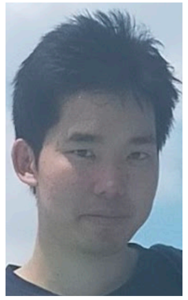
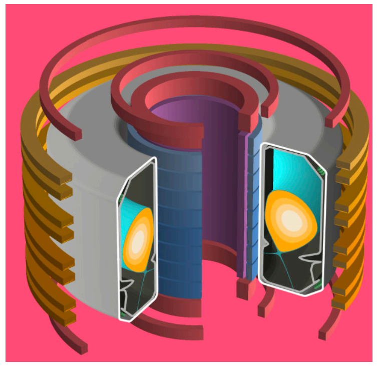
On the Curses of Future and History in Partially Observed Off-policy Evaluation

Nan Jiang
University of Illinois at Urbana-Champaign
March 2024

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



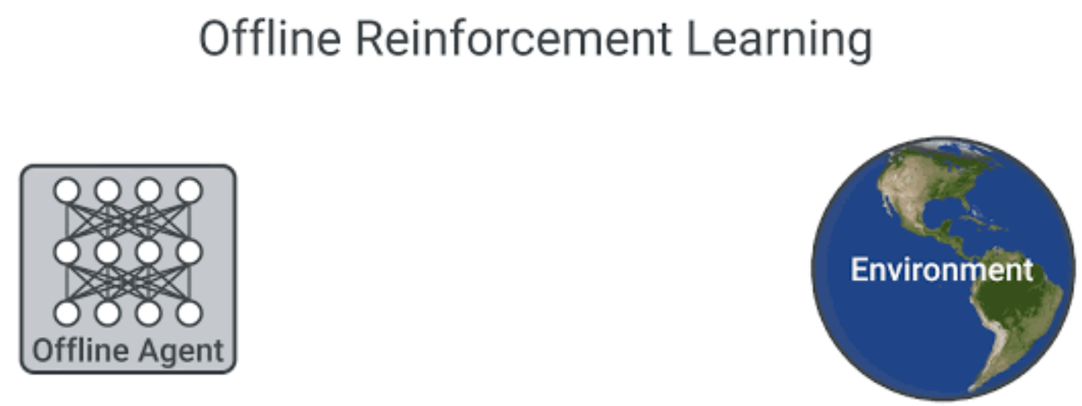
Key ingredient: **simulator**

- Unlimited data
- Decision w/o real consequences
- Can easily evaluate new strategy

Key ingredient: **simulator**

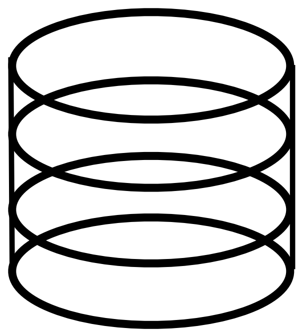
- Unlimited data X
- Decision w/o real consequences X
- Can easily evaluate new strategy X



Key ingredient: **simulator**

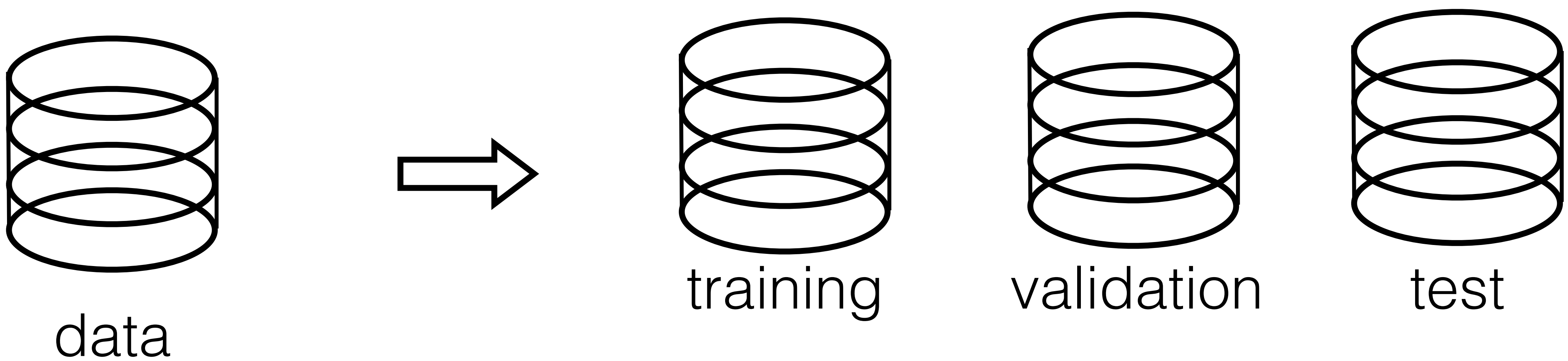
- Unlimited data X
- Decision w/o real consequences X
- Can easily evaluate new strategy X

