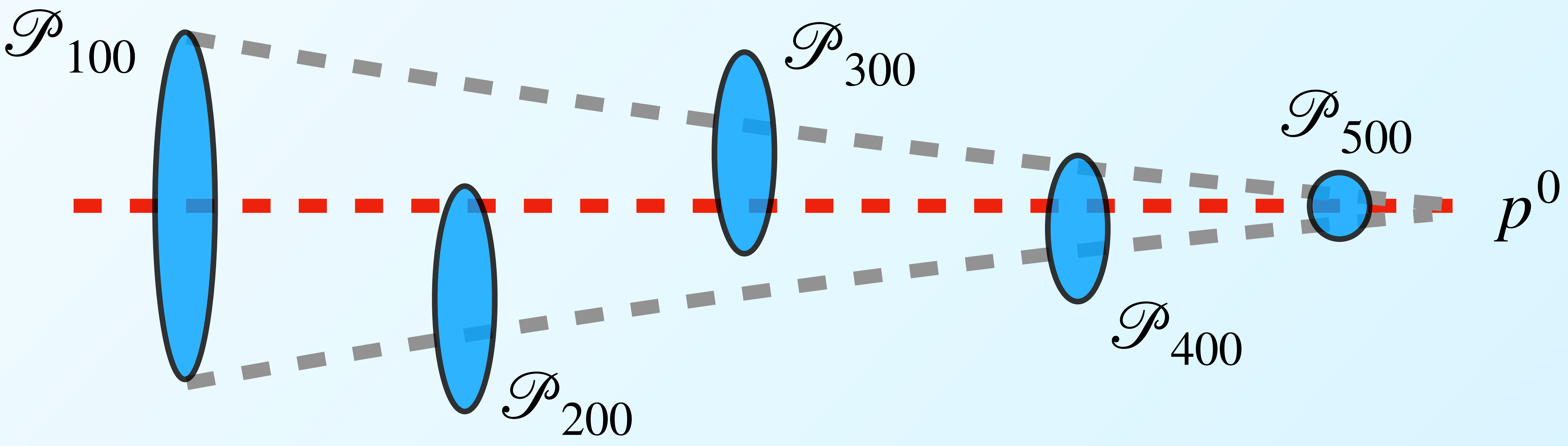
Theorem

$\mathscr{P}_n = \mathscr{P}^0 \cap \mathscr{P}(\mathscr{H}_n)$ with $\delta = (1-\beta)$-quantile of χ^2-distribution with κ degrees of freedom

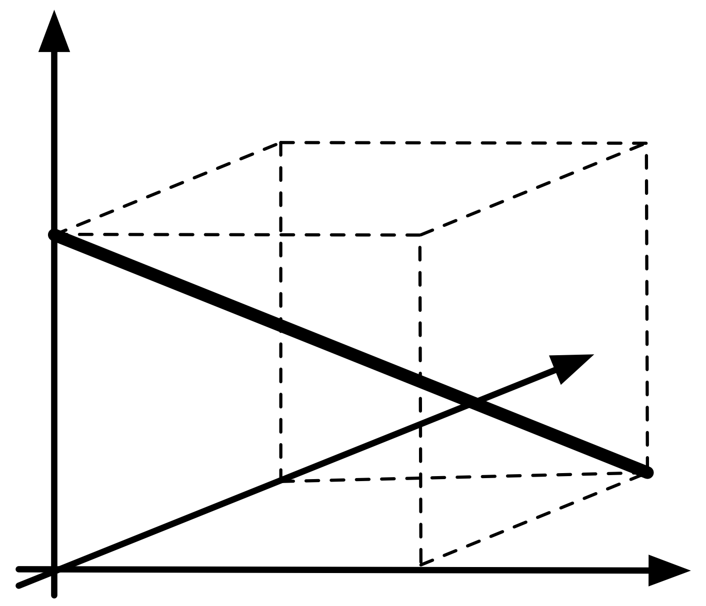
Assumption: Historical policy π^0 visits every (s,a) infinitely often as $n \longrightarrow \infty$

1. $\lim_{n\longrightarrow\infty} \mathbb{P}\left[p^0 \in \mathscr{P}_n\right] = 1-\beta$

2. $\operatorname{plim}_{n\longrightarrow\infty} \left[\sqrt{n}\cdot d^{\mathsf{H}}(\mathscr{P}_n, \{p^0\})\right] = 0$



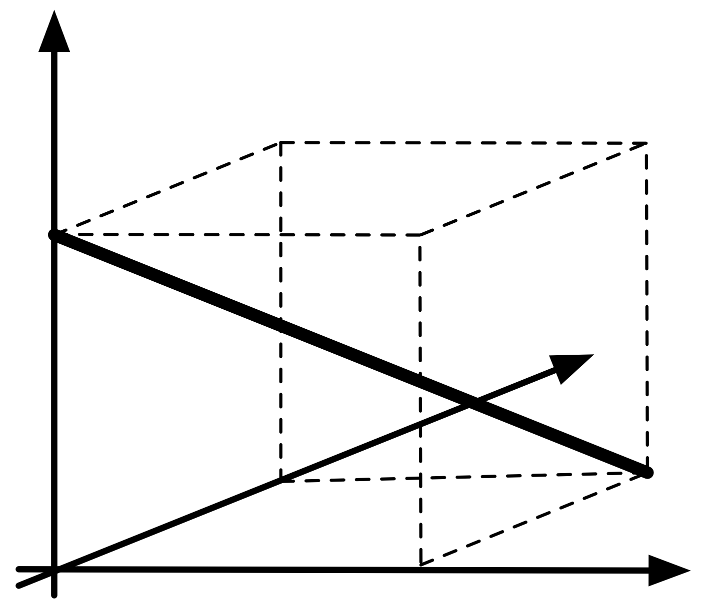
Wiesemann et al. (2013), *Robust Markov Decision Processes*.



General (non-rectangular) ambiguity sets

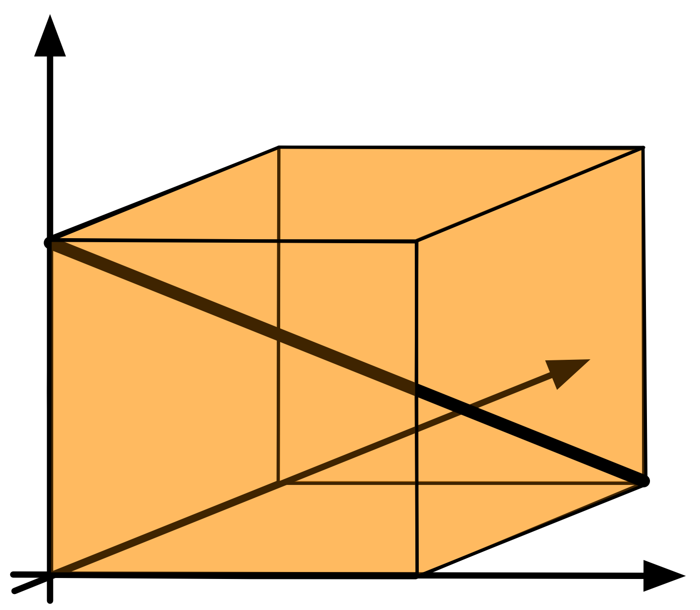
- Optimal policy can be randomized & history-dependent
- Bellman optimality principle violated; NP-hard

Rectangularity



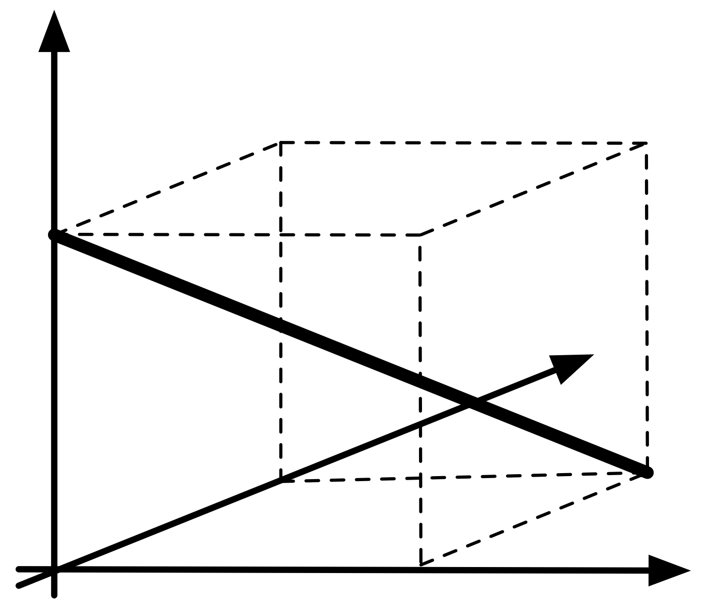
General (non-rectangular) ambiguity sets

- Optimal policy can be randomized & history-dependent
- Bellman optimality principle violated; NP-hard



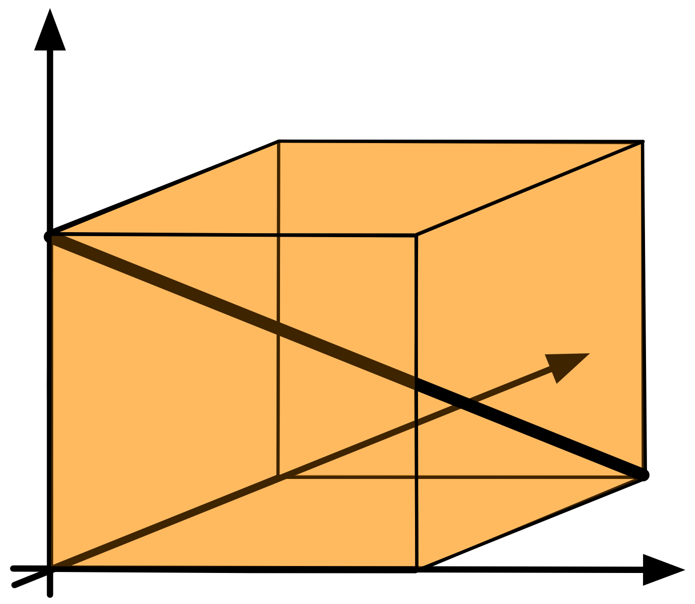
(*s*,*a*)-rectangular ambiguity sets

$$\mathcal{P} = \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \mathcal{P}_{s,a} \quad \text{with} \quad \mathcal{P}_{s,a} \subseteq \Delta(\mathcal{S})$$



General (non-rectangular) ambiguity sets

- Optimal policy can be randomized & history-dependent
- Bellman optimality principle violated; NP-hard



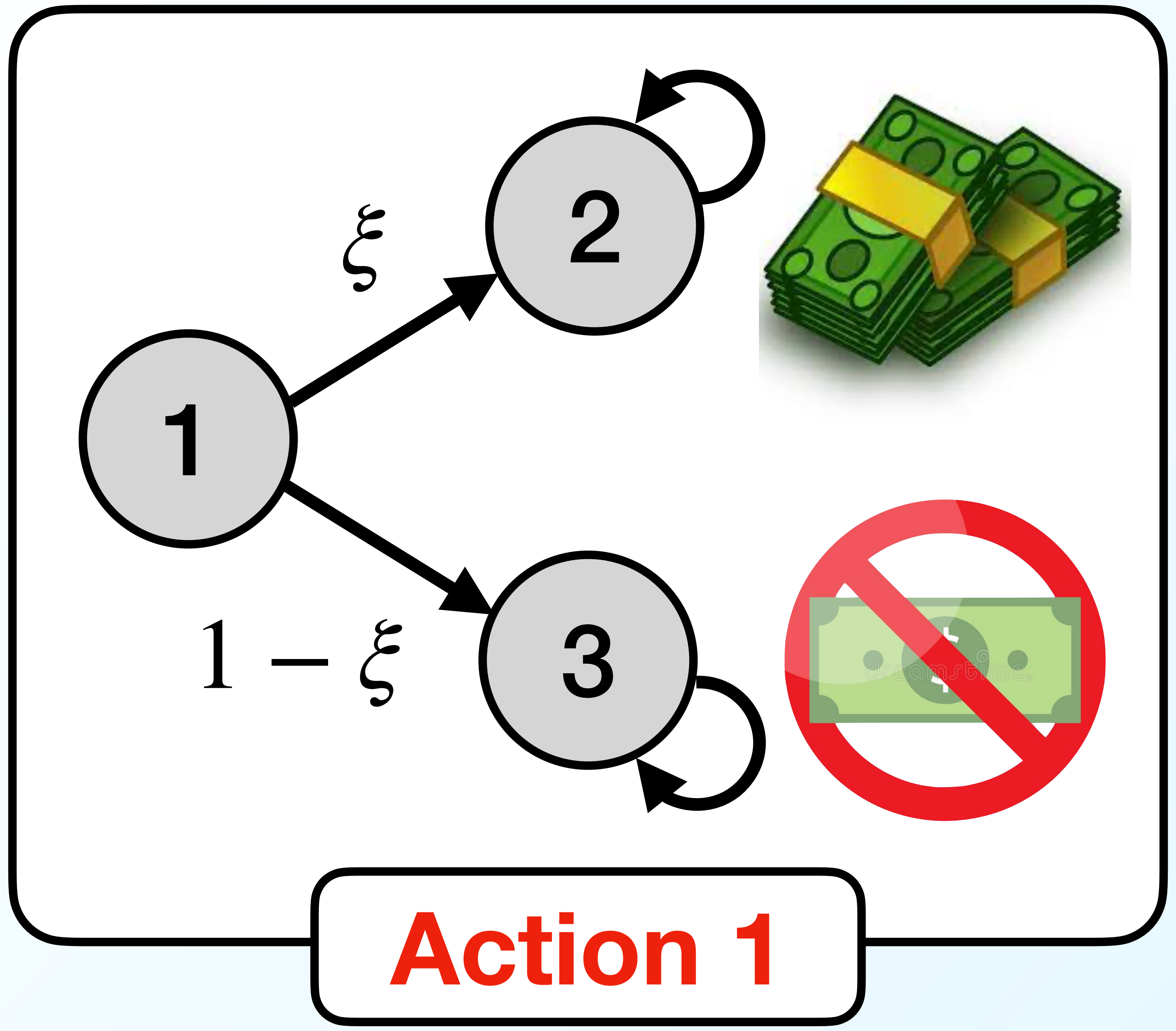
(*s*,*a*)-rectangular ambiguity sets

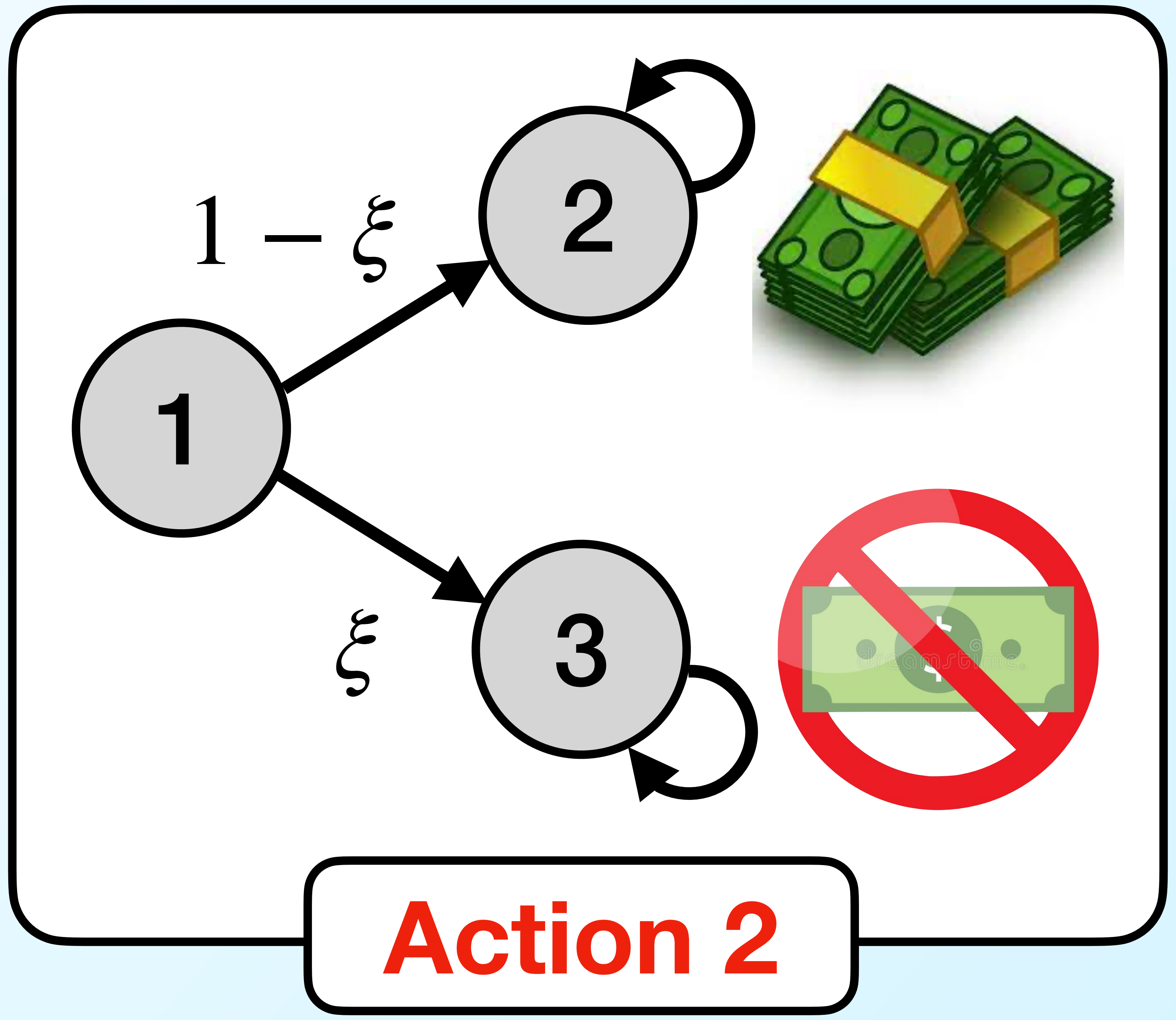
- Optimal policy stationary and deterministic
- Bellman optimality principle holds

