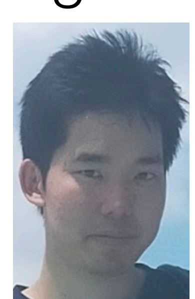
On the Curses of Horizon in Off-policy Evaluation in non-Markov Environments

Nan Jiang
University of Illinois at Urbana-Champaign
September 2024
@Simons Institute

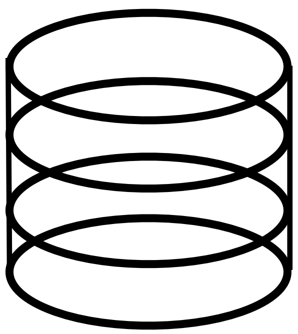
Based on: (1) Uehara et al. NeurIPS 2023. https://arxiv.org/pdf/2207.13081.pdf
(2) Zhang and Jiang. 2024. https://arxiv.org/pdf/2402.14703.pdf


Masatoshi Uehara

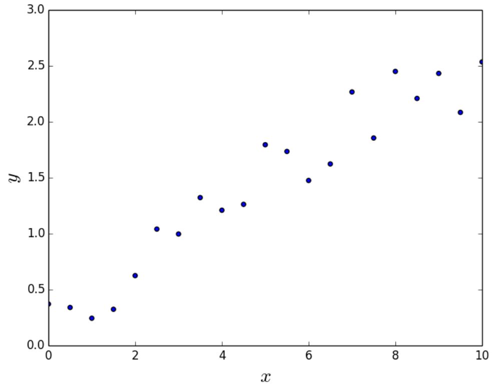

Yuheng Zhang

Generalization in Prediction

- How do we know if an algorithm generalizes?

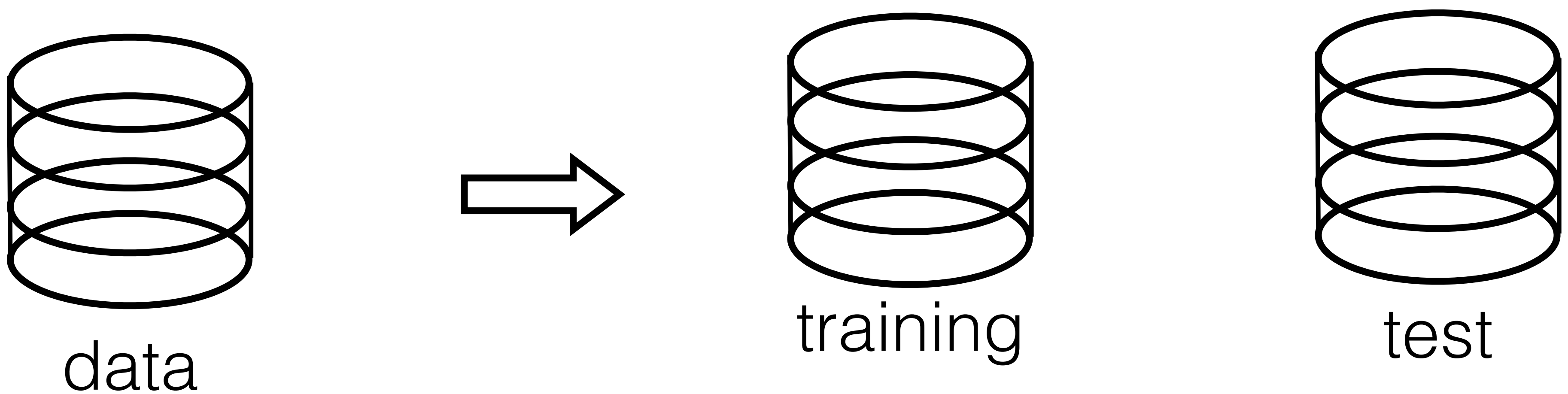


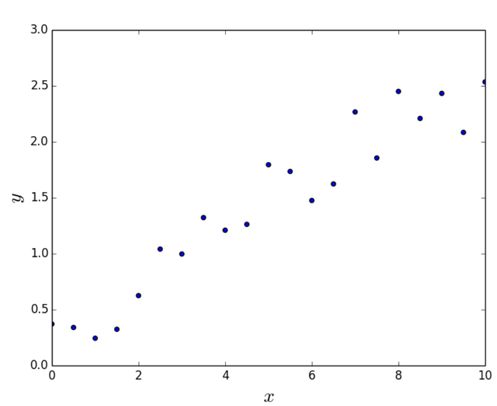
data



Generalization in Prediction

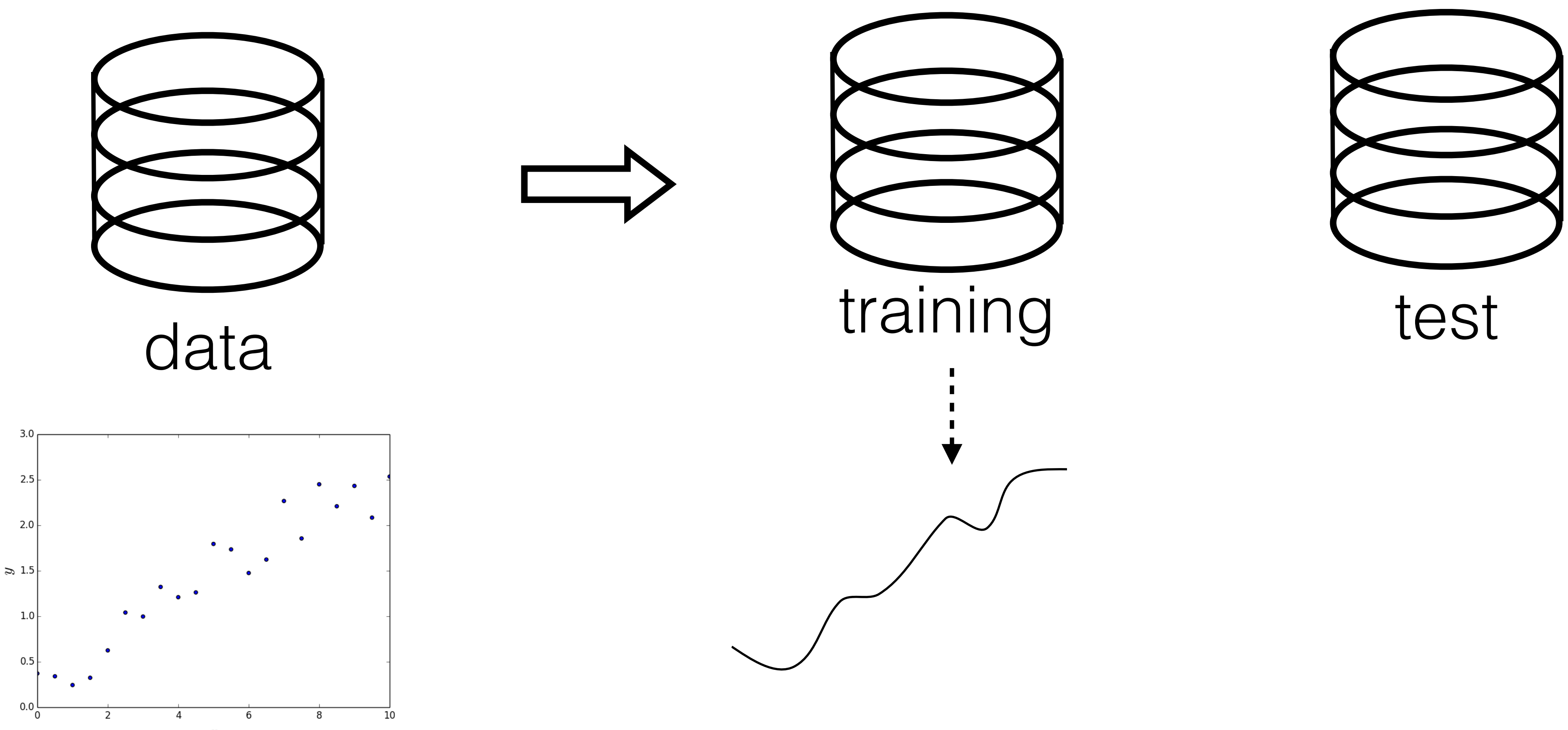
- How do we know if an algorithm generalizes?





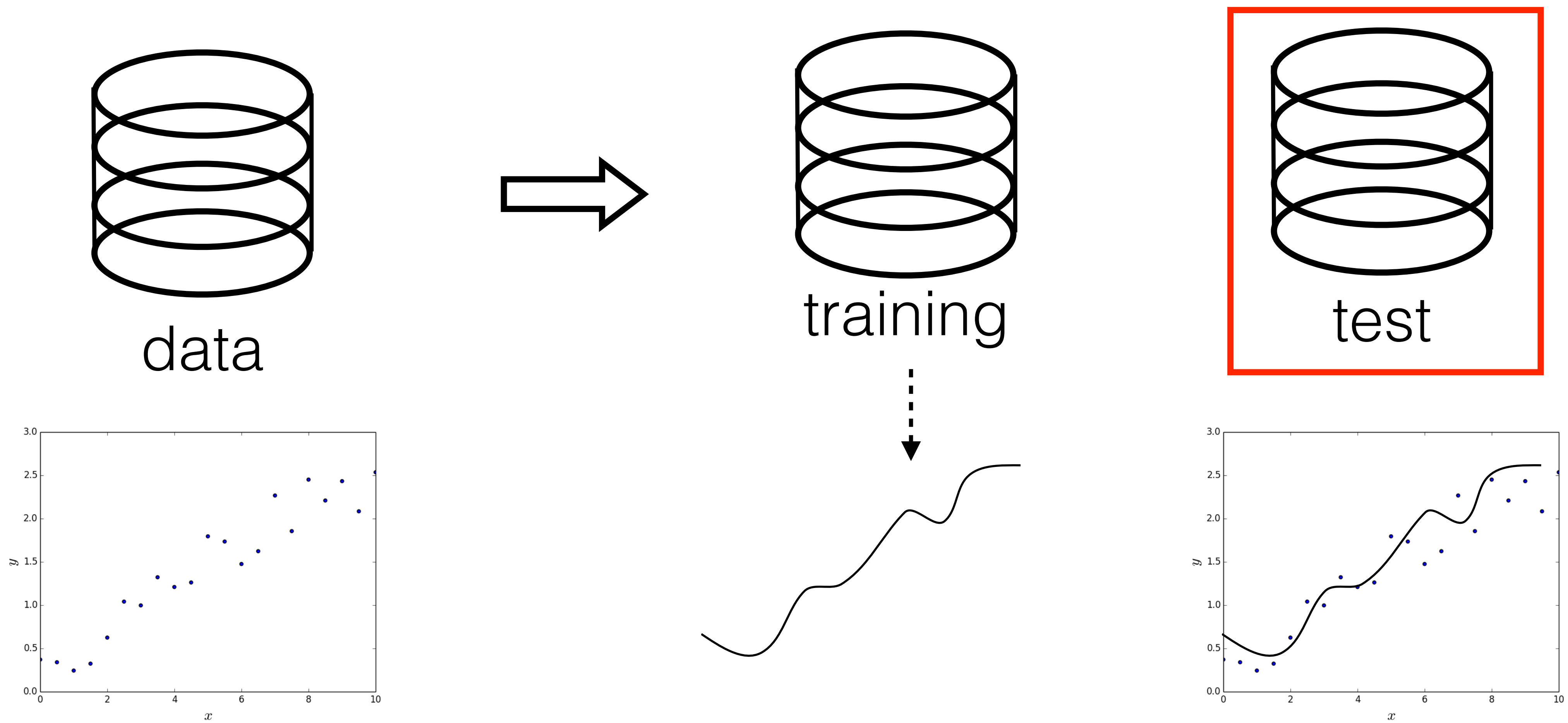
Generalization in Prediction

- How do we know if an algorithm generalizes?



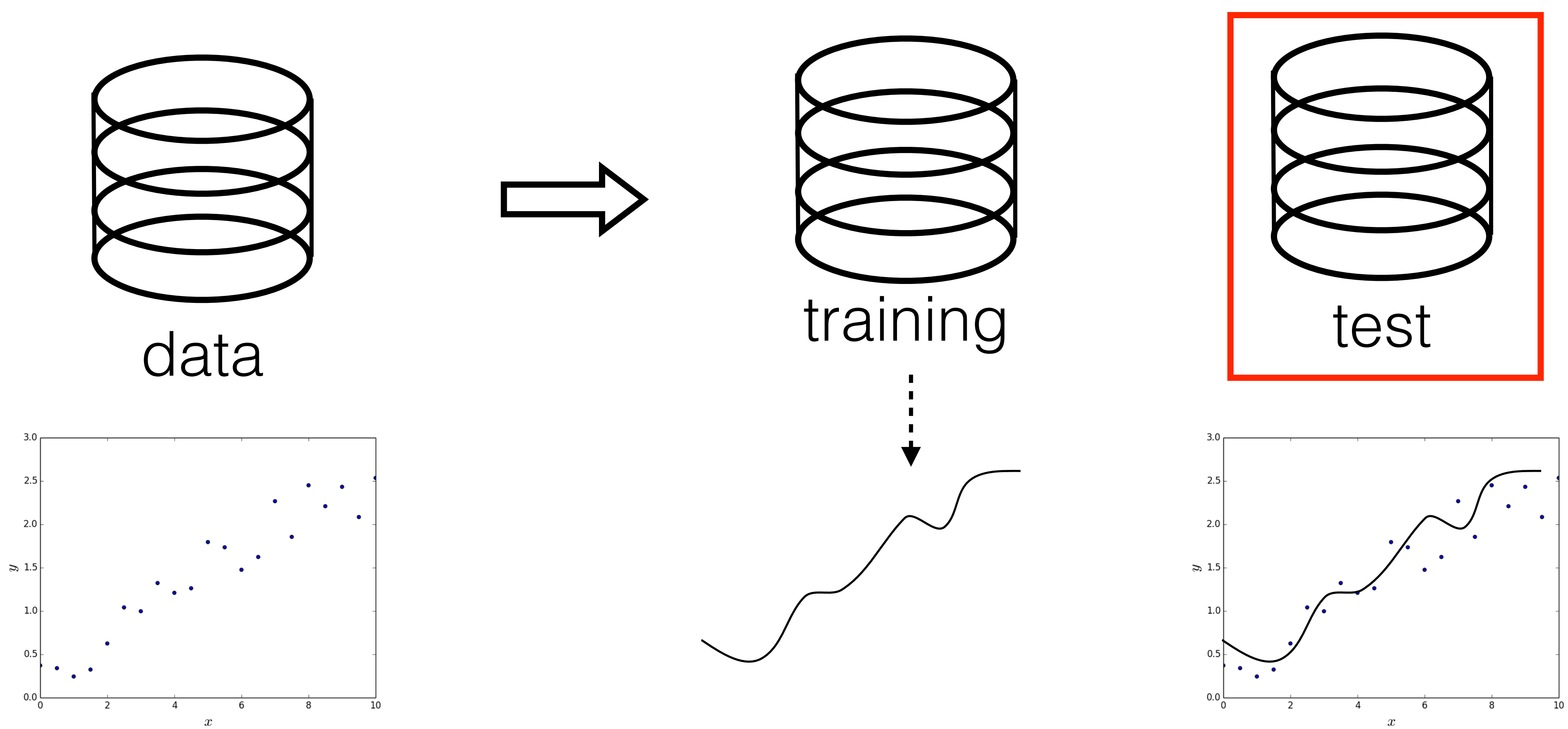
Generalization in Prediction

- How do we know if an algorithm generalizes?



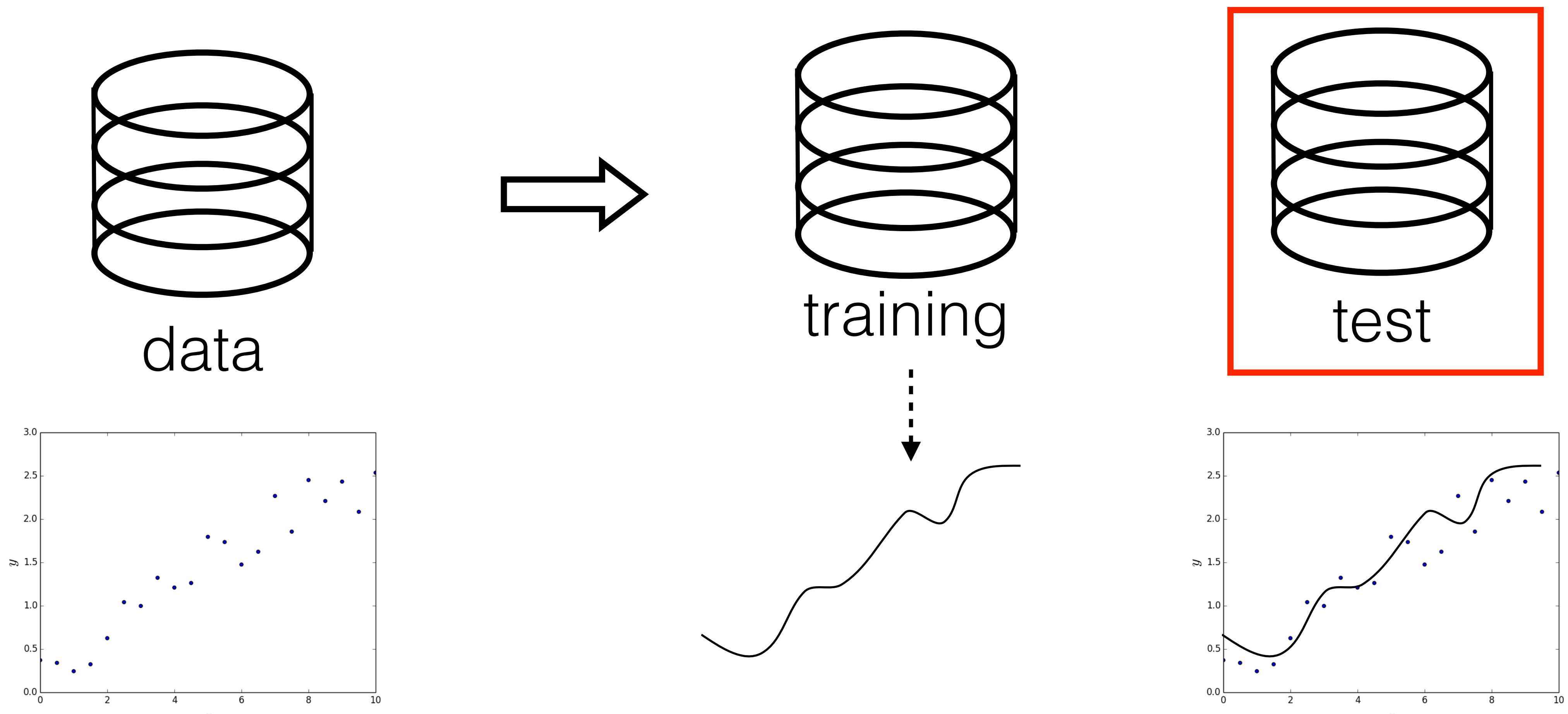
Generalization in Prediction

- How do we know if an algorithm generalizes?
    - Training error ≈ test error (calculated from test data)



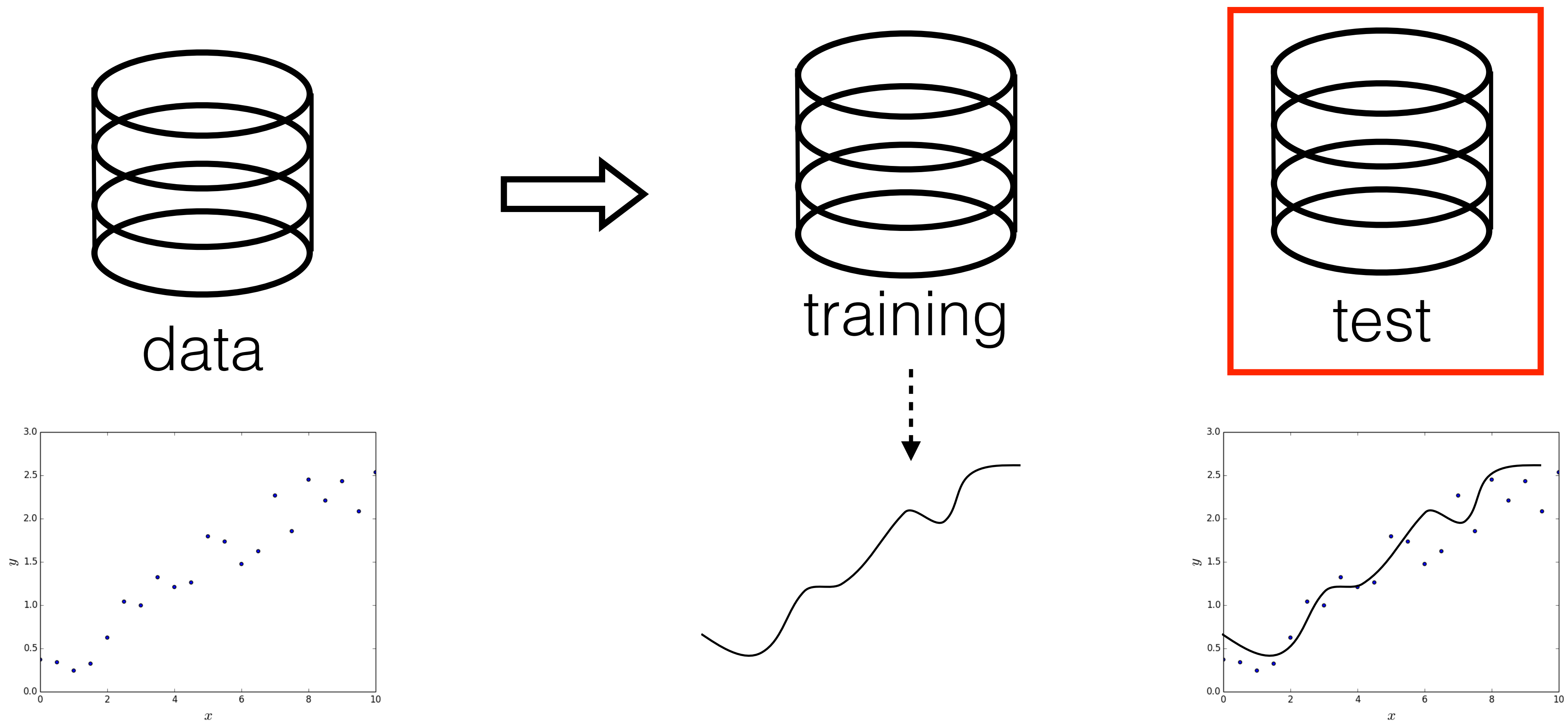
Generalization in Prediction

- How do we know if an algorithm generalizes?
  - Training error ≈ test error (calculated from test data)
- When generalization happens?

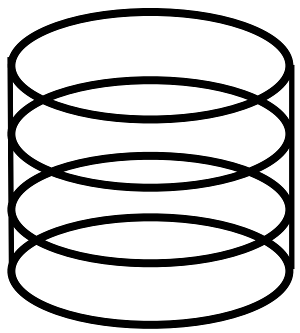


Generalization in Prediction

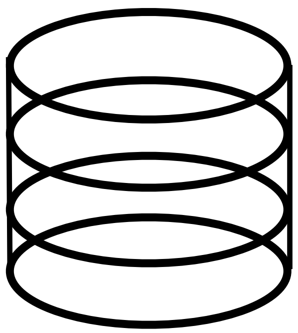
- How do we know if an algorithm generalizes?
  - Training error ≈ test error (calculated from test data)
- When generalization happens?
  - Sufficient training data (> capacity of function class)



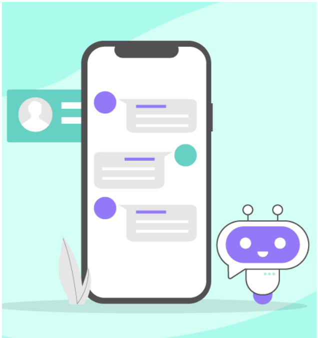
Generalization in Decision-Making

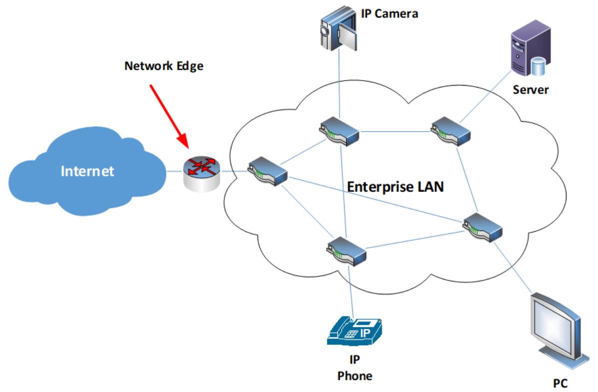


Generalization in Decision-Making



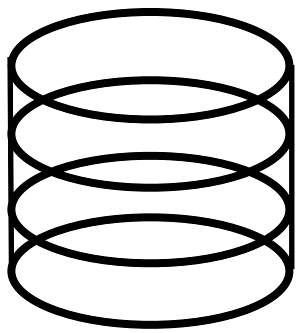
data





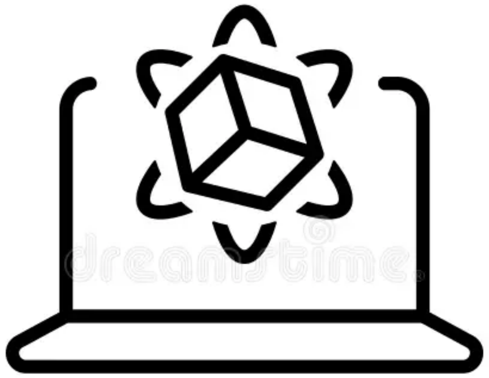
system

Generalization in Decision-Making

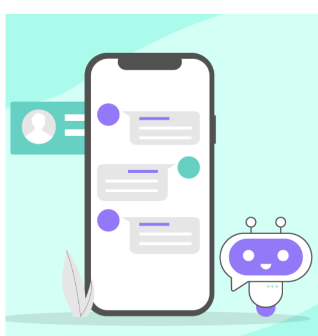


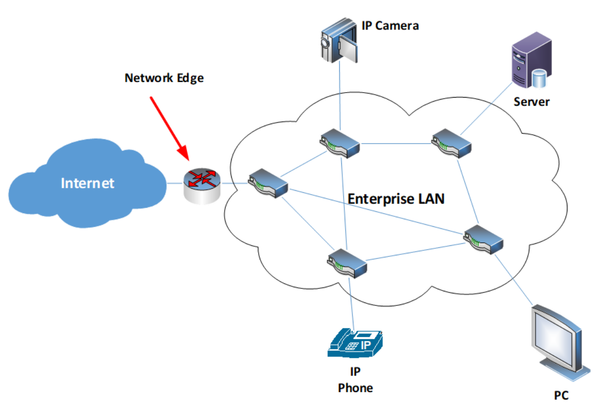
data

fit



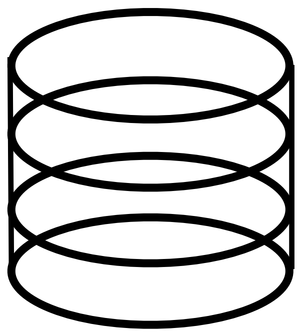
simulator





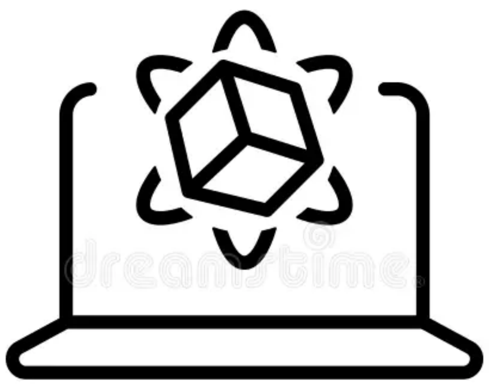
system

Generalization in Decision-Making



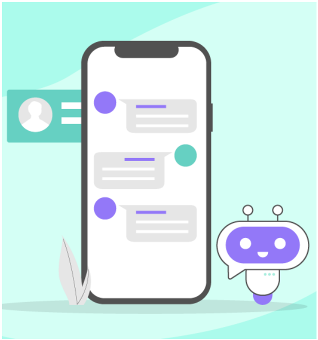
data

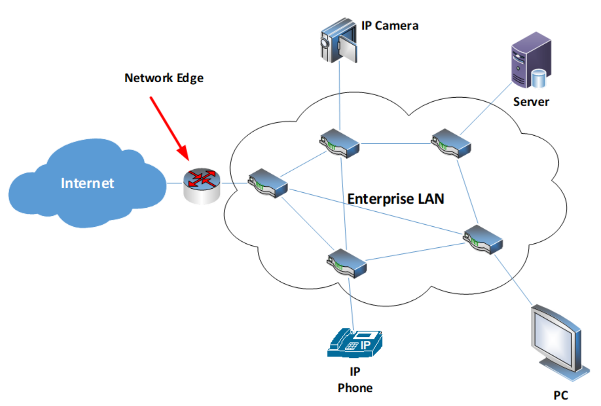
fit



simulator

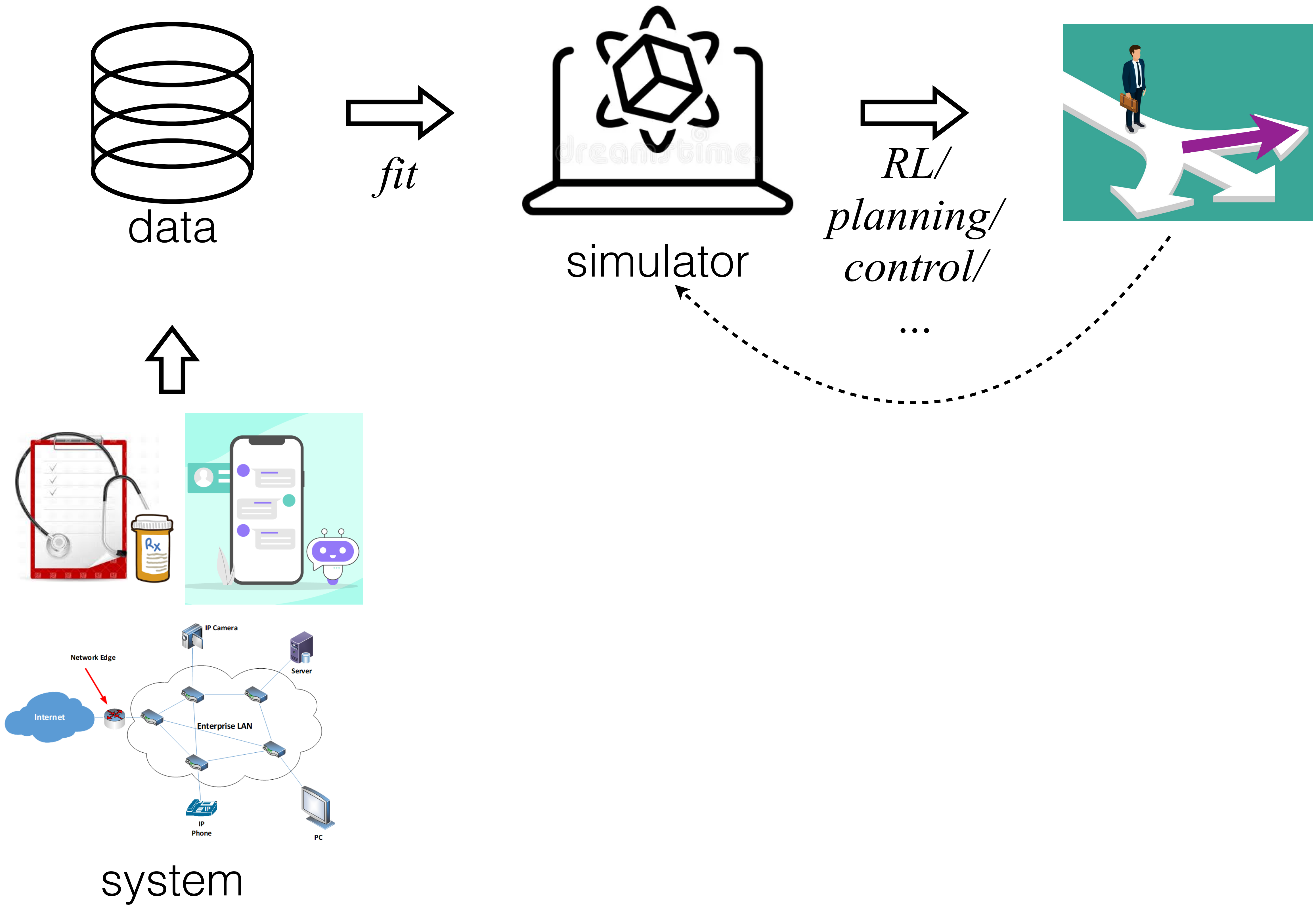
RL/
planning/
control/
...



