Data-Driven Markov Decision Processes

Wolfram Wiesemann
Imperial College Business School

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 $p(s' \mid s, a)$;
- $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are the expected one-step rewards;
- $\lambda \in (0,1)$ is the discount factor.

Markov decision process

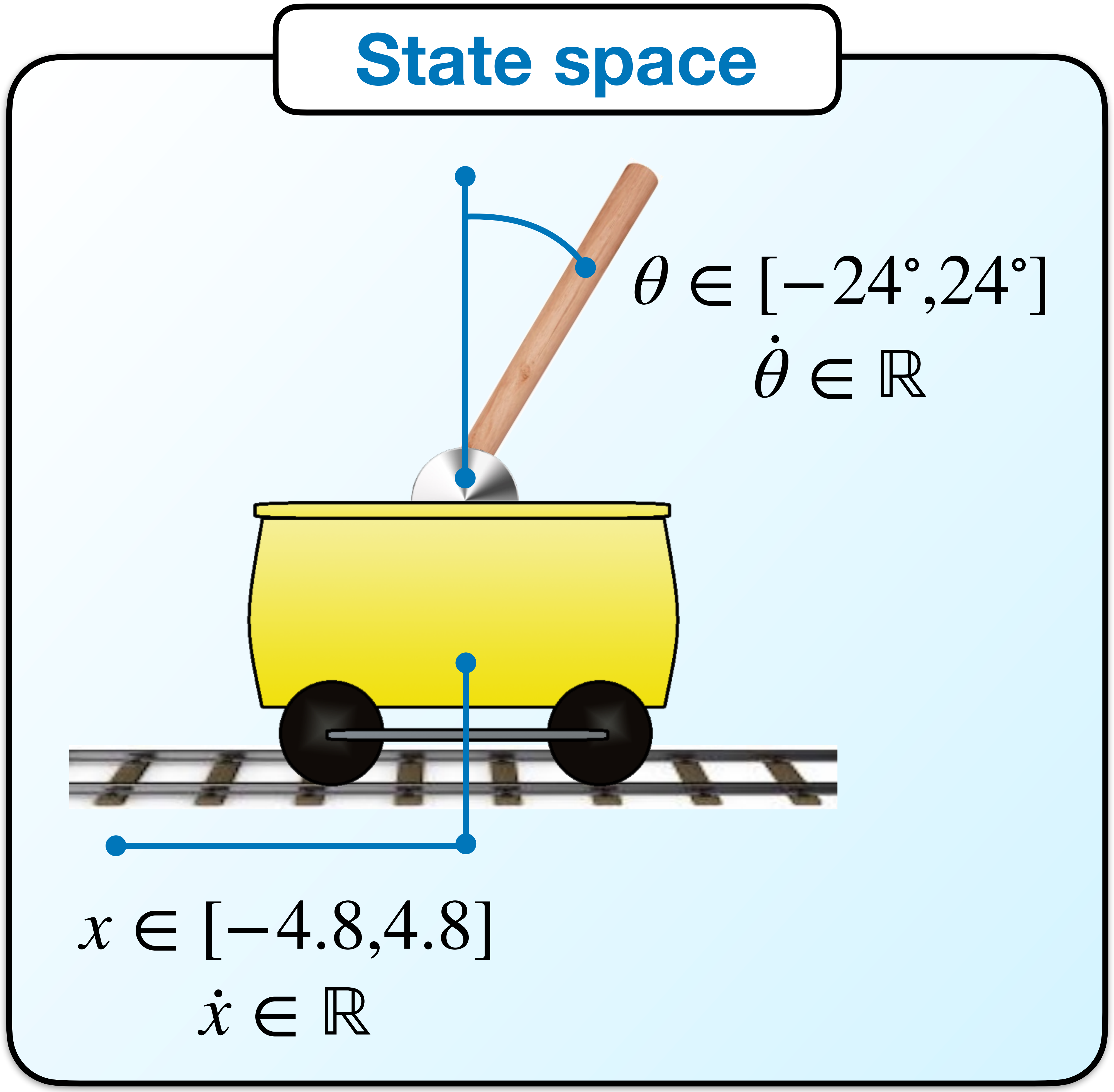
Tuple $(\mathscr{S}, \mathscr{A}, q, p, r, \lambda)$ where

- $\mathscr{S} = \{1, \ldots, S\}$ 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

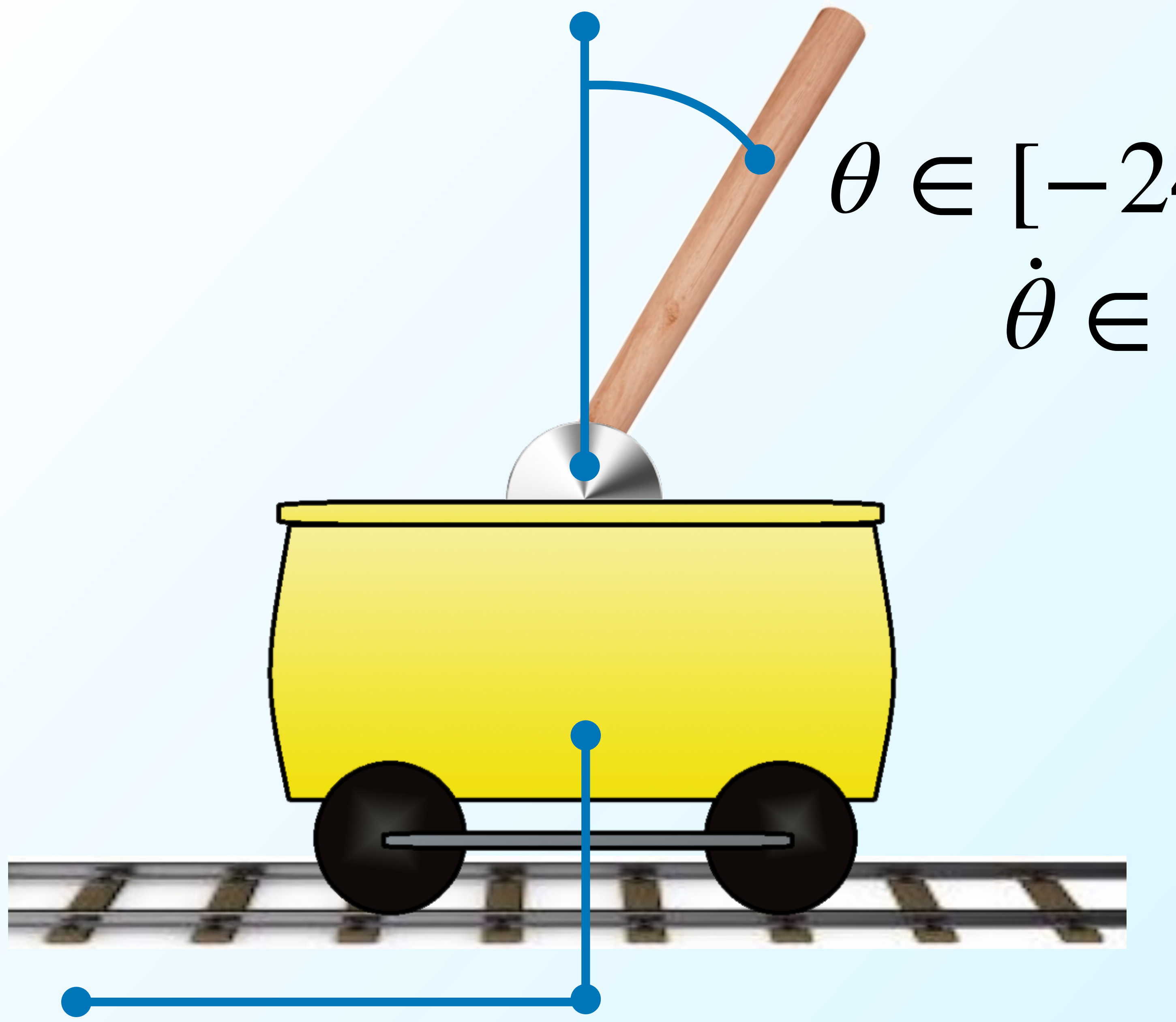
find policy $\pi : \mathscr{S} \to \mathscr{A}$ that maximizes the expected total discounted rewards:

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

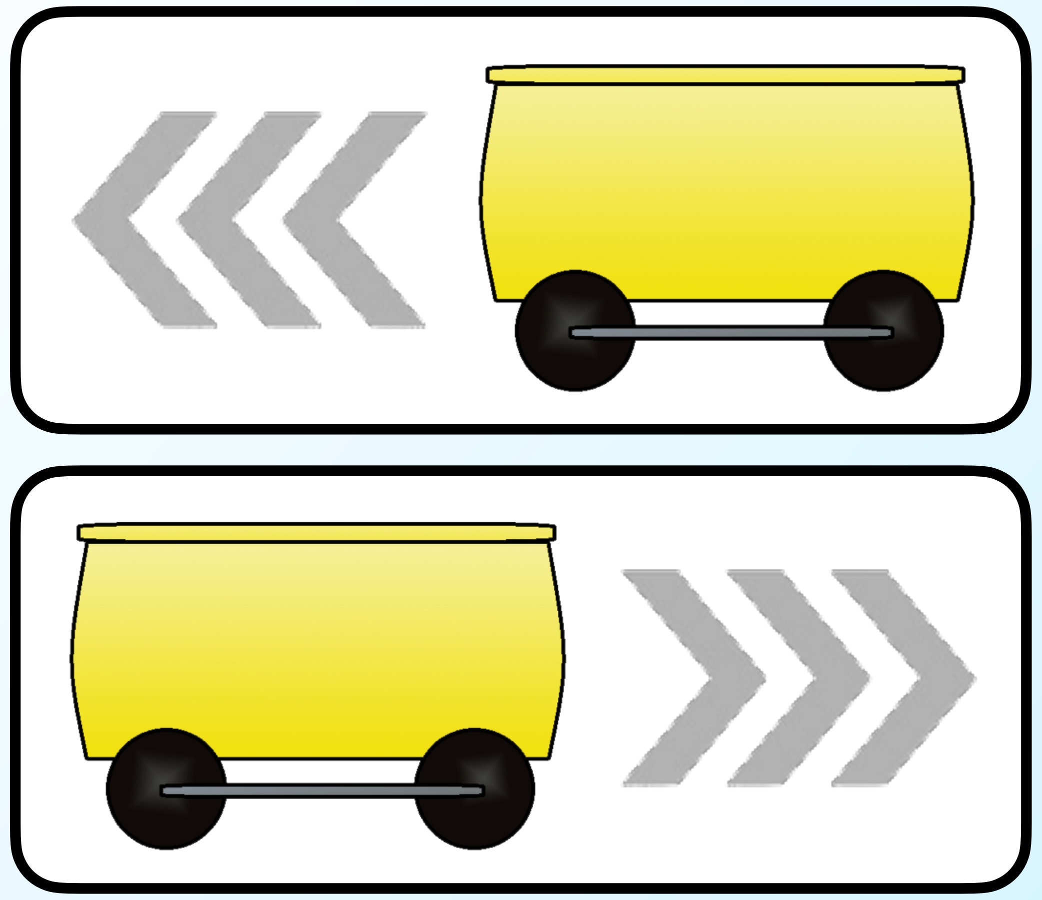


Barto et al. (1983), *Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems*.