General (non-rectangular) ambiguity sets

endent

Example





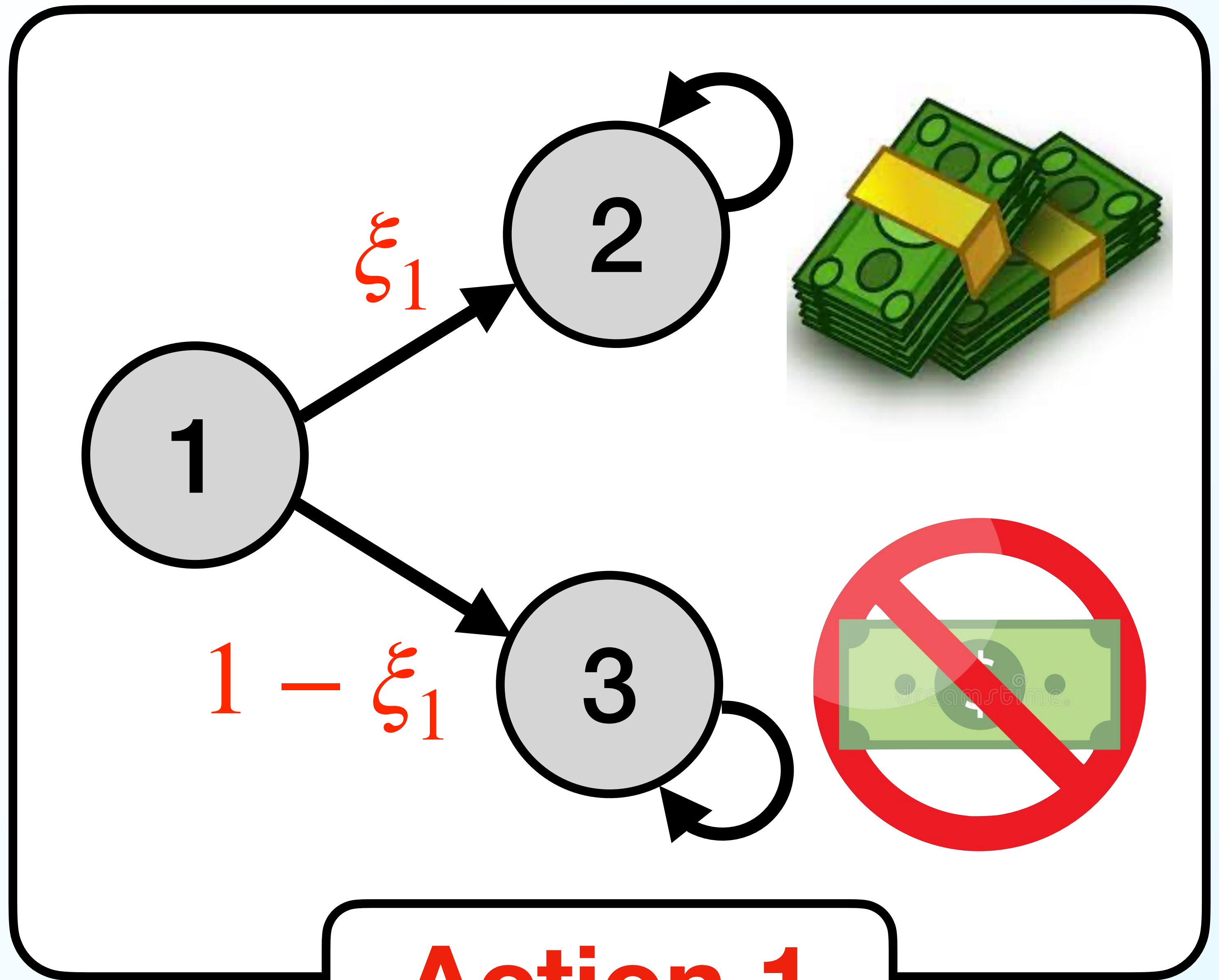
for some unknown $\xi \in [0,1]$

Bellman optimality principle holds

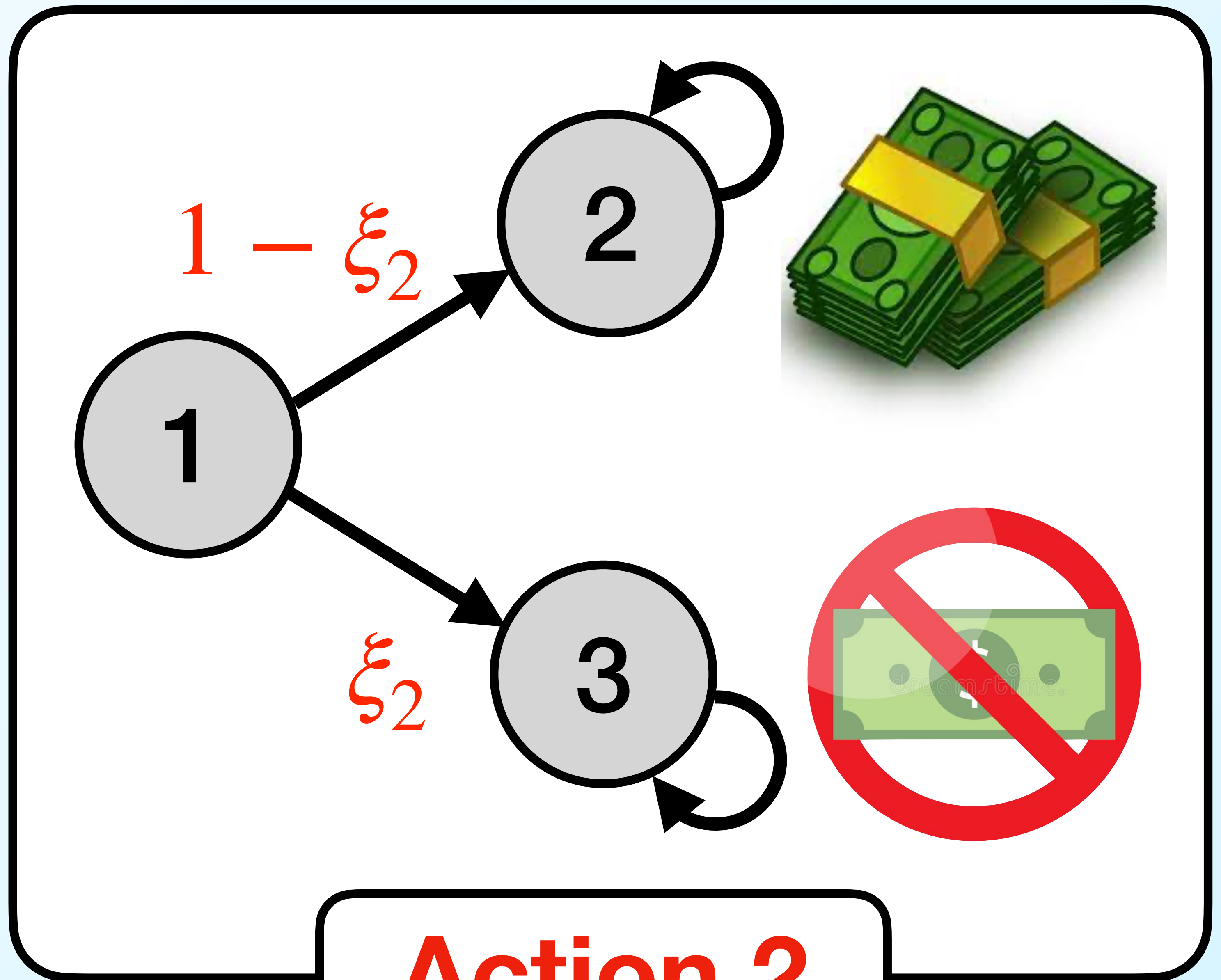
General (non-rectangular) ambiguity sets

endent

Example



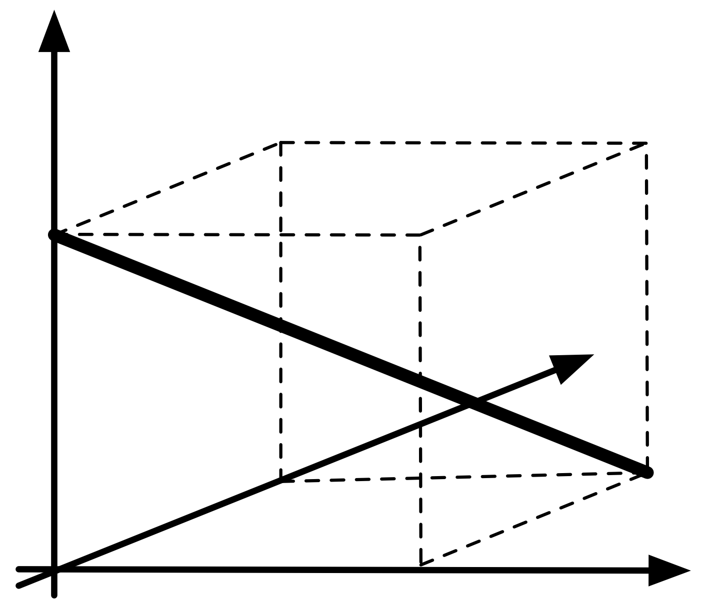
Action 1



Action 2

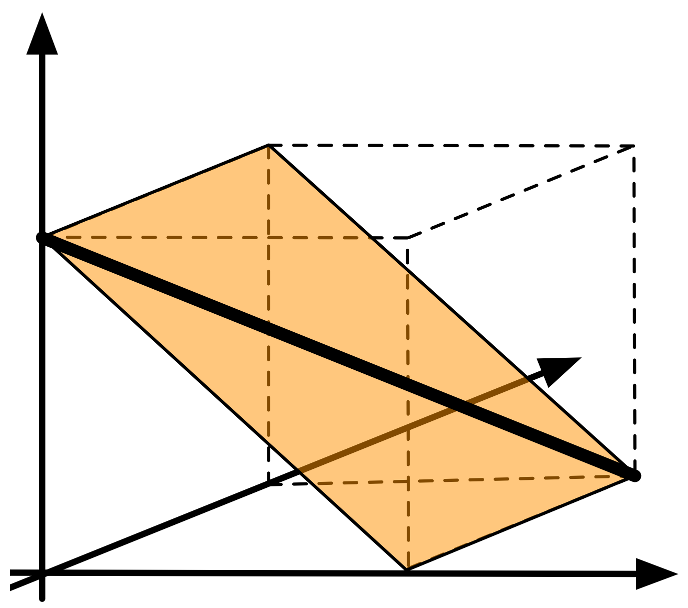
for some unknown $\xi_1, \xi_2 \in [0,1]$

Bellman optimality principle holds



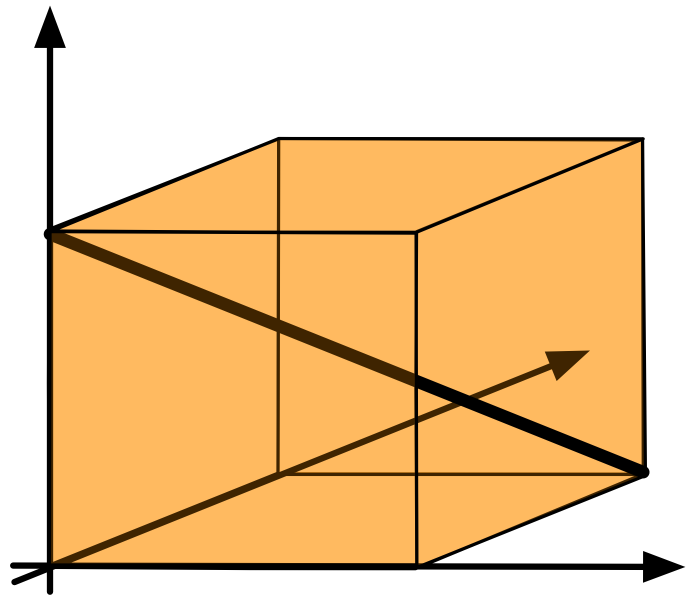
General (non-rectangular) ambiguity sets

- Optimal policy can be randomized & history-dependent
- Bellman optimality principle violated; NP-hard



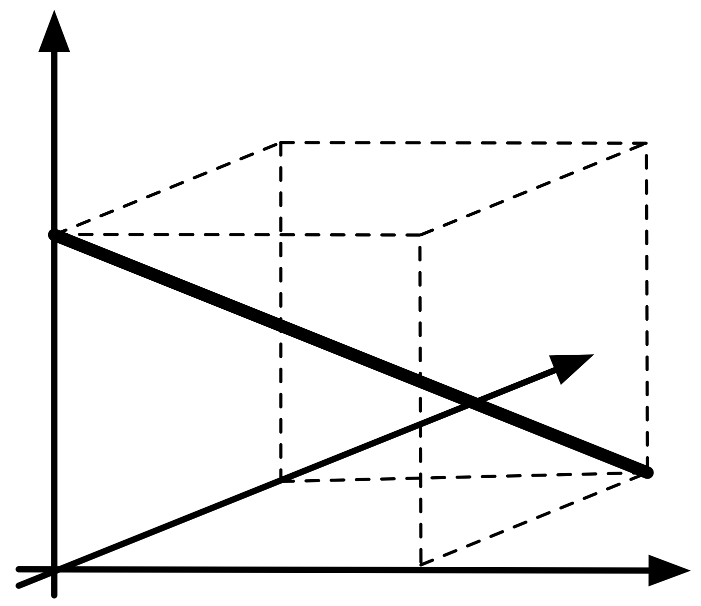
s-rectangular ambiguity sets

$$\mathscr{P} = \prod_{s \in \mathscr{S}} \mathscr{P}_s \quad \text{with} \quad \mathscr{P}_s \subseteq [\Delta(\mathscr{S})]^A$$



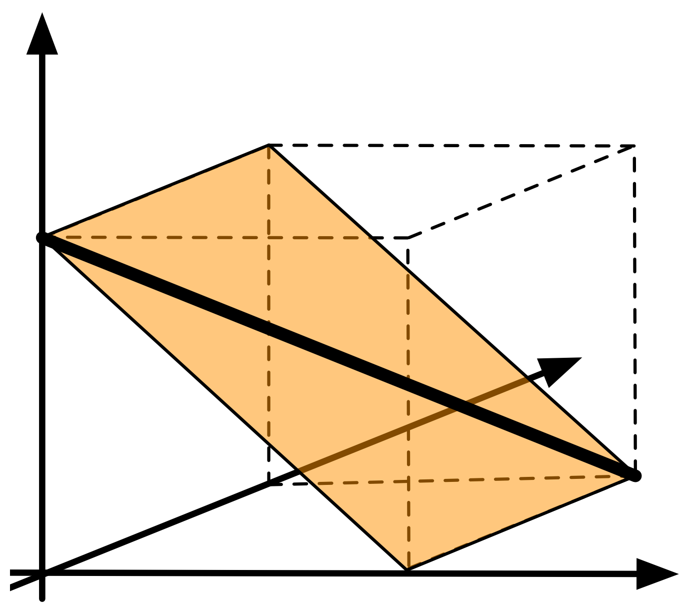
(*s*,*a*)-rectangular ambiguity sets

- Optimal policy stationary and deterministic
- Bellman optimality principle holds



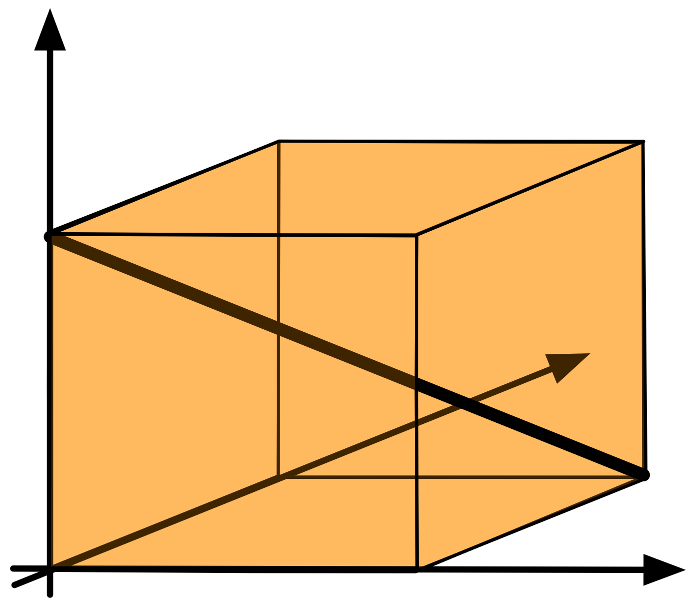
General (non-rectangular) ambiguity sets

- 👎 Optimal policy can be randomized & history-dependent
- 👎 Bellman optimality principle violated; NP-hard



s-rectangular ambiguity sets

- 👍 Optimal policy stationary but can be *randomized*
- 👍 Bellman optimality principle holds