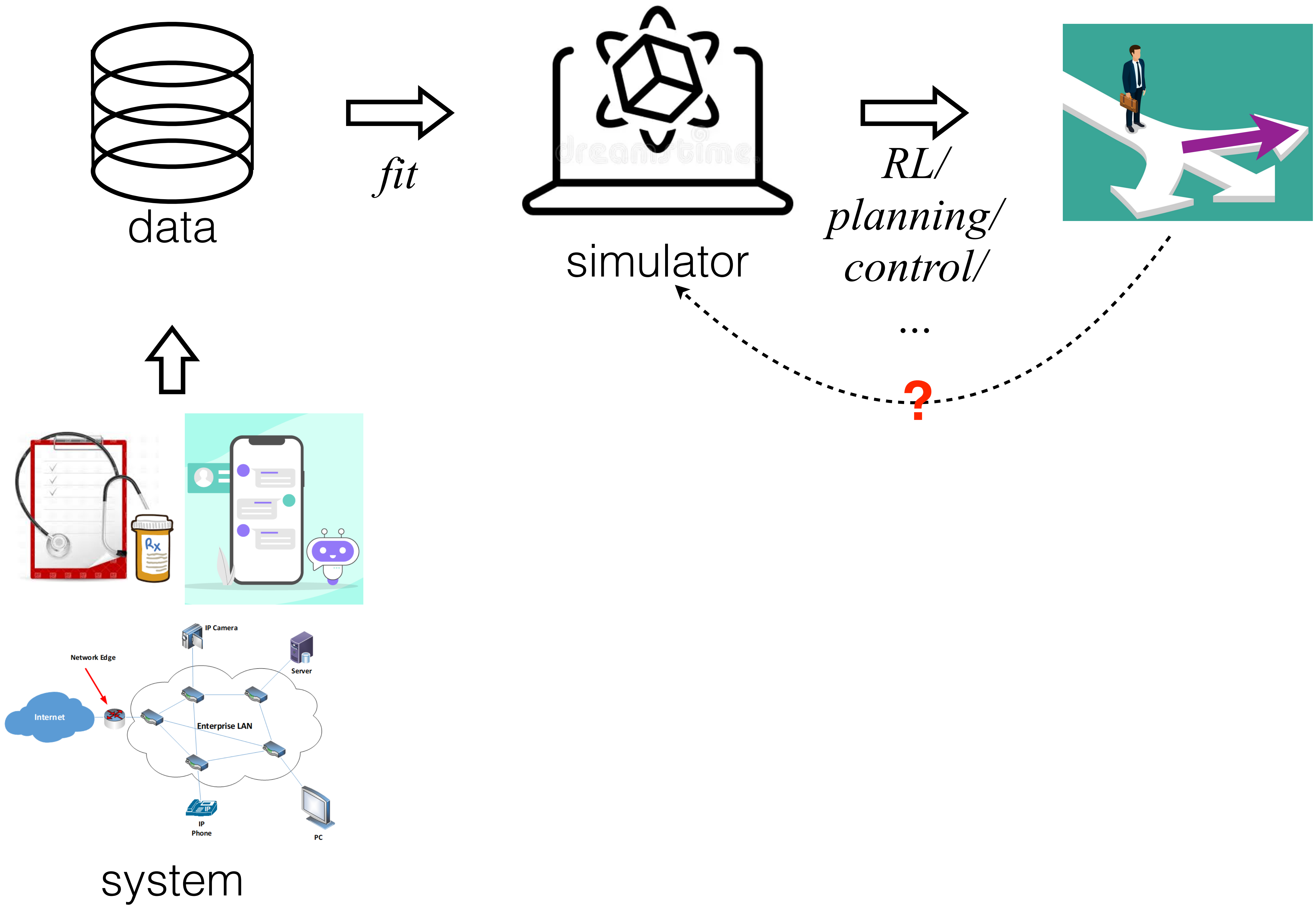


system

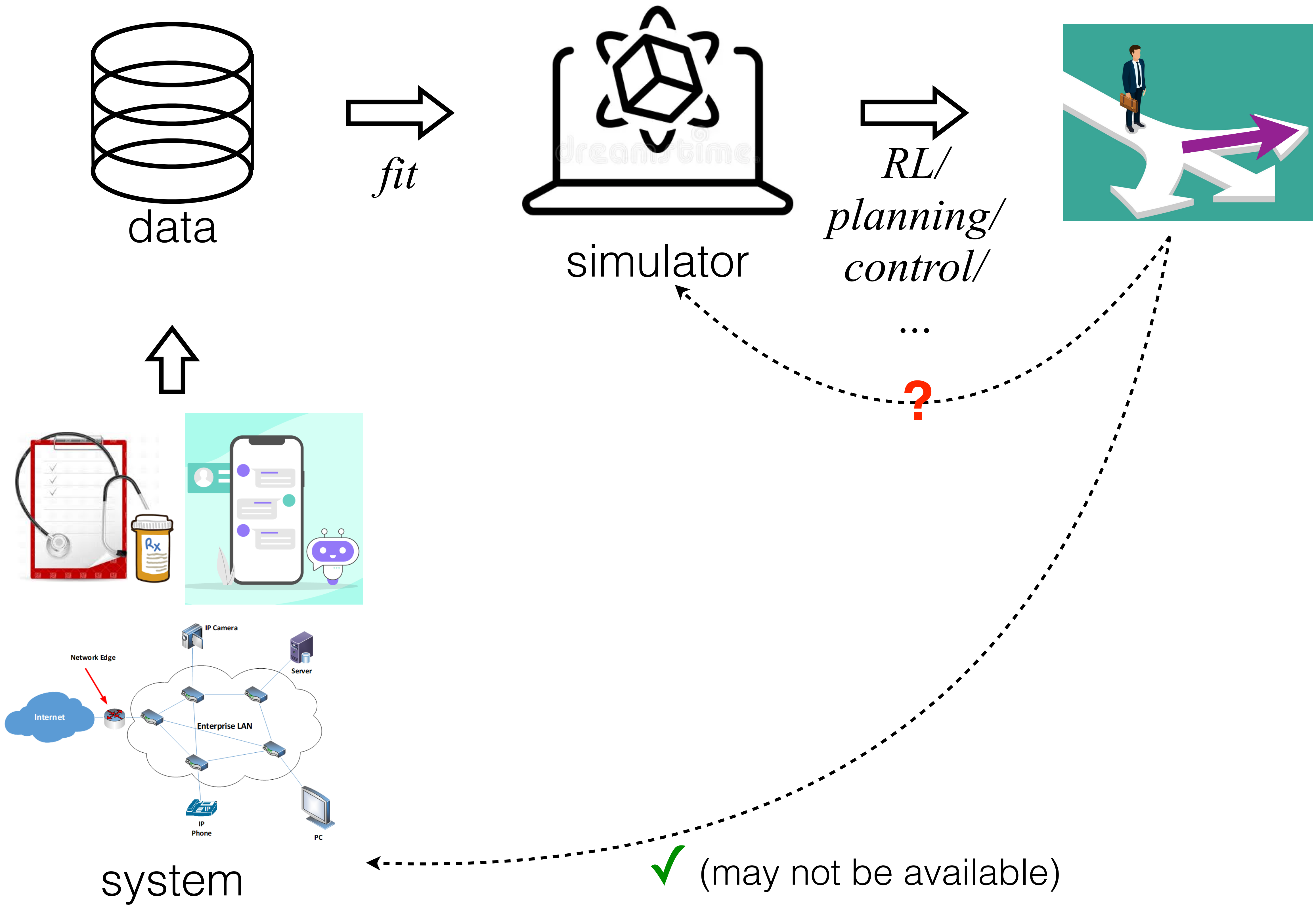
Generalization in Decision-Making



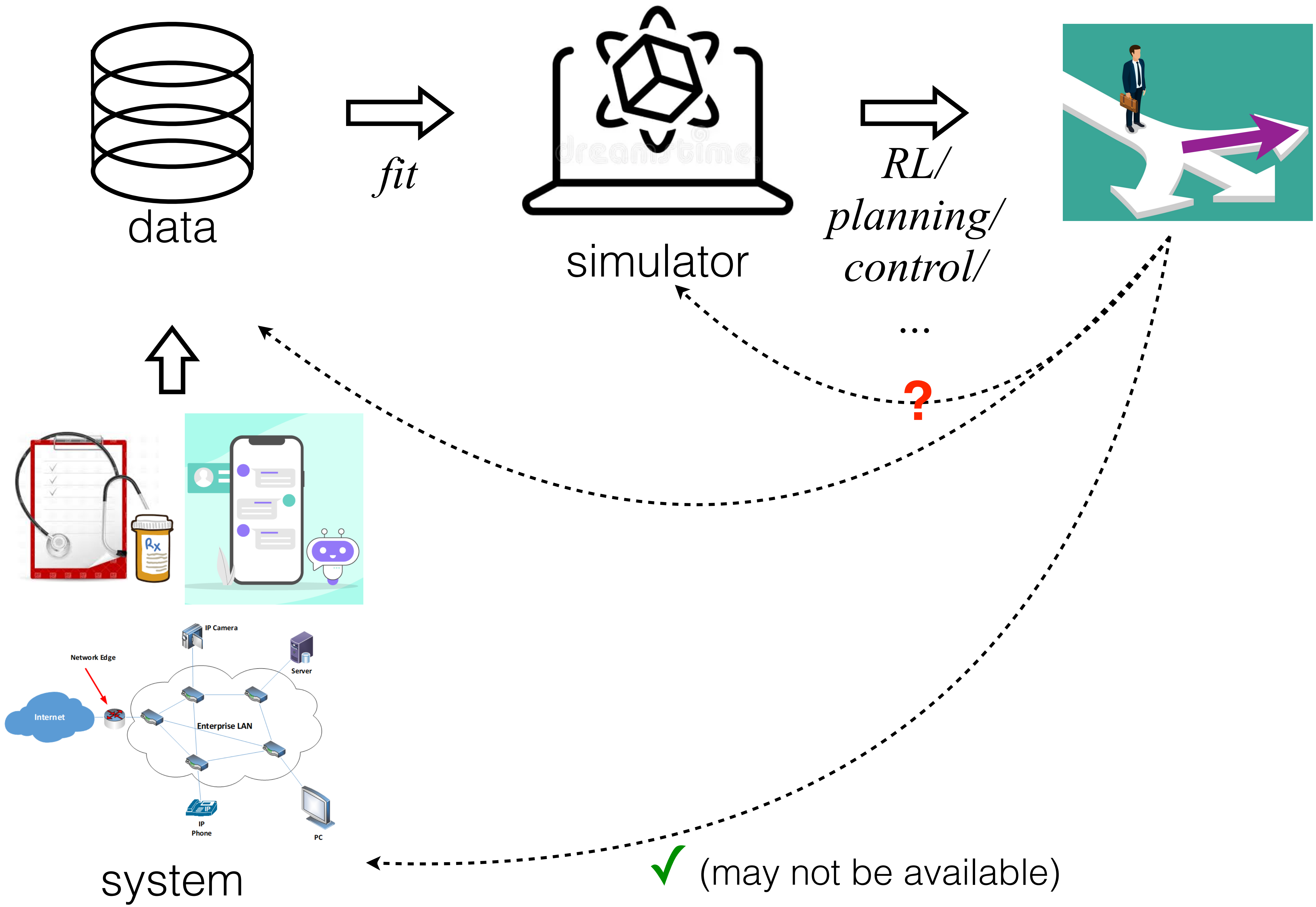
Generalization in Decision-Making



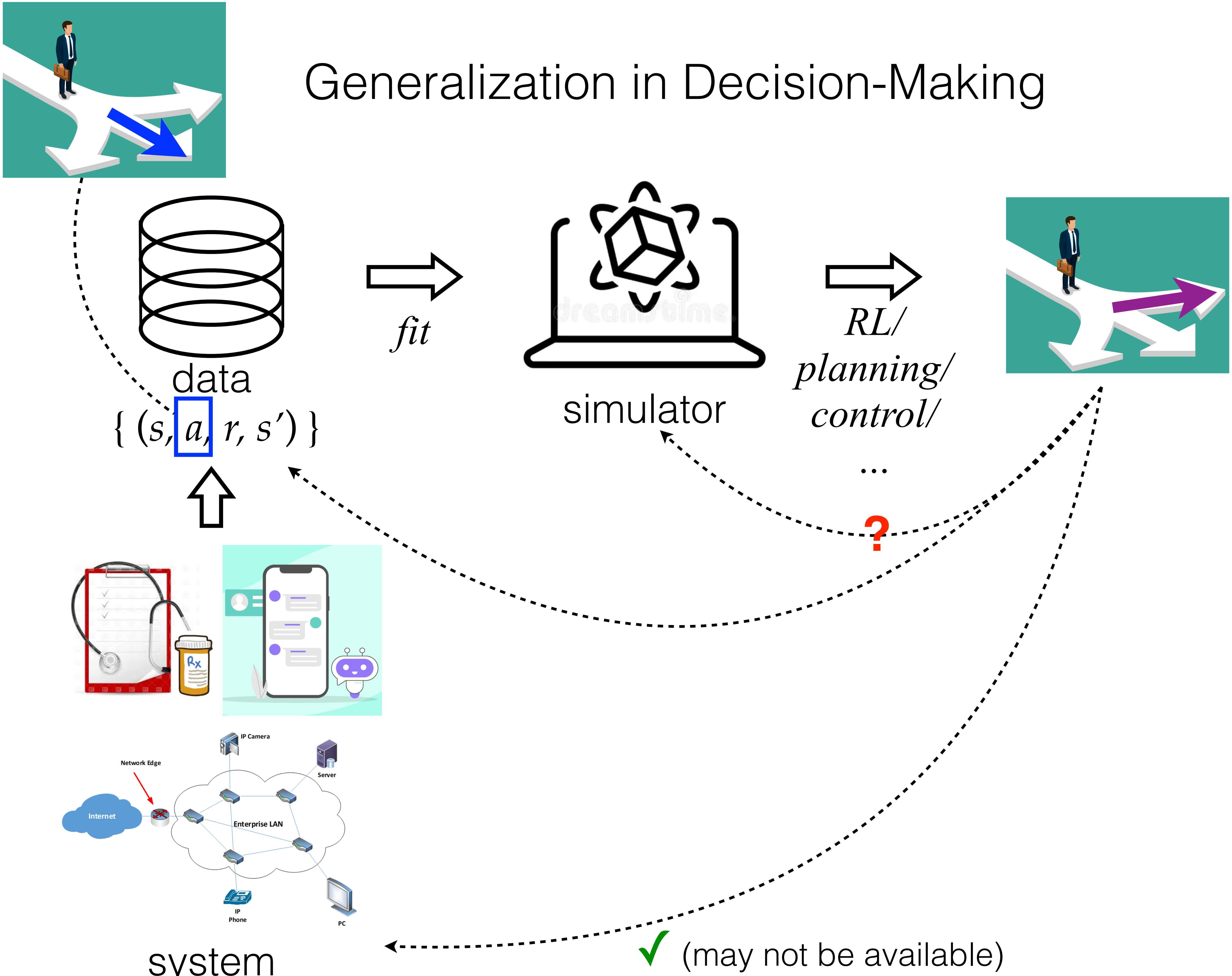
Generalization in Decision-Making



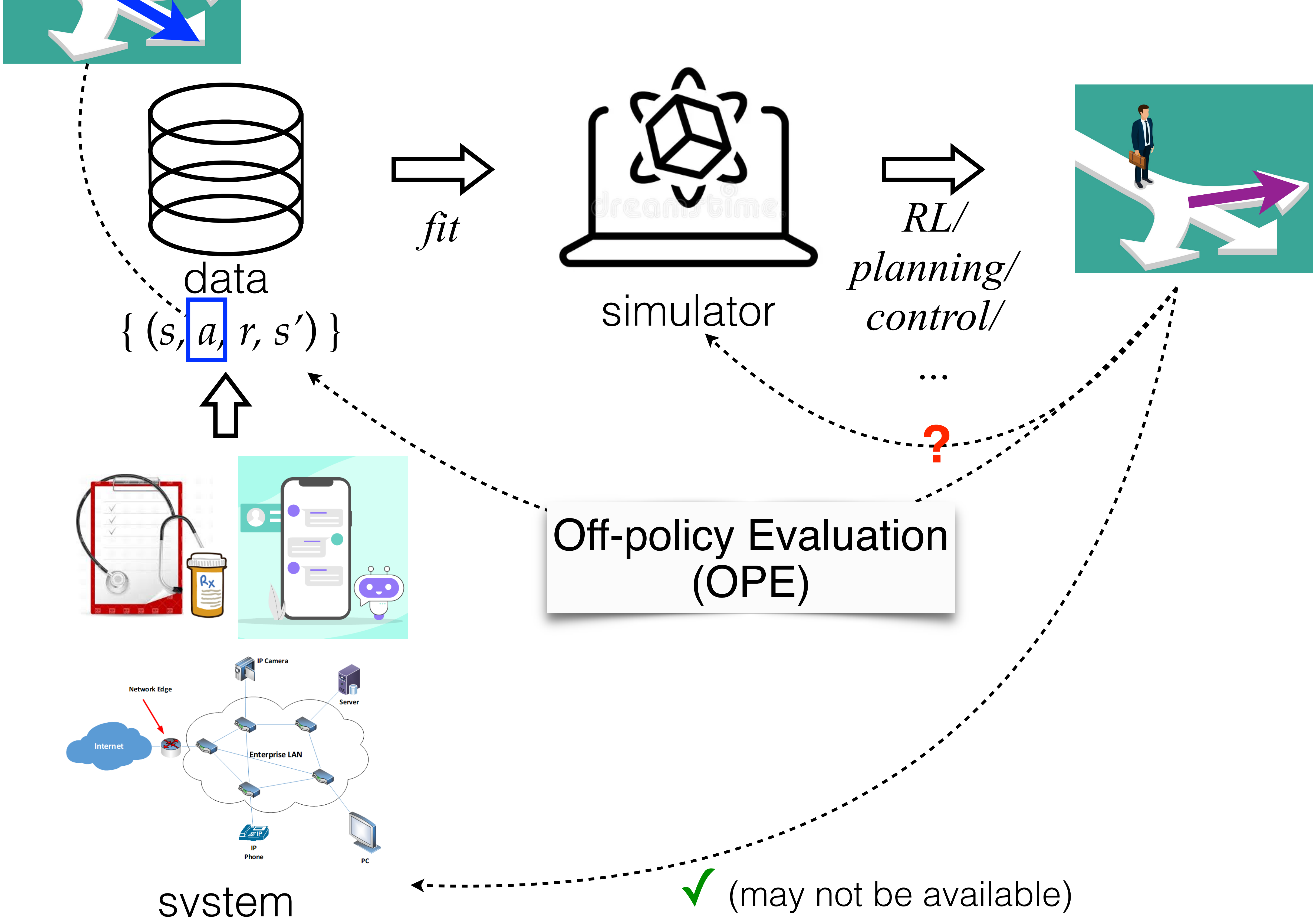
Generalization in Decision-Making



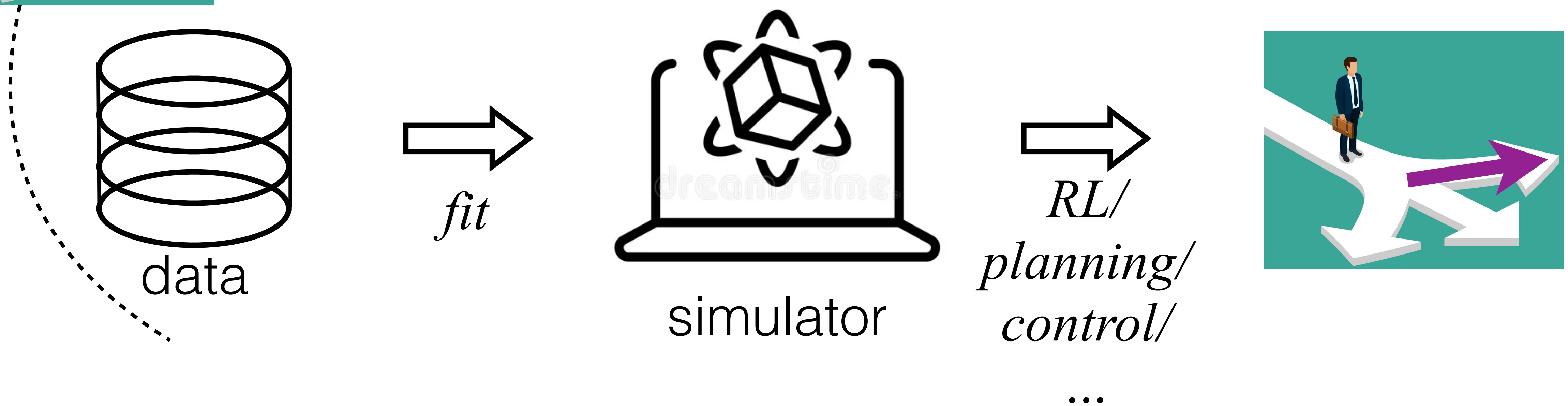
Generalization in Decision-Making



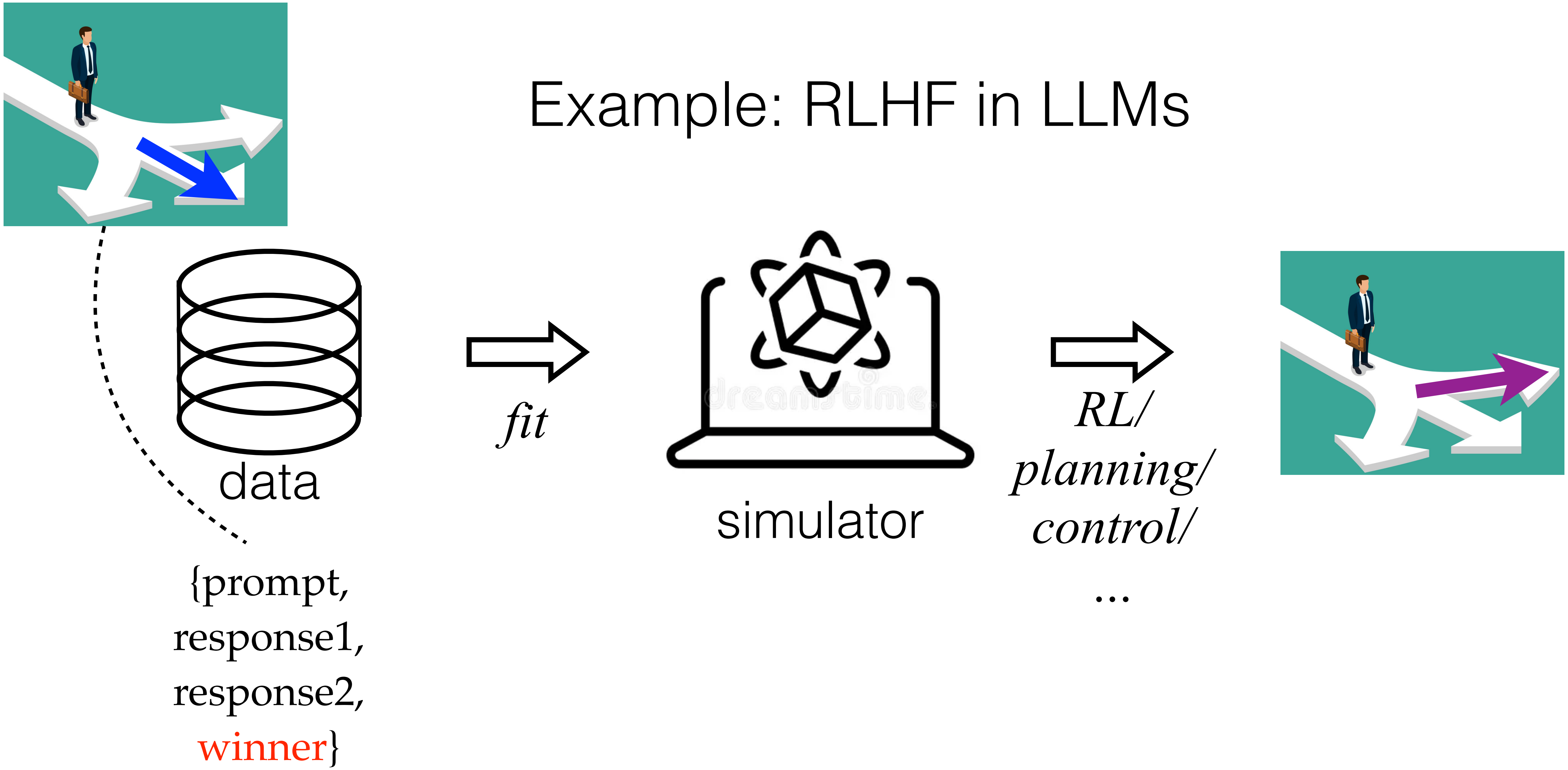
Generalization in Decision-Making



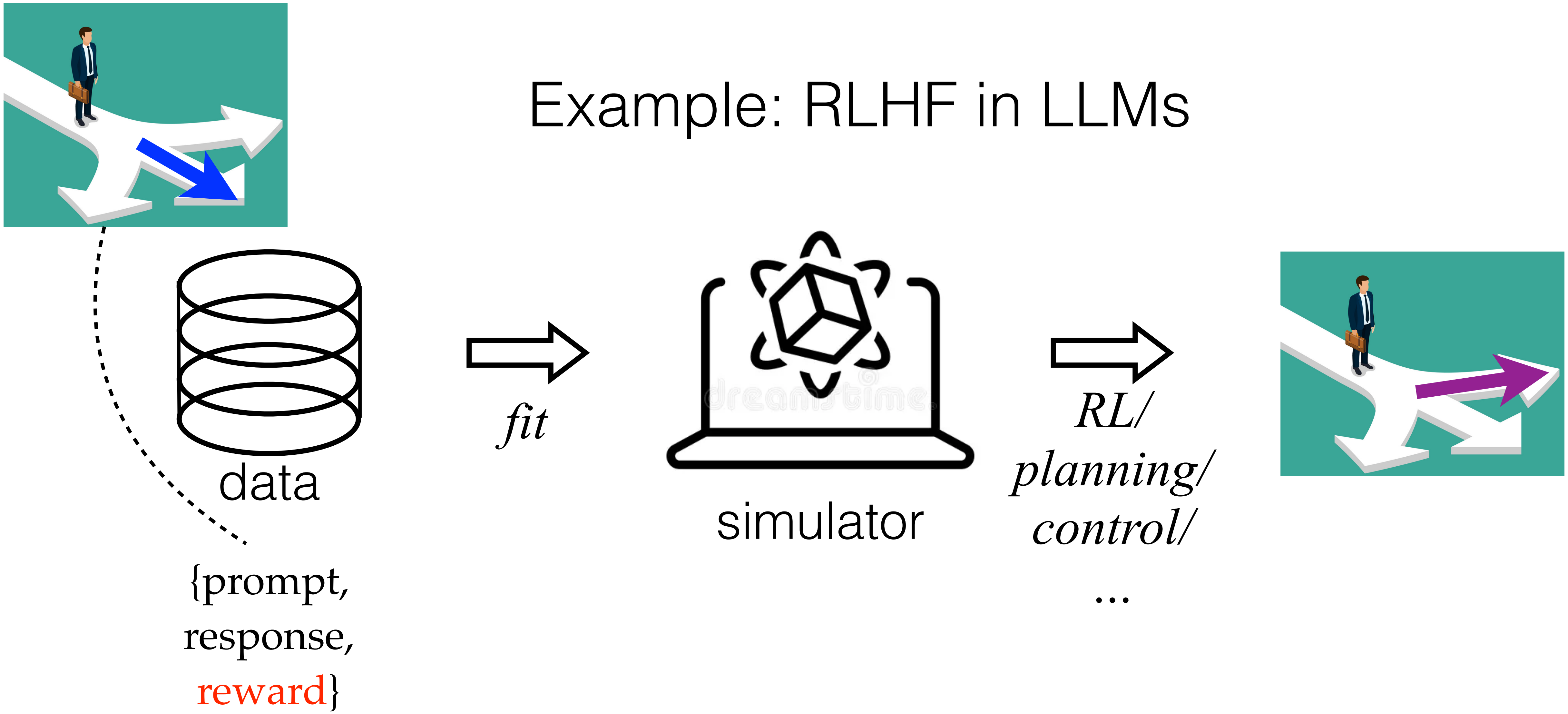
Example: RLHF in LLMs



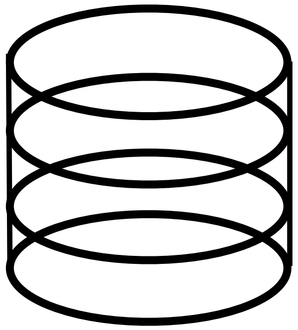
Example: RLHF in LLMs



Example: RLHF in LLMs



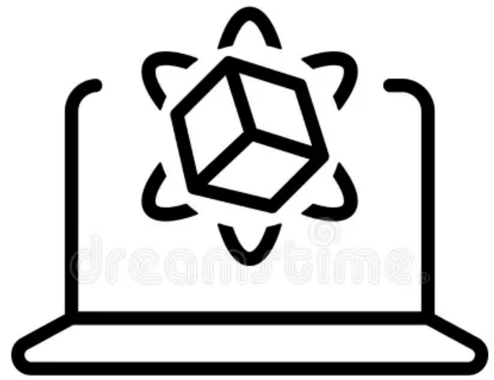
Example: RLHF in LLMs



data

{prompt, response, reward}

fit

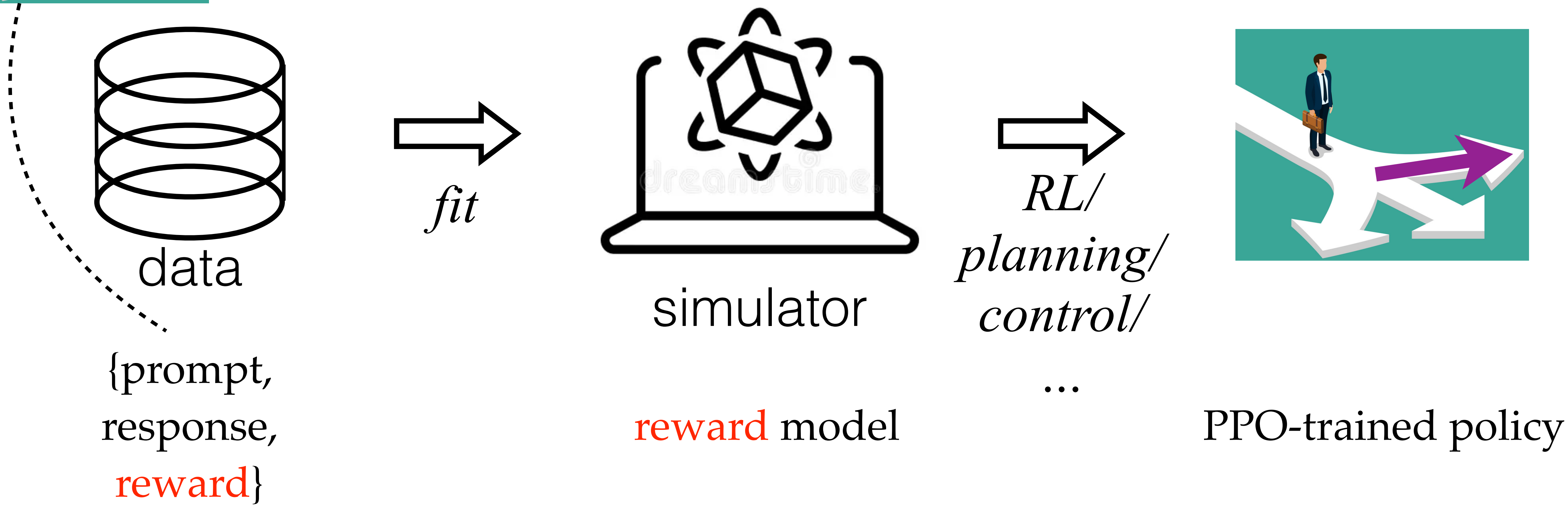


simulator

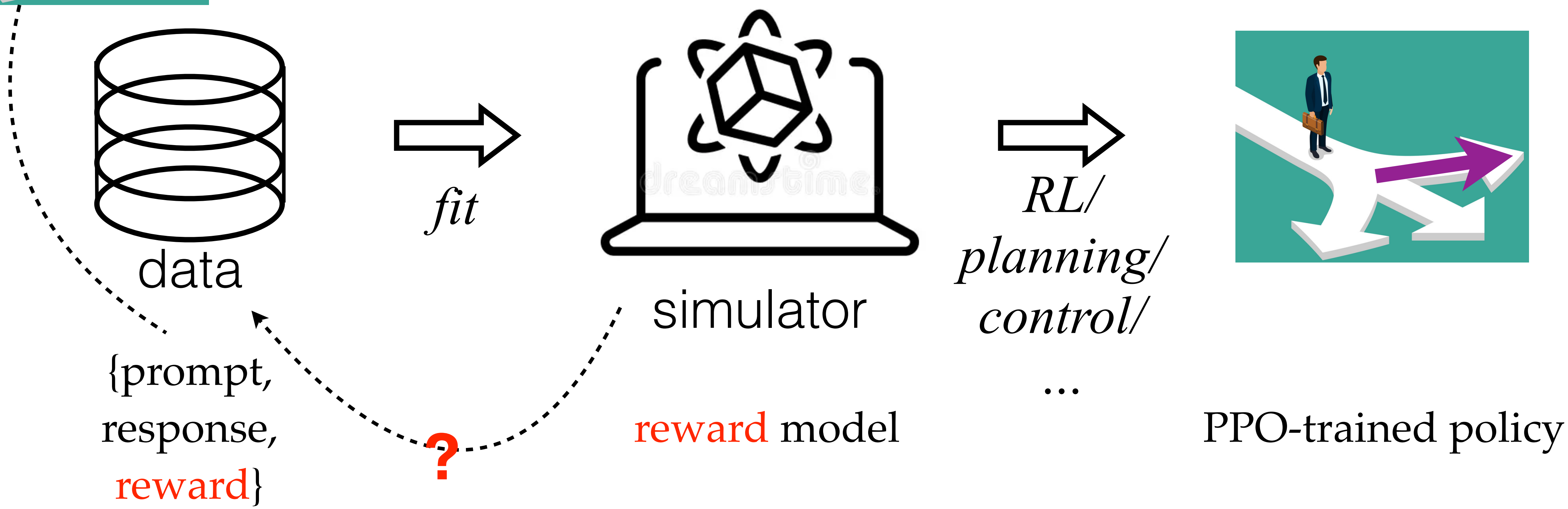
reward model

RL/ planning/ control/ …

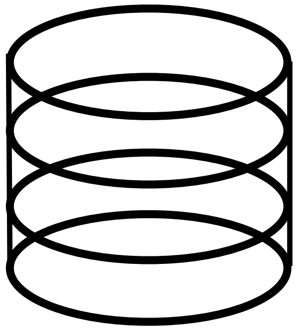
Example: RLHF in LLMs



Example: RLHF in LLMs



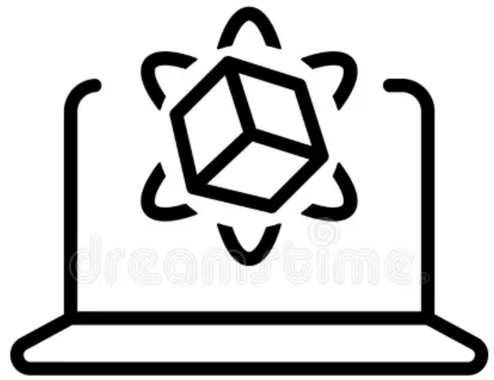
Example: RLHF in LLMs



data

{prompt, response, reward}

fit



simulator

reward model

RL/ planning/ control/

...

PPO-trained policy

?

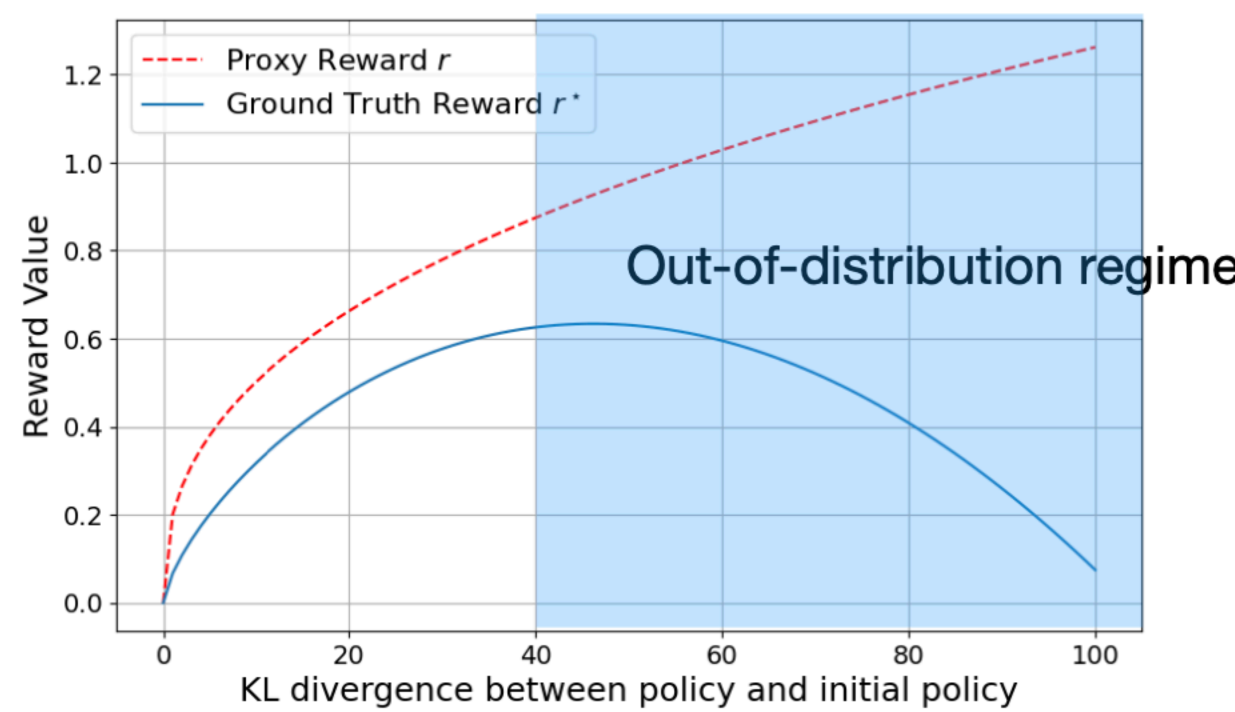
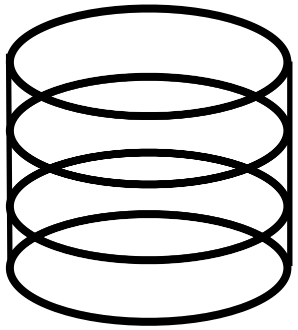


Figure from Gao et al. 2023

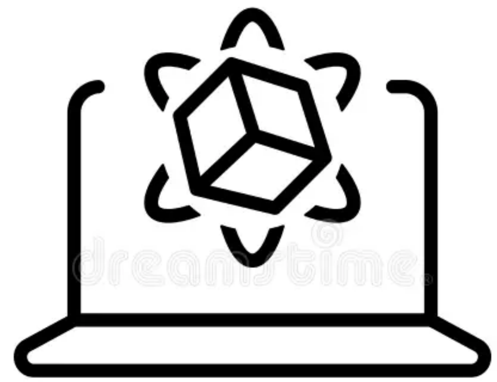
Example: RLHF in LLMs



data

{prompt, response, reward}

fit



simulator

reward model

?

RL/ planning/ control/ …

PPO-trained policy

- Generalization only happens to policies "*covered by*" behavior policy
- How to define *coverage*, and what's its interplay with algorithms?

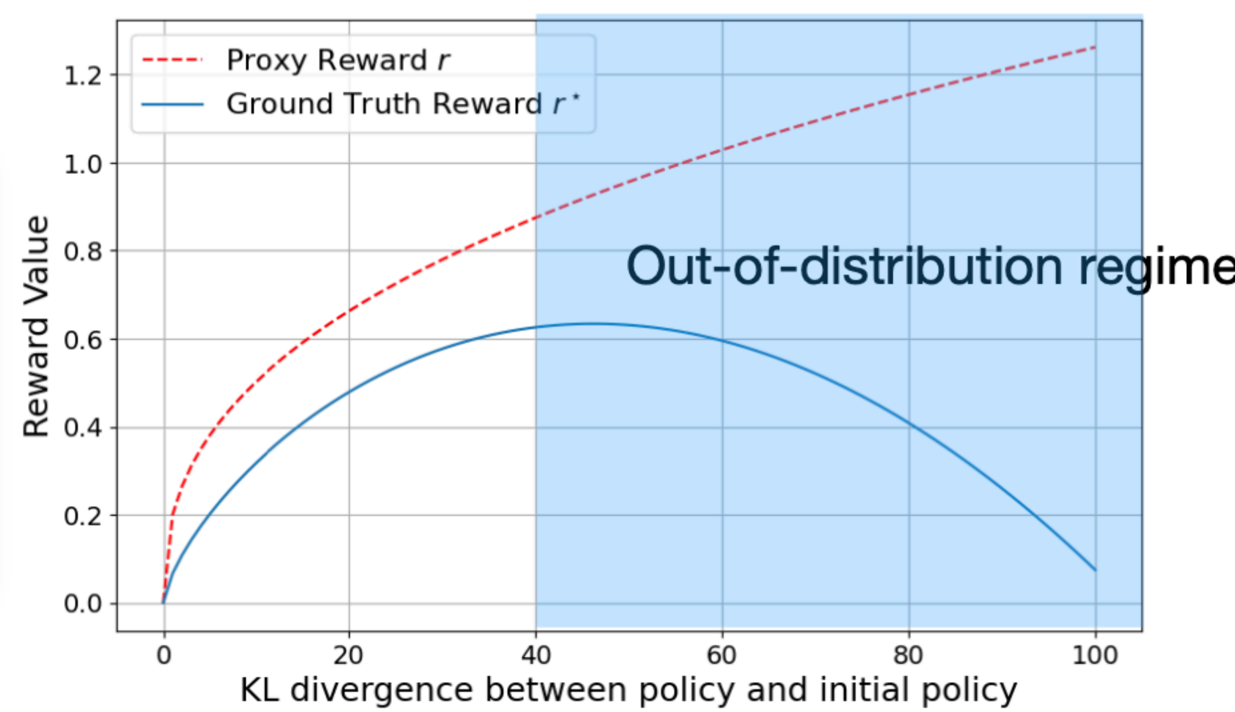


Figure from Gao et al. 2023

Framework for decision-making

- Episodic RL:

$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H$$

Framework for decision-making

- Episodic RL:

$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H$$

- RLHF
    - o_1 is prompt, each a_h is a token

Framework for decision-making

- Episodic RL:

$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H$$

- RLHF
  - o_1 is prompt, each a_h is a token
  - $r_h = 0$ except for $h = H$
  - o_h is constant symbol (nothing happens between tokens)

Framework for decision-making

- Episodic RL:

$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H$$

- RLHF
  - o_1 is prompt, each a_h is a token
  - $r_h = 0$ except for $h{=}H$
  - o_h is constant symbol (nothing happens between tokens)
  - alternatively, bandit ($H{=}1$) with combinatorial action space

Framework for decision-making

- Episodic RL:

$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H$$

- RLHF
    - o_1 is prompt, each a_h is a token
    - $r_h = 0$ except for h=H
    - o_h is constant symbol (nothing happens between tokens)
    - alternatively, bandit (H=1) with combinatorial action space
- A policy π maps history

$$\tau_h = (o_1, a_1, r_1, \ldots, o_h)$$

to action distribution: $a_h \sim \pi(\cdot \mid \tau_h)$

Framework for decision-making

- Episodic RL:

$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H$$

- RLHF
  - o_1 is prompt, each a_h is a token
  - $r_h = 0$ except for $h{=}H$
  - o_h is constant symbol (nothing happens between tokens)
  - alternatively, bandit ($H{=}1$) with combinatorial action space
- A policy π maps history

$$\tau_h = (o_1, a_1, r_1, \ldots, o_h)$$

to action distribution: $a_h \sim \pi(\cdot \mid \tau_h)$

- Performance measured by $J(\pi) := \mathbb{E}_\pi[\sum_{h=1}^H r_h]$

Framework for decision-making

- Episodic RL:

$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H$$

- RLHF
  - o_1 is prompt, each a_h is a token
  - $r_h = 0$ except for $h{=}H$
  - o_h is constant symbol (nothing happens between tokens)
  - alternatively, bandit ($H{=}1$) with combinatorial action space
- A policy π maps history

$$\tau_h = (o_1, a_1, r_1, \ldots, o_h)$$

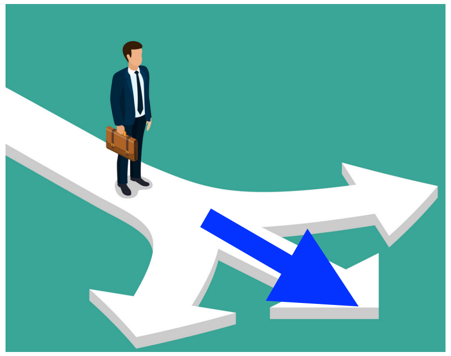
  to action distribution: $a_h \sim \pi(\cdot \mid \tau_h)$

- Performance measured by $J(\pi) := \mathbb{E}_\pi[\sum_{h=1}^{H} r_h]$
- OPE: estimate $J(\pi)$ using data episodes collected with π_b

Unbiased OPE

Importance sampling (IS) [Precup'00]

Behavior



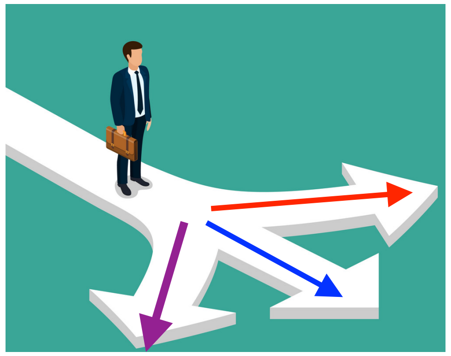
Target

Precup. 2000. Eligibility traces for off-policy policy evaluation.

Unbiased OPE

Importance sampling (IS) [Precup'00]

Behavior



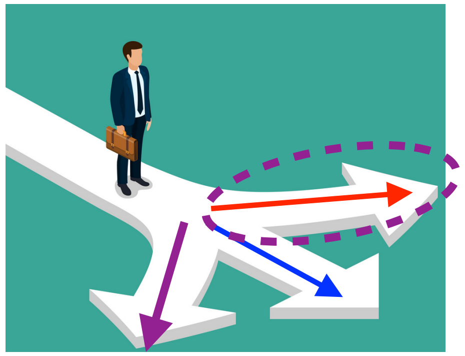
Target

Precup. 2000. Eligibility traces for off-policy policy evaluation.

Unbiased OPE

Importance sampling (IS) [Precup'00]

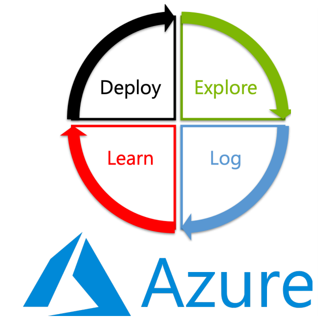
Behavior



Target

Precup. 2000. Eligibility traces for off-policy policy evaluation.

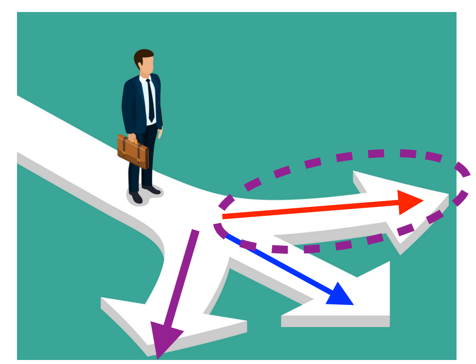
Unbiased OPE



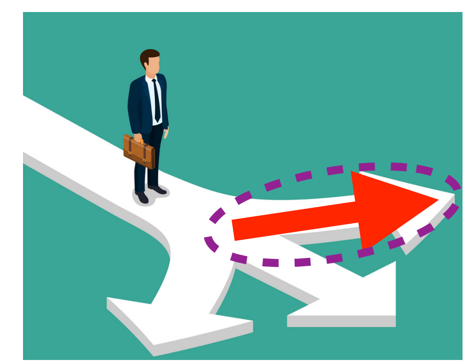
Importance sampling (IS) [Precup'00]

- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓

Behavior

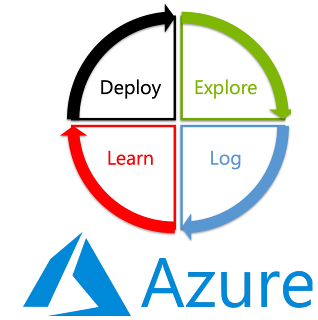


Target



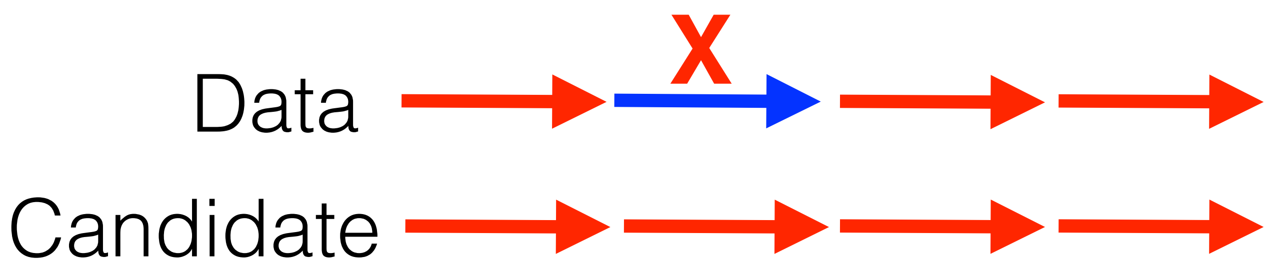
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Unbiased OPE

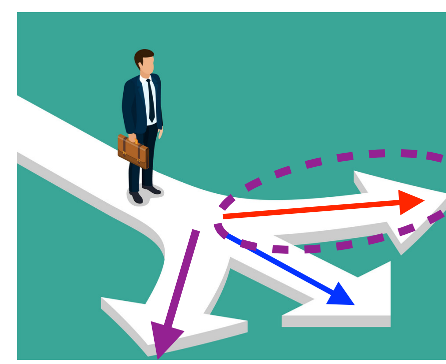


Importance sampling (IS) [Precup'00]

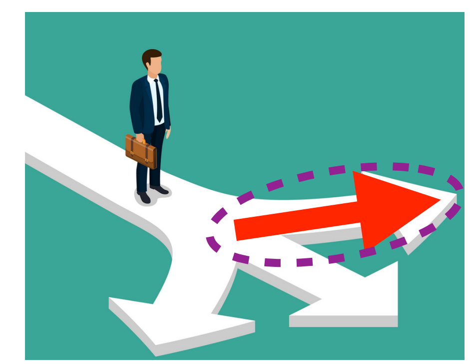
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!



Behavior

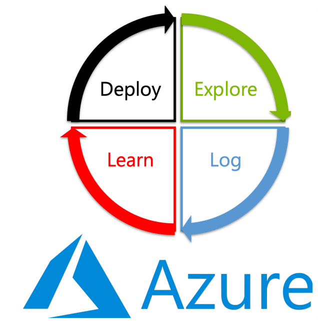


Target



Precup. 2000. Eligibility traces for off-policy policy evaluation.

Unbiased OPE

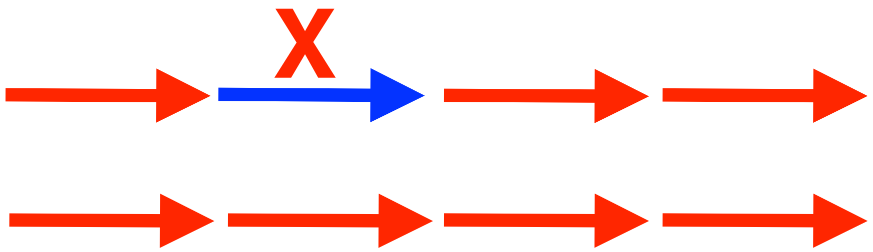


Importance sampling (IS) [Precup'00]

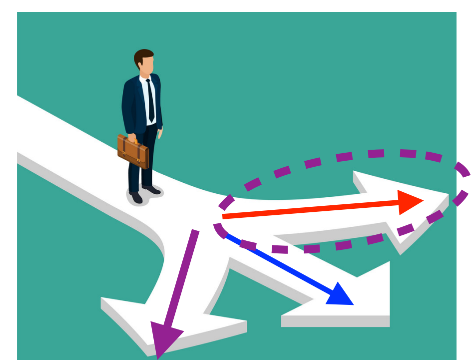
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!

Data

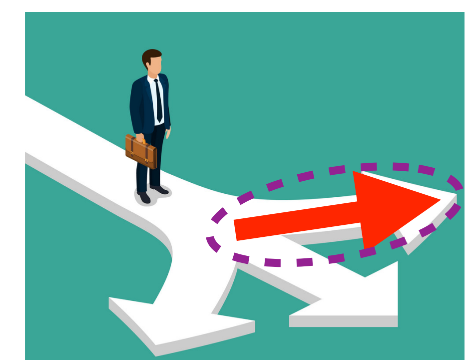
Candidate



Behavior



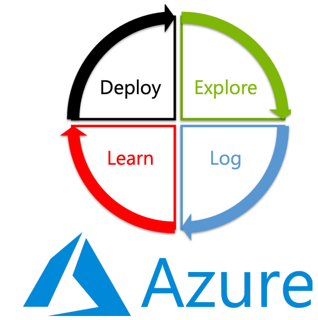
Target



$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H \quad \Rightarrow \quad \left(\prod_{h=1}^{H} \frac{\pi(a_h|\tau_h)}{\pi_b(a_h|\tau_h)} \right) \left(\sum_{h=1}^{H} r_h \right)$$

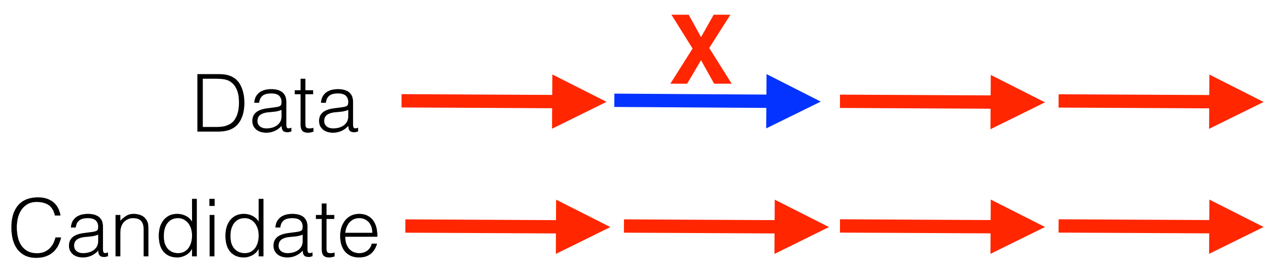
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Unbiased OPE

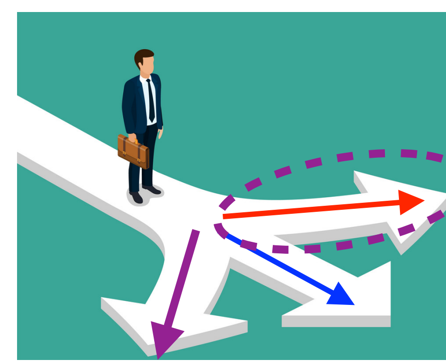


Importance sampling (IS) [Precup'00]

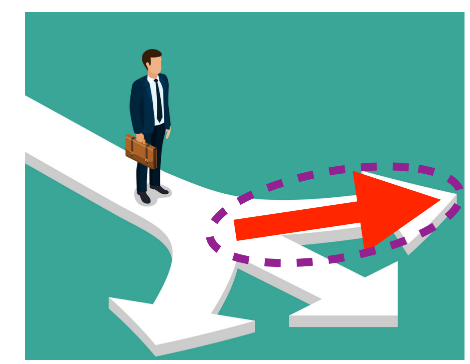
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!



Behavior



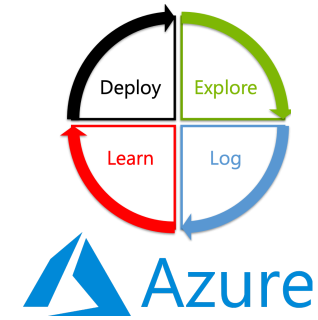
Target



$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H \quad \Rightarrow \quad \left(\prod_{h=1}^{H} \frac{\pi(a_h|\tau_h)}{\pi_b(a_h|\tau_h)}\right)\left(\sum_{h=1}^{H} r_h\right)$$

Precup. 2000. Eligibility traces for off-policy policy evaluation.

Unbiased OPE

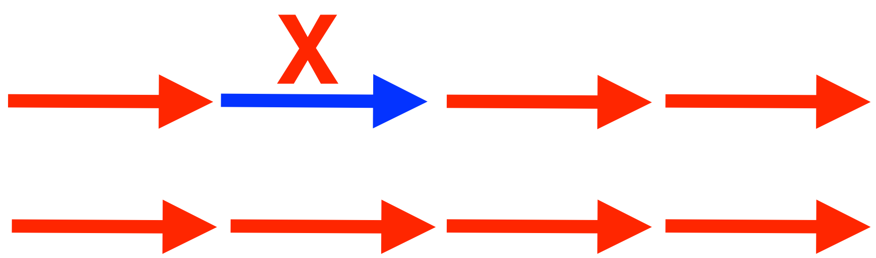


Importance sampling (IS) [Precup'00]

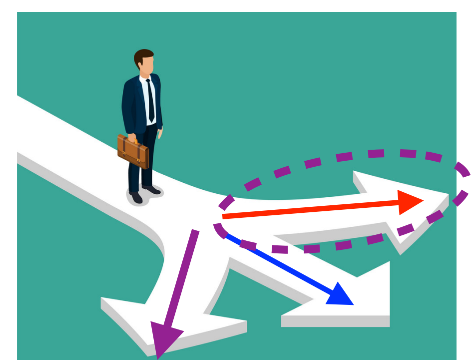
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!

Data

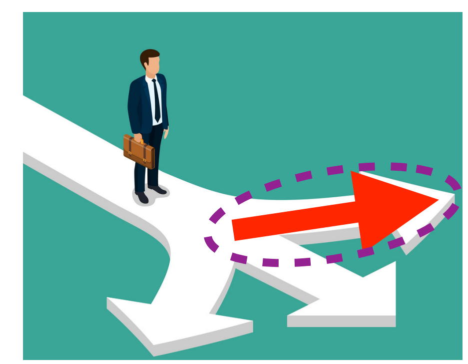
Candidate



Behavior



Target

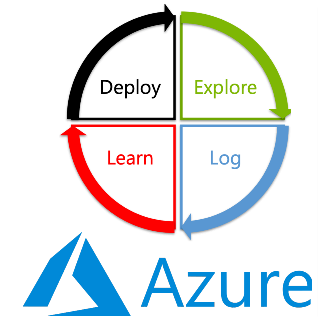


$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H \quad \Rightarrow \quad \left(\prod_{h=1}^{H} \frac{\pi(a_h|\tau_h)}{\pi_b(a_h|\tau_h)} \right) \left(\sum_{h=1}^{H} r_h \right)$$

- Or, can only evaluate π when $\prod_{h=1}^{H} \frac{\pi(a_h|\tau_h)}{\pi_b(a_h|\tau_h)}$ small

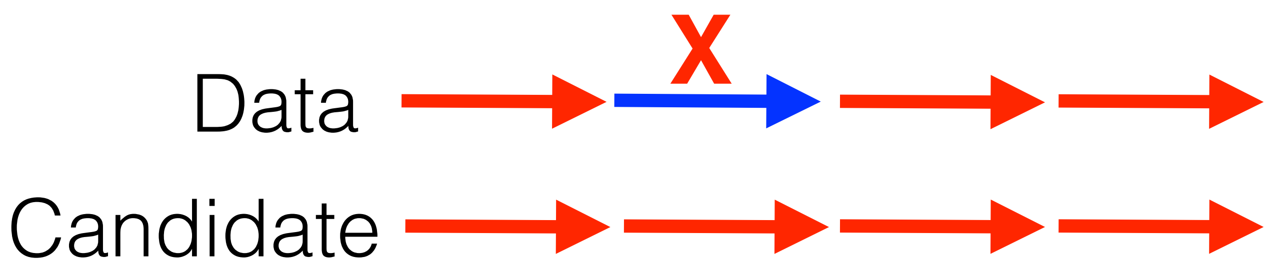
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Unbiased OPE

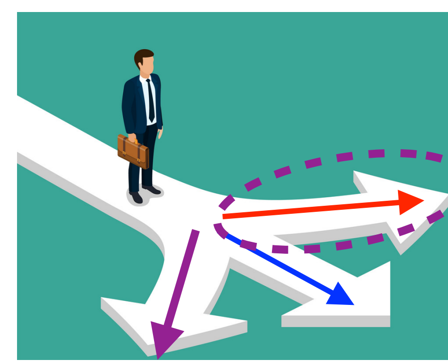


Importance sampling (IS) [Precup'00]

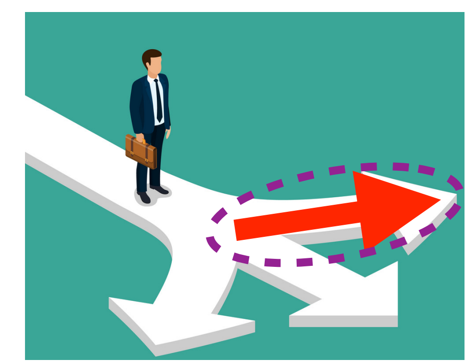
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!



Behavior



Target



$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H \Rightarrow \left(\prod_{h=1}^{H} \frac{\pi(a_h|\tau_h)}{\pi_b(a_h|\tau_h)} \right) \left(\sum_{h=1}^{H} r_h \right)$$

- Or, can only evaluate π when $\prod_{h=1}^{H} \frac{\pi(a_h|\tau_h)}{\pi_b(a_h|\tau_h)}$ small

IS' measure of *coverage*

Precup. 2000. Eligibility traces for off-policy policy evaluation.