Supervised learning pipeline

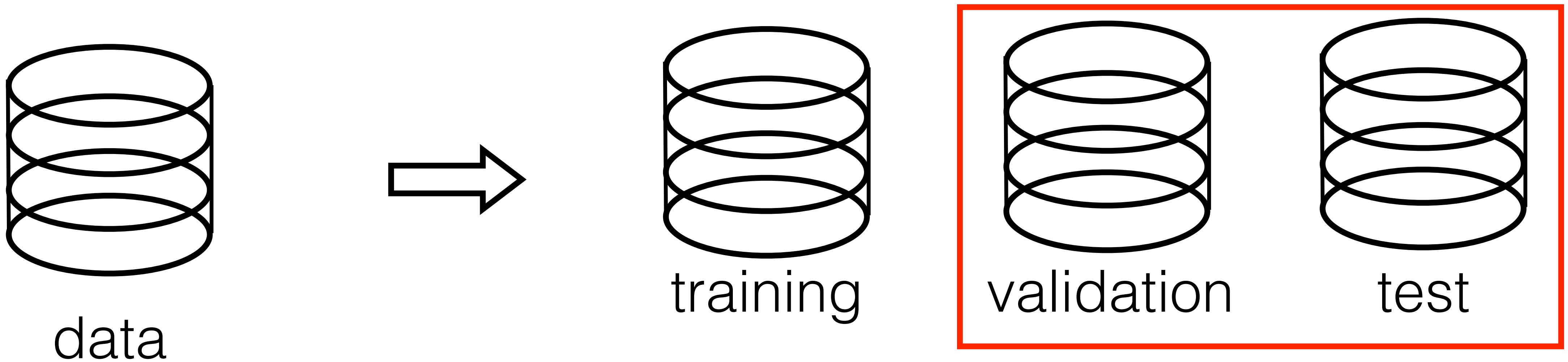


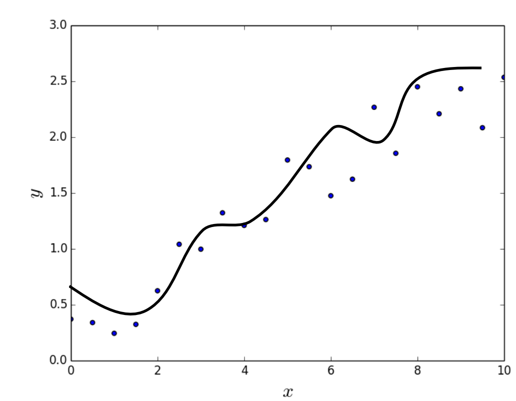
data

Supervised learning pipeline

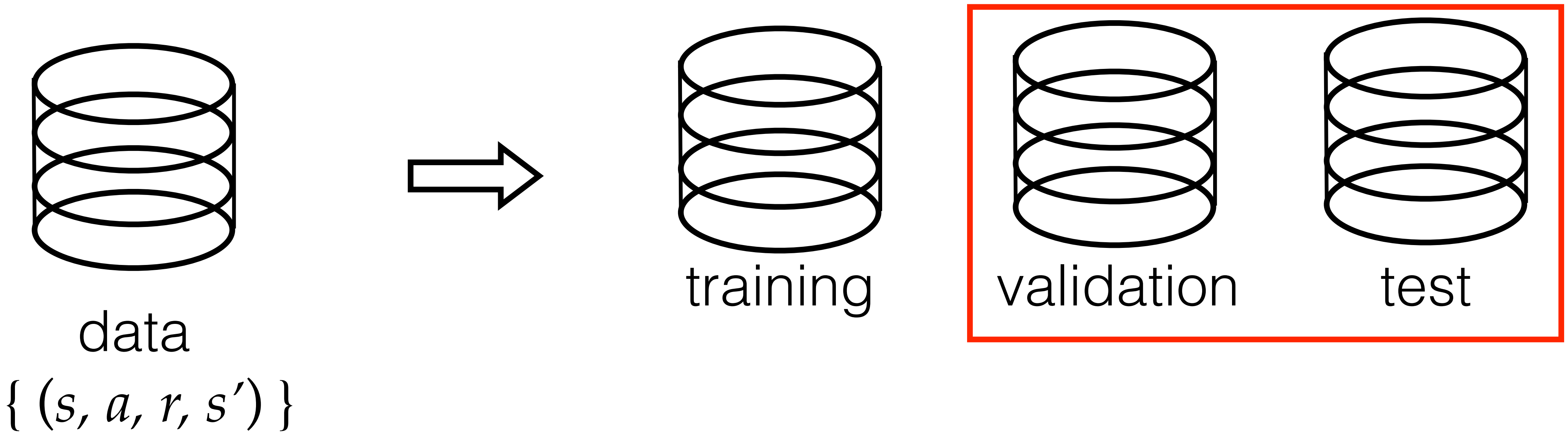


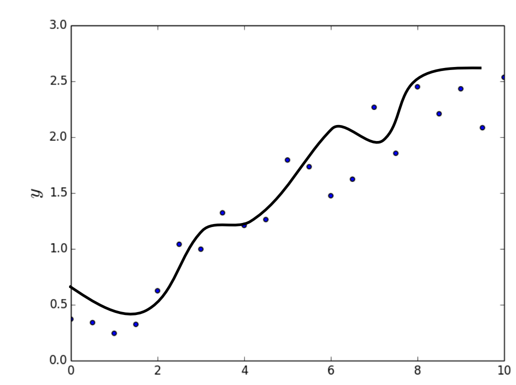
Supervised learning pipeline



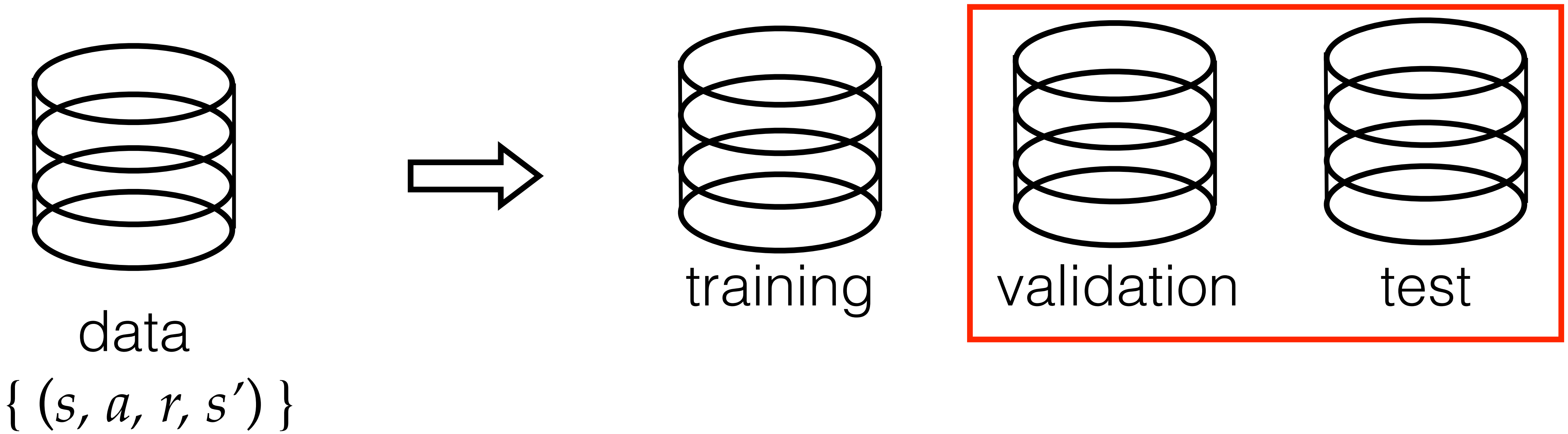


Offline RL pipeline

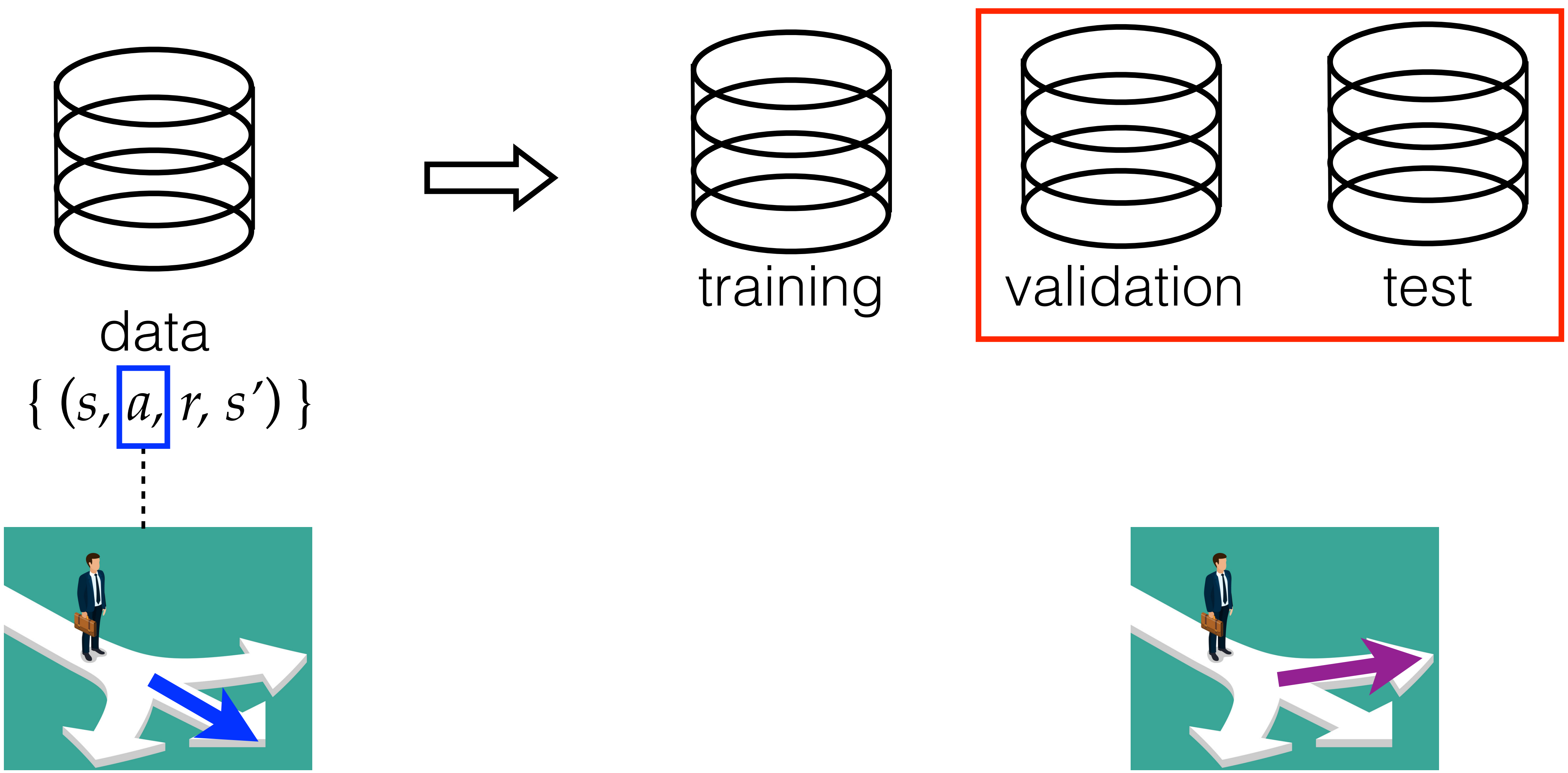




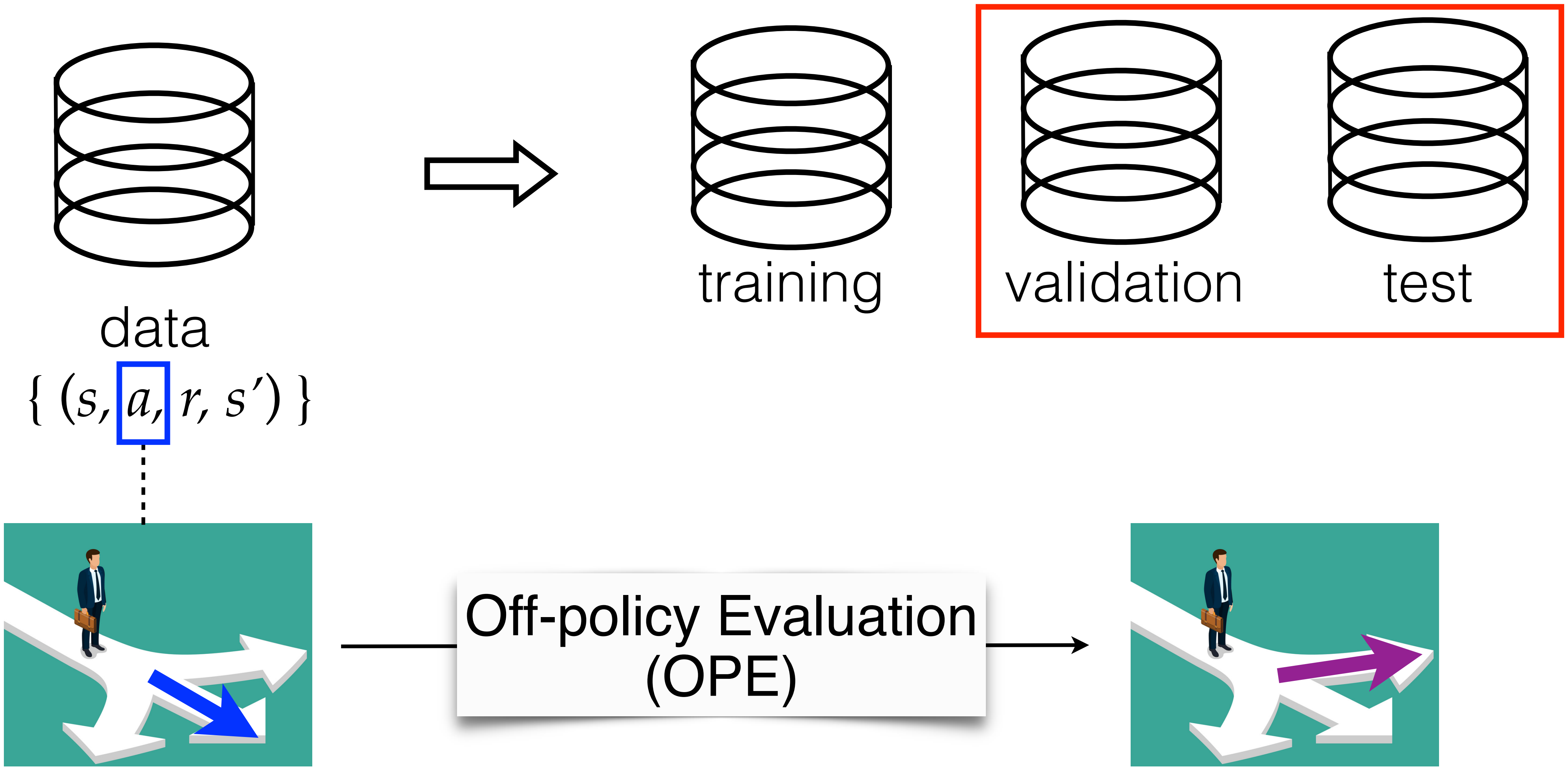
Offline RL pipeline



Offline RL pipeline



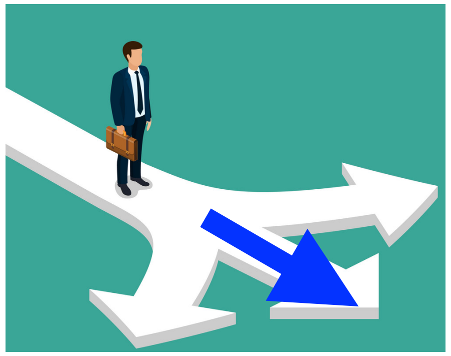
Offline RL pipeline



Unbiased OPE

Importance sampling (IS) [Precup'00]

Behavior



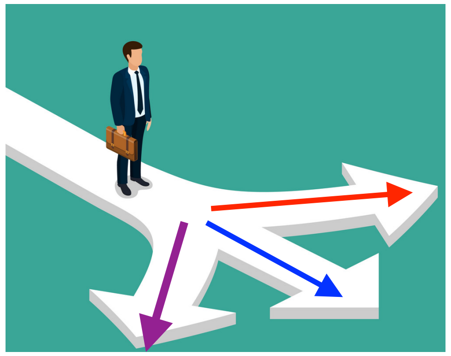
Target

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

Unbiased OPE

Importance sampling (IS) [Precup'00]

Behavior



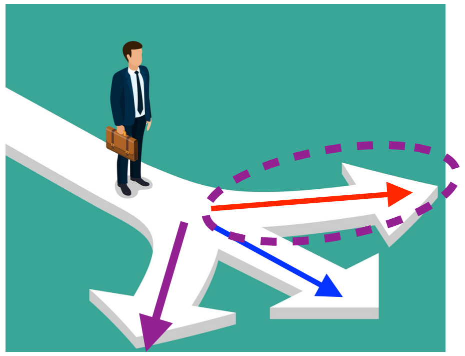
Target

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

Unbiased OPE

Importance sampling (IS) [Precup'00]

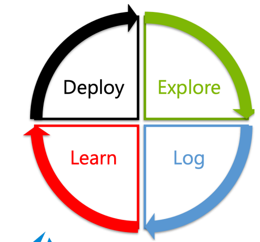
Behavior



Target

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

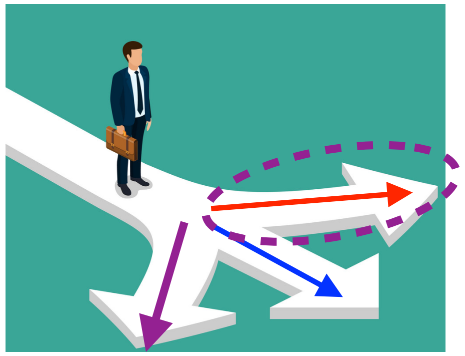
Unbiased OPE



Importance sampling (IS) [Precup'00]

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

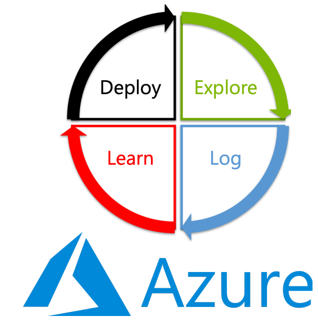
Behavior



Target

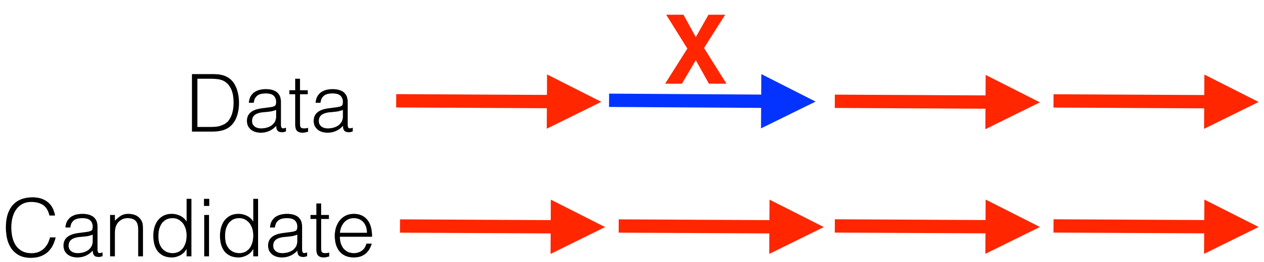
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Unbiased OPE

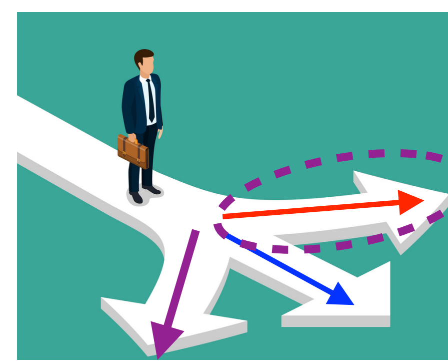


Importance sampling (IS) [Precup'00]

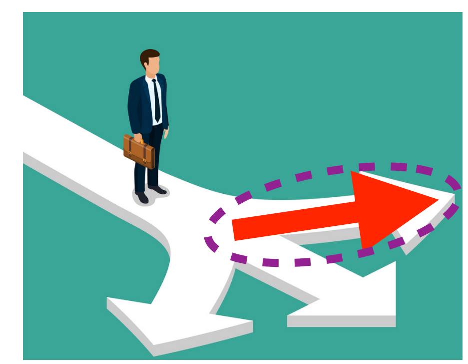
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!



Behavior

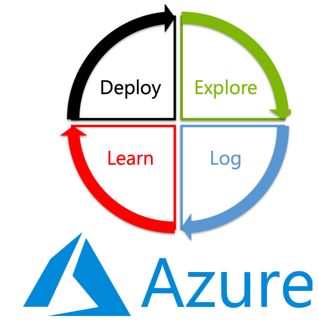


Target



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

Unbiased OPE



Importance sampling (IS) [Precup'00]

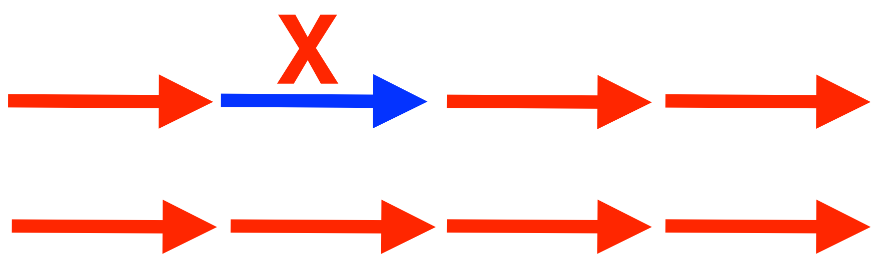
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!