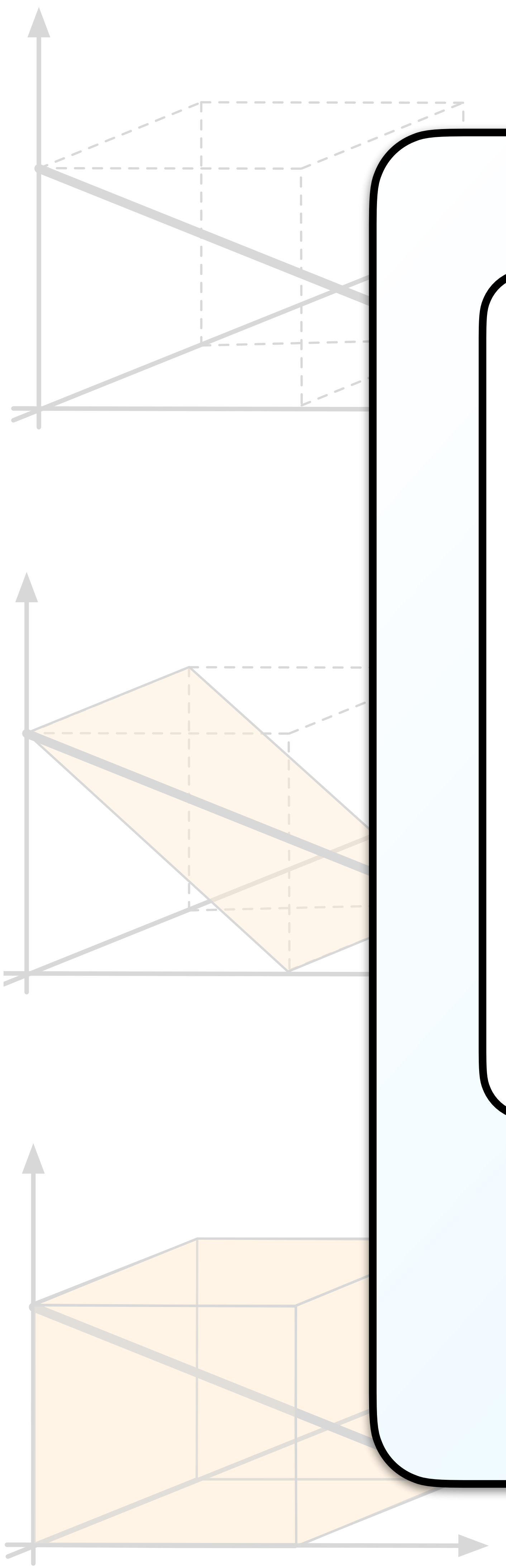


(*s*,*a*)-rectangular ambiguity sets

- 👍 Optimal policy stationary and deterministic
- 👍 Bellman optimality principle holds

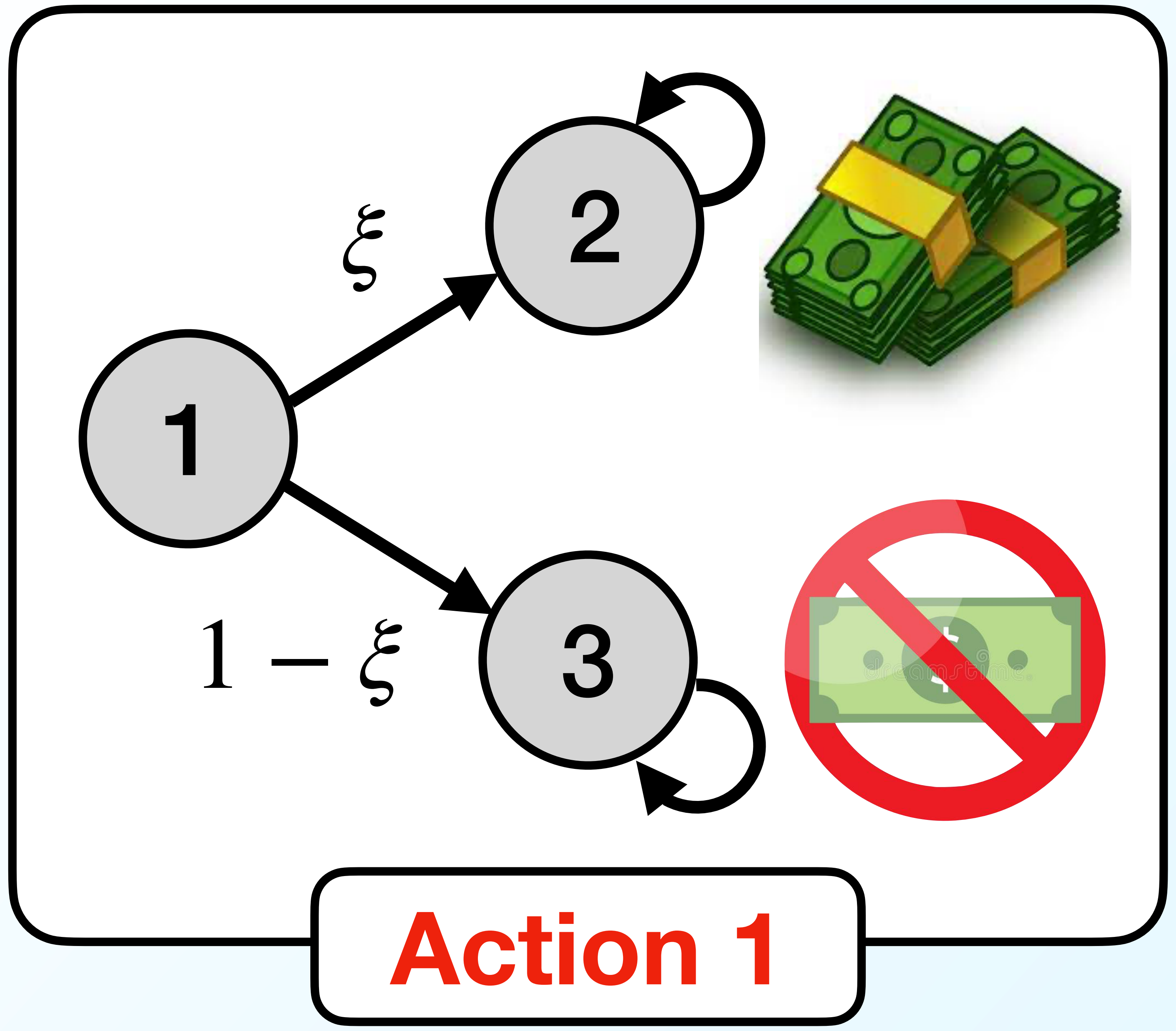
Rectangularity

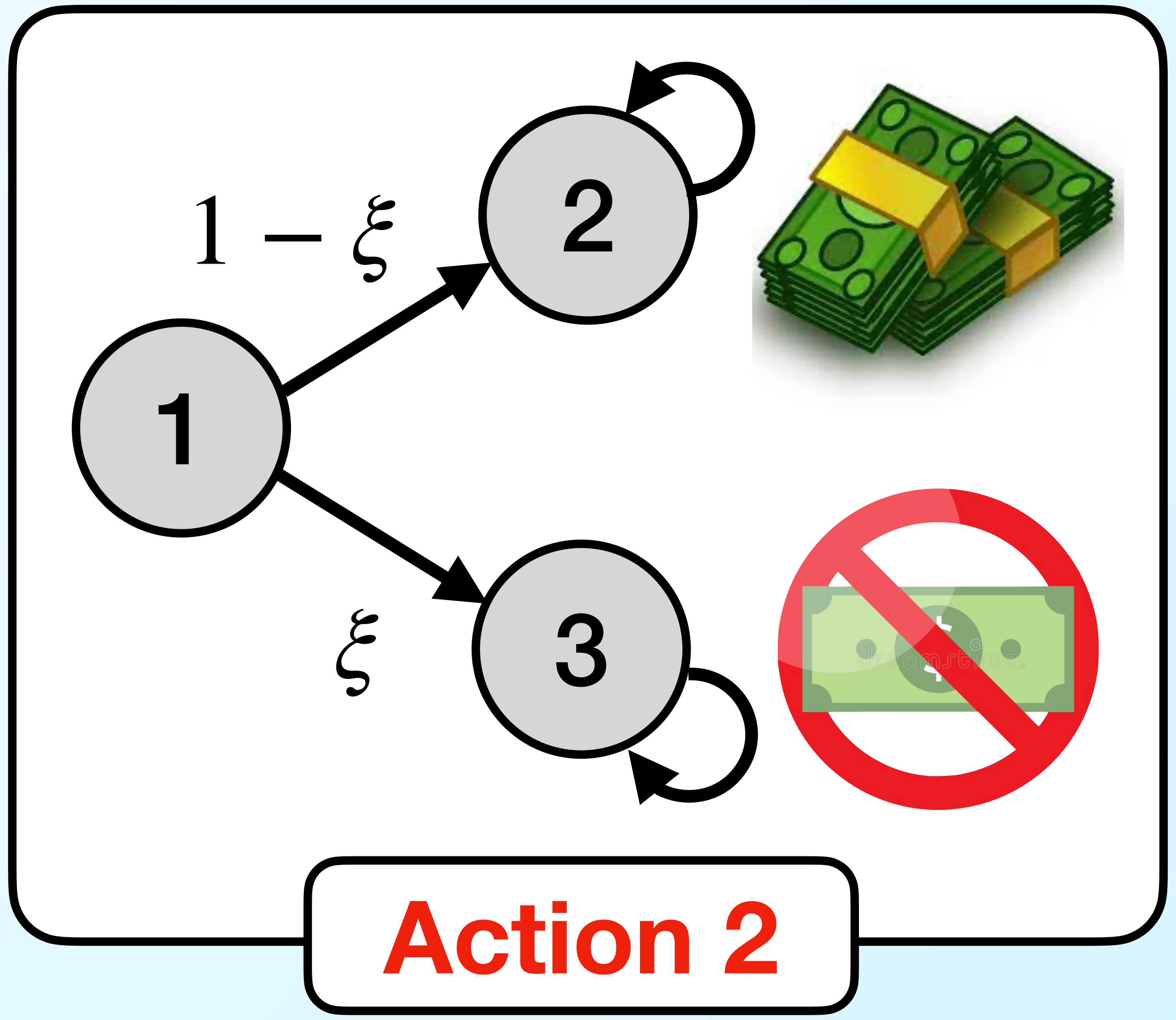
General (non-rectangular) ambiguity sets



endent

Example





for some unknown $\xi \in [0,1]$

Bellman optimality principle holds

s-Rectangular Ambiguity Sets: Bellman Operator

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

Classical (non-robust) Bellman equations

$$v^\star(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v^\star(s') \right\}$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

Robust Bellman equations

$$v^\star(s) = \max_{\pi \in \Delta(\mathscr{A})} \min_{p \in \mathscr{P}_s} \left\{ \sum_{a \in \mathscr{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s, a) \cdot v^\star(s') \right] \right\}$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

Robust Bellman operator

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathscr{A})} \min_{p \in \mathscr{P}_s} \left\{ \sum_{a \in \mathscr{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \,|\, s,a) \cdot v(s') \right] \right\}$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

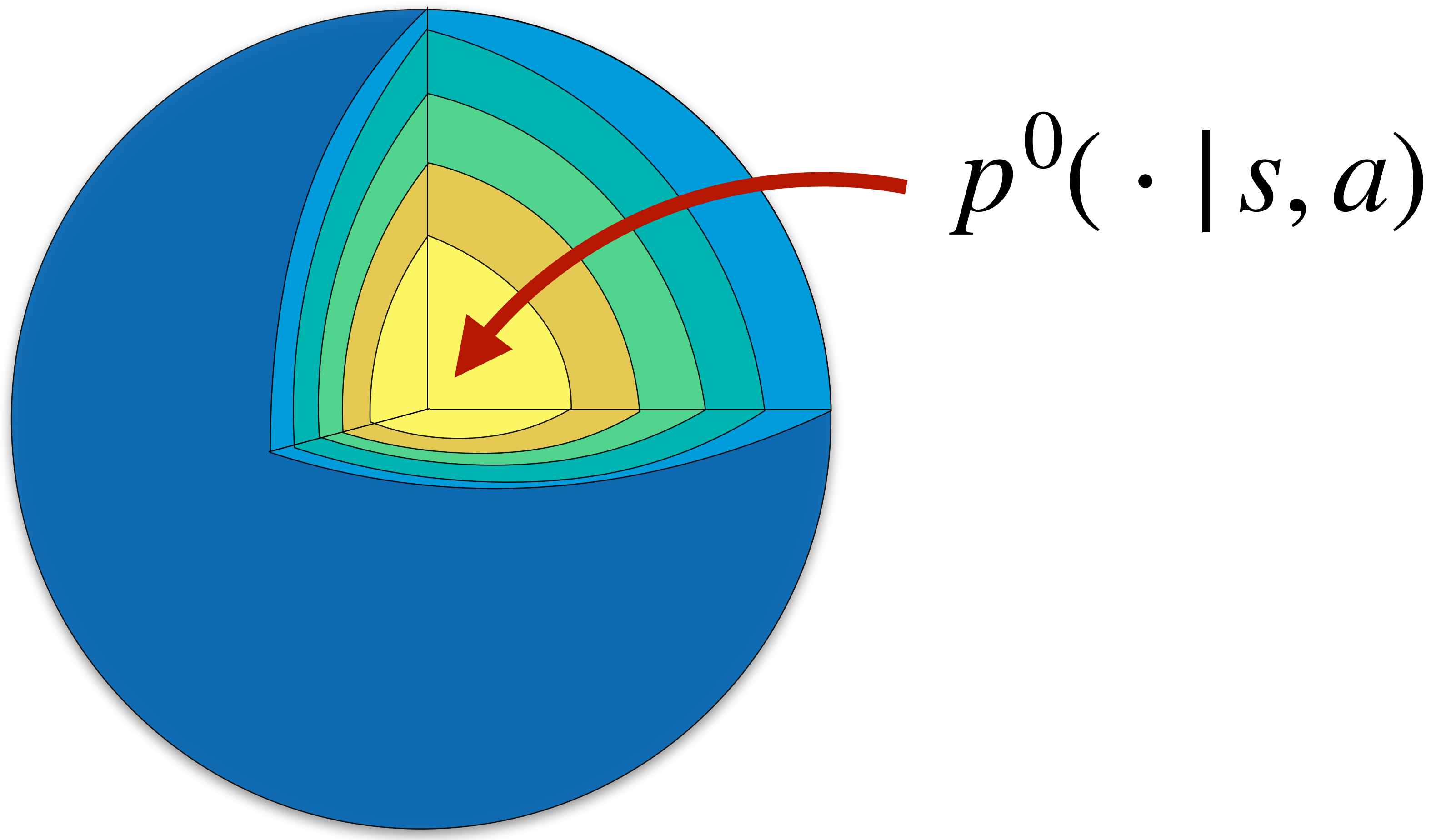
s-Rectangular Ambiguity Sets: Bellman Operator

Robust Bellman operator

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathscr{A})} \min_{p \in \mathscr{P}_s} \left\{ \sum_{a \in \mathscr{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right] \right\}$$

Distance-constrained *s*-rectangular ambiguity set

$$\mathscr{P} = \prod_{s \in \mathscr{S}} \mathscr{P}_s \quad \text{with} \quad \mathscr{P}_s = \left\{ p(\,\cdot \mid s, \cdot\,) \;:\; \sum_{a \in \mathscr{A}} d\left[p(\,\cdot \mid s,a), p^0(\,\cdot \mid s,a) \right] \le \kappa \right\}$$



Ho et al. (2023); Ho et al. (2026).

s-Rectangular Ambiguity Sets: Bellman Operator

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathscr{A})} \min_{p \in \mathscr{P}_s} \left\{ \sum_{a \in \mathscr{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s,a) \cdot v(s') \right] \right\}$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathcal{A})} \min_{p \in \mathcal{P}_s} \left\{ \sum_{a \in \mathcal{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right] \right\}$$

$$= \min_{p \in \mathcal{P}_s} \max_{\pi \in \Delta(\mathcal{A})} \left\{ \sum_{a \in \mathcal{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right] \right\}$$

Minimax theorem: exchange order of min and max

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathcal{A})} \min_{p \in \mathcal{P}_s} \left\{ \sum_{a \in \mathcal{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right] \right\}$$

$$= \min_{p \in \mathcal{P}_s} \max_{\pi \in \Delta(\mathcal{A})} \left\{ \sum_{a \in \mathcal{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right] \right\}$$

$$= \min_{p \in \mathcal{P}_s} \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right\}$$

Linearity: we only need to consider ext $\Delta(\mathcal{A}) = \mathcal{A}$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

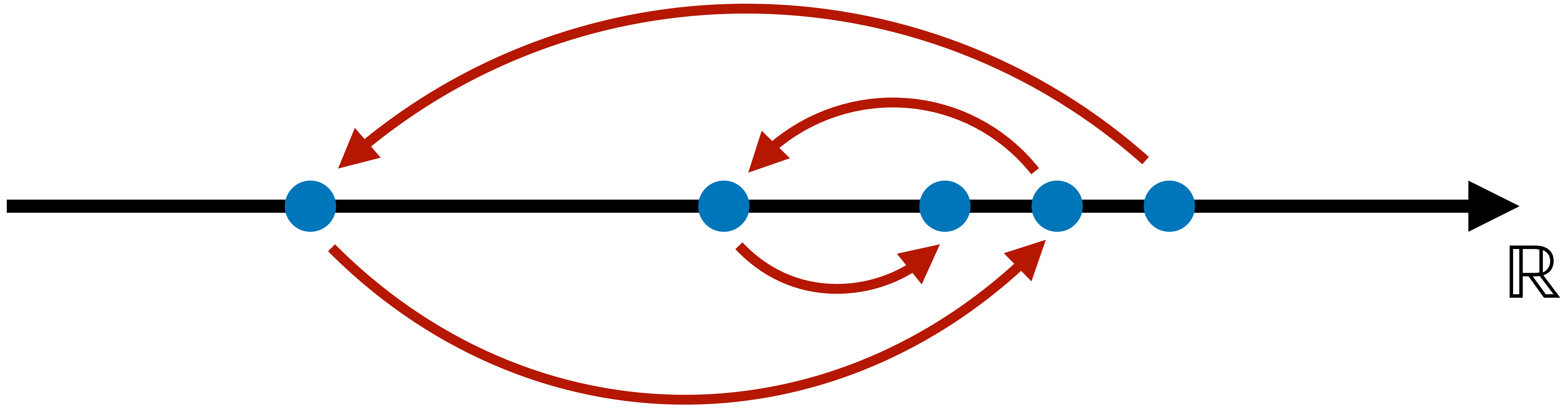
s-Rectangular Ambiguity Sets: Bellman Operator