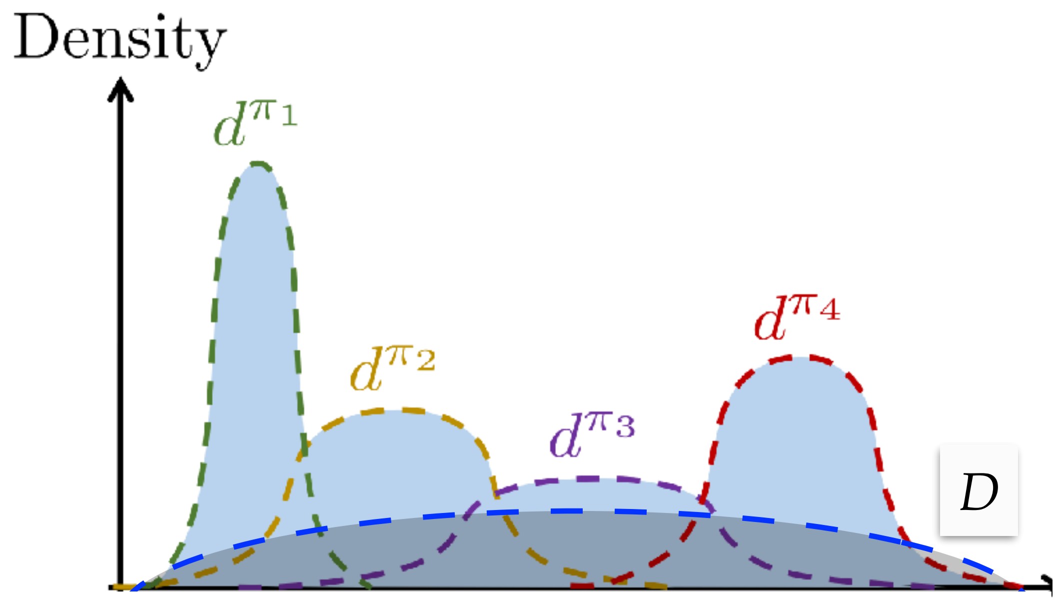
Why OPE & Coverage?

Evaluation is the basis of optimization

Why OPE & Coverage?

Evaluation is the basis of optimization

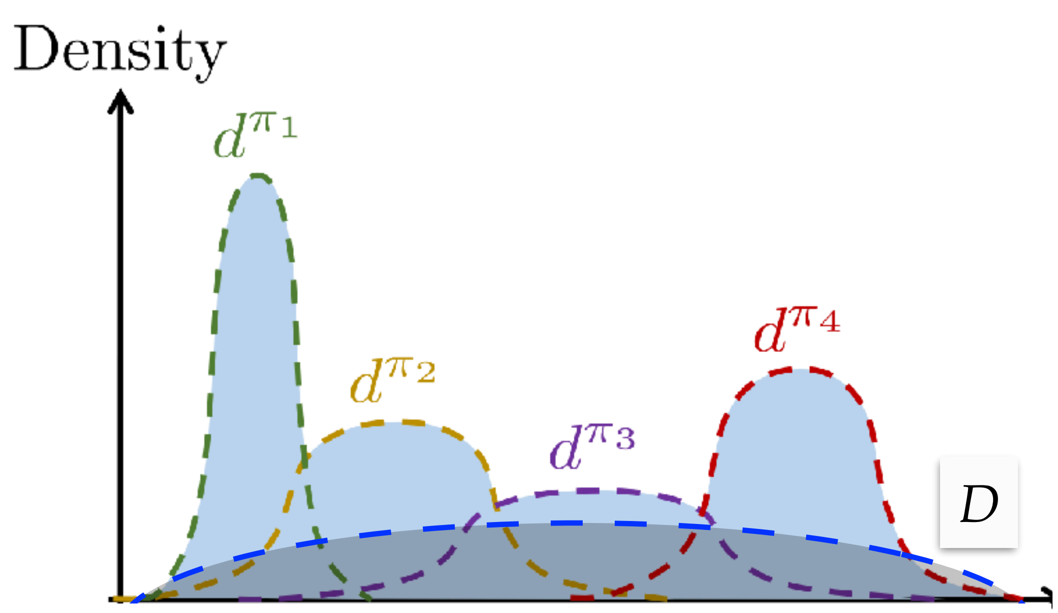
- If all policies are covered & accurately evaluated, we can pick the approximate best

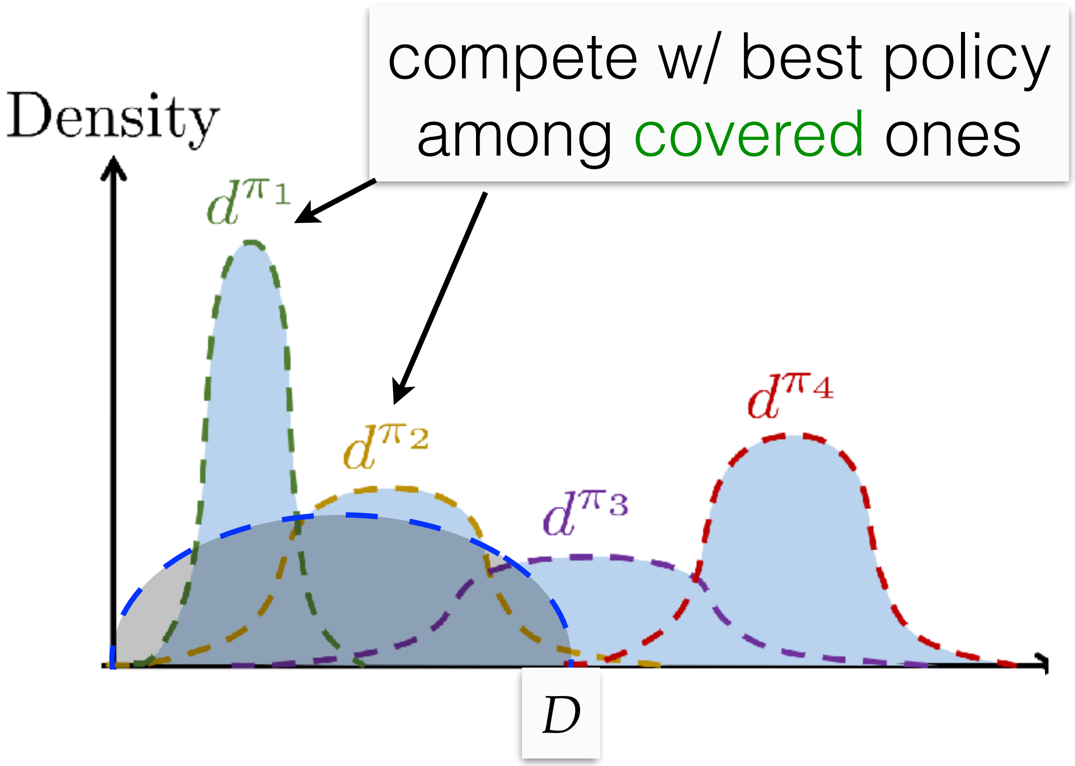


Why OPE & Coverage?

Evaluation is the basis of optimization

- If all policies are covered & accurately evaluated, we can pick the approximate best
- Doesn't have to! Can constrain learned policy to be covered => compete with best covered policy

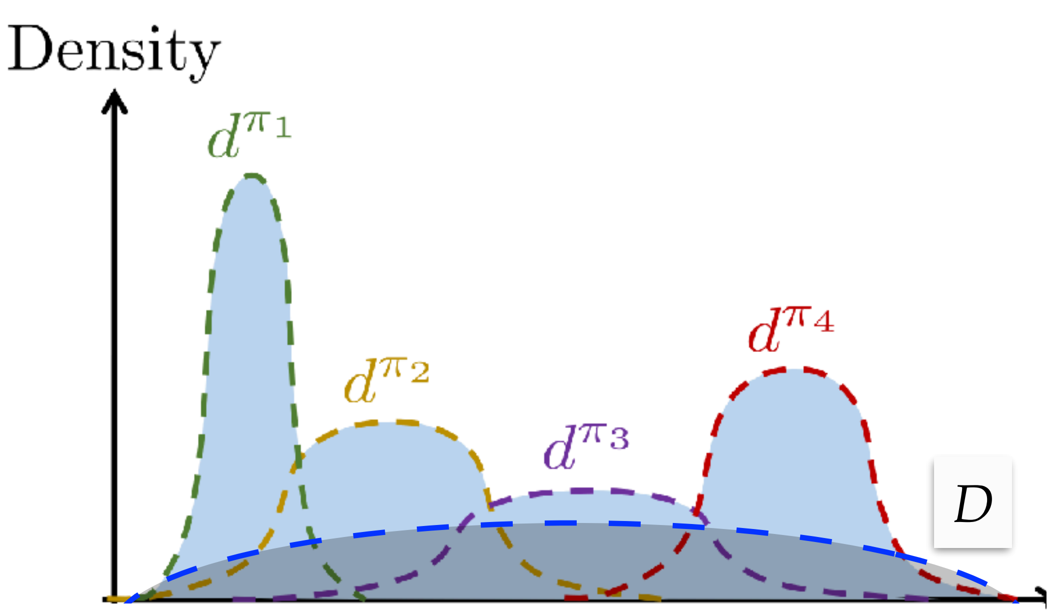


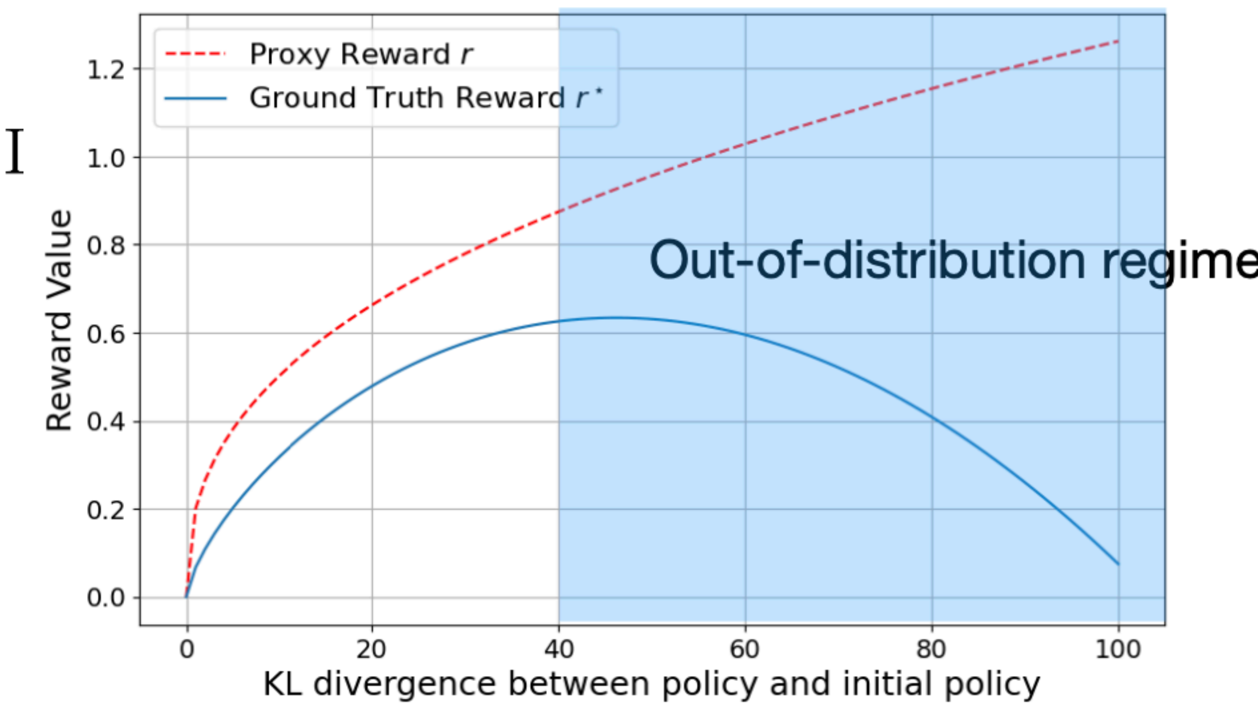


Why OPE & Coverage?

Evaluation is the basis of optimization

- If all policies are covered & accurately evaluated, we can pick the approximate best
- Doesn't have to! Can constrain learned policy to be covered => compete with best covered policy
  - behavior regularization, pessimism in face of uncertainty, etc.



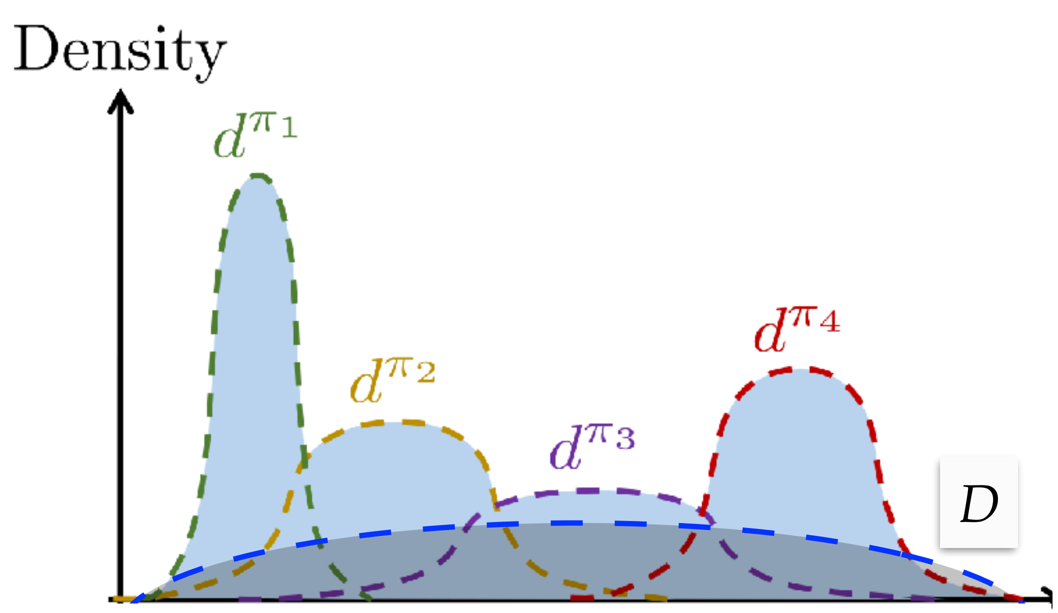


D Figure from Gao et al. 2023

Why OPE & Coverage?

Evaluation is the basis of optimization

- If all policies are covered & accurately evaluated, we can pick the approximate best
- Doesn't have to! Can constrain learned policy to be covered => compete with best covered policy
  - behavior regularization, pessimism in face of uncertainty, etc.
- More lenient coverage => stronger competing target



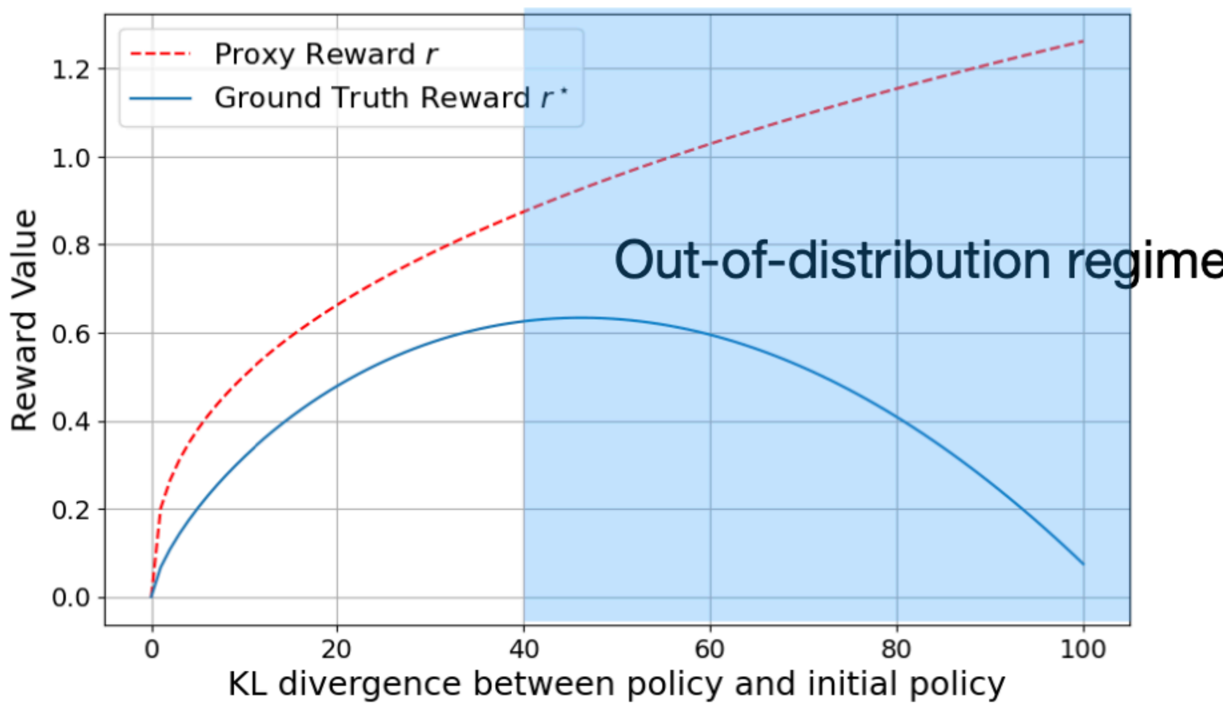
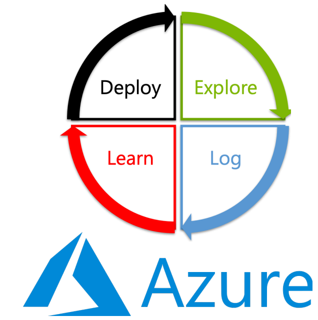


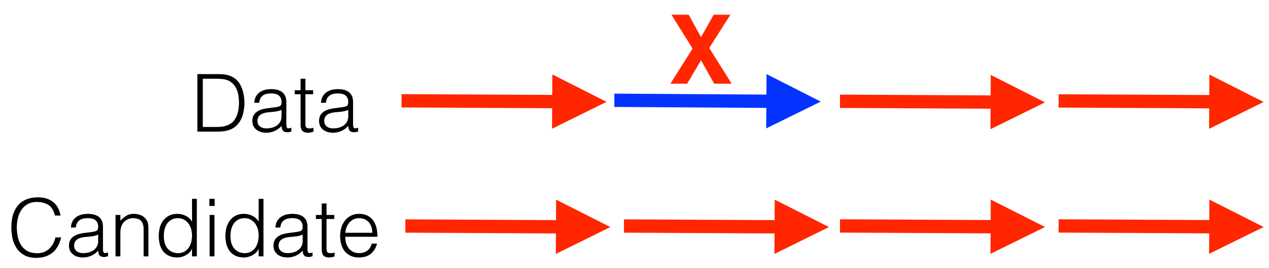
Figure from Gao et al. 2023

Unbiased OPE

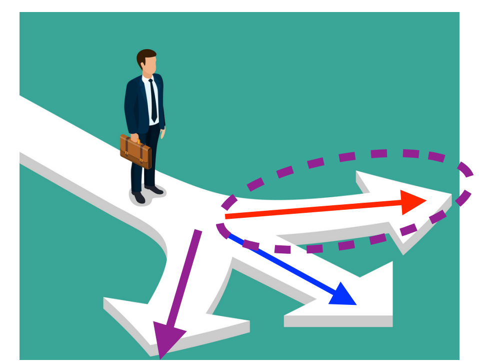


Importance sampling (IS) [Precup'00]

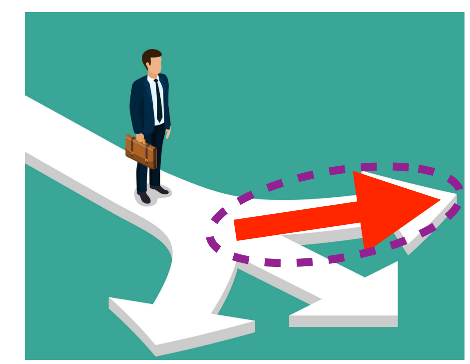
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!



Behavior



Target



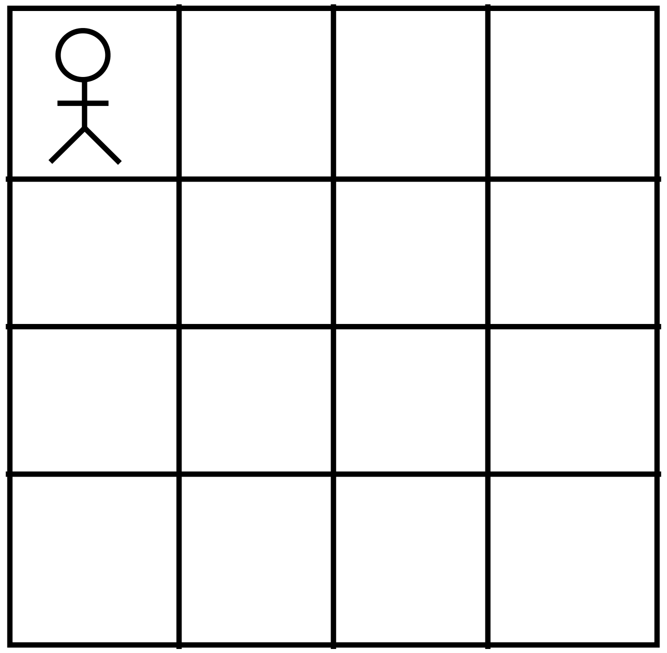
$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H \Rightarrow \left(\prod_{h=1}^{H} \frac{\pi(a_h|\tau_h)}{\pi_b(a_h|\tau_h)}\right)\left(\sum_{h=1}^{H} r_h\right)$$

- Or, can only evaluate π when $\prod_{h=1}^{H} \frac{\pi(a_h|\tau_h)}{\pi_b(a_h|\tau_h)}$ small

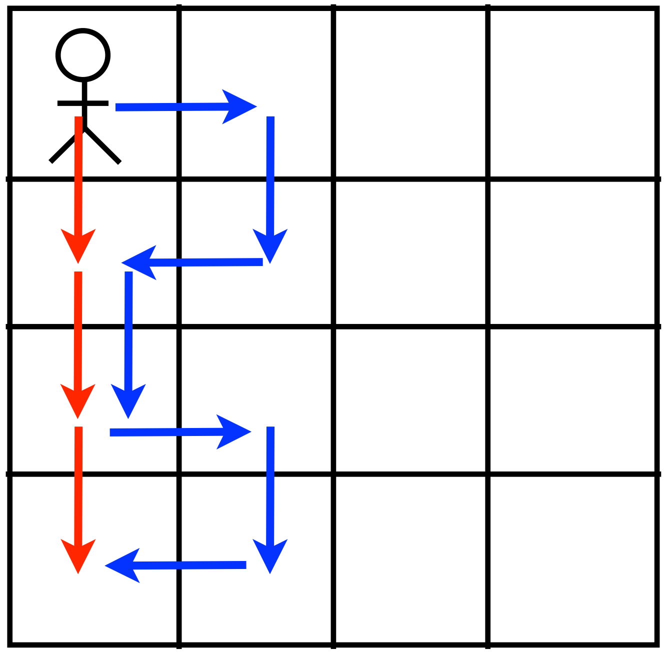
IS' measure of *coverage*

Precup. 2000. Eligibility traces for off-policy policy evaluation.

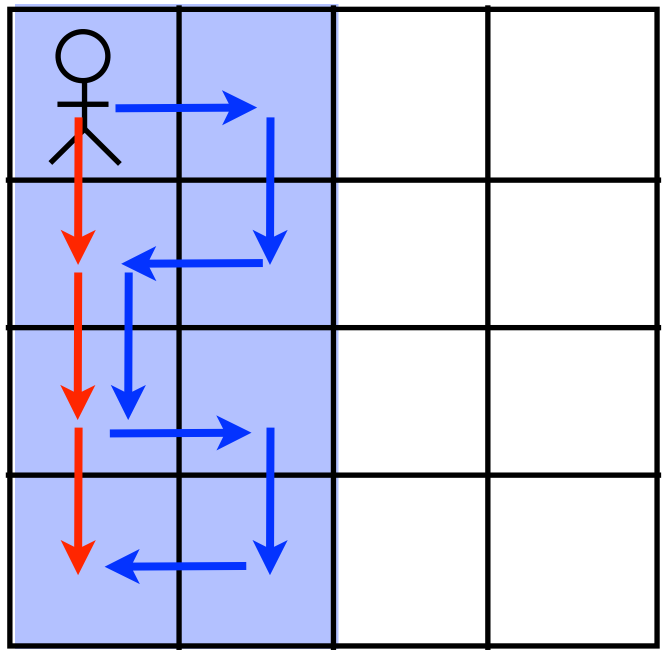
Better coverage?



Better coverage?



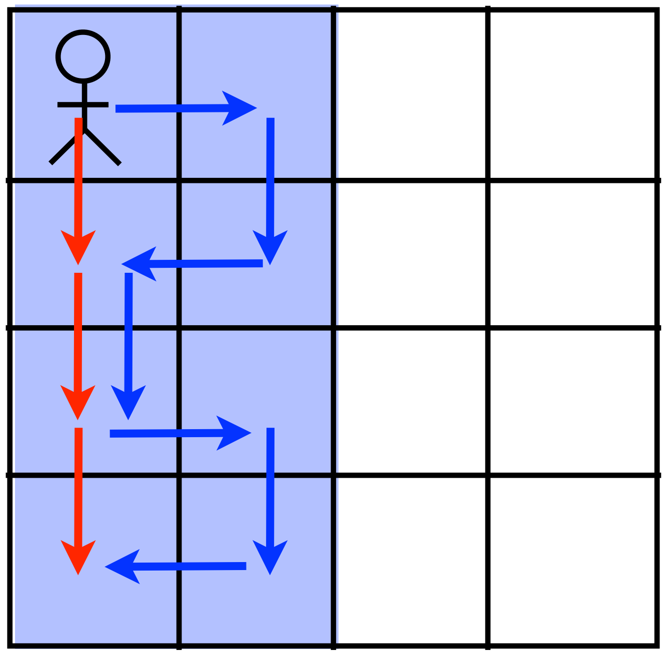
Better coverage?



Better coverage?

FQE [Munos, Szepesvari… CJ'19, …] / **MIS** [Liu et al'18, Nachum et al'19, UHJ'20, …]

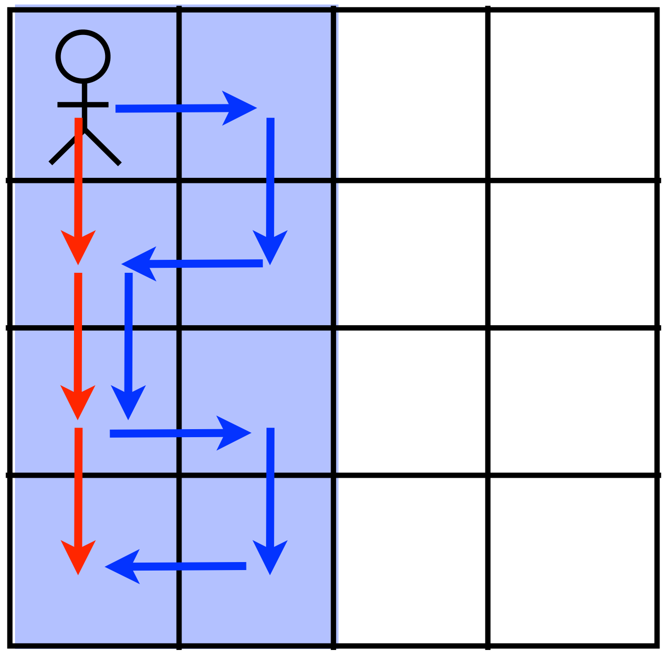
- Assume MDPs: $o_1(=s_1), a_1, r_1, \ldots, o_H(=s_H), a_H, r_H$



Better coverage?

FQE [Munos, Szepesvari… CJ'19, …] / **MIS** [Liu et al'18, Nachum et al'19, UHJ'20, …]

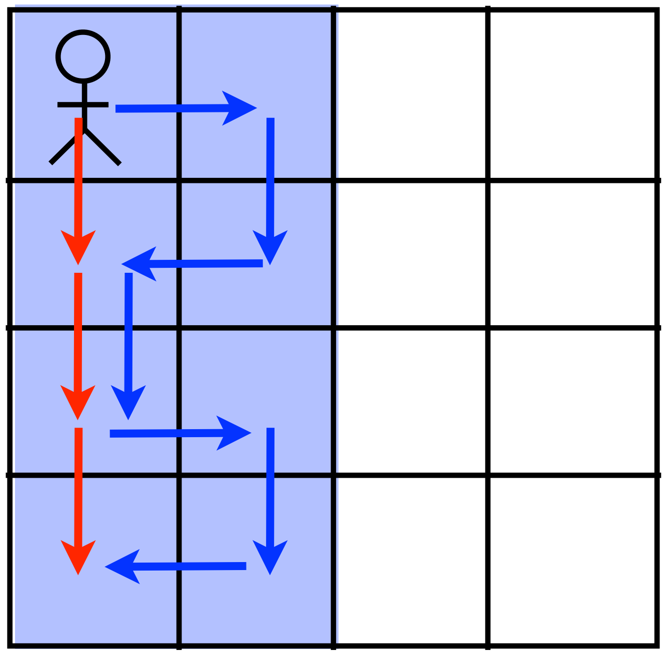
- Assume MDPs: $o_1 (= s_1), a_1, r_1, \ldots, o_H (= s_H), a_H, r_H$
- Learn value functions: $V^\pi(s_h) := \mathbb{E}_\pi[\sum_{h'=h}^{H} r_{h'} | s_h]$
  - Satisfies $V^\pi(s_h) = (\mathcal{T}^\pi V^\pi)(s_h) := R(s_h, \pi) + \mathbb{E}_{s_{h+1} \sim P(s_h, \pi)}[V^\pi(s_{h+1})]$.



Better coverage?

FQE [Munos, Szepesvari… CJ'19, …] / MIS [Liu et al'18, Nachum et al'19, UHJ'20, …]

- Assume MDPs: $o_1 (= s_1), a_1, r_1, \ldots, o_H (= s_H), a_H, r_H$
- Learn value functions: $V^\pi(s_h) := \mathbb{E}_\pi[\sum_{h'=h}^{H} r_{h'} | s_h]$
  - Satisfies $V^\pi(s_h) = (\mathcal{T}^\pi V^\pi)(s_h) := R(s_h, \pi) + \mathbb{E}_{s_{h+1} \sim P(s_h, \pi)}[V^\pi(s_{h+1})]$.
  - Estimate from class $\mathcal{V}$ by $\arg\min_{V \in \mathcal{V}} \mathbb{E}_{\pi_b}[(V - \mathcal{T}^\pi V)^2]$

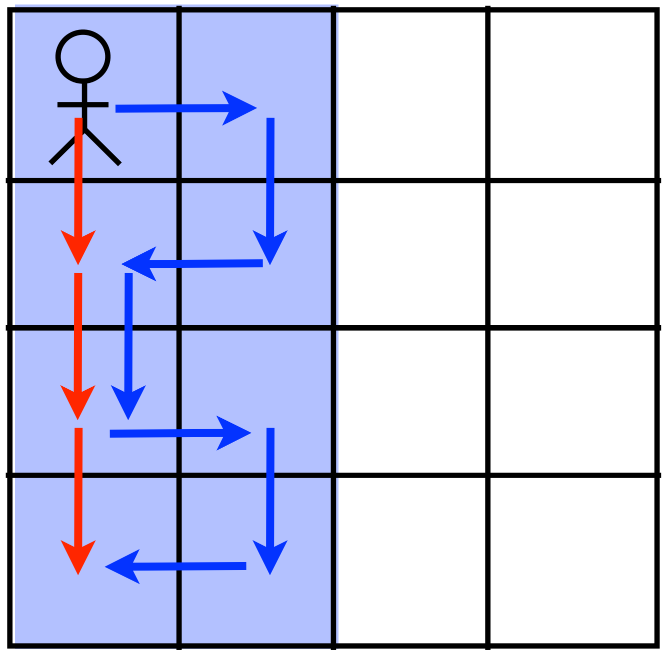


* Also needs Bellman completeness

Better coverage?

FQE [Munos, Szepesvari… CJ'19, …] / MIS [Liu et al'18, Nachum et al'19, UHJ'20, …]

- Assume MDPs: $o_1(=s_1), a_1, r_1, \ldots, o_H(=s_H), a_H, r_H$
- Learn value functions: $V^\pi(s_h) := \mathbb{E}_\pi[\sum_{h'=h}^{H} r_{h'} | s_h]$
  - Satisfies $V^\pi(s_h) = (\mathcal{T}^\pi V^\pi)(s_h) := R(s_h, \pi) + \mathbb{E}_{s_{h+1} \sim P(s_h, \pi)}[V^\pi(s_{h+1})]$.
  - Estimate from class $\mathcal{V}$ by $\arg\min_{V \in \mathcal{V}} \mathbb{E}_{\pi_b}[(V - \mathcal{T}^\pi V)^2]$
- Coverage: *marginal* state distribution
  - require $\frac{d^\pi(s_h)}{d^{\pi_b}(s_h)}$ and $\frac{\pi(a_h|o_h)}{\pi_b(a_h|o_h)}$ small



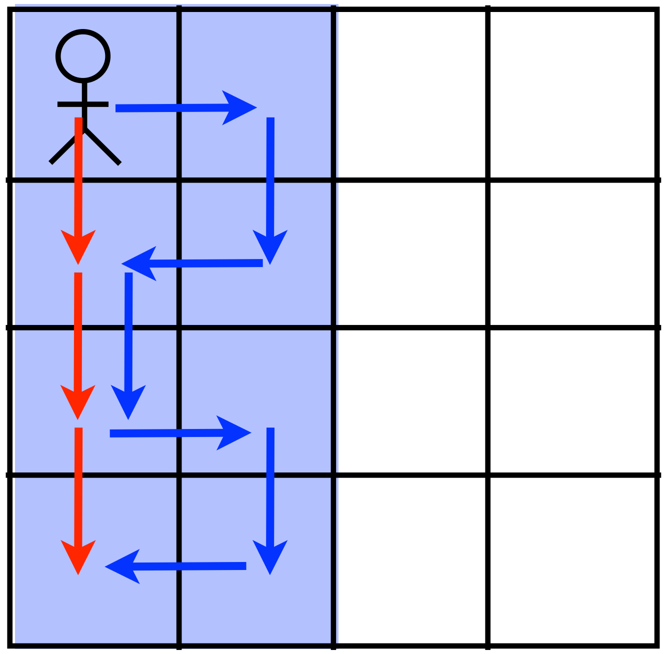
* Also needs Bellman completeness

Better coverage?

FQE [Munos, Szepesvari… CJ'19, …] / **MIS** [Liu et al'18, Nachum et al'19, UHJ'20, …]

- Assume MDPs: $o_1 (= s_1), a_1, r_1, \ldots, o_H (= s_H), a_H, r_H$
- Learn value functions: $V^\pi(s_h) := \mathbb{E}_\pi[\sum_{h'=h}^{H} r_{h'} | s_h]$
  - Satisfies $V^\pi(s_h) = (\mathcal{T}^\pi V^\pi)(s_h) := R(s_h, \pi) + \mathbb{E}_{s_{h+1} \sim P(s_h, \pi)}[V^\pi(s_{h+1})]$.
  - Estimate from class $\mathcal{V}$ by $\arg\min_{V \in \mathcal{V}} \mathbb{E}_{\pi_b}[(V - \mathcal{T}^\pi V)^2]$
- Coverage: *marginal* state distribution
  - require $\frac{d^\pi(s_h)}{d^{\pi_b}(s_h)}$ and $\frac{\pi(a_h|o_h)}{\pi_b(a_h|o_h)}$ small
  - Can further improve to e.g.,

$$\sup_{V \in \mathcal{V}} \frac{|\mathbb{E}_\pi[V - \mathcal{T}^\pi V]|}{\sqrt{\mathbb{E}_{\pi_b}[(V - \mathcal{T}^\pi V)^2]}}$$



* Also needs Bellman completeness

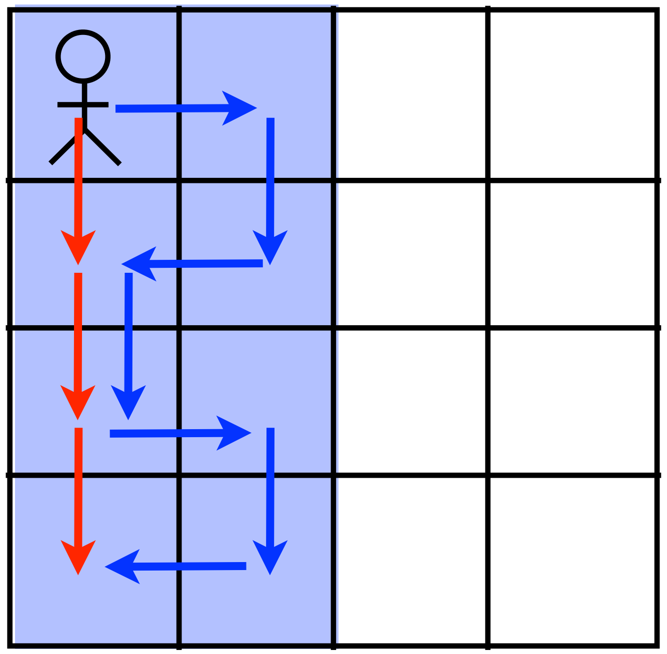
Better coverage?

FQE [Munos, Szepesvari… CJ'19, …] / MIS [Liu et al'18, Nachum et al'19, UHJ'20, …]

- Assume MDPs: $o_1 (= s_1), a_1, r_1, \ldots, o_H (= s_H), a_H, r_H$
- Learn value functions: $V^\pi(s_h) := \mathbb{E}_\pi[\sum_{h'=h}^{H} r_{h'} | s_h]$
    - Satisfies $V^\pi(s_h) = (\mathcal{T}^\pi V^\pi)(s_h) := R(s_h, \pi) + \mathbb{E}_{s_{h+1} \sim P(s_h, \pi)}[V^\pi(s_{h+1})]$.
    - Estimate from class $\mathcal{V}$ by $\arg\min_{V \in \mathcal{V}} \mathbb{E}_{\pi_b}[(V - \mathcal{T}^\pi V)^2]$
- Coverage: *marginal* state distribution
    - require $\frac{d^\pi(s_h)}{d^{\pi_b}(s_h)}$ and $\frac{\pi(a_h|o_h)}{\pi_b(a_h|o_h)}$ small
    - Can further improve to e.g.,

$$\sup_{V \in \mathcal{V}} \frac{|\mathbb{E}_\pi[V - \mathcal{T}^\pi V]|}{\sqrt{\mathbb{E}_{\pi_b}[(V - \mathcal{T}^\pi V)^2]}}$$

- Fundamental to offline training & online exploration



* Also needs Bellman completeness

How do value functions help in MDPs?

$$V(s_h) - (\mathcal{T}^{\pi} V)(s_h)$$
$$\text{“}\underbrace{\qquad\qquad}\text{”}$$

Bellman error

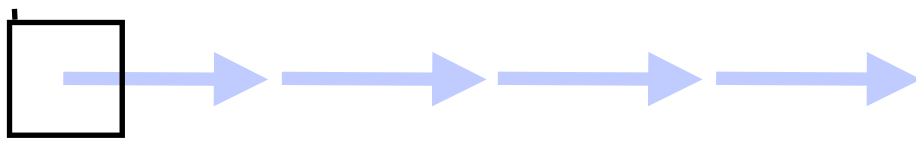
How do value functions help in MDPs?

$\mathbb{E}[V(s_1)] - J(\pi)$ **?**

Prediction | Groundtruth

$$V(s_h) - (\mathcal{T}^{\pi} V)(s_h)$$

"Bellman error"



$V(s_1)$

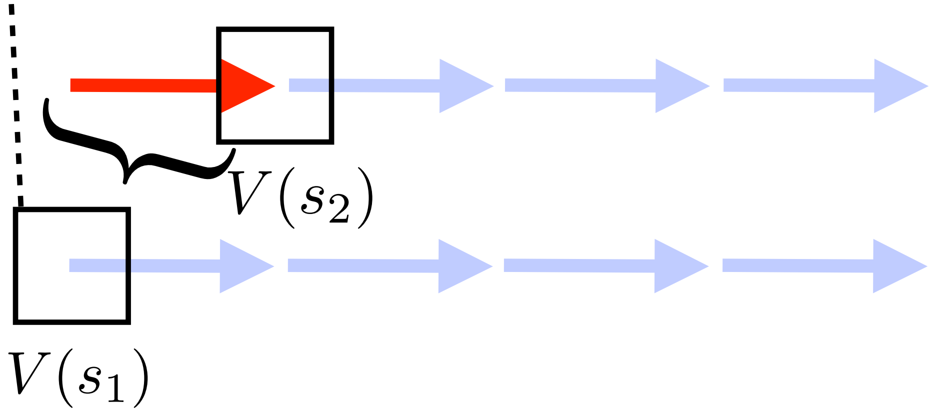
How do value functions help in MDPs?

$\mathbb{E}[V(s_1)] - J(\pi)$ **?**

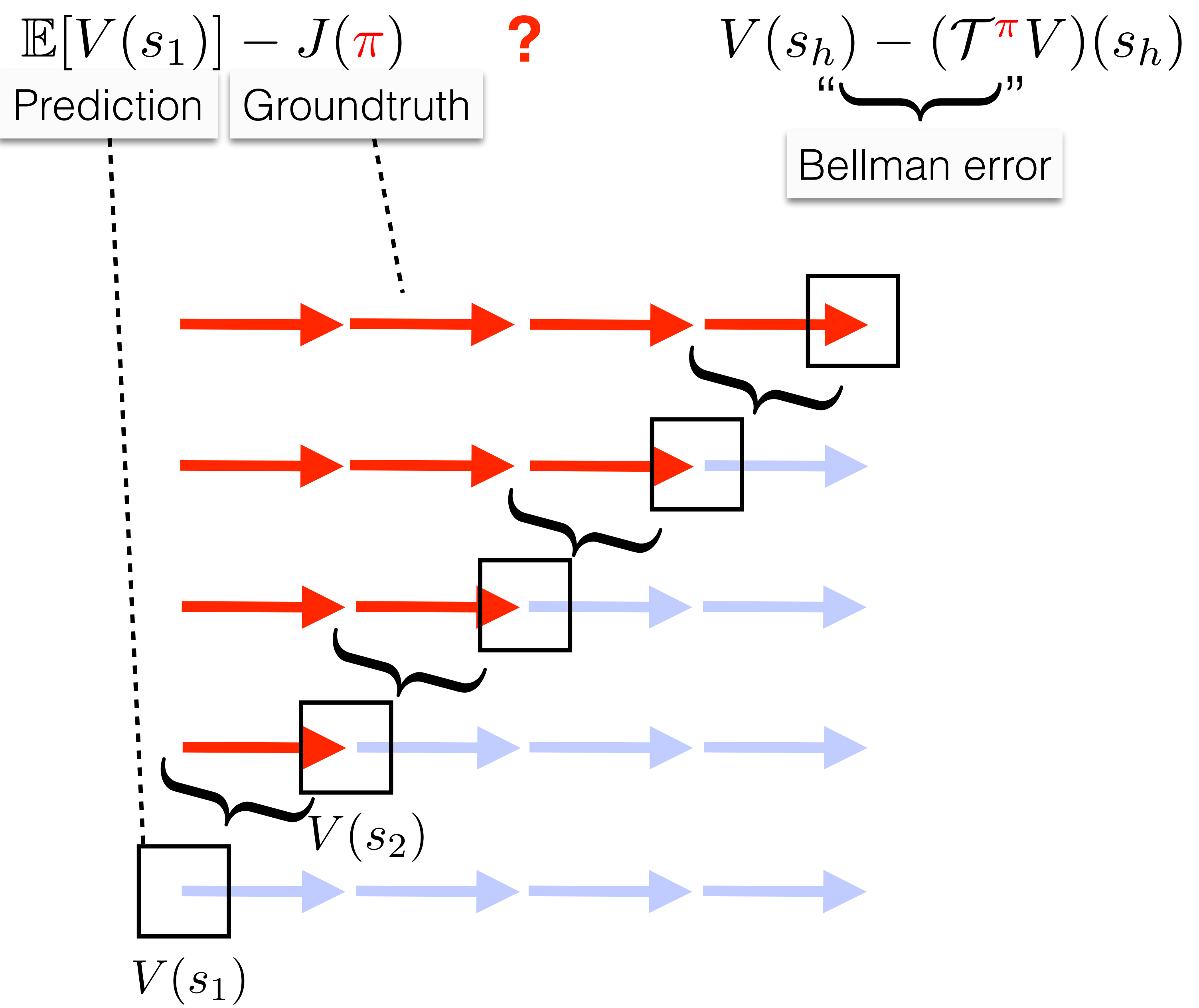
Prediction | Groundtruth

$$V(s_h) - (\mathcal{T}^{\pi} V)(s_h)$$

"Bellman error"



How do value functions help in MDPs?



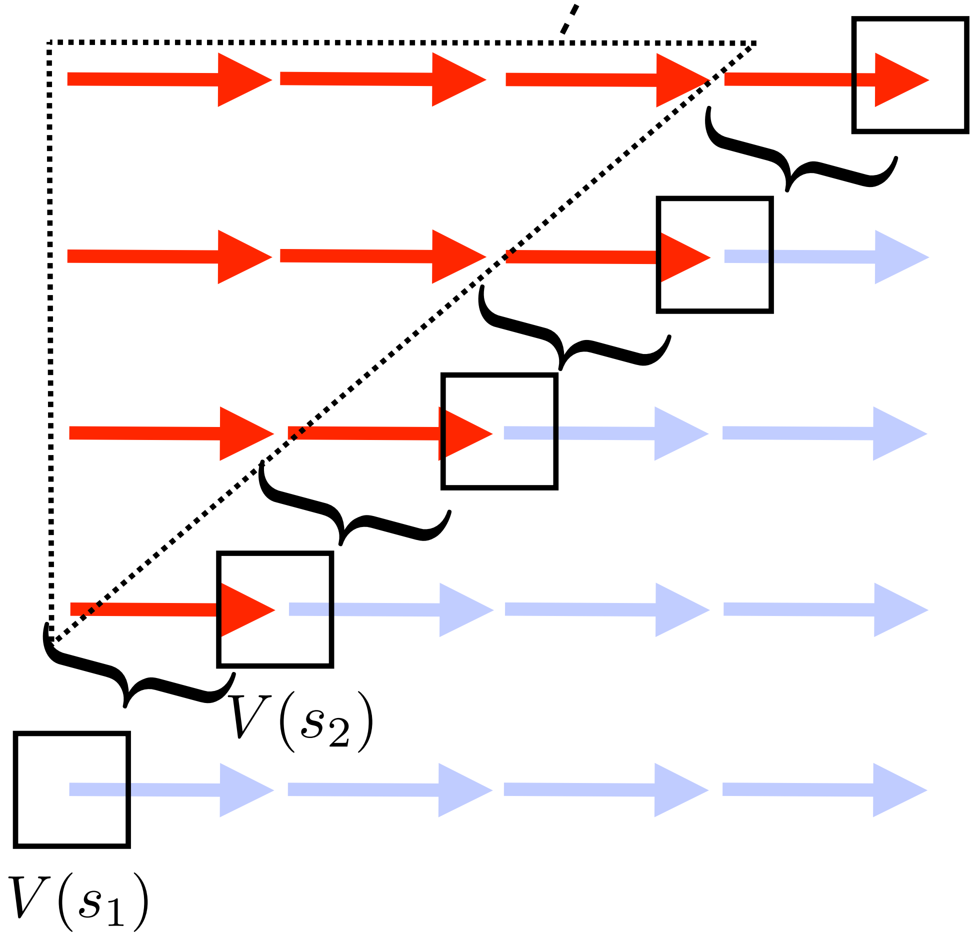
How do value functions help in MDPs?