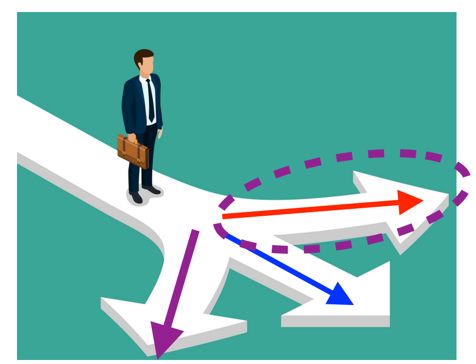
Behavior

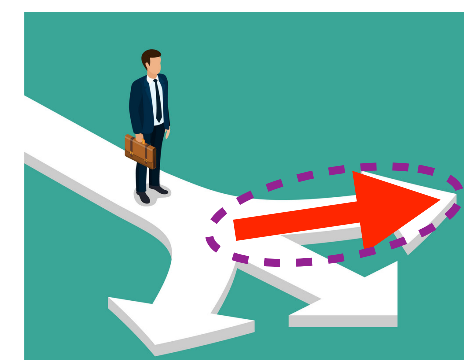
Target

Data

Candidate



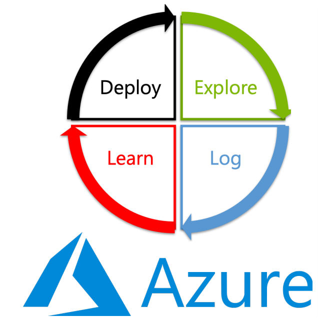




$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H \quad \Rightarrow \quad \left(\prod_{h=1}^{H} \frac{\pi(a_h|o_h)}{\pi_b(a_h|o_h)}\right)\left(\sum_{h=1}^{H} r_h\right)$$

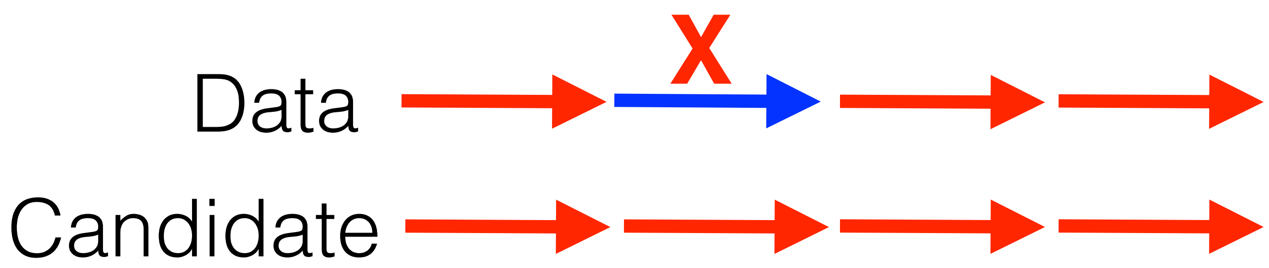
Precup. 2000. Eligibility traces for off-policy policy evaluation.

Unbiased OPE

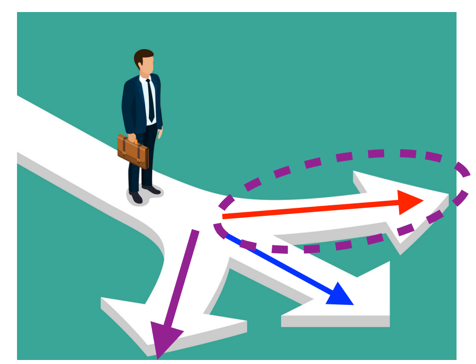


Importance sampling (IS) [Precup'00]

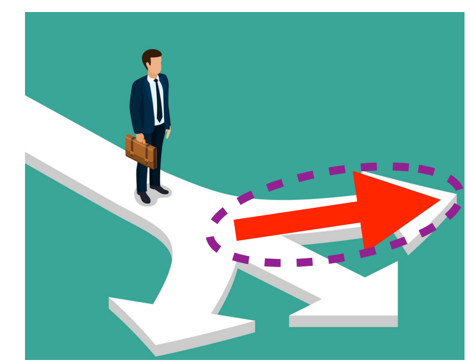
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!



Behavior



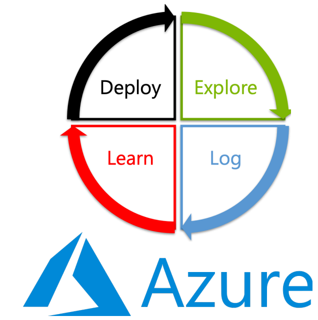
Target



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

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

Unbiased OPE



Importance sampling (IS) [Precup'00]

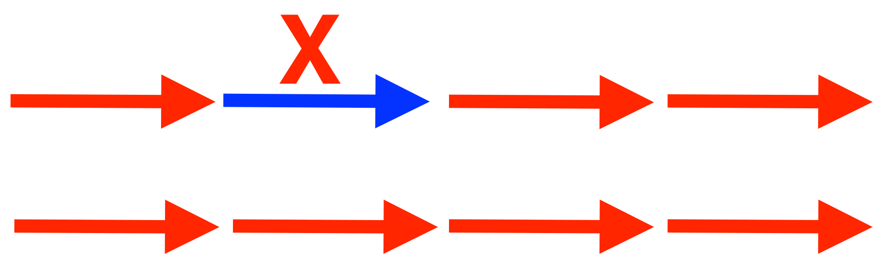
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!

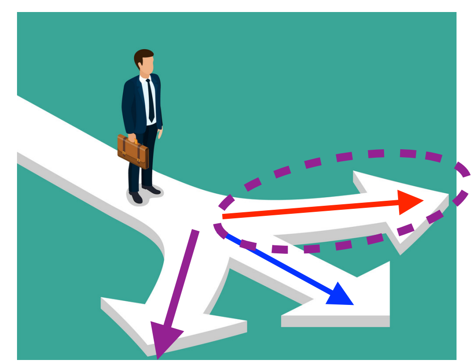
Behavior

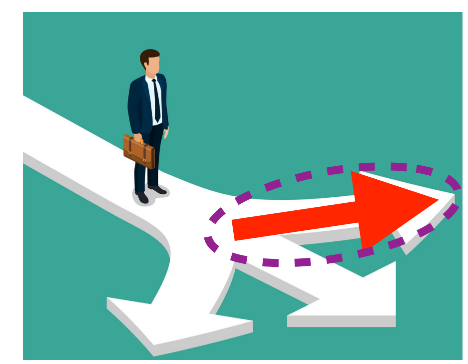
Target

Data

Candidate





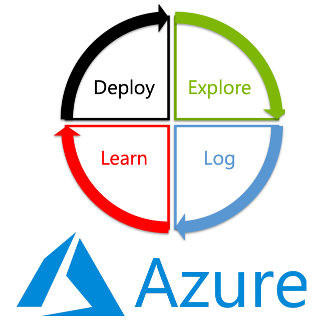


$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H \quad \Rightarrow \quad \boxed{\left(\prod_{h=1}^{H} \frac{\pi(a_h|o_h)}{\pi_b(a_h|o_h)}\right)} \left(\sum_{h=1}^{H} r_h\right)$$

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

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

Unbiased OPE

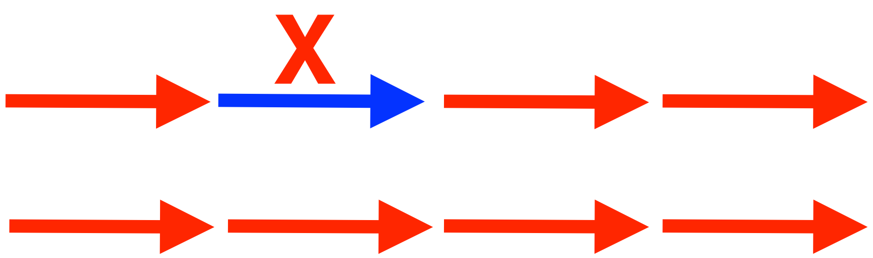


Importance sampling (IS) [Precup'00]

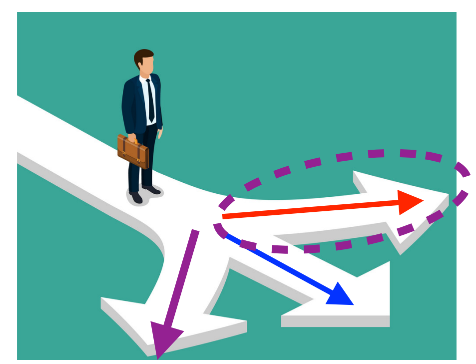
- Industry deployment (ctx. bandit, horizon=1)
- No Markovianity required ✓
- Exponential-in-horizon variance!

Data

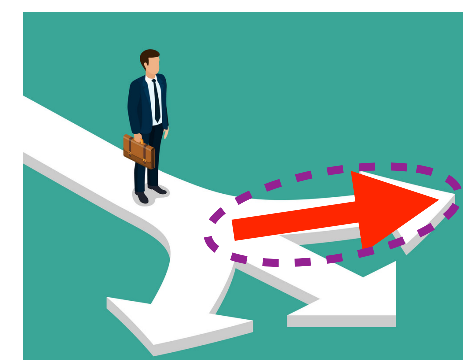
Candidate



Behavior



Target



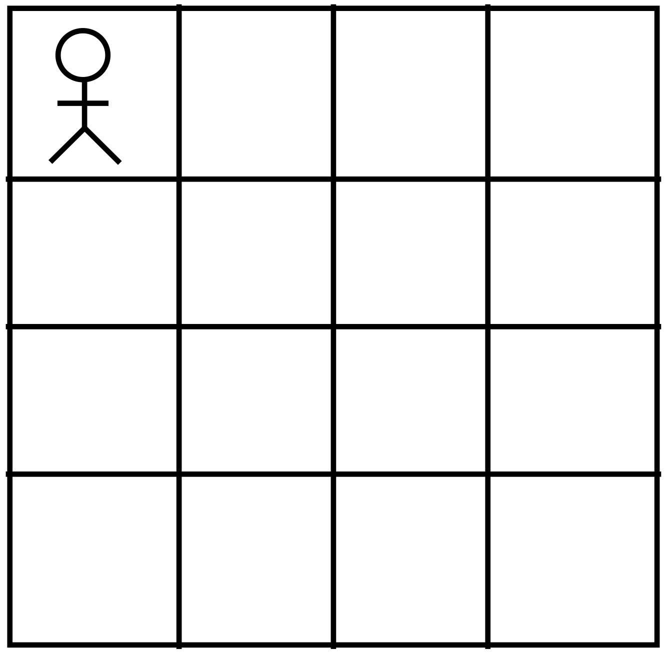
$$o_1, a_1, r_1, \ldots, o_H, a_H, r_H \Rightarrow \left(\prod_{h=1}^{H} \frac{\pi(a_h|o_h)}{\pi_b(a_h|o_h)}\right)\left(\sum_{h=1}^{H} r_h\right)$$

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

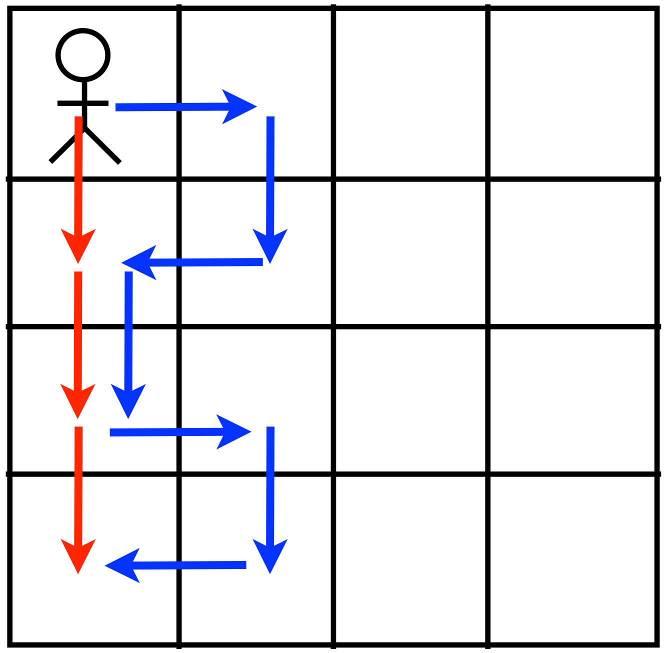
IS' measure of *coverage*

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

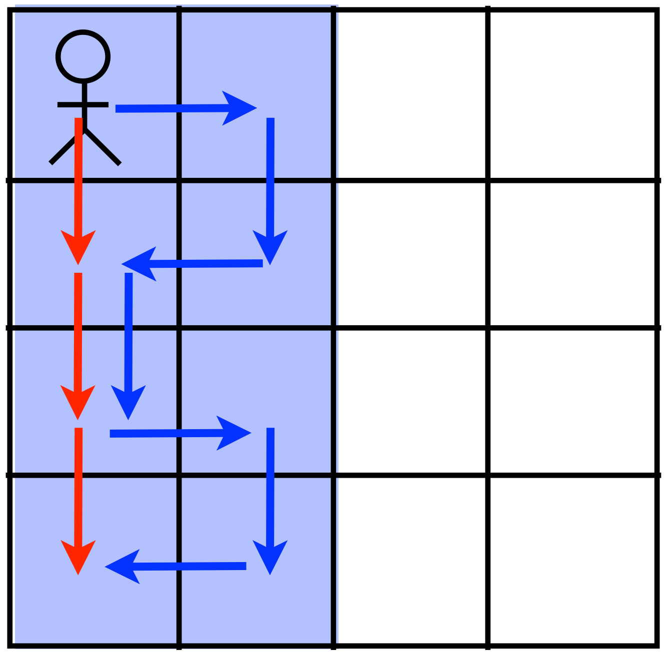
Better coverage?



Better coverage?



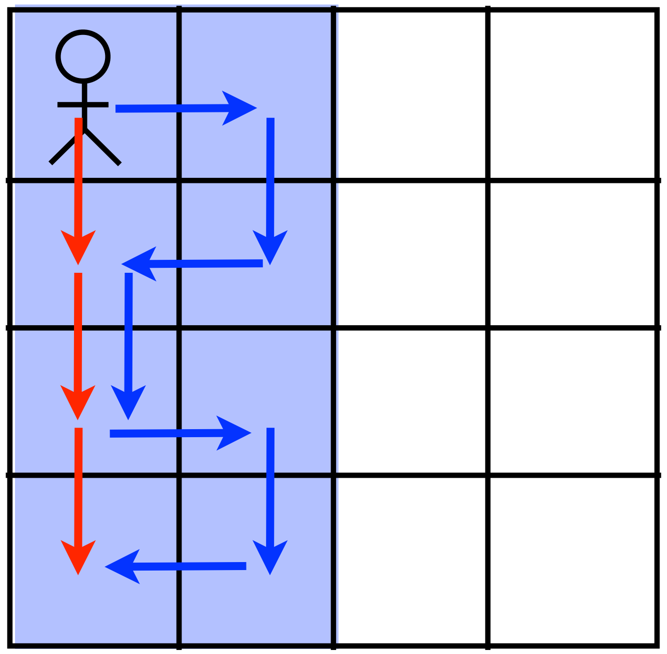
Better coverage?