State space

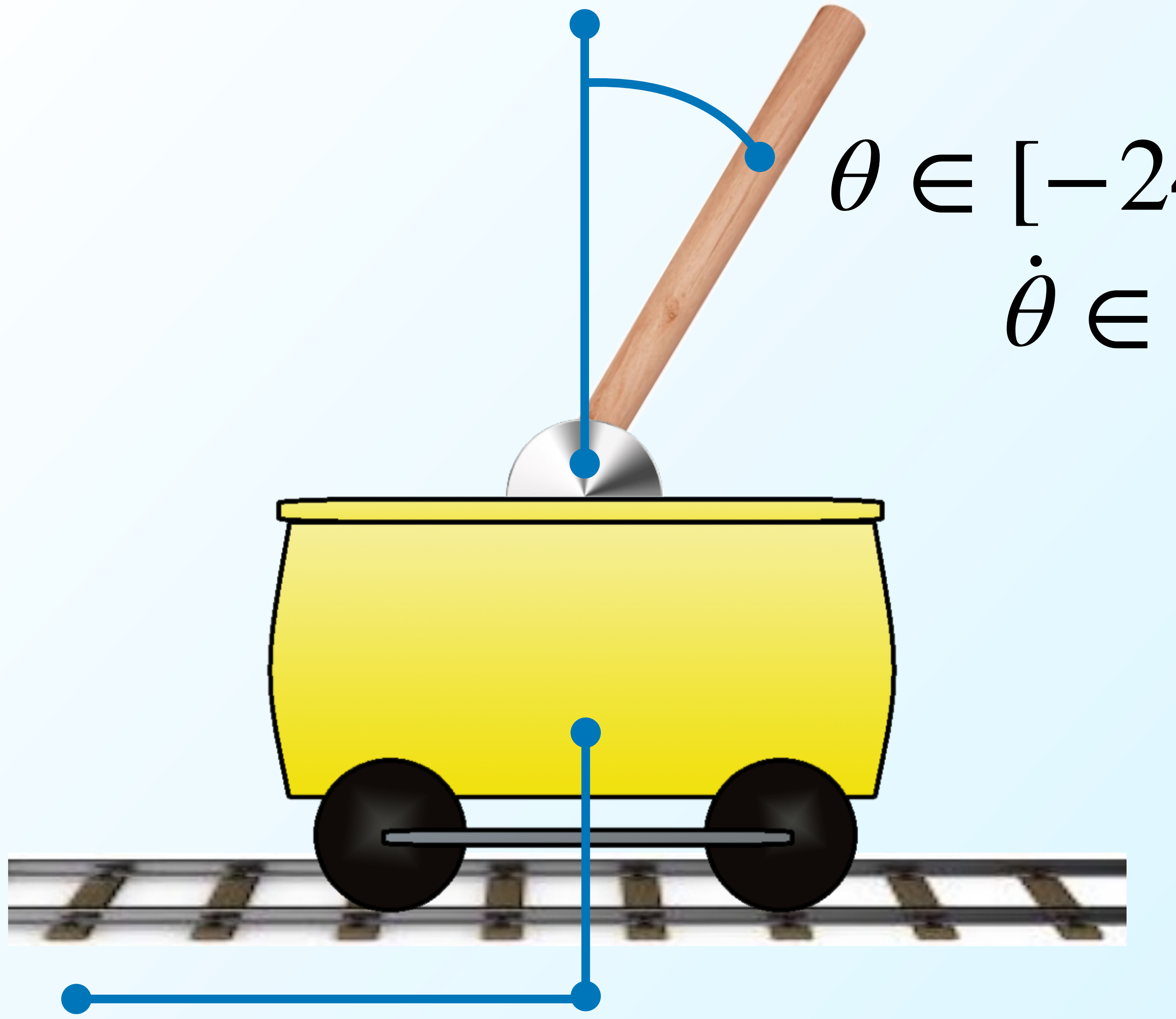


Action space



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

State space



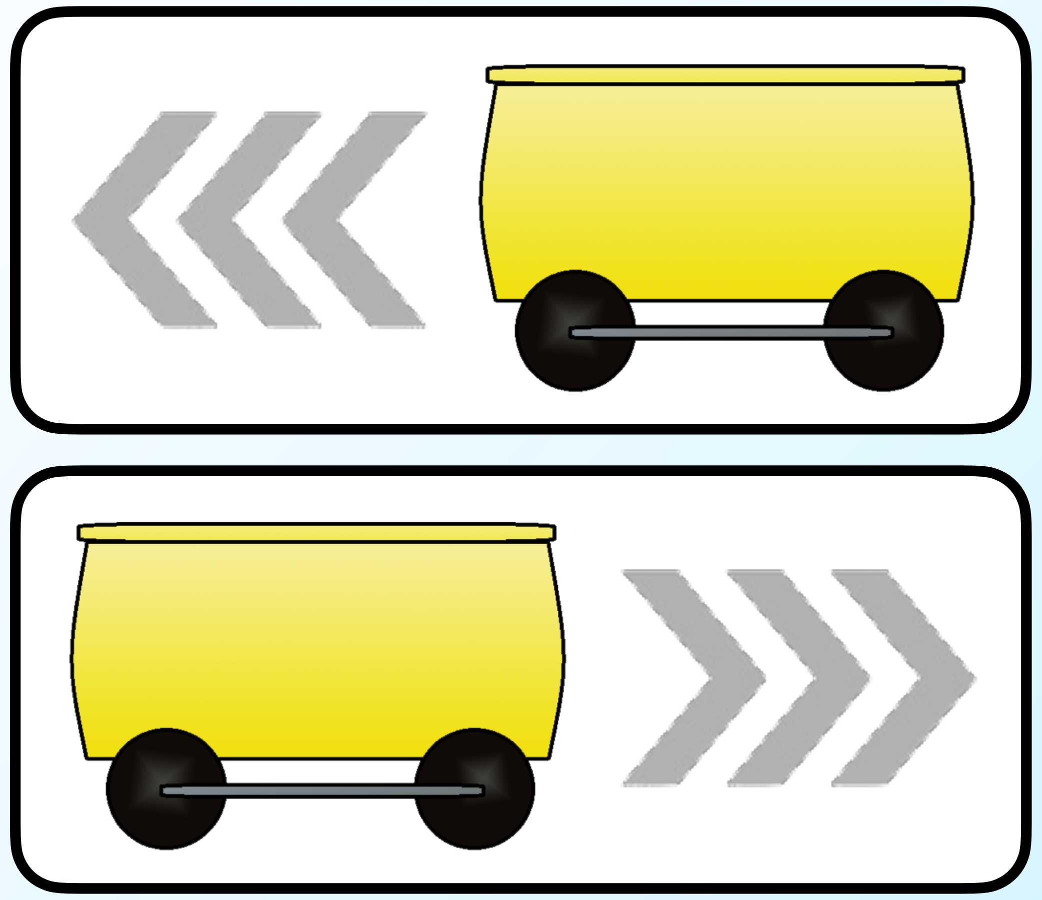
$\theta \in [-24°, 24°]$

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

$x \in [-4.8, 4.8]$

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

Action space



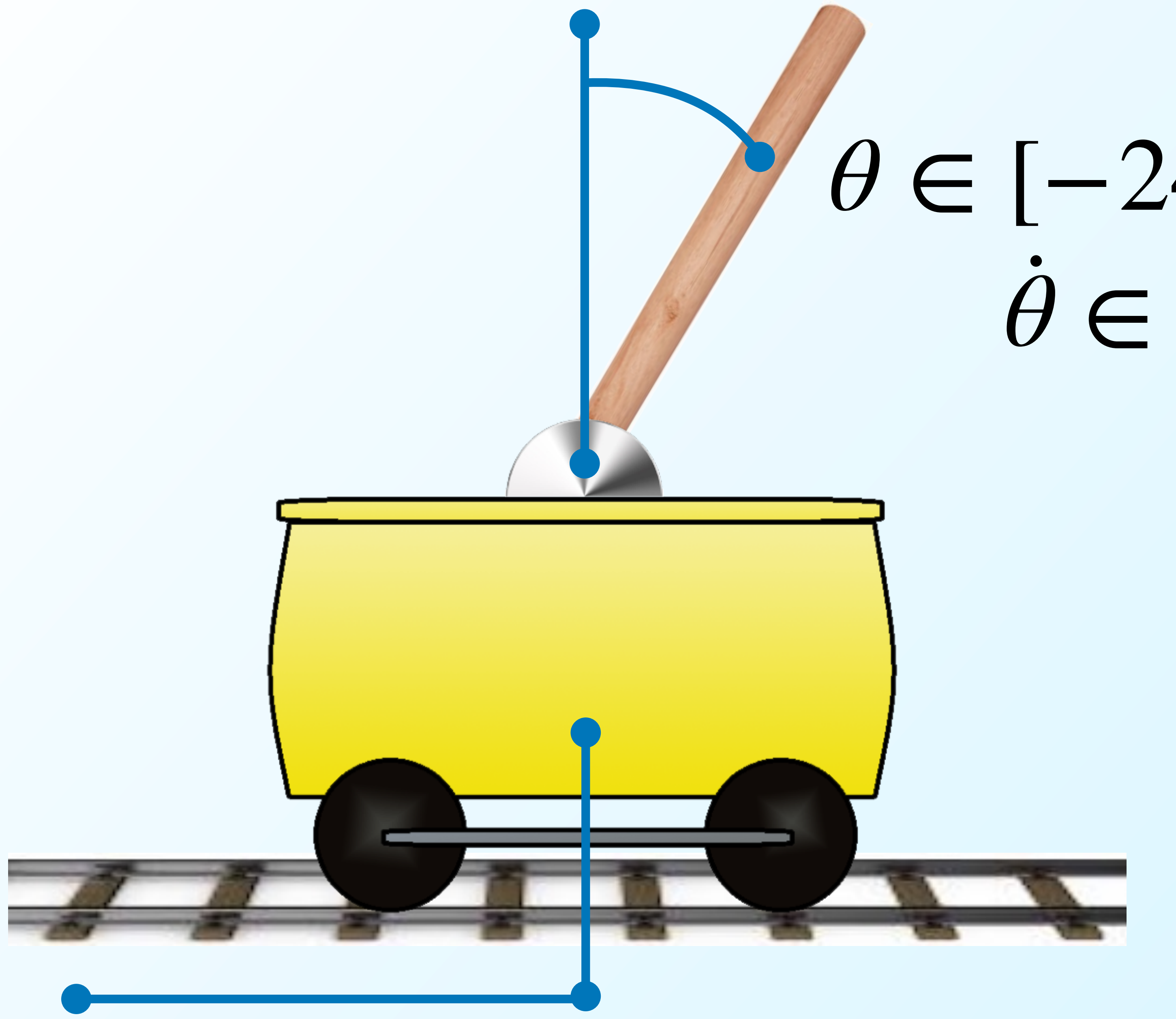
Initial state

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

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

Cart Pole Example

State space



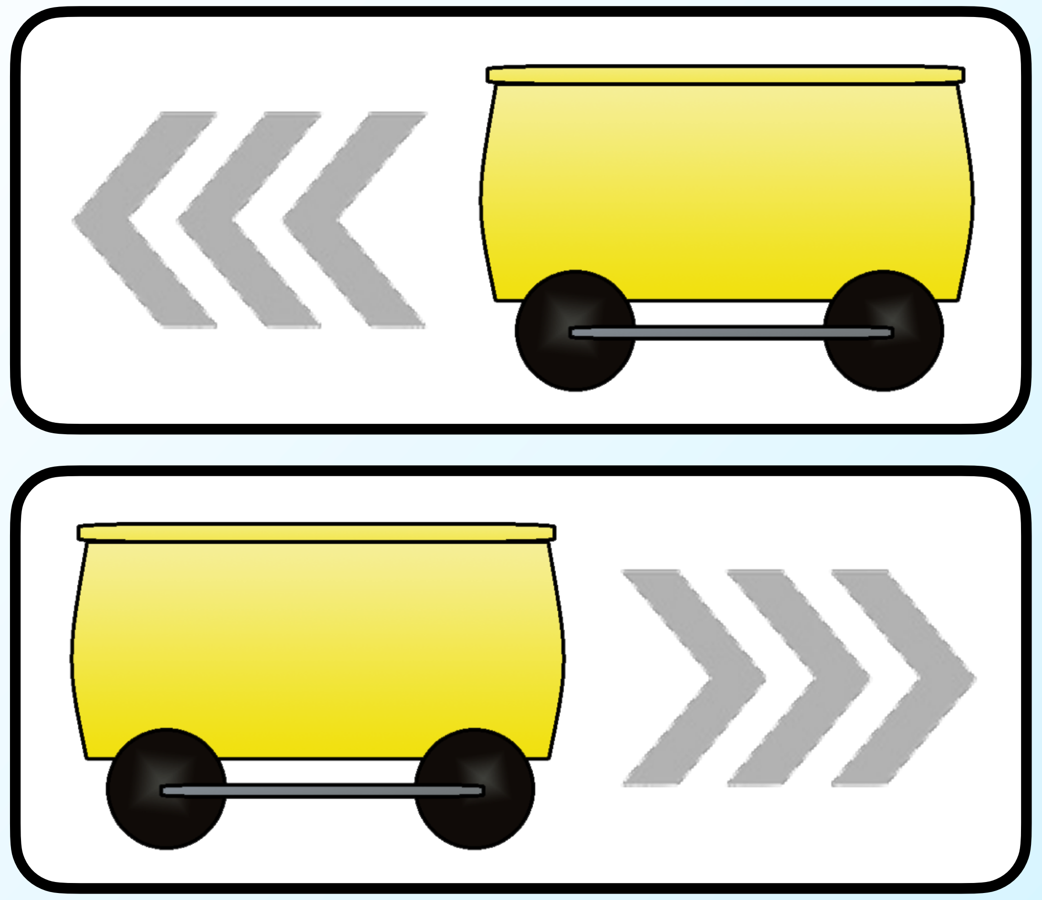
$\theta \in [-24°, 24°]$

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

$x \in [-4.8, 4.8]$

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

Action space



Transitions

- deterministic via laws of mechanics
- terminate if

$x \notin [-2.4, 2.4]$

or $\theta \notin [-12°, 12°]$

Initial state

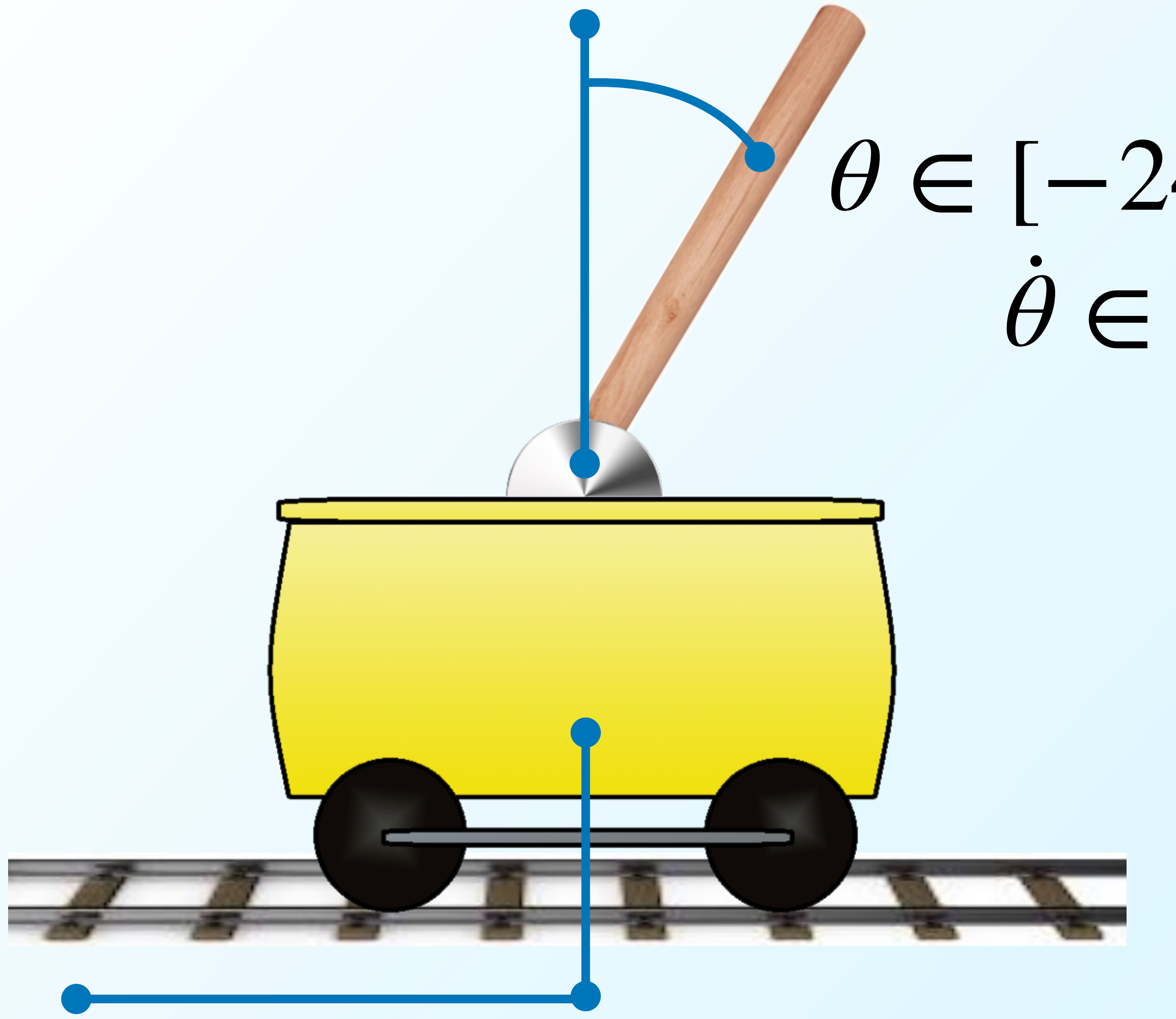
$x, \dot{x}, \theta, \dot{\theta} \sim$

$\mathcal{U}[-0.05, 0.05]$

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

Cart Pole Example

State space



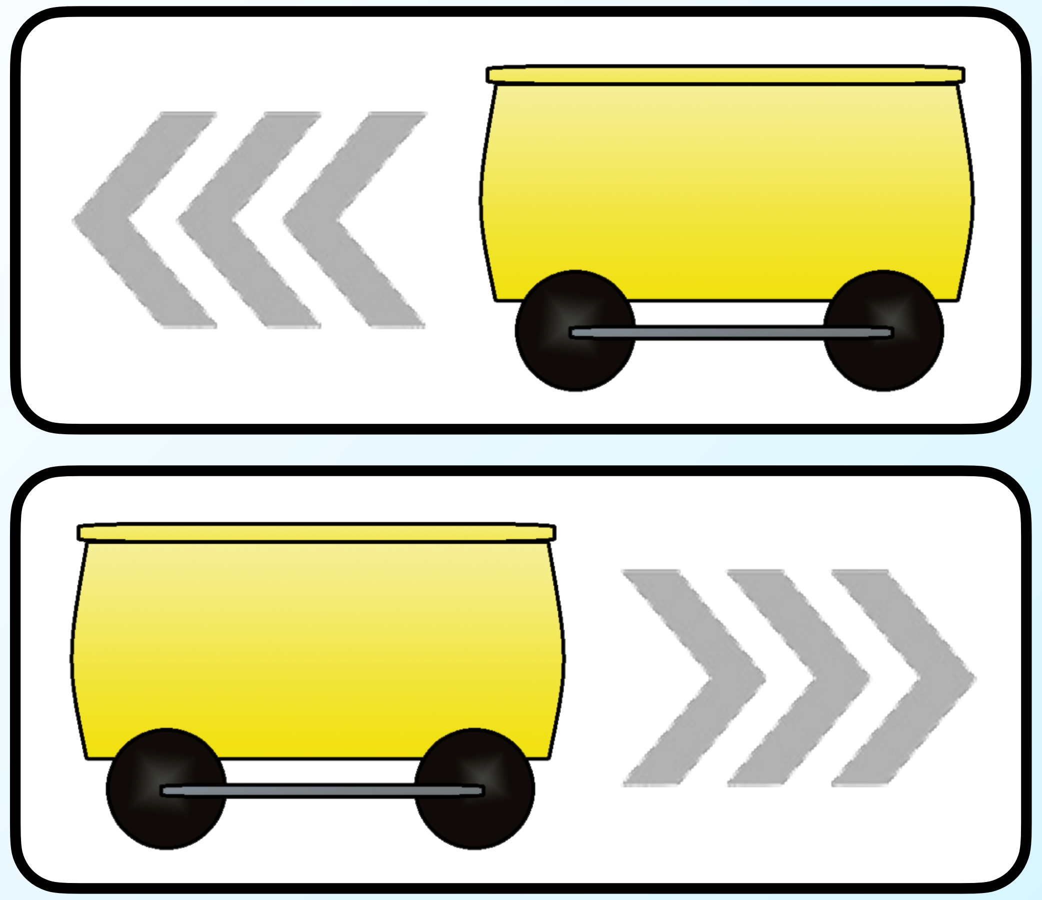
$\theta \in [-24°, 24°]$

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

$x \in [-4.8, 4.8]$

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

Action space



Transitions

- deterministic via laws of mechanics
- terminate if

$x \notin [-2.4, 2.4]$

or $\theta \notin [-12°, 12°]$

Initial state

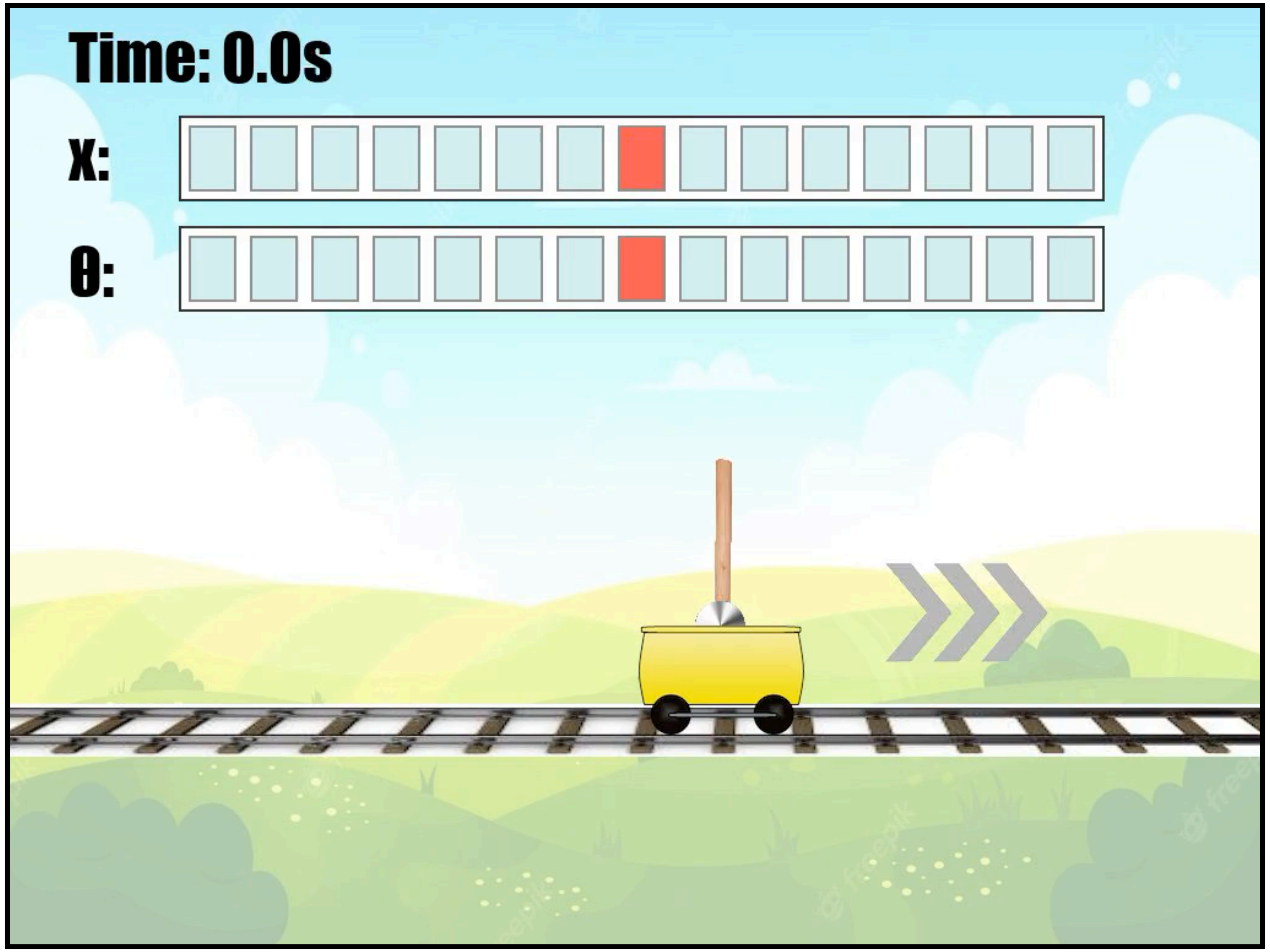
$x, \dot{x}, \theta, \dot{\theta} \sim \mathcal{U}[-0.05, 0.05]$

Rewards

+1/non-terminated time step

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

Cart Pole Example



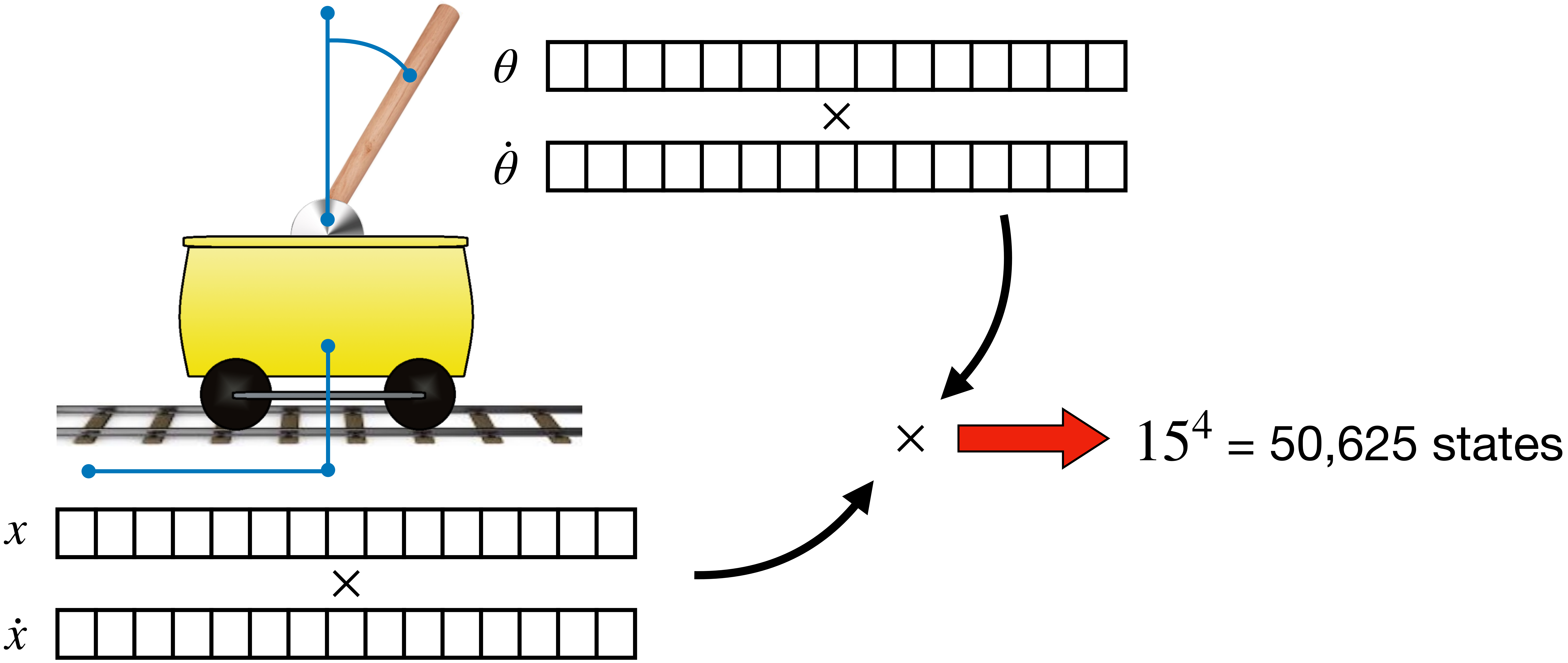
Barto et al. (1983), *Neuronlike Adaptive Elements that can Solve Difficult Learning Control Problems*.

STOCHASTICITY AND AMBIGUITY

Ambiguity and Robust MDPs

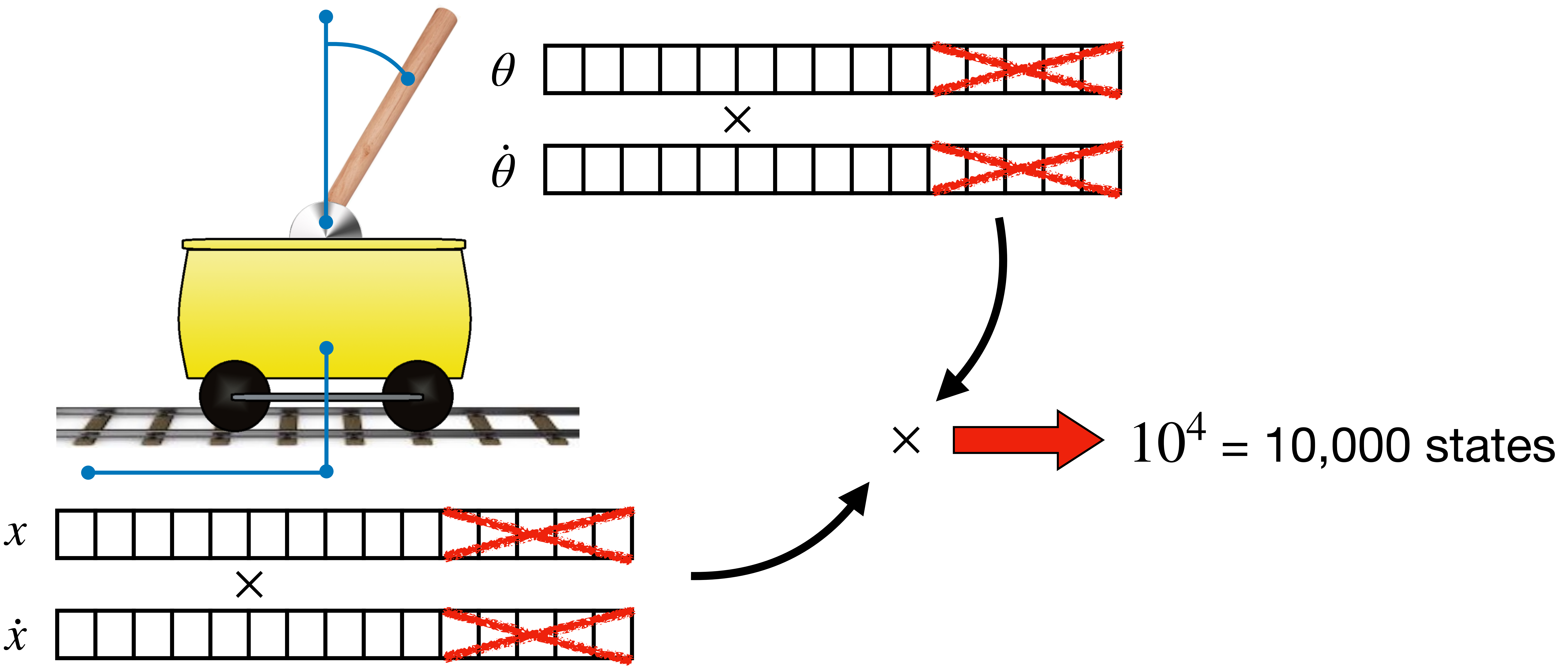
Two common sources of ambiguity:

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



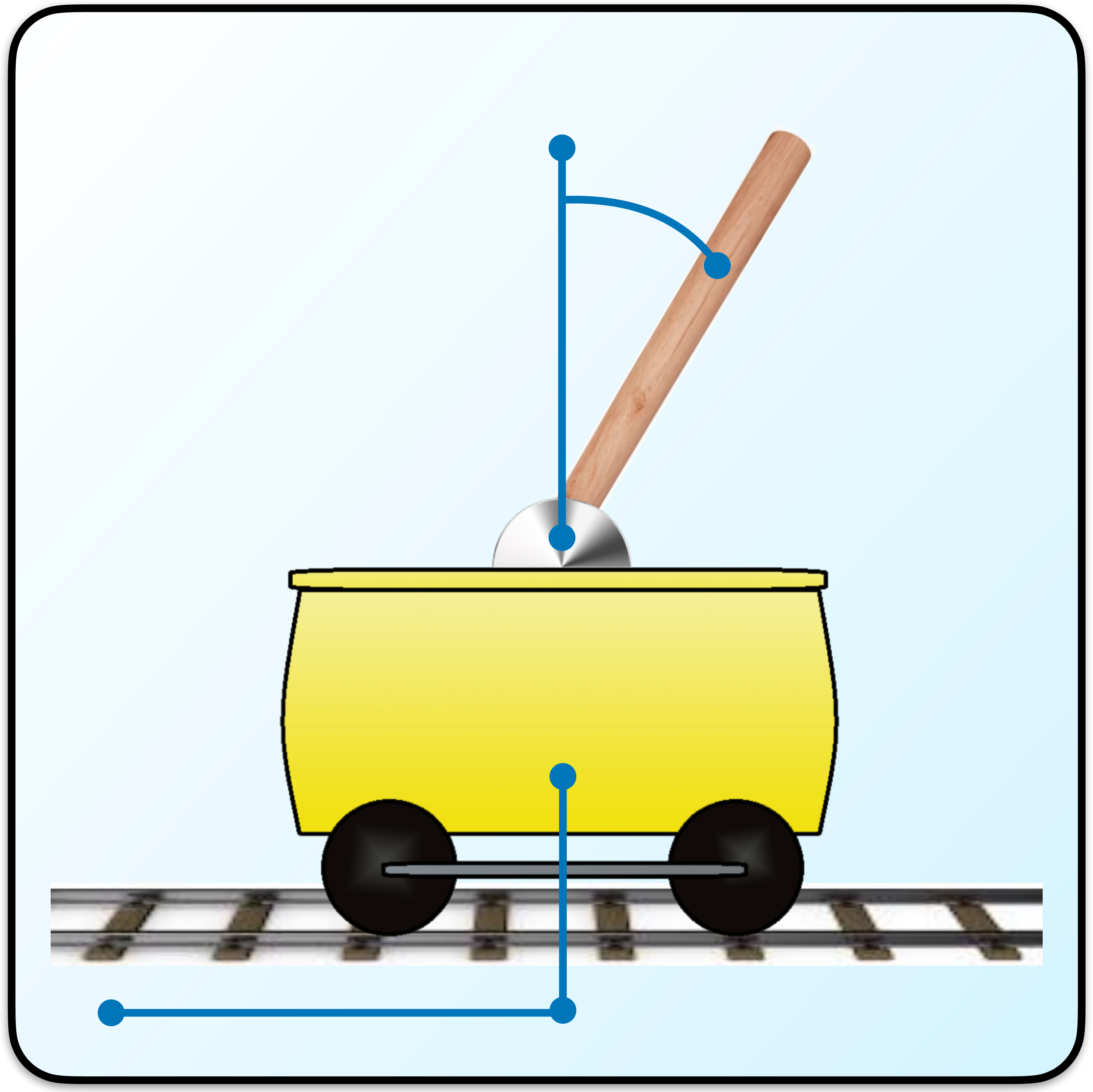
Two common sources of ambiguity:

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



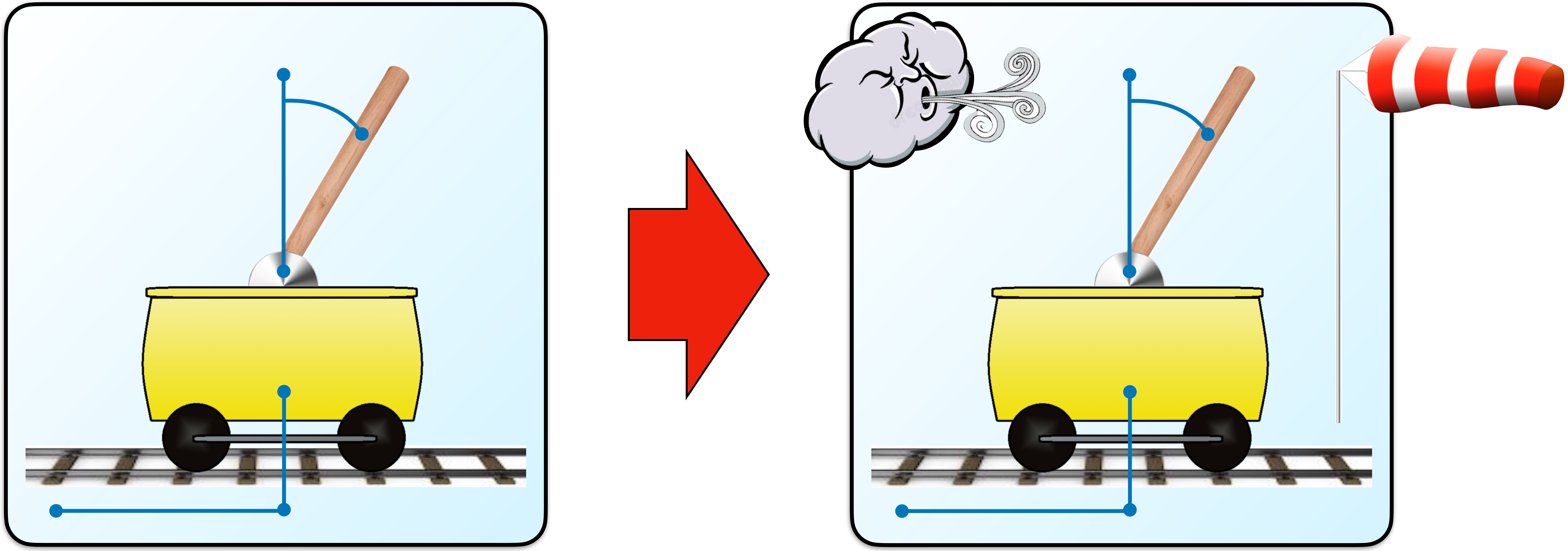
Two common sources of ambiguity:

- **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

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

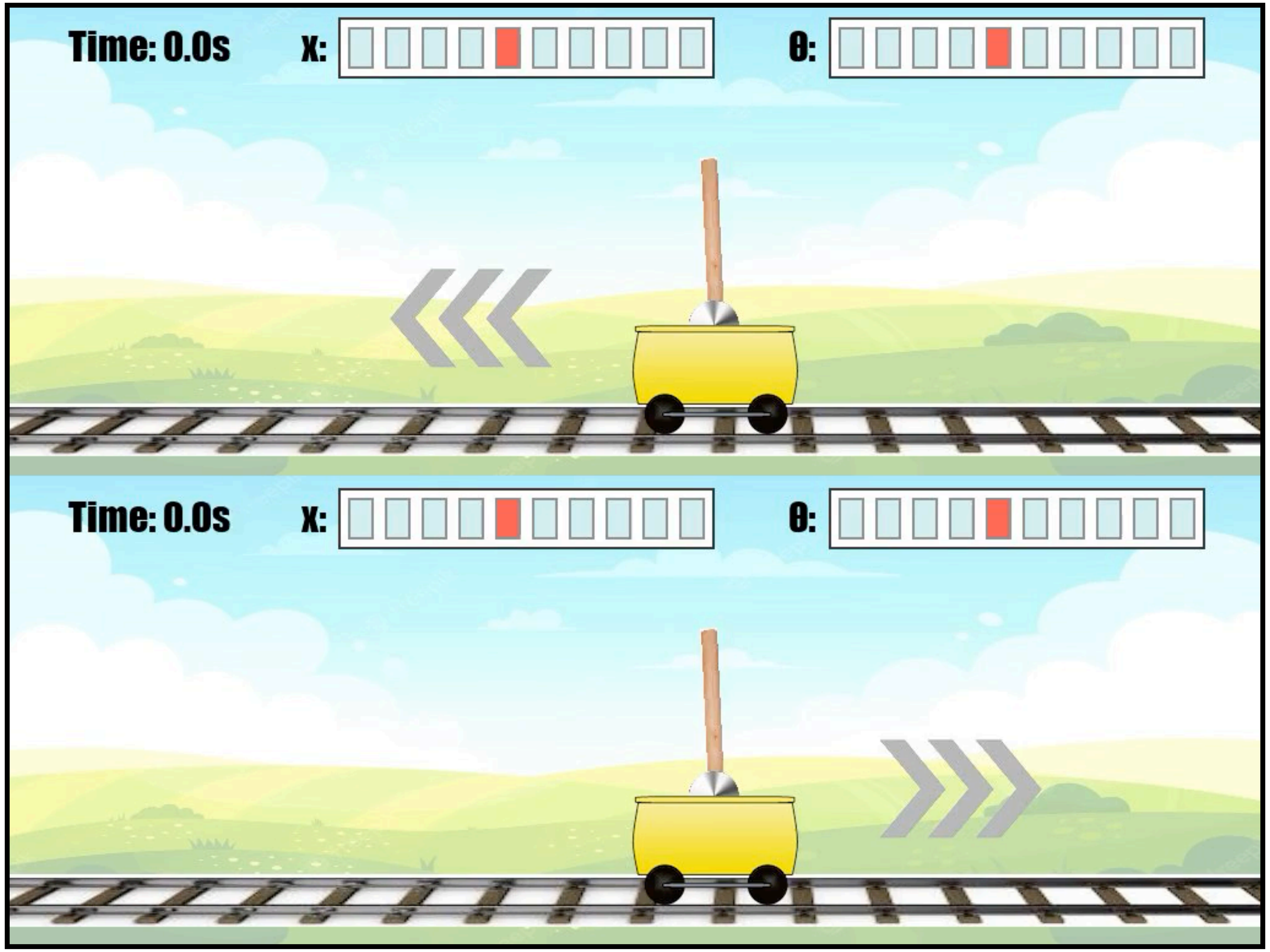
$\mathscr{P}$: ***ambiguity set***

Robust MDPs admit interpretation as regularized MDPs!