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathcal{A})} \min_{p \in \mathcal{P}_s} \left\{ \sum_{a \in \mathcal{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right] \right\}$$

$$= \min_{p \in \mathcal{P}_s} \max_{\pi \in \Delta(\mathcal{A})} \left\{ \sum_{a \in \mathcal{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right] \right\}$$

$$\min_{p \in \mathcal{P}_s} \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right\} \leq \theta \, ?$$

Bisection search:



Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

$$\min_{p \in \mathscr{P}_s} \quad \max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right\} \quad \leq \quad \theta \text{ ?}$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

$$\min_{p \in \boxed{\mathscr{P}_s}} \max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right\} \leq \theta \ ?$$

$$\min_{p \in [\Delta(\mathscr{S})]^A} \left\{ \max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right\} : \boxed{\sum_{a \in \mathscr{A}} d\left[p(\cdot \mid s,a), p^0(\cdot \mid s,a)\right] \leq \kappa} \right\} \leq \theta$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

$$\underbrace{\min_{p \in \mathscr{P}_s} \quad \max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s, a) \cdot v(s') \right\}}_{} \quad \leq \quad \theta \text{ ?}$$

$$\min_{p \in [\Delta(\mathscr{S})]^A} \left\{ \boxed{\max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s, a) \cdot v(s') \right\}} : \boxed{\sum_{a \in \mathscr{A}} d\left[p(\,\cdot \mid s, a),\, p^0(\,\cdot \mid s, a)\right]} \leq \kappa \right\} \quad \leq \quad \theta$$

$f\ (p)$ $\qquad\qquad$ $g\ (p)$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

$$\min_{p \in \mathscr{P}_s} \ \max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right\} \ \leq \ \theta \ ?$$

$$\min_{p \in [\Delta(\mathscr{S})]^A} \left\{ \underbrace{\max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right\}}_{f\,(p)} : \underbrace{\sum_{a \in \mathscr{A}} d\left[p(\,\cdot \mid s,a),\, p^0(\,\cdot \mid s,a)\right]}_{g\,(p)} \leq \kappa \right\} \ \leq \ \theta$$

$$\min_{p \in [\Delta(\mathscr{S})]^A} \left\{ \underbrace{\sum_{a \in \mathscr{A}} d\left[p(\,\cdot \mid s,a),\, p^0(\,\cdot \mid s,a)\right]}_{g\,(p)} : \underbrace{\max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right\}}_{f\,(p)} \leq \theta \right\} \ \leq \ \kappa$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

$$\underbrace{\min_{p \in \mathscr{P}_s} \max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right\}} \leq \theta \ ?$$

$$\min_{p \in [\Delta(\mathscr{S})]^A} \left\{ \underbrace{\boxed{\sum_{a \in \mathscr{A}} d\left[p(\,\cdot \mid s,a), p^0(\,\cdot \mid s,a)\right]}}_{g\ (p)} : \underbrace{\boxed{\max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right\}}}_{f\ (p)} \leq \theta \right\} \leq \kappa$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

$$\underbrace{\min_{p \in \mathscr{P}_s} \max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right\}}\ \le\ \theta\ ?$$

$$\min_{p \in [\Delta(\mathscr{S})]^A} \left\{ \boxed{\sum_{a \in \mathscr{A}} d\left[p(\,\cdot \mid s,a),\, p^0(\,\cdot \mid s,a)\right]} : \boxed{\max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \right\} \le \theta} \right\} \ \le\ \kappa$$

g (*p*) *f* (*p*)

$$\Longleftrightarrow \quad \sum_{a \in \mathscr{A}} \min_{p_a \in \Delta(\mathscr{S})} \left\{ d\left[p(\,\cdot \mid s,a),\, p^0(\,\cdot \mid s,a)\right] \ :\ r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s' \mid s,a) \cdot v(s') \le \theta \right\} \ \le\ \kappa$$

Separability: of both objective and constraints in $a \in \mathscr{A}$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

$$\underbrace{\min_{p \in \mathcal{P}_s} \; \max_{a \in \mathcal{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s,a) \cdot v(s') \right\}} \;\leq\; \theta \; ?$$

$$\sum_{a \in \mathcal{A}} \min_{p_a \in \Delta(\mathcal{S})} \left\{ d\left[p(\,\cdot\,|\,s,a),\, p^0(\,\cdot\,|\,s,a)\right] \;:\; r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s,a) \cdot v(s') \leq \theta \right\} \;\leq\; \kappa$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

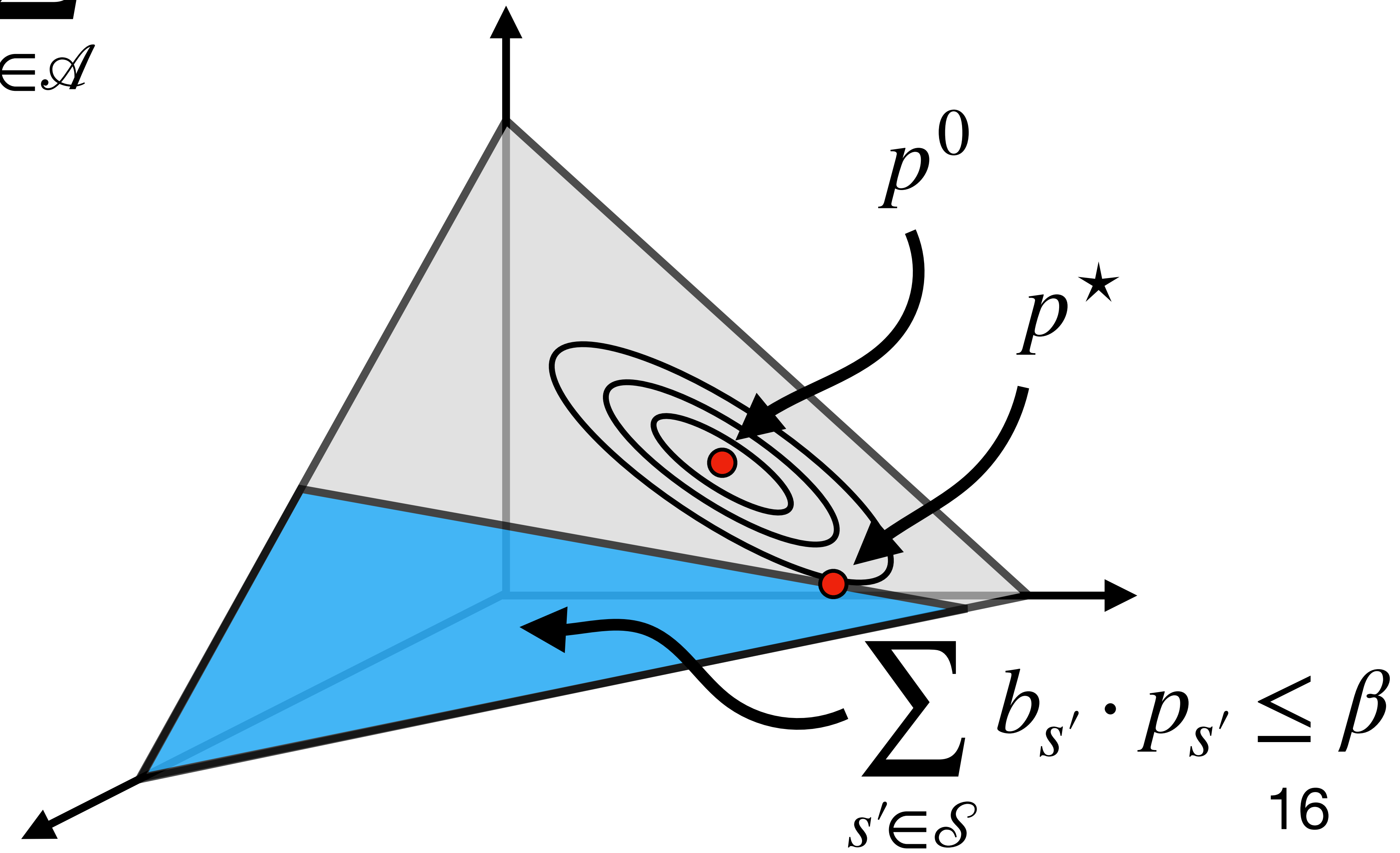
s-Rectangular Ambiguity Sets: Bellman Operator

$$\min_{p \in \mathscr{P}_s} \; \max_{a \in \mathscr{A}} \left\{ r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s,a) \cdot v(s') \right\} \;\leq\; \theta \,?$$

$$\sum_{a \in \mathscr{A}} \min_{p_a \in \Delta(\mathcal{S})} \left\{ d\left[p(\,\cdot\,|\, s,a),\, p^0(\,\cdot\,|\, s,a)\right] \;:\; r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s,a) \cdot v(s') \leq \theta \right\} \;\leq\; \kappa$$

$$\Longleftrightarrow \; \sum_{a \in \mathscr{A}} \mathfrak{P}(p^0; \lambda v, \theta - r(s \,|\, a)) \;\leq\; \kappa$$

with $\mathfrak{P}(p^0; b, \beta) \; = \left[\begin{array}{ll} \underset{p}{\text{minimize}} & d\left[p, p^0\right] \\ \text{subject to} & \sum_{s' \in \mathcal{S}} b_{s'} \cdot p_{s'} \leq \beta \\ & p \in \Delta(\mathcal{S}) \end{array} \right]$



Ho et al. (2023); Ho et al. (2026).

s-Rectangular Ambiguity Sets: Bellman Operator

Distance-constrained *s*-rectangular ambiguity set

$$\mathscr{P} = \prod_{s \in \mathscr{S}} \mathscr{P}_s \quad \text{with} \quad \mathscr{P}_s = \left\{ p(\cdot \mid s, \cdot) \; : \; \sum_{a \in \mathscr{A}} d\left[p(\cdot \mid s, a), p^0(\cdot \mid s, a)\right] \leq \kappa \right\}$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

Distance-constrained *s*-rectangular ambiguity set

$$\mathscr{P} = \prod_{s \in \mathscr{S}} \mathscr{P}_s \quad \text{with} \quad \mathscr{P}_s = \left\{ p(\,\cdot\,|\,s,\cdot\,) \;:\; \sum_{a \in \mathscr{A}} d\left[p(\,\cdot\,|\,s,a),\, p^0(\,\cdot\,|\,s,a)\right] \leq \kappa \right\}$$

Robust Bellman operator

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathscr{A})} \min_{p \in \mathscr{P}_s} \left\{ \sum_{a \in \mathscr{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s'\,|\,s,a) \cdot v(s') \right] \right\}$$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

Distance-constrained *s*-rectangular ambiguity set

$$\mathscr{P} = \prod_{s \in \mathscr{S}} \mathscr{P}_s \quad \text{with} \quad \mathscr{P}_s = \left\{ p(\,\cdot\,|\,s,\cdot\,) \;:\; \sum_{a \in \mathscr{A}} d\left[p(\,\cdot\,|\,s,a),\, p^0(\,\cdot\,|\,s,a)\right] \leq \kappa \right\}$$

Robust Bellman operator

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathscr{A})} \min_{p \in \mathscr{P}_s} \left\{ \sum_{a \in \mathscr{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathscr{S}} p(s'\,|\,s,a) \cdot v(s') \right] \right\}$$

Projection problem

$$\mathfrak{P}(p^0; b, \beta) \;=\; \left[\begin{array}{ll} \underset{p}{\text{minimize}} & d\left[p, p^0\right] \\ \text{subject to} & \displaystyle\sum_{s' \in \mathscr{S}} b_{s'} \cdot p_{s'} \leq \beta \\ & p \in \Delta(\mathscr{S}) \end{array} \right]$$

s-Rectangular Ambiguity Sets: Bellman Operator

Distance-constrained s-rectangular ambiguity set

$$\mathscr{P} = \prod_{s \in \mathscr{S}} \mathscr{P}_s \quad \text{with} \quad \mathscr{P}_s = \left\{ p(\,\cdot\,|\,s,\cdot\,) \;:\; \sum_{a \in \mathscr{A}} d\left[p(\,\cdot\,|\,s,a),\, p^0(\,\cdot\,|\,s,a)\right] \leq \kappa \right\}$$

Theorem

Assume $\mathfrak{P}$ can be computed exactly in time $\mathcal{O}(h(S))$.
Then $\mathfrak{B}$ can be computed to accuracy $\epsilon > 0$ in time $\mathcal{O}(AS \cdot h(S) \cdot \log[\overline{R}/\epsilon])$.

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

Distance-constrained s-rectangular ambiguity set

$$\mathscr{P} = \prod_{s \in \mathcal{S}} \mathscr{P}_s \quad \text{with} \quad \mathscr{P}_s = \left\{ p(\cdot \mid s, \cdot) \, : \, \sum_{a \in \mathcal{A}} d\left[p(\cdot \mid s, a), p^0(\cdot \mid s, a) \right] \leq \kappa \right\}$$

Theorem

Assume $\mathfrak{P}$ can be computed exactly in time $\mathcal{O}(h(S))$. Then $\mathfrak{B}$ can be computed to accuracy $\epsilon > 0$ in time $\mathcal{O}(AS \cdot h(S) \cdot \log[\overline{R}/\epsilon])$.

Assume $\mathfrak{P}$ can be computed to any accuracy $\delta > 0$ in time $\mathcal{O}(h(\delta))$. Then $\mathfrak{B}$ can be computed to accuracy $\epsilon > 0$ in time $\mathcal{O}(AS \cdot h(\epsilon\kappa/[2A\overline{R} + A\epsilon]) \cdot \log[\overline{R}/\epsilon])$.

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Bellman Operator

Divergence	$d_a(\cdot, p^0)$	Ours	Previous
KL-Divergence	$\sum_{s' \in \mathcal{S}} p(s' \| s, a) \cdot \log \left(\frac{p(s' \| s, a)}{p^0(s' \| s, a)} \right)$	$\mathcal{O}(S^2 A \cdot \log A)$	$\mathcal{O}(\ell^2 \cdot S^2 \cdot A)$
Burg Entropy	$\sum_{s' \in \mathcal{S}} p^0(s' \| s, a) \cdot \log \left(\frac{p^0(s' \| s, a)}{p(s' \| s, a)} \right)$	$\mathcal{O}(S^2 A \cdot \log A)$	(none)
Variation Distance	$\sum_{s' \in \mathcal{S}} \| p(s' \| s, a) - p^0(s' \| s, a) \|$	$\mathcal{O}(S^2 A \cdot \log S)$	$\mathcal{O}(S^2 A \cdot \log S)$
χ^2-Distance	$\sum_{s' \in \mathcal{S}} \frac{\left[p(s' \| s, a) - p^0(s' \| s, a) \right]^2}{p^0(s' \| s, a)}$	$\mathcal{O}(S^2 A \cdot \log S)$	$\mathcal{O}(S^{4.5} \cdot A)$

Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

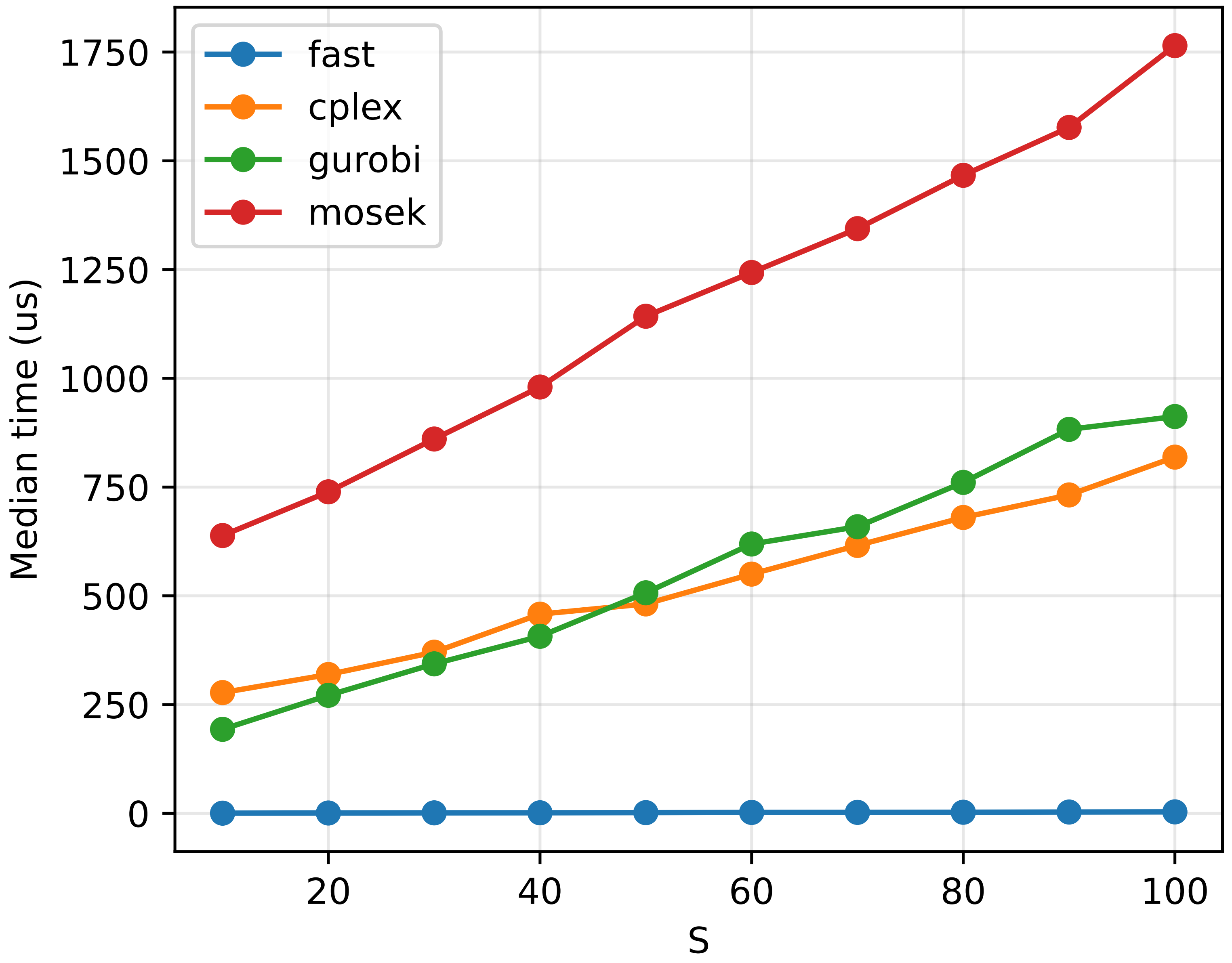
s-Rectangular Ambiguity Sets: Bellman Operator