$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Prediction

Groundtruth

"Bellman error"



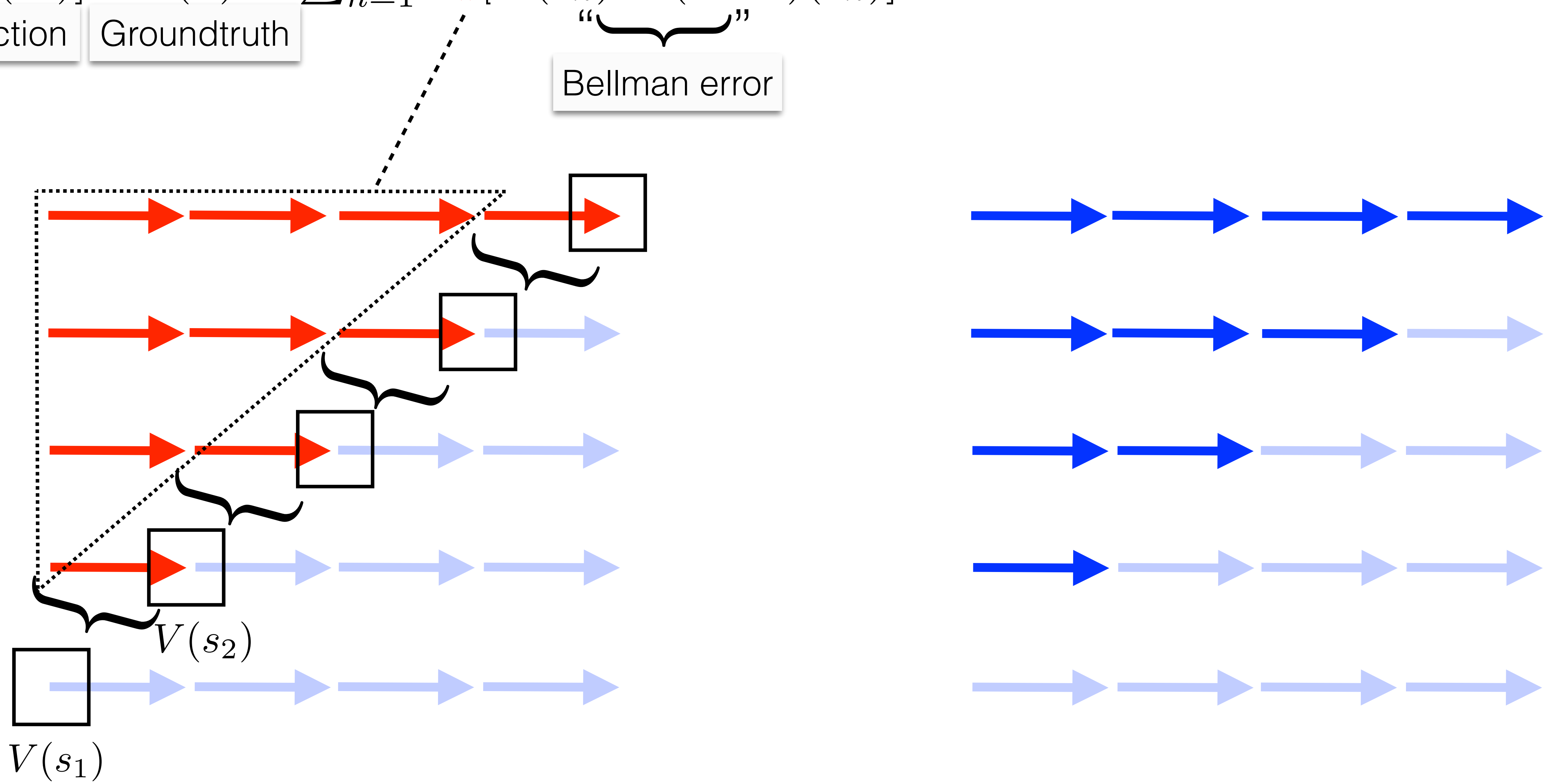
How do value functions help in MDPs?

$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Prediction

Groundtruth

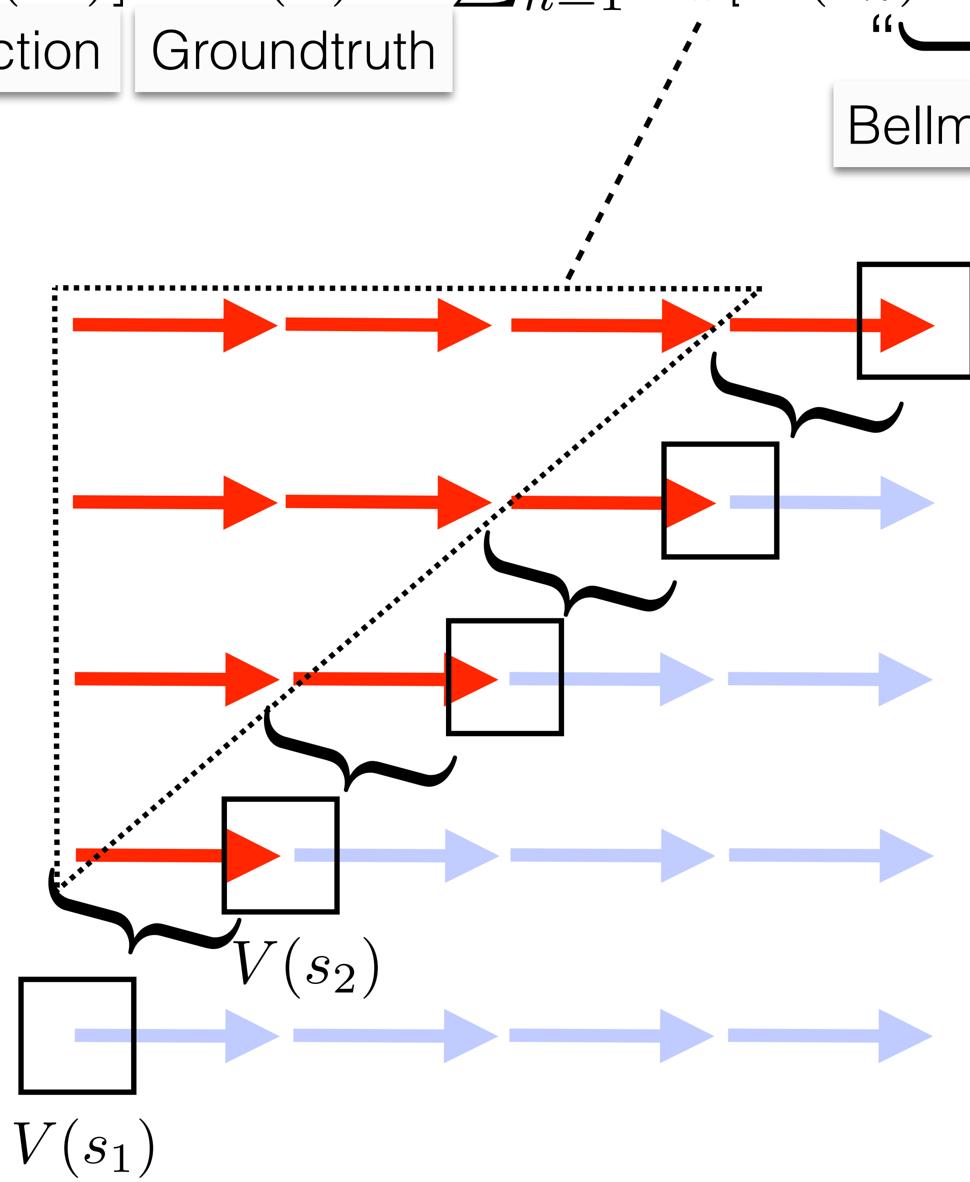
"Bellman error"

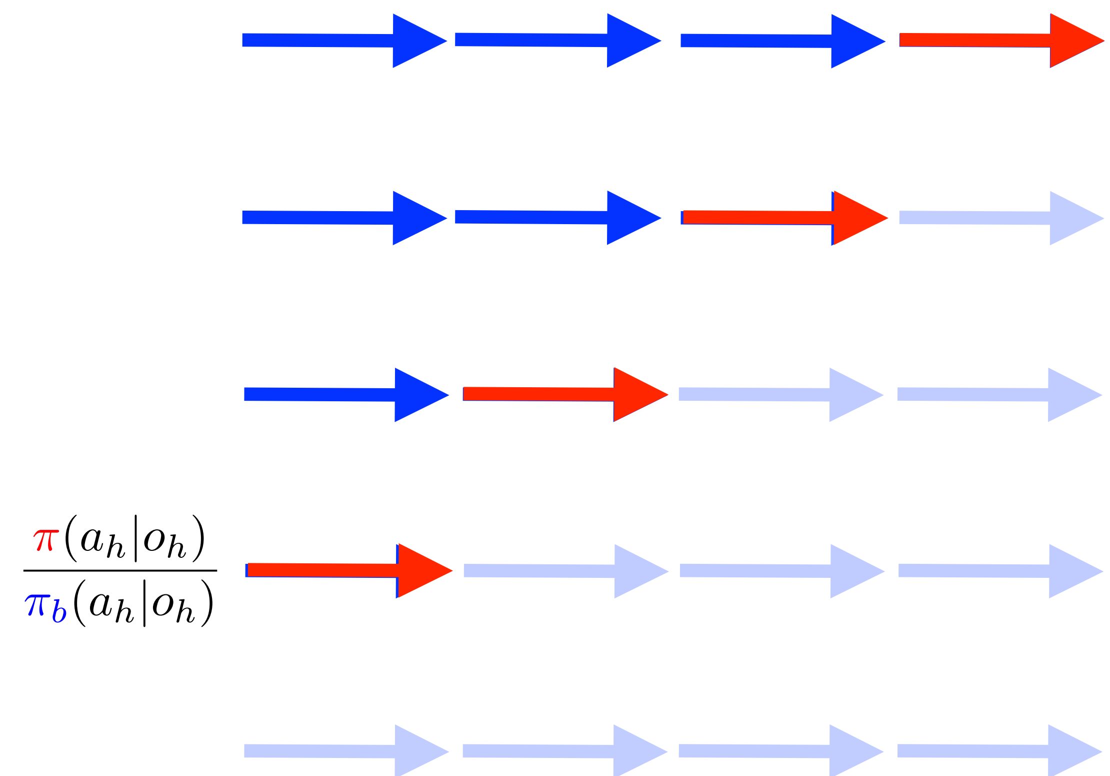


How do value functions help in MDPs?

$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Prediction | Groundtruth | "Bellman error"





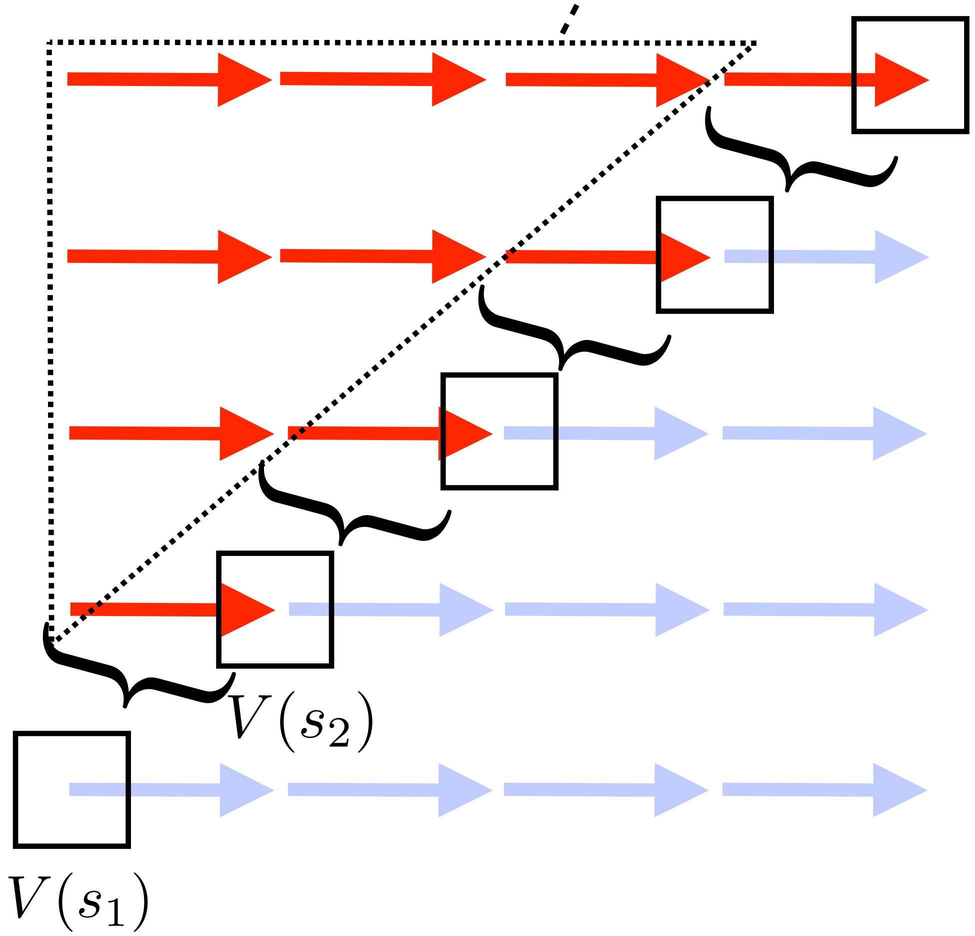
How do value functions help in MDPs?

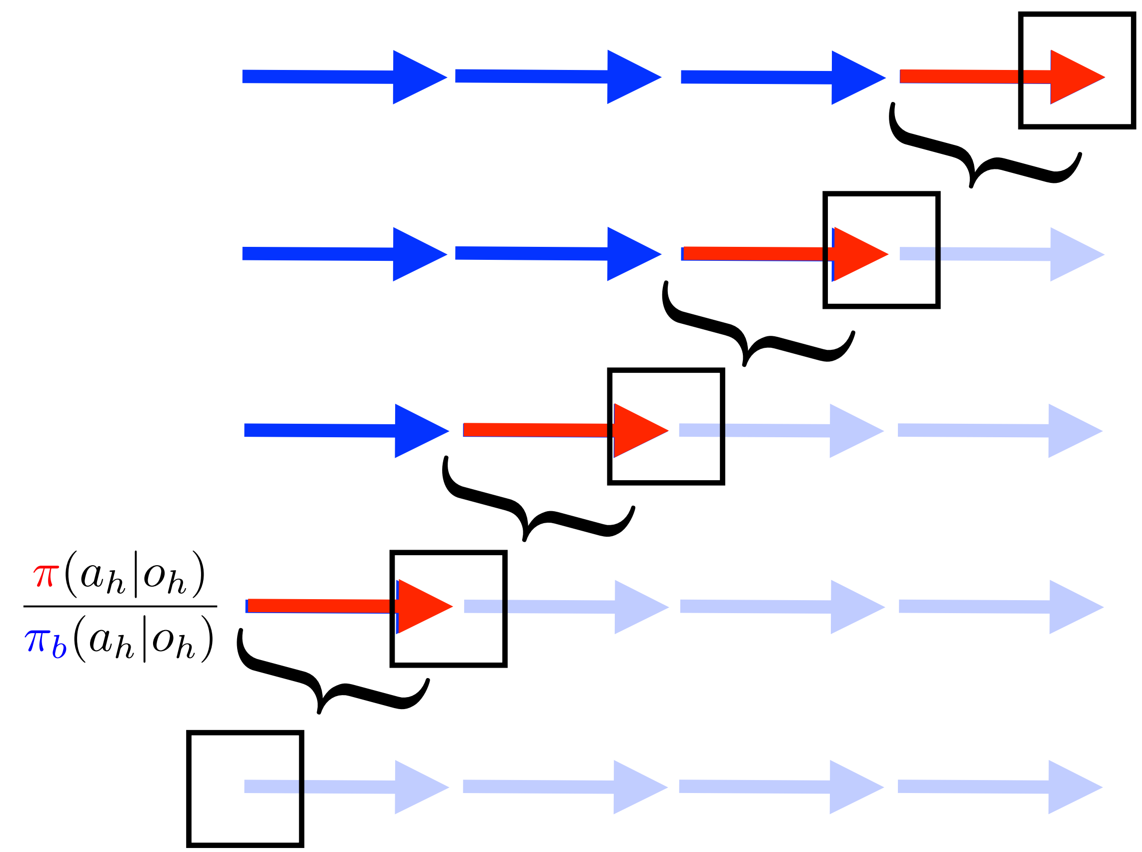
$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Prediction Groundtruth

"Bellman error"

$$\arg\min_V \sum_{h=1}^{H} \mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$



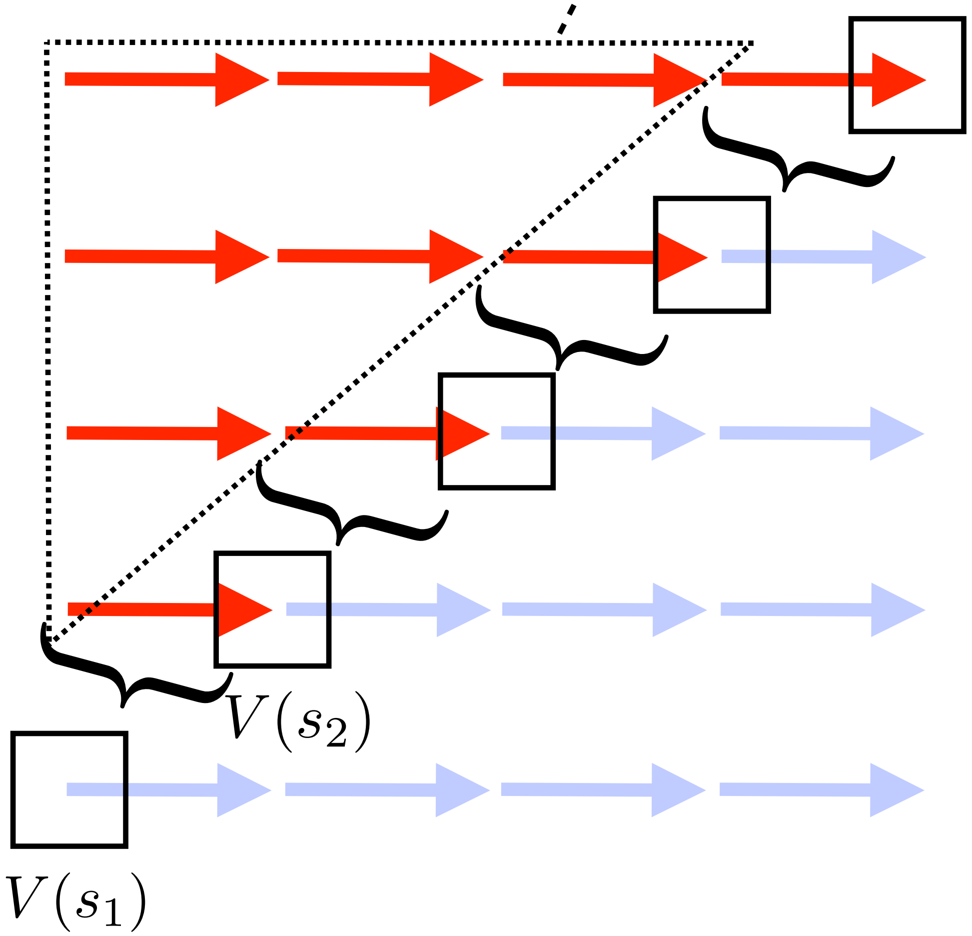


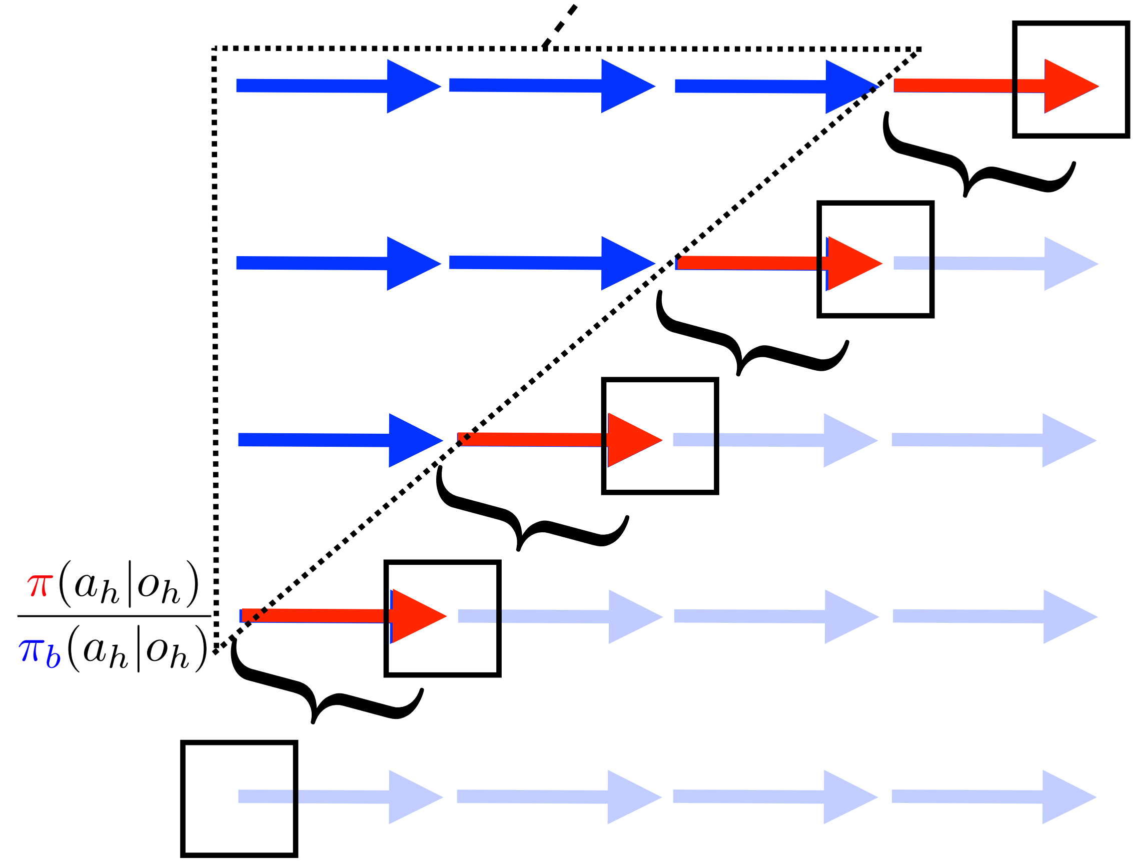
How do value functions help in MDPs?

$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Prediction Groundtruth "Bellman error"

$$\arg\min_{V} \sum_{h=1}^{H} \mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$





How do value functions help in MDPs?

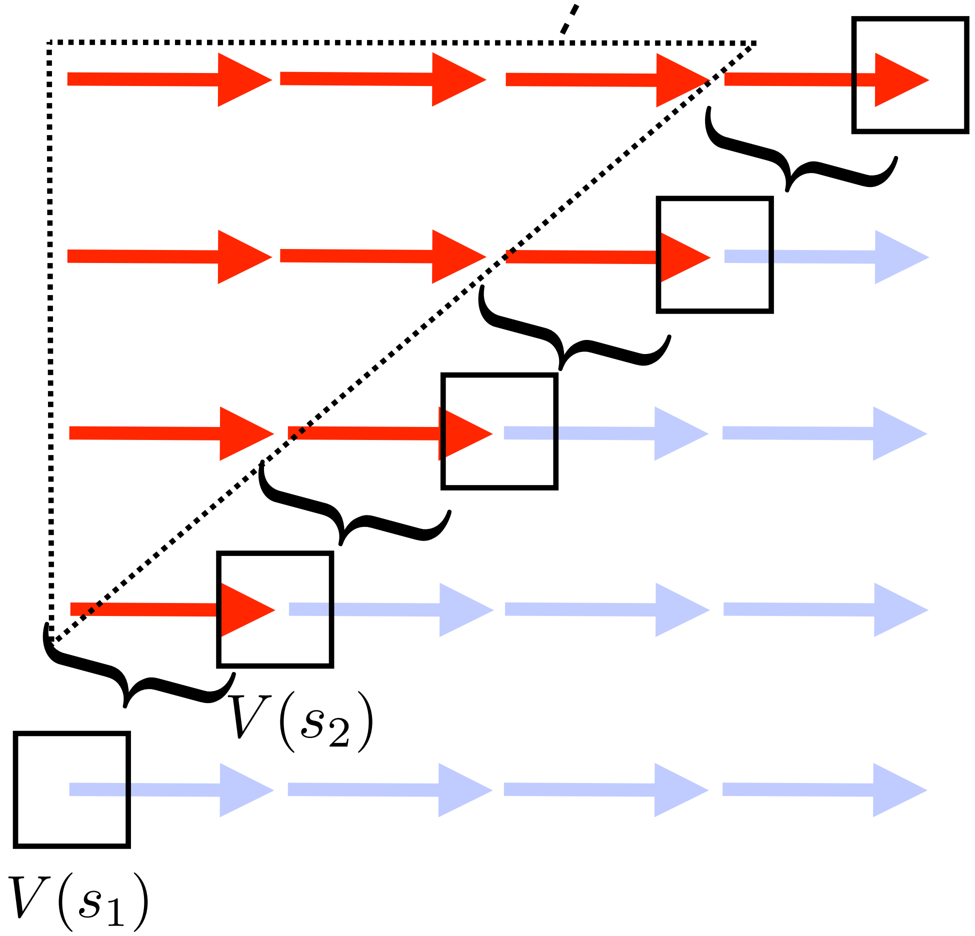
$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

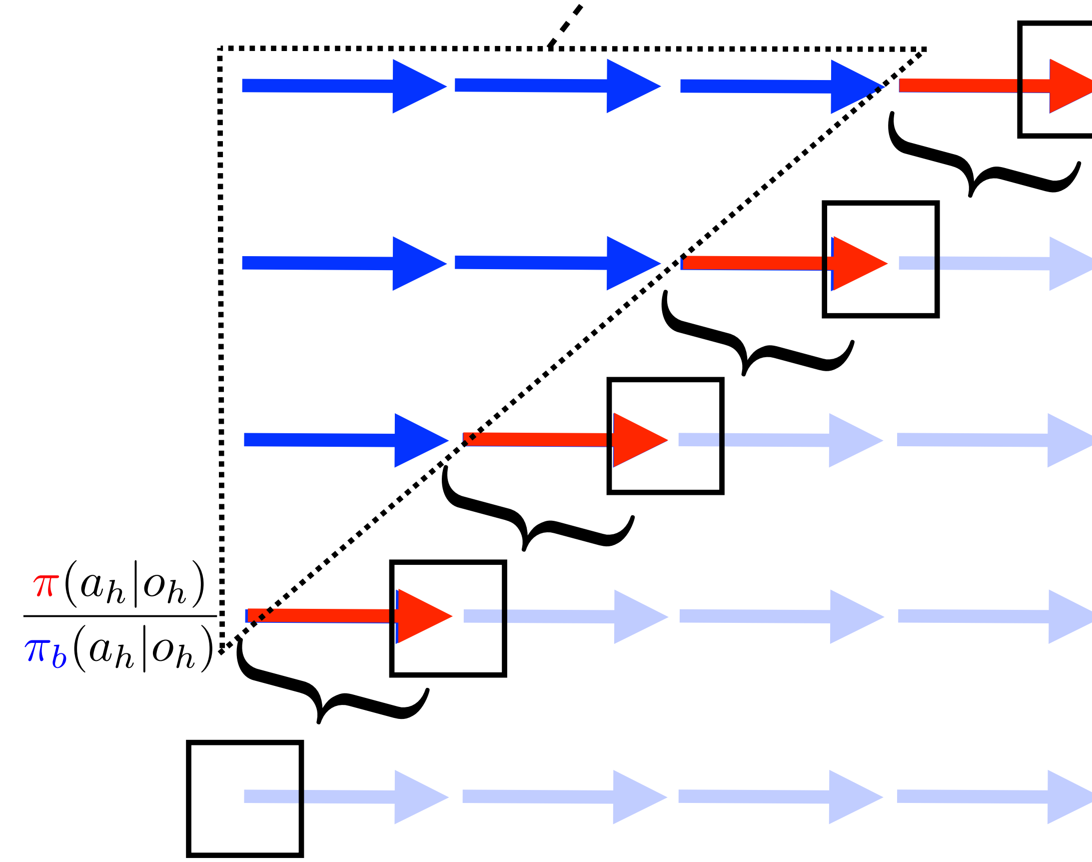
Prediction Groundtruth

"Bellman error"

$$\mathbb{E}_p[g^2] \leq \|p/q\|_\infty \cdot \mathbb{E}_q[g^2]$$

$$\arg\min_V \sum_{h=1}^{H} \mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$





How do value functions help in MDPs?

$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

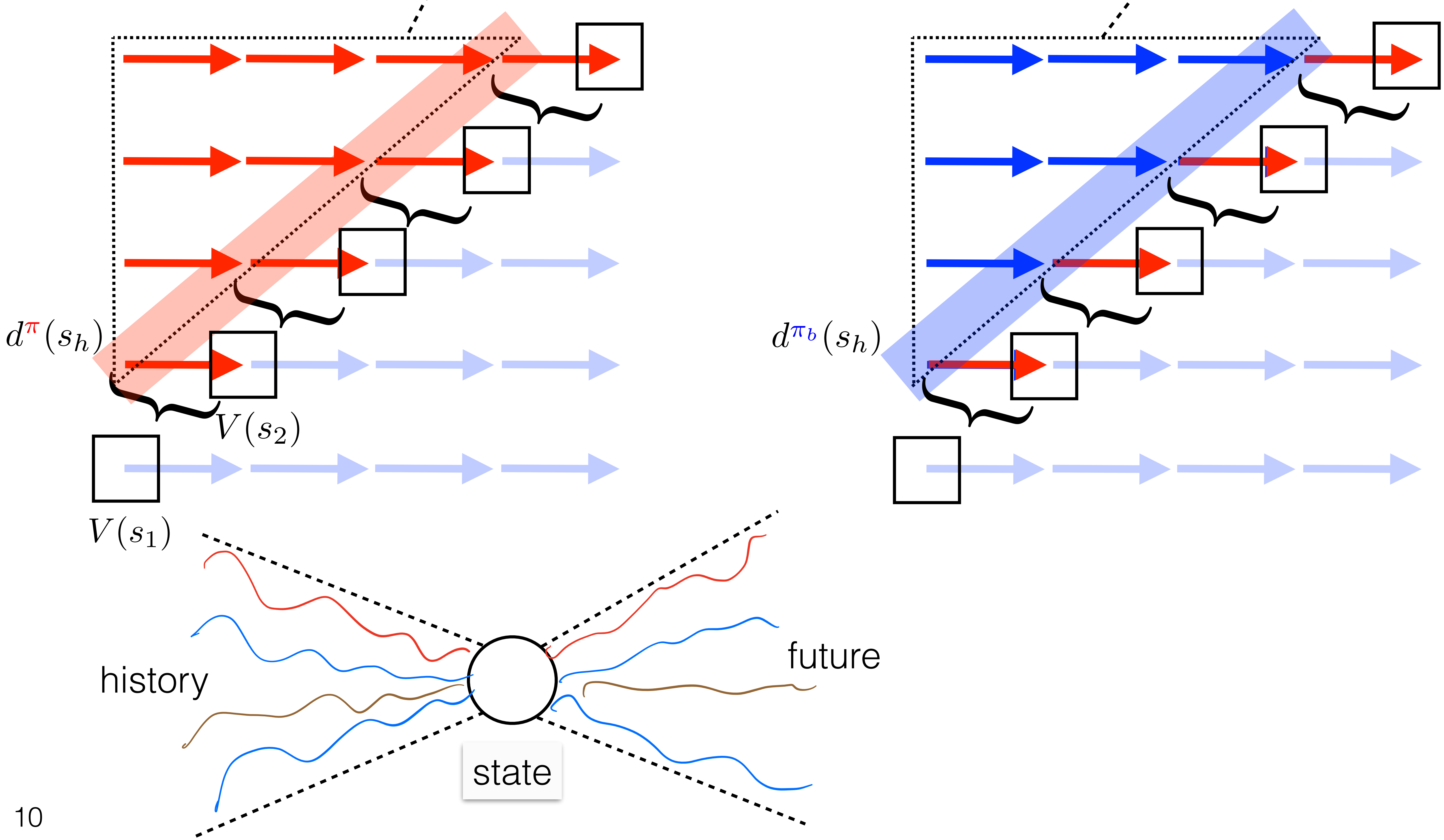
Prediction

Groundtruth

“ ” Bellman error

$$\mathbb{E}_p[g^2] \le \|p/q\|_\infty \cdot \mathbb{E}_q[g^2]$$

$$\arg\min_V \sum_{h=1}^{H} \mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$



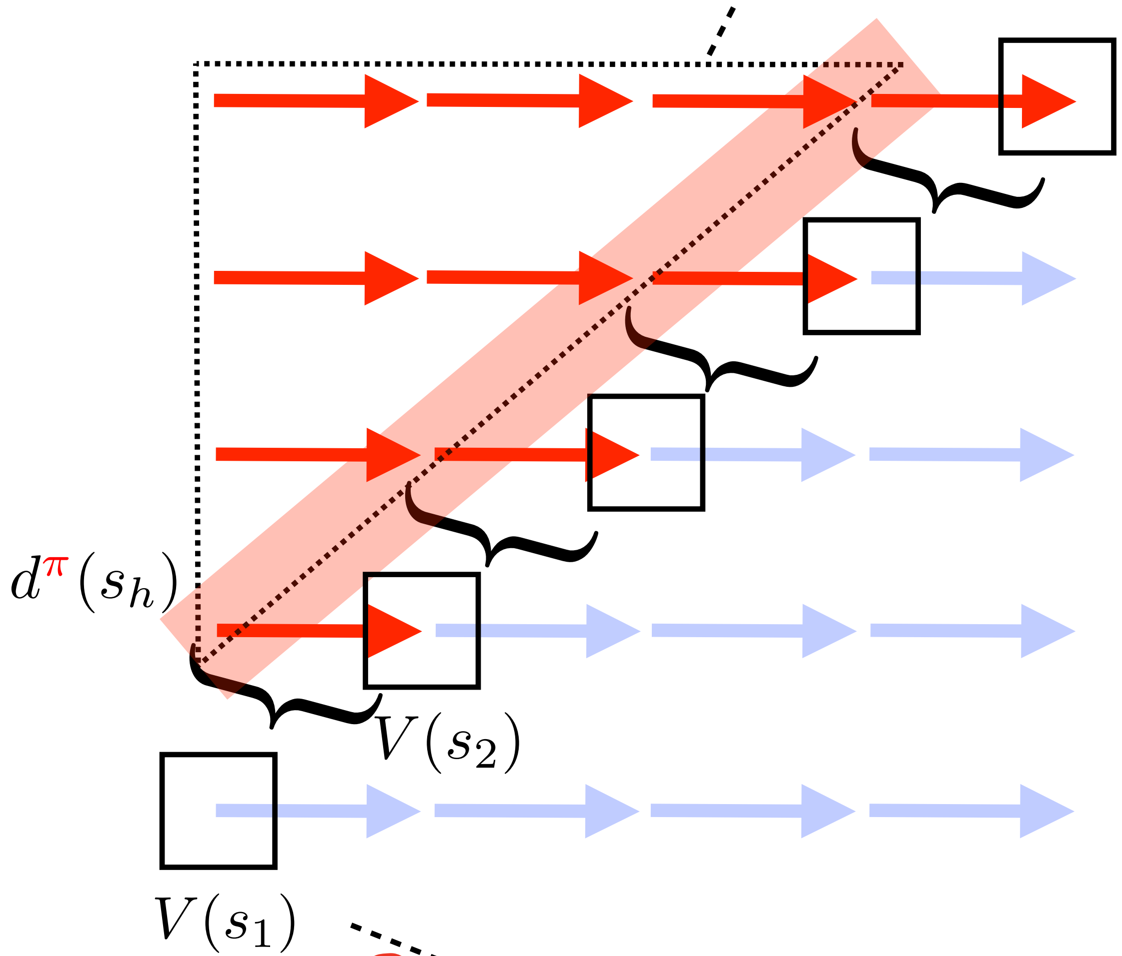
How do value functions help in MDPs?

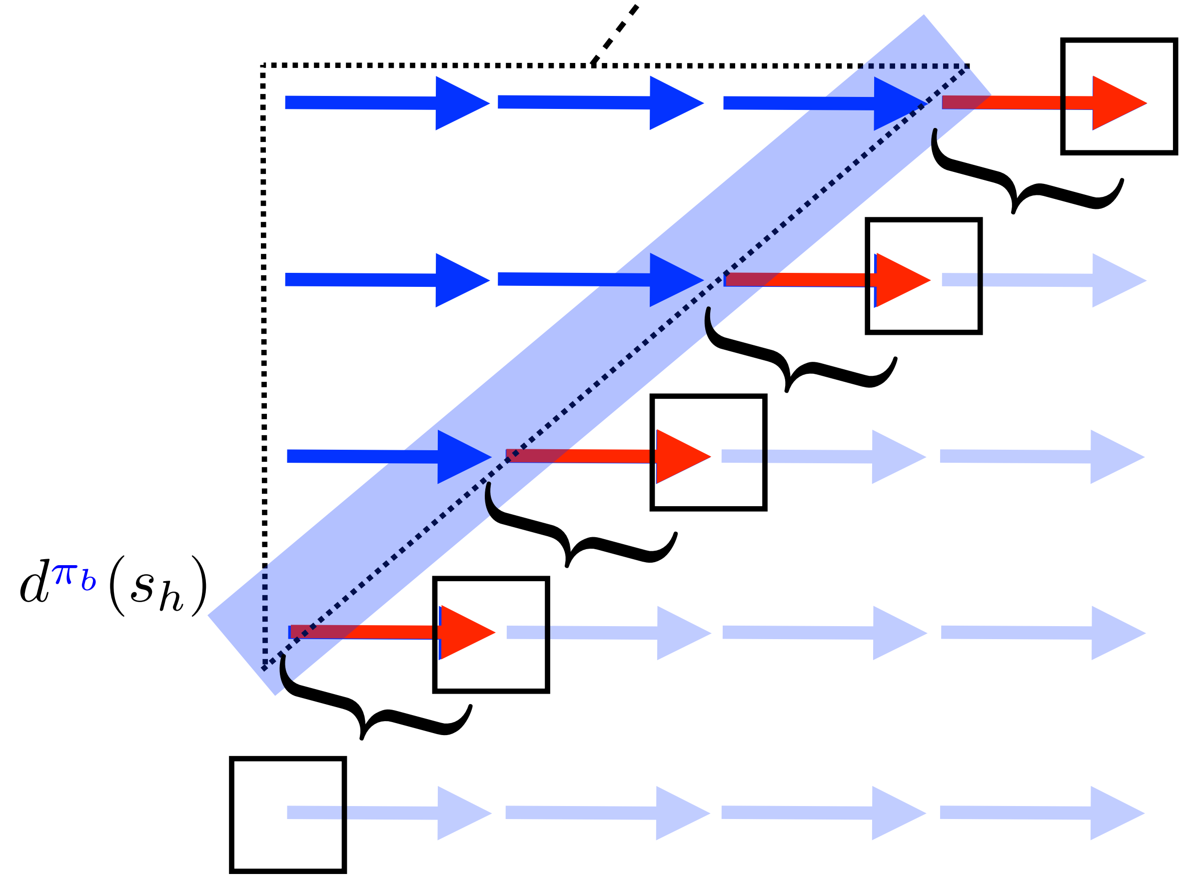
$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi}V)(s_h)]$$

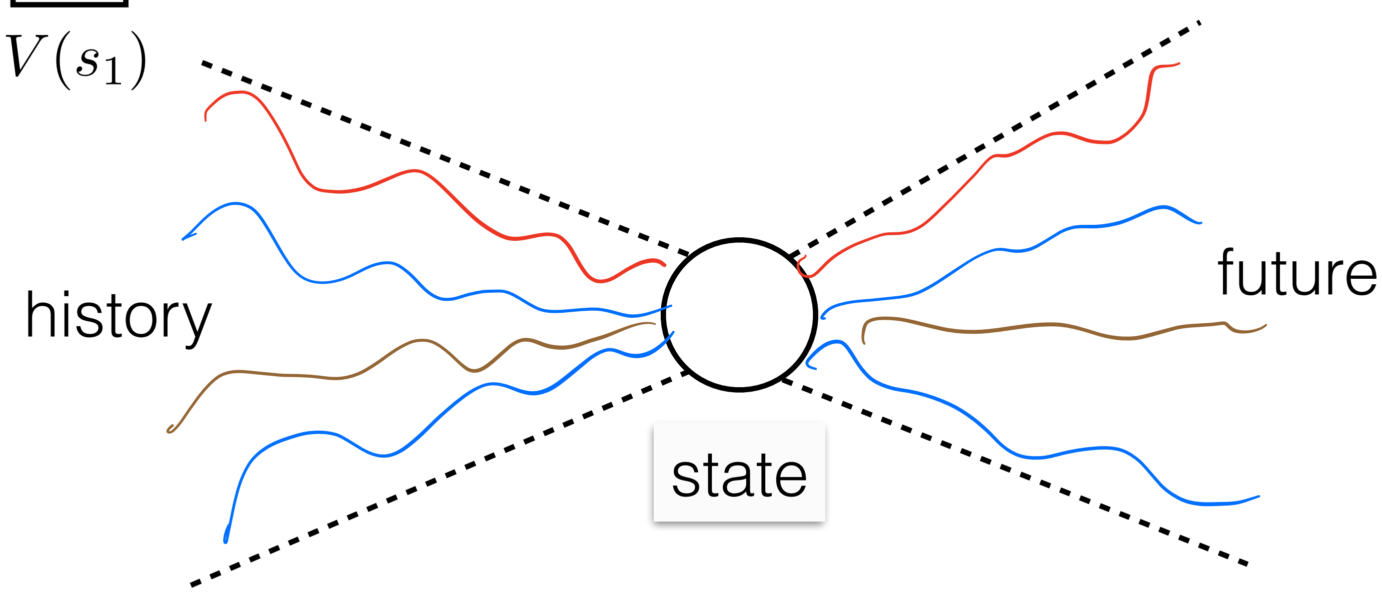
Prediction | Groundtruth | "Bellman error"

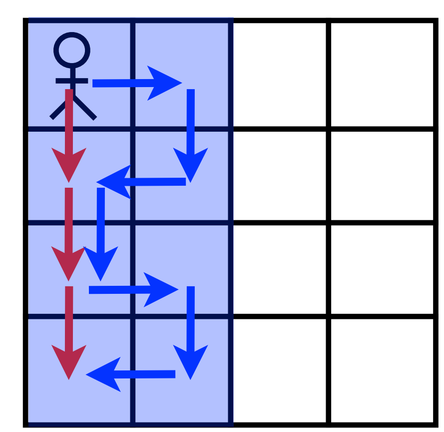
$$\mathbb{E}_p[g^2] \le \|p/q\|_\infty \cdot \mathbb{E}_q[g^2]$$

$$\arg\min_V \sum_{h=1}^{H} \mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi}V)(s_h))^2]$$

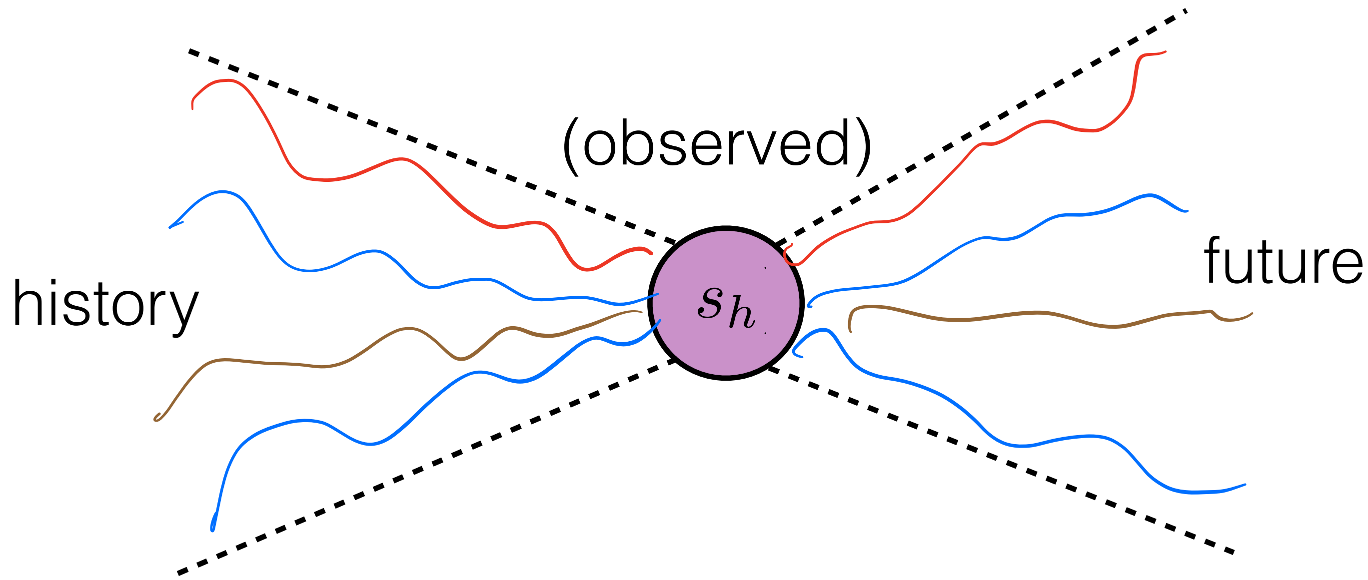




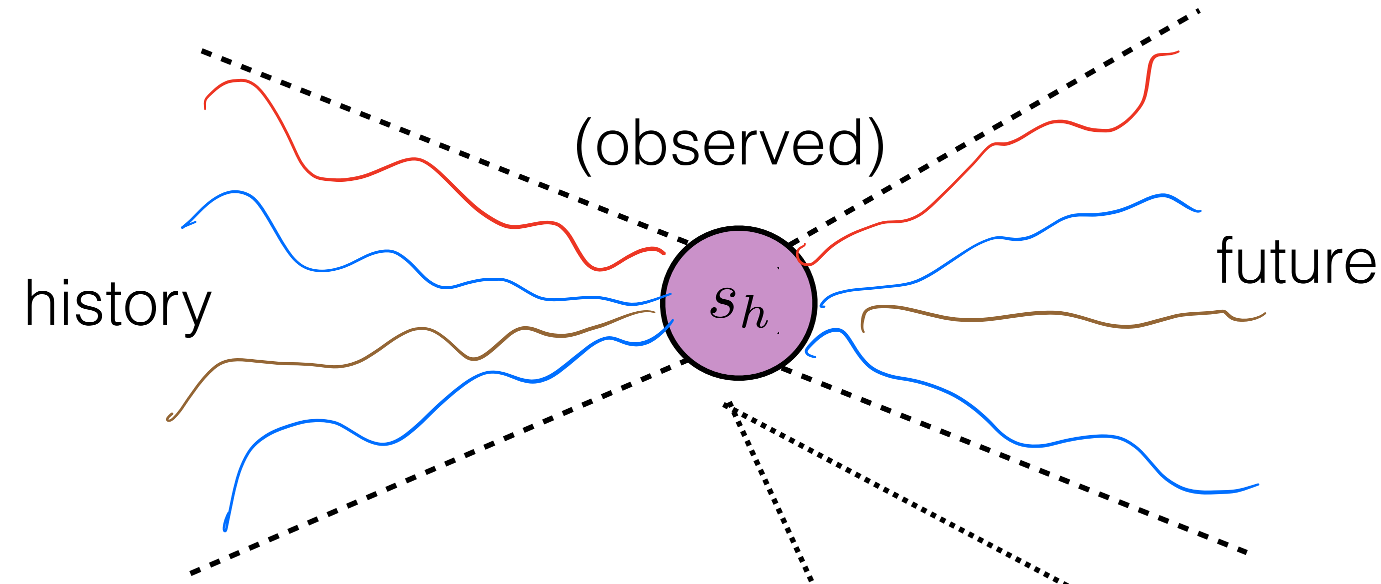




How do value functions help in MDPs?