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

nominal

robust

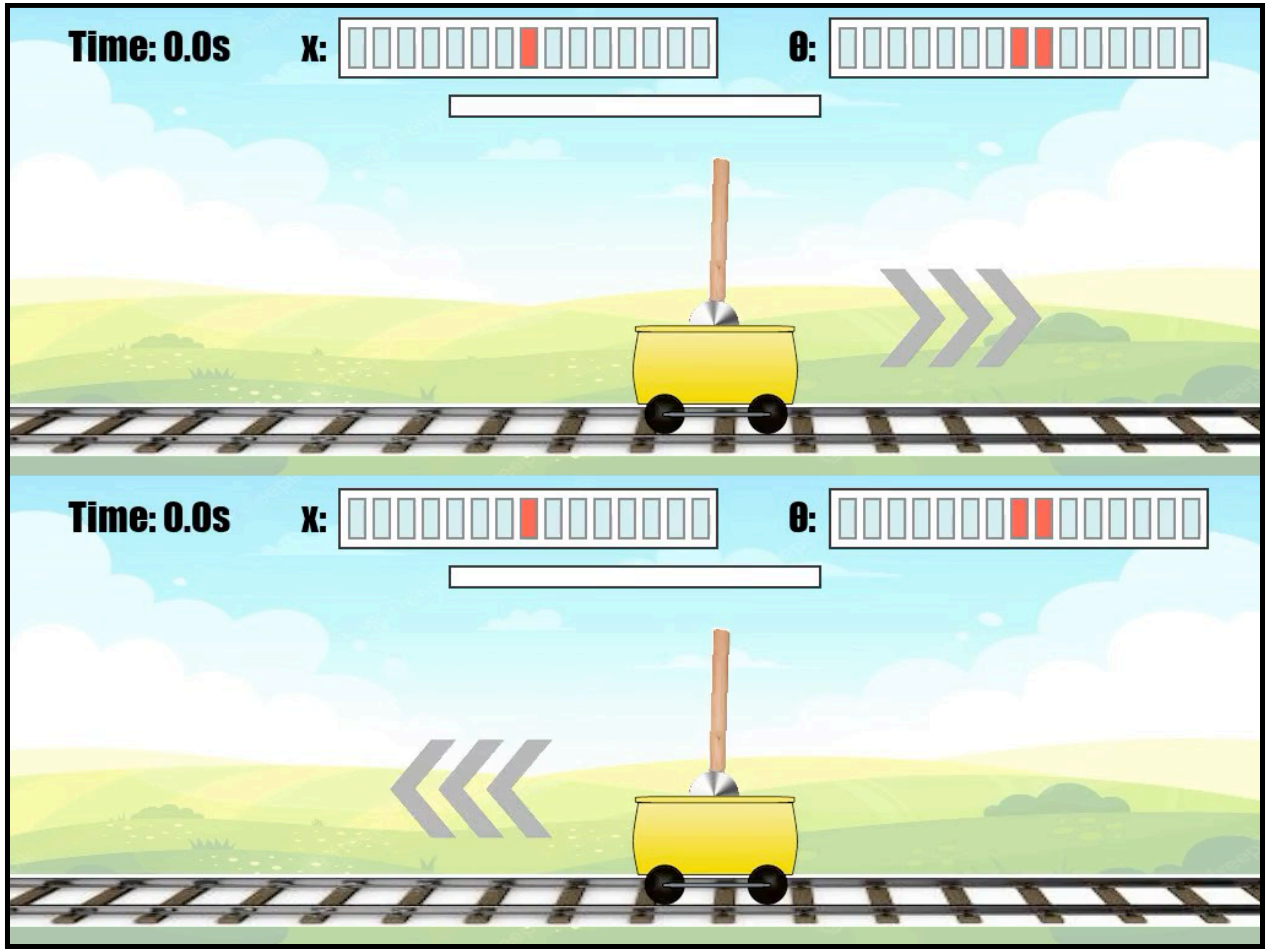


Modelling errors: 32.67 secs/run → 2.45 secs/run → 15.77 secs/run

Ambiguity: Estimation Errors

nominal

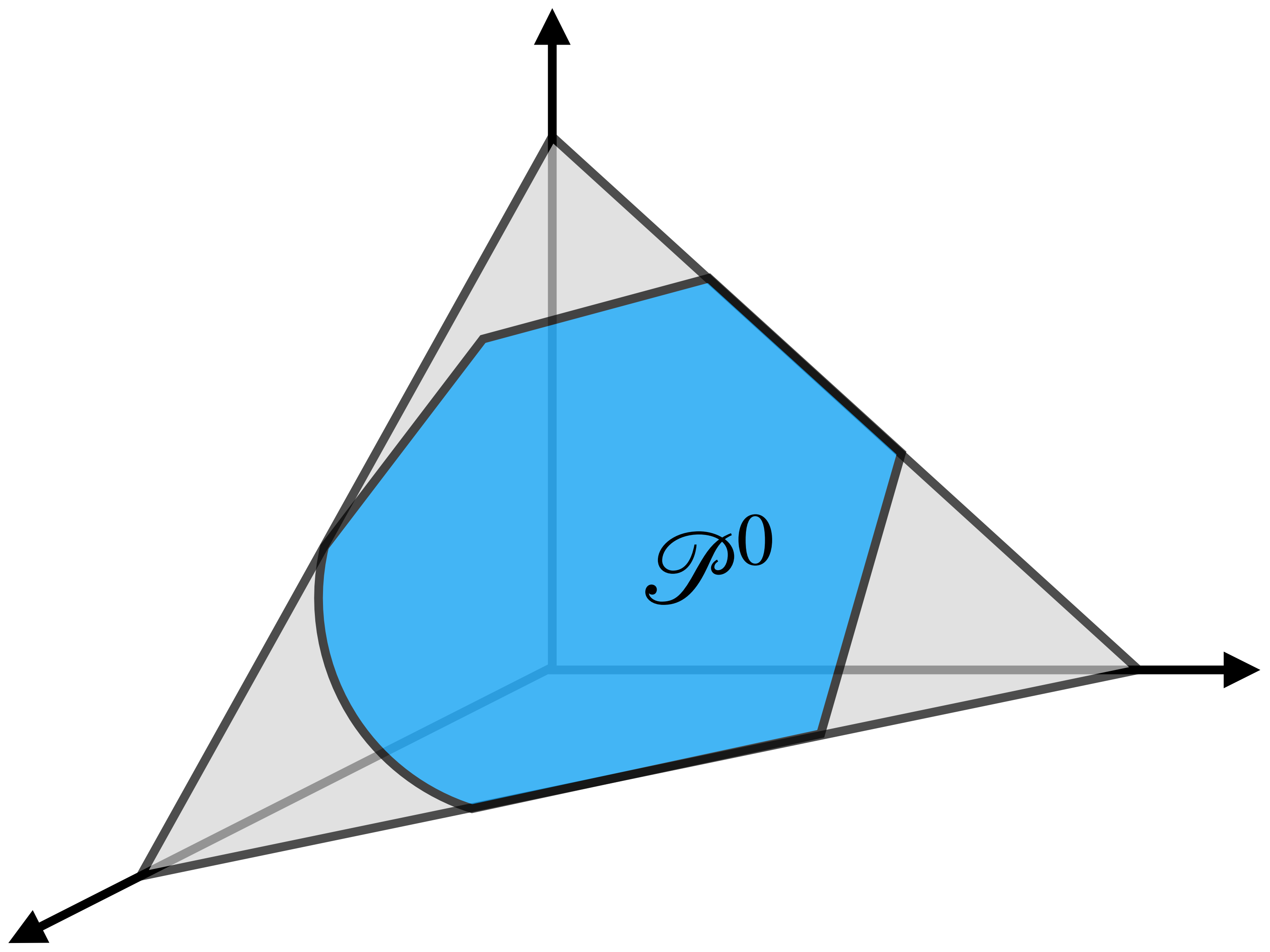
robust



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

Ambiguity Sets

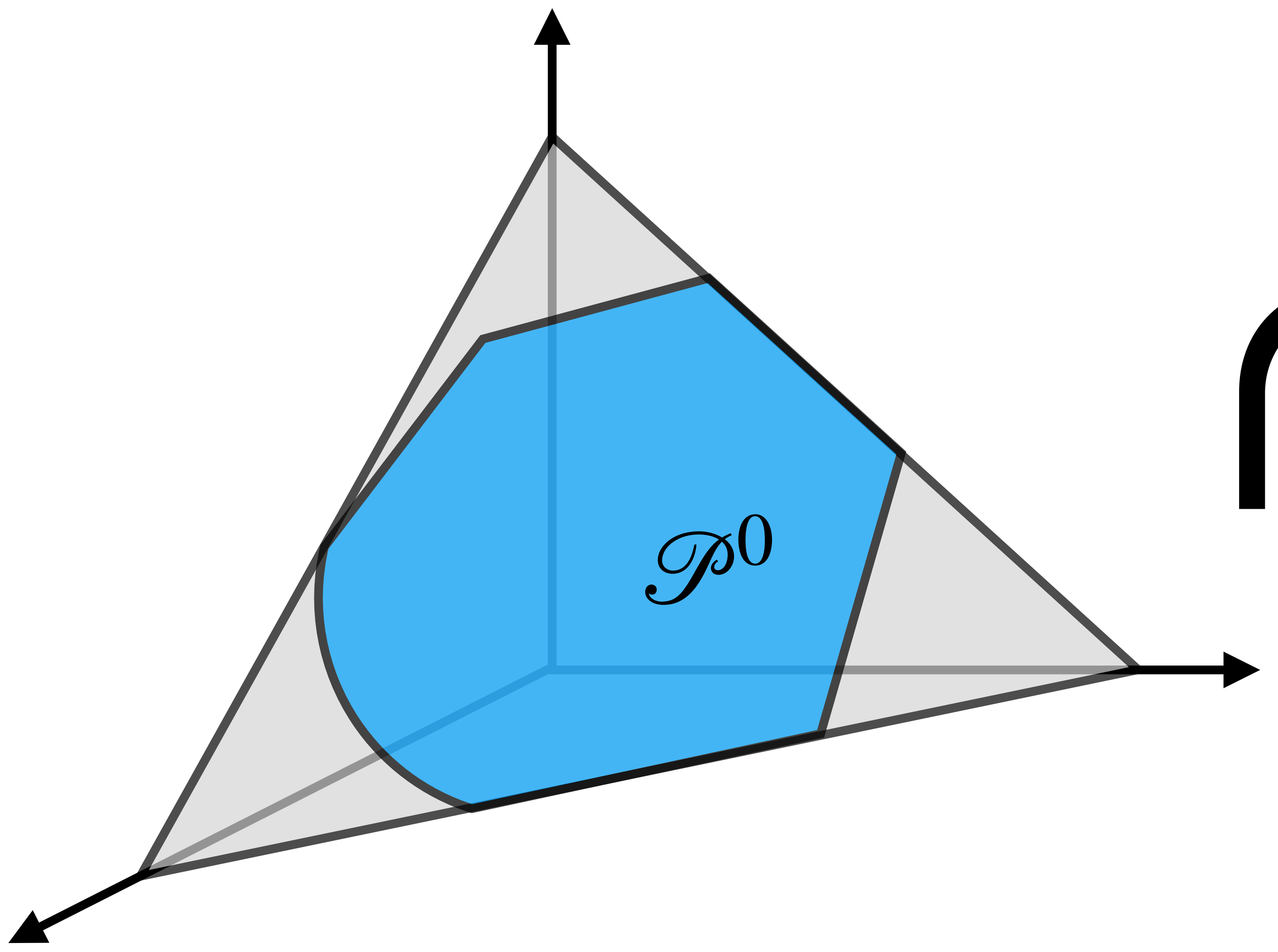
Structural ambiguity set



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

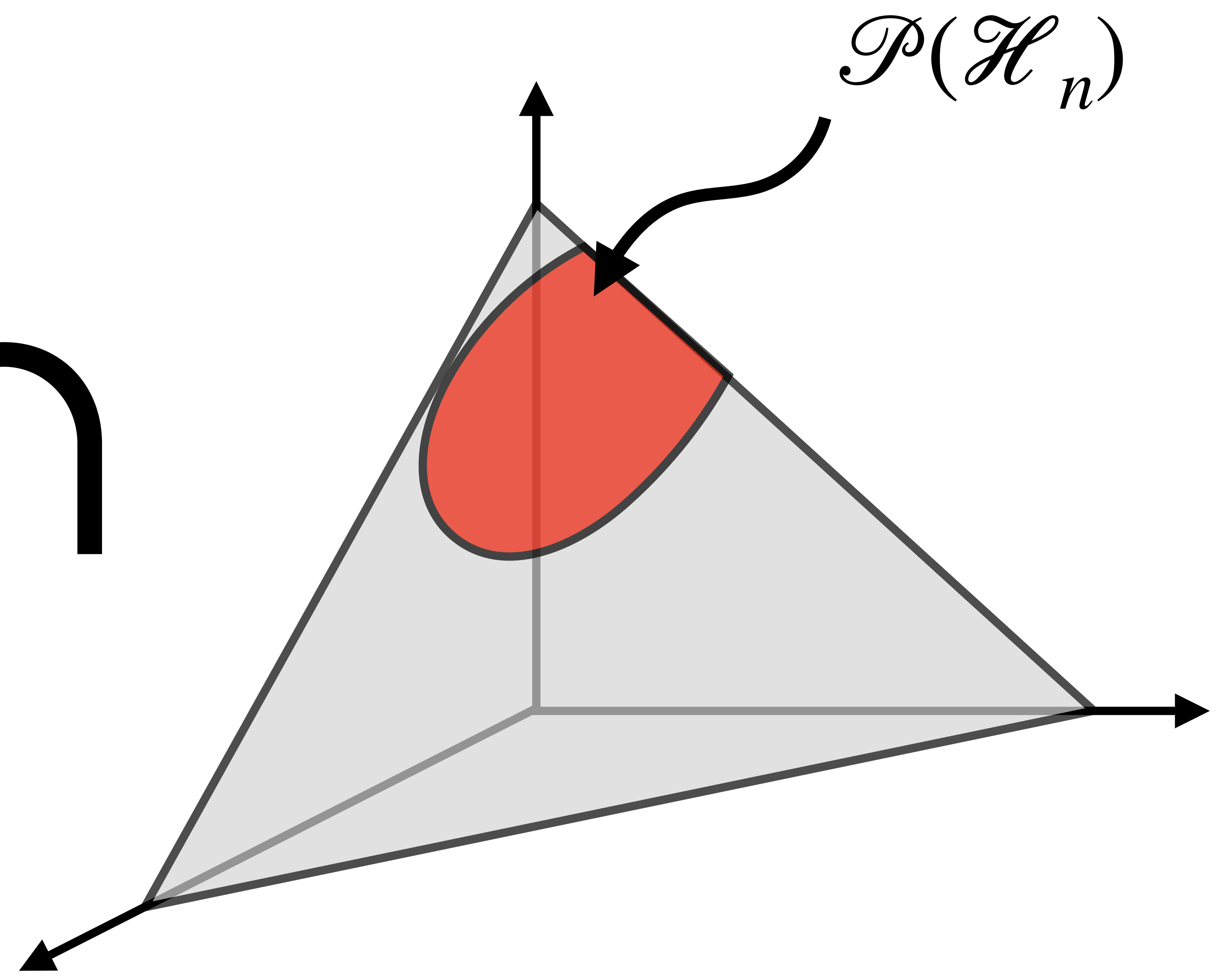
Ambiguity Sets

Structural ambiguity set



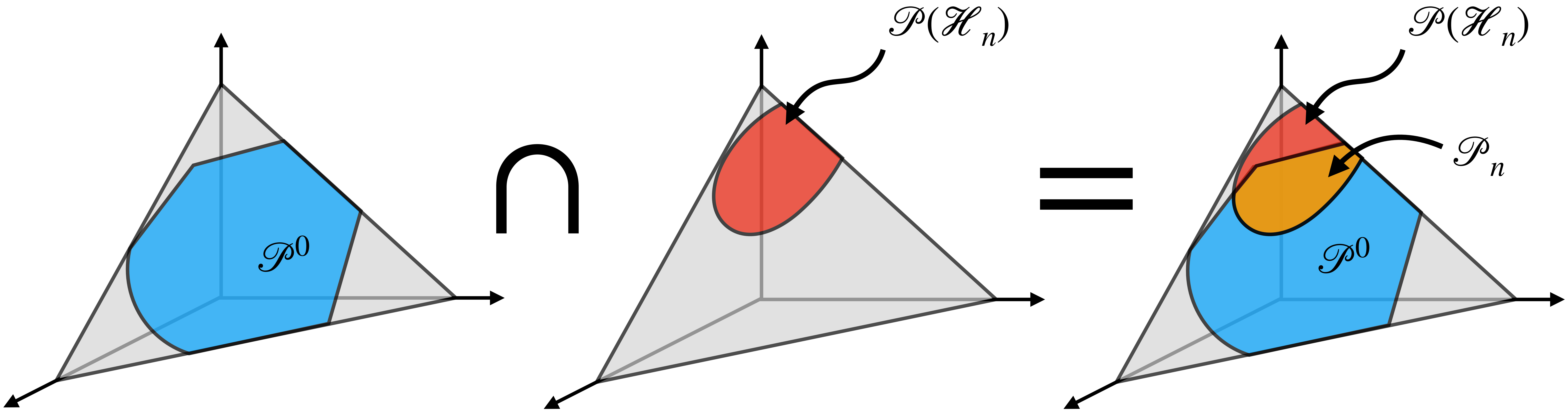
$\cap$

Historical sample



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

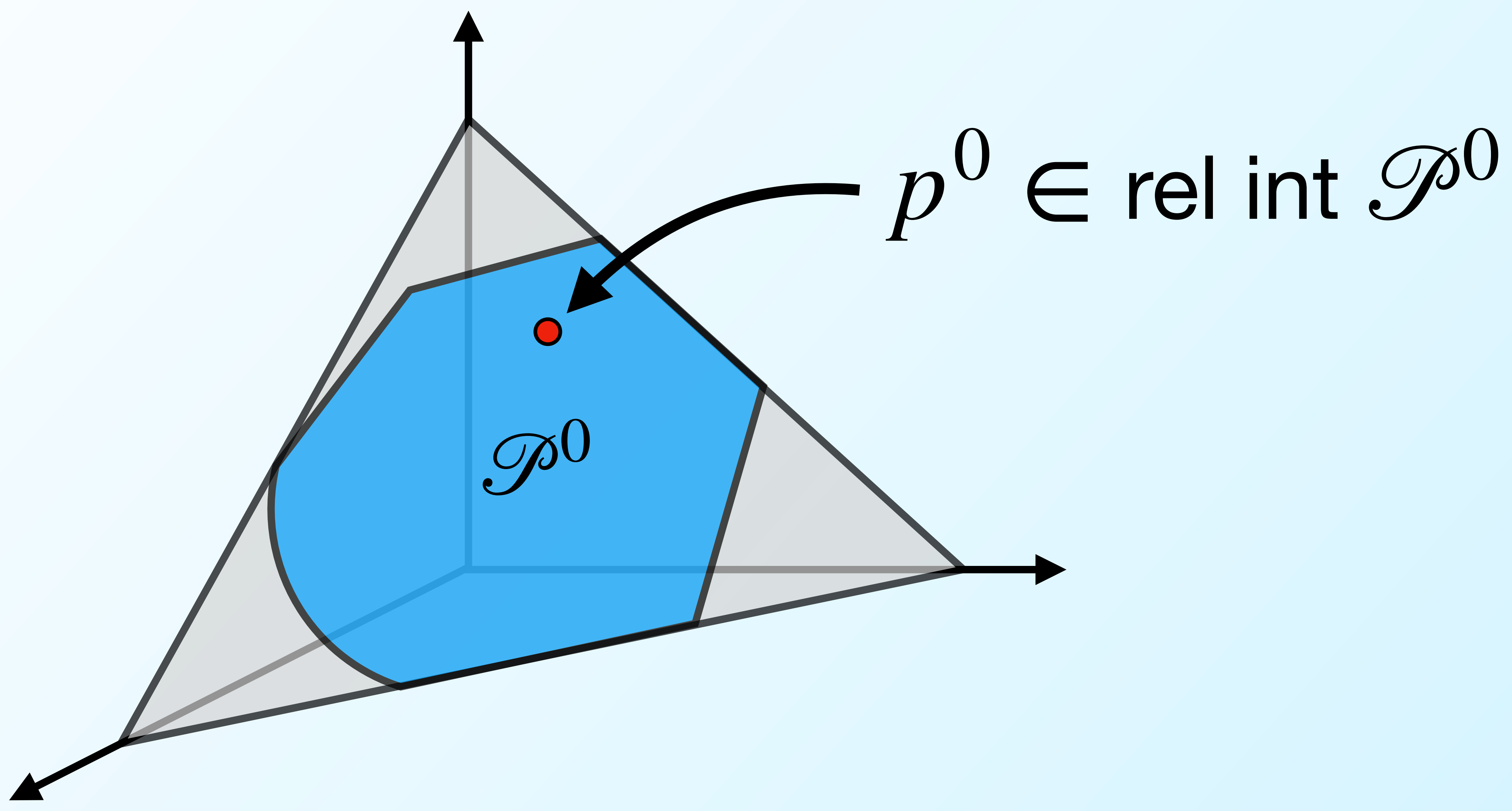
Structural ambiguity set | **Historical sample** | **Out-of-sample guarantee**



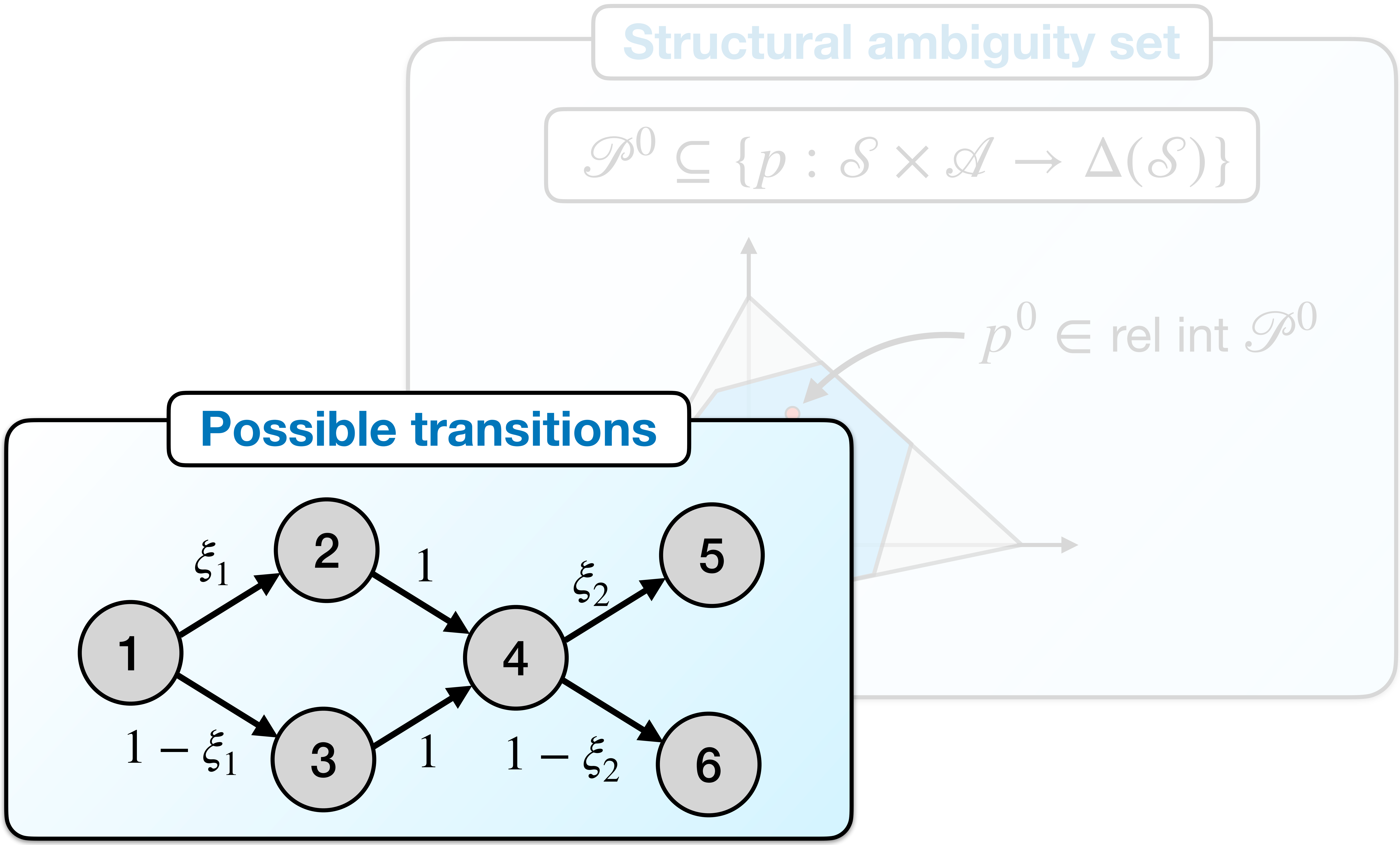
Wiesemann et al. (2013), *Robust Markov Decision Processes*.