Divergence	$d_a(\cdot, p^0)$	Ours	Previous
KL-Divergence	$\sum_{s' \in \mathcal{S}} p(s' \mid s, a) \cdot \log\left(\frac{p(s' \mid s, a)}{p^0(s' \mid s, a)}\right)$	$\mathcal{O}(S^2 A \cdot \log A)$	$\mathcal{O}(\ell^2 \cdot S^2 \cdot A)$
Burg Entropy	$\sum_{s' \in \mathcal{S}} p^0(s' \mid s, a) \cdot \log\left(\frac{p^0(s' \mid s, a)}{p(s' \mid s, a)}\right)$	$\mathcal{O}(S^2 A \cdot \log A)$	(none)
Variation Distance	$\sum_{s' \in \mathcal{S}} \lvert p(s' \mid s, a) - p^0(s' \mid s, a) \rvert$	$\mathcal{O}(S^2 A \cdot \log S)$	$\mathcal{O}(S^2 A \cdot \log S)$
χ^2-Distance	$\sum_{s' \in \mathcal{S}} \frac{\left[p(s' \mid s, a) - p^0(s' \mid s, a)\right]^2}{p^0(s' \mid s, a)}$	$\mathcal{O}(S^2 A \cdot \log S)$	$\mathcal{O}(S^{4.5} \cdot A)$

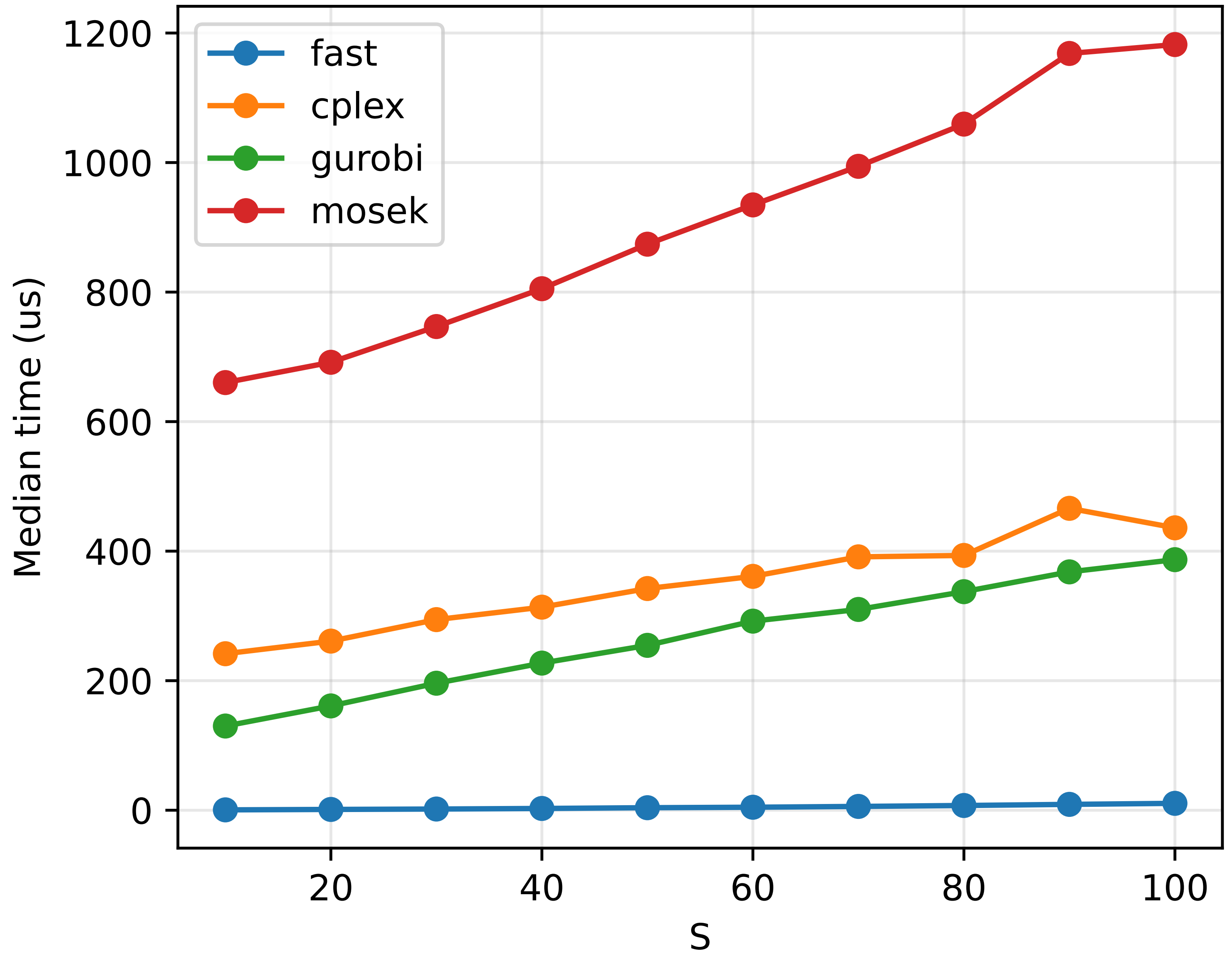
Ho et al. (2023), *Robust Phi-Divergence MDPs;* Ho et al. (2026), *Efficient Algorithms for Robust MDPs with s-Rectangular Ambiguity Sets*.