Better coverage?

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

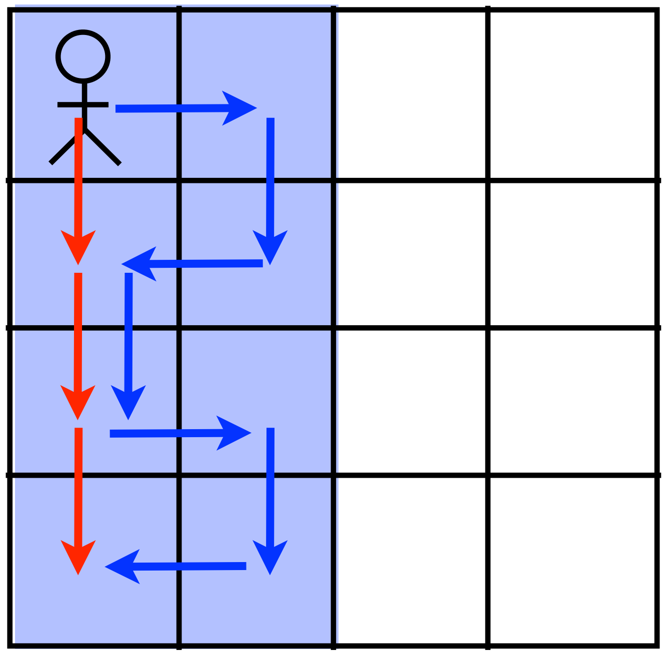
- Assume MDPs: $o_1 (= s_1), a_1, r_1, \ldots, o_H (= s_H), a_H, r_H$



Better coverage?

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

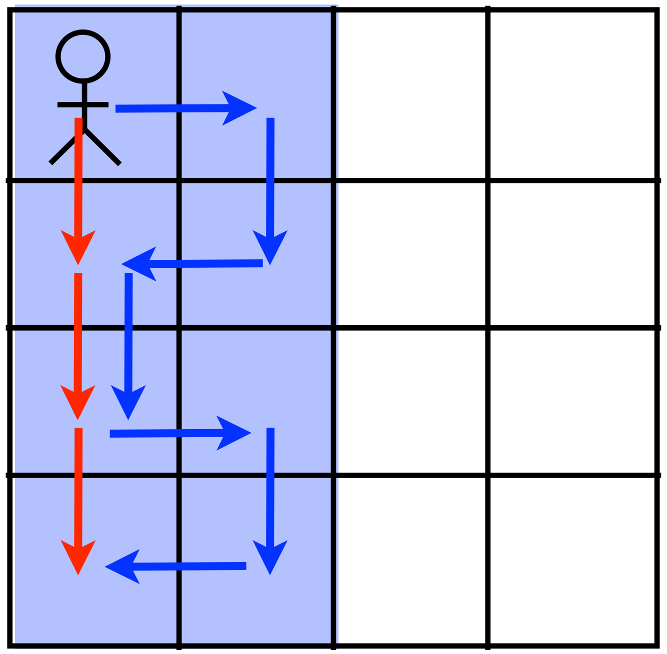
- Assume MDPs: $o_1(= s_1), a_1, r_1, \ldots, o_H(= s_H), a_H, r_H$
- Learn value functions: $V^\pi(s_h) := \mathbb{E}_\pi[\sum_{h'=h}^{H} r_{h'} | s_h]$
  - Satisfies $V^\pi(s_h) = (\mathcal{T}^\pi V^\pi)(s_h) := R(s_h, \pi) + \mathbb{E}_{s_{h+1} \sim P(s_h, \pi)}[V^\pi(s_{h+1})]$.



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

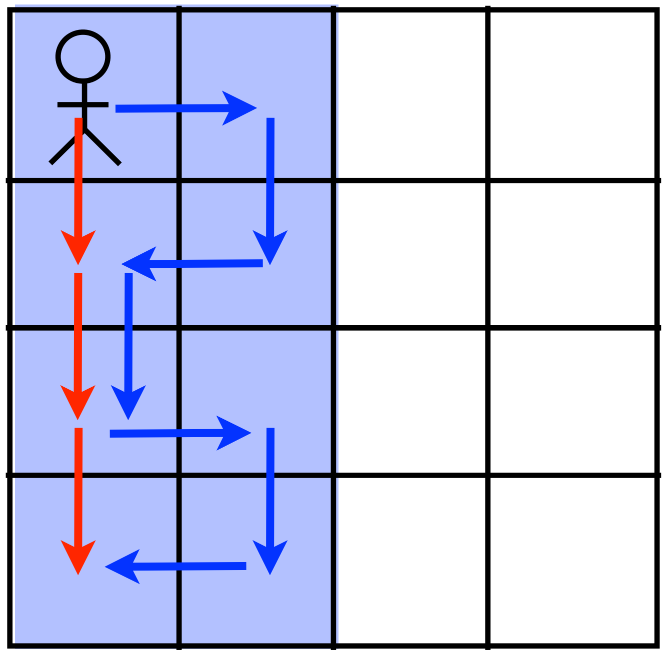


* Also needs Bellman completeness

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



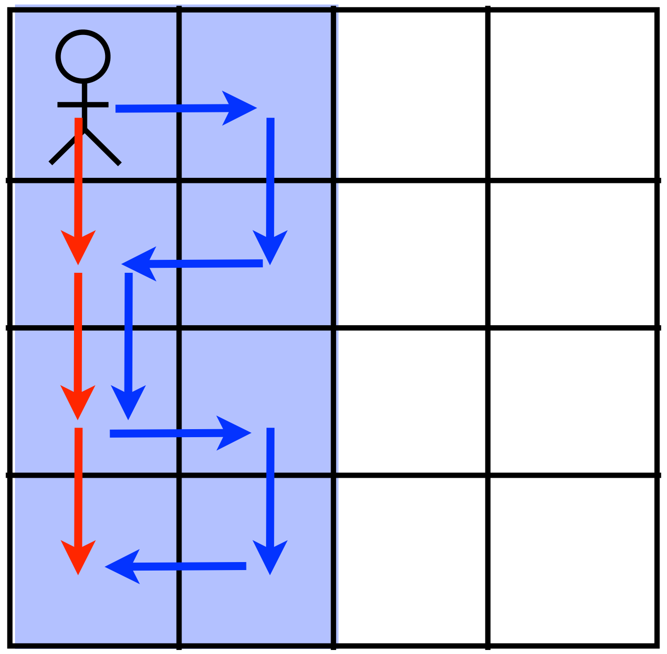
* Also needs Bellman completeness

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?

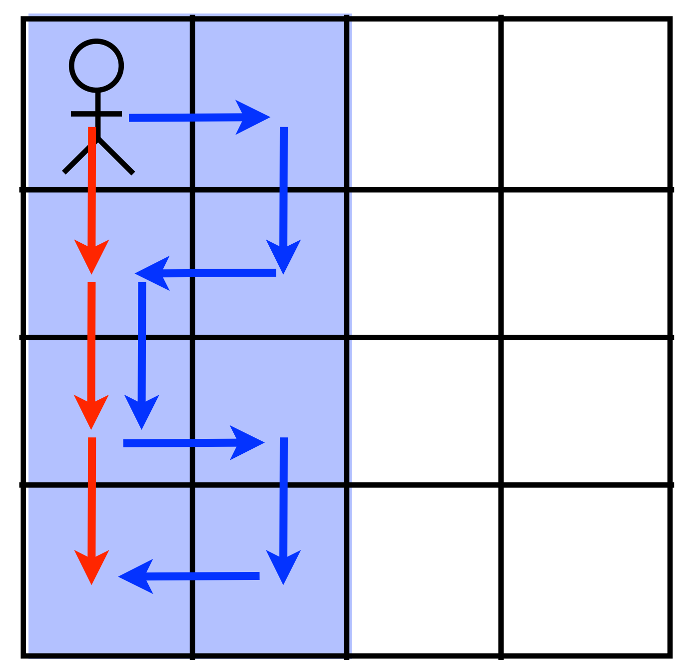
FQE [Munos, Szepesvari… CJ'19, …] / MIS [Liu et al'18, Nachum et al'19, UHJ'20, …]

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

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, o_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, o_h). Problem solved?

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

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, o_h). Problem solved?

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

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, o_h). Problem solved?

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

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

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, o_h). Problem solved?

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

- Structured 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, o_h). Problem solved?

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

- Structured 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, o_h). Problem solved?

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

- Structured 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, o_h). Problem solved?

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

- Structured 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
  - connection to known empirical evidence (LLMs, RLHF, etc.)

* 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, o_h). Problem solved?

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

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

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

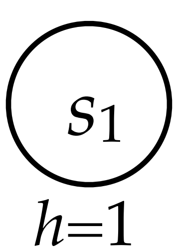
  - this assumes given low-dim linear feature to encode history…

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?)

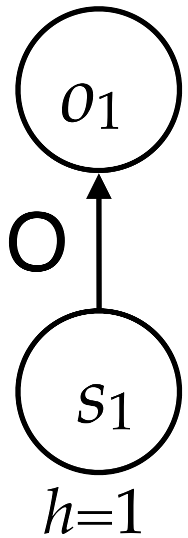


$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

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), $o_h \sim$ O$(\cdot \mid s_h)$

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

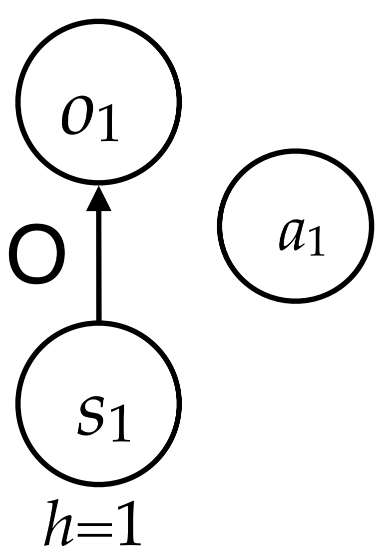


$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

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), $o_h \sim$ O$(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)

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



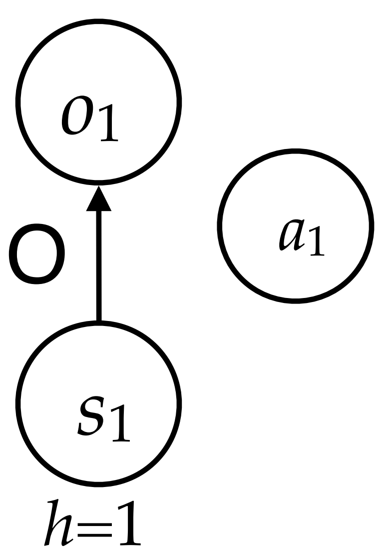
$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

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
O: $S \rightarrow \Delta(O)$

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



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

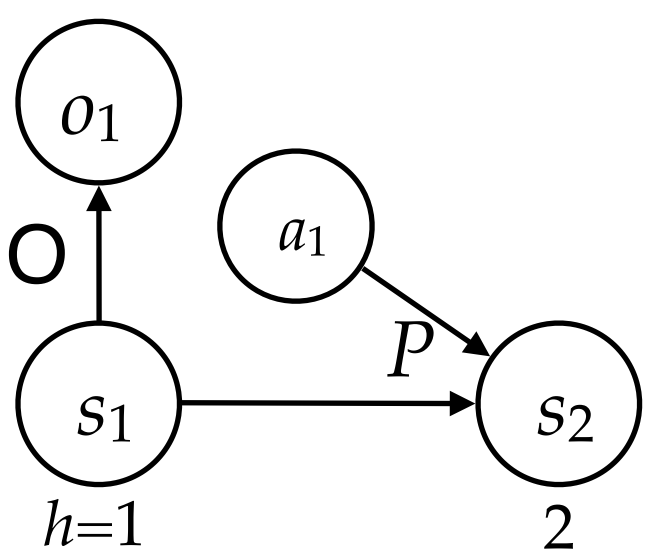
Partially Observable MDPs (POMDPs)

- 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 $r_h = R(o_h, a_h)$

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

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

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



Partially Observable MDPs (POMDPs)

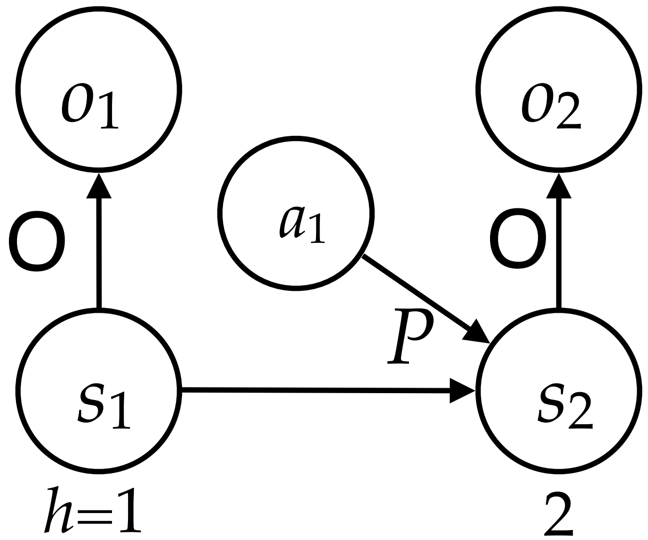
$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

- 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), $o_h \sim$ 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
O: $S \to \Delta(O)$

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



Partially Observable MDPs (POMDPs)