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*.



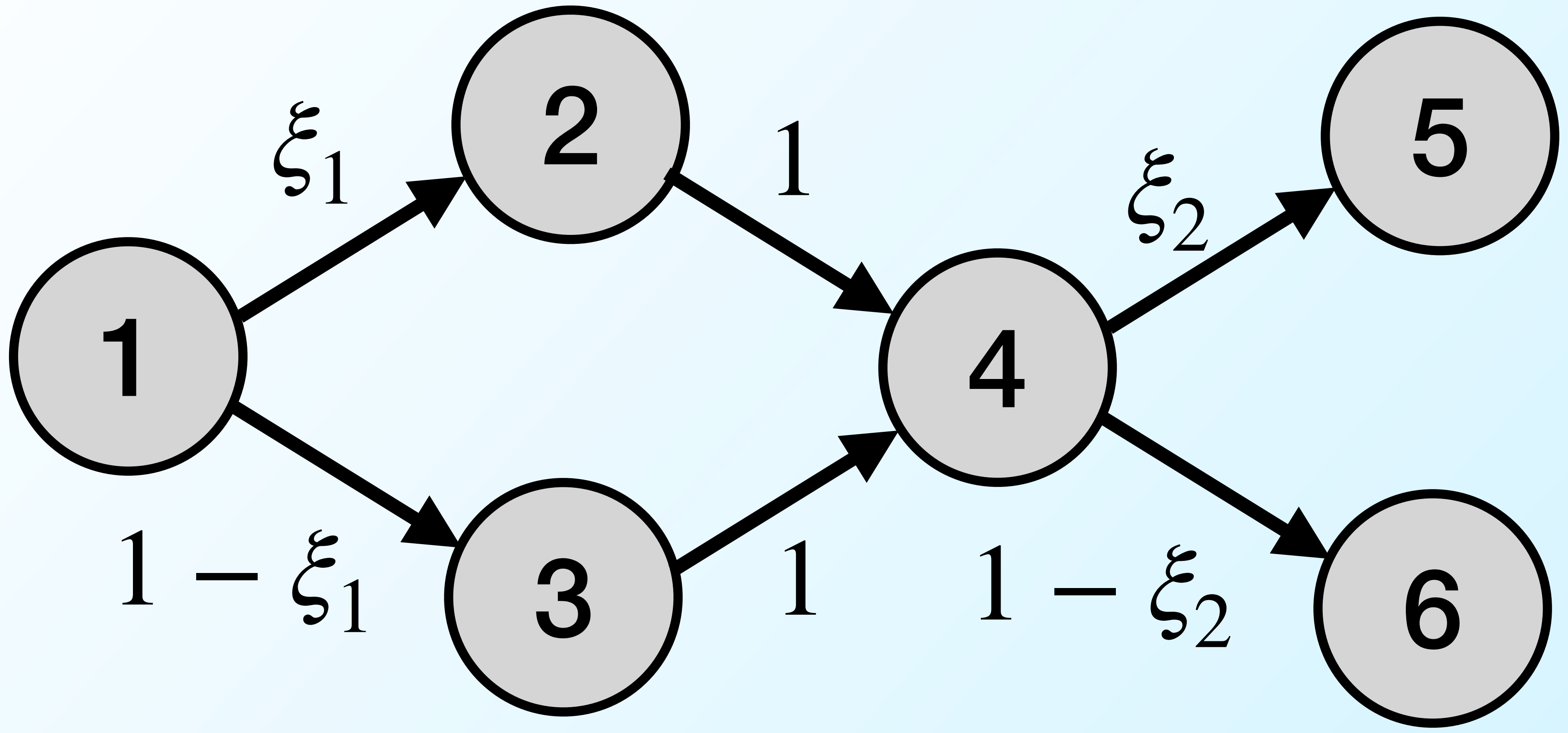
Wiesemann et al. (2013), *Robust Markov Decision Processes*.

Structural ambiguity set

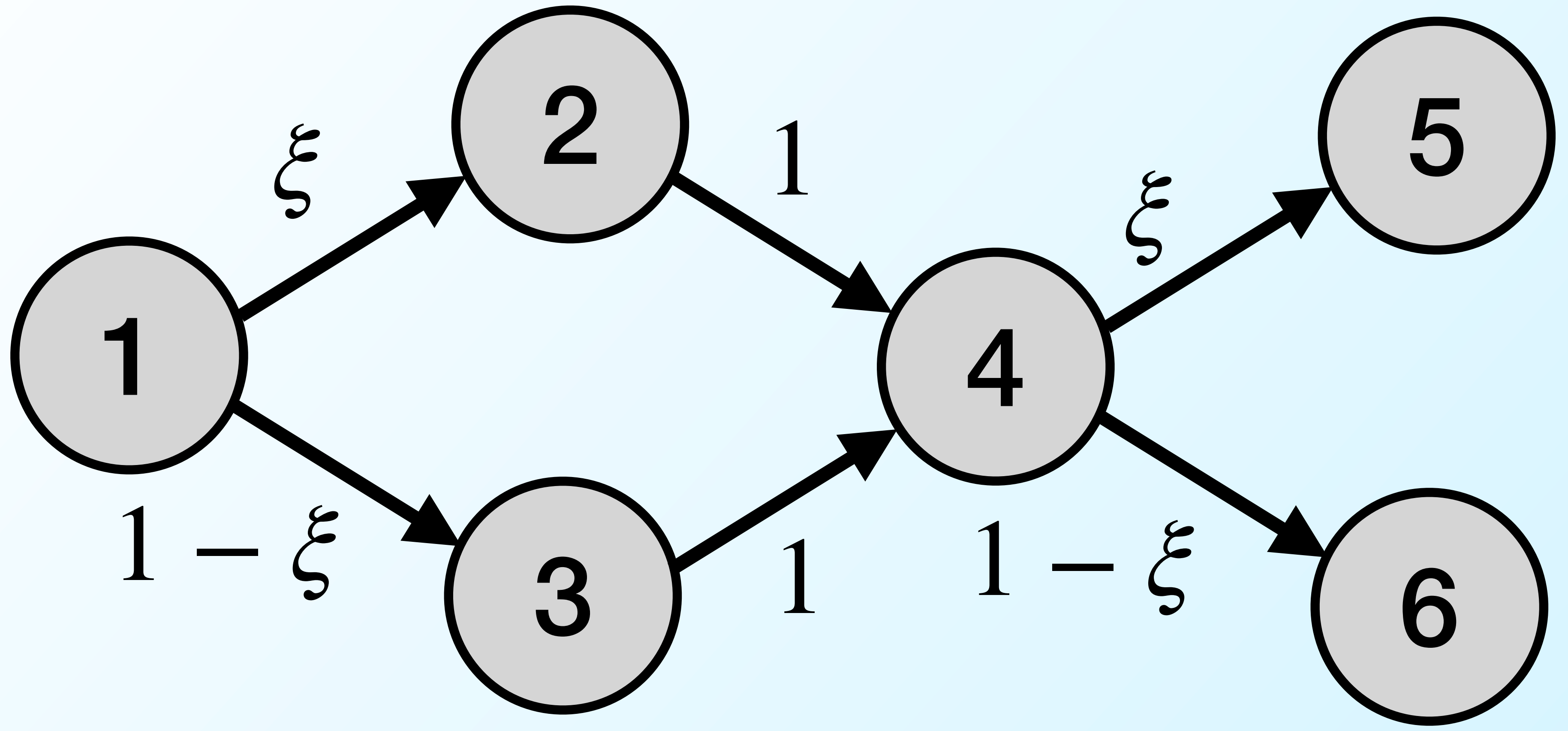
$\mathscr{P}^0 \subseteq \{p : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}$

$p^0 \in$ rel int $\mathscr{P}^0$

Possible transitions

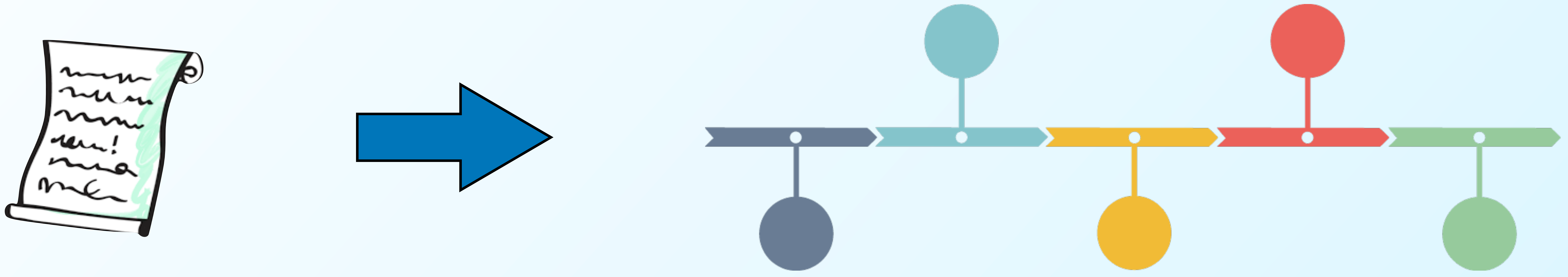


Equal probabilities



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

Historical sample



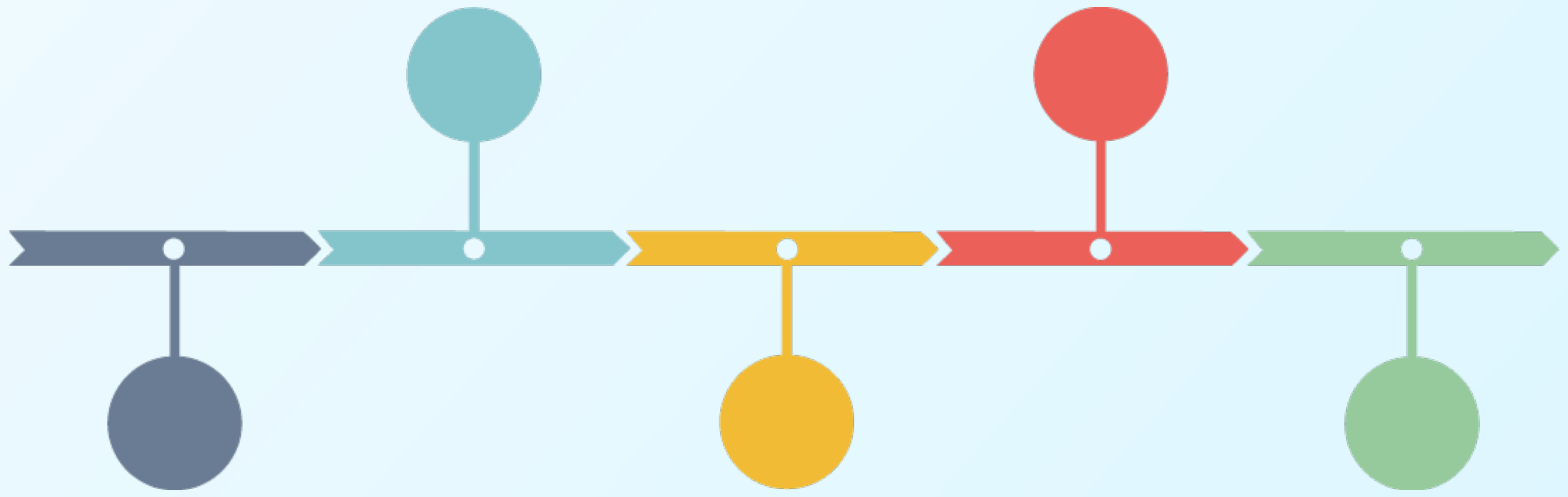
historical policy π^0
(stationary, randomized)

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

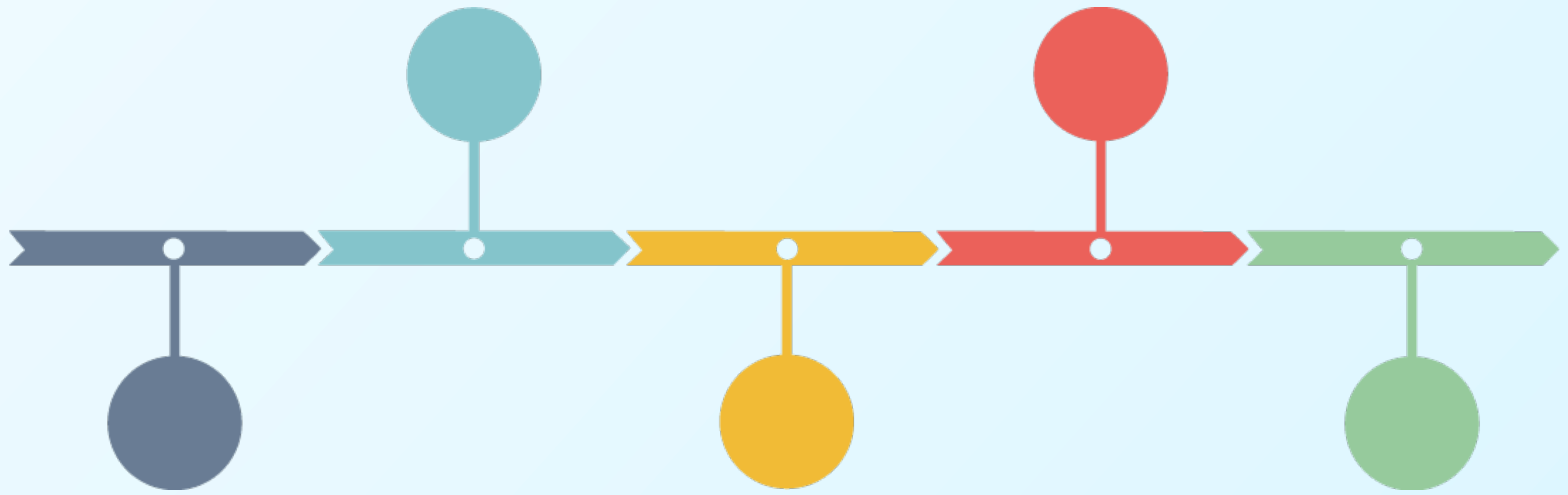
$$\mathscr{L}_n(p) \;=\; q(s_1) \cdot \pi^0(a_n \,|\, s_n) \cdot \prod_{t=1}^{n-1} \left[\pi^0(a_t \,|\, s_t) \cdot p(s_{t+1} \,|\, s_t, a_t)\right]$$

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

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 $s \in \mathcal{S}$ infinite 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 δ = $(1-\beta)$-quantile of χ^2-distribution with κ degrees of freedom

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

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

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

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

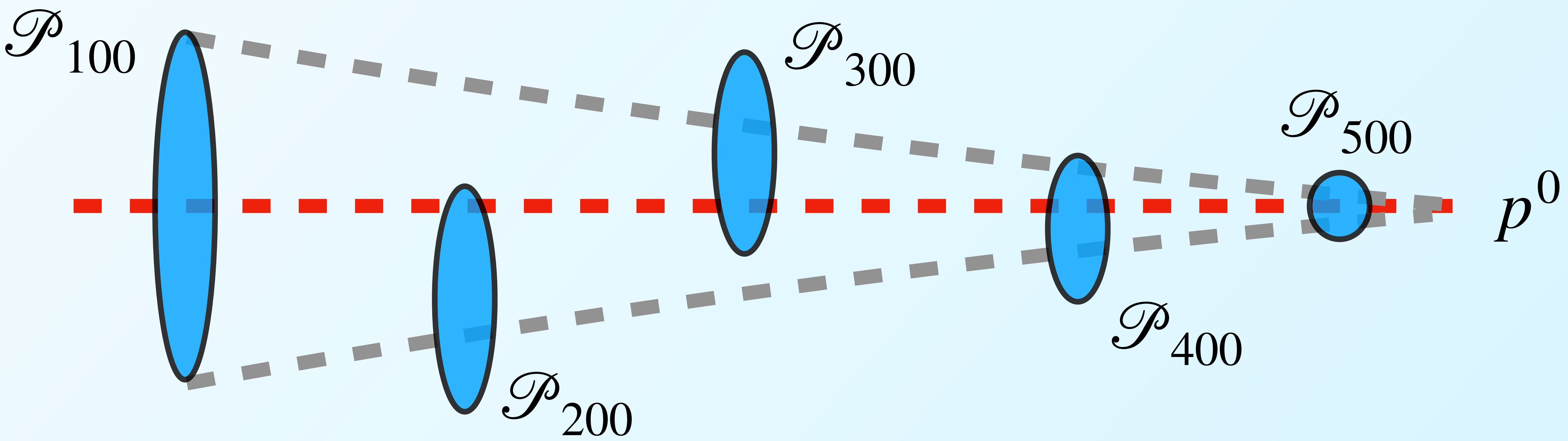
Theorem

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

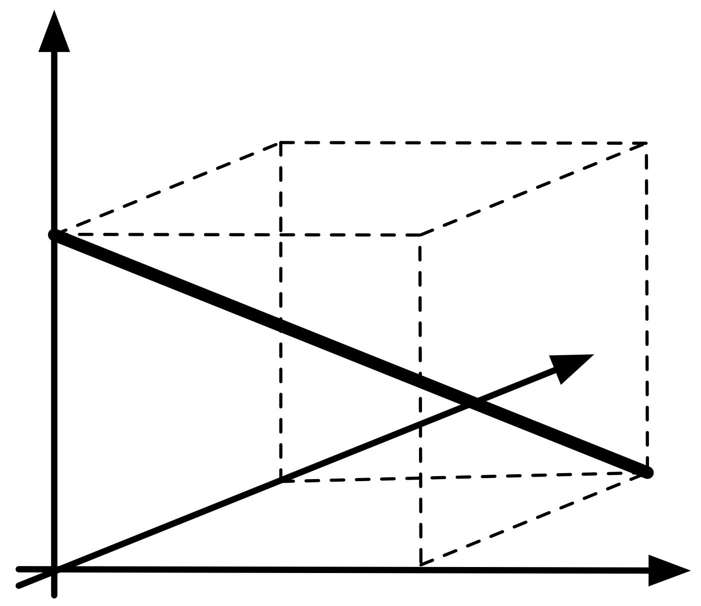
Assumption: Historical policy π^0 visits every (s, a) infinite often as $n \longrightarrow \infty$

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

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

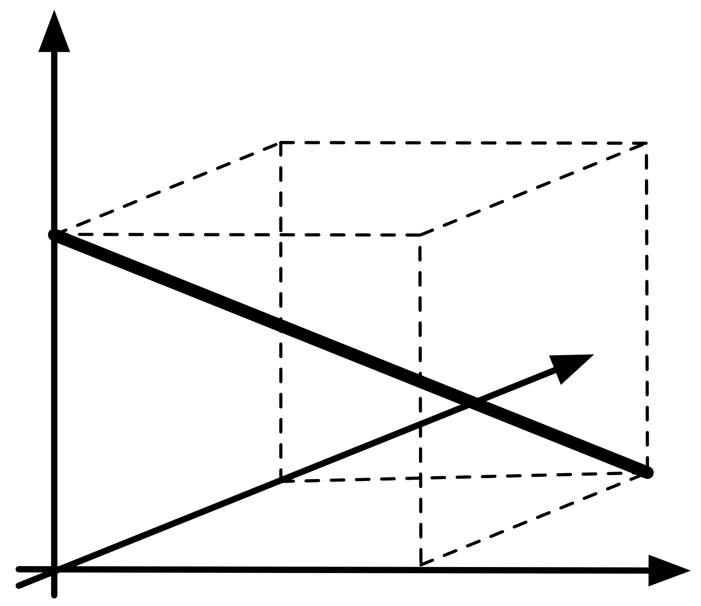


Wiesemann et al. (2013), *Robust Markov Decision Processes*.



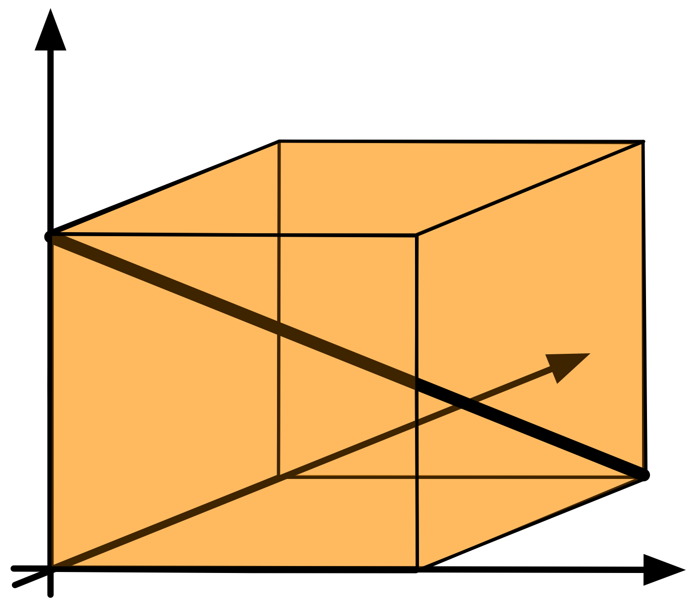
General (non-rectangular) ambiguity sets

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



General (non-rectangular) ambiguity sets

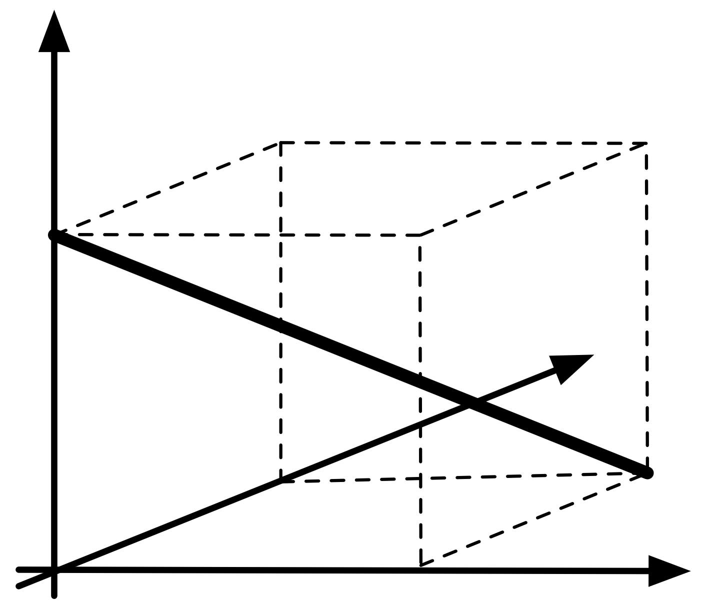
- Optimal policy can be randomized & history-dependent
- Bellman optimality principle violated; NP-hard



(s,a)-rectangular ambiguity sets

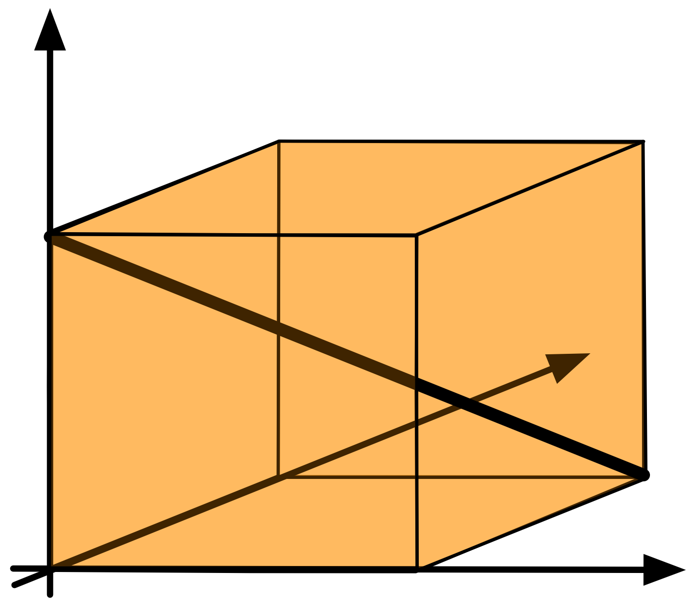
$$\mathscr{P} = \prod_{(s,a)\in\mathscr{S}\times\mathscr{A}} \mathscr{P}_{s,a} \quad \text{with} \quad \mathscr{P}_{s,a} \subseteq \Delta(\mathscr{S})$$

Rectangularity



General (non-rectangular) ambiguity sets

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

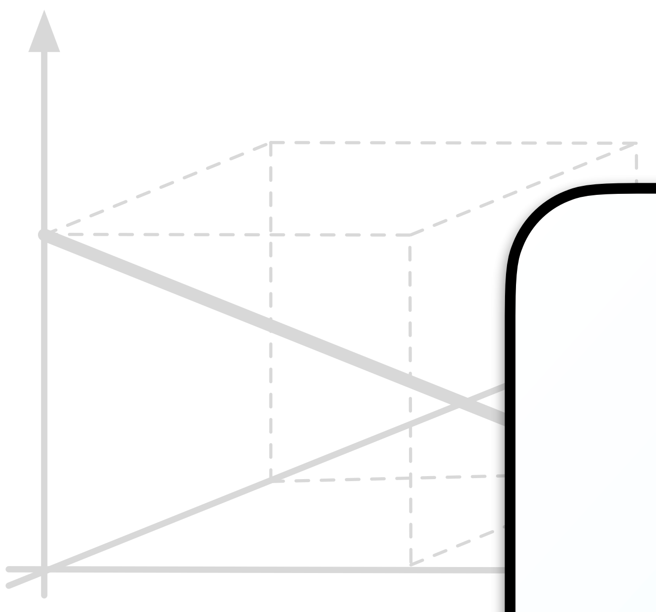


(*s*,*a*)-rectangular ambiguity sets

- Optimal policy stationary and deterministic
- Bellman optimality principle holds