s-Rectangular Ambiguity Sets: Projection Problem

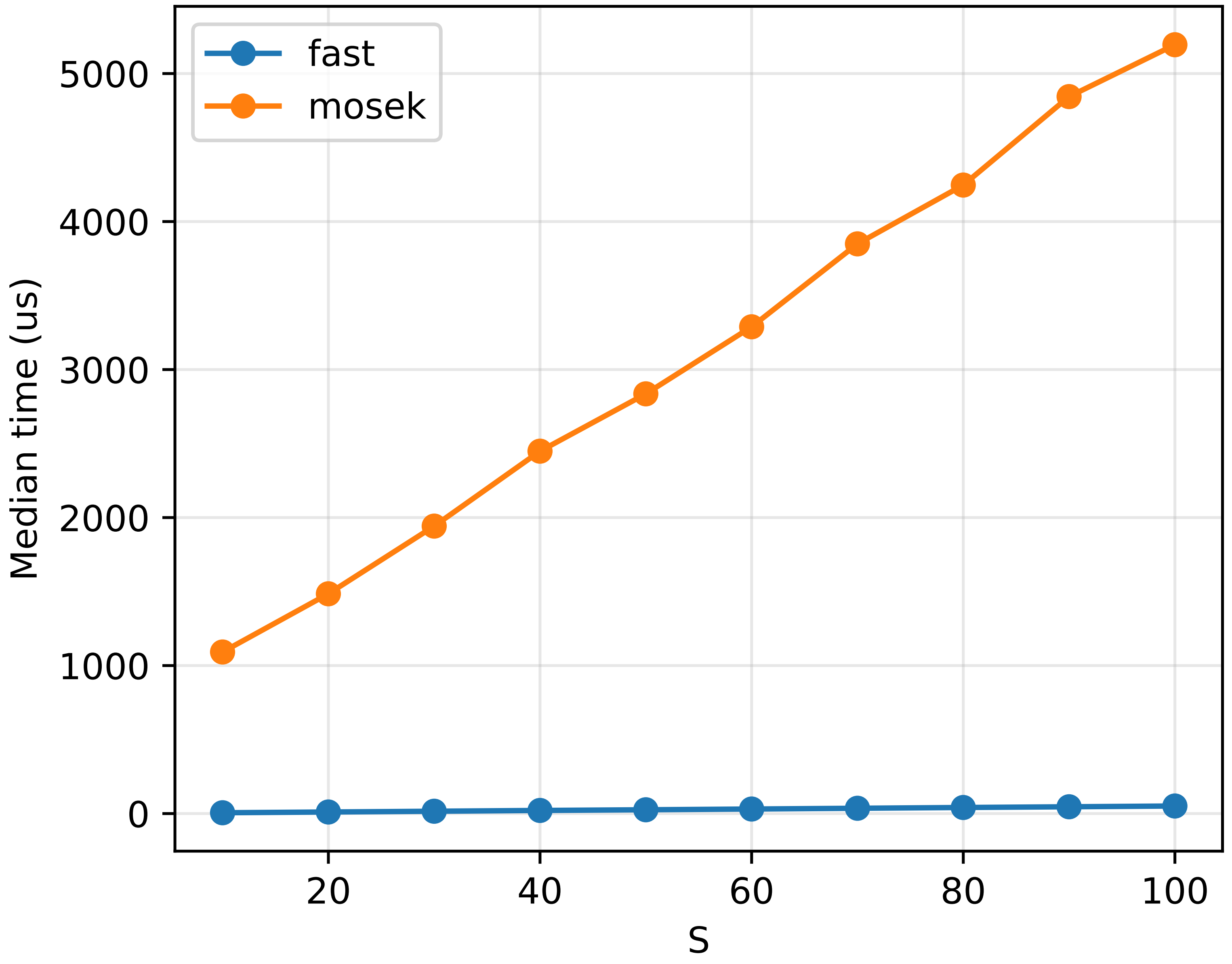
1-norm



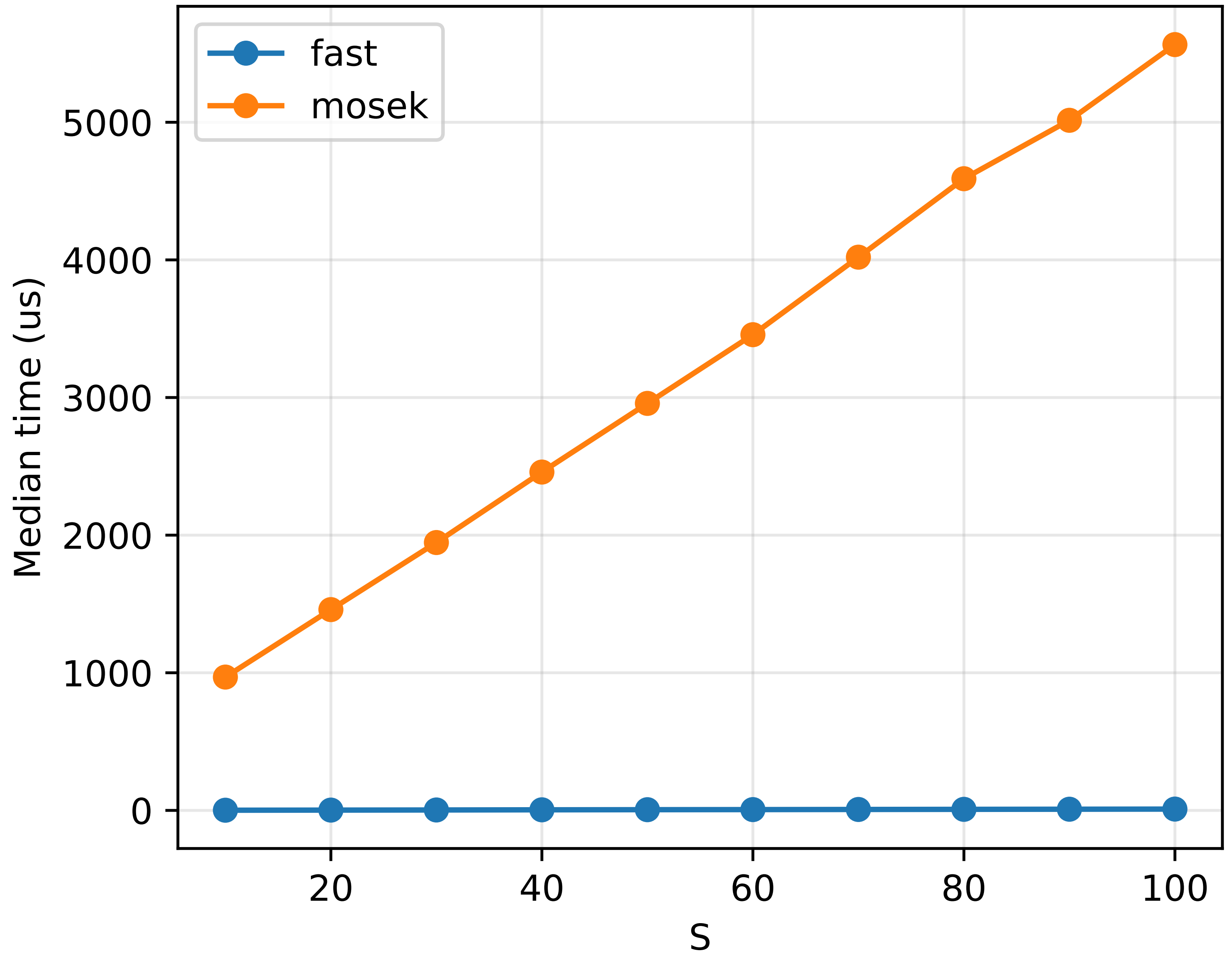
2-norm



KL-Div

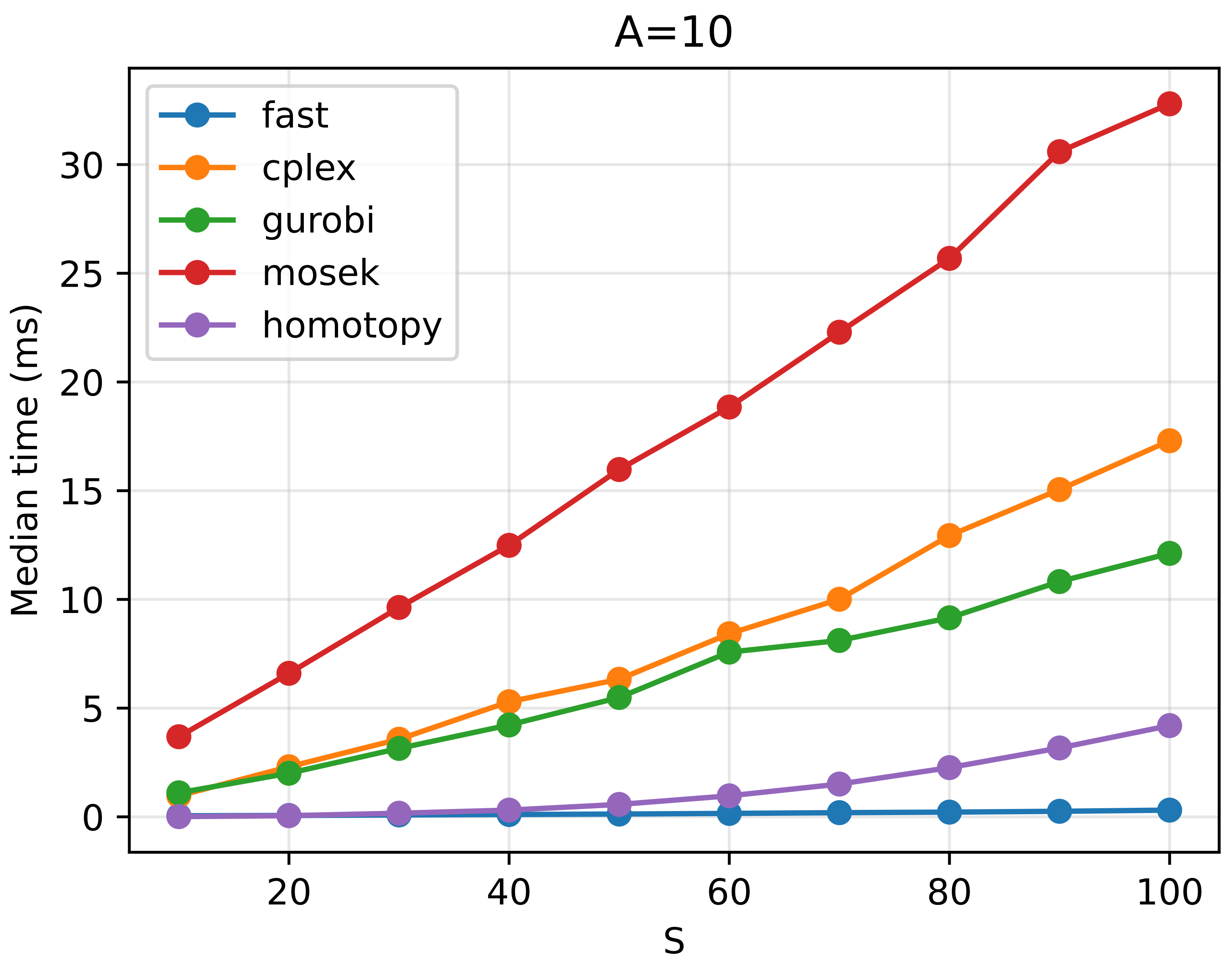


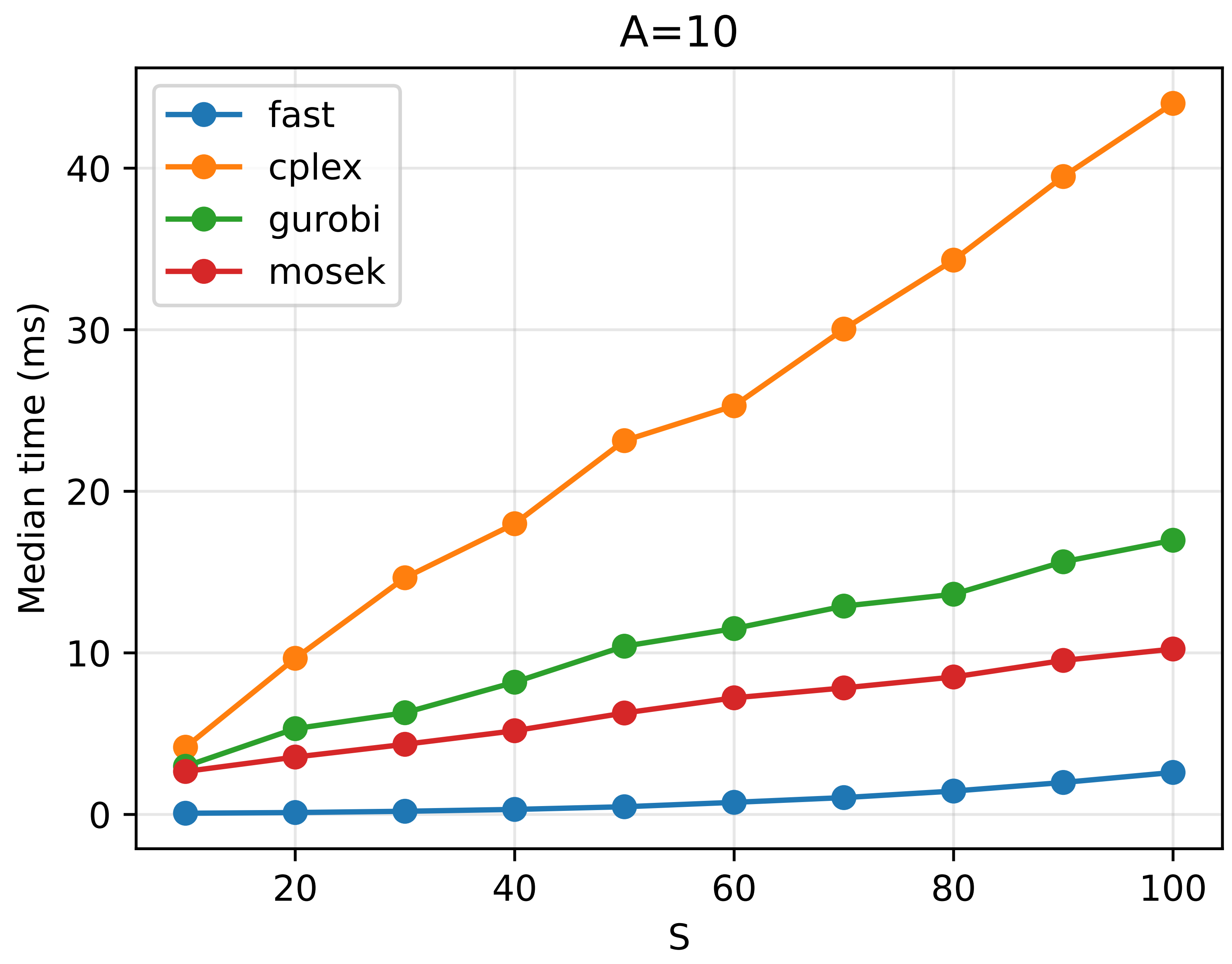
Burg



s-Rectangular Ambiguity Sets: Bellman Operator

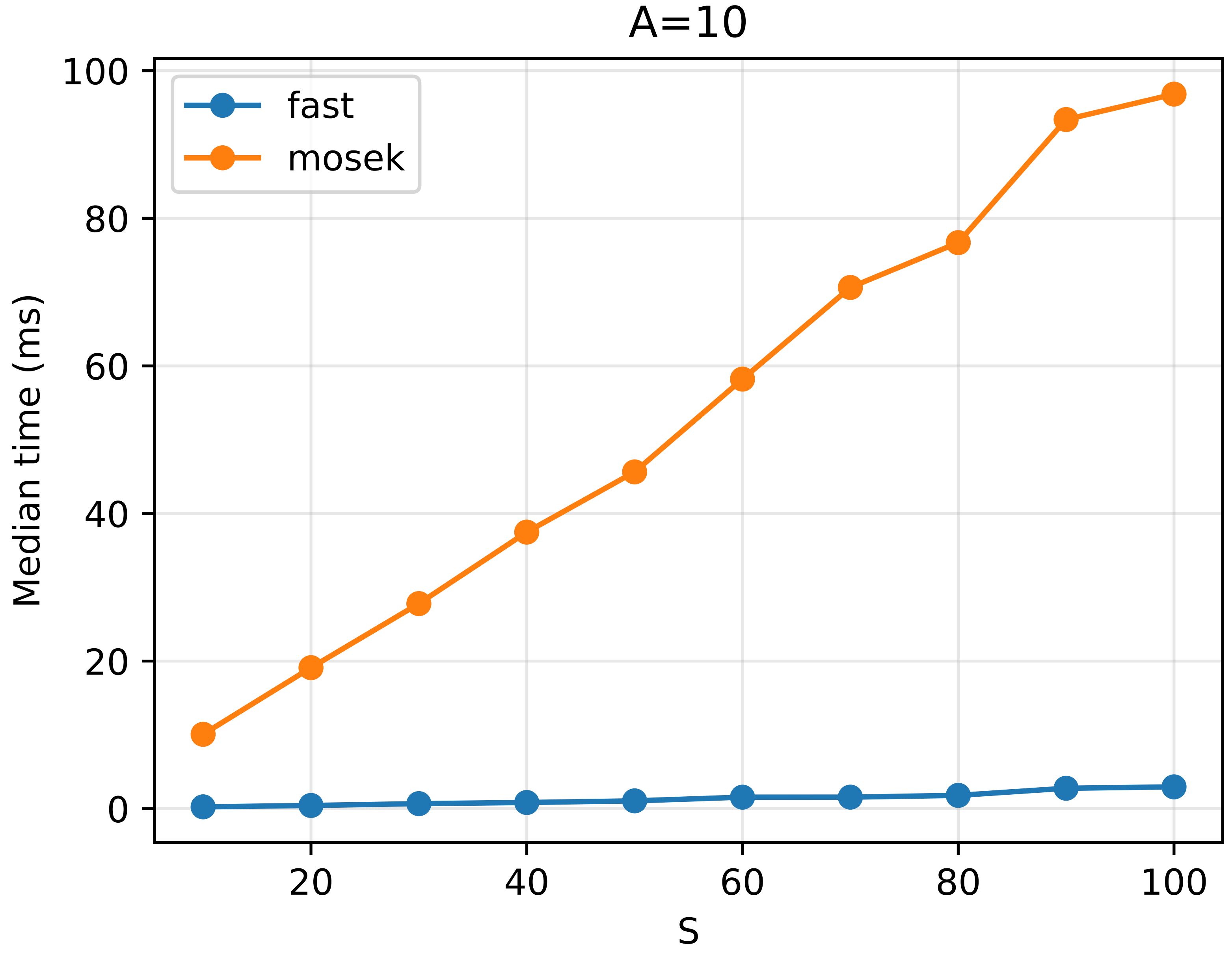
1-norm

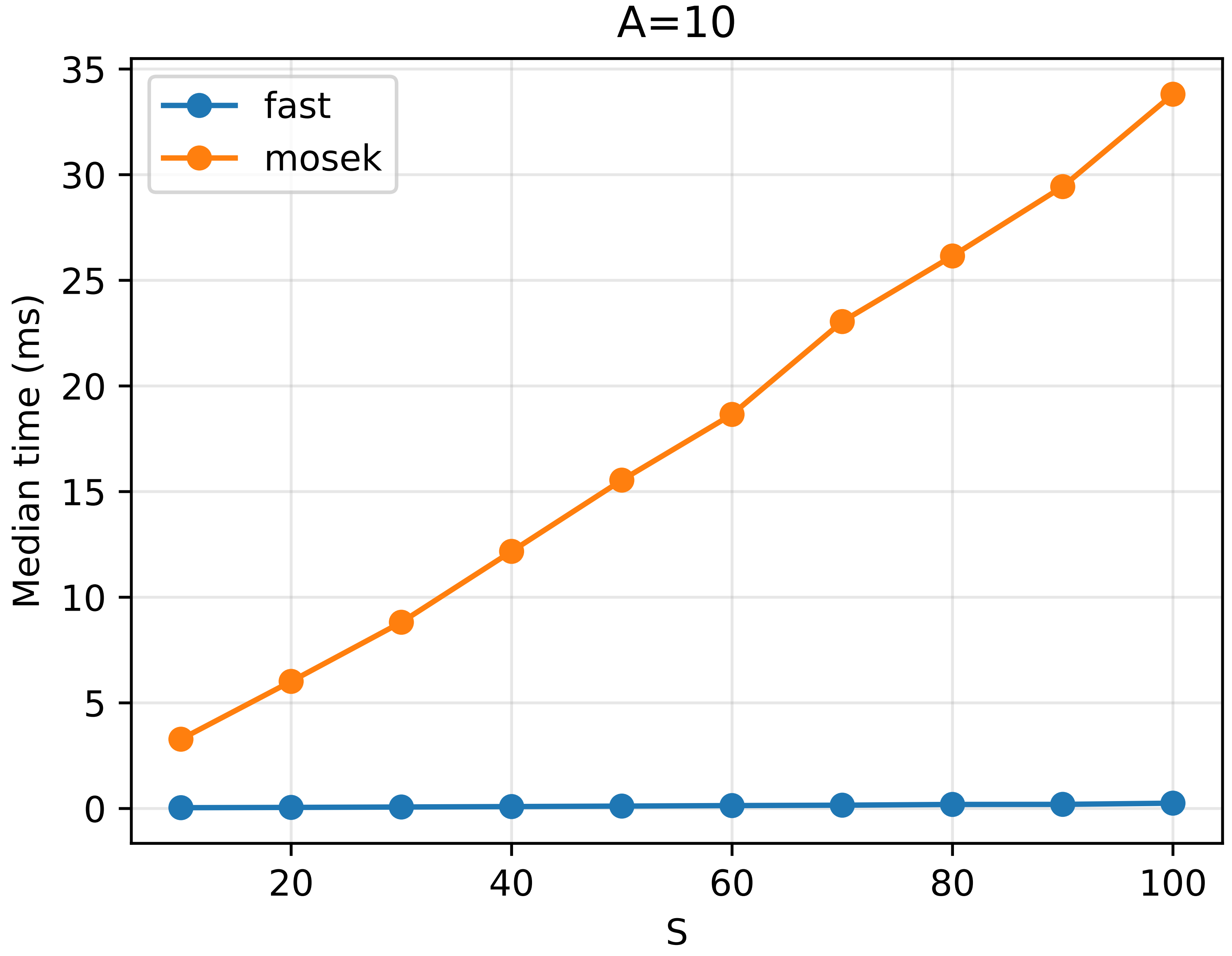




2-norm

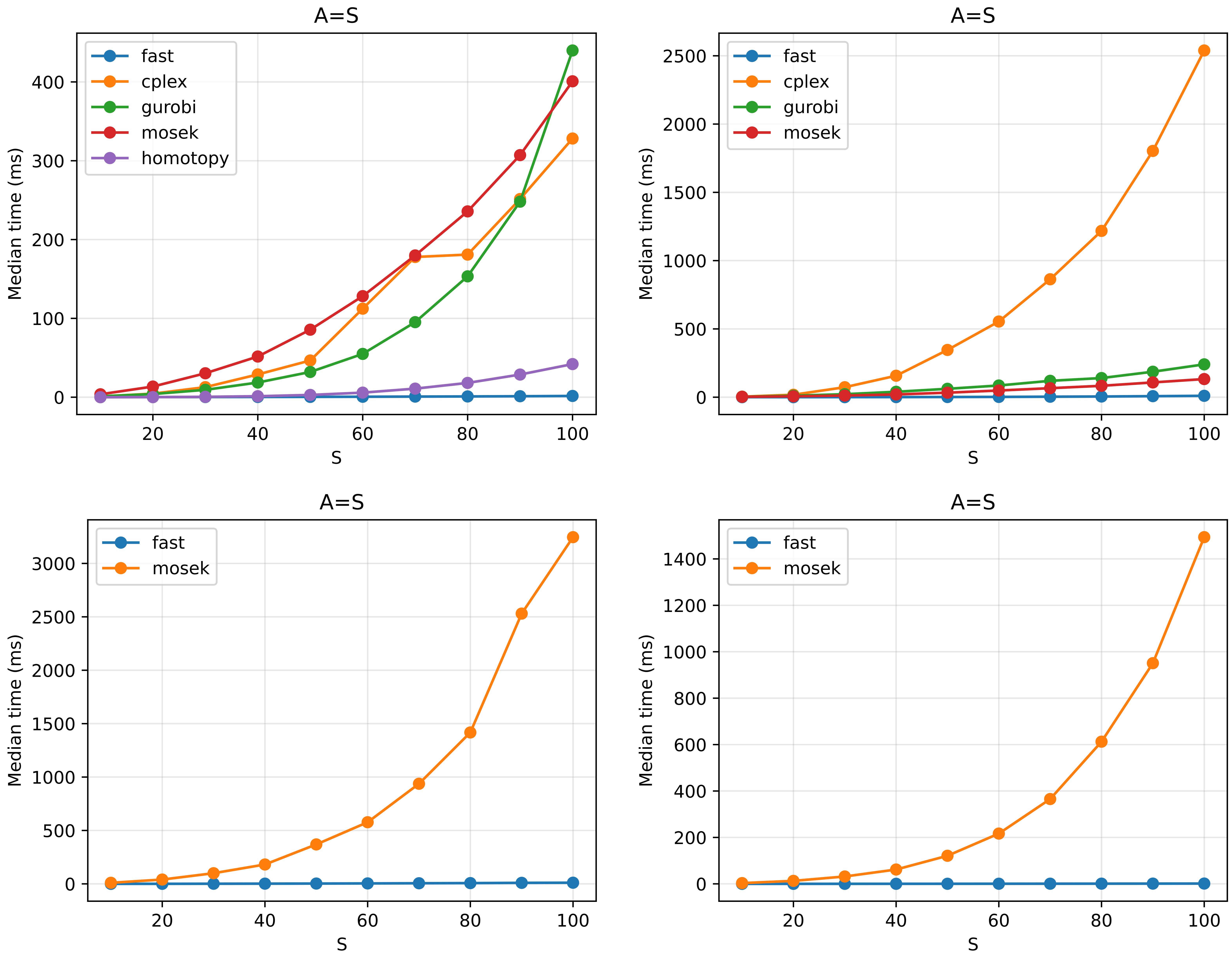
KL-Div





Burg

s-Rectangular Ambiguity Sets: Bellman Operator



1-norm

2-norm

KL-Div

Burg

Reconsider idea of (modified) policy iteration:

Ho et al. (2021), *Partial Policy Iteration for L1-Robust Markov Decision Processes*.

s-Rectangular Ambiguity Sets: Partial Policy Iteration

Reconsider idea of (modified) policy iteration:

Policy improvement

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathcal{A})} \min_{p \in \mathcal{P}_s} \left\{ \sum_{a \in \mathcal{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \,|\, s,a) \cdot v(s') \right] \right\}$$

expensive operator: requires robust Bellman operator

Ho et al. (2021), *Partial Policy Iteration for L1-Robust Markov Decision Processes*.

s-Rectangular Ambiguity Sets: Partial Policy Iteration

Reconsider idea of (modified) policy iteration:

Policy improvement

$$[\mathfrak{B}v](s) = \max_{\pi \in \Delta(\mathscr{A})} \min_{p \in \mathscr{P}_s} \left\{ \sum_{a \in \mathscr{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right] \right\}$$

expensive operator: requires robust Bellman operator

Policy evaluation

$$[\mathfrak{B}(\pi)\, v](s) = \min_{p \in \mathscr{P}_s} \left\{ \sum_{a \in \mathscr{A}} \pi(a) \cdot \left[r(s,a) + \lambda \sum_{s' \in \mathcal{S}} p(s' \mid s,a) \cdot v(s') \right] \right\}$$

cheaper operator: can be recast as Bellman operator of a nominal MDP

s-Rectangular Ambiguity Sets: Robust Modified Policy Iteration

Kaufman & Schaefer (2013), *Robust Modified Policy Iteration*.

s-Rectangular Ambiguity Sets: Robust Modified Policy Iteration

Repeat: starting with $i = 0$, v^0 arbitrary

Policy improvement

Compute $w^{i+1} = \mathfrak{B} v^i$ and let π^{i+1} be the corresponding greedy policy

expensive operator:
robust value iteration

$i = i + 1$

Until $\|w^{i+1} - \vartheta^{i+1,N}\|_\infty < \frac{1-\lambda}{2} \cdot \delta$

Kaufman & Schaefer (2013), *Robust Modified Policy Iteration*.

s-Rectangular Ambiguity Sets: Robust Modified Policy Iteration

Repeat: starting with $i = 0$, v^0 arbitrary

Policy improvement

Compute $w^{i+1} = \mathfrak{B} v^i$ and let π^{i+1} be the corresponding greedy policy

Policy evaluation

Compute sequence $\vartheta^{i+1,j+1} = \mathfrak{B}(\pi^{i+1})\, \vartheta^{i+1,j}$ with $\vartheta^{i+1,0} = w^{i+1}$ until $\|\vartheta^{i+1,j+1} - \vartheta^{i+1,j}\|_\infty \leq (1-\lambda)\epsilon_{i+1}$

cheaper operator: no maximum involved

$i = i + 1$

Until $\|w^{i+1} - \vartheta^{i+1,N}\|_\infty < \dfrac{1-\lambda}{2} \cdot \delta$

Kaufman & Schaefer (2013), *Robust Modified Policy Iteration*.

s-Rectangular Ambiguity Sets: Partial Policy Iteration

Repeat: starting with $i = 0$, v^0 arbitrary

Policy improvement

Compute $w^{i+1} = \mathfrak{B} v^i$ and let π^{i+1} be the corresponding greedy policy

Policy evaluation

Alternate between single robust Bellman evaluation $\mathfrak{B}(\pi^{i+1})$ and multiple nominal Bellman evaluations under worst-case p.

cheap (!) operator: nominal evaluations

$i = i + 1$

Until $\|w^{i+1} - \vartheta^{i+1,N}\|_\infty < \frac{1-\lambda}{2} \cdot \delta$

Ho et al. (2021), *Partial Policy Iteration for L1-Robust Markov Decision Processes*.