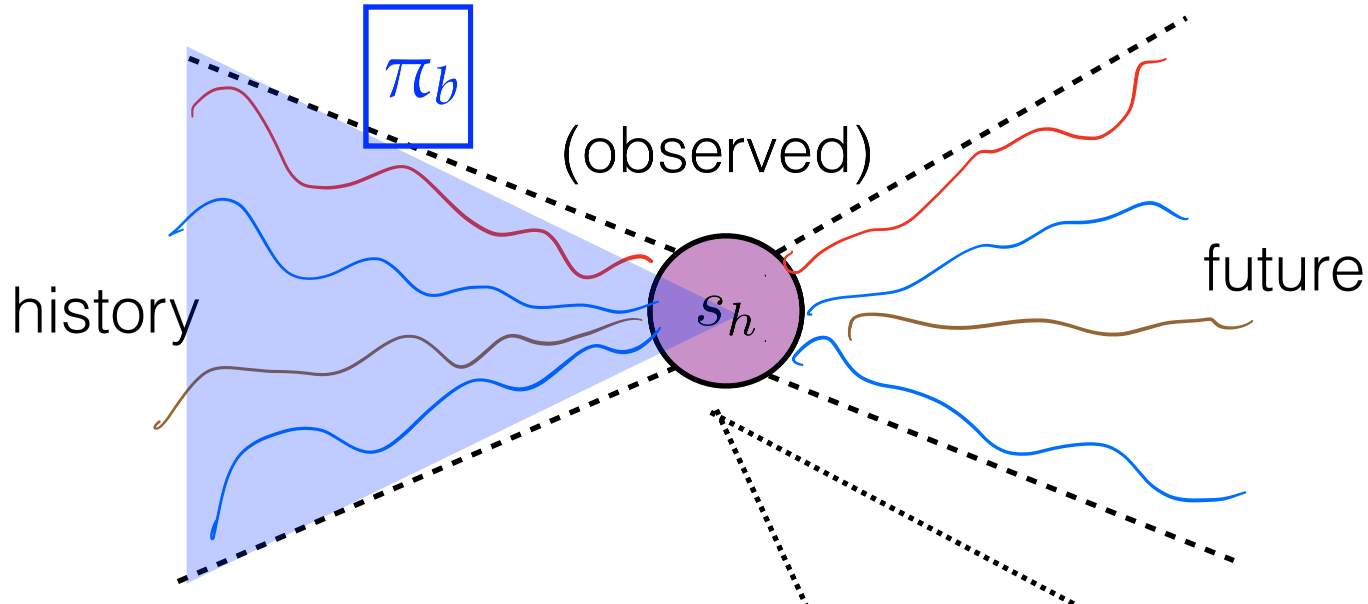


How do value functions help in MDPs?



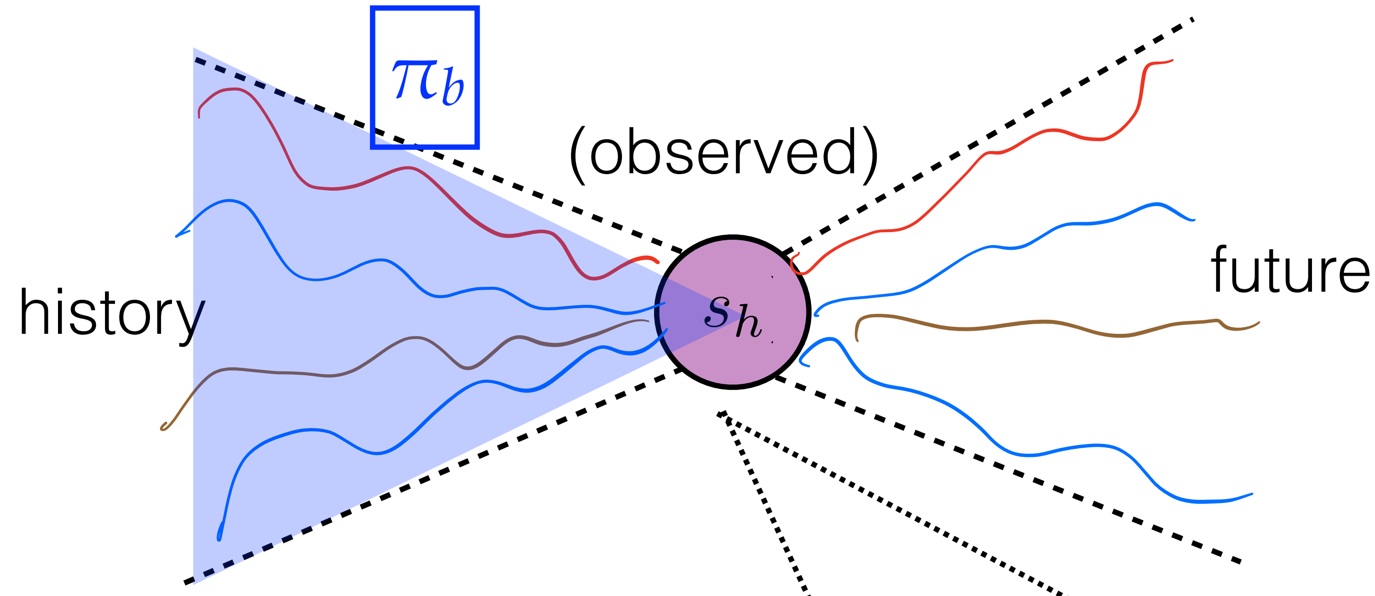
$\mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi}V)(s_h))^2]$

How do value functions help in MDPs?

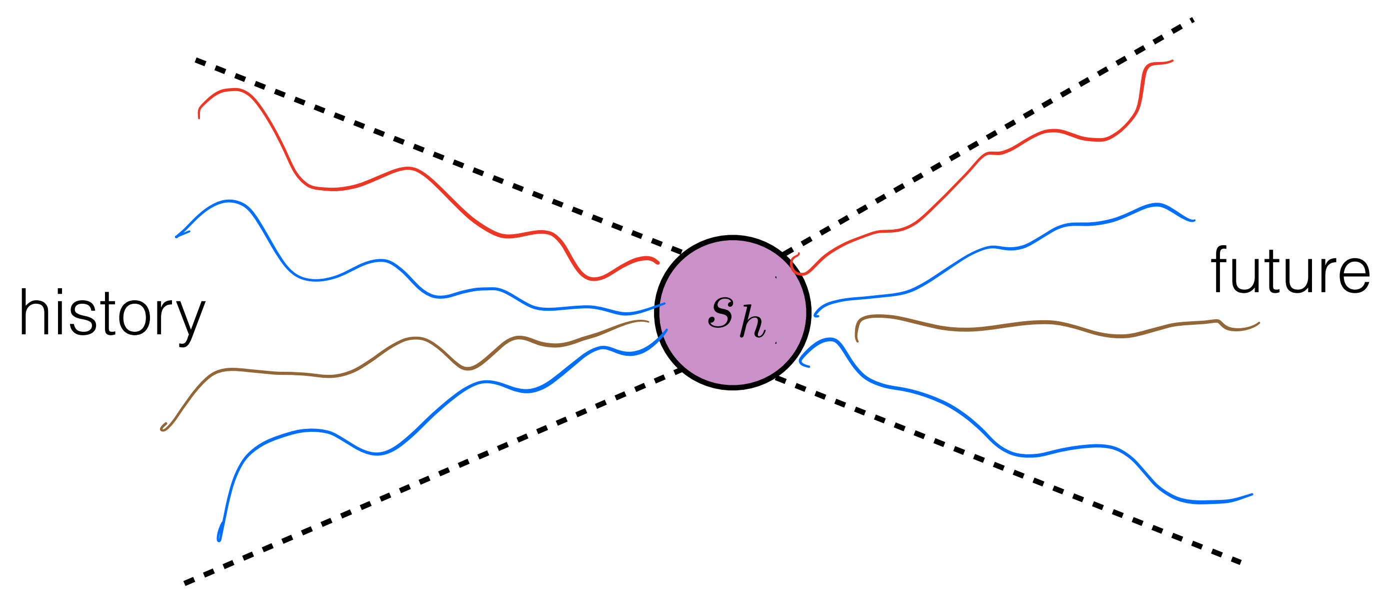


$$\mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$

How do value functions help in MDPs?



$$\mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$

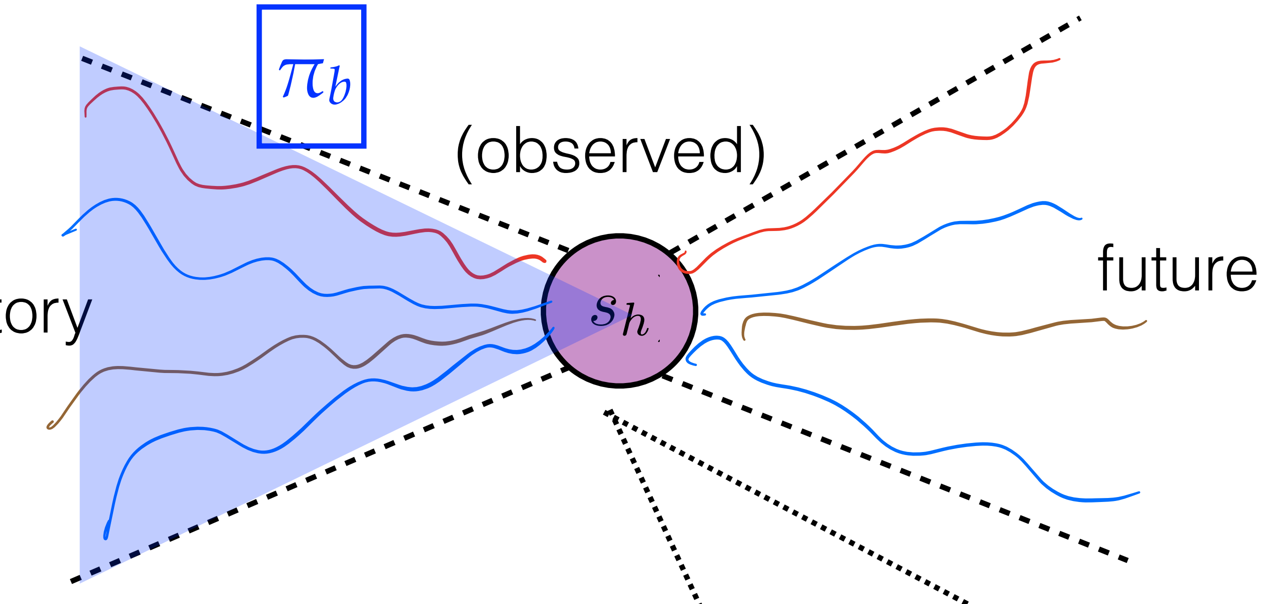


Prediction

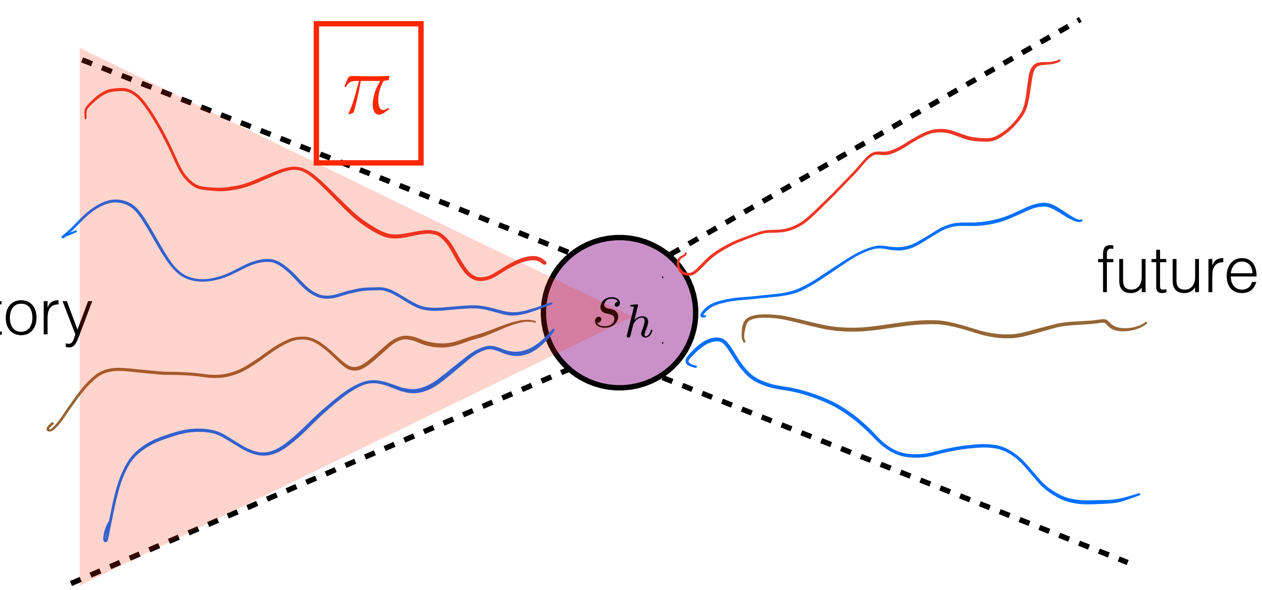
$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Groundtruth

How do value functions help in MDPs?



$$\mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$

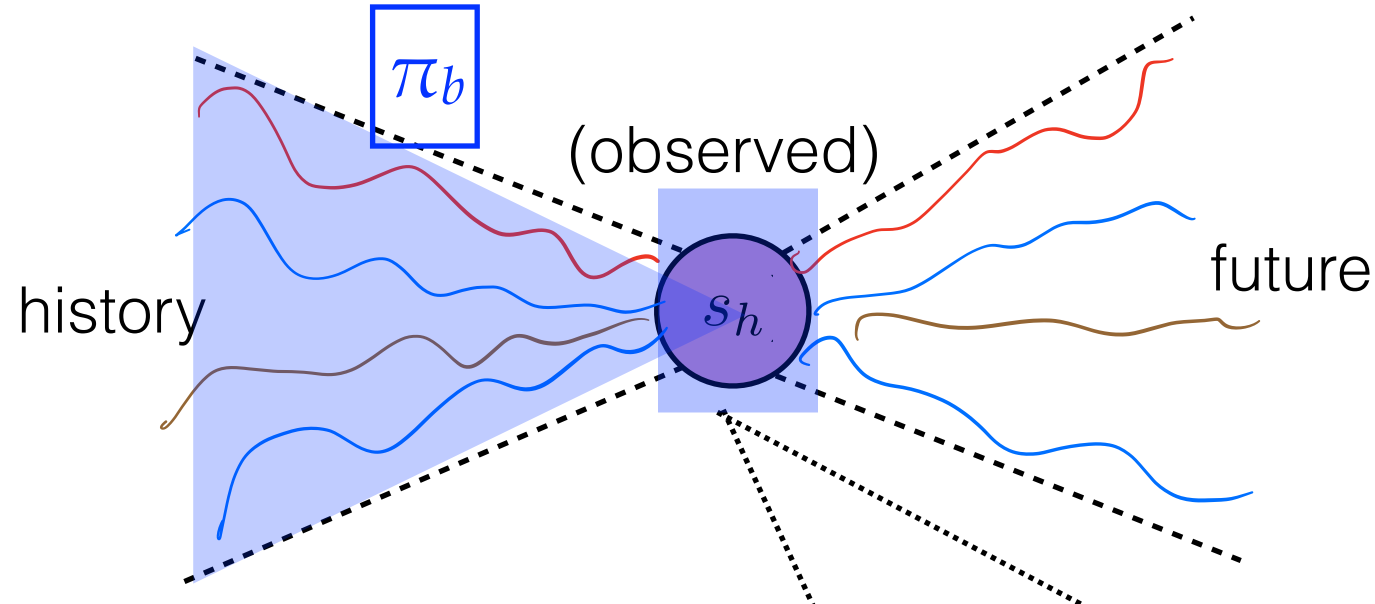


Prediction

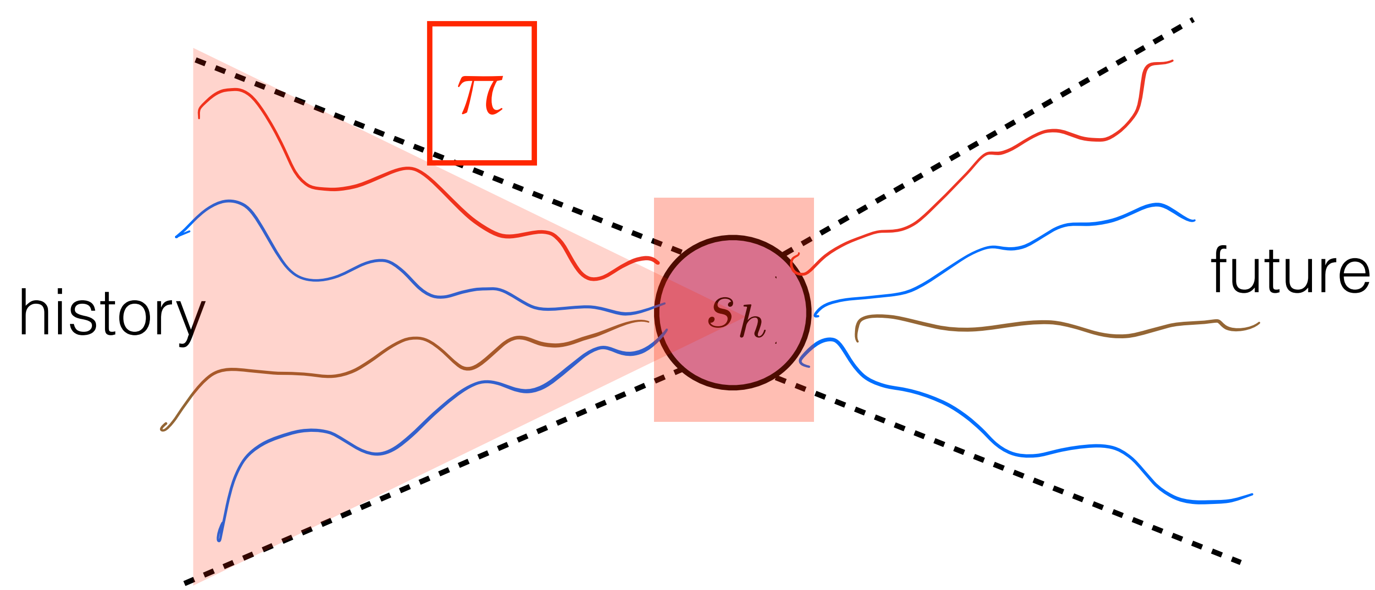
$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Groundtruth

How do value functions help in MDPs?



$$\mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$

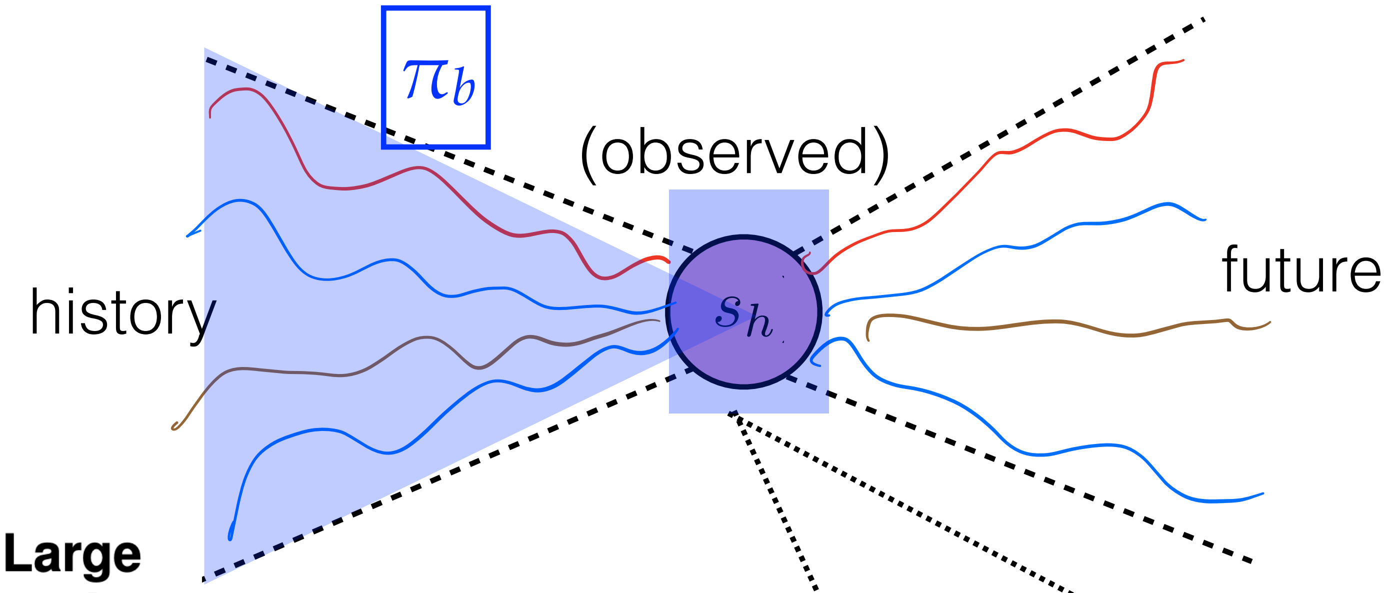


Prediction

$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Groundtruth

How do value functions help in MDPs?



Submitted to Statistical Science

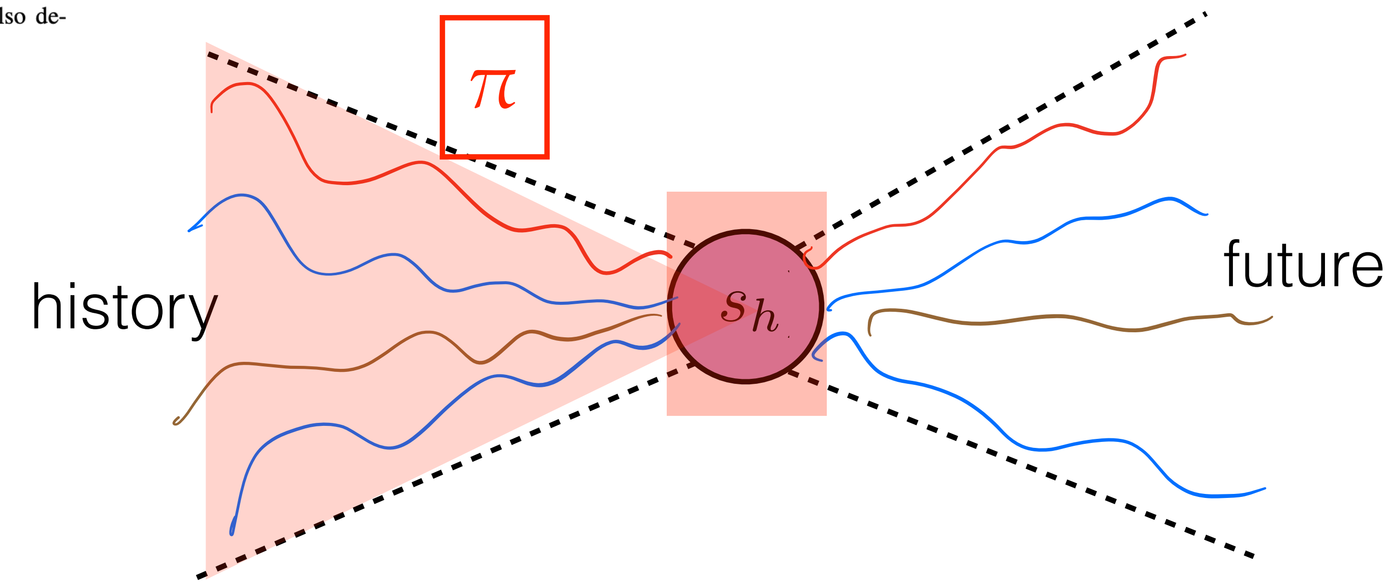
Offline Reinforcement Learning in Large State Spaces: Algorithms and Guarantees

Nan Jiang and Tengyang Xie (draft version; under review)

Abstract. This article introduces the theory of offline reinforcement learning in large state spaces, where good policies are learned from historical data without online interactions with the environment. Key concepts introduced include expressivity assumptions on function approximation (e.g., Bellman-completeness vs. realizability) and data coverage (e.g., all-policy vs. single-policy coverage), and a rich landscape of algorithms and results is described, depending on the assumptions one is willing to make and the sample and computational complexity guarantees one wishes to achieve. We also describe open questions and connections to adjacent areas.

Key words and phrases: offline reinforcement learning.

$$\mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$



Prediction

$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Groundtruth

Partially Observed (non-Markov) Problems

$$o_1, a_1, r_1, \ldots, o_h, a_h, r_h, \ldots o_H, a_H, r_H$$

Partially Observed (non-Markov) Problems

- Can always convert to MDP

$$\underbrace{o_1, a_1, r_1, \ldots, o_h}, a_h, r_h, \ldots o_H, a_H, r_H$$

  - Define new state τ_h. Problem solved?

Partially Observed (non-Markov) Problems

- Can always convert to MDP

$$\underbrace{o_1, a_1, r_1, \ldots, o_h}, a_h, r_h, \ldots o_H, a_H, r_H$$

  - Define new state τ_h. Problem solved?

- State density ratio: $\frac{d^{\pi}(o_1, a_1, \ldots, o_h)}{d^{\pi_b}(o_1, a_1, \ldots, o_h)}$

Partially Observed (non-Markov) Problems

- Can always convert to MDP

$$\underbrace{o_1, a_1, r_1, \ldots, o_h}, a_h, r_h, \ldots o_H, a_H, r_H$$

  - Define new state τ_h. Problem solved?

- State density ratio: $\frac{d^{\pi}(o_1, a_1, \ldots, o_h)}{d^{\pi_b}(o_1, a_1, \ldots, o_h)} = \prod_{h'=1}^{h-1} \frac{\pi(a_{h'}|o_{h'})}{\pi_b(a_{h'}|o_{h'})}$ **!!**

Partially Observed (non-Markov) Problems

- Can always convert to MDP

$$\underbrace{o_1, a_1, r_1, \ldots, o_h}, a_h, r_h, \ldots o_H, a_H, r_H$$

  - Define new state τ_h. Problem solved?

- State density ratio: $\dfrac{d^{\pi}(o_1, a_1, \ldots, o_h)}{d^{\pi_b}(o_1, a_1, \ldots, o_h)} = \displaystyle\prod_{h'=1}^{h-1} \frac{\pi(a_{h'}|o_{h'})}{\pi_b(a_{h'}|o_{h'})}$ **!!**

- Value in RLHF: > exp(25)!!!

Bai et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Partially Observed (non-Markov) Problems

- Can always convert to MDP

$$\underbrace{o_1, a_1, r_1, \ldots, o_h}, a_h, r_h, \ldots o_H, a_H, r_H$$

  - Define new state τ_h. Problem solved?

- State density ratio: $\frac{d^{\pi}(o_1, a_1, \ldots, o_h)}{d^{\pi_b}(o_1, a_1, \ldots, o_h)} = \prod_{h'=1}^{h-1} \frac{\pi(a_{h'}|o_{h'})}{\pi_b(a_{h'}|o_{h'})}$ **!!**

- Structure can help: e.g., if $\mathcal{V}$ is linear in feature $\phi : \mathcal{S} \to \mathbb{R}^d$

Partially Observed (non-Markov) Problems

- Can always convert to MDP

$$\underbrace{o_1, a_1, r_1, \ldots, o_h}, a_h, r_h, \ldots o_H, a_H, r_H$$

  - Define new state τ_h. Problem solved?

- State density ratio: $\frac{d^{\pi}(o_1, a_1, \ldots, o_h)}{d^{\pi_b}(o_1, a_1, \ldots, o_h)} = \prod_{h'=1}^{h-1} \frac{\pi(a_{h'}|o_{h'})}{\pi_b(a_{h'}|o_{h'})}$ **!!**

- Structure can help: e.g., if $\mathcal{V}$ is linear in feature $\phi : \mathcal{S} \to \mathbb{R}^d$

$$\sup_{V \in \mathcal{V}} \frac{|\mathbb{E}_{\pi}[V - \mathcal{T}^{\pi} V]|}{\sqrt{\mathbb{E}_{\pi_b}[(V - \mathcal{T}^{\pi} V)^2]}} \leq \mathbb{E}_{\pi}[\phi]^{\top} \mathbb{E}_{\pi_b}[\phi\phi^{\top}]^{-1} \mathbb{E}_{\pi}[\phi]$$

* Also needs Bellman completeness

Partially Observed (non-Markov) Problems

- Can always convert to MDP

$$\underbrace{o_1, a_1, r_1, \ldots, o_h}, a_h, r_h, \ldots o_H, a_H, r_H$$

  - Define new state τ_h. Problem solved?
- State density ratio: $\frac{d^{\pi}(o_1, a_1, \ldots, o_h)}{d^{\pi_b}(o_1, a_1, \ldots, o_h)} = \prod_{h'=1}^{h-1} \frac{\pi(a_{h'}|o_{h'})}{\pi_b(a_{h'}|o_{h'})}$ **!!**
- Structure can help: e.g., if $\mathcal{V}$ is linear in feature $\phi : \mathcal{S} \to \mathbb{R}^d$

$$\sup_{V \in \mathcal{V}} \frac{|\mathbb{E}_{\pi}[V - \mathcal{T}^{\pi} V]|}{\sqrt{\mathbb{E}_{\pi_b}[(V - \mathcal{T}^{\pi} V)^2]}} \leq \mathbb{E}_{\pi}[\phi]^{\top} \mathbb{E}_{\pi_b}[\phi\phi^{\top}]^{-1} \mathbb{E}_{\pi}[\phi]$$

  - this assumes given low-dim linear feature to encode history…

* Also needs Bellman completeness

Partially Observed (non-Markov) Problems

- Can always convert to MDP

$$\underbrace{o_1, a_1, r_1, \ldots, o_h}, a_h, r_h, \ldots o_H, a_H, r_H$$

  - Define new state τ_h. Problem solved?

- State density ratio: $\dfrac{d^{\pi}(o_1, a_1, \ldots, o_h)}{d^{\pi_b}(o_1, a_1, \ldots, o_h)} = \displaystyle\prod_{h'=1}^{h-1} \frac{\pi(a_{h'}|o_{h'})}{\pi_b(a_{h'}|o_{h'})}$ **!!**

- Structure can help: e.g., if $\mathcal{V}$ is linear in feature $\phi : \mathcal{S} \to \mathbb{R}^d$

$$\sup_{V \in \mathcal{V}} \frac{|\mathbb{E}_{\pi}[V - \mathcal{T}^{\pi} V]|}{\sqrt{\mathbb{E}_{\pi_b}[(V - \mathcal{T}^{\pi} V)^2]}} \leq \mathbb{E}_{\pi}[\phi]^{\top} \mathbb{E}_{\pi_b}[\phi \phi^{\top}]^{-1} \mathbb{E}_{\pi}[\phi]$$

  - this assumes given low-dim linear feature to encode history…
  - side q: what structure in $\mathcal{V}$ balances expressivity and coverage

* Also needs Bellman completeness

Partially Observed (non-Markov) Problems

- Can always convert to MDP

$$\underbrace{o_1, a_1, r_1, \ldots, o_h}, a_h, r_h, \ldots o_H, a_H, r_H$$

  - Define new state τ_h. Problem solved?

- State density ratio: $\dfrac{d^{\pi}(o_1, a_1, \ldots, o_h)}{d^{\pi_b}(o_1, a_1, \ldots, o_h)} = \prod_{h'=1}^{h-1} \dfrac{\pi(a_{h'}|o_{h'})}{\pi_b(a_{h'}|o_{h'})}$ **!!**

- Structure can help: e.g., if $\mathcal{V}$ is linear in feature $\phi : \mathcal{S} \to \mathbb{R}^d$

$$\sup_{V \in \mathcal{V}} \frac{|\mathbb{E}_{\pi}[V - \mathcal{T}^{\pi} V]|}{\sqrt{\mathbb{E}_{\pi_b}[(V - \mathcal{T}^{\pi} V)^2]}} \leq \mathbb{E}_{\pi}[\phi]^{\top} \mathbb{E}_{\pi_b}[\phi\phi^{\top}]^{-1} \mathbb{E}_{\pi}[\phi]$$

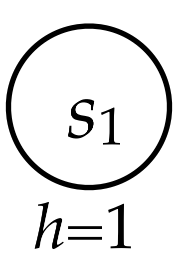
  - this assumes given low-dim linear feature to encode history…

Can we avoid the exponentials in OPE in PO settings, without relying on structured function classes?

* Also needs Bellman completeness

Partially Observable MDPs (POMDPs)

- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)

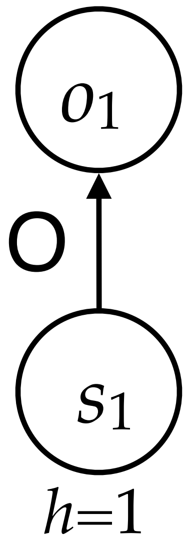


Partially Observable MDPs (POMDPs)

$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)
  - agent observes $o_h \in O_h$ (large), $o_h \sim$ O$(\cdot \mid s_h)$

emission process
O: $S \to \Delta(O)$

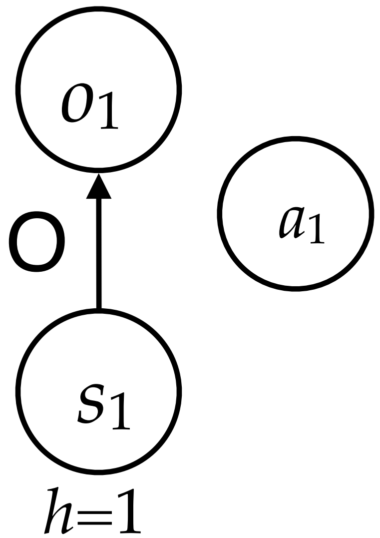


Partially Observable MDPs (POMDPs)

$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)
  - agent observes $o_h \in O_h$ (large), $o_h \sim \mathbb{O}(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)

emission process
$\mathbb{O}: S \to \Delta(O)$



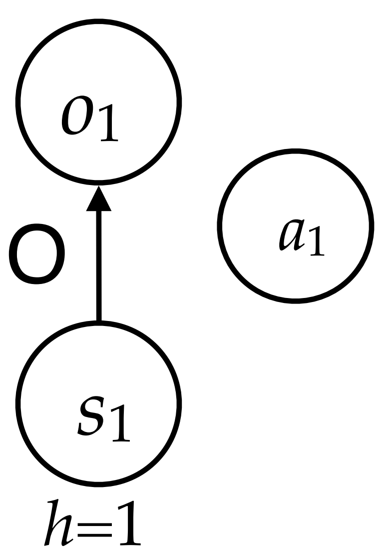
$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

Partially Observable MDPs (POMDPs)

- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)
  - agent observes $o_h \in O_h$ (large), $o_h \sim \mathrm{O}(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)
  - receives reward $r_h = R(o_h, a_h)$

emission process
$\mathrm{O}: S \to \Delta(O)$

reward function
$R: S \times A \to [0,1]$



Partially Observable MDPs (POMDPs)

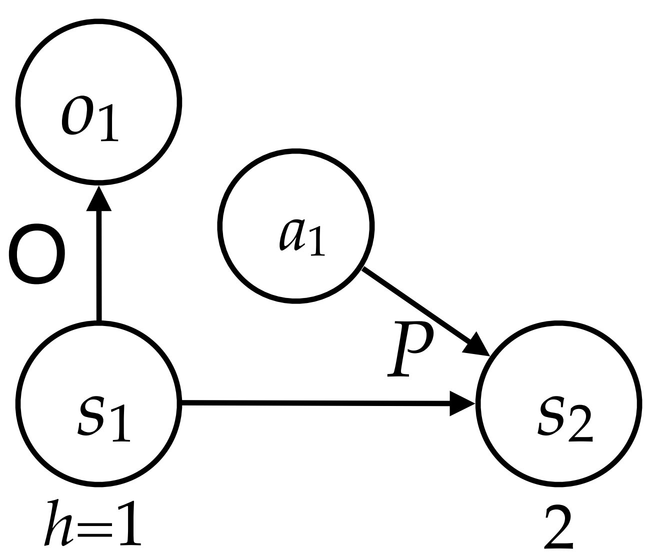
$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)
    - $s_h \sim P(\cdot \mid s_{h\text{-}1}, a_{h\text{-}1})$ for $h \geq 2$
  - agent observes $o_h \in O_h$ (large), $o_h \sim \mathrm{O}(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)
  - receives reward $r_h = R(o_h, a_h)$

transition dynamics
$P: S \times A \to \Delta(S)$

emission process
$\mathrm{O}: S \to \Delta(O)$

reward function
$R: S \times A \to [0,1]$



$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

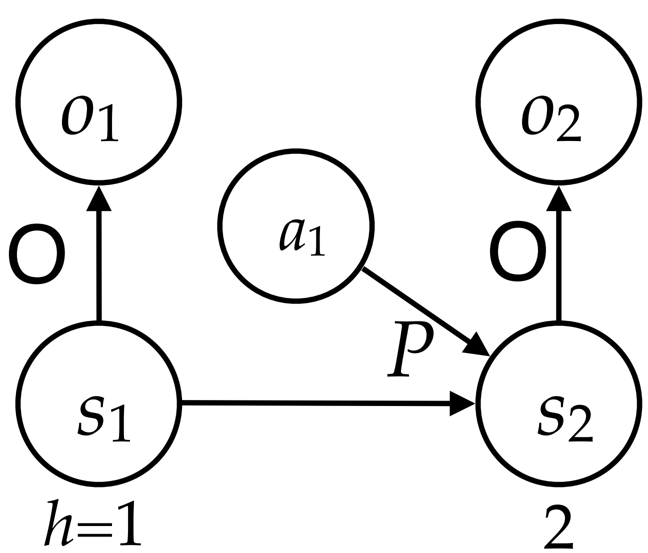
Partially Observable MDPs (POMDPs)

- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)
    - $s_h \sim P(\cdot \mid s_{h-1}, a_{h-1})$ for $h \geq 2$
  - agent observes $o_h \in O_h$ (large), $o_h \sim \mathrm{O}(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)
  - receives reward $r_h = R(o_h, a_h)$

transition dynamics
$P: S \times A \to \Delta(S)$

emission process
$\mathrm{O}: S \to \Delta(O)$

reward function
$R: S \times A \to [0,1]$



$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

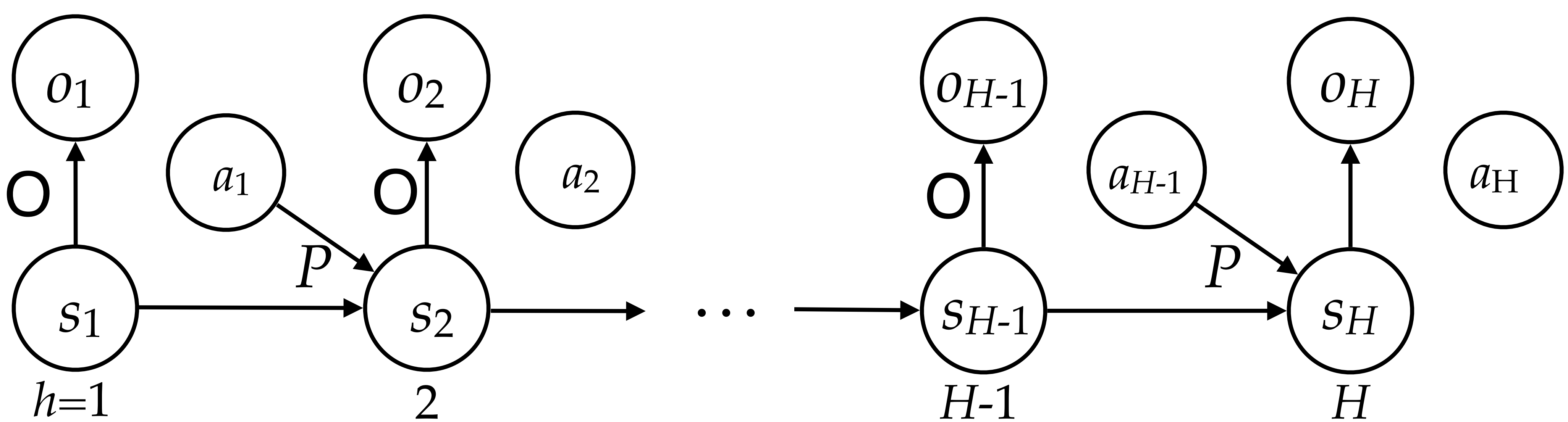
Partially Observable MDPs (POMDPs)

- For $h = 1, 2, \ldots, H,$
  - nature generates *latent* state $s_h \in S_h$ (small?)
    - $s_h \sim P(\cdot \mid s_{h\text{-}1}, a_{h\text{-}1})$ for $h \geq 2$
  - agent observes $o_h \in O_h$ (large), $o_h \sim \mathrm{O}(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)
  - receives reward $r_h = R(o_h, a_h)$

transition dynamics
$P: S{\times}A \rightarrow \Delta(S)$

emission process
$\mathrm{O}: S \rightarrow \Delta(O)$

reward function
$R: S{\times}A \rightarrow [0,1]$



$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

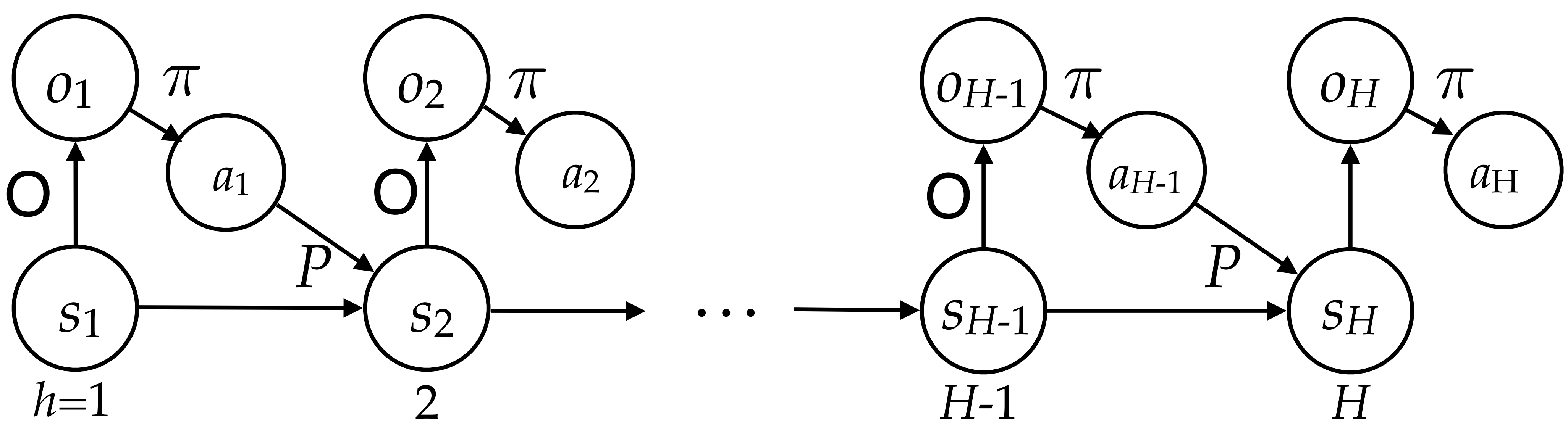
Partially Observable MDPs (POMDPs)

- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)
    - $s_h \sim P(\cdot \mid s_{h-1}, a_{h-1})$ for $h \geq 2$
  - agent observes $o_h \in O_h$ (large), $o_h \sim \mathrm{O}(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)
  - receives reward $r_h = R(o_h, a_h)$
- *Memoryless* policies $\pi : \mathcal{O} \to \Delta(\mathcal{A})$

transition dynamics
$P: S \times A \to \Delta(S)$

emission process
$\mathrm{O}: S \to \Delta(O)$

reward function
$R: S \times A \to [0,1]$



$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

Partially Observable MDPs (POMDPs)

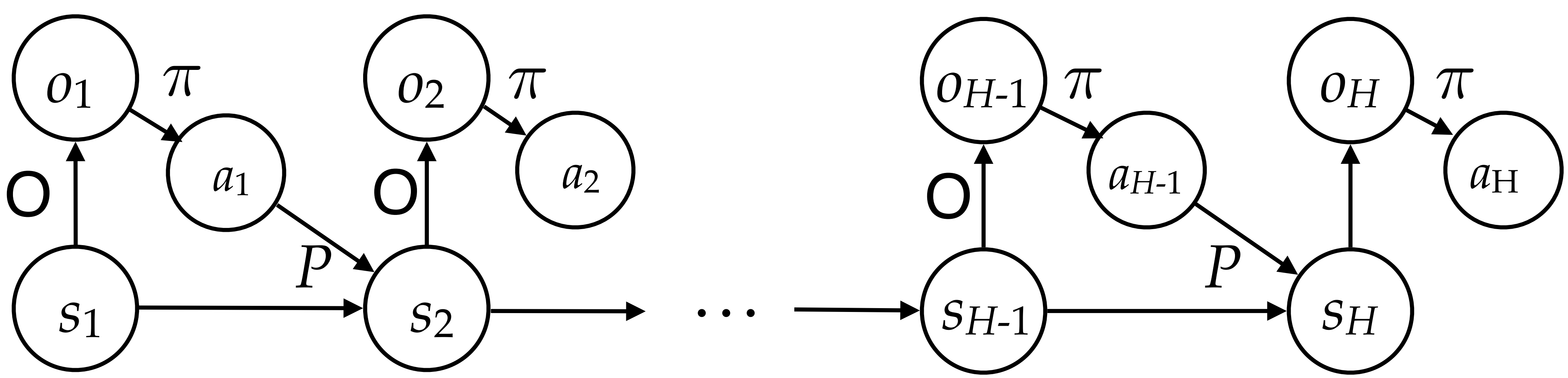
- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)
    - $s_h \sim P(\cdot \mid s_{h\text{-}1}, a_{h\text{-}1})$ for $h \geq 2$
  - agent observes $o_h \in O_h$ (large), $o_h \sim \mathrm{O}(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)
  - receives reward $r_h = R(o_h, a_h)$

transition dynamics
$P: S \times A \to \Delta(S)$

emission process
$\mathrm{O}: S \to \Delta(O)$

reward function
$R: S \times A \to [0,1]$

- *Memoryless* policies $\pi : \mathcal{O} \to \Delta(\mathcal{A})$
- Goal: estimate $J(\pi) := \mathbb{E}_\pi[\sum_{h=1}^{H} r_h]$ using episodes collected by π_b



Partially Observable MDPs (POMDPs)

$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

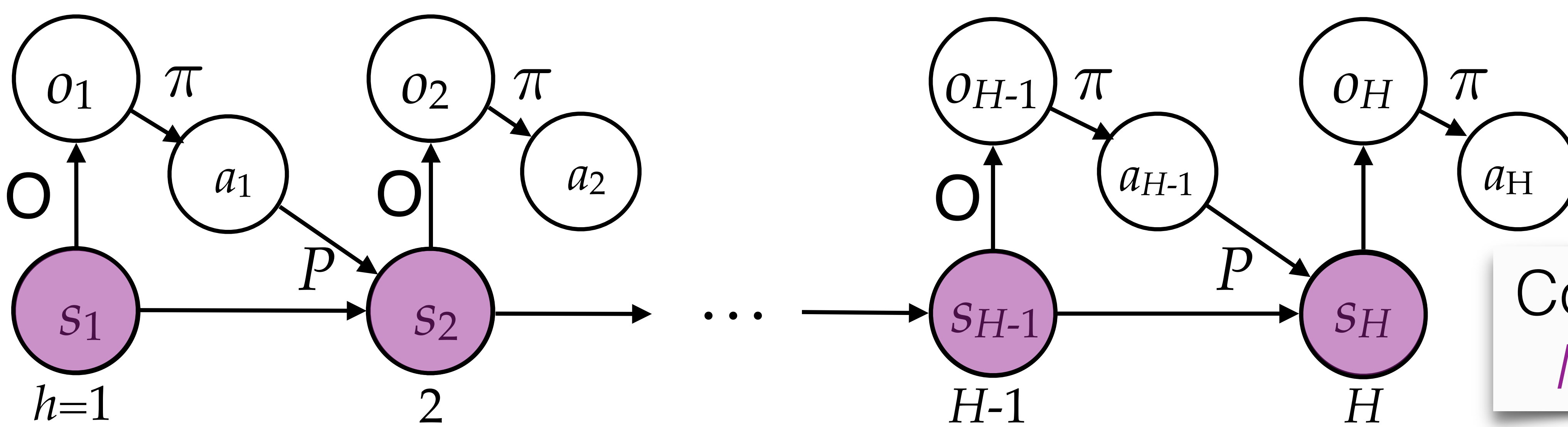
- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)
    - $s_h \sim P(\cdot \mid s_{h-1}, a_{h-1})$ for $h \geq 2$
  - agent observes $o_h \in O_h$ (large), $o_h \sim \mathbb{O}(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)
  - receives reward $r_h = R(o_h, a_h)$

transition dynamics
$P: S \times A \to \Delta(S)$

emission process
$\mathbb{O}: S \to \Delta(O)$

reward function
$R: S \times A \to [0,1]$

- *Memoryless* policies $\pi : \mathcal{O} \to \Delta(\mathcal{A})$
- Goal: estimate $J(\pi) := \mathbb{E}_\pi[\sum_{h=1}^{H} r_h]$ using episodes collected by π_b

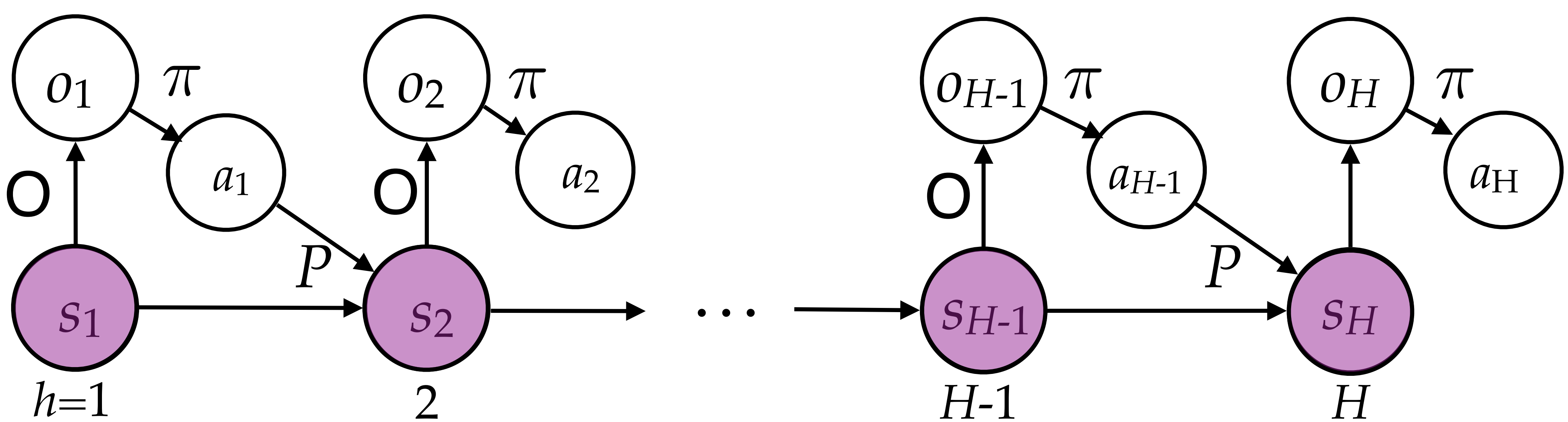


Coverage over *latent* state?