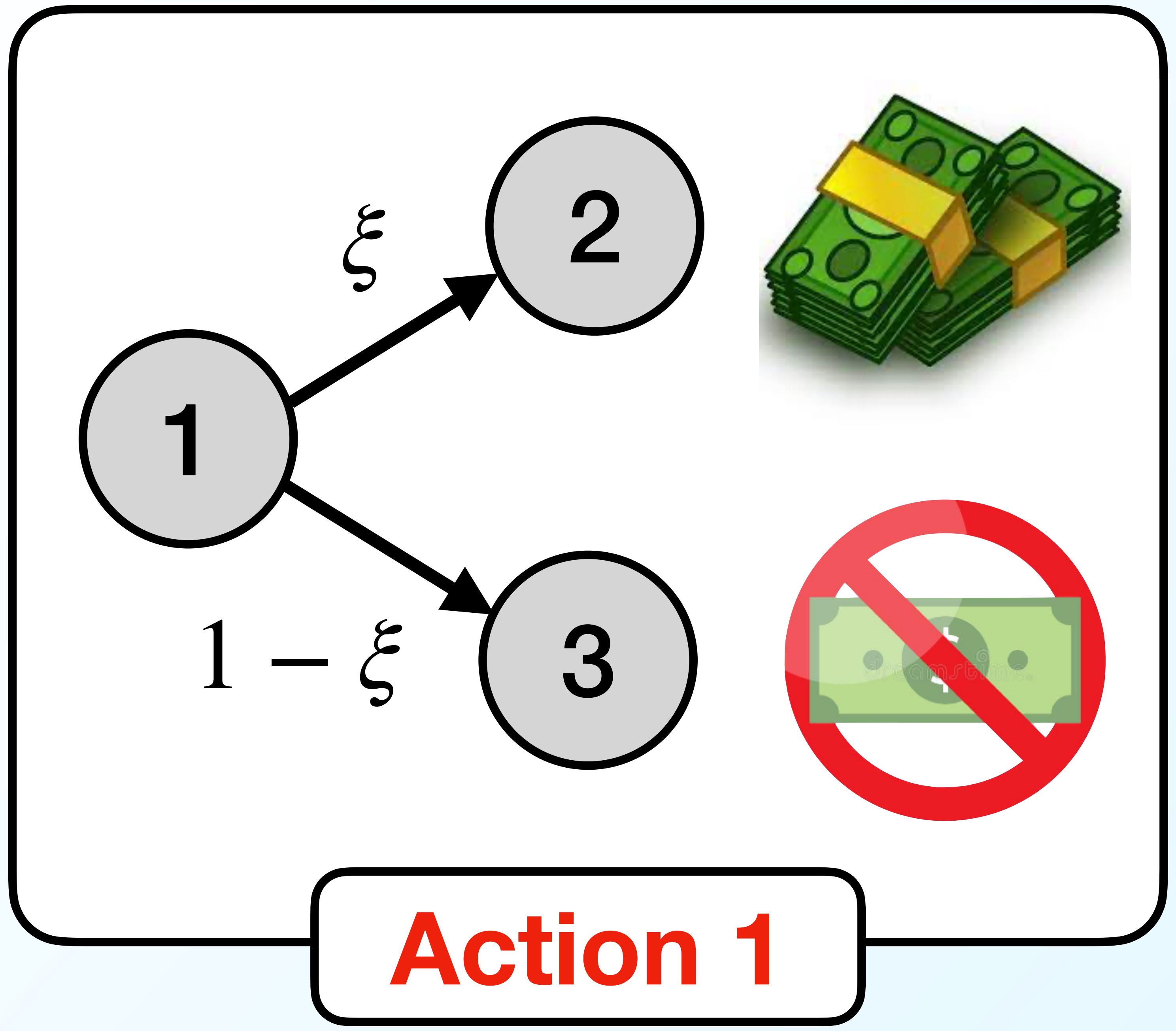
Rectangularity

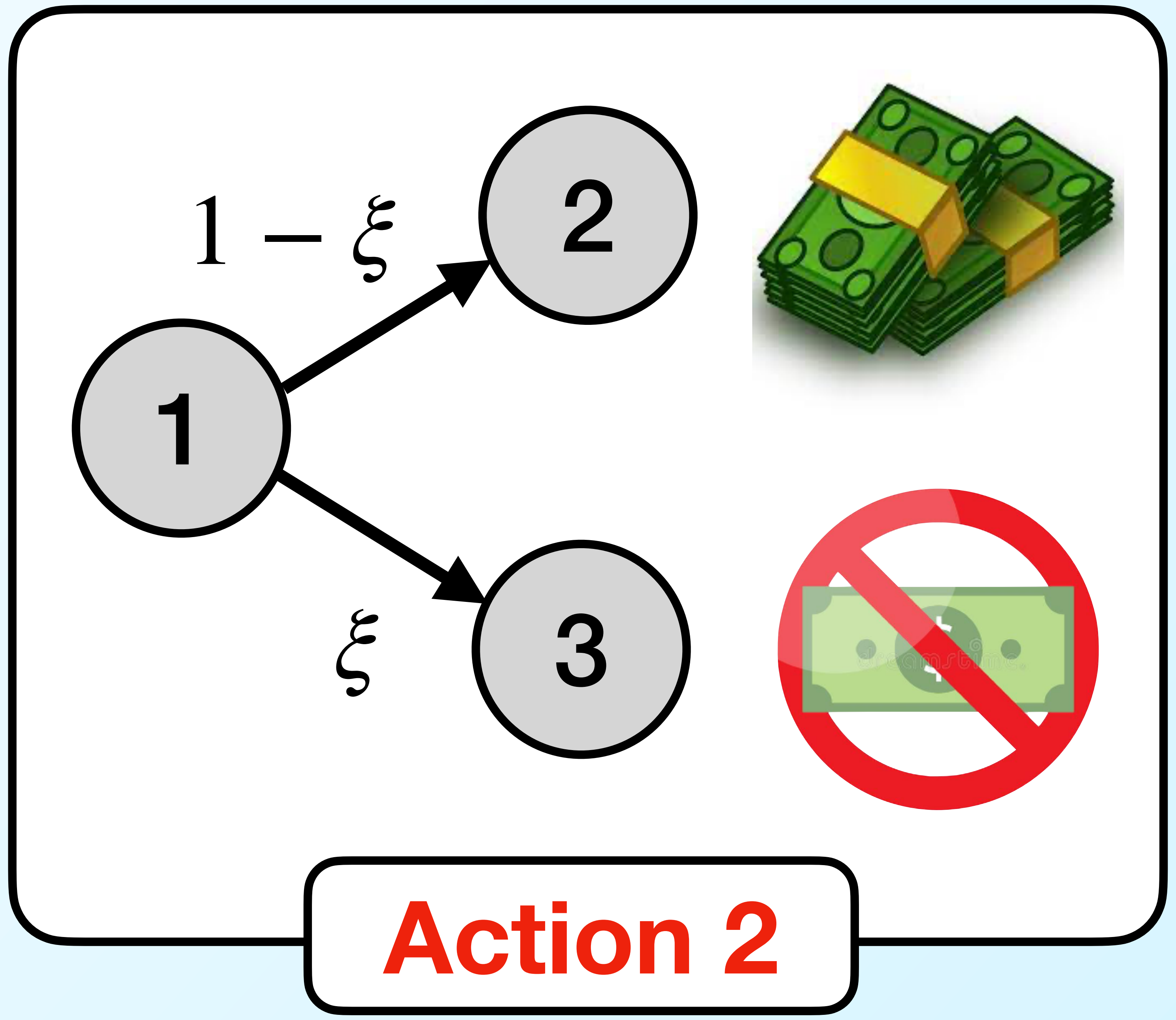
General (non-rectangular) ambiguity sets



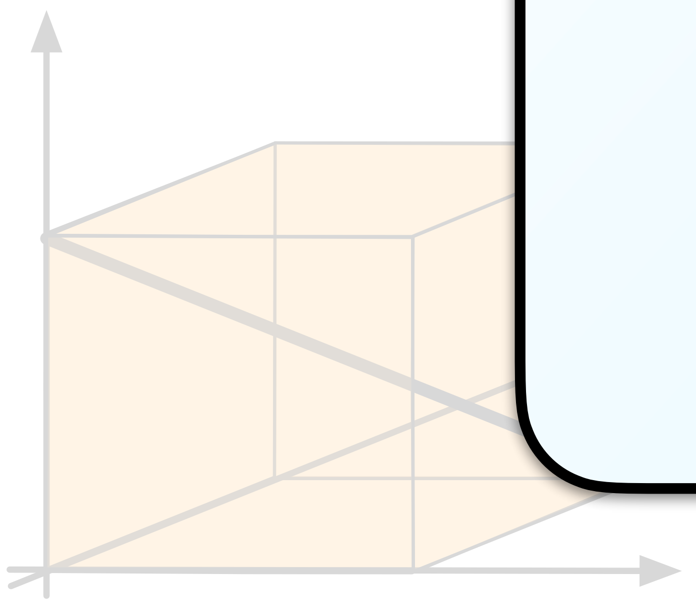
endent

Example



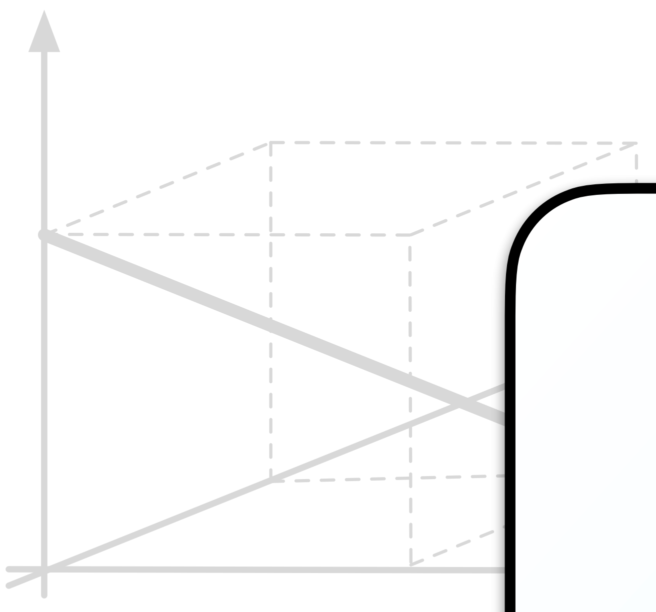


for some unknown $\xi \in [0,1]$



Bellman optimality principle holds

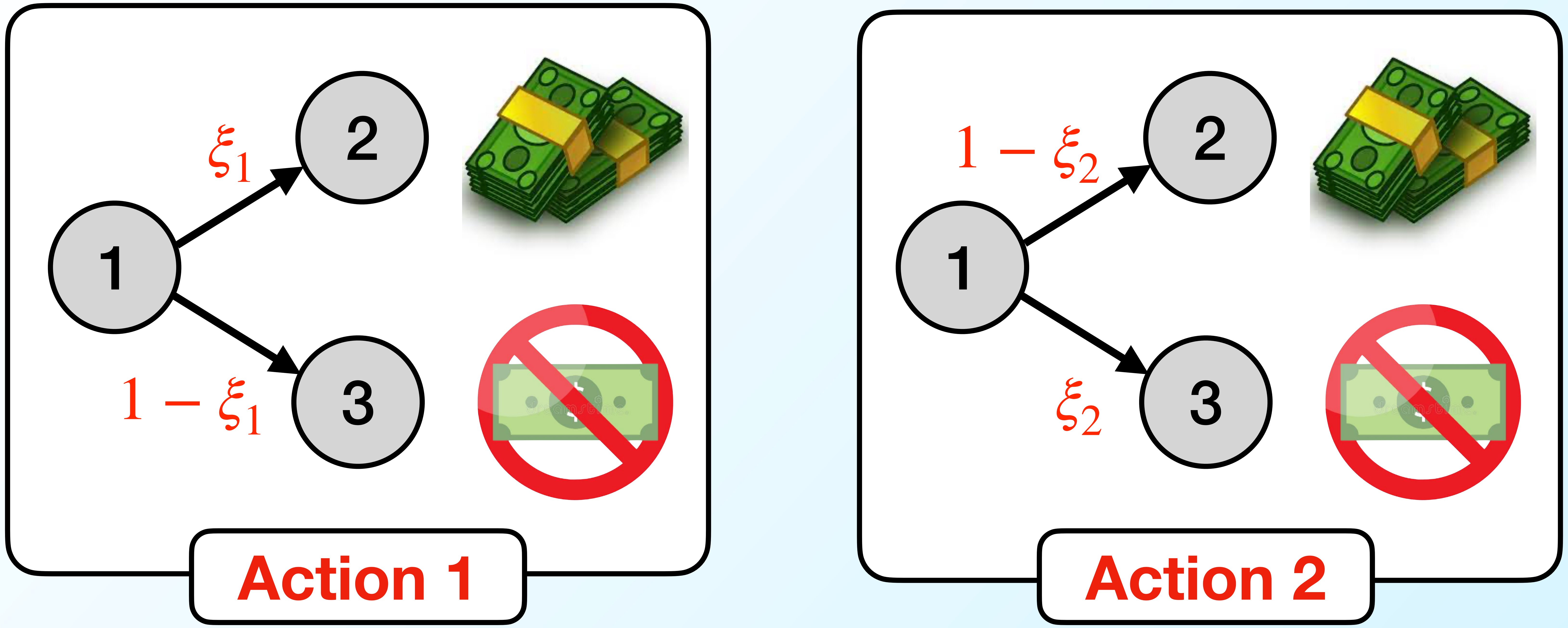
Rectangularity



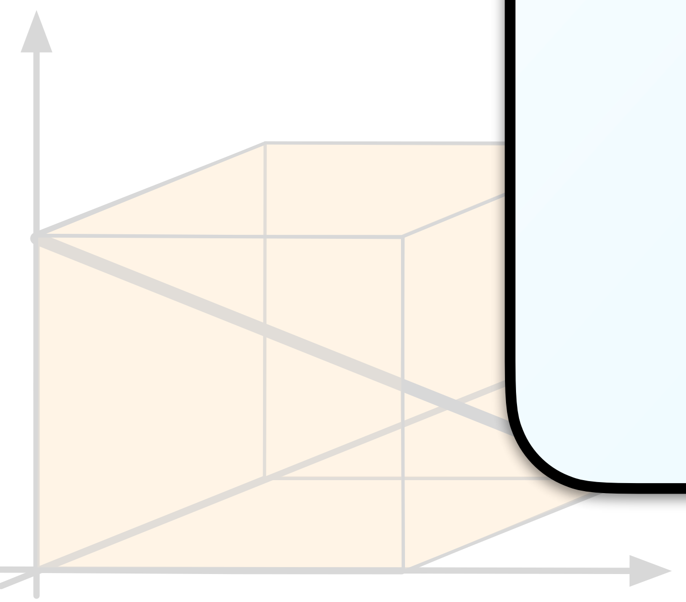
General (non-rectangular) ambiguity sets

…endent

Example

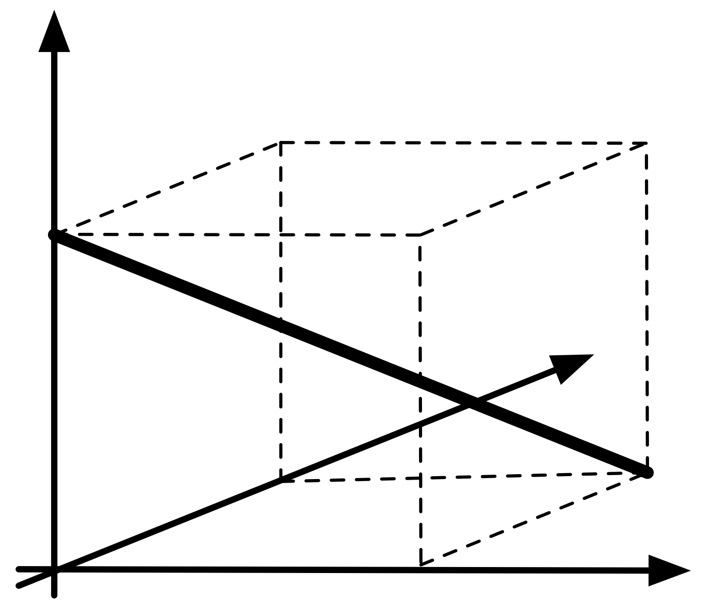


for some unknown $\xi_1, \xi_2 \in [0,1]$



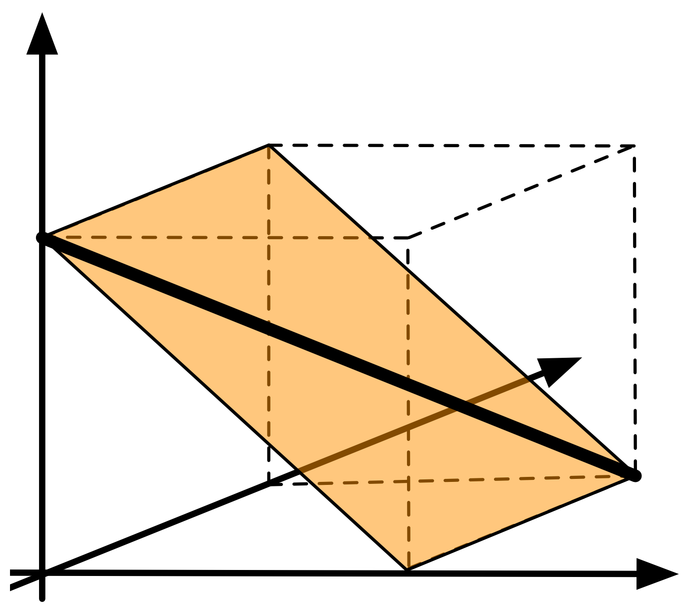
Bellman optimality principle holds

Rectangularity



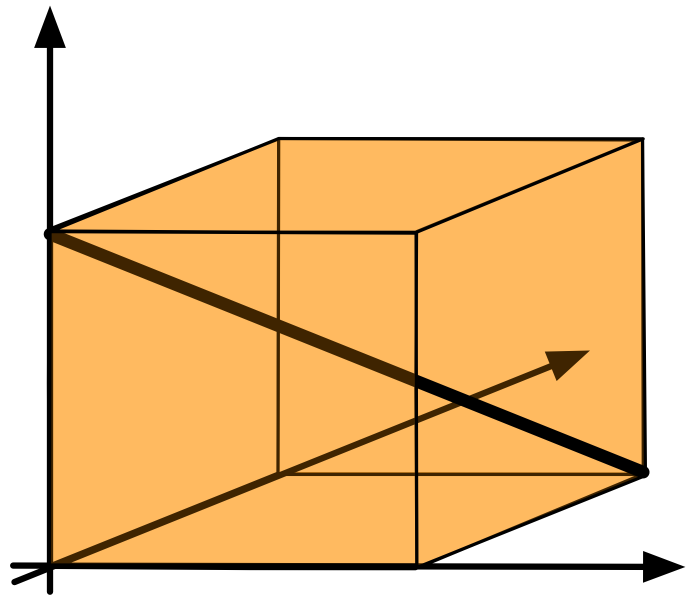
General (non-rectangular) ambiguity sets

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



s-rectangular ambiguity sets

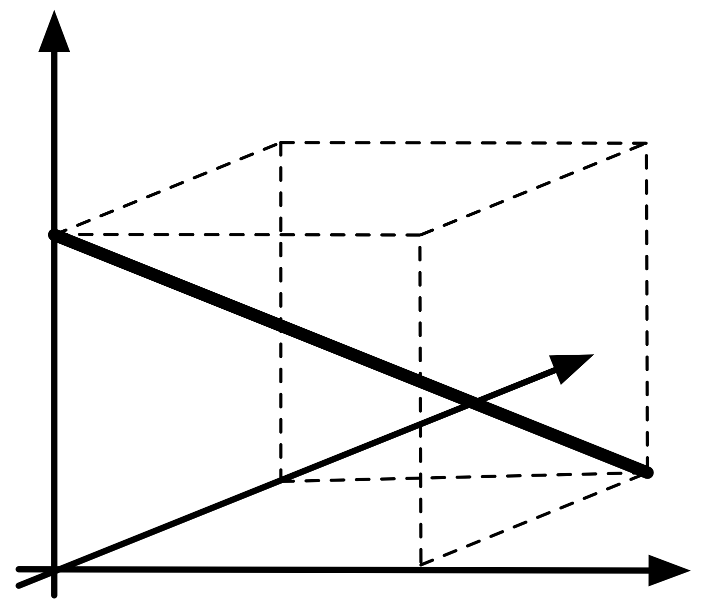
$$\mathscr{P} = \prod_{s \in \mathscr{S}} \mathscr{P}_s \quad \text{with} \quad \mathscr{P}_s \subseteq [\Delta(\mathscr{S})]^A$$



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

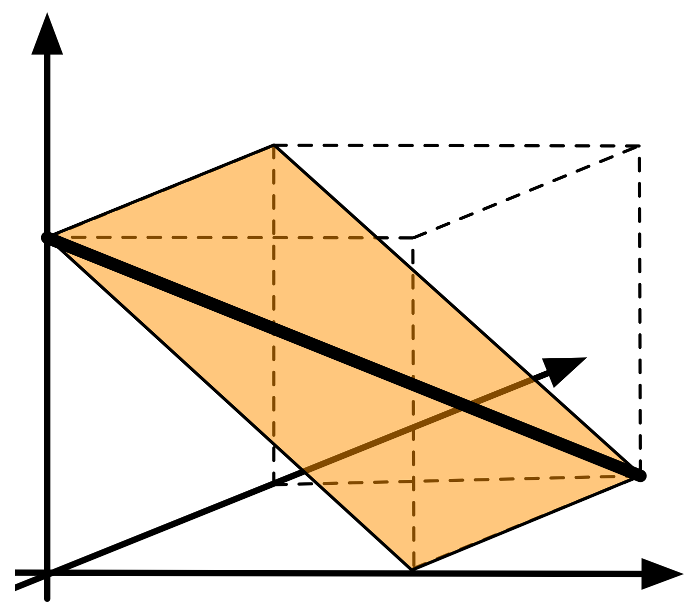
- Optimal policy stationary and deterministic
- Bellman optimality principle holds

Rectangularity



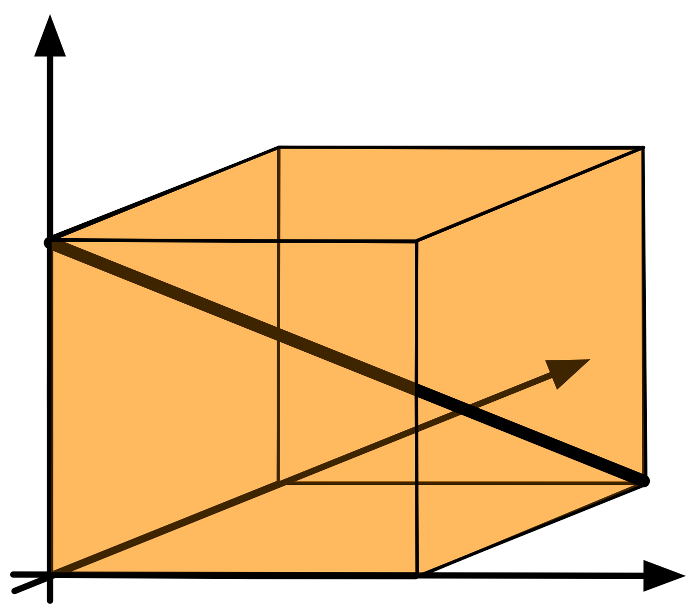
General (non-rectangular) ambiguity sets

- 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



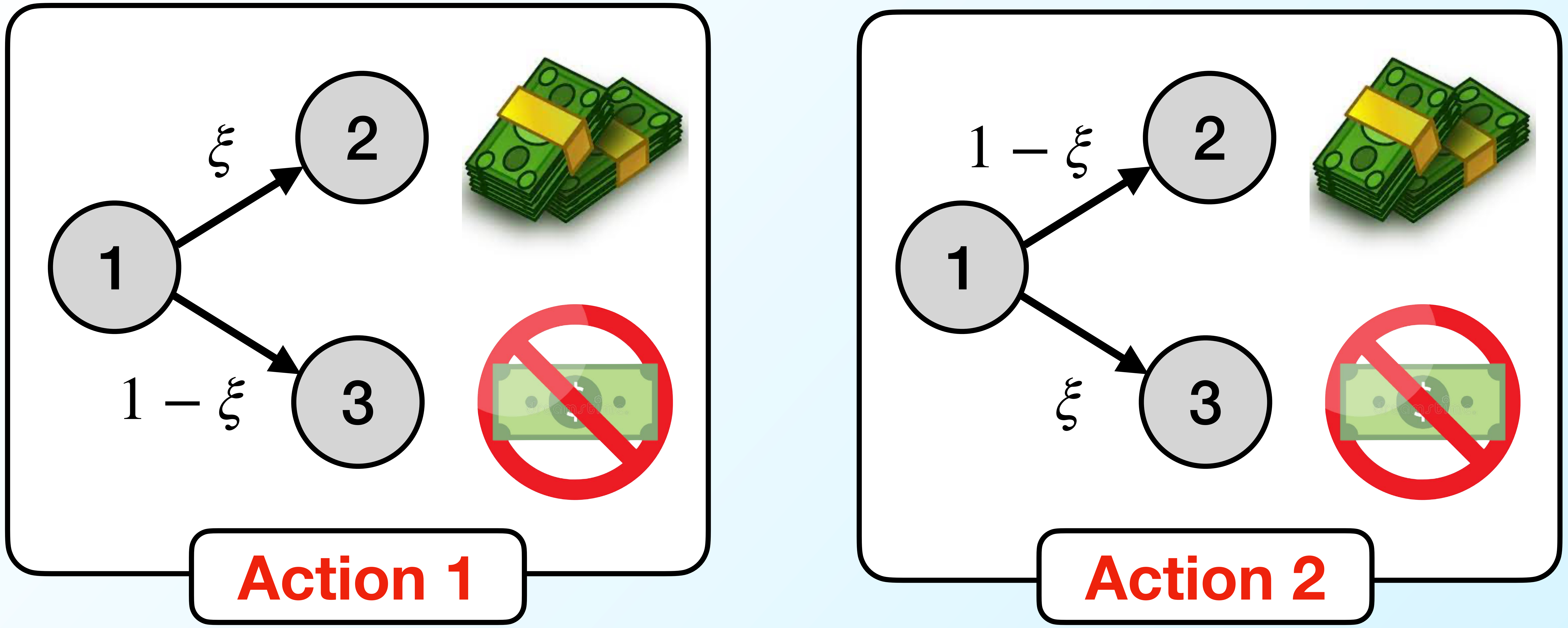
(*s*,*a*)-rectangular ambiguity sets

- Optimal policy stationary and deterministic
- Bellman optimality principle holds