Repeat: starting with $i = 0$, v^0 arbitrary

Policy improvement

Compute $w^{i+1} = \mathfrak{B} v^i$ and ... he corresponding greedy policy

C ... until

$\|$

$i = i + 1$

Until $\|w^{i+1} - \vartheta^{i+1,N}\|_\infty < \frac{1-\lambda}{2} \cdot \delta$

Theorem

Assume $\epsilon_{i+1} < \lambda^c \cdot \epsilon_i$ for some $c > 1$. Then the optimality gap of partial policy iteration satisfies:

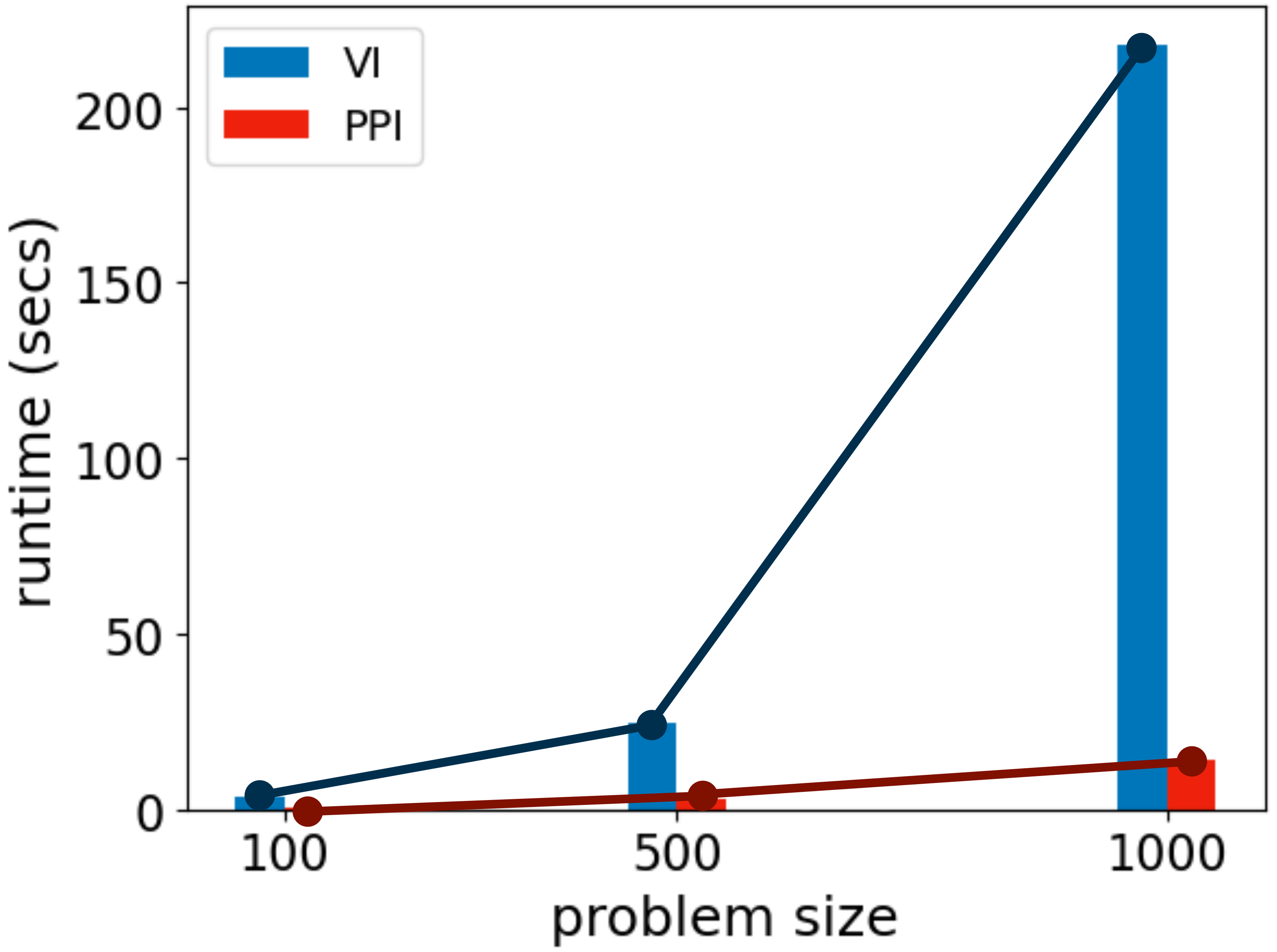
$$\|v(\pi^{i+1}) - v^\star\|_\infty \leq \lambda^i \left(\|v(\pi^1) - v^\star\|_\infty + \frac{2\epsilon_1}{(1 - \lambda^{c-1})(1-\lambda)} \right)$$

Ho et al. (2021), *Partial Policy Iteration for L1-Robust Markov Decision Processes*.

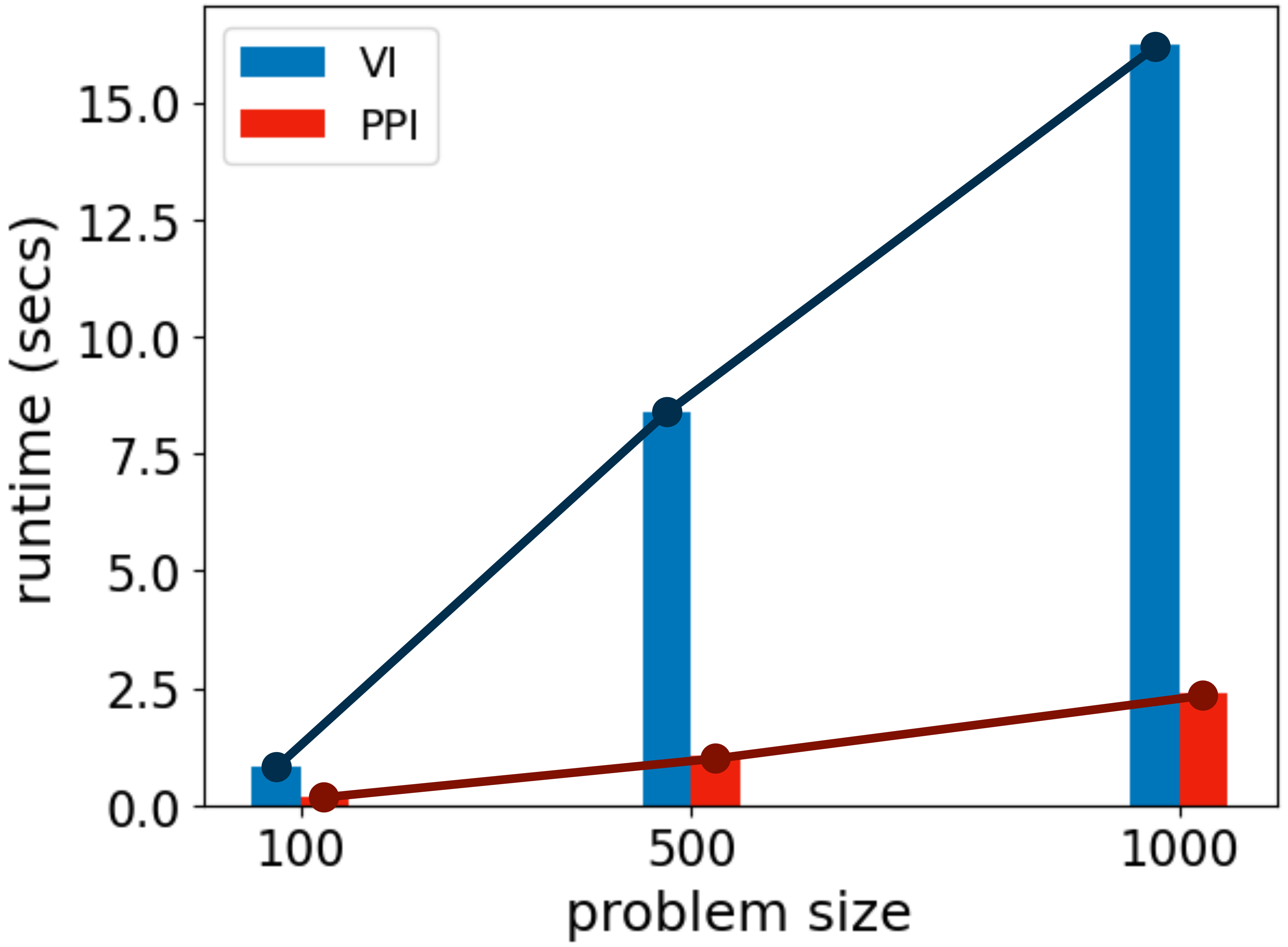
s-Rectangular Ambiguity Sets: Partial Policy Iteration

Inventory

uniform

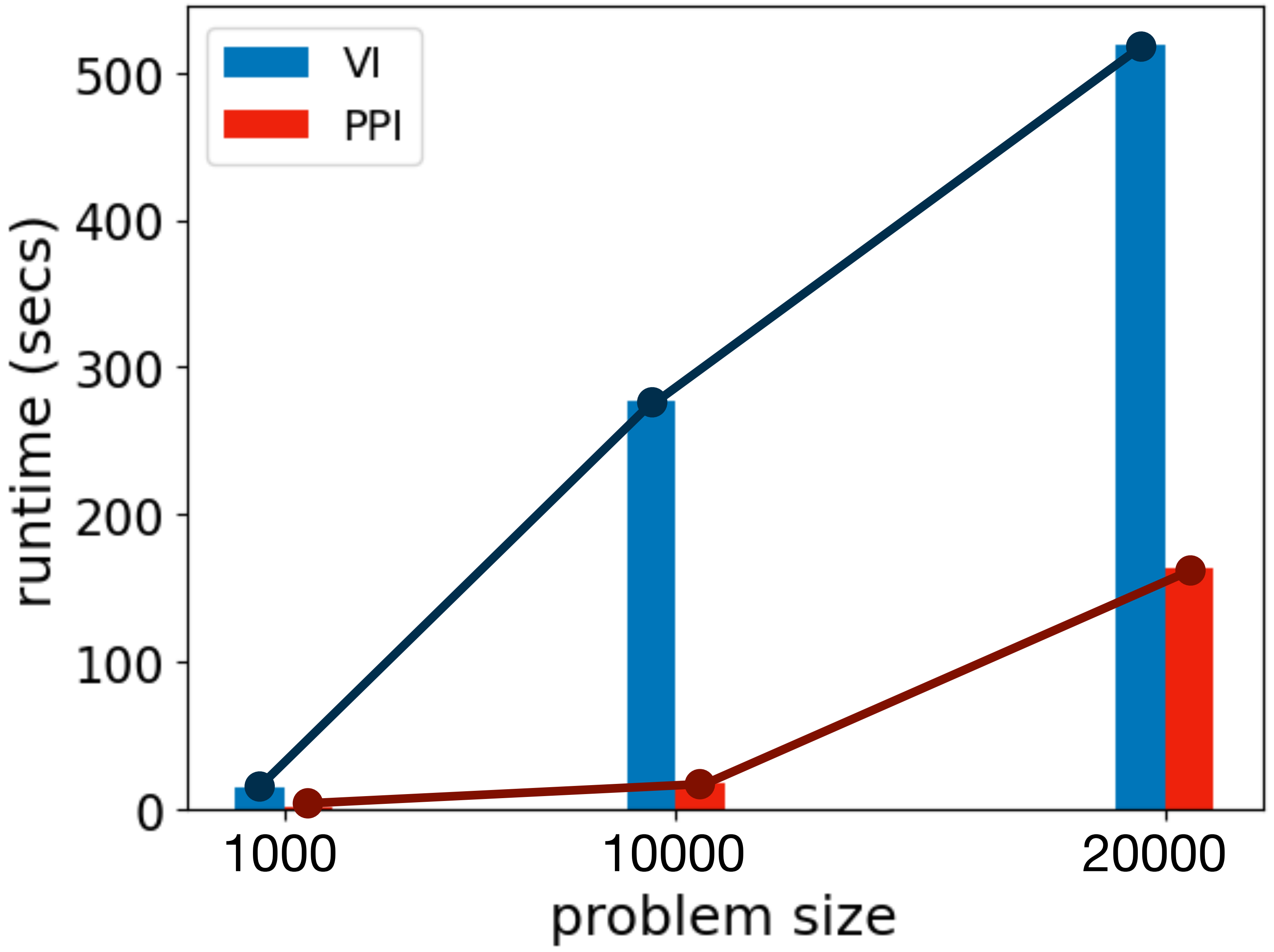


weighted

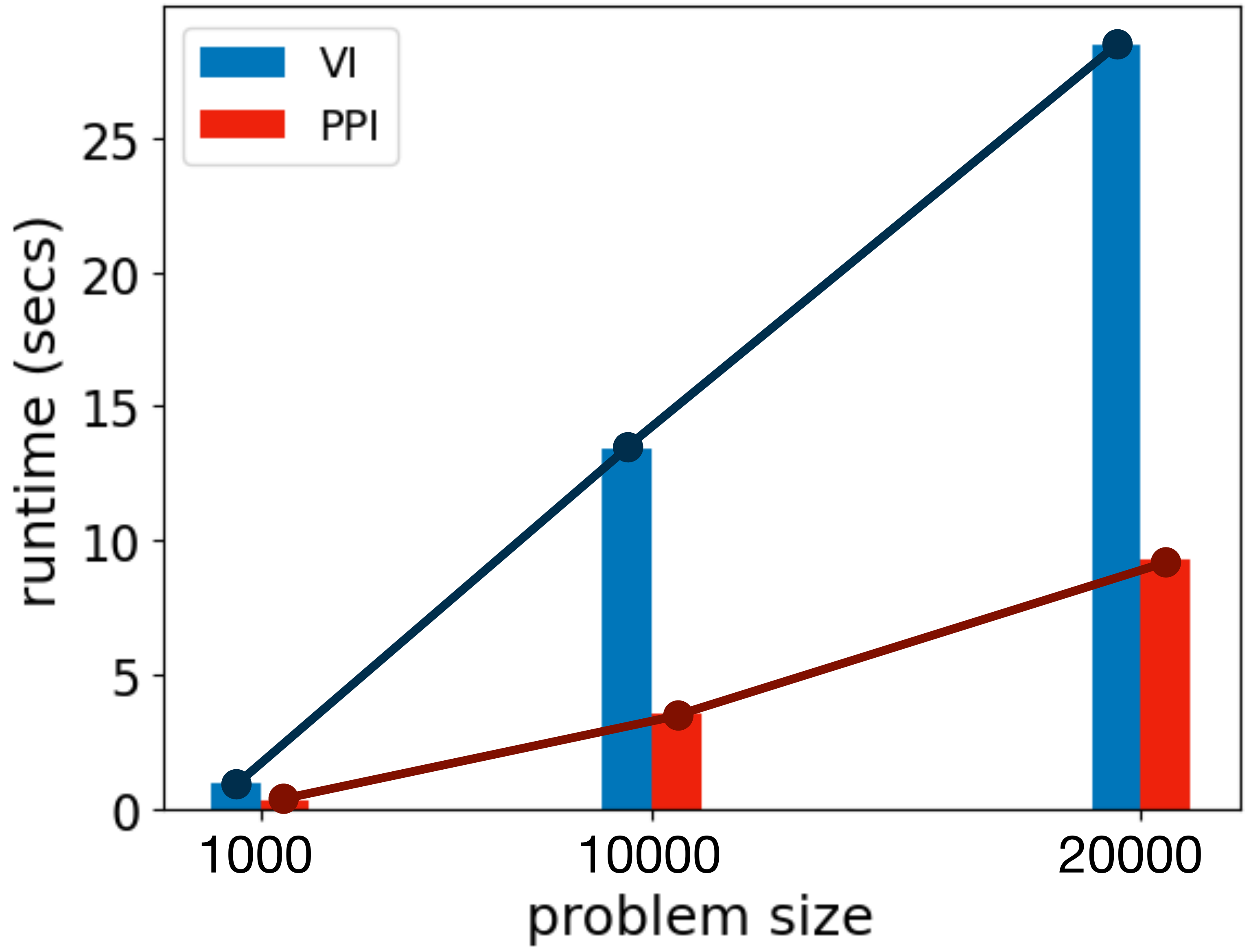


Cart Pole

uniform



weighted



Conclusions: MDPs — Now More Important Than Ever!

Conclusions: MDPs — Now More Important Than Ever!

- Reinforcement learning is changing the world:

Conclusions: MDPs — Now More Important Than Ever!