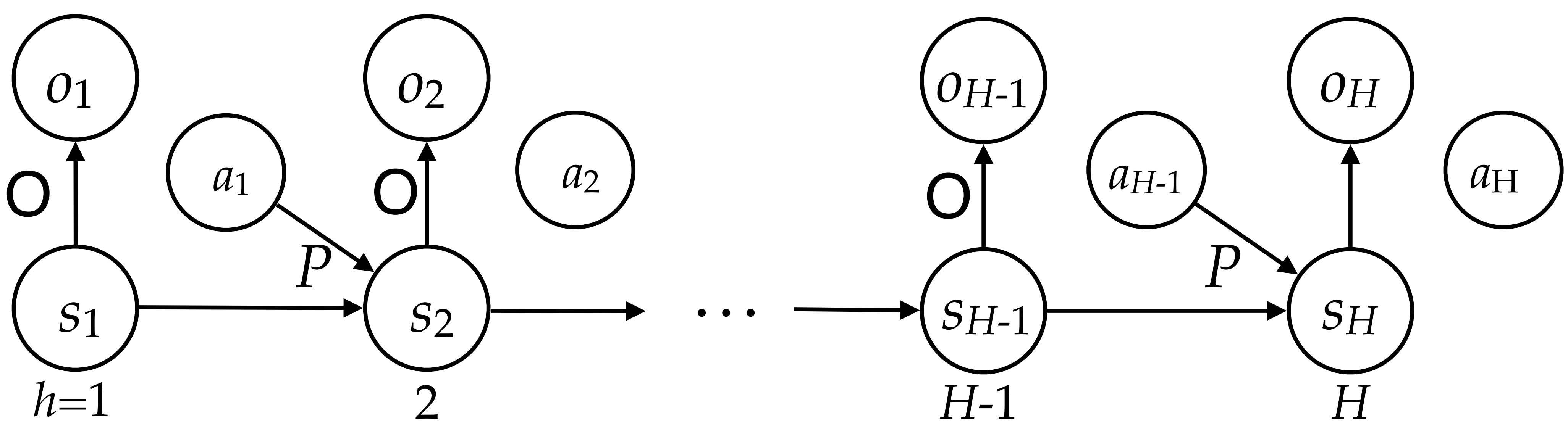
$\mathcal{S} = \bigcup_h \mathcal{S}_h$
$\mathcal{O} = \bigcup_h \mathcal{O}_h$

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



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

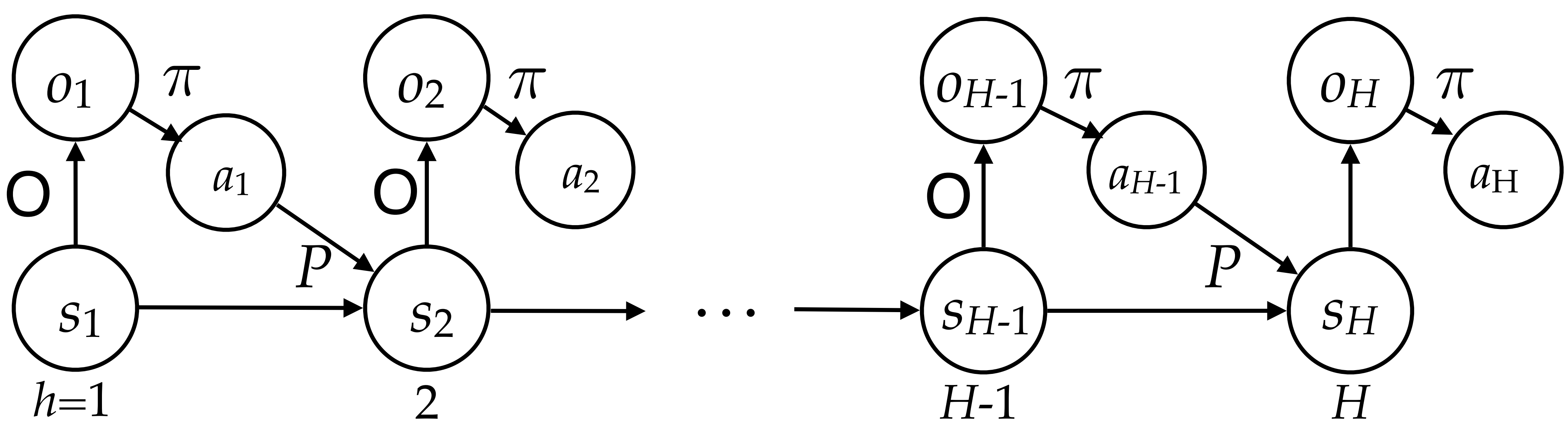
Partially Observable MDPs (POMDPs)

- 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 $r_h = R(o_h, a_h)$
- *Memoryless* policies $\pi : \mathcal{O} \to \Delta(\mathcal{A})$

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)

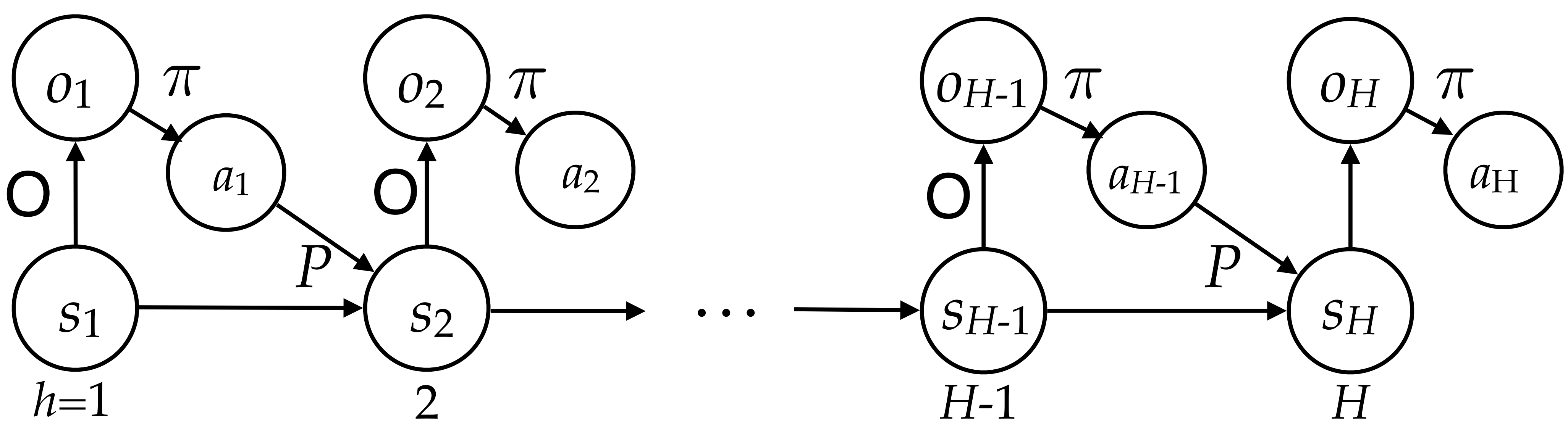
- For $h = 1, 2, \ldots, H$,
  - nature generates *latent* state $s_h \in S_h$ (small?)
    - $s_h \sim P(\cdot \mid s_{h\text{-}1}, a_{h\text{-}1})$ for $h \geq 2$
  - agent observes $o_h \in O_h$ (large), $o_h \sim \mathrm{O}(\cdot \mid s_h)$
  - chooses action $a_h \in A$ (small)
  - receives reward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

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

emission process
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



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

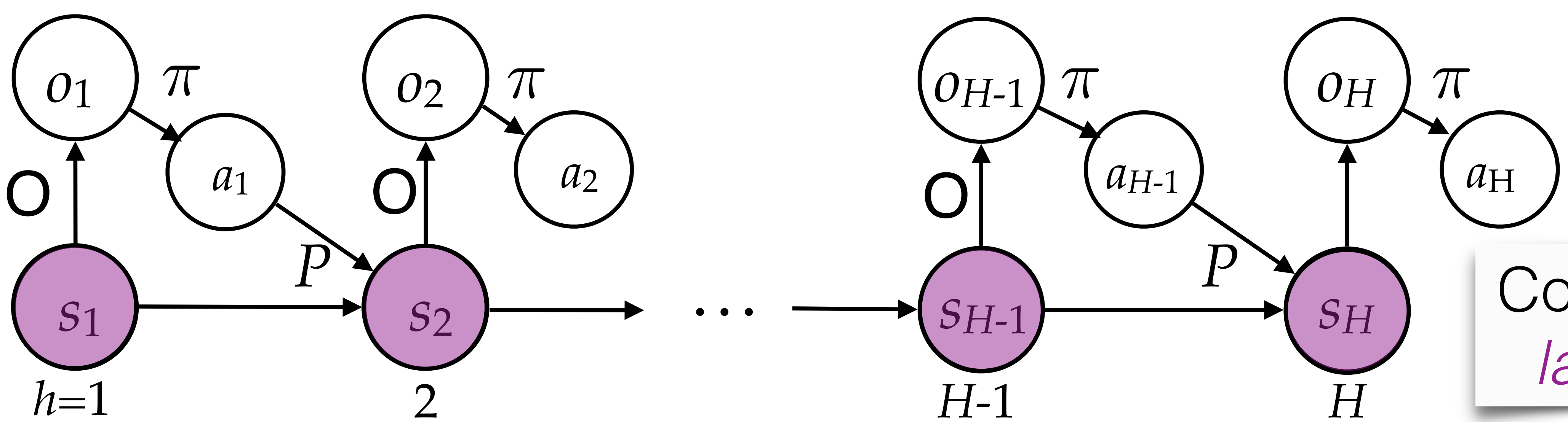
Partially Observable MDPs (POMDPs)

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

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

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

- 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 $r_h = R(o_h, a_h)$
- *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!

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!