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*.

s-Rectangular Ambiguity Sets: Bellman Operator

Classical (non-robust) Bellman equations

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

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*.

s-Rectangular Ambiguity Sets: Bellman Operator

Robust Bellman operator

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

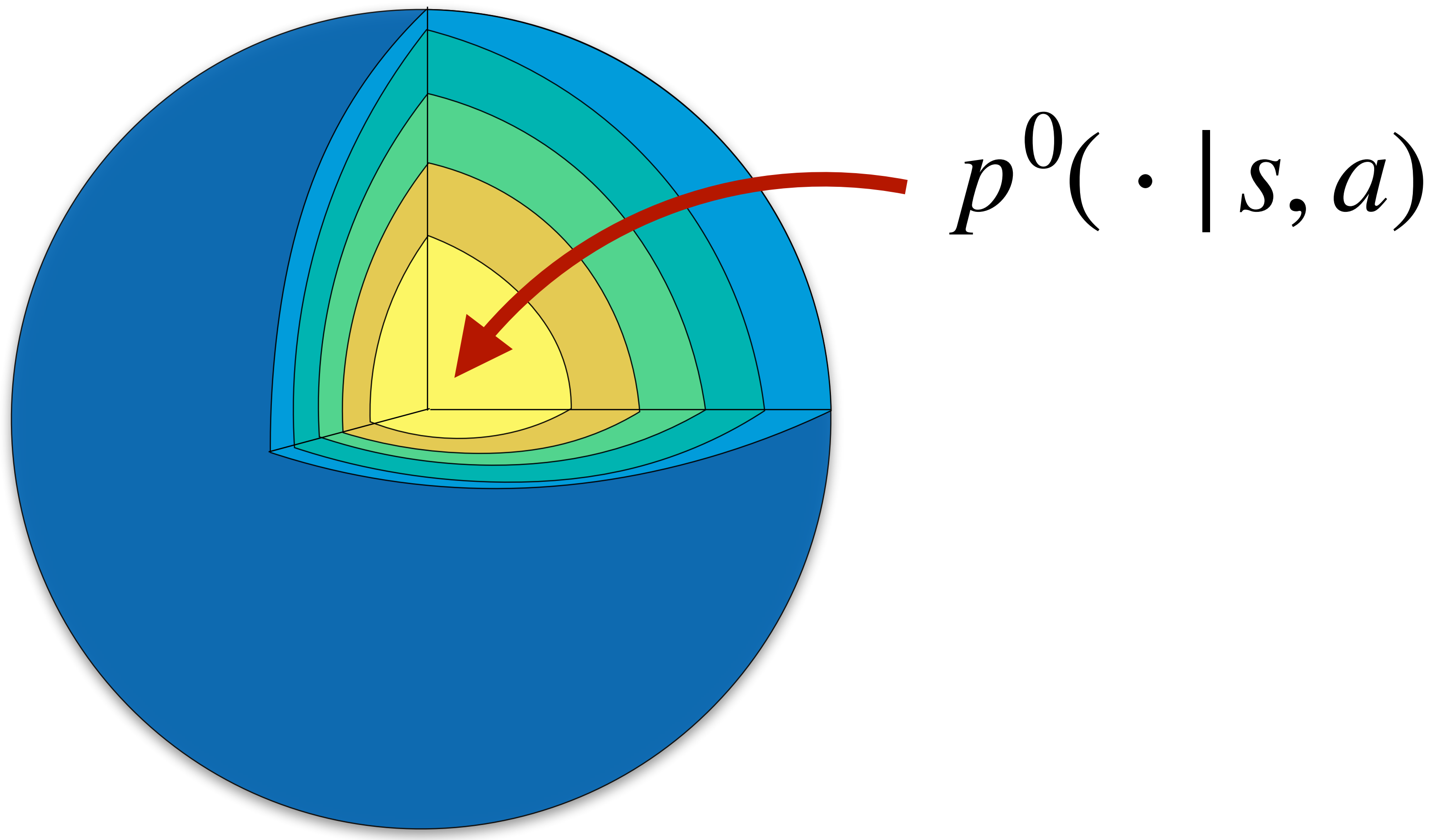
s-Rectangular Ambiguity Sets: Bellman Operator

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

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



Ho et al. (2023), *Robust Phi-Divergence MDPs*.

s-Rectangular Ambiguity Sets: Bellman Operator

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

s-Rectangular Ambiguity Sets: Bellman Operator

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

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

Minimax theorem: exchange order of min and max

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

s-Rectangular Ambiguity Sets: Bellman Operator

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

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

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

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

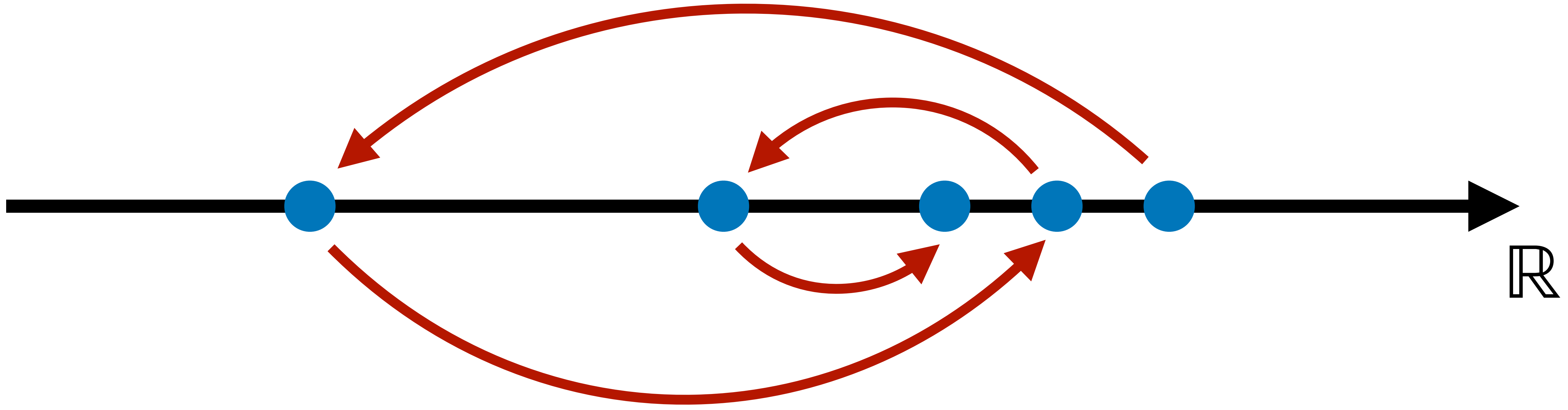
s-Rectangular Ambiguity Sets: Bellman Operator

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

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

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

Bisection search:



Ho et al. (2023), *Robust Phi-Divergence MDPs*.

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*.

s-Rectangular Ambiguity Sets: Bellman Operator

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

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

s-Rectangular Ambiguity Sets: Bellman Operator

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

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

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

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

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

$f(p)$ $g(p)$

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

$g(p)$ $f(p)$

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

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

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

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

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

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

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

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

s-Rectangular Ambiguity Sets: Bellman Operator

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

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

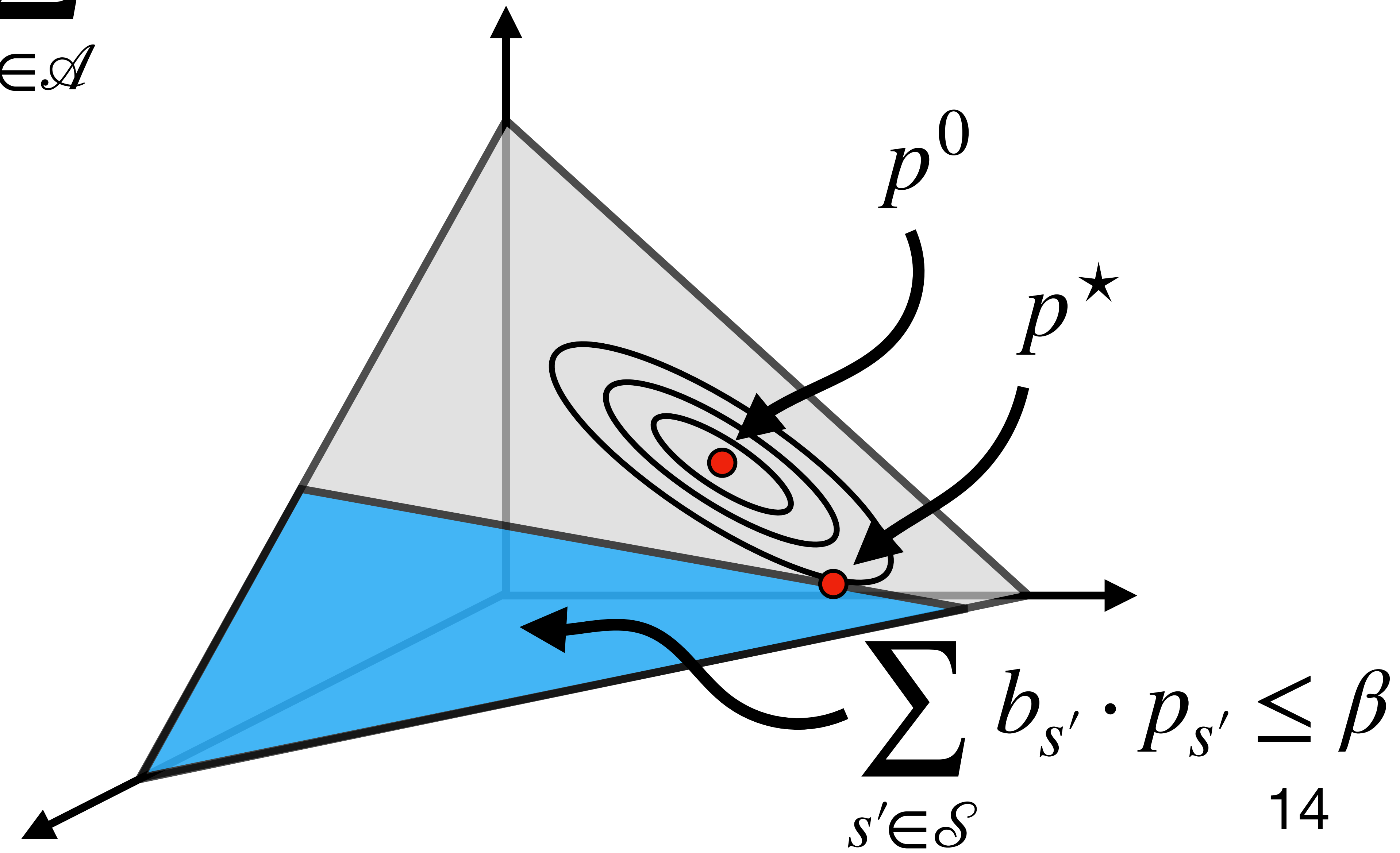
s-Rectangular Ambiguity Sets: Bellman Operator

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

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

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

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



Ho et al. (2023), *Robust Phi-Divergence MDPs*.

s-Rectangular Ambiguity Sets: Bellman Operator

Distance-constrained *s*-rectangular ambiguity set

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

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

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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

s-Rectangular Ambiguity Sets: Bellman Operator

Distance-constrained s-rectangular ambiguity set

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

Robust Bellman operator

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

Projection problem

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

Distance-constrained *s*-rectangular ambiguity set

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

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*.

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*.

s-Rectangular Ambiguity Sets: Bellman Operator

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

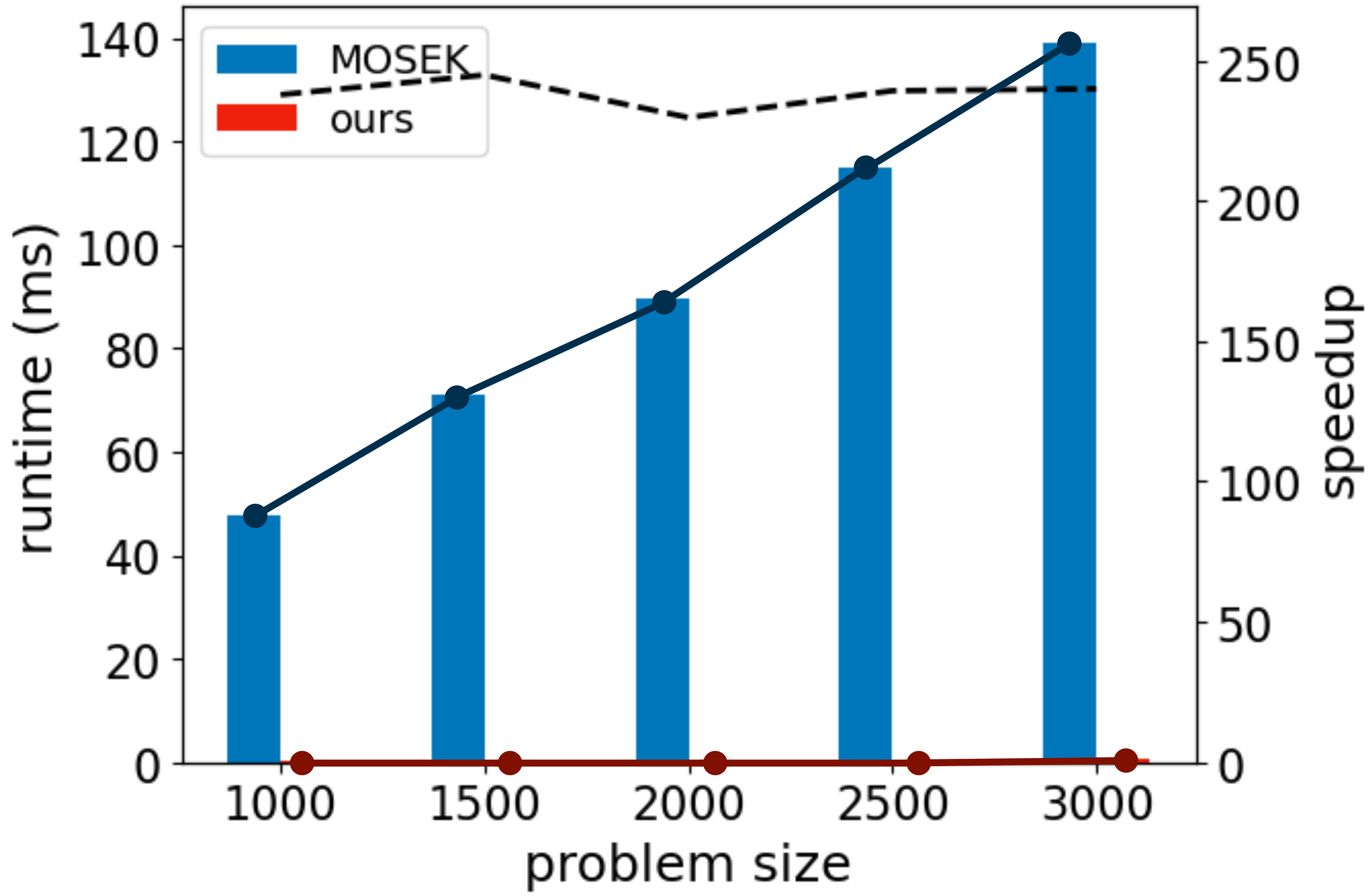
Ho et al. (2023), *Robust Phi-Divergence MDPs*.

s-Rectangular Ambiguity Sets: Bellman Operator

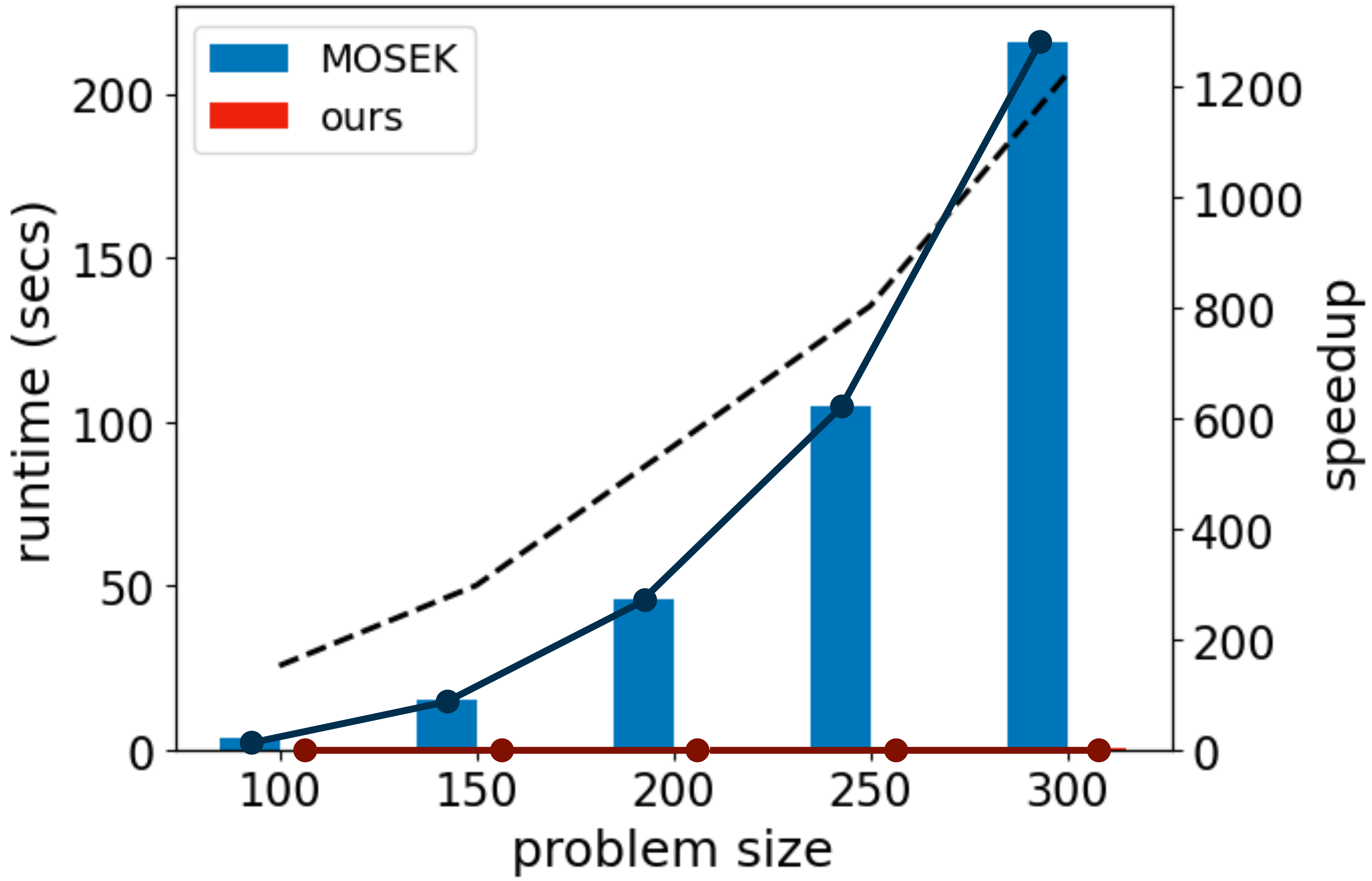
Divergence	$d_a(\cdot, p^0)$	Ours	Previous
KL-Divergence	$\sum_{s' \in \mathcal{S}} p(s' \mid s, a) \cdot \log\left(\frac{p(s' \mid s, a)}{p^0(s' \mid s, a)}\right)$	$\mathcal{O}(S^2 \cdot A \log A)$	$\mathcal{O}(\ell^2 \cdot S^2 \cdot A)$
Burg Entropy	$\sum_{s' \in \mathcal{S}} p^0(s' \mid s, a) \cdot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

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

s-Rectangular Ambiguity Sets: Bellman Operator



Projection problem



Bellman operator

Ho et al. (2023), *Robust Phi-Divergence MDPs*.

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.

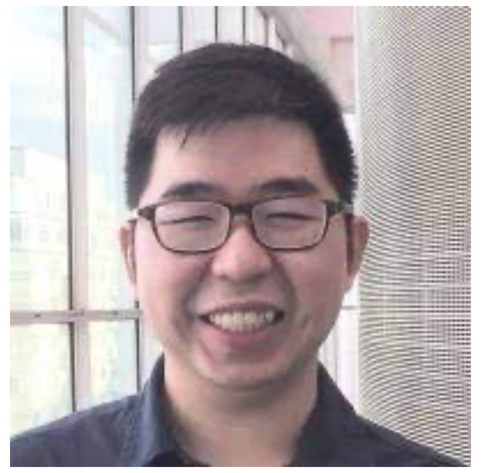
[3] J. C. D'Aeth, WW et al. Optimal National Prioritization Policies for Hospital Care During the SARS-CoV-2 Pandemic, *Nature Computational Science* 1(8):521-531, 2021.

[4] J. C. D'Aeth, WW et al. Optimal Hospital Care Scheduling During the SARS-CoV-2 Pandemic, *Management Science* (Online First), 2023.

[5] 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.

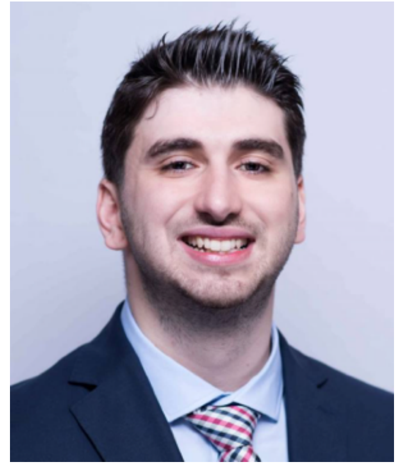
[6] C. Ho, M. Petrik, WW, Robust Phi-Divergence MDPs, *Advances in Neural Information Processing Systems 35 (NeurIPS Proceedings)*, 2022.