Future-Dependent Value Function

- Define: value function of latent state

$$V_{\mathcal{S}}^{\pi}(s_h) := \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r_{h'} | s_h] \in [0, H]$$

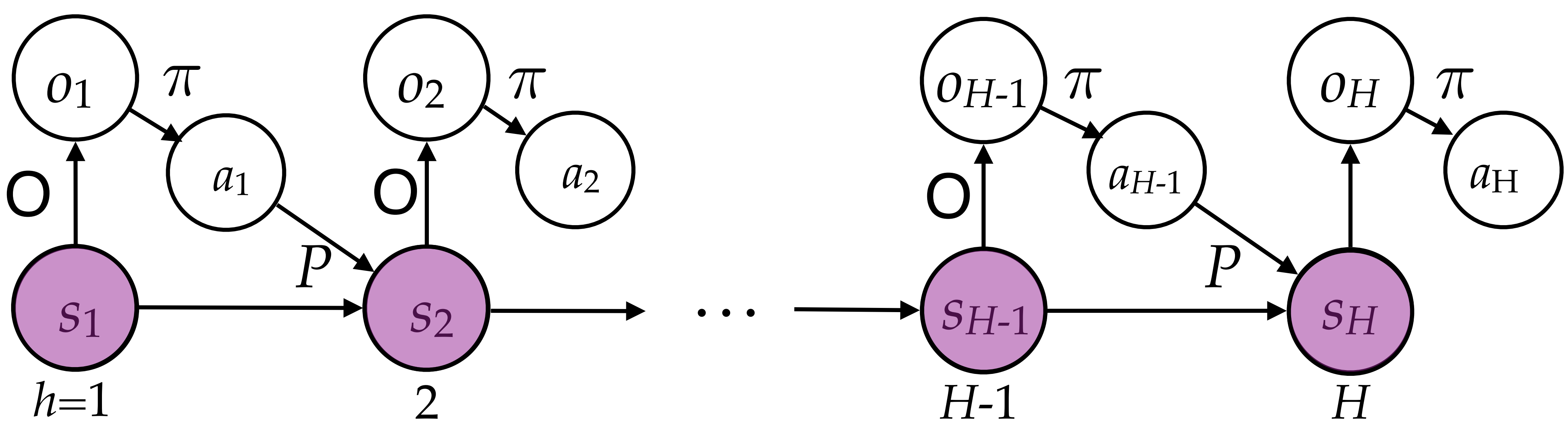


Future-Dependent Value Function

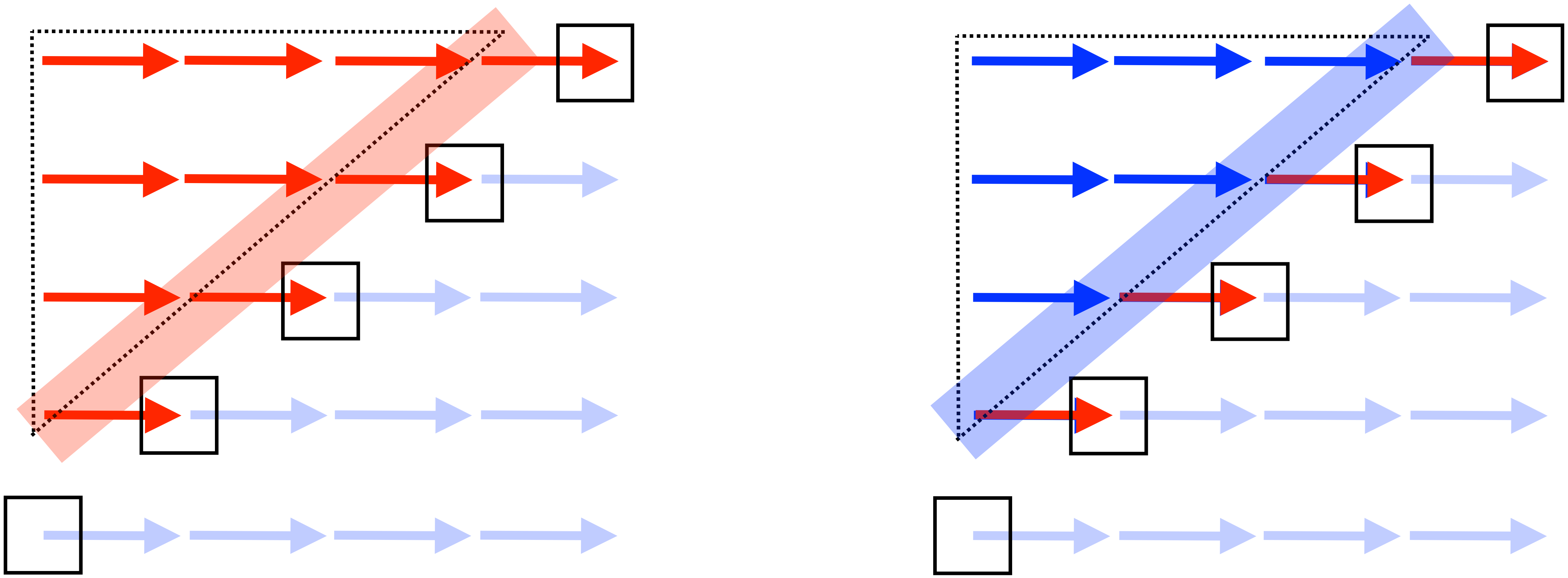
- Define: value function of latent state

$$V_{\mathcal{S}}^{\pi}(s_h) := \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r_{h'} | s_h] \in [0, H]$$

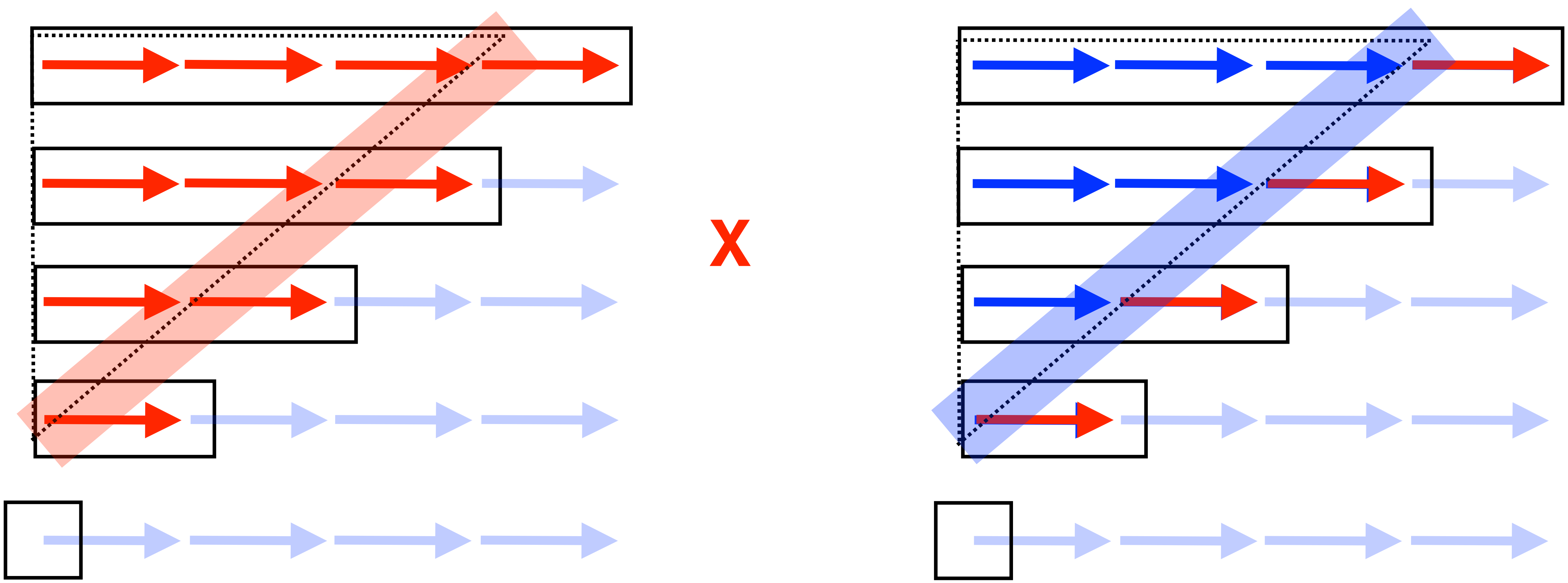
  - Problem: s_h is latent — can't even use this function!



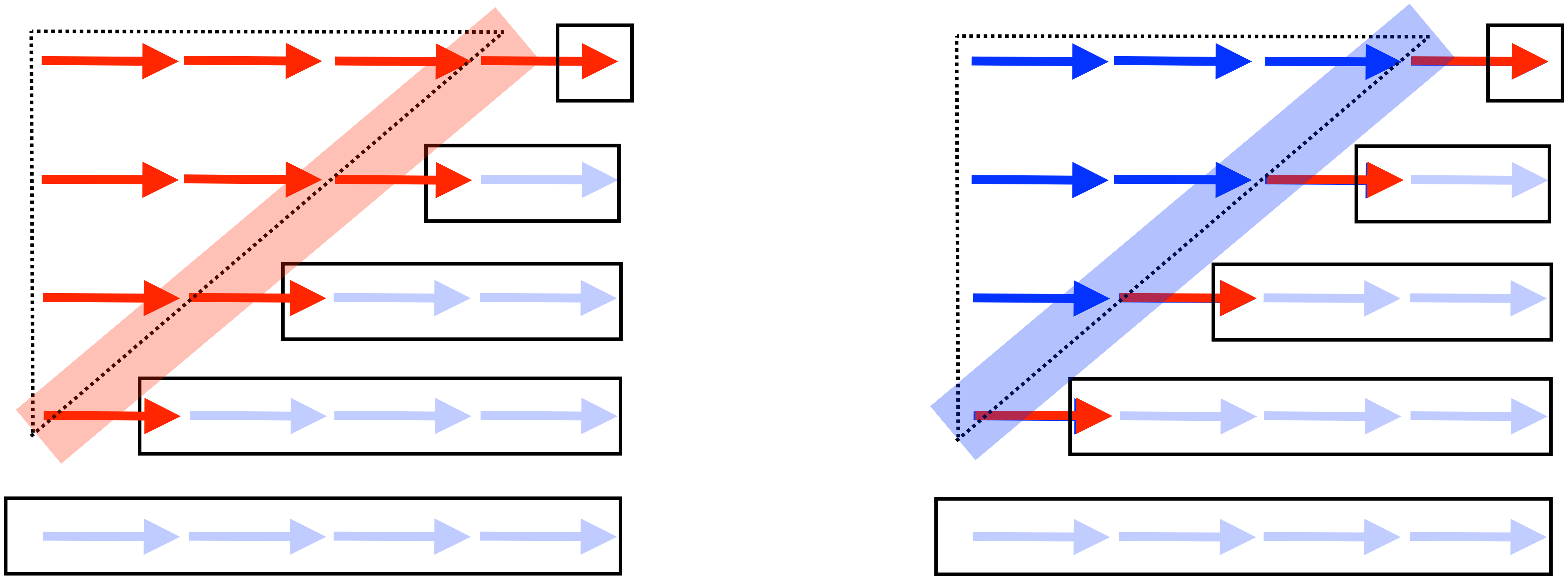
Value function?

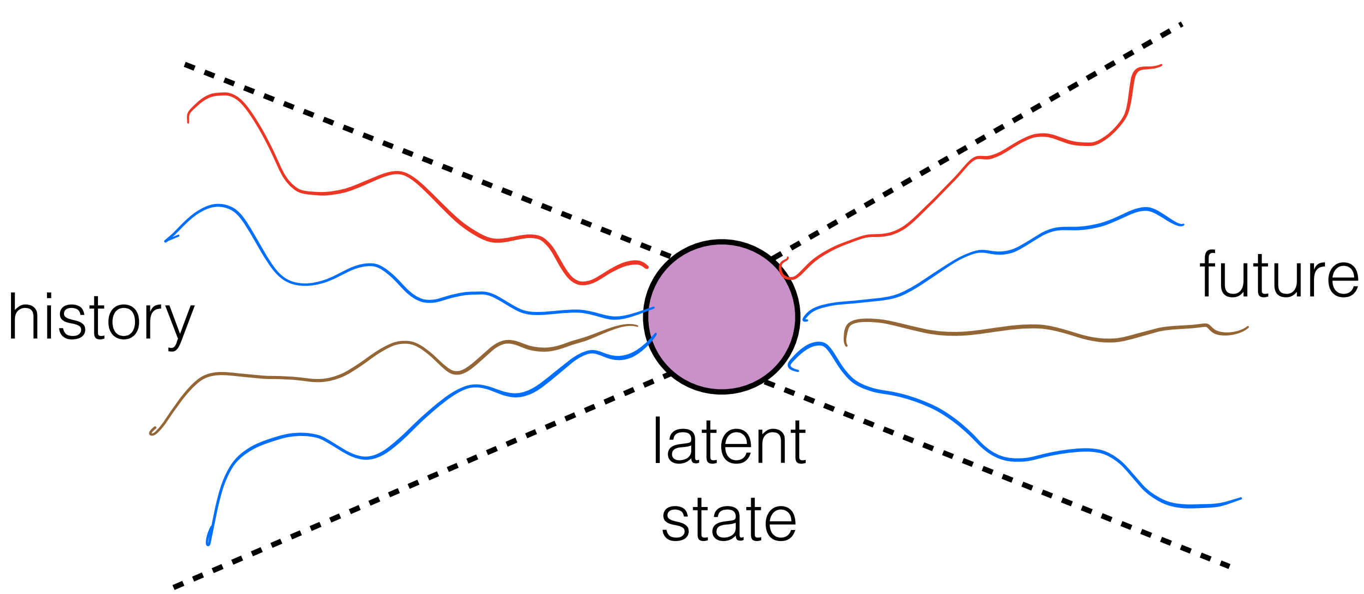


Value function?



Value function?





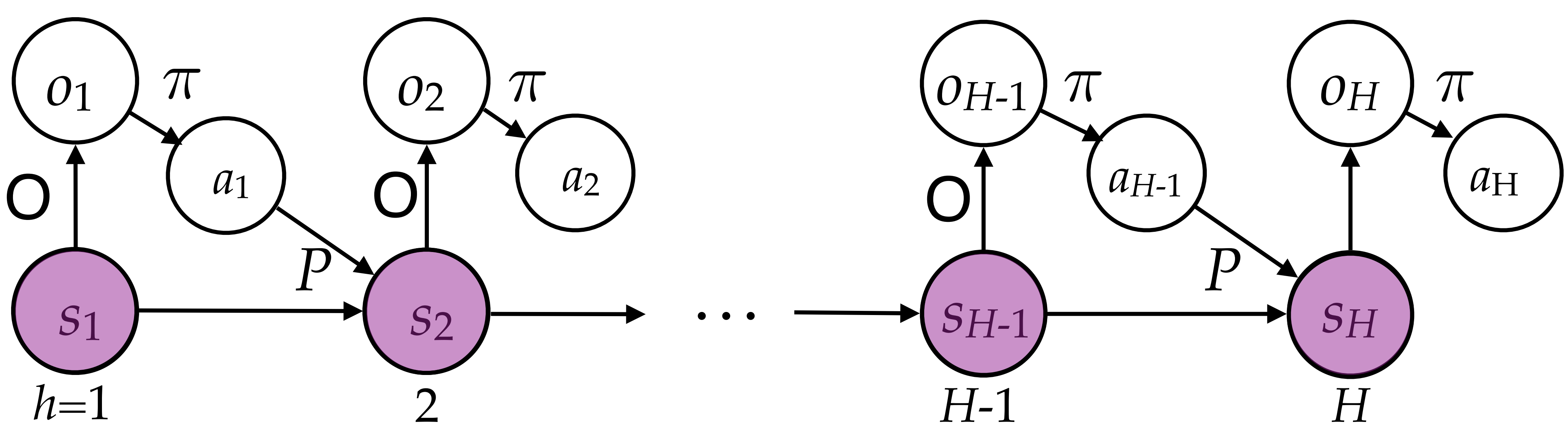
Future-Dependent Value Function

- Define: value function of latent state

$$V_{\mathcal{S}}^{\pi}(s_h) := \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r_{h'} | s_h] \in [0, H]$$

  - Problem: s_h is latent — can't even use this function!

- Solution: $V_{\mathcal{F}}^{\pi}$ as proxy of $V_{\mathcal{S}}^{\pi}$, using *future* as input!

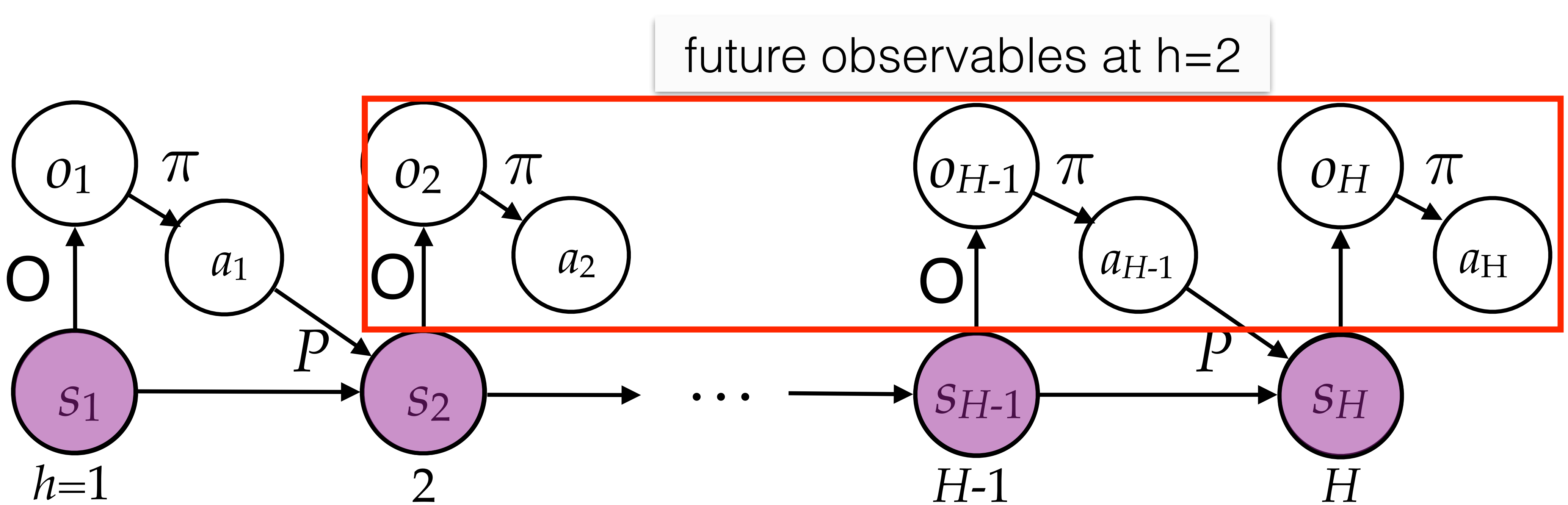


Future-Dependent Value Function

- Define: value function of latent state

$$V_{\mathcal{S}}^{\pi}(s_h) := \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r_{h'} | s_h] \in [0, H]$$

  - Problem: s_h is latent — can't even use this function!
- Solution: $V_{\mathcal{F}}^{\pi}$ as proxy of $V_{\mathcal{S}}^{\pi}$, using *future* as input!
  - $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)|s_h] = V_{\mathcal{S}}^{\pi}(s_h)$

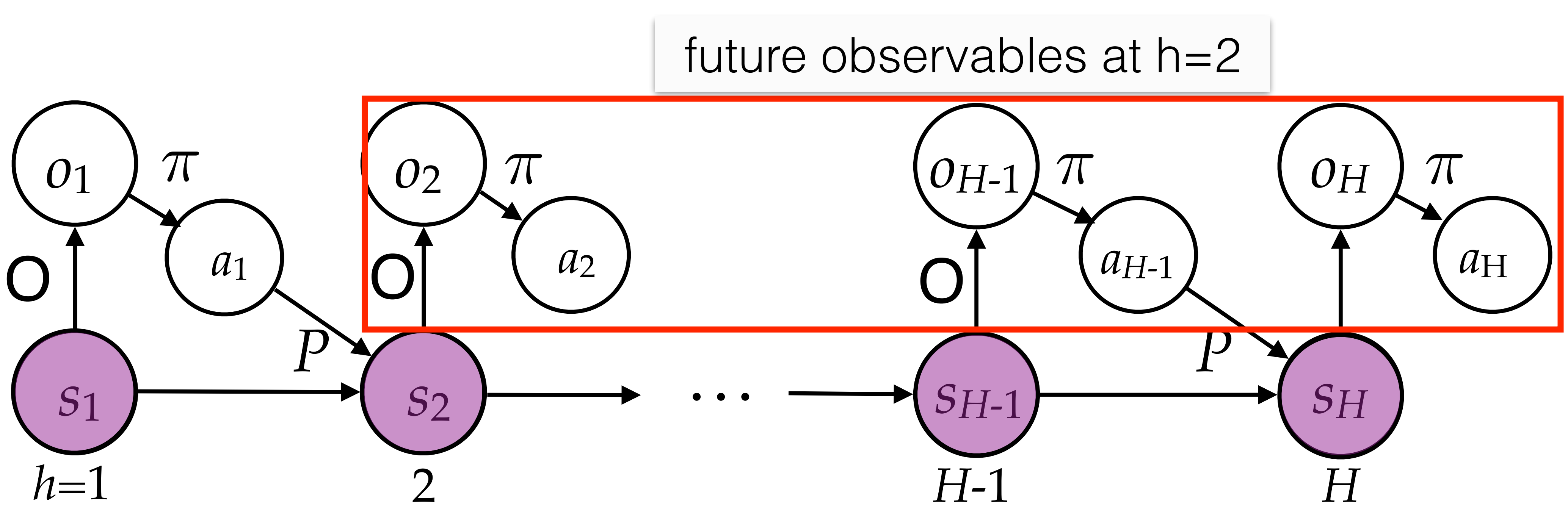


Future-Dependent Value Function

- Define: value function of latent state

$$V_{\mathcal{S}}^{\pi}(s_h) := \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r_{h'} | s_h] \in [0, H]$$

  - Problem: s_h is latent — can't even use this function!
- Solution: $V_{\mathcal{F}}^{\pi}$ as proxy of $V_{\mathcal{S}}^{\pi}$, using *future* as input!
  - $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)|s_h] = V_{\mathcal{S}}^{\pi}(s_h)$
  - Long history for using future (prediction) as state
    - OOM, SMA, PSR, etc - see Thon & Jaeger'15

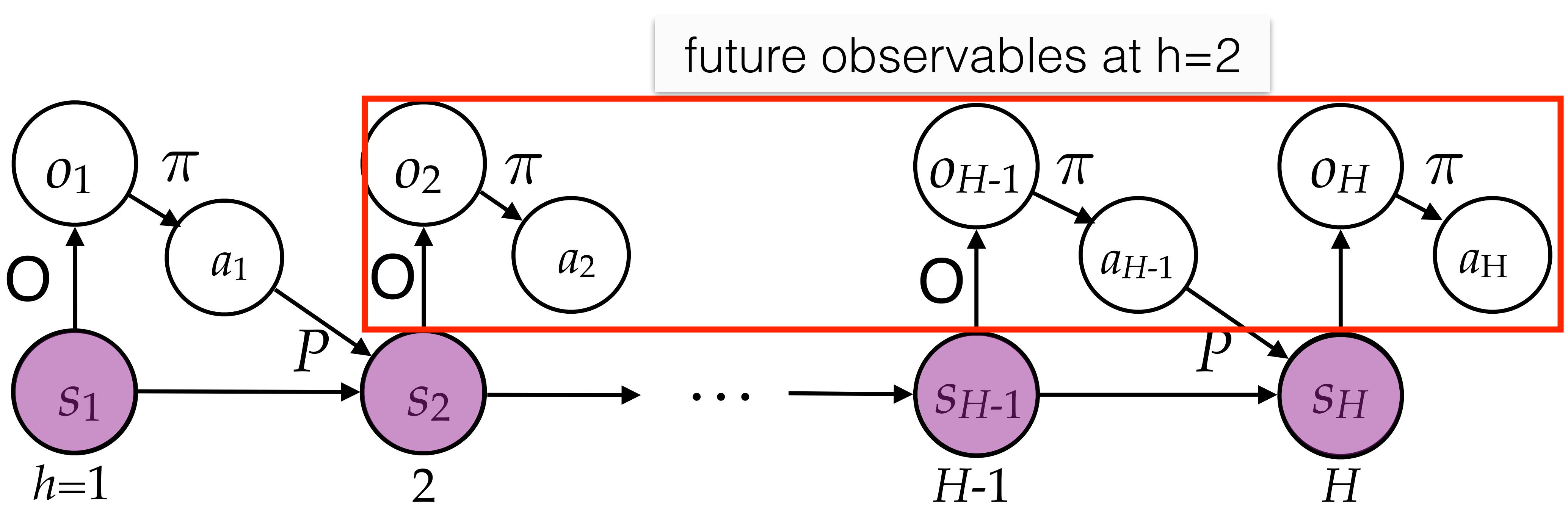


Future-Dependent Value Function

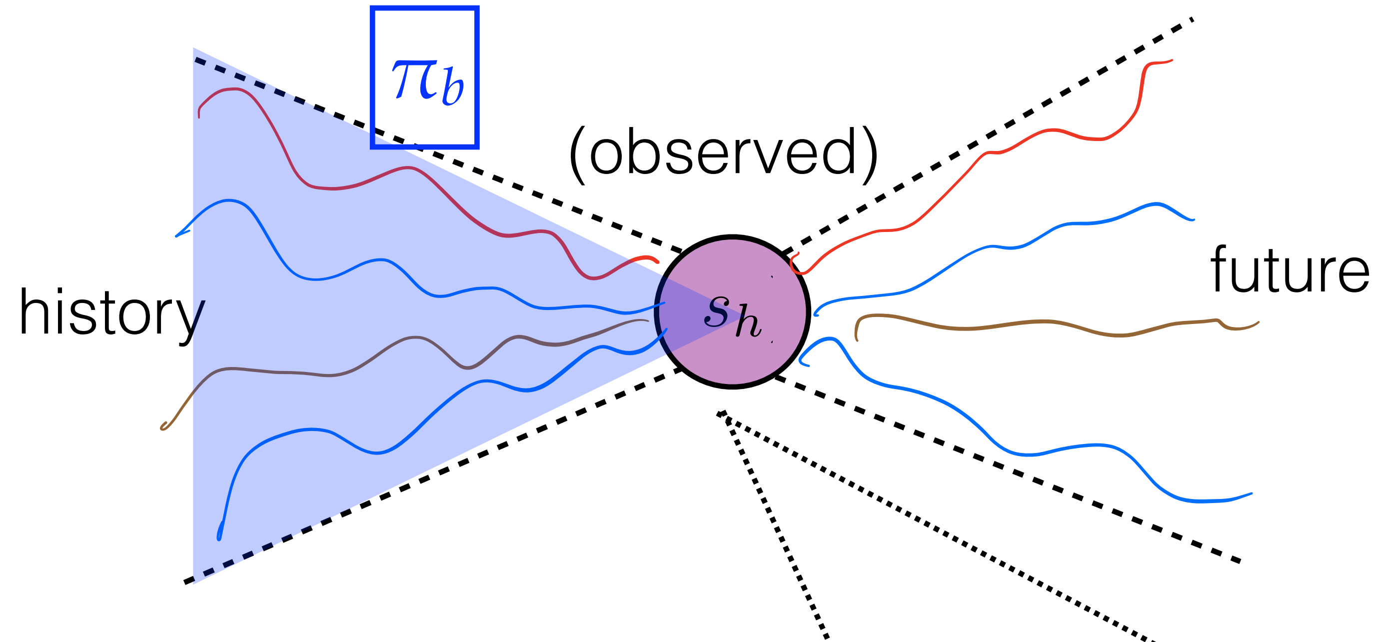
- Define: value function of latent state

$$V_{\mathcal{S}}^{\pi}(s_h) := \mathbb{E}_{\pi}[\sum_{h'=h}^{H} r_{h'} | s_h] \in [0, H]$$

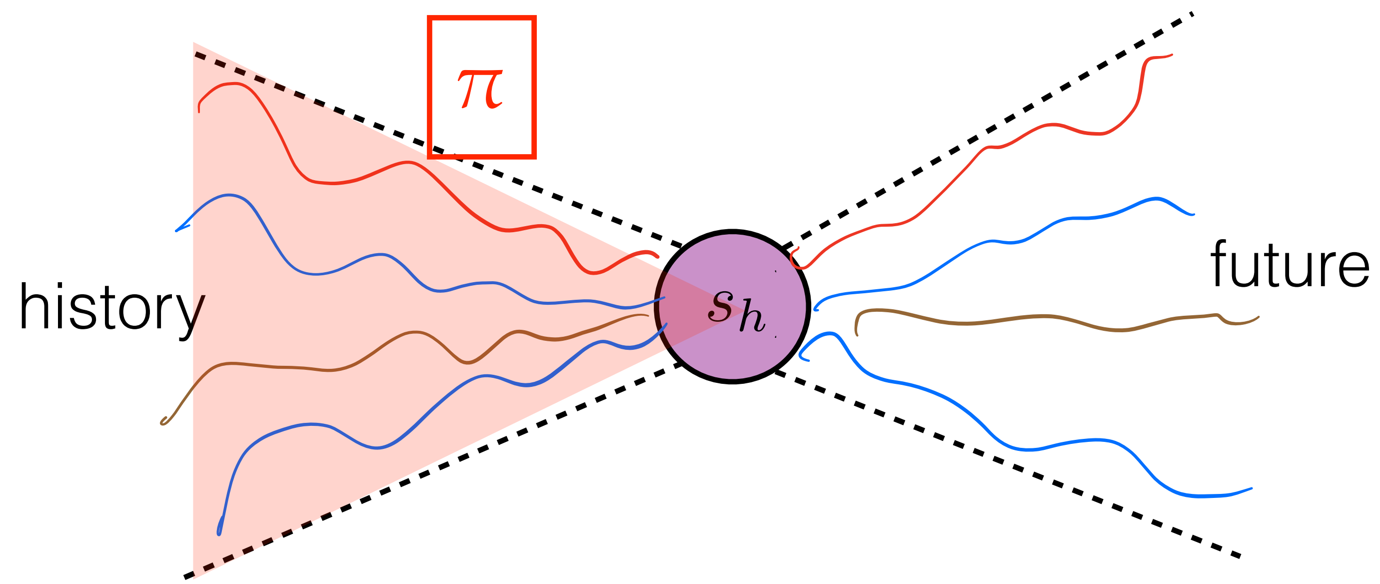
  - Problem: s_h is latent — can't even use this function!
- Solution: $V_{\mathcal{F}}^{\pi}$ as proxy of $V_{\mathcal{S}}^{\pi}$, using *future* as input!
  - $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h) | s_h] = V_{\mathcal{S}}^{\pi}(s_h)$
  - Long history for using future (prediction) as state
    - OOM, SMA, PSR, etc - see Thon & Jaeger'15
    - Distribution of future observables is low-rank (≤ |S|)



Markov case

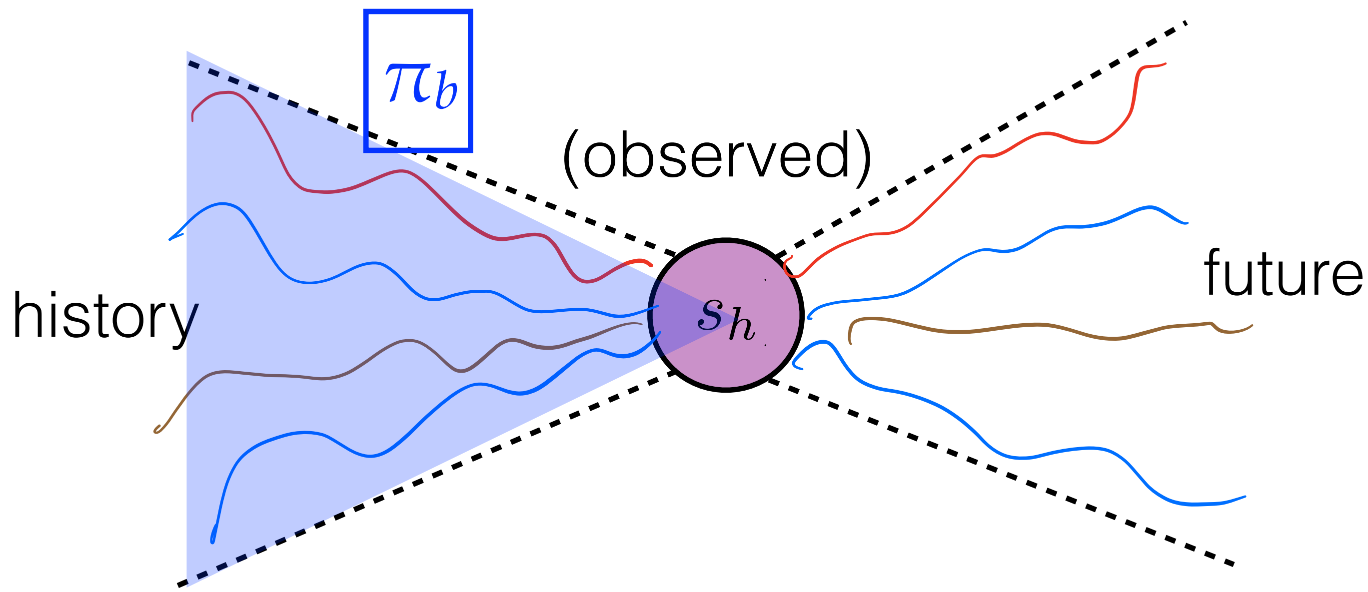


$$\mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$

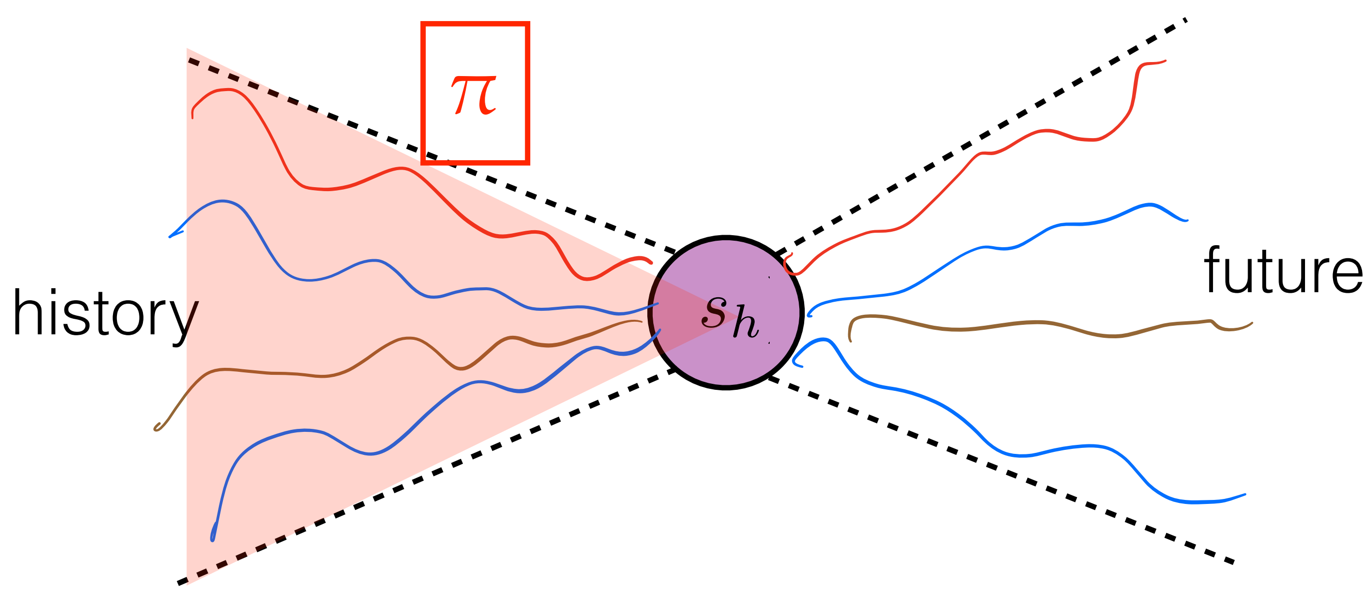


$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

Markov case

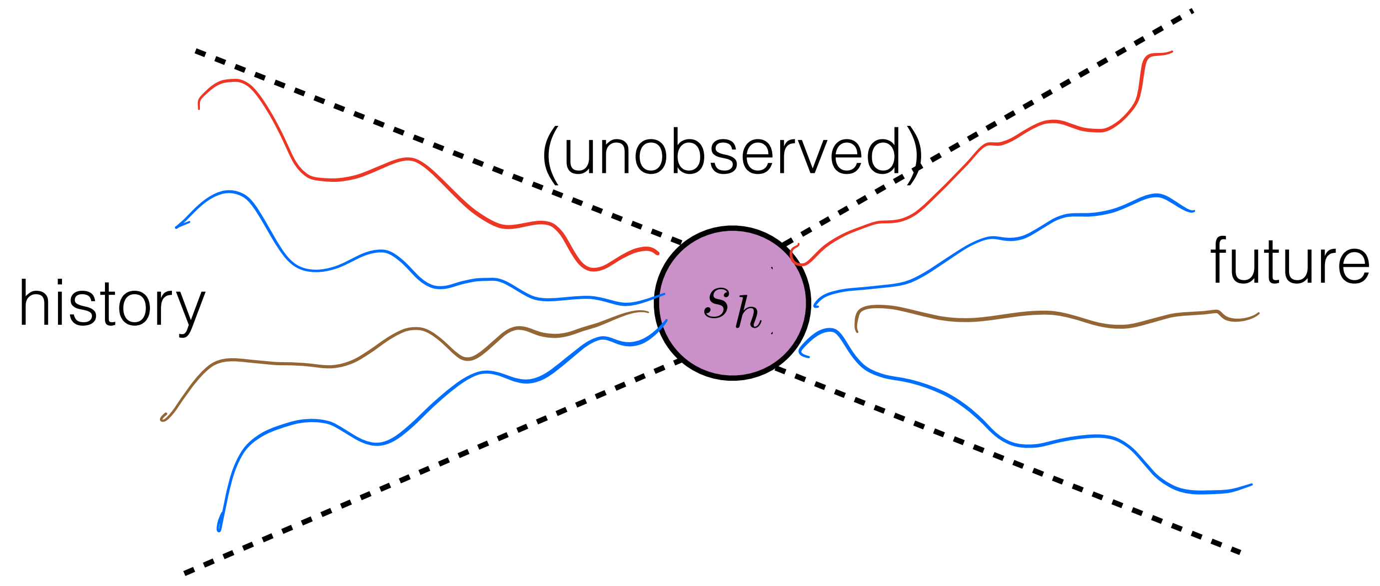


$$\mathbb{E}_{\pi_b}[(V(s_h) - (\mathcal{T}^{\pi} V)(s_h))^2]$$

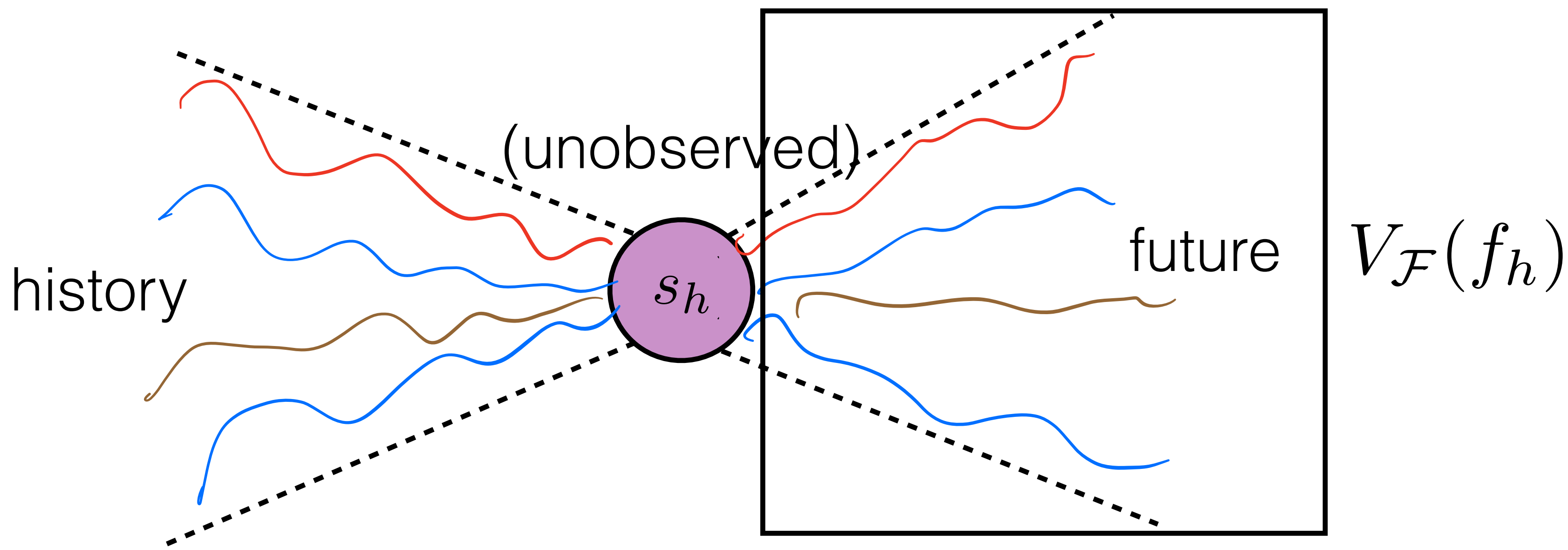


$$\mathbb{E}[V(s_1)] - J(\pi) = \sum_{h=1}^{H} \mathbb{E}_{\pi}[V(s_h) - (\mathcal{T}^{\pi} V)(s_h)]$$