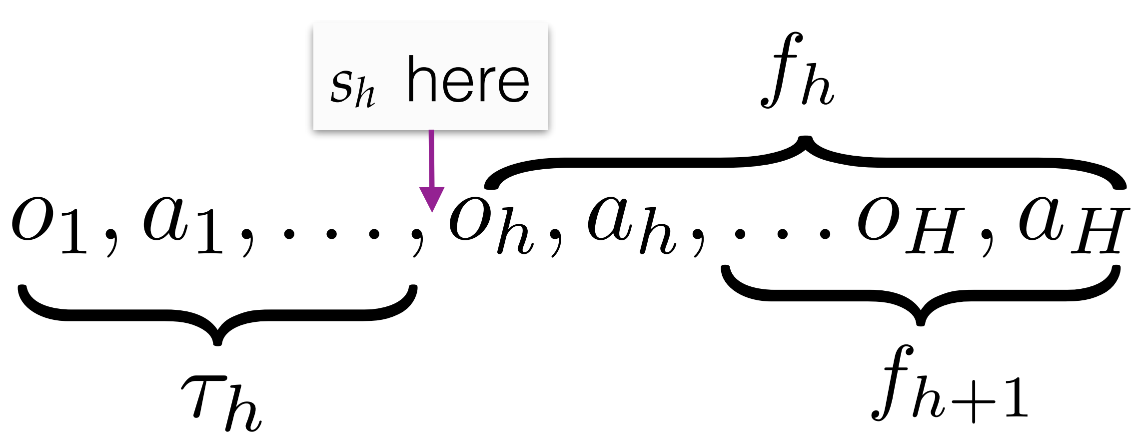
Future-Dependent Value Function

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



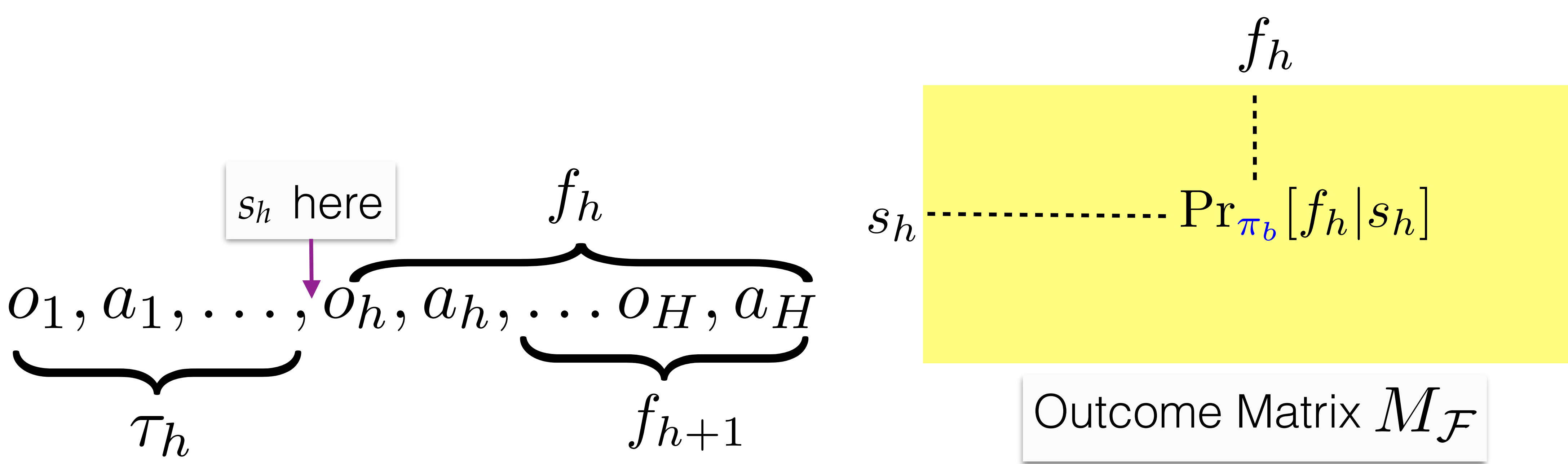
Future-Dependent Value Function

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

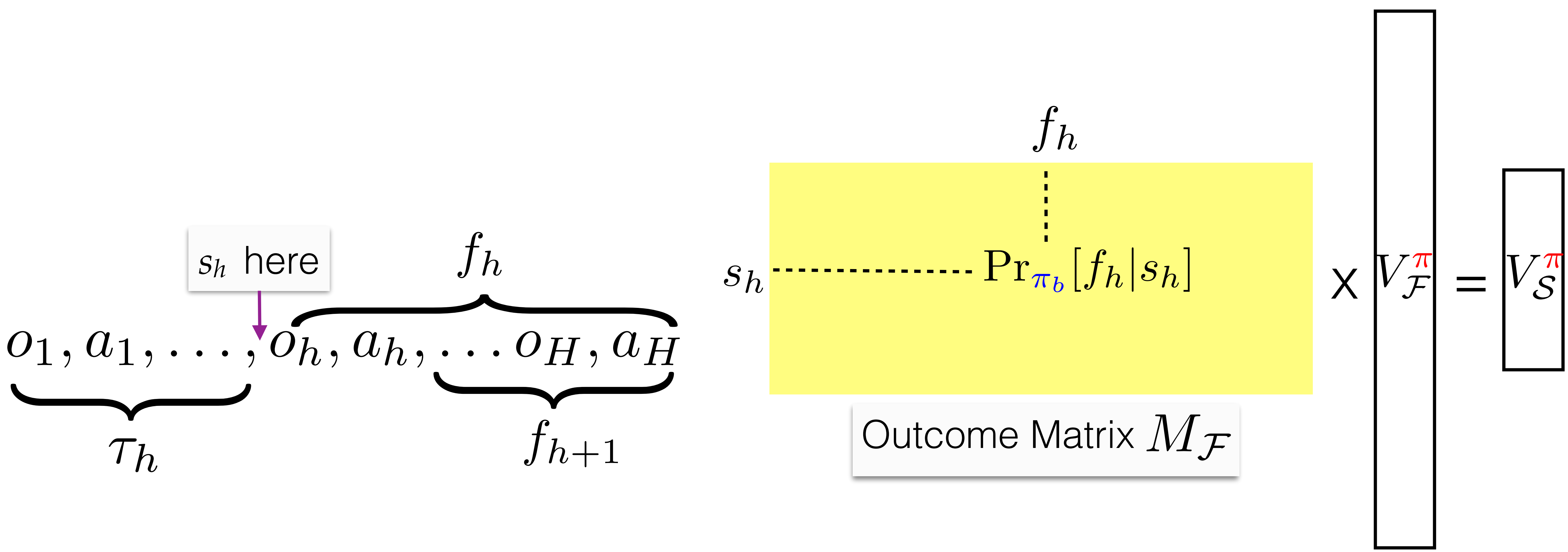


Future-Dependent Value Function

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



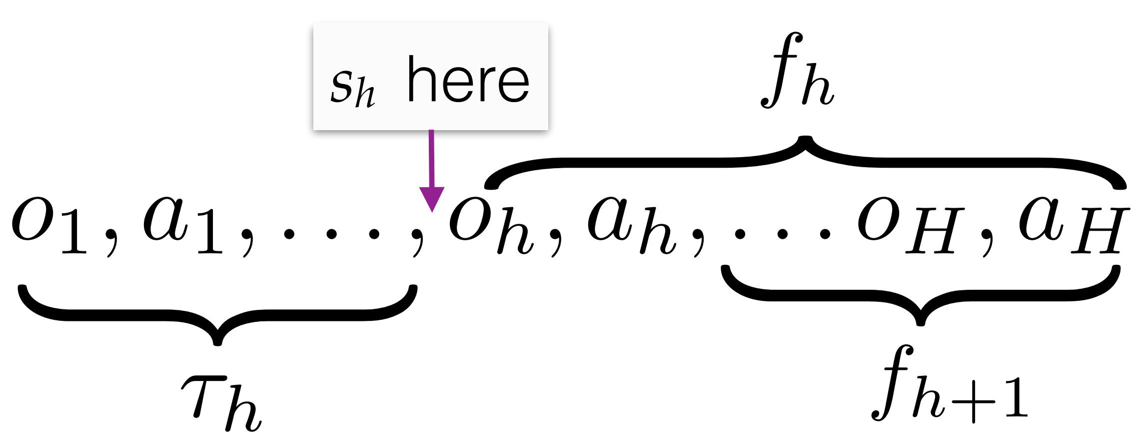
Future-Dependent Value Function

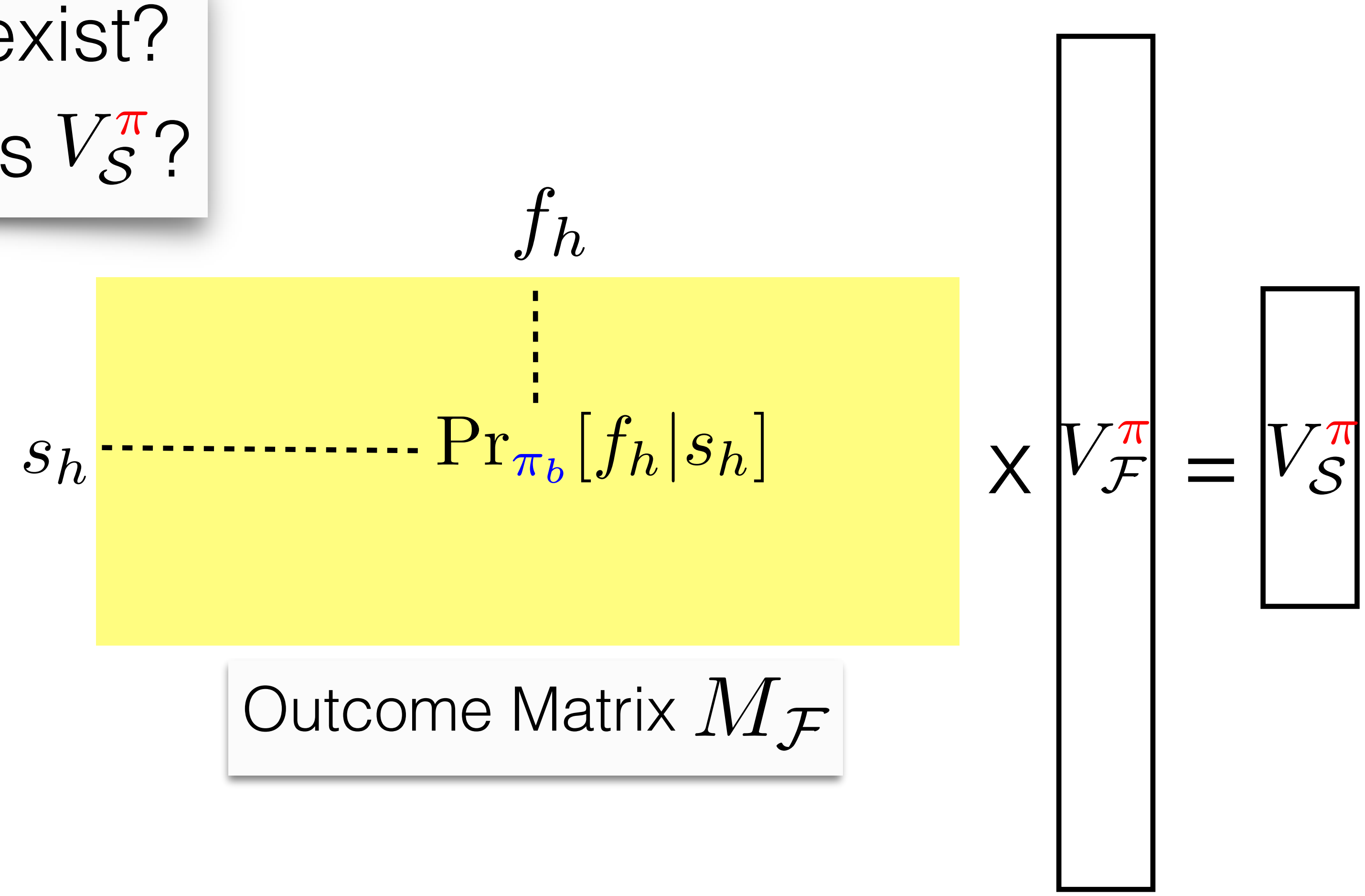
- Define: value function of latent state

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

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

1. Does (well-behaved) $V_{\mathcal{F}}^{\pi}$ even exist?
2. Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?





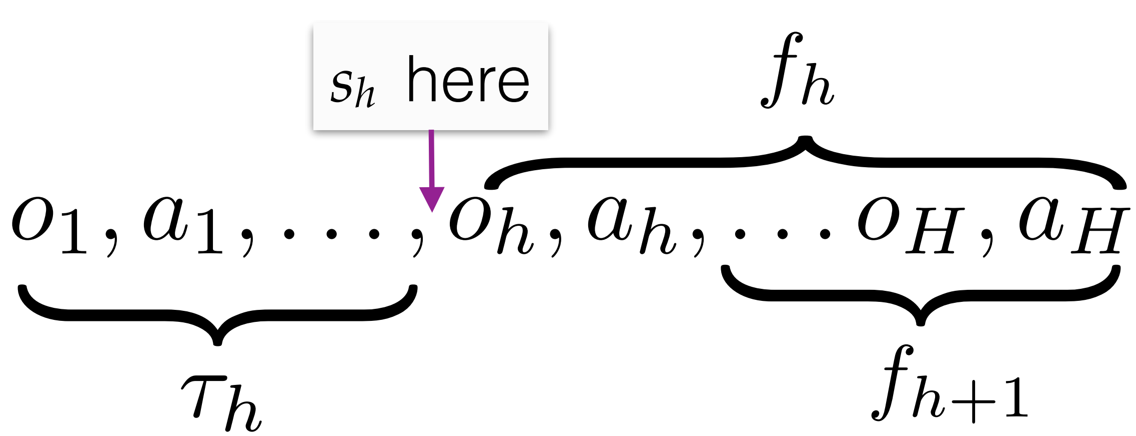
Future-Dependent Value Function

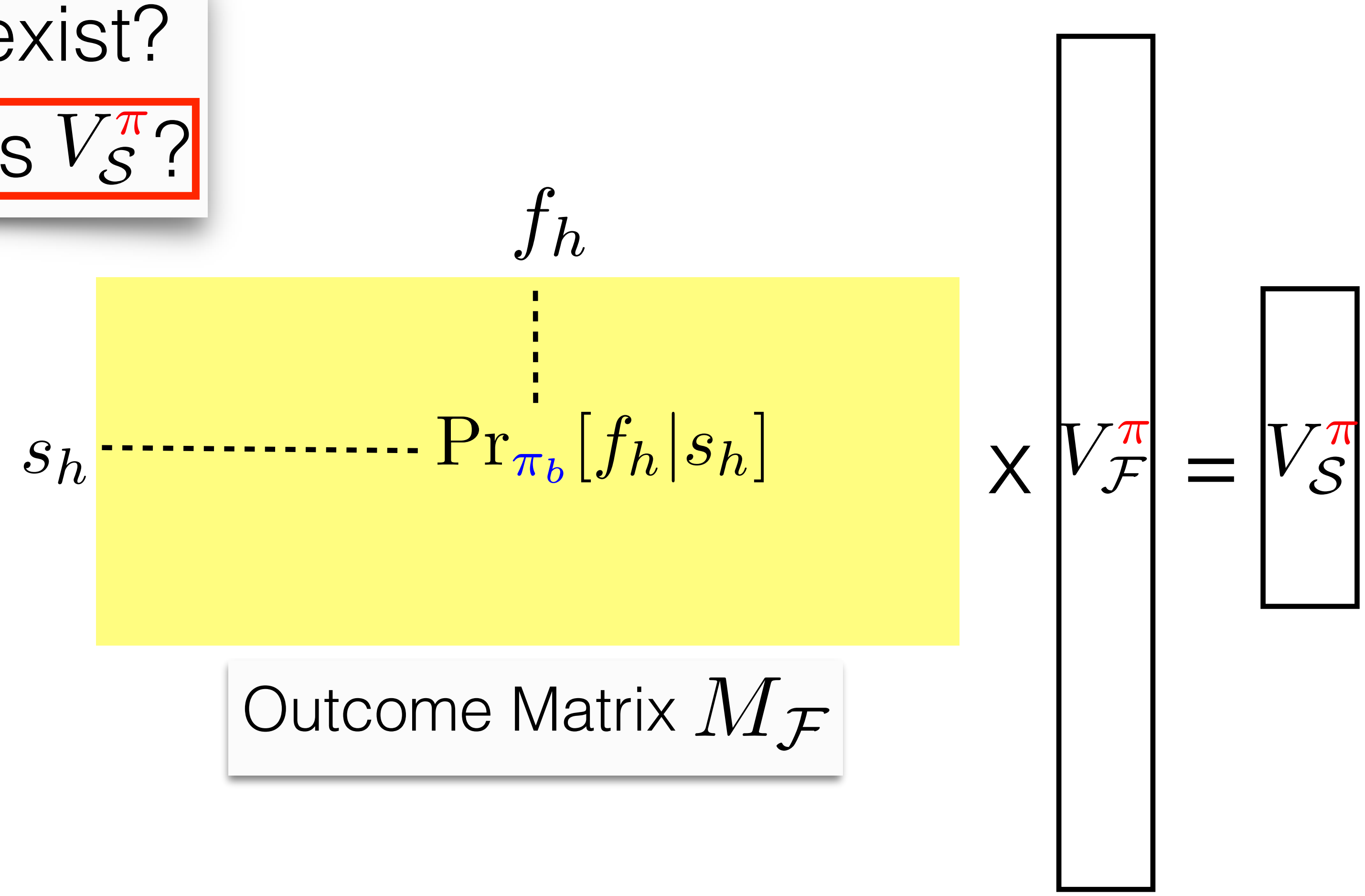
- Define: value function of latent state

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

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

1. Does (well-behaved) $V_{\mathcal{F}}^{\pi}$ even exist?
2. Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?





Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

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

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

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

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

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

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$	$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)\|s_h]$
Candidate	$V_{\mathcal{F}}$	$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h)\|s_h]$
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$		$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h) \mid s_h]$
Candidate	$V_{\mathcal{F}}$		$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h) \mid s_h]$
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$	=	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$		$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h) \vert s_h]$
Candidate	$V_{\mathcal{F}}$		$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h) \vert s_h]$
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$	=	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$
Error	$\sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h} \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\Delta_h V_{\mathcal{F}}]$ $V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})$		$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$	$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h) \mid s_h]$
Candidate	$V_{\mathcal{F}}$	$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h) \mid s_h]$
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$ **=**	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$
Error	$\sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h} \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\Delta_h V_{\mathcal{F}}]$ **=** $\Delta_h V_{\mathcal{F}}$: $V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})$	$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$		$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h) \vert s_h]$
Candidate	$V_{\mathcal{F}}$		$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h) \vert s_h]$
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$	=	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$
Error	$\sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h} \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}}]$ where $\Delta_h V_{\mathcal{F}} = V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})$	=	$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$
Learning objective			$\mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$	$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h) \mid s_h]$
Candidate	$V_{\mathcal{F}}$	$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h) \mid s_h]$
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$ **=**	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$
Error	$\sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h} \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\Delta_h V_{\mathcal{F}}]$ **=** $V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})$	$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$
Learning objective	$\sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_b} \left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\Delta_h V_{\mathcal{F}} \mid s_h]^2 \right]$ **=**	$\mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$		$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)\|s_h]$
Candidate	$V_{\mathcal{F}}$		$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h)\|s_h]$ ↓
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$	=	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$
Error	$\sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h}\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}}]$ ↓ $V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})$	=	$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$
Learning objective	$\sum_{h=1}^{H} \mathbb{E}_{s_h\sim\pi_b}\left[\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}}\|s_h]^2\right]$	= X	$\mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$		$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h)\vert s_h]$
Candidate	$V_{\mathcal{F}}$		$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h)\vert s_h]$
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$	=	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$
Error	$\sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h}\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}}]$ ↓ $V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})$	=	$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$
Learning objective	$\sum_{h=1}^{H} \mathbb{E}_{s_h\sim\pi_b}\left[\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}}\vert s_h]^2\right]$	= X	$\mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

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

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$		$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h) \mid s_h]$
Candidate	$V_{\mathcal{F}}$		$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h) \mid s_h]$ ↓
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$	=	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$
Error	$\sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h} \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}}]$ ↓ $V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})$	=	$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$
Learning objective	$\sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_b}\left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} \mid s_h]^2\right]$	= X	$\mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

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

$$= \sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim \pi_b}\left[\left(\sum_{s_h} \mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}}[\Delta_h V_{\mathcal{F}} \mid s_h] \cdot \Pr[s_h \mid \tau_h]\right)^2\right]$$

Does it work in the same way as $V_\mathcal{S}^\pi$?