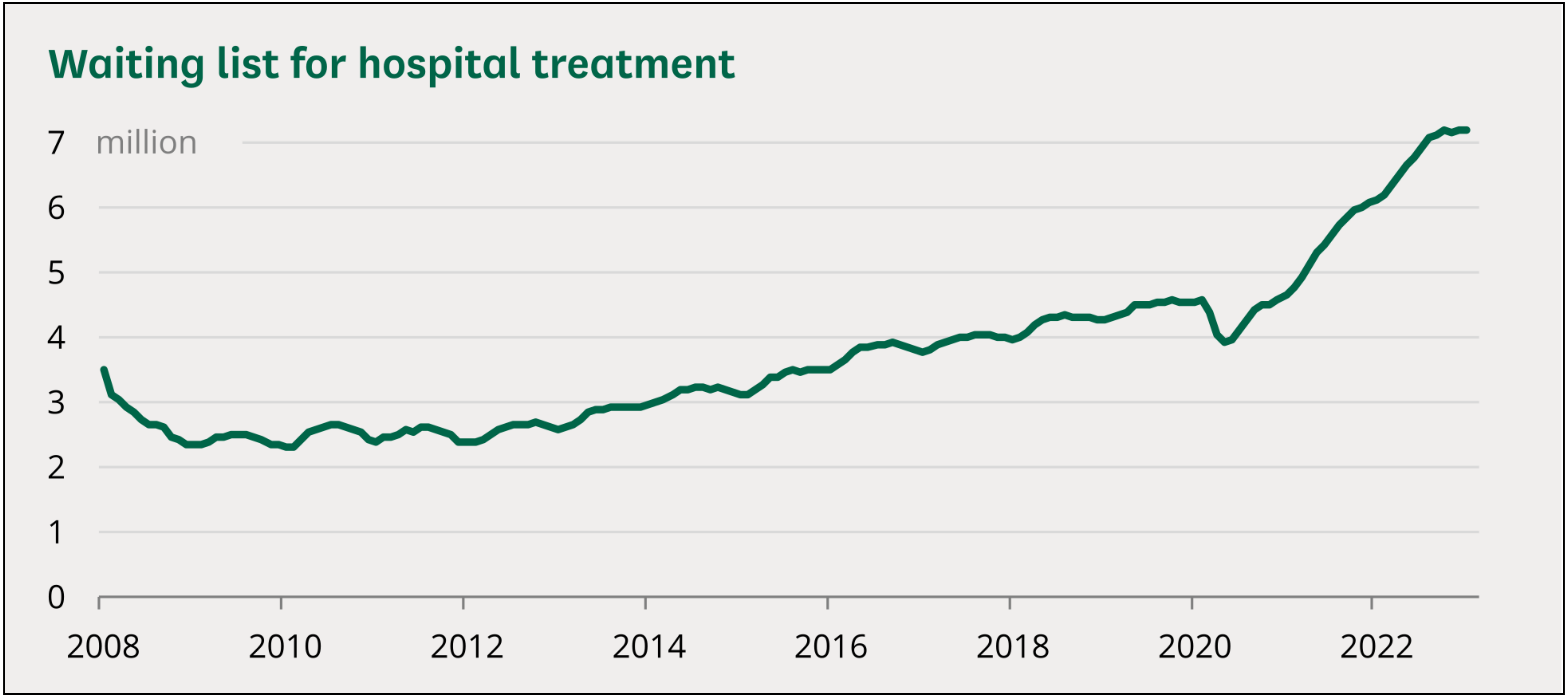



ww@imperial.ac.uk

— BACKUP —

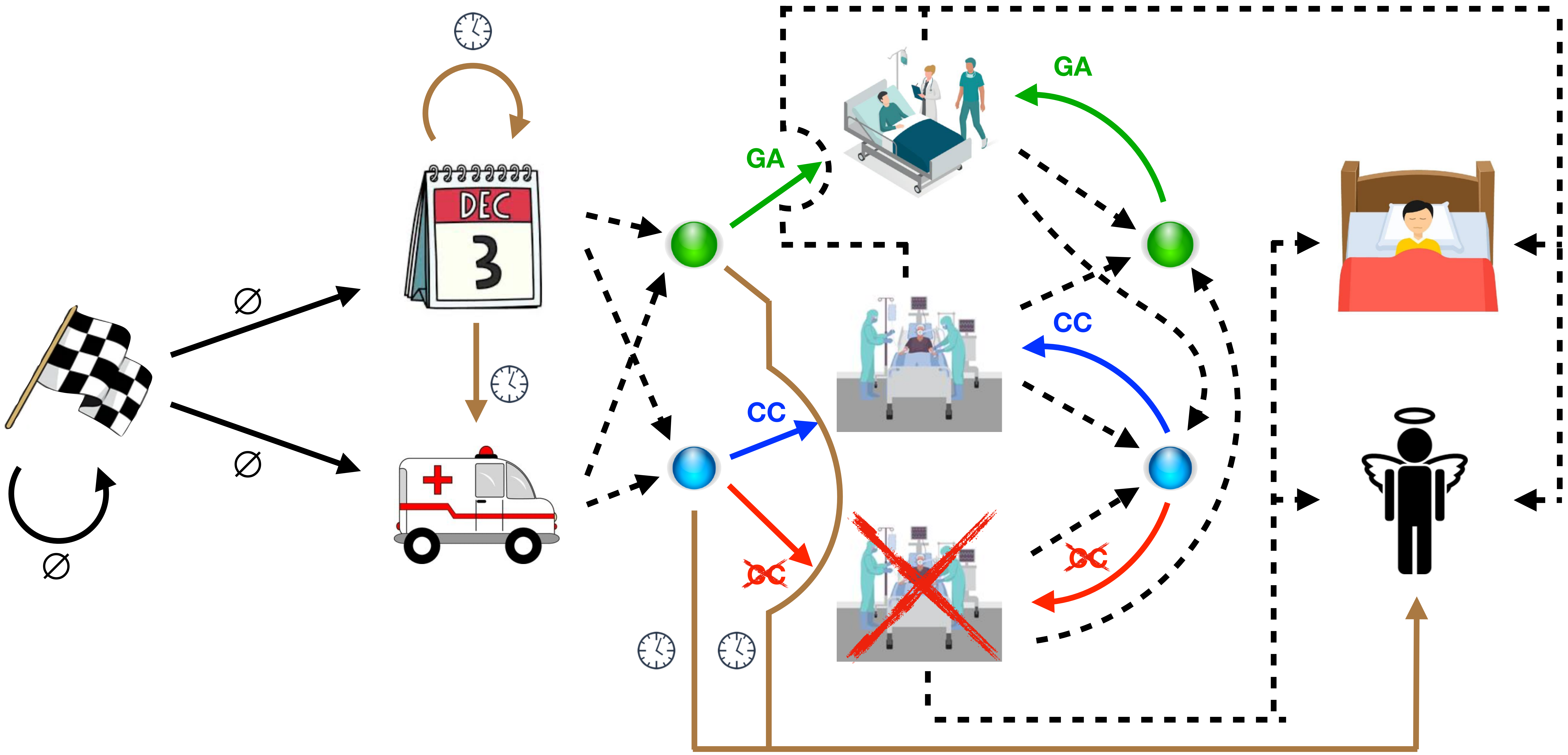
LARGE-SCALE PROBLEMS

Case Study: National Patient Prioritization



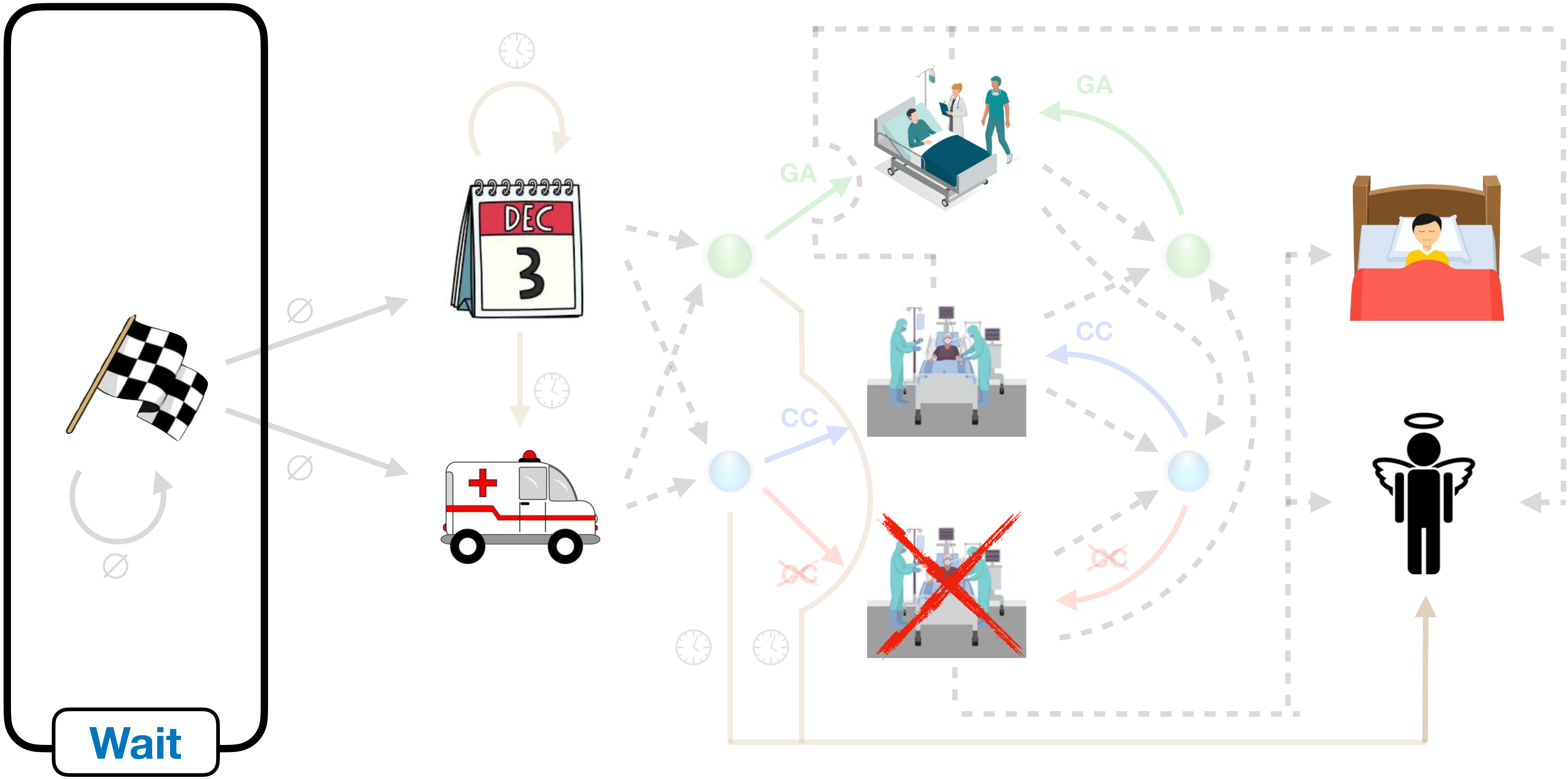
House of Commons Research Briefing: *NHS Key Statistics, England, March 2023*.

MDP model of an **individual patient**:

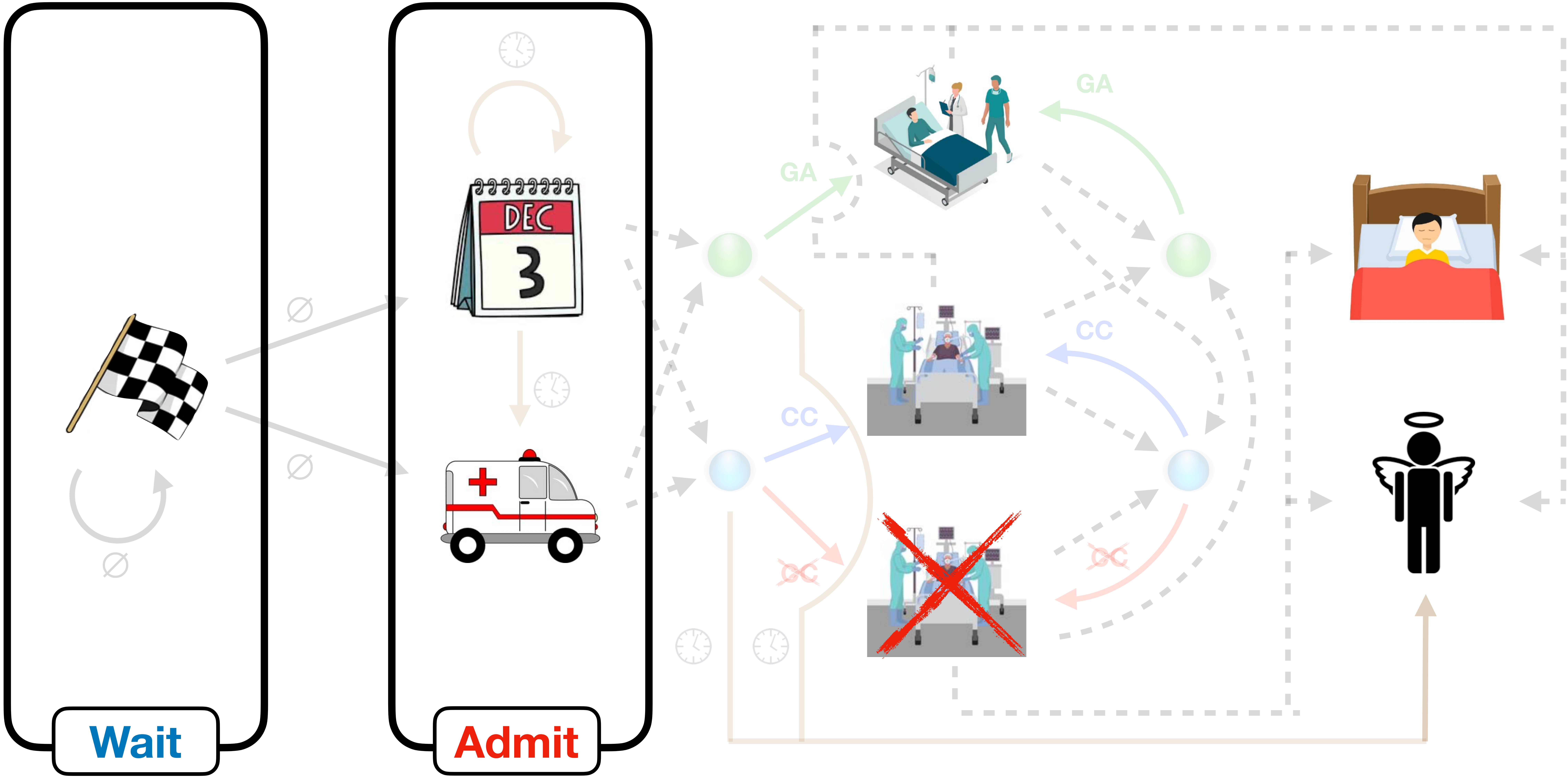


D'Aeth et al. (2021), *Optimal National Prioritization Policies for Hospital Care During the SARS-CoV-2 Pandemic*.

MDP model of an individual patient:

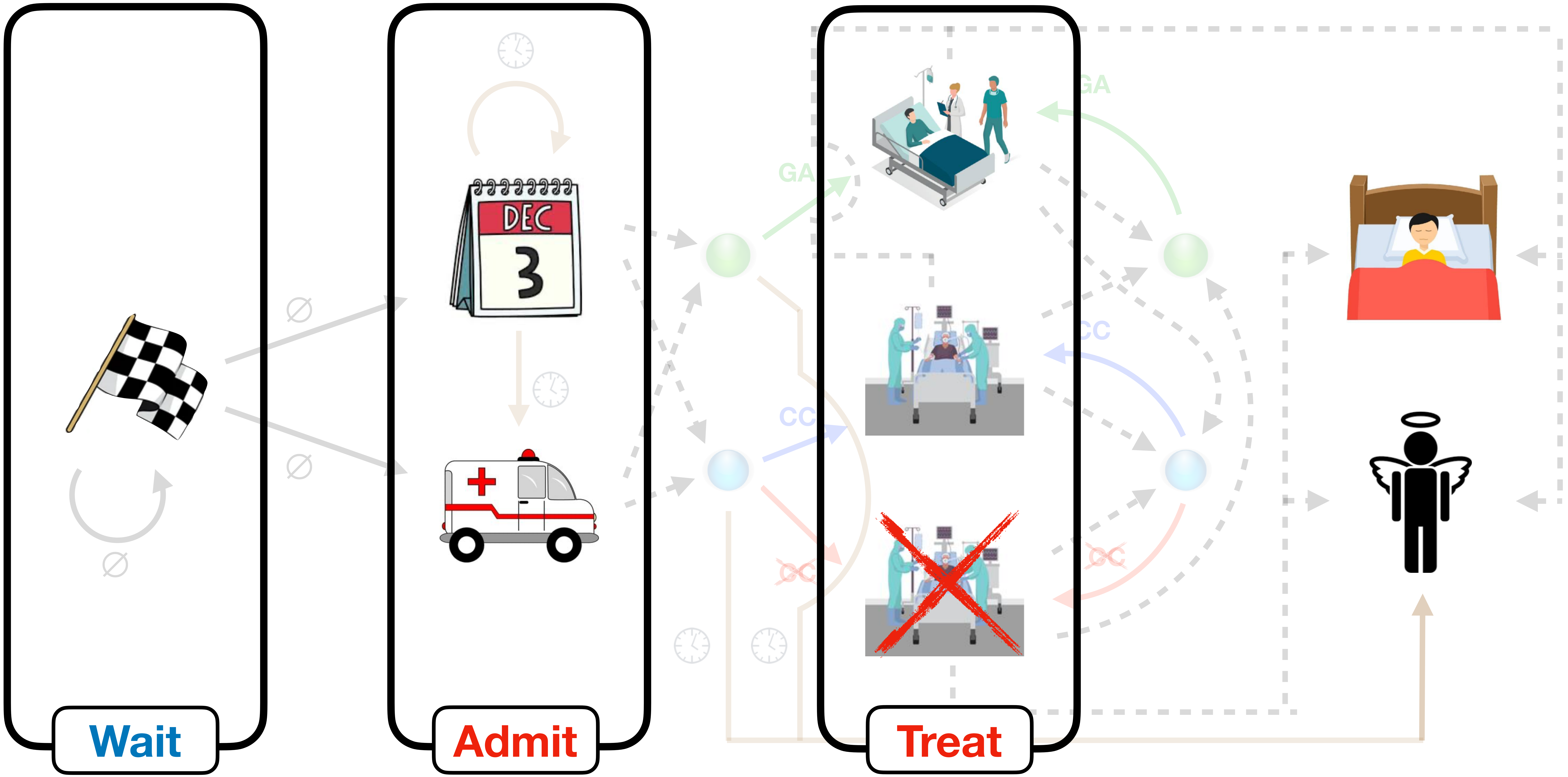


MDP model of an individual patient:

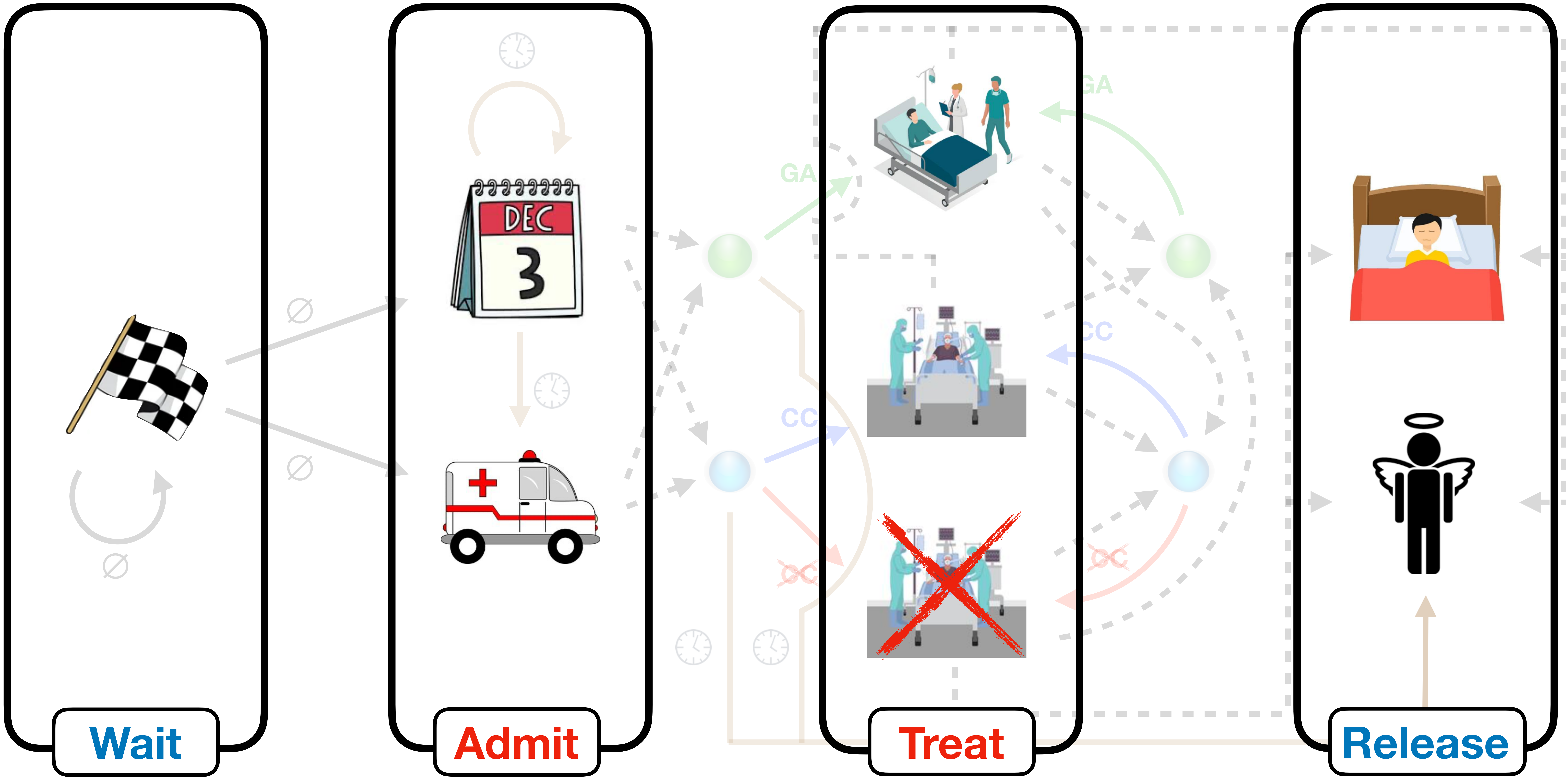


Case Study: National Patient Prioritization

MDP model of an individual patient:



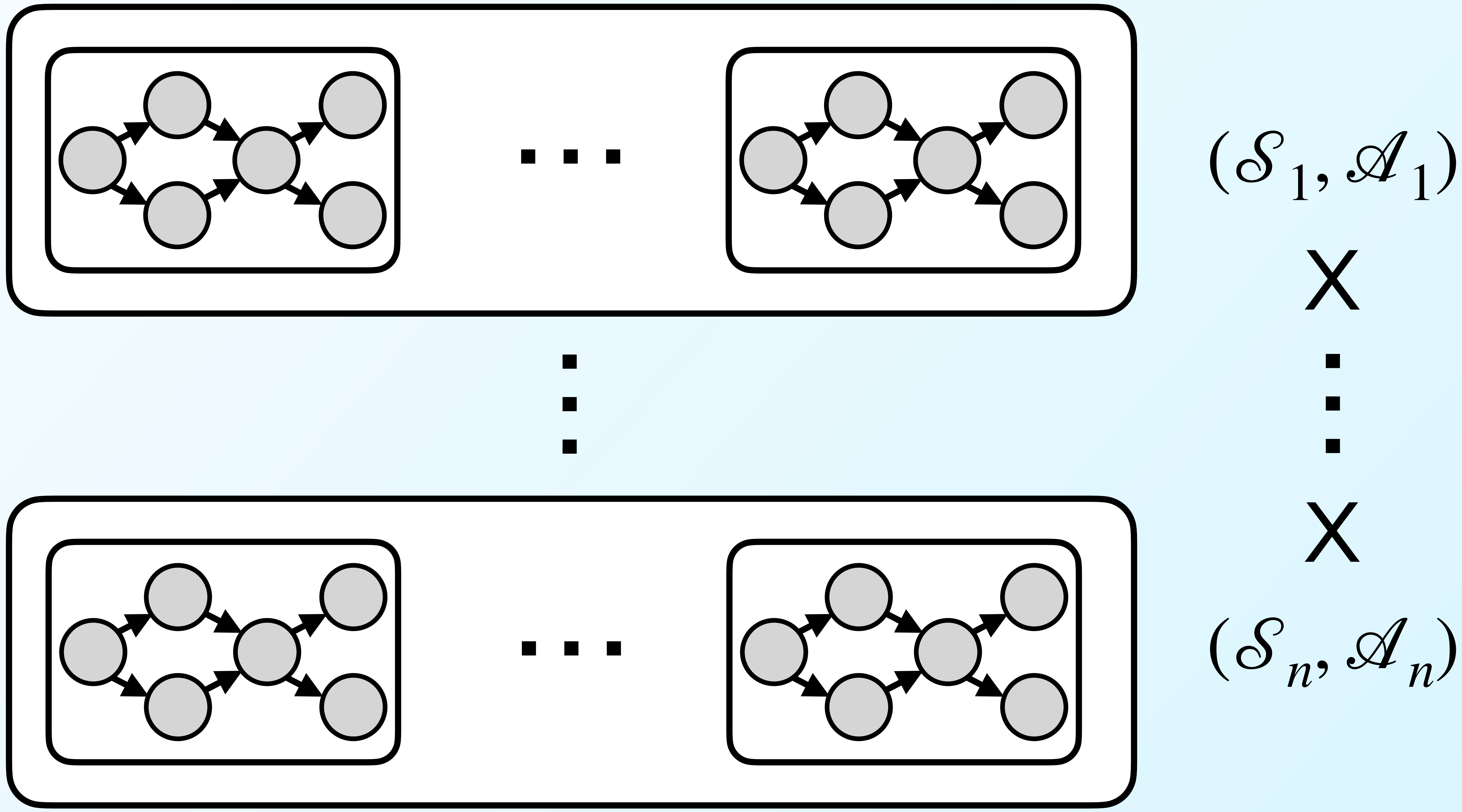
MDP model of an individual patient:



Markov decision process

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

- $\mathcal{S} = \prod_{i=1}^{n} \mathcal{S}_i$ with $\mathcal{S}_i = \{1, \ldots, S_i\}$ is the (finite) state space;
- $\mathcal{A} = \prod_{i=1}^{n} \mathcal{A}_i$ with $\mathcal{A}_i = \{1, \ldots, A_i\}$ is the (finite) action space;



Markov decision process

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

- $\mathcal{S} = \prod_{i=1}^{n} \mathcal{S}_i$ with $\mathcal{S}_i = \{1, \ldots, S_i\}$ is the (finite) state space;
- $\mathcal{A} = \prod_{i=1}^{n} \mathcal{A}_i$ with $\mathcal{A}_i = \{1, \ldots, A_i\}$ is the (finite) action space;
- $q(s) = \prod_{i=1}^{n} q_i(s_i)$ is the initial state distribution;
- $p_t(s' \mid s, a) = \prod_{i=1}^{n} p_{ti}(s_i' \mid s_i, a_i)$ is the transition kernel;
- $r_t(s, a) = \sum_{i=1}^{n} r_{ti}(s_i, a_i)$ are the expected one-step rewards;
- $T \in \mathbb{N}$ is the (finite) time horizon

Weakly coupled Markov decision process

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

- $\mathcal{S} = \prod_{i=1}^{n} \mathcal{S}_i$ with $\mathcal{S}_i = \{1, \ldots, S_i\}$ is the (finite) state space;
- $\mathcal{A} = \prod_{i=1}^{n} \mathcal{A}_i$ with $\mathcal{A}_i = \{1, \ldots, A_i\}$ is the (finite) action space;
- $q(s) = \prod_{i=1}^{n} q_i(s_i)$ is the initial state distribution;
- $p_t(s' \mid s, a) = \prod_{i=1}^{n} p_{ti}(s_i' \mid s_i, a_i)$ is the transition kernel;
- $r_t(s, a) = \sum_{i=1}^{n} r_{ti}(s_i, a_i)$ are the expected one-step rewards;
- $T \in \mathbb{N}$ is the (finite) time horizon

and

$a \in \mathcal{A}$ admissible only if $\sum_{i=1}^{n} c_{tli}(s_i, a_i) \leq b_{tl}$ for all $l \in \mathcal{L}$