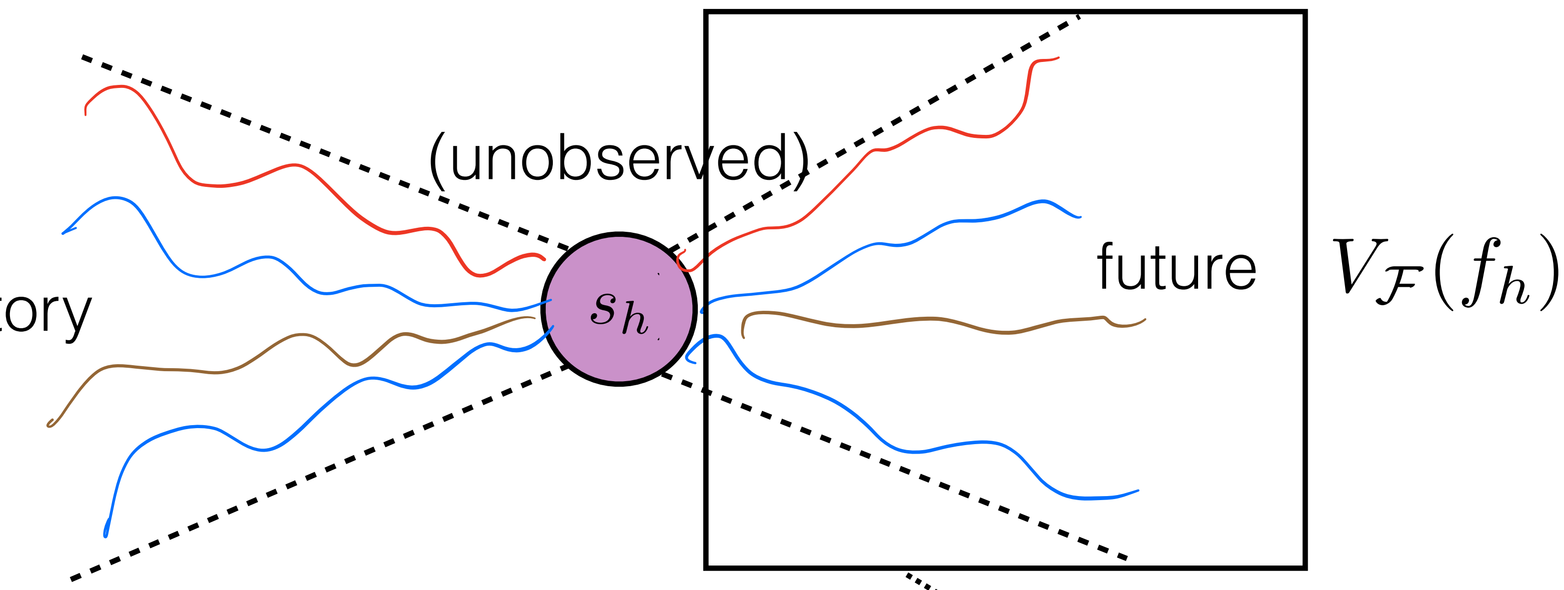
POMDP case



POMDP case



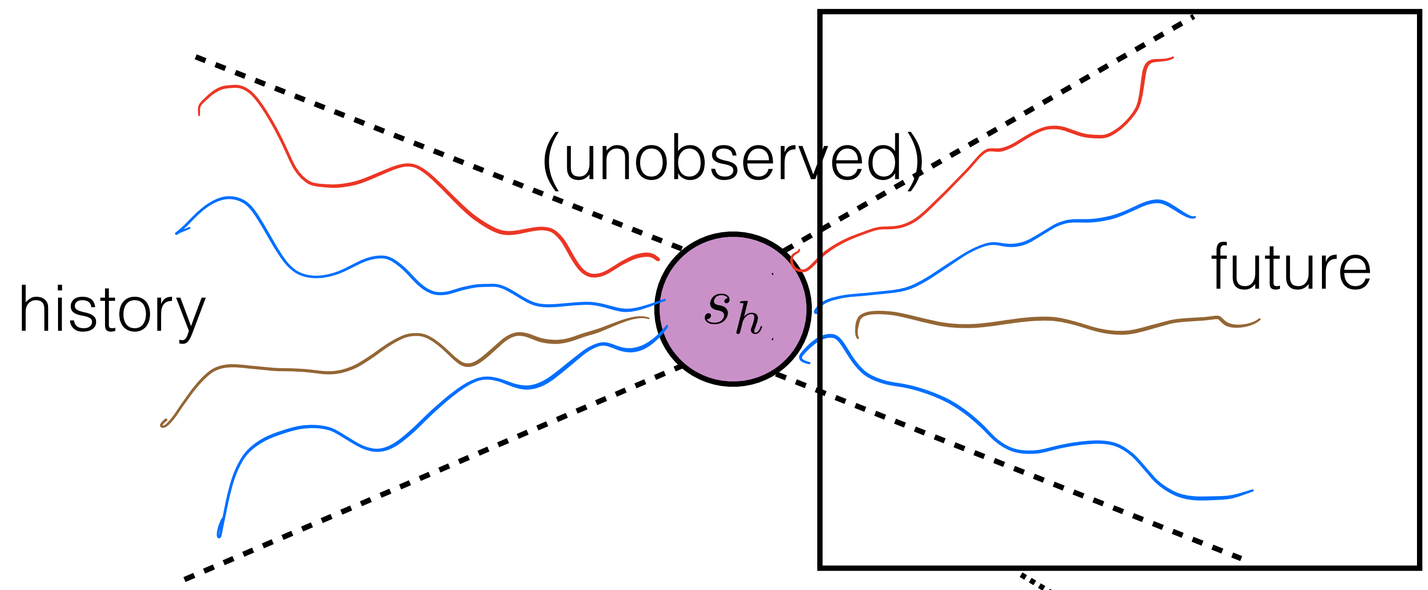
POMDP case



Ideal loss:

$$\mathbb{E}_{s_h\sim\pi_b}\left[\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h)-r_h-V_{\mathcal{F}}(f_{h+1})}|s_h])^2\right]$$

POMDP case



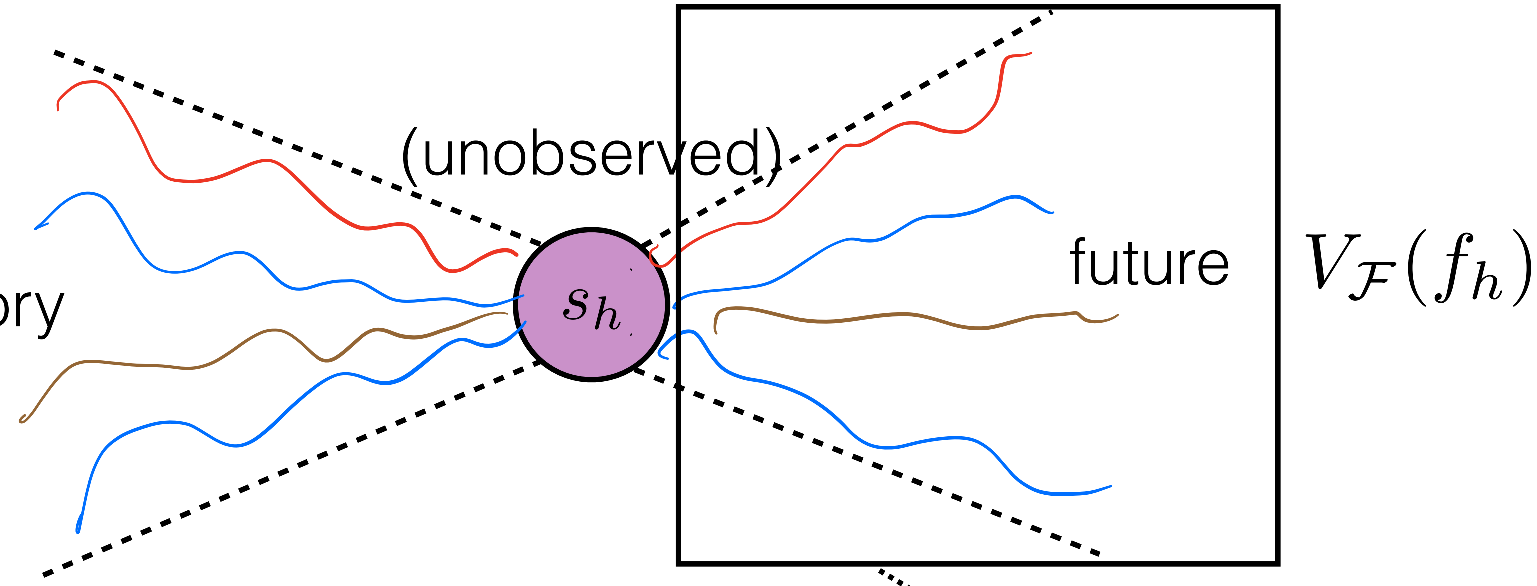
$V_{\mathcal{F}}(f_h)$

Ideal loss:

$$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h]^2\right]$$

$$\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h) | s_h] = V_{\mathcal{S}}^{\pi}(s_h)$$

POMDP case

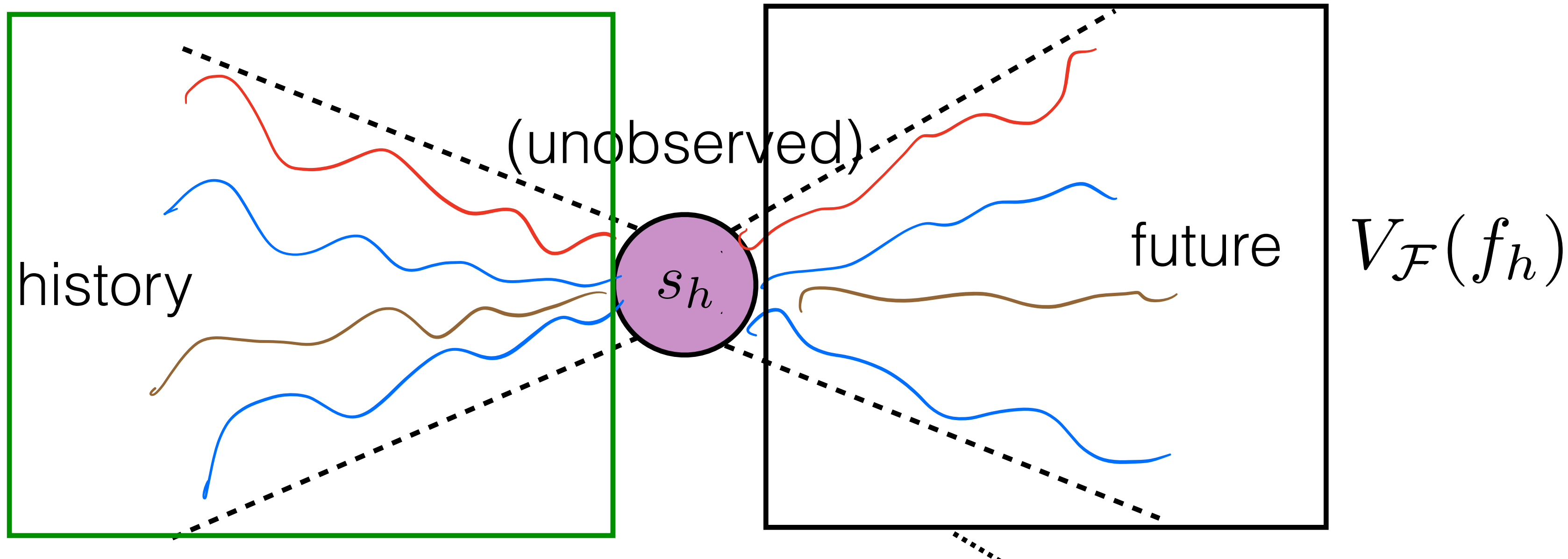


Ideal loss:

$$\mathbb{E}_{s_h \sim \pi_b} \left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h])^2 \right]$$

X

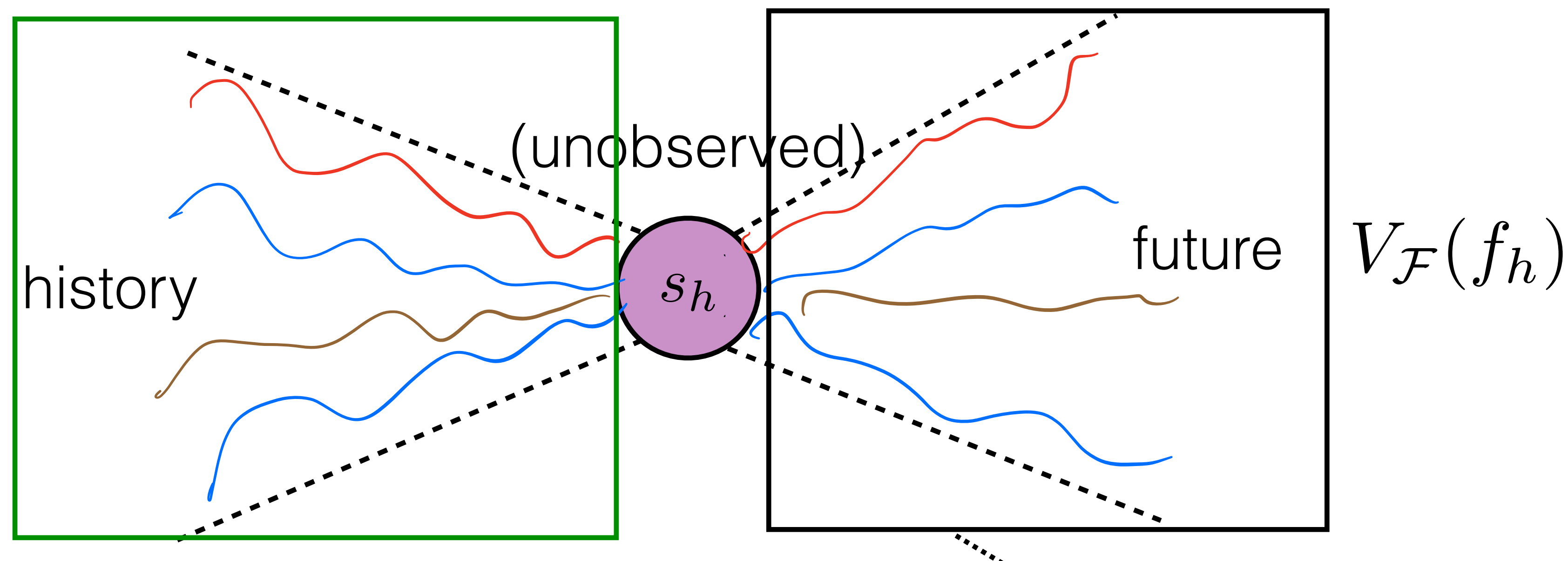
POMDP case



Ideal loss: $$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h])^2\right]$$ X

Actual loss: $$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | \tau_h]^2\right]$$

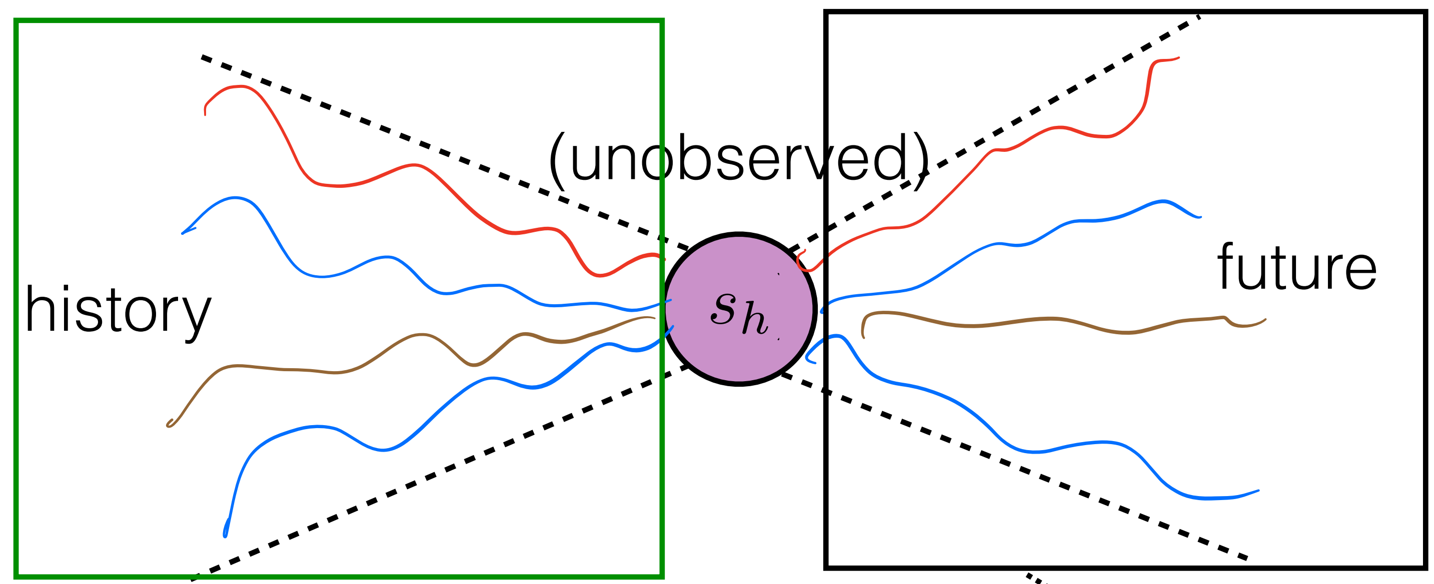
POMDP case



Ideal loss: $$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h])^2\right]$$ ✗

Actual loss: $$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | \tau_h]^2\right]$$ ✓

POMDP case



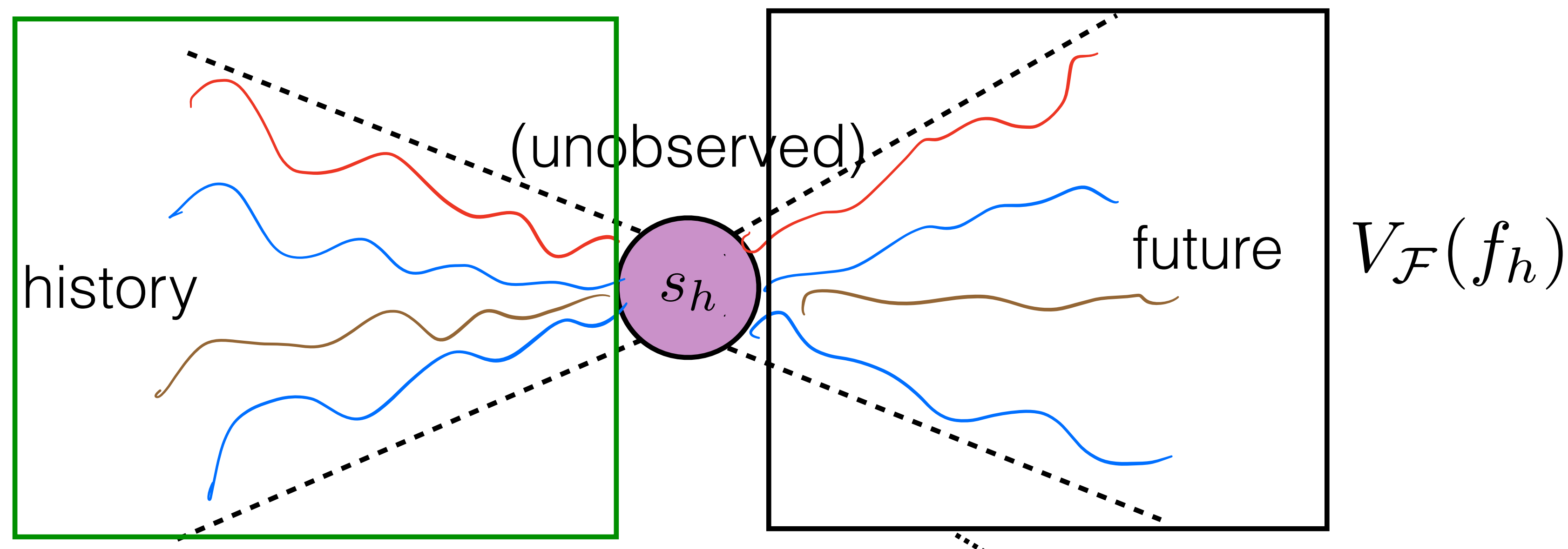
Ideal loss: $$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h])^2\right]$$ ✗

$$\mathbb{E}_{s_h \sim \pi_b}\left[\langle \mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | s_h], \mathbf{b}(\cdot | \tau_h)\rangle)^2\right]$$

||

Actual loss: $$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | \tau_h]^2\right]$$ ✓

POMDP case



Ideal loss: $\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | s_h])^2\right]$ X

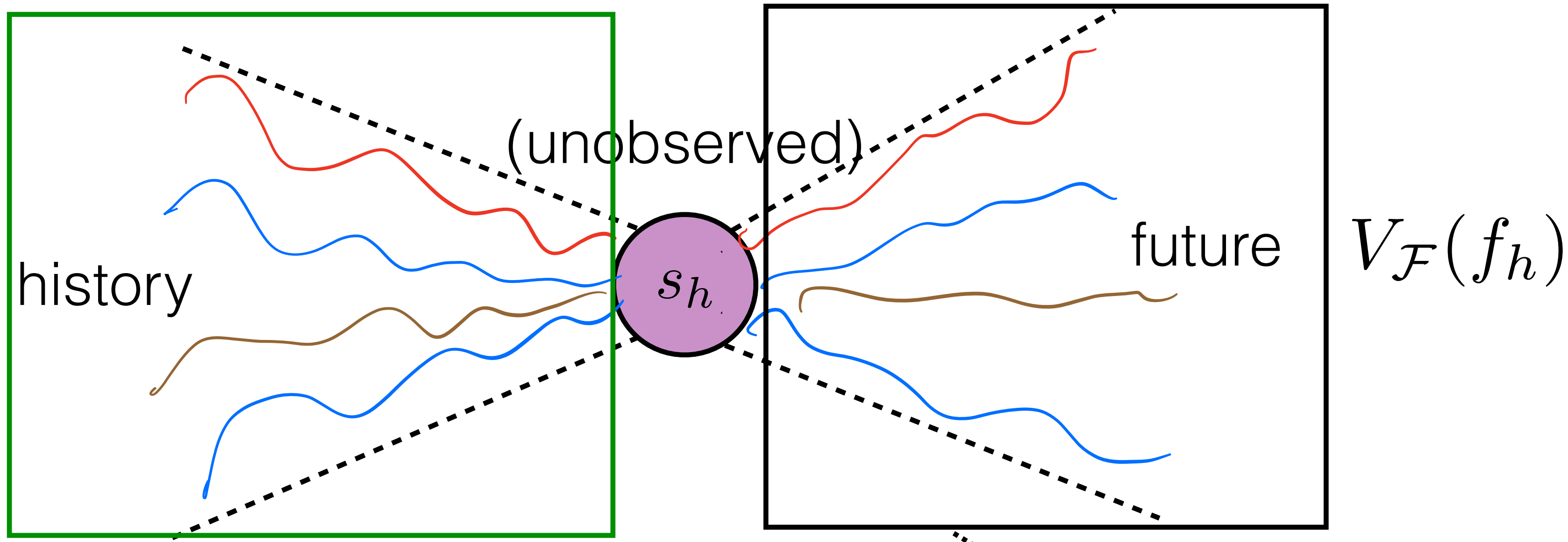
$\Delta_h V_{\mathcal{F}} = V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})$

$$\mathbb{E}_{s_h \sim \pi_b}\left[\langle \mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | s_h], \mathbf{b}(\cdot | \tau_h)\rangle)^2\right]$$

||

Actual loss: $\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | \tau_h]^2\right]$ ✓

POMDP case



Ideal loss: $$\mathbb{E}_{s_h\sim\pi_b}\left[\left(\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h]\right)^2\right]$$ ✗

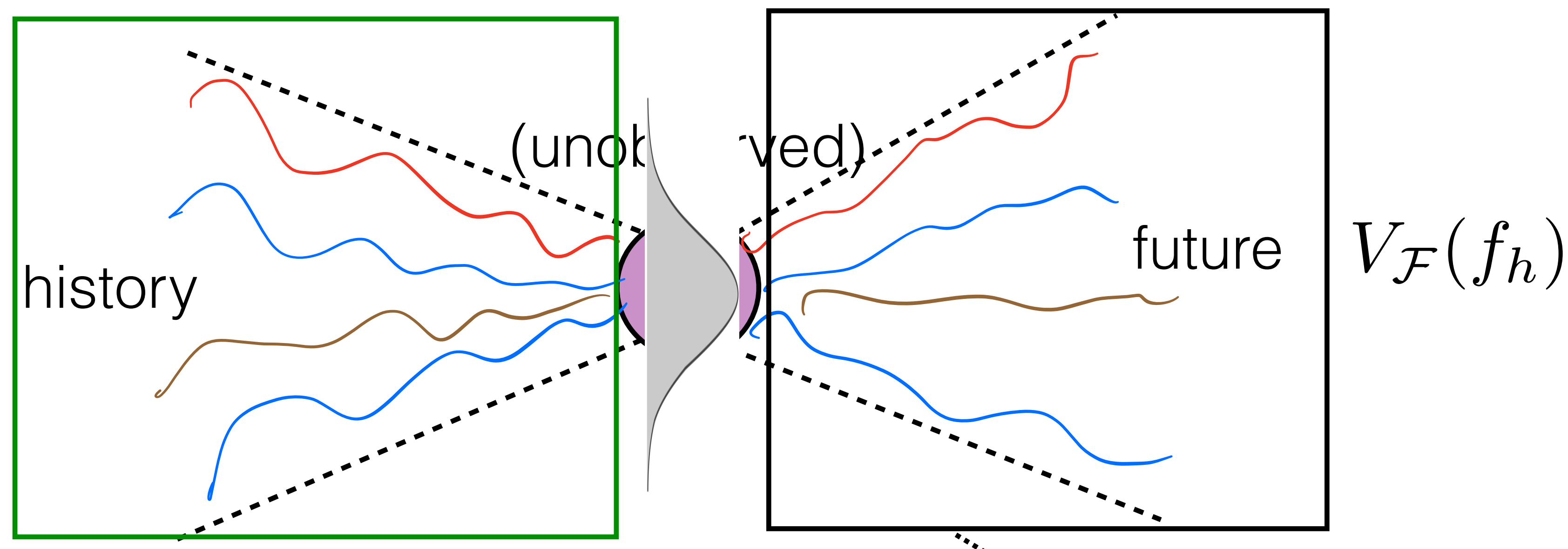
belief state $\Pr[s_h|\tau_h]$

$$\mathbb{E}_{s_h\sim\pi_b}\left[\left(\left\langle\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}} | s_h], \mathbf{b}(\cdot|\tau_h)\right\rangle\right)^2\right]$$

||

Actual loss: $$\mathbb{E}_{s_h\sim\pi_b}\left[\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}} | \tau_h]^2\right]$$ ✓

POMDP case



Ideal loss: $\mathbb{E}_{s_h \sim \pi_b} \left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h])^2 \right]$ ✗

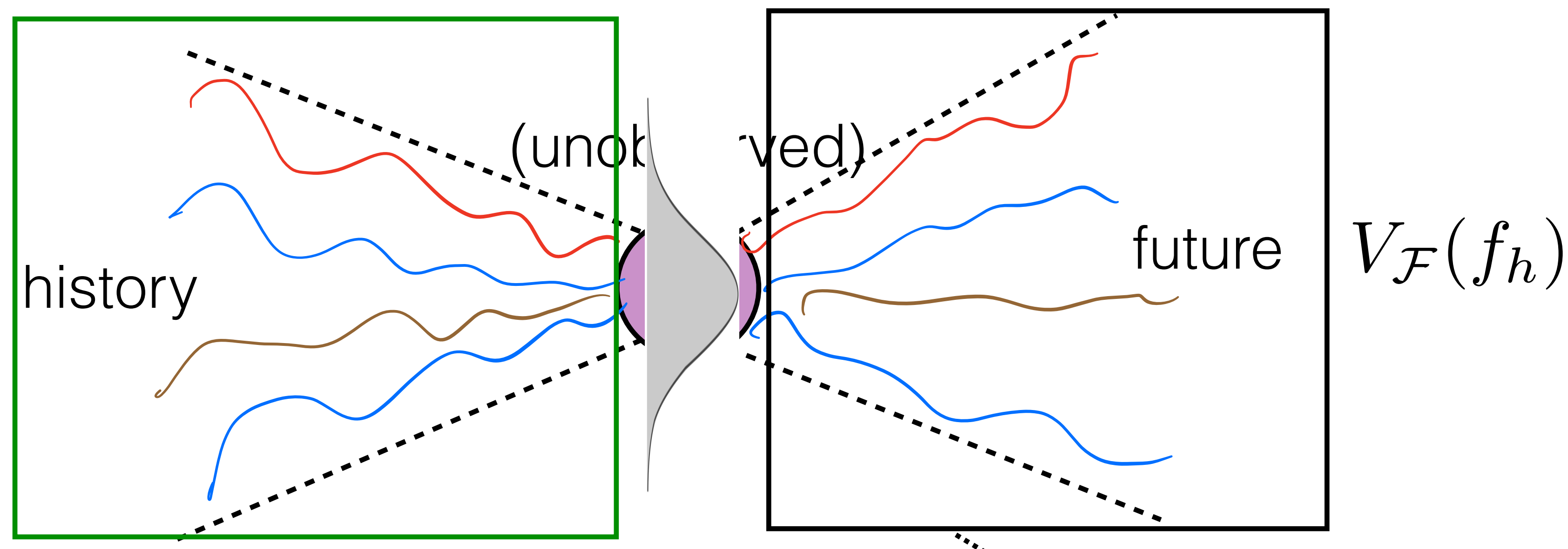
belief state $\Pr[s_h | \tau_h]$

$$\mathbb{E}_{s_h \sim \pi_b} \left[\langle \mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\Delta_h V_{\mathcal{F}} | s_h], \mathbf{b}(\cdot | \tau_h) \rangle)^2 \right]$$

||

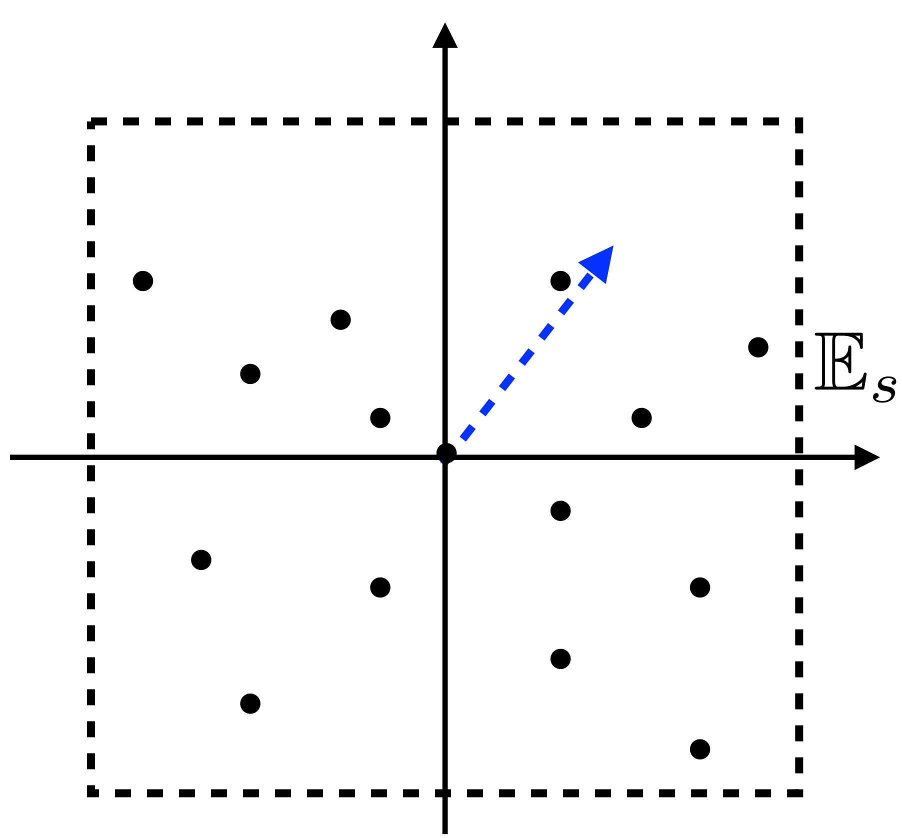
Actual loss: $\mathbb{E}_{s_h \sim \pi_b} \left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\Delta_h V_{\mathcal{F}} | \tau_h]^2 \right]$ ✓

POMDP case



Ideal loss:

$$\mathbb{E}_{s_h\sim\pi_b}\left[\mathbb{E}_{\substack{a_h\sim\pi\\ a_{h+1:H}\sim\pi_b}}[\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h]^2\right]$$ X



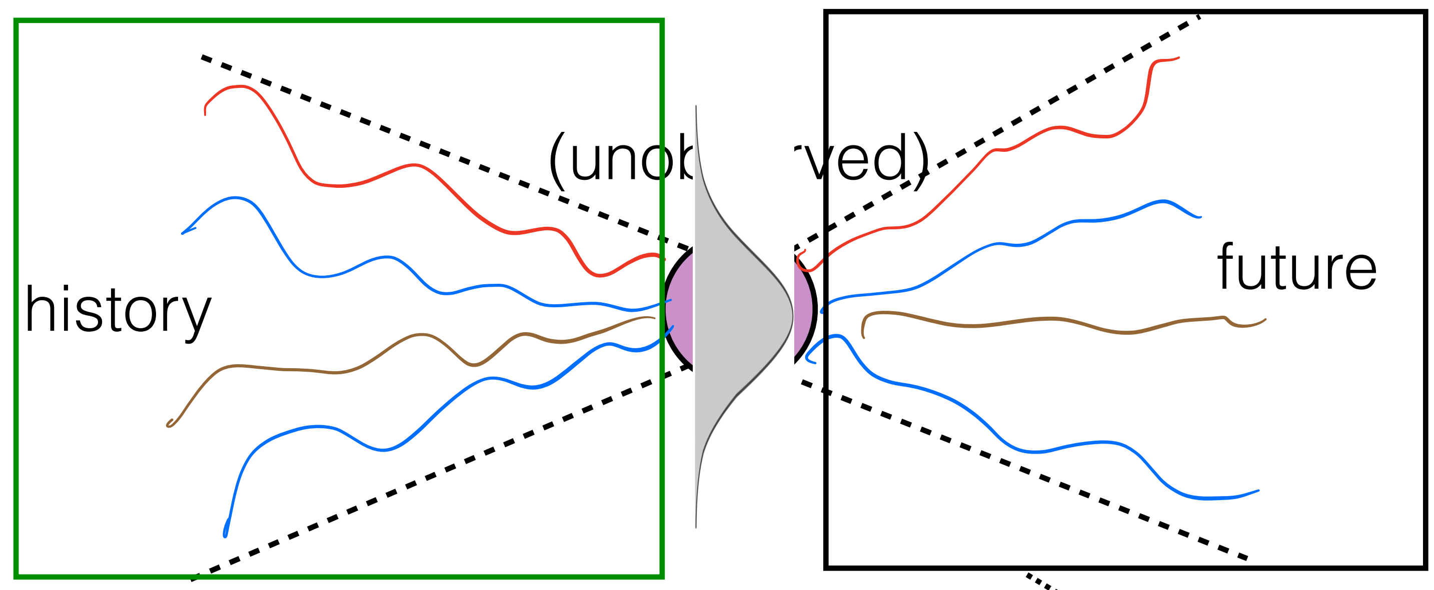
belief state $\Pr[s_h | \tau_h]$

$$\mathbb{E}_{s_h\sim\pi_b}\left[\langle \mathbb{E}_{\substack{a_h\sim\pi\\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}} | s_h], \mathbf{b}(\cdot|\tau_h)\rangle^2\right]$$

$$=$$

$$\mathbb{E}_{s_h\sim\pi_b}\left[\mathbb{E}_{\substack{a_h\sim\pi\\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}} | \tau_h]^2\right]$$ ✓

POMDP case



$V_\mathcal{F}(f_h)$

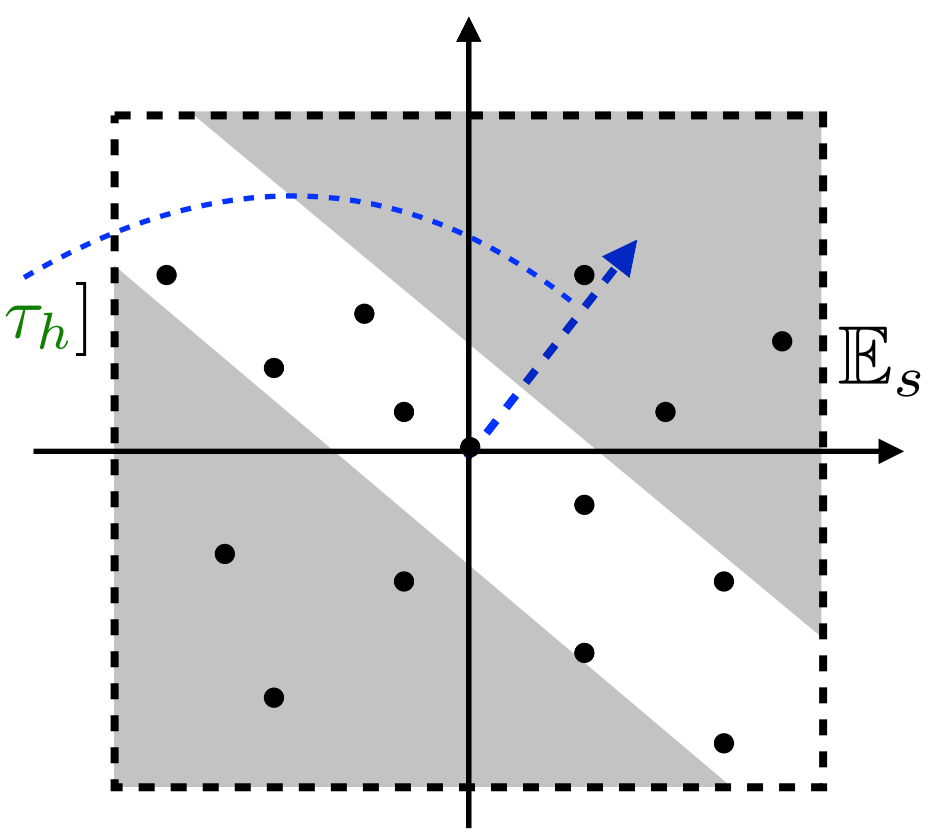
Ideal loss:

$$\mathbb{E}_{s_h\sim\pi_b}\left[\left(\underset{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}{\mathbb{E}}\Big[\underbrace{\Delta_h V_\mathcal{F}}_{V_\mathcal{F}(f_h)-r_h-V_\mathcal{F}(f_{h+1})}\Big| s_h\Big]\right)^2\right]$$

✗

belief state $\Pr[s_h|\tau_h]$

$\Pr[s_h|\tau_h]$



$S=2$

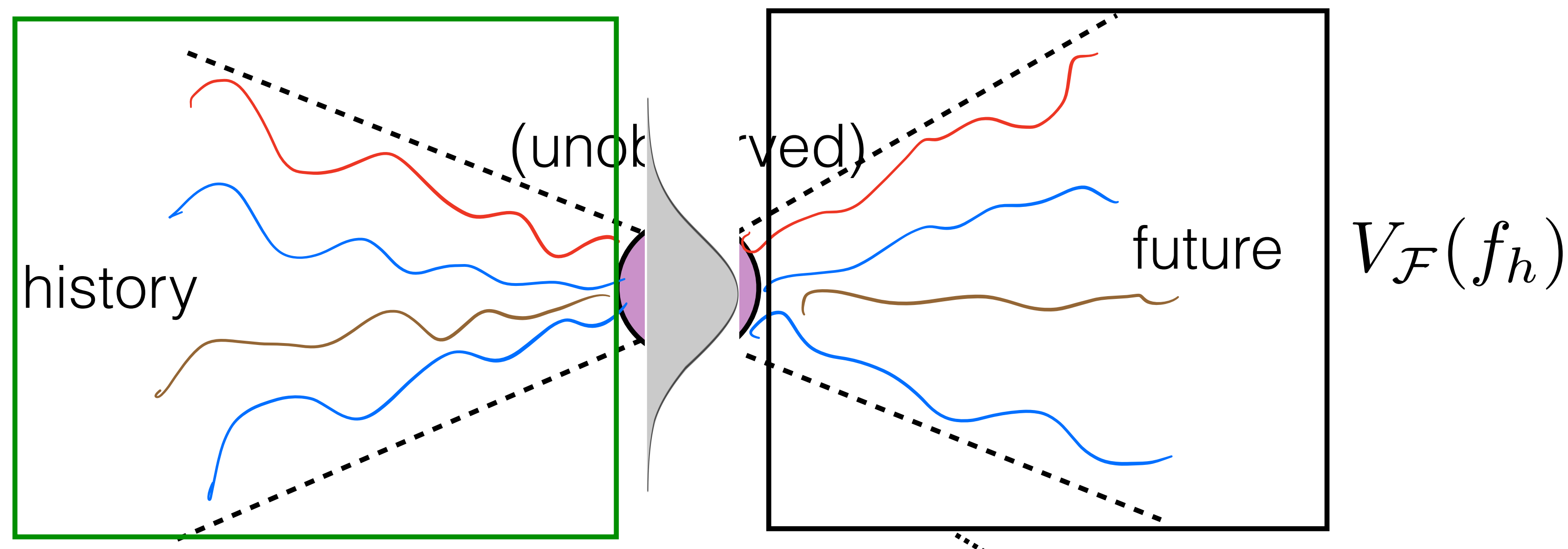
$$\mathbb{E}_{s_h\sim\pi_b}\left[\left\langle\underset{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}{\mathbb{E}}[\Delta_h V_\mathcal{F}|s_h], \mathbf{b}(\cdot|\tau_h)\right\rangle^2\right]$$

$$=$$

$$\mathbb{E}_{s_h\sim\pi_b}\left[\underset{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}{\mathbb{E}}[\Delta_h V_\mathcal{F}|\tau_h]^2\right]$$

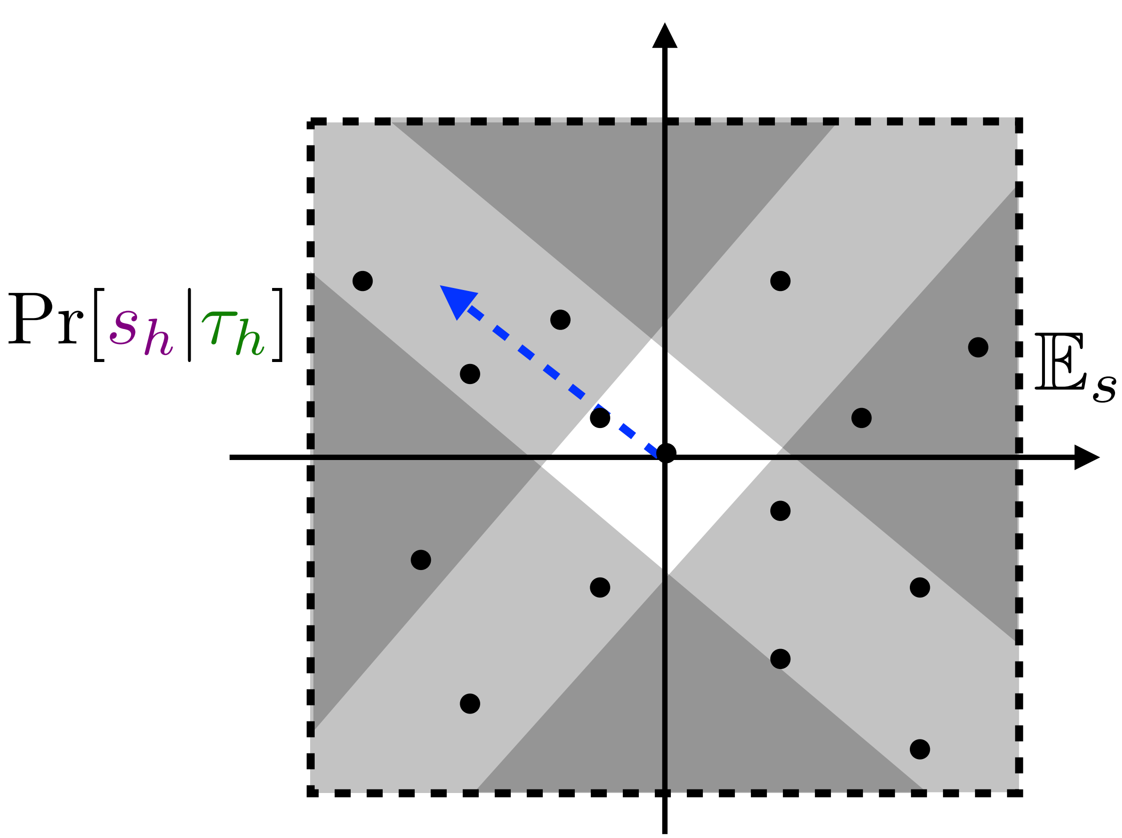
✓

POMDP case



Ideal loss:

$$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h]^2\right]$$ ✗



$S = 2$

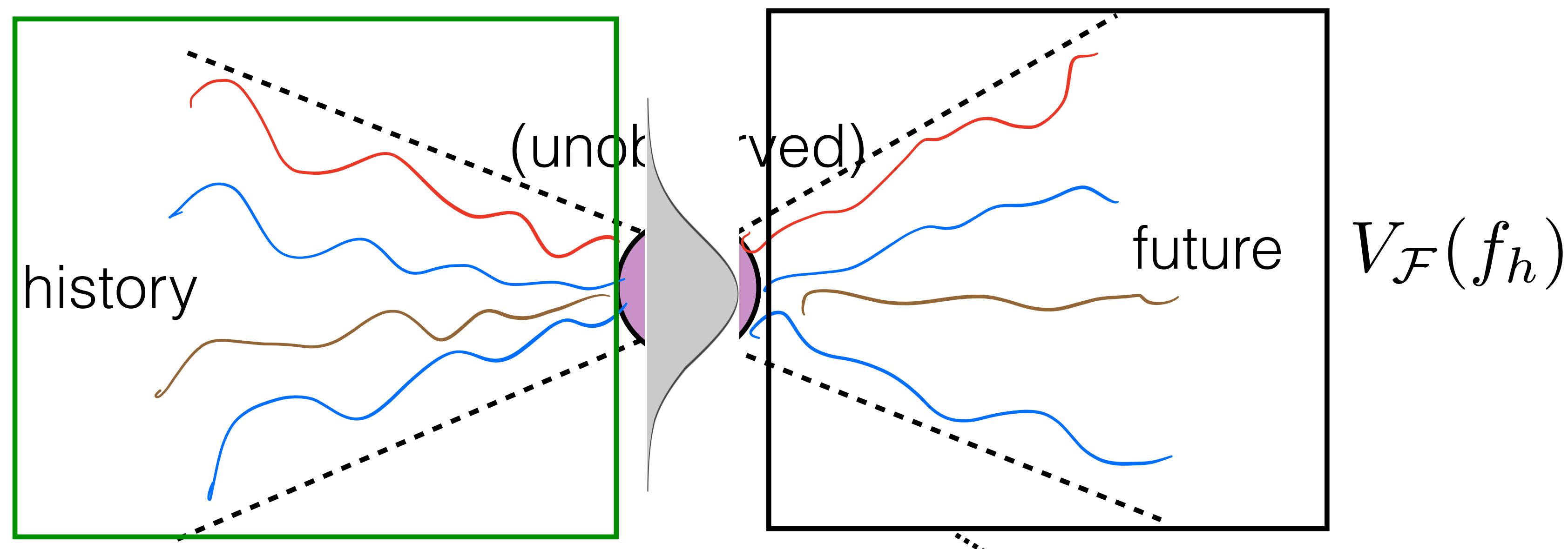
belief state $\Pr[s_h | \tau_h]$

$$\mathbb{E}_{s_h \sim \pi_b}\left[\langle \mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | s_h], \mathbf{b}(\cdot | \tau_h)\rangle^2\right]$$

$$\|$$

$$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | \tau_h]^2\right]$$ ✓

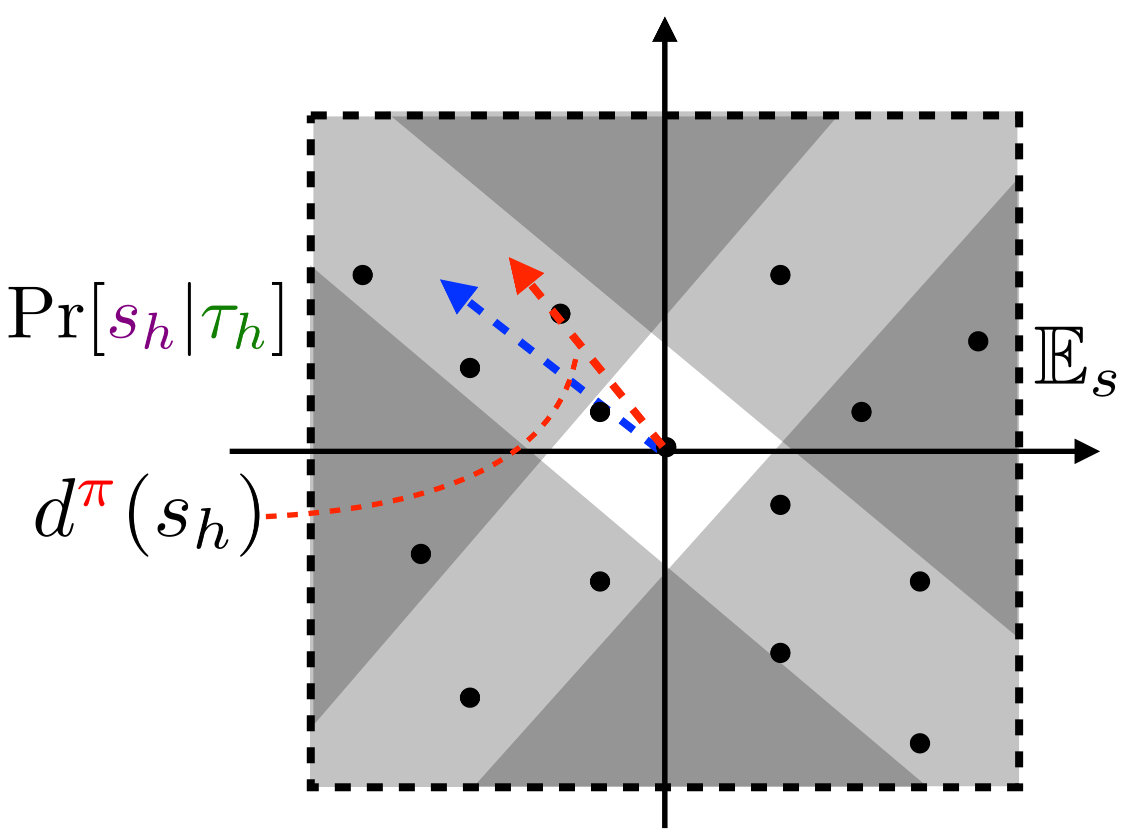
POMDP case



Ideal loss:

$$\mathbb{E}_{s_h \sim \pi_b}\left[\left(\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\underbrace{\Delta_h V_{\mathcal{F}}}_{V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})} | s_h]\right)^2\right]$$