	$V_\mathcal{F}^\pi$	$V_\mathcal{S}^\pi(s_h) = \mathbb{E}_{\pi_b}[V_\mathcal{F}^\pi(f_h) \mid s_h]$
Candidate	$V_\mathcal{F}$	$V_\mathcal{S}(s_h) := \mathbb{E}_{\pi_b}[V_\mathcal{F}(f_h) \mid s_h]$
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_\mathcal{F}(f_1)]$ =	$\mathbb{E}[V_\mathcal{S}(s_1)]$
Error	$\sum_{h=1}^H \mathbb{E}_{\substack{a_{1:h}\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_\mathcal{F}]$ = $\Delta_h V_\mathcal{F} = V_\mathcal{F}(f_h) - r_h - V_\mathcal{F}(f_{h+1})$	$\sum_{h=1}^H \mathbb{E}_{\pi}[V_\mathcal{S}(s_h) - r_h - V_\mathcal{S}(s_{h+1})]$
Learning objective	$\sum_{h=1}^H \mathbb{E}_{s_h\sim\pi_b}\left[\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_\mathcal{F} \mid s_h]^2\right]$, where $g(s_h) = \mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_\mathcal{F} \mid s_h]$ = X	$\mathbb{E}_{\pi_b}[(V_\mathcal{S}(s_h) - (\mathcal{T}^\pi V_\mathcal{S})(s_h))^2]$

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

$$= \sum_{h=1}^H \mathbb{E}_{\tau_h\sim\pi_b}\left[\left(\sum_{s_h} g(s_h) \cdot \Pr[s_h \mid \tau_h]\right)^2\right]$$

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

	$V_{\mathcal{F}}^{\pi}$		$V_{\mathcal{S}}^{\pi}(s_h) = \mathbb{E}_{\pi_b}[V_{\mathcal{F}}^{\pi}(f_h) \mid s_h]$
Candidate	$V_{\mathcal{F}}$		$V_{\mathcal{S}}(s_h) := \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_h) \mid s_h]$
Prediction	$J(\pi) \approx \mathbb{E}_{\pi_b}[V_{\mathcal{F}}(f_1)]$	=	$\mathbb{E}[V_{\mathcal{S}}(s_1)]$
Error	$\sum_{h=1}^{H} \mathbb{E}_{\substack{a_{1:h}\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}}]$, $\Delta_h V_{\mathcal{F}} = V_{\mathcal{F}}(f_h) - r_h - V_{\mathcal{F}}(f_{h+1})$	=	$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$
Learning objective	$\sum_{h=1}^{H} \mathbb{E}_{s_h\sim\pi_b}\left[\underbrace{\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}} \mid s_h]}_{g(s_h)}{}^2\right]$	= X	$\mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

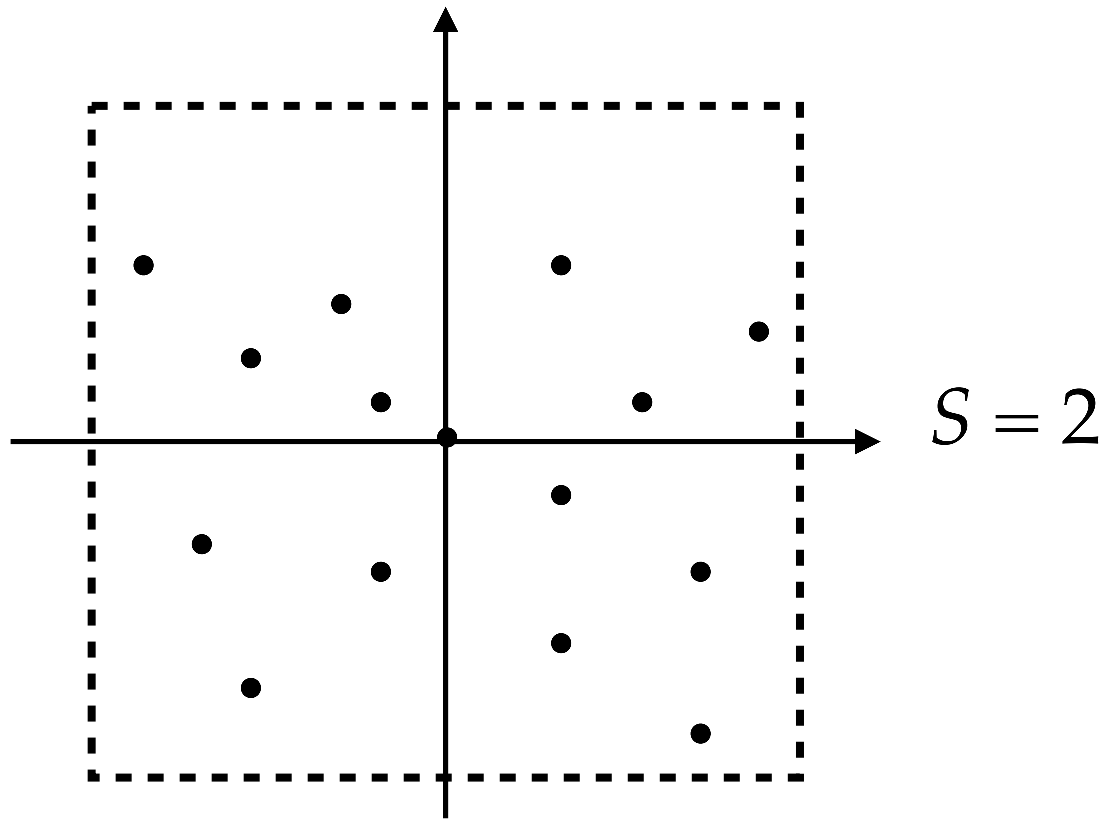
$$\sum_{h=1}^{H} \mathbb{E}_{\tau_h\sim\pi_b}\left[\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}} \mid \tau_h]^2\right]$$

$$= \sum_{h=1}^{H} \mathbb{E}_{\tau_h\sim\pi_b}\left[\left(\sum_{s_h} \underbrace{\mathbb{E}_{\substack{a_h\sim\pi \\ a_{h+1:H}\sim\pi_b}}[\Delta_h V_{\mathcal{F}} \mid s_h]}_{g(s_h)} \cdot \Pr[s_h \mid \tau_h]\right)^2\right]$$

belief state

linear measure

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?



$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$

Learning objective $\sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_b} \left[g(s_h)^2 \right]$

$\left[\mathbb{E}_{a_h \sim \pi, \, a_{h+1:H} \sim \pi_b} [\Delta_h V_{\mathcal{F}} | s_h]^2 \right] = \mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

X

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

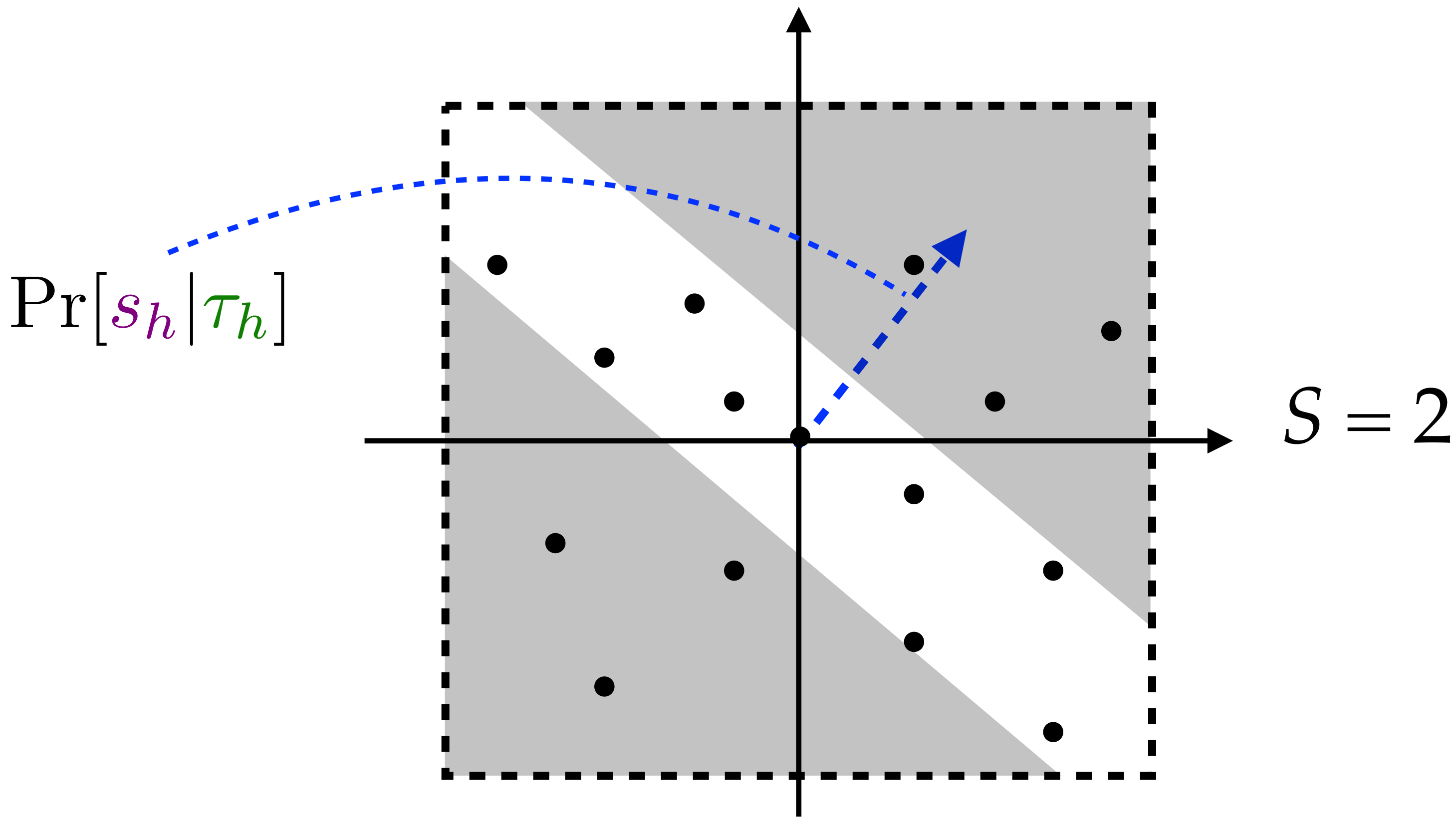
$= \sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim \pi_b} \left[\left(\sum_{s_h} g(s_h) \cdot \Pr[s_h | \tau_h] \right)^2 \right]$

$\mathbb{E}_{a_h \sim \pi, \, a_{h+1:H} \sim \pi_b} [\Delta_h V_{\mathcal{F}} | s_h]$

belief state

linear measure

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?