Weakly coupled Markov decision process

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

- $\mathscr{S} = \prod_{i=1}^{n} \mathscr{S}_i$ with $\mathscr{S}_i = \{1, \dots, S_i\}$ is the (finite) state space;
- $\mathscr{A} = \prod_{i=1}^{n} \mathscr{A}_i$ with $\mathscr{A}_i = \{1, \dots, A_i\}$ is the (finite) action space;
- $q(s) = \prod_{i=1}^{n} q_i(s_i)$ is the initial state distribution;
- $p_t(s' \mid s, a) = \prod^{n} p_{t,i}$ … e transition kernel;
- $r_t(s,$ … ewards;
- $T \in$ …

and

$a \in$ … $\in \mathscr{L}$

Objective

find policy $\pi = \mathscr{S} \rightarrow \mathscr{A}$ that

maximizes the expected total rewards:

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

Fluid Linear Program

$$\underset{\sigma,\pi\geq 0}{\text{maximize}} \quad \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{s_i\in\mathcal{S}_i}\sum_{a_i\in\mathcal{A}_i} r_{ti}(s_i,a_i)\cdot\pi_{ti}(s_i,a_i)$$

subject to

$$\sigma_{1i}(s_i) = q_i(s_i) \qquad \forall i, \forall s_i \in \mathcal{S}_i$$

$$\sigma_{t+1,i}(s_i') = \sum_{s_i\in\mathcal{S}_i}\sum_{a_i\in\mathcal{A}_i} p_{ti}(s_i' \mid s_i, a_i)\cdot\pi_{ti}(s_i,a_i) \qquad \forall i, \forall s_i' \in \mathcal{S}_i, \forall t$$

$$\sum_{i=1}^{n}\sum_{s_i\in\mathcal{S}_i}\sum_{a_i\in\mathcal{A}_i} c_{tli}(s_i,a_i)\cdot\pi_{ti}(s_i,a_i) \leq b_{tl} \qquad \forall l, \forall t$$

$$\sum_{a_i\in\mathcal{A}_i} \pi_{ti}(s_i,a_i) = \sigma_{ti}(s_i) \qquad \forall i, \forall s \in \mathcal{S}_i, \forall t$$

Fluid Linear Program

maximize $\sum_{t=1}^{T} \sum_{i=1}^{n} \sum \sum r_{ti}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i)$

$\sigma, \pi \geq 0$

$\sigma_{ti}(s_i)$: % of MDP *i* that is in state s_i in stage *t*

subject to

$\forall i, \forall s_i \in \mathcal{S}_i$

$$\sigma_{t+1,i}(s'_i) = \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} p_{ti}(s'_i \mid s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \quad \forall i, \forall s'_i \in \mathcal{S}_i, \forall t$$

$$\sum_{i=1}^{n} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} c_{tli}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \leq b_{tl} \quad \forall l, \forall t$$

$$\sum_{a_i \in \mathcal{A}_i} \pi_{ti}(s_i, a_i) = \sigma_{ti}(s_i) \quad \forall i, \forall s \in \mathcal{S}_i, \forall t$$

Fluid Linear Program

maximize $\sigma, \pi \geq 0$

$$\sum^{T} \sum^{n} \sum \sum r_{ti}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i)$$

$\sigma_{ti}(s_i)$: % of MDP *i* that is in state s_i in stage *t*

$\pi_{ti}(s_i, a_i)$: % of MDP from $\sigma_{ti}(s_i)$ that we apply action a_i to

subject to

$\forall i, \forall s_i \in \mathcal{S}_i$

$i, \forall s_i' \in \mathcal{S}_i, \forall t$

$$\sum_{i=1} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} c_{tli}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \leq b_{tl} \qquad \forall l, \forall t$$

$$\sum_{a_i \in \mathcal{A}_i} \pi_{ti}(s_i, a_i) = \sigma_{ti}(s_i) \qquad \forall i, \forall s \in \mathcal{S}_i, \forall t$$

Fluid Linear Program

$$\underset{\sigma,\pi\geq 0}{\text{maximize}} \quad \sum_{t=1}^{T}\sum_{i=1}^{n}\sum_{s_i\in\mathcal{S}_i}\sum_{a_i\in\mathcal{A}_i} r_{ti}(s_i,a_i)\cdot\pi_{ti}(s_i,a_i)$$

objective function:
maximize rewards

subject to

$$\sigma_{1i}(s_i) = q_i(s_i) \qquad \forall i, \forall s_i \in \mathcal{S}_i$$

$$\sigma_{t+1,i}(s_i') = \sum_{s_i\in\mathcal{S}_i}\sum_{a_i\in\mathcal{A}_i} p_{ti}(s_i' \mid s_i, a_i)\cdot\pi_{ti}(s_i,a_i) \qquad \forall i, \forall s_i' \in \mathcal{S}_i, \forall t$$

$$\sum_{i=1}^{n}\sum_{s_i\in\mathcal{S}_i}\sum_{a_i\in\mathcal{A}_i} c_{tli}(s_i,a_i)\cdot\pi_{ti}(s_i,a_i) \leq b_{tl} \qquad \forall l, \forall t$$

$$\sum_{a_i\in\mathcal{A}_i} \pi_{ti}(s_i,a_i) = \sigma_{ti}(s_i) \qquad \forall i, \forall s \in \mathcal{S}_i, \forall t$$

Fluid Linear Program

$$\underset{\sigma, \pi \geq 0}{\text{maximize}} \quad \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} r_{ti}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i)$$

initial states: must follow q

subject to

$$\sigma_{1i}(s_i) = q_i(s_i) \qquad \forall i, \forall s_i \in \mathcal{S}_i$$

$$\sigma_{t+1,i}(s_i') = \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} p_{ti}(s_i' \mid s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \qquad \forall i, \forall s_i' \in \mathcal{S}_i, \forall t$$

$$\sum_{i=1}^{n} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} c_{tli}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \leq b_{tl} \qquad \forall l, \forall t$$

$$\sum_{a_i \in \mathcal{A}_i} \pi_{ti}(s_i, a_i) = \sigma_{ti}(s_i) \qquad \forall i, \forall s \in \mathcal{S}_i, \forall t$$

Fluid Linear Program

$$\underset{\sigma,\pi \geq 0}{\text{maximize}} \quad \sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} r_{ti}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i)$$

subject to

$$\sigma_{1i}(s_i) = q_i(s_i) \qquad \forall i, \forall s_i \in \mathcal{S}_i$$

$$\sigma_{t+1,i}(s_i') = \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} p_{ti}(s_i' \mid s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \qquad \forall i, \forall s_i' \in \mathcal{S}_i, \forall t$$

$$\sum_{i=1}^{n} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} c_{tli}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \leq b_{tl} \qquad \forall l, \forall t$$

$$\sum_{a_i \in \mathcal{A}_i} \pi_{ti}(s_i, a_i) = \sigma_{ti}(s_i) \qquad \forall i, \forall s \in \mathcal{S}_i, \forall t$$

transitions:
must follow p

Fluid Linear Program

resources: budgets must be kept

maximize $\sigma, \pi \geq 0$

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} r_{ti}(s_i, a_i) \cdot \pi_{ti}(s_i, a$$

subject to

$$\sigma_{1i}(s_i) = q_i(s_i) \qquad \forall i, \forall s_i \in \mathcal{S}_i$$

$$\sigma_{t+1,i}(s'_i) = \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} p_{ti}(s'_i \mid s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \qquad \forall i, \forall s'_i \in \mathcal{S}_i, \forall t$$

$$\sum_{i=1}^{n} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} c_{tli}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \leq b_{tl} \qquad \forall l, \forall t$$

$$\sum_{a_i \in \mathcal{A}_i} \pi_{ti}(s_i, a_i) = \sigma_{ti}(s_i) \qquad \forall i, \forall s \in \mathcal{S}_i, \forall t$$

Fluid Linear Program

maximize $\sigma, \pi \geq 0$

$$\sum_{t=1}^{T} \sum_{i=1}^{n} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} r_{ti}(s_i, a_i) \cdot \pi_{ti}(s_i,$$

"flow preservation":
we cannot "drop" MDPs

subject to

$$\sigma_{1i}(s_i) = q_i(s_i) \qquad \forall i, \forall s_i \in \mathcal{S}_i$$

$$\sigma_{t+1,i}(s_i') = \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} p_{ti}(s_i' \mid s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \qquad \forall i, \forall s_i' \in \mathcal{S}_i, \forall t$$

$$\sum_{i=1}^{n} \sum_{s_i \in \mathcal{S}_i} \sum_{a_i \in \mathcal{A}_i} c_{tli}(s_i, a_i) \cdot \pi_{ti}(s_i, a_i) \leq b_{tl} \qquad \forall l, \forall t$$

$$\sum_{a_i \in \mathcal{A}_i} \pi_{ti}(s_i, a_i) = \sigma_{ti}(s_i) \qquad \forall i, \forall s \in \mathcal{S}_i, \forall t$$

Observation

The fluid LP constitutes a relaxation of the weakly coupled MDP.

Observation

The fluid LP constitutes a relaxation of the weakly coupled MDP.

Randomized policy

For each MDP i, take action a_i in state s_i with probability $\dfrac{\pi_{ti}(s_i, a_i)}{\sigma_{ti}(s_i)}$ at time t.

Observation

The fluid LP constitutes a relaxation of the weakly coupled MDP.

Randomized policy

For each MDP i, take action a_i in state s_i with probability $\dfrac{\pi_{ti}(s_i, a_i)}{\sigma_{ti}(s_i)}$ at time t.

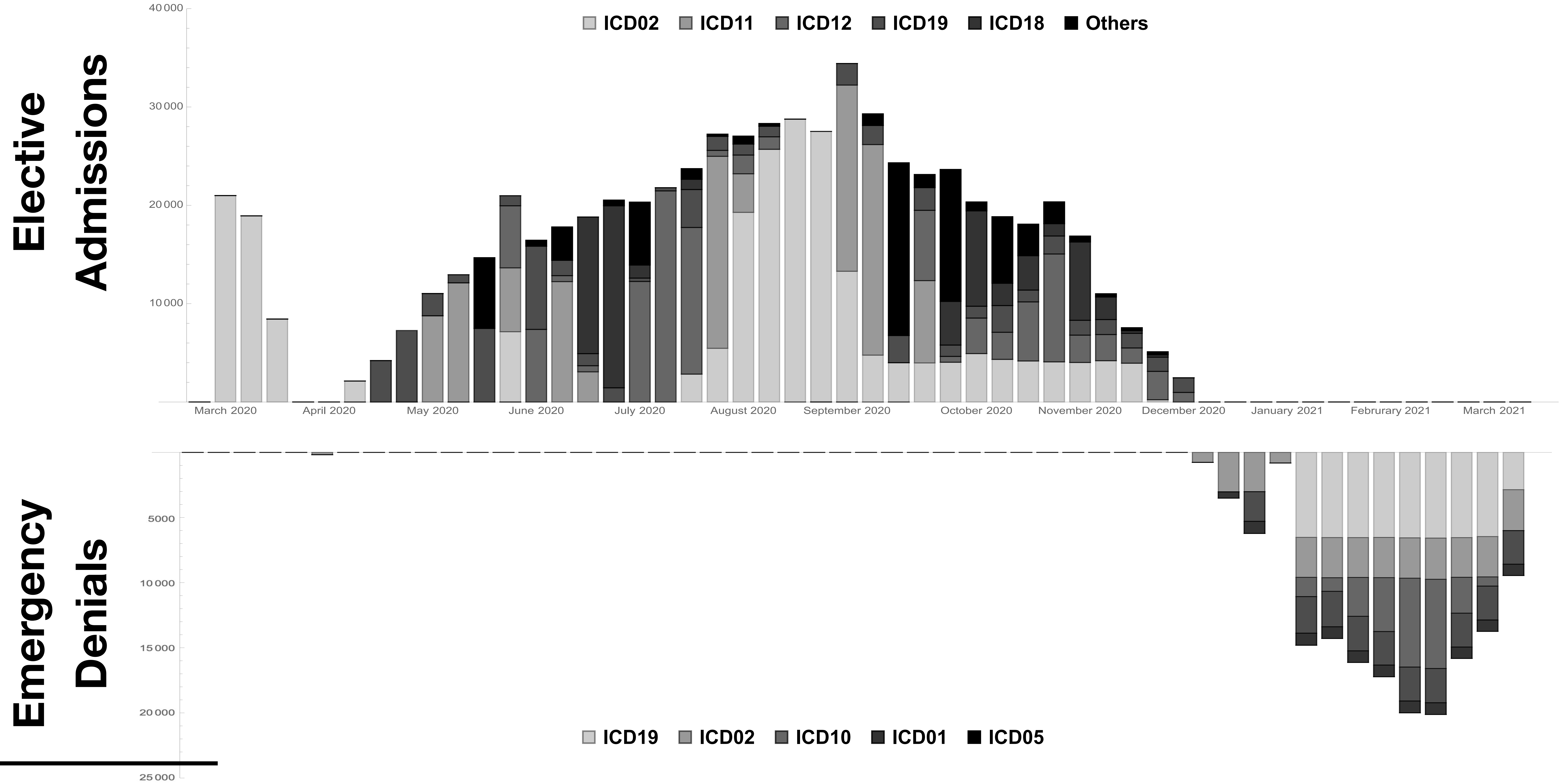
Performance guarantee

For suitably adapted b_{tl}, the randomized policy is guaranteed to be feasible in the weakly coupled MDP. Moreover, the relative optimality gap for large MDPs is:

$$T \cdot \sqrt{\frac{\log n}{n}} + \frac{T^2 L}{n^2} \quad \underset{n\to\infty}{\longrightarrow} \quad 0$$

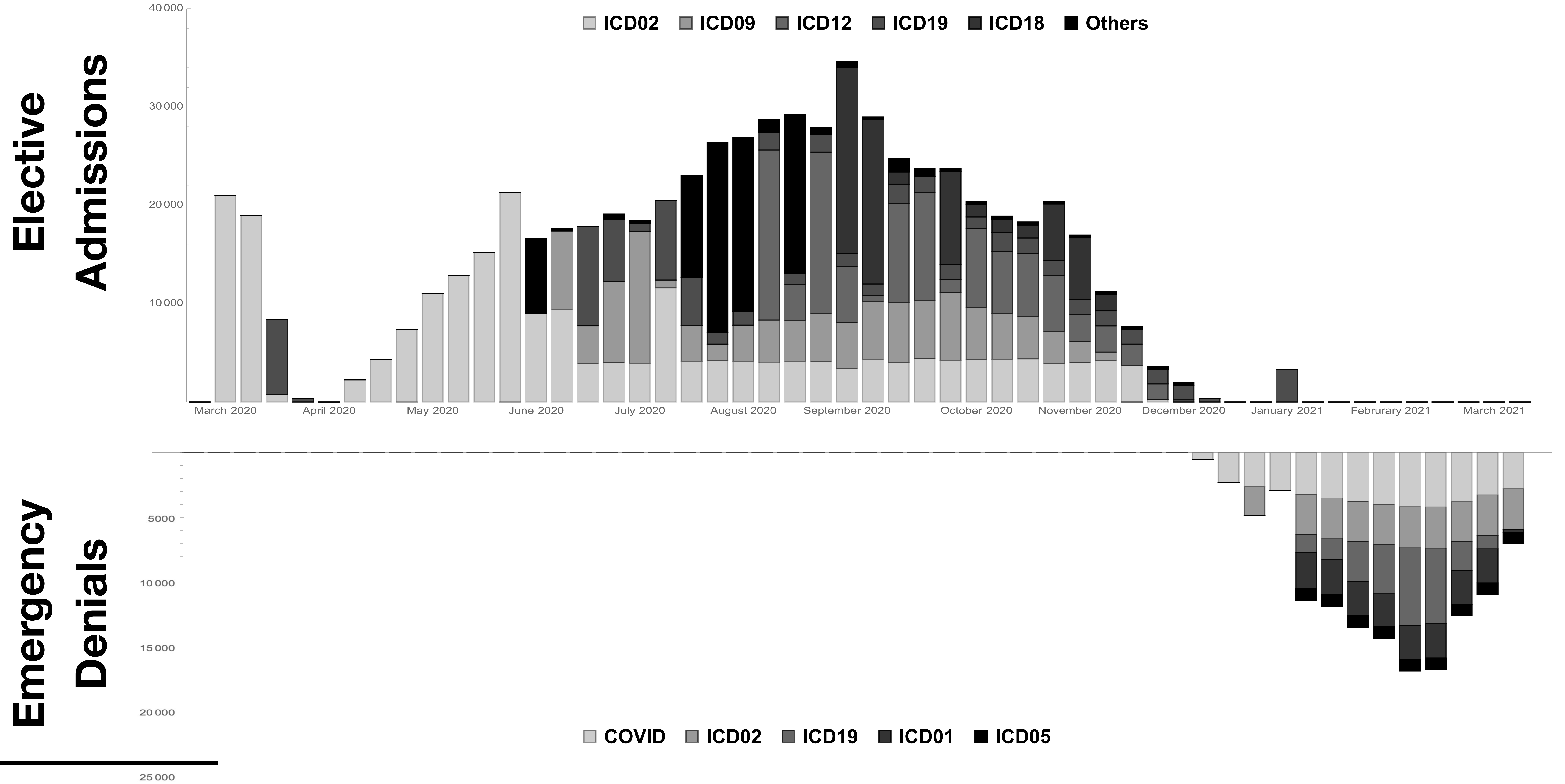
Case Study: National Patient Prioritization

Simulation of Government Policy



D'Aeth et al. (2021), *Optimal National Prioritization Policies for Hospital Care During the SARS-CoV-2 Pandemic*.

Optimized Schedule

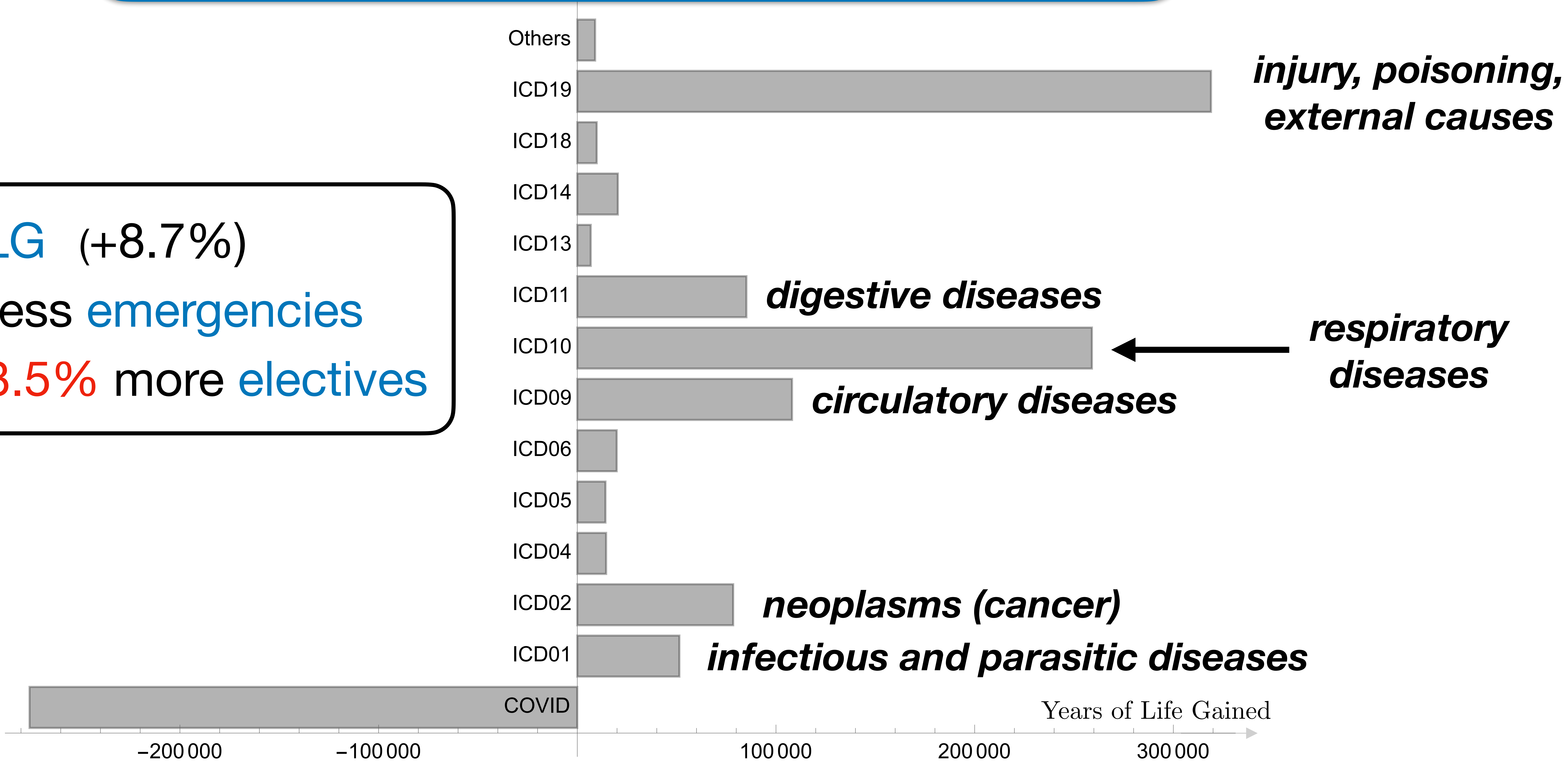


D'Aeth et al. (2021), *Optimal National Prioritization Policies for Hospital Care During the SARS-CoV-2 Pandemic*.

Case Study: National Patient Prioritization

Years of Life Gained by Optimized Schedule

- 720k YLG (+8.7%)
- 22.1% less emergencies
- up to 53.5% more electives



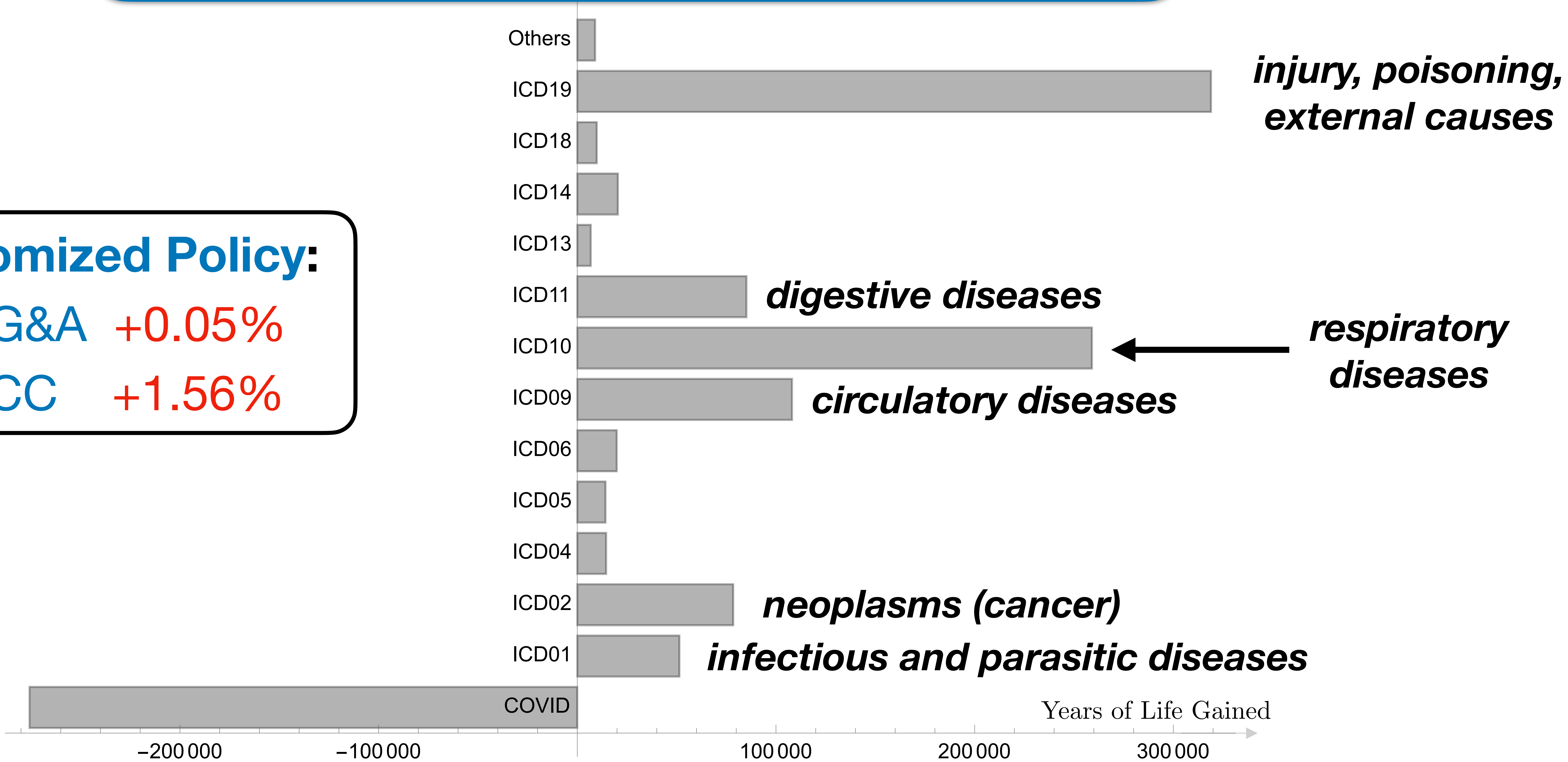
D'Aeth et al. (2021), *Optimal National Prioritization Policies for Hospital Care During the SARS-CoV-2 Pandemic*.

Case Study: National Patient Prioritization

Years of Life Gained by Optimized Schedule

Randomized Policy:

- G&A +0.05%
- CC +1.56%



D'Aeth et al. (2021), *Optimal National Prioritization Policies for Hospital Care During the SARS-CoV-2 Pandemic*.