- Reinforcement learning is changing the world:

ChatGPT Gemini Claude

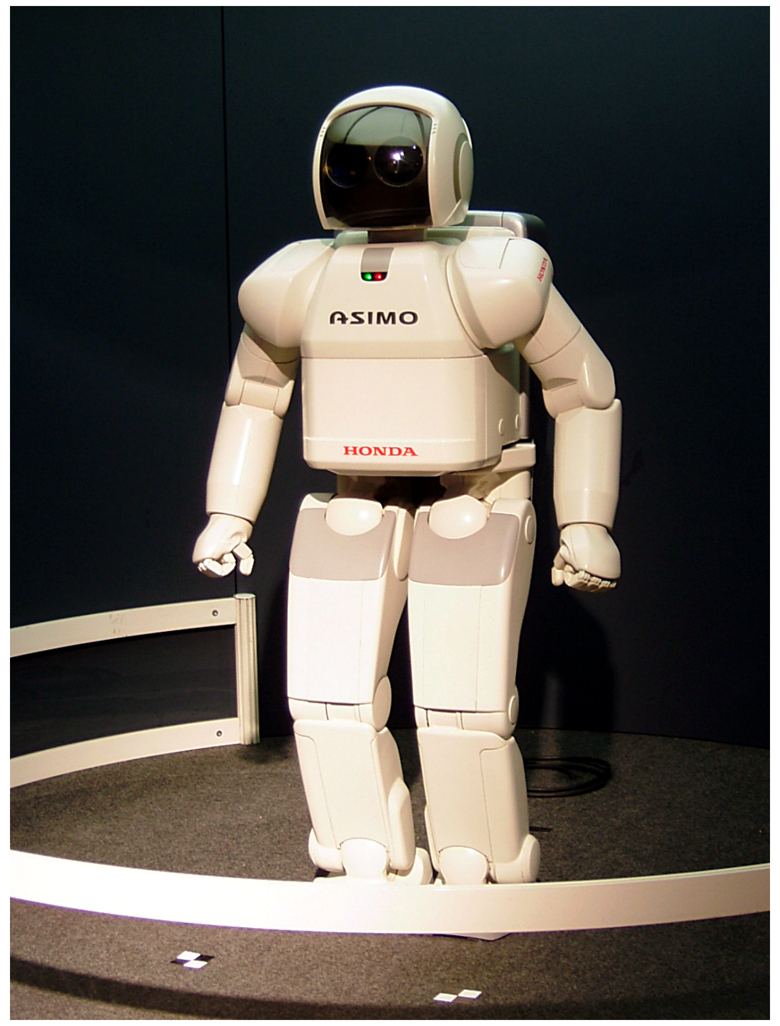
Reinforcement learning with human feedback

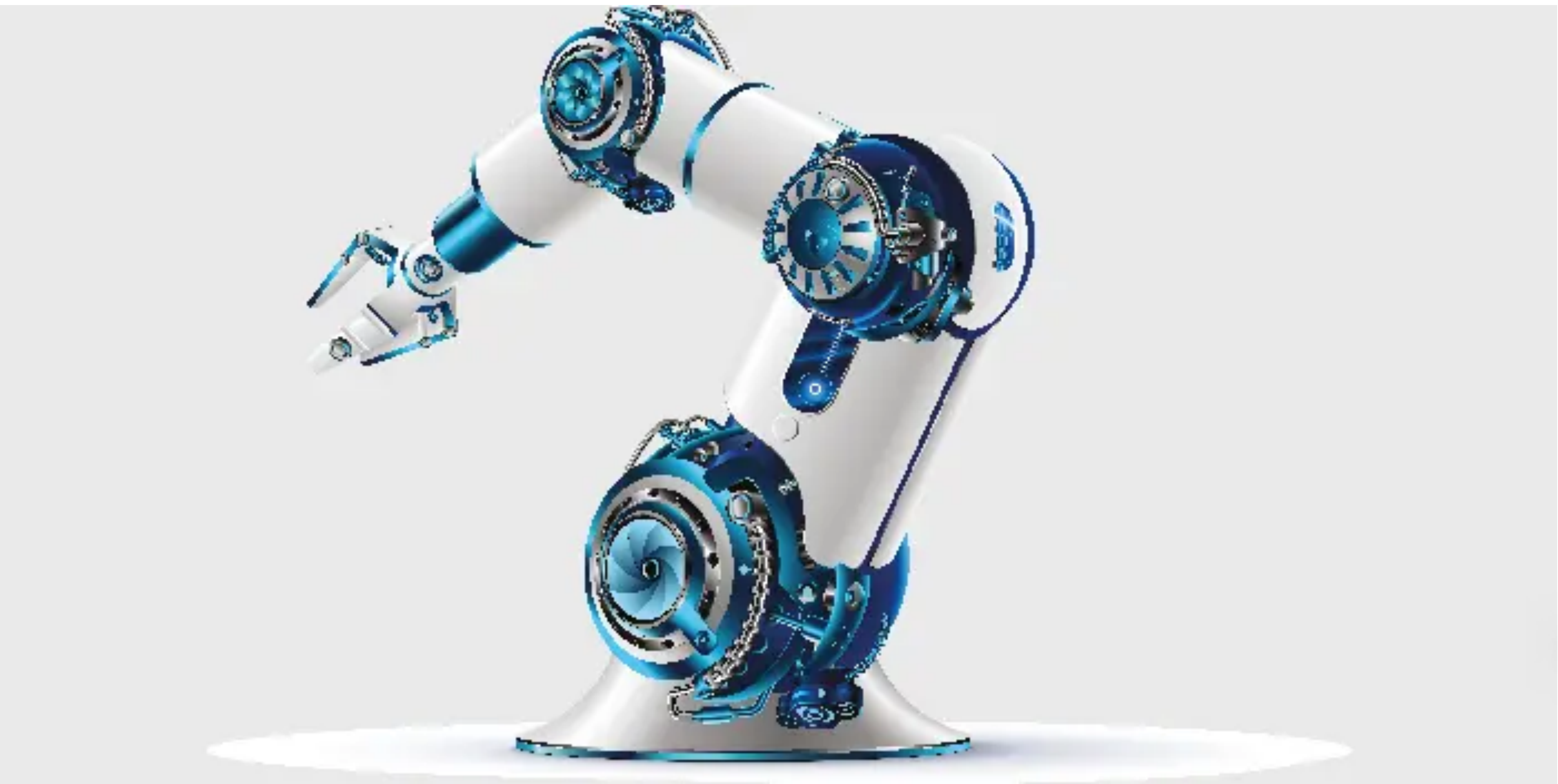
Conclusions: MDPs — Now More Important Than Ever!

- Reinforcement learning is changing the world:

ChatGPT Gemini Claude

Reinforcement learning with human feedback





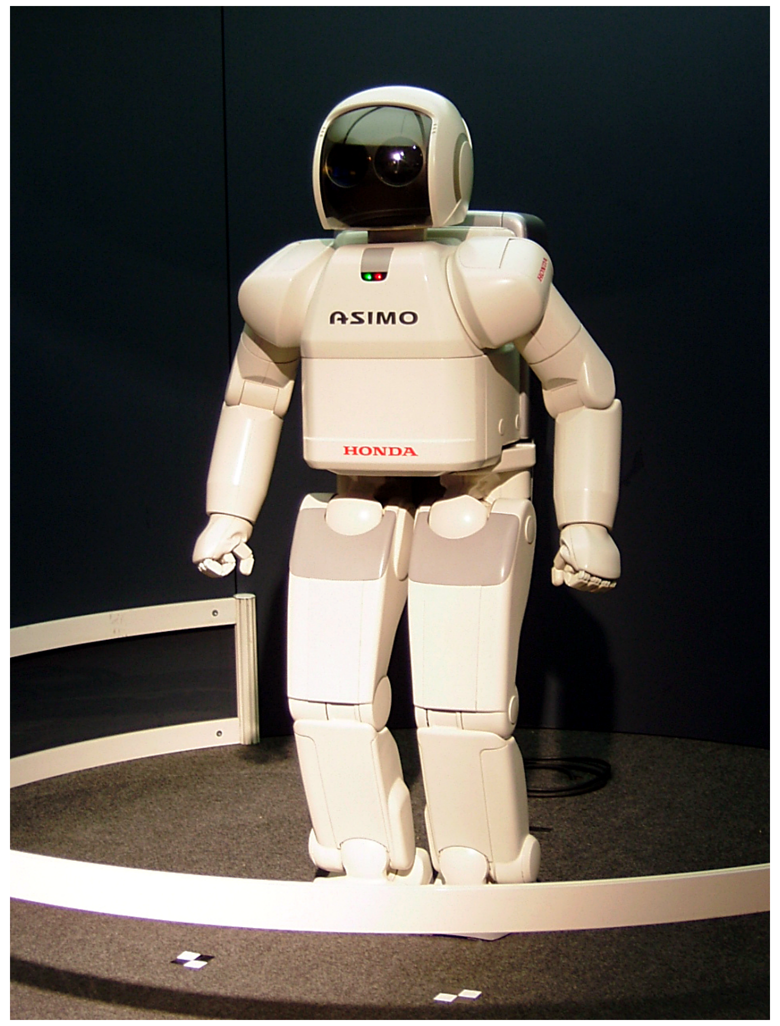
Robotics

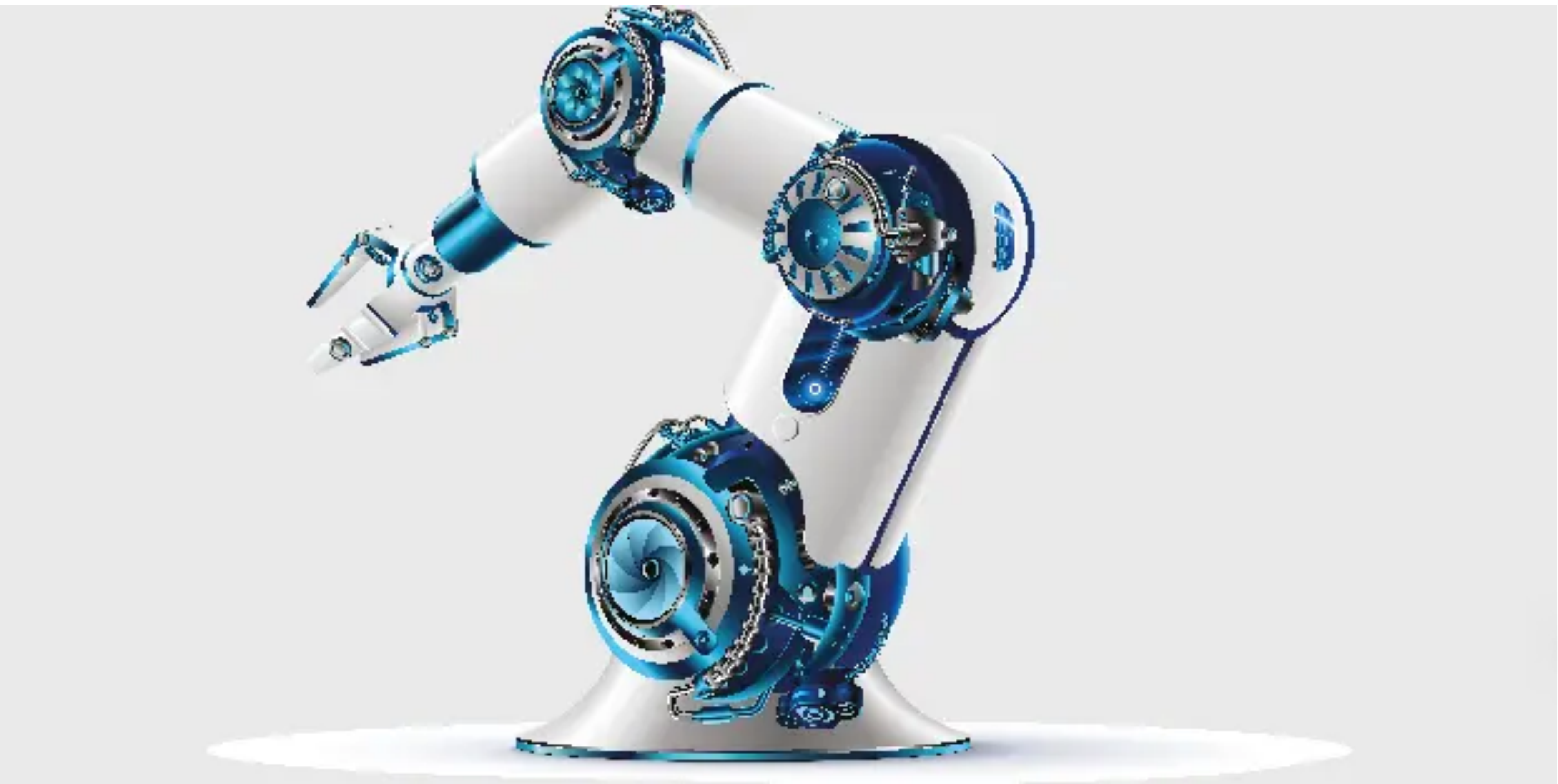
Conclusions: MDPs — Now More Important Than Ever!

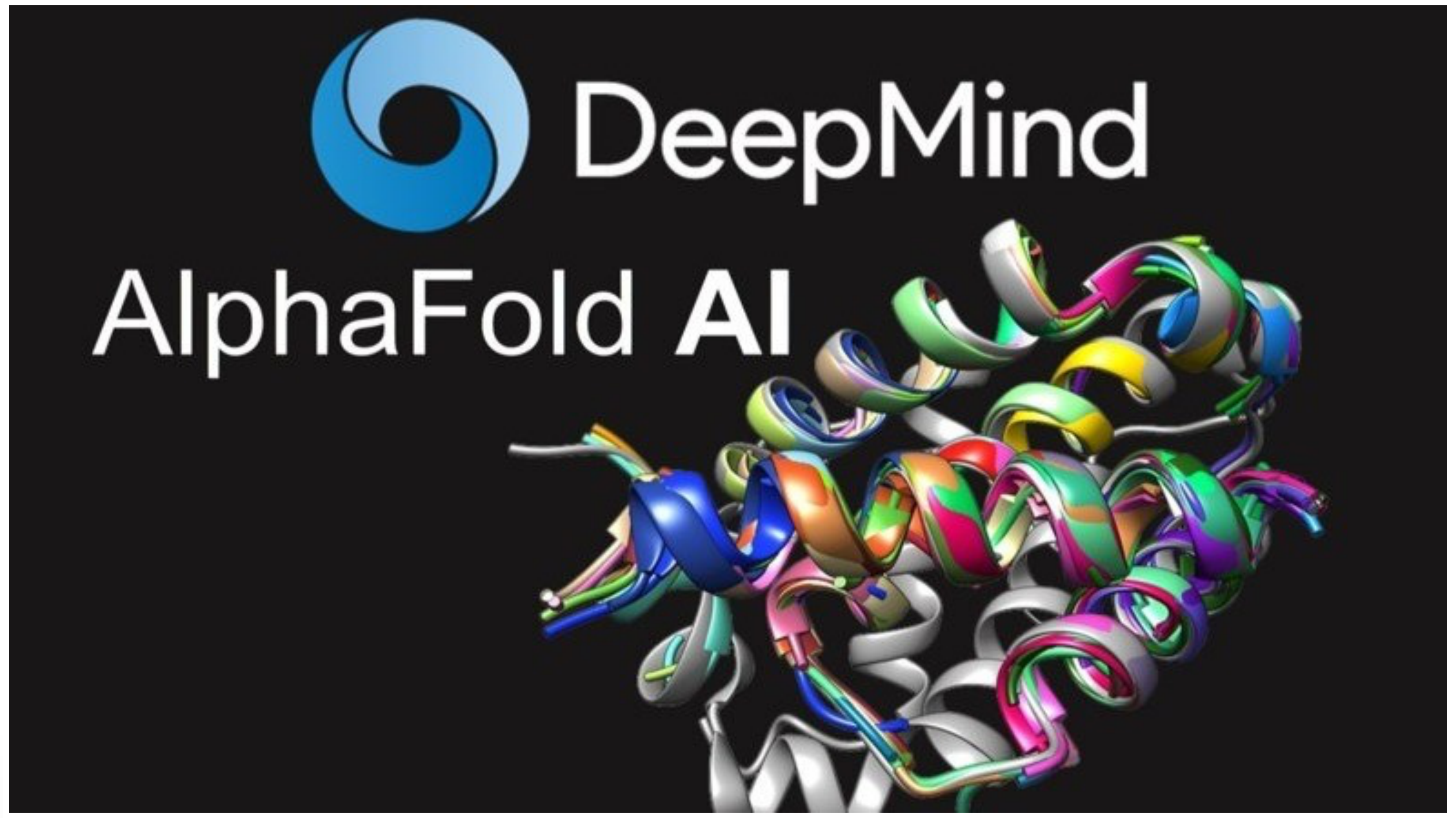
- Reinforcement learning is changing the world:

ChatGPT Gemini Claude

Reinforcement learning with human feedback







Robotics

Medical discovery

Conclusions: MDPs — Now More Important Than Ever!

- Reinforcement learning is changing the world.
- Reinforcement learning ≈ MDPs with learning & value function approximation.

Conclusions: MDPs — Now More Important Than Ever!

- Reinforcement learning is changing the world.
- Reinforcement learning ≈ MDPs with learning & value function approximation.
- Real-world value function approximations (e.g. via deep learning) are largely “black magic” without rigorous analysis.

Conclusions: MDPs — Now More Important Than Ever!

- Reinforcement learning is changing the world.
- Reinforcement learning ≈ MDPs with learning & value function approximation.
- Real-world value function approximations (e.g. via deep learning) are largely "black magic" without rigorous analysis.
- MDPs give us the mathematical understanding that underpins all this.

Conclusions: MDPs — Now More Important Than Ever!

- Reinforcement learning is changing the world.
- Reinforcement learning ≈ MDPs with learning & value function approximation.
- Real-world value function approximations (e.g. via deep learning) are largely "black magic" without rigorous analysis.
- MDPs give us the mathematical understanding that underpins all this.
- There are exciting new developments that bring both fields closer:
  - Linear MDPs: Reinforcement learning with linear dynamics
  - Weakly Coupled MDPs, Factored MDPs, Constrained MDPs, Safe RL, …
  - covered extensively in top AI/ML conferences in recent years!

Bibliography

[1] WW, D. Kuhn, B. Rustem, Robust Markov Decision Processes, *Mathematics of Operations Research* 38(1):153-183, 2013.

[2] C. Ho, M. Petrik, WW, Fast Bellman Updates for Robust MDPs, *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

[3] C. Ho, M. Petrik, WW, Partial Policy Iteration for L1-Robust Markov Decision Processes, *The Journal of Machine Learning Research* 22(1):12612-12657, 2021.

[4] C. Ho, M. Petrik, WW, Robust Phi-Divergence MDPs, *Advances in Neural Information Processing Systems 35 (NeurIPS Proceedings)*, 2023.

[5] C. Ho, M. Petrik, WW, Efficient Algorithms for Robust Markov Decision Processes with *s*-Rectangular Ambiguity Sets. *Under Review*, 2026.

ww@imperial.ac.uk