✗



$S = 2$

belief state $\Pr[s_h|\tau_h]$

$$\mathbb{E}_{s_h \sim \pi_b}\left[\langle \mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | s_h], \mathbf{b}(\cdot|\tau_h)\rangle)^2\right]$$

$$\|$$

$$\mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} | \tau_h]^2\right]$$

✓

Guarantee under belief coverage

Guarantee under belief coverage

Theorem (Informal): Assume

$$(d_h^{\pi})^\top \mathbb{E}_{\tau_h \sim \pi_b}[\mathbf{b}(\tau_h)\mathbf{b}(\tau_h)^\top]^{-1} d_h^{\pi} \leq C_{\mathcal{H}}$$

Guarantee under belief coverage

Theorem (Informal): Assume

$$(d_h^{\pi})^\top \mathbb{E}_{\tau_h \sim \pi_b}[\mathbf{b}(\tau_h)\mathbf{b}(\tau_h)^\top]^{-1} d_h^{\pi} \leq C_{\mathcal{H}}$$

and standard representation assumptions (realizability & Bellman-completeness), the sample complexity of OPE is poly in

- Coverage parameters: $C_{\mathcal{H}}$ and $C_{\mathcal{A}} := \max_{s_h, a_h} \frac{\pi(a_h|s_h)}{\pi_b(a_h|s_h)}$
- Ranges & complexities of function classes (e.g., that of $\mathcal{V}$)

Guarantee under belief coverage

Theorem (Informal): Assume

$$(d_h^{\pi})^{\top}\mathbb{E}_{\tau_h \sim \pi_b}[\mathbf{b}(\tau_h)\mathbf{b}(\tau_h)^{\top}]^{-1} d_h^{\pi} \leq C_{\mathcal{H}}$$

and standard representation assumptions (realizability & Bellman-completeness), the sample complexity of OPE is poly in

- Coverage parameters: $C_{\mathcal{H}}$ and $C_{\mathcal{A}} := \max_{s_h, a_h} \frac{\pi(a_h|s_h)}{\pi_b(a_h|s_h)}$
- Ranges & complexities of function classes (e.g., that of $\mathcal{V}$)

- Similar to $\mathbb{E}_{\pi}[\phi]^{\top}\mathbb{E}_{\pi_b}[\phi\phi^{\top}]^{-1}\mathbb{E}_{\pi}[\phi]$

Guarantee under belief coverage

Theorem (Informal): Assume

$$(d_h^{\pi})^\top \mathbb{E}_{\tau_h \sim \pi_b}[\mathbf{b}(\tau_h)\mathbf{b}(\tau_h)^\top]^{-1} d_h^{\pi} \leq C_{\mathcal{H}}$$

and standard representation assumptions (realizability & Bellman-completeness), the sample complexity of OPE is poly in

- Coverage parameters: $C_{\mathcal{H}}$ and $C_{\mathcal{A}} := \max_{s_h, a_h} \frac{\pi(a_h|s_h)}{\pi_b(a_h|s_h)}$
- Ranges & complexities of function classes (e.g., that of $\mathcal{V}$)

- Similar to $\mathbb{E}_{\pi}[\phi]^\top \mathbb{E}_{\pi_b}[\phi\phi^\top]^{-1}\mathbb{E}_{\pi}[\phi]$
- $\pi = \pi_b$: $C_{\mathcal{H}}$ =1

Guarantee under belief coverage

Theorem (Informal): Assume

$$(d_h^{\pi})^\top \mathbb{E}_{\tau_h \sim \pi_b}[\mathbf{b}(\tau_h)\mathbf{b}(\tau_h)^\top]^{-1} d_h^{\pi} \le C_{\mathcal{H}}$$

and standard representation assumptions (realizability & Bellman-completeness), the sample complexity of OPE is poly in

- Coverage parameters: $C_{\mathcal{H}}$ and $C_{\mathcal{A}} := \max_{s_h, a_h} \frac{\pi(a_h|s_h)}{\pi_b(a_h|s_h)}$
- Ranges & complexities of function classes (e.g., that of $\mathcal{V}$)

- Similar to $\mathbb{E}_{\pi}[\phi]^\top \mathbb{E}_{\pi_b}[\phi\phi^\top]^{-1}\mathbb{E}_{\pi}[\phi]$
- $\pi = \pi_b$: $C_{\mathcal{H}}$ =1
- 1-hot $\mathbf{b}(\tau_h)$: $\mathbb{E}_{\pi_b}\left[(d_h^{\pi}/d_h^{\pi_b})^2\right]$

Guarantee under belief coverage

Theorem (Informal): Assume

$$(d_h^{\pi})^\top \mathbb{E}_{\tau_h \sim \pi_b}[\mathbf{b}(\tau_h)\mathbf{b}(\tau_h)^\top]^{-1} d_h^{\pi} \leq C_{\mathcal{H}}$$

and standard representation assumptions (realizability & Bellman-completeness), the sample complexity of OPE is poly in

- Coverage parameters: $C_{\mathcal{H}}$ and $C_{\mathcal{A}} := \max_{s_h, a_h} \frac{\pi(a_h|s_h)}{\pi_b(a_h|s_h)}$
- Ranges & complexities of function classes (e.g., that of $\mathcal{V}$)

- Similar to $\mathbb{E}_{\pi}[\phi]^\top \mathbb{E}_{\pi_b}[\phi\phi^\top]^{-1}\mathbb{E}_{\pi}[\phi]$
- $\pi = \pi_b$: $C_{\mathcal{H}}$ =1
- 1-hot $\mathbf{b}(\tau_h)$: $\mathbb{E}_{\pi_b}\left[(d_h^{\pi}/d_h^{\pi_b})^2\right]$
- $\|d_h^{\pi}/d_h^{\pi_b}\|_\infty$ => $\|\mathbb{E}_{\pi_b}[\mathbf{b}(\tau_h)\mathbf{b}(\tau_h)^\top]^{-1} d_h^{\pi}\|_\infty$

Does FDVF exist…?

- $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)|s_h] = V_{\mathcal{S}}^{\pi}(s_h)$

Does FDVF exist…?

- $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)|s_h] = V_{\mathcal{S}}^{\pi}(s_h)$
- Equation has no solution (!) if
  - two latent states have the same future distribution under π_b
  - … but the have different values in $V_{\mathcal{S}}^{\pi}(s_h)$

Does FDVF exist…?

- $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)|s_h] = V_{\mathcal{S}}^{\pi}(s_h)$
- Equation has no solution (!) if
  - two latent states have the same future distribution under π_b
  - … but the have different values in $V_{\mathcal{S}}^{\pi}(s_h)$
- New condition: outcome coverage
  - Let $\Sigma_{\mathcal{F}}$ be the confusion matrix of predicting s_h from future

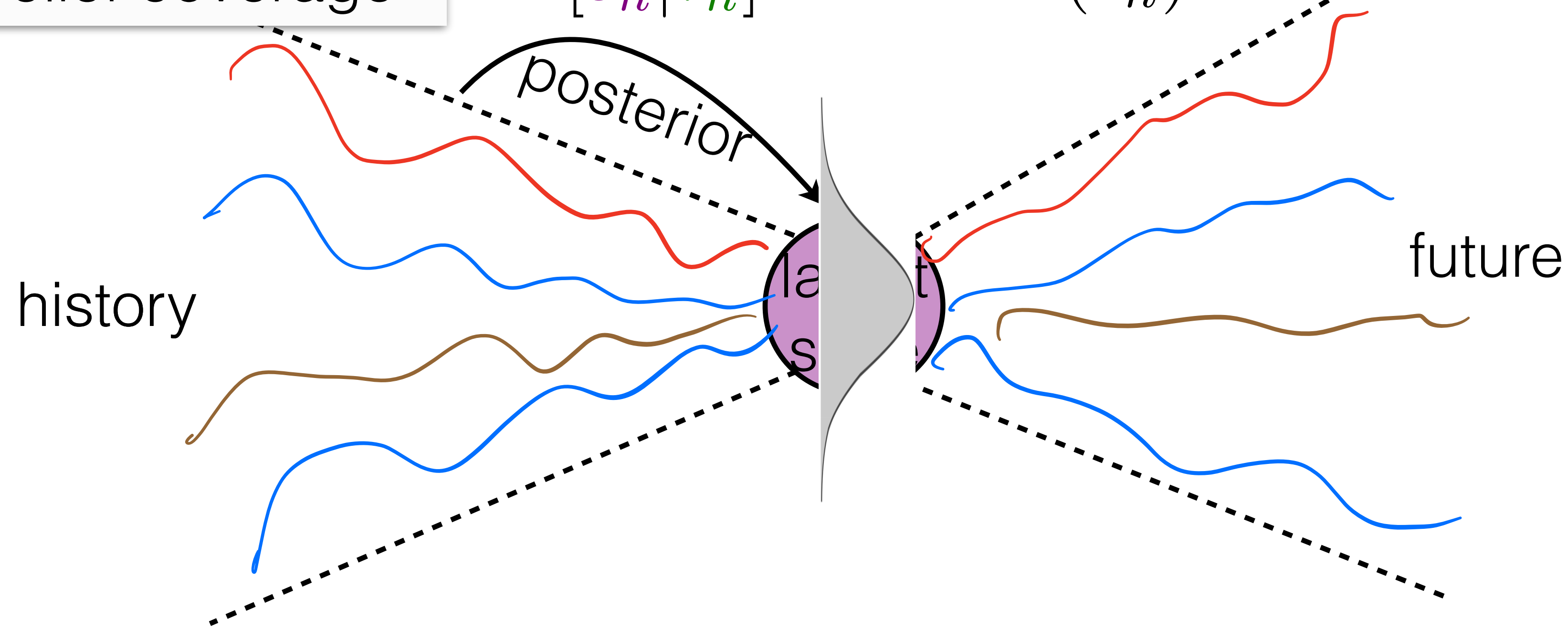
Does FDVF exist…?

- $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)|s_h] = V_{\mathcal{S}}^{\pi}(s_h)$
- Equation has no solution (!) if
  - two latent states have the same future distribution under π_b
  - … but the have different values in $V_{\mathcal{S}}^{\pi}(s_h)$
- New condition: outcome coverage
  - Let $\Sigma_{\mathcal{F}}$ be the confusion matrix of predicting s_h from future
  - Finite & bounded $\|\Sigma_{\mathcal{F}}^{-1} V_{\mathcal{S}}^{\pi}\|_{\infty}$

Does FDVF exist…?

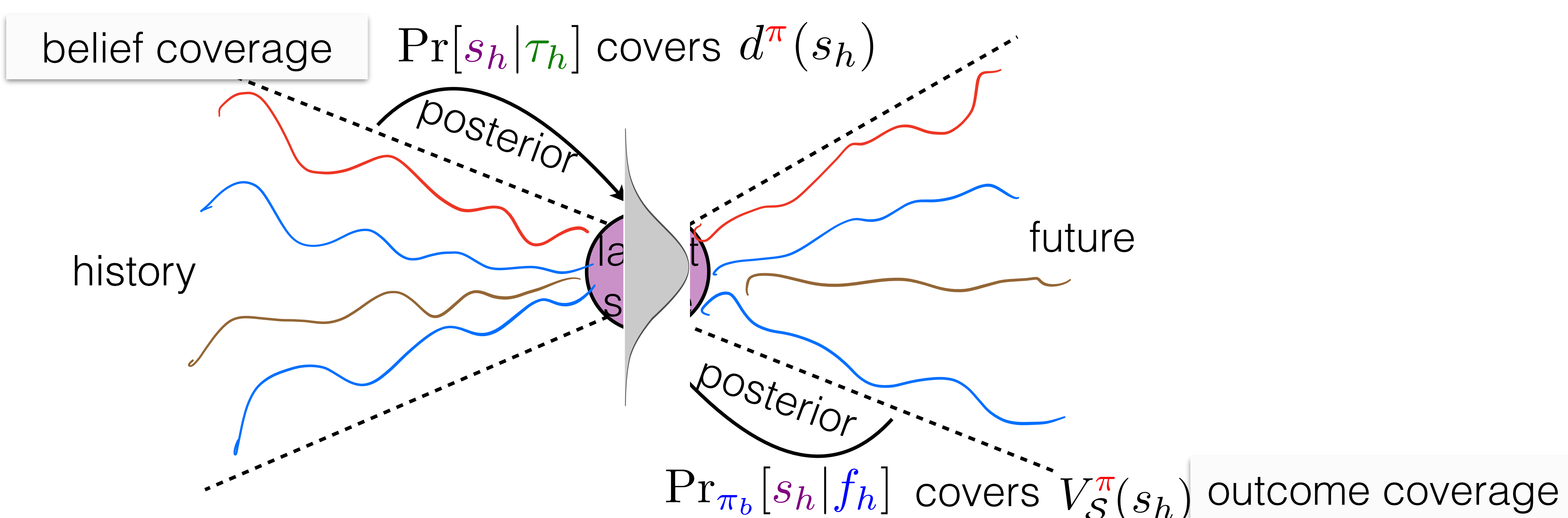
- $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)|s_h] = V_{\mathcal{S}}^{\pi}(s_h)$
- Equation has no solution (!) if
  - two latent states have the same future distribution under π_b
  - … but the have different values in $V_{\mathcal{S}}^{\pi}(s_h)$
- New condition: outcome coverage
  - Let $\Sigma_{\mathcal{F}}$ be the confusion matrix of predicting s_h from future
  - Finite & bounded $\|\Sigma_{\mathcal{F}}^{-1} V_{\mathcal{S}}^{\pi}\|_\infty$

belief coverage $\Pr[s_h|\tau_h]$ covers $d^{\pi}(s_h)$



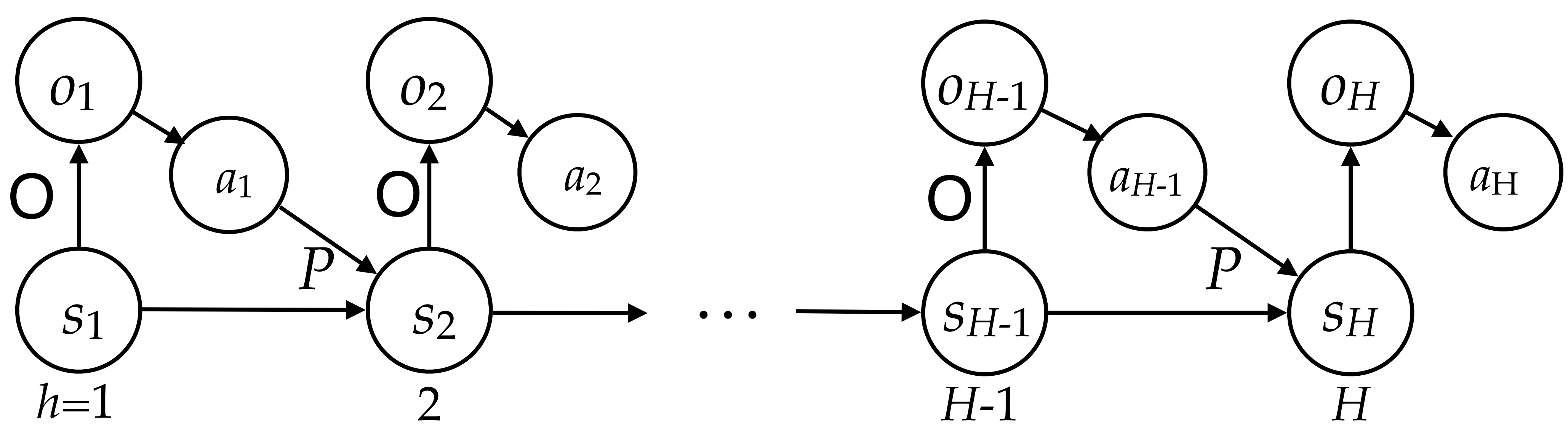
Does FDVF exist…?

- $\mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)|s_h] = V_{\mathcal{S}}^{\pi}(s_h)$
- Equation has no solution (!) if
  - two latent states have the same future distribution under π_b
  - … but the have different values in $V_{\mathcal{S}}^{\pi}(s_h)$
- New condition: outcome coverage
  - Let $\Sigma_{\mathcal{F}}$ be the confusion matrix of predicting s_h from future
  - Finite & bounded $\|\Sigma_{\mathcal{F}}^{-1} V_{\mathcal{S}}^{\pi}\|_{\infty}$



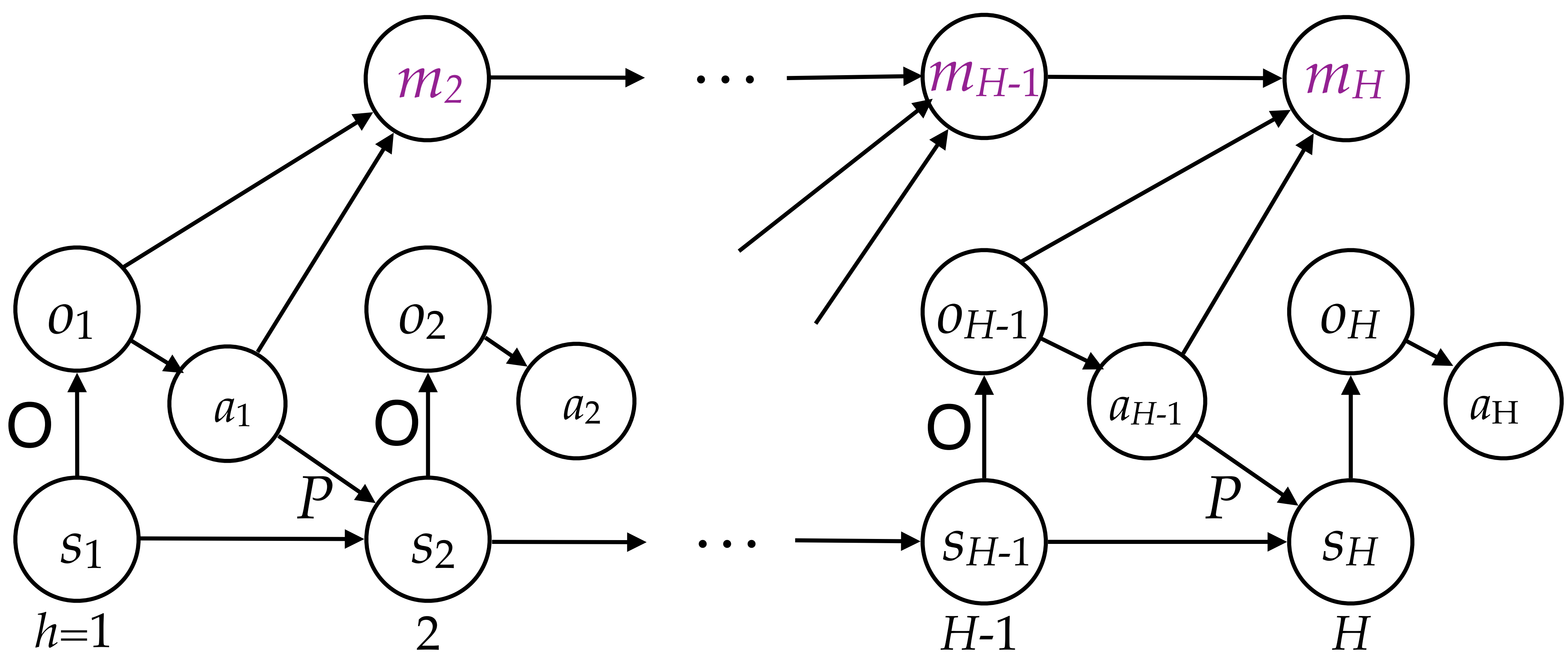
Reduction from FSM to Memoryless Policies

- Recurrent memory: $m_h = \text{update}(m_{h-1}, o_{h-1}, a_{h-1})$



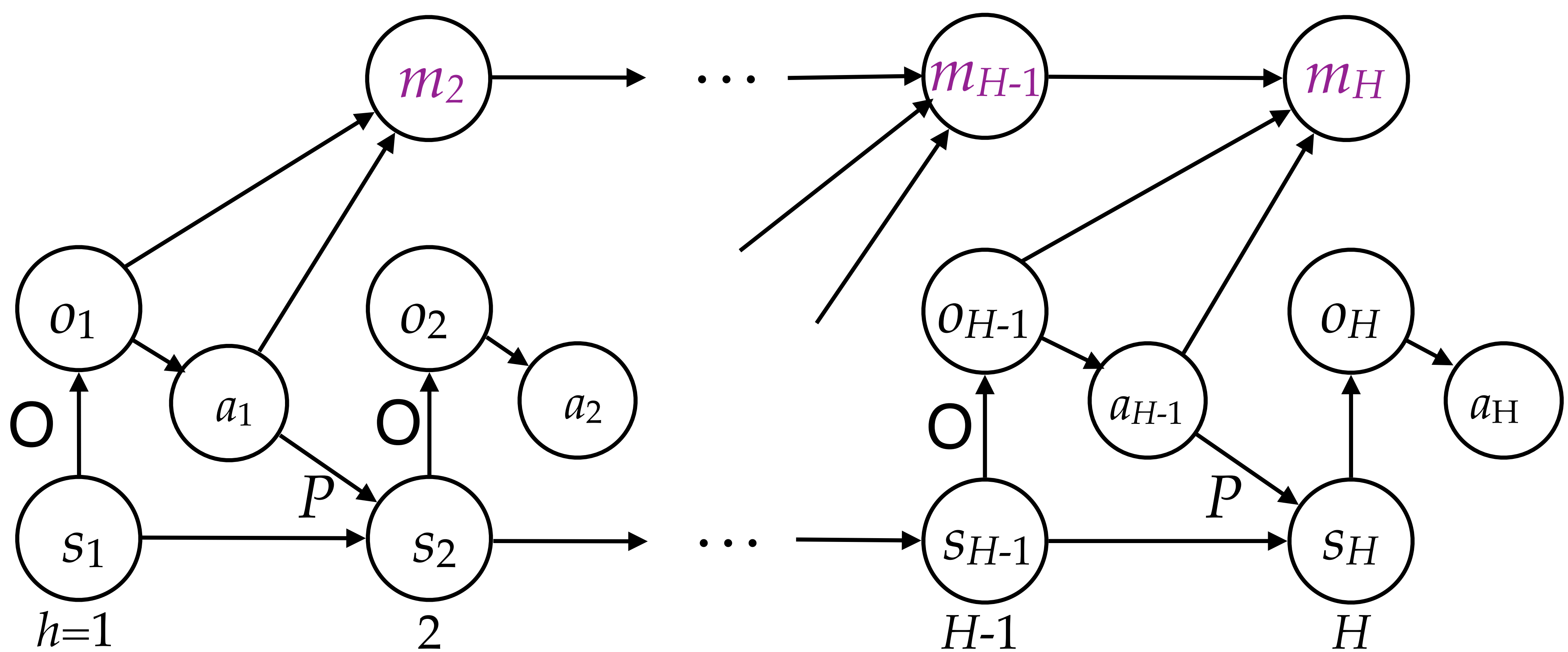
Reduction from FSM to Memoryless Policies

- Recurrent memory: $m_h = \text{update}(m_{h-1}, o_{h-1}, a_{h-1})$



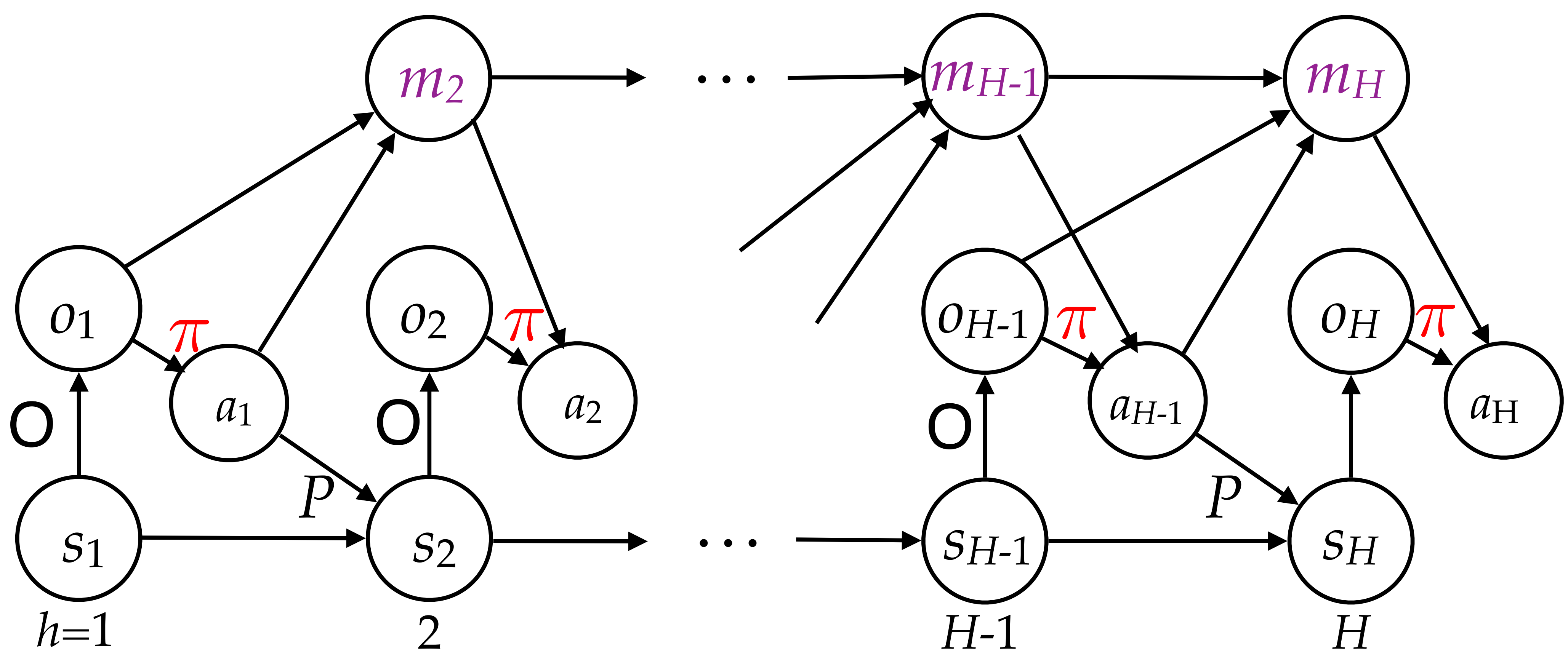
Reduction from FSM to Memoryless Policies

- Recurrent memory: $m_h = \mathrm{update}(m_{h-1}, o_{h-1}, a_{h-1})$
  - Subsume $m_h = \tau_h$ as a special case



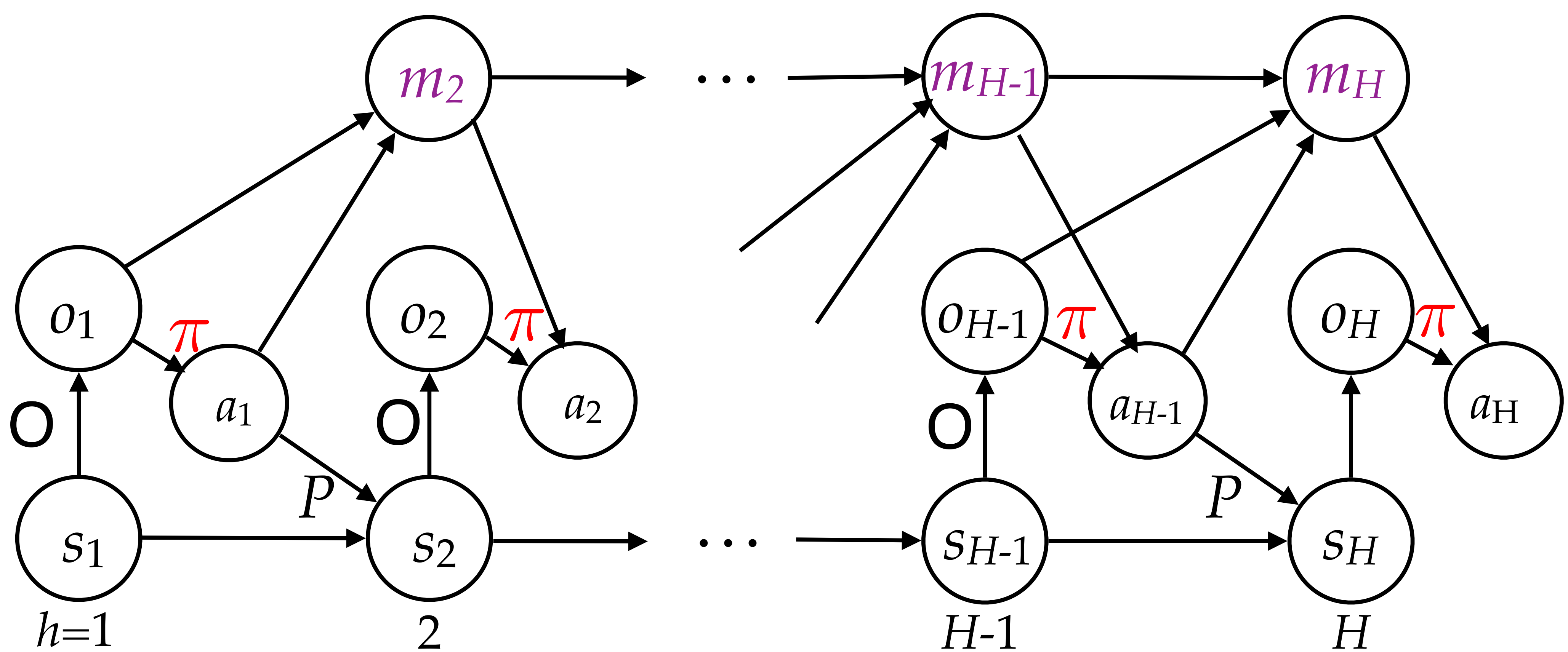
Reduction from FSM to Memoryless Policies

- Recurrent memory: $m_h = \text{update}(m_{h-1}, o_{h-1}, a_{h-1})$
  - Subsume $m_h = \tau_h$ as a special case
- FSM policy: $\pi(a_h | o_h, m_h)$



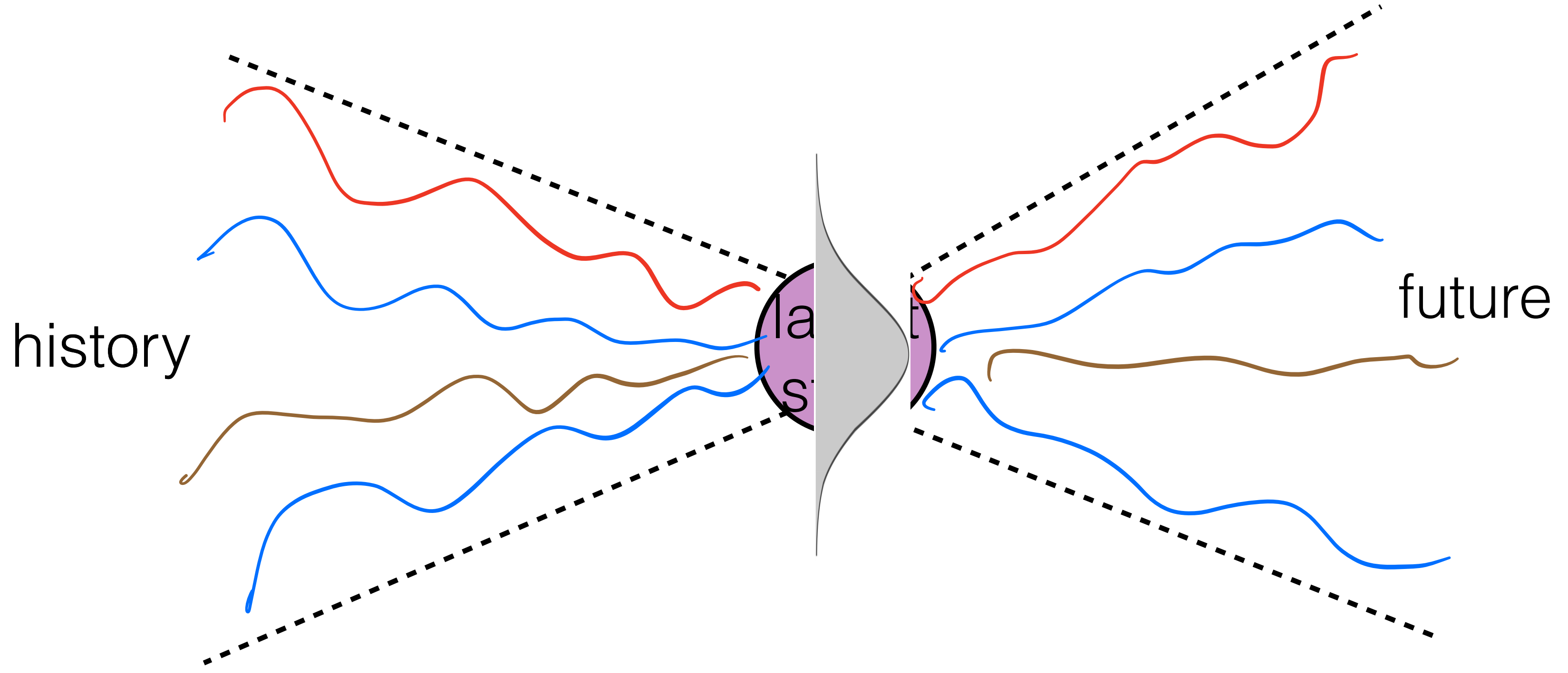
Reduction from FSM to Memoryless Policies

- Recurrent memory: $m_h = \mathrm{update}(m_{h-1}, o_{h-1}, a_{h-1})$
  - Subsume $m_h = \tau_h$ as a special case
- FSM policy: $\pi(a_h | o_h, m_h)$
- Reduction: define an augmented POMDP
  - latent state $\tilde{s}_h = (s_h, m_h)$
  - observation $\tilde{o}_h = (o_h, m_h)$



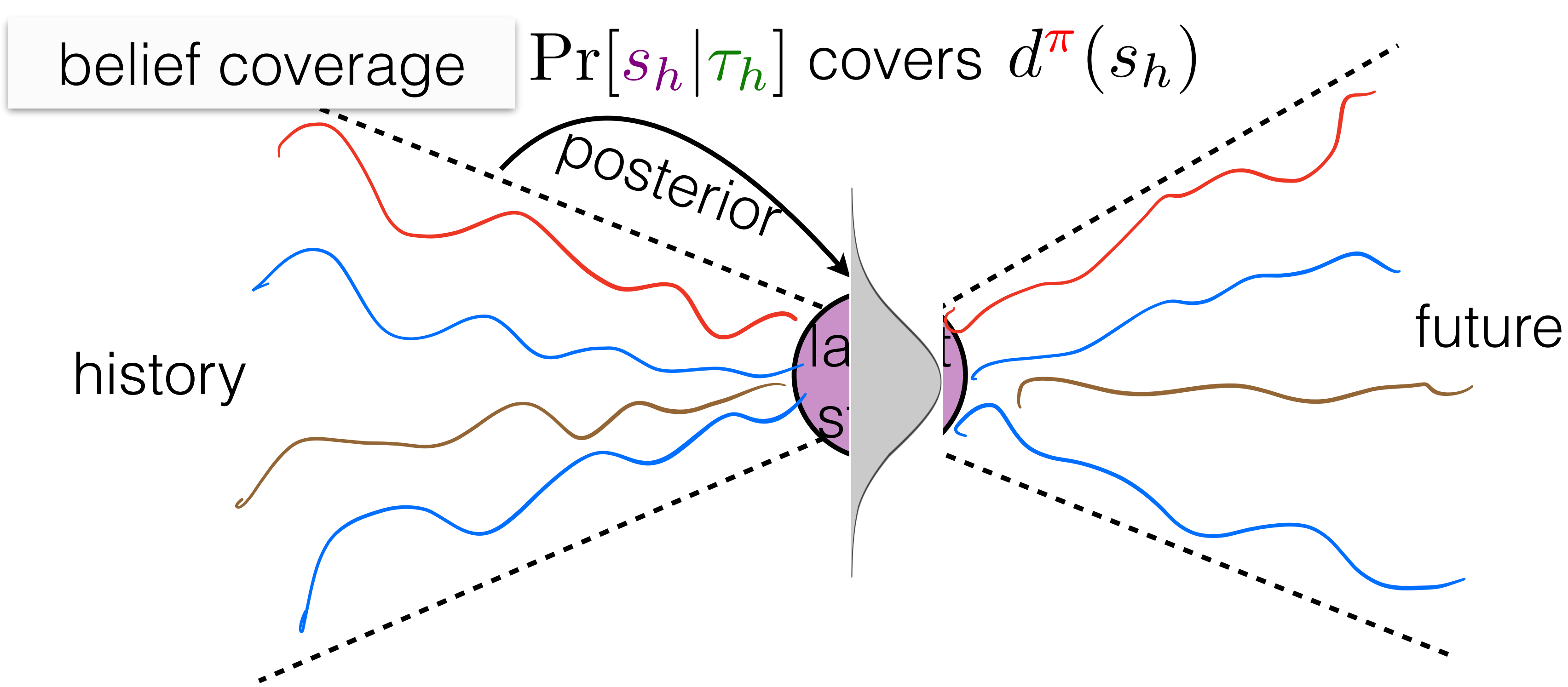
Conclusion

- Problem: OPE in POMDPs
- New framework: future-dependent value function $V_{\mathcal{F}}^{\pi}$



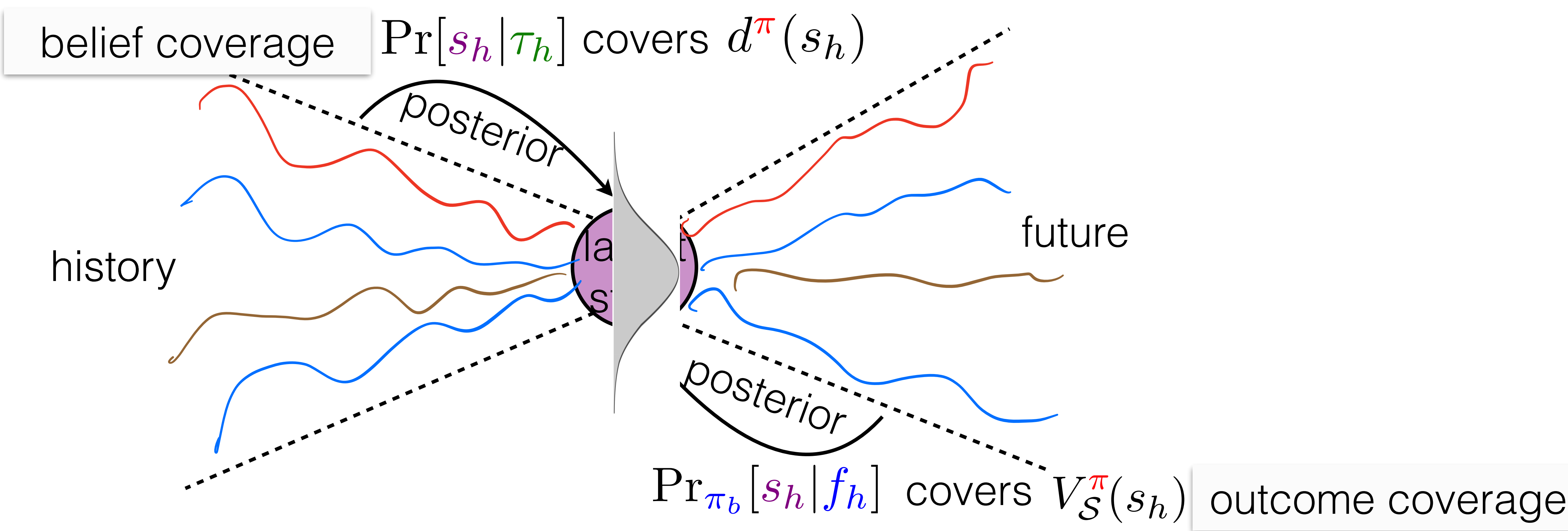
Conclusion

- Problem: OPE in POMDPs
- New framework: future-dependent value function $V_{\mathcal{F}}^{\pi}$
- New assumptions:
  - **Belief coverage =>** error transfer



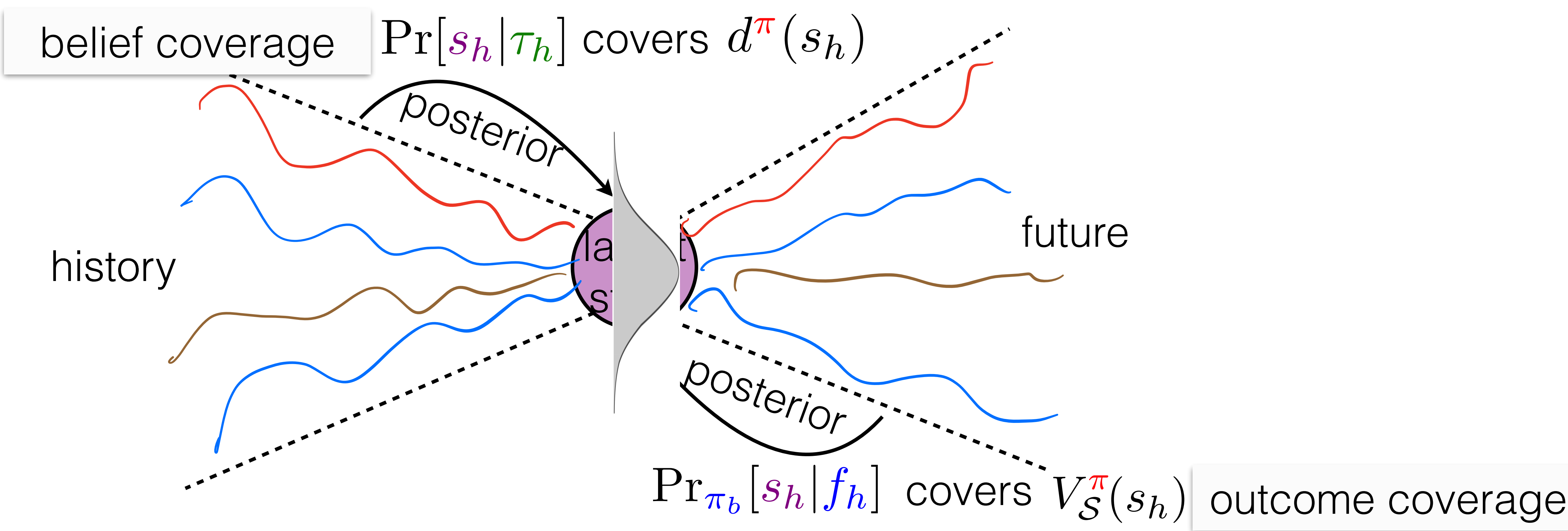
Conclusion

- Problem: OPE in POMDPs
- New framework: future-dependent value function $V_{\mathcal{F}}^{\pi}$
- New assumptions:
  - **Belief coverage =>** error transfer
  - **Outcome coverage =>** bounded $V_{\mathcal{F}}^{\pi}$

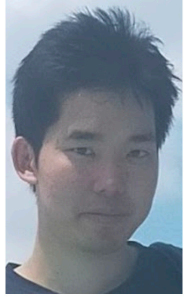


Conclusion

- Problem: OPE in POMDPs
- New framework: future-dependent value function $V^{\pi}_{\mathcal{F}}$
- New assumptions:
  - **Belief coverage =>** error transfer
  - **Outcome coverage =>** bounded $V^{\pi}_{\mathcal{F}}$
- Open question: beyond memoryless & FSM policies



Conclusion



Masatoshi Uehara

Yuheng Zhang

- Problem: OPE in POMDPs
- New framework: future-dependent value function $V_{\mathcal{F}}^{\pi}$
- New assumptions:
  - **Belief coverage =>** error transfer
  - **Outcome coverage =>** bounded $V_{\mathcal{F}}^{\pi}$
- Open question: beyond memoryless & FSM policies