$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$

Learning objective $\sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_b} \left[g(s_h) \right]^2$ = $\mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

X

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

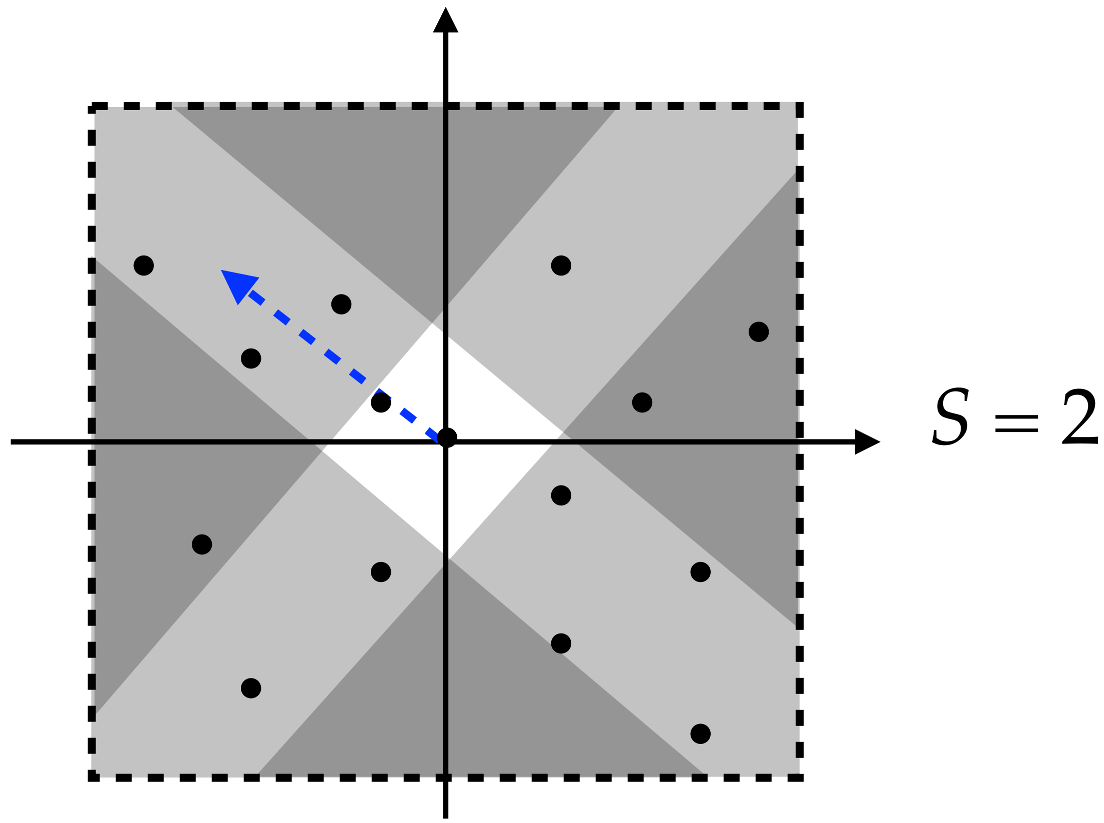
$= \sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim \pi_b} \left[\left(\sum_{s_h} g(s_h) \cdot \Pr[s_h|\tau_h] \right)^2 \right]$

belief state

linear measure

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

$\Pr[s_h|\tau_h]$



$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$

Learning objective $\sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_b} \left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\Delta_h V_{\mathcal{F}} | s_h]^2 \right] = \mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

$g(s_h)$ X

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

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

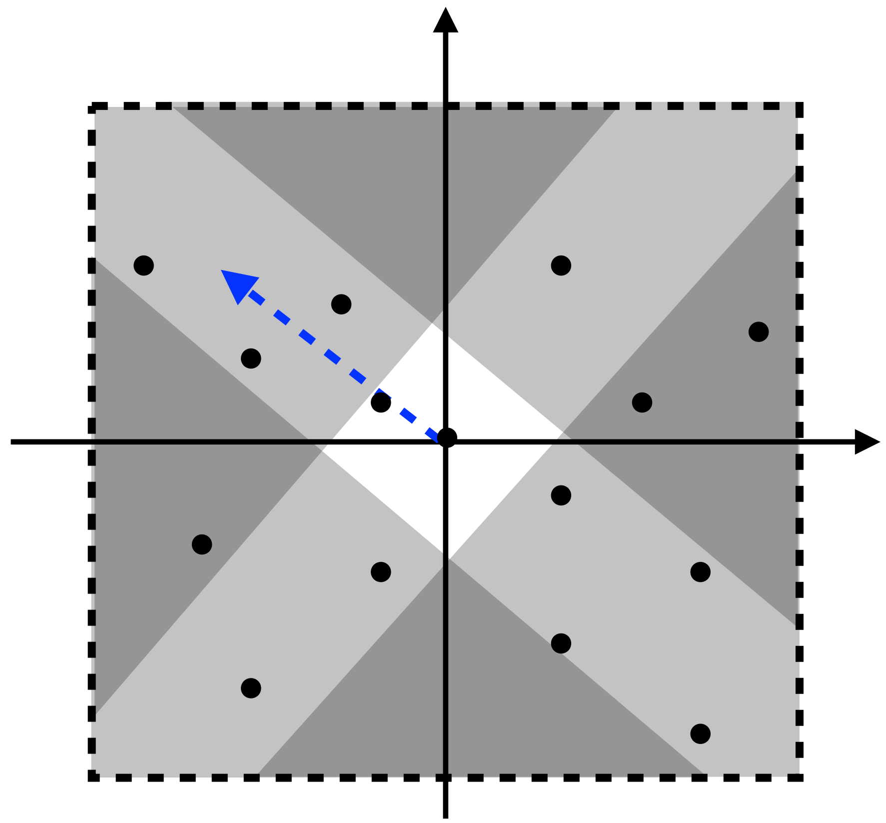
$g(s_h)$

belief state

linear measure

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

$\Pr[s_h|\tau_h]$



$S = 2$

$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$

$= \sum_{h=1}^{H} \sum_{s_h} g(s_h) d^{\pi}(s_h)$

Learning objective $\sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_b} \left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\Delta_h V_{\mathcal{F}} | s_h]^2 \right]$ = $\mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

$g(s_h)$

X

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

belief state

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

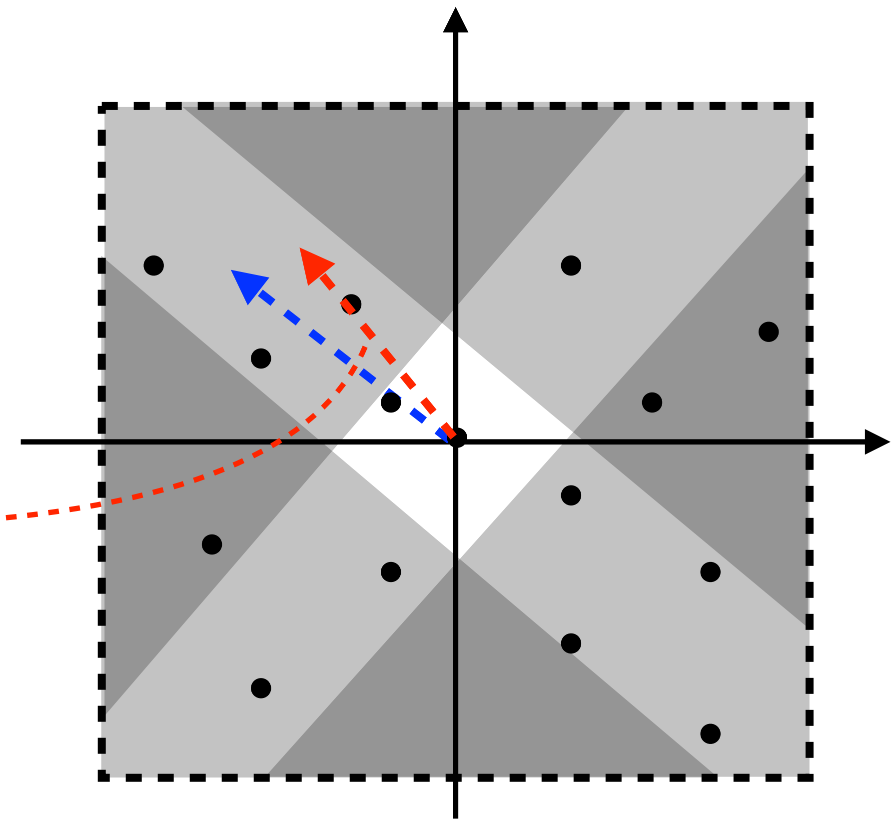
$g(s_h)$

linear measure

Does it work in the same way as $V_{\mathcal{S}}^{\pi}$?

$\Pr[s_h|\tau_h]$

$d^{\pi}(s_h)$



$S = 2$

$\sum_{h=1}^{H} \mathbb{E}_{\pi}[V_{\mathcal{S}}(s_h) - r_h - V_{\mathcal{S}}(s_{h+1})]$

$= \sum_{h=1}^{H} \sum_{s_h} g(s_h) d^{\pi}(s_h)$

Learning objective $\sum_{h=1}^{H} \mathbb{E}_{s_h \sim \pi_b} \left[\mathbb{E}_{\substack{a_h \sim \pi \\ a_{h+1:H} \sim \pi_b}} [\Delta_h V_{\mathcal{F}} | s_h]^2 \right]$ $g(s_h)$ $= \mathbb{E}_{\pi_b}[(V_{\mathcal{S}}(s_h) - (\mathcal{T}^{\pi} V_{\mathcal{S}})(s_h))^2]$

X

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

belief state

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

linear measure

$$= \sum_{h=1}^{H} \sum_{s_h} g(s_h) d^{\pi}(s_h)$$

cover

$$= \sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim \pi_b} \left[\left(\sum_{s_h} g(s_h) \cdot \Pr[s_h | \tau_h] \right)^2 \right]$$

linear measure

$$\sum_{h=1}^{H} \sum_{s_h} g(s_h) d^{\pi}(s_h)$$

cover

$$\sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim \pi_b} \left[\left(\sum_{s_h} g(s_h) \cdot \Pr[s_h | \tau_h] \right)^2 \right]$$

$$\mathbb{E}_{a_h \sim \pi, a_{h+1:H} \sim \pi_b} [\Delta_h V_{\mathcal{F}} | s_h]$$

linear measure

Theorem (Informal): Assume

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

$$\sum_{h=1}^{H}\sum_{s_h} g(s_h) d^{\pi}(s_h)$$

cover

$$\sum_{h=1}^{H}\mathbb{E}_{\tau_h\sim\pi_b}\left[\left(\sum_{s_h} g(s_h)\cdot \Pr[s_h|\tau_h]\right)^2\right]$$

linear measure

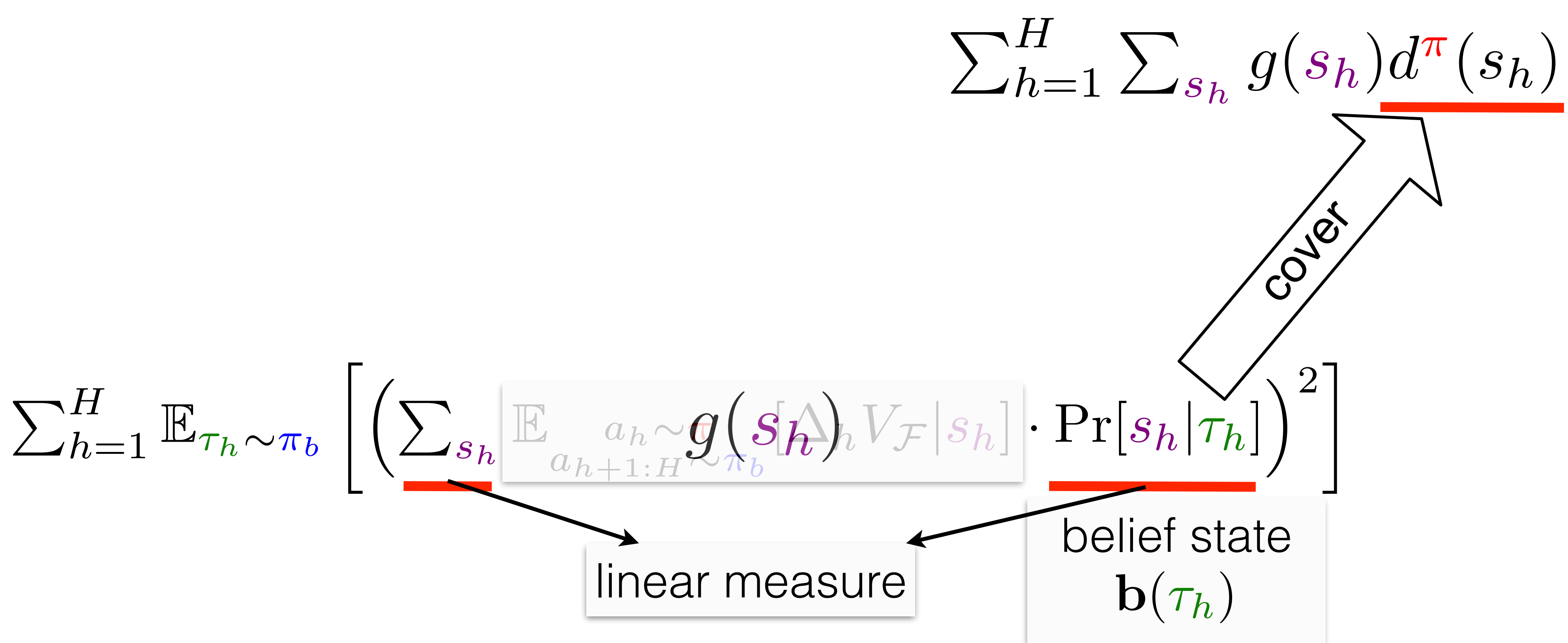
belief state $\mathbf{b}(\tau_h)$

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 $\mathcal{V}$)



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

$\mathbb{E}_{\tau_h \sim \pi}[\Pr[s_h|\tau_h]]$

$$\sum_{h=1}^{H} \sum_{s_h} g(s_h) d^{\pi}(s_h)$$

cover

$$\sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim \pi_b}\left[\left(\sum_{s_h} g(s_h) \cdot \Pr[s_h|\tau_h]\right)^2\right]$$

$\mathbb{E}_{a_h \sim \pi, a_{h+1:H} \sim \pi_b}[(\Delta)_h V_{\mathcal{F}} | s_h]$

linear measure

belief state $\mathbf{b}(\tau_h)$

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

$\mathbb{E}_{\tau_h \sim \pi}[\Pr[s_h|\tau_h]]$

$$\sum_{h=1}^{H} \sum_{s_h} g(s_h) d^{\pi}(s_h)$$

cover

$$\sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim \pi_b}\left[\left(\sum_{s_h} g(s_h) \cdot \Pr[s_h|\tau_h]\right)^2\right]$$

$\mathbb{E}_{a_h \sim \pi, a_{h+1:H} \sim \pi_b}[\Delta_h V_{\mathcal{F}} | s_h]$

linear measure

belief state $\mathbf{b}(\tau_h)$

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 $\mathbf{b}(\tau_h)$: $\mathbb{E}_{\pi_b}\left[(d_h^{\pi}/d_h^{\pi_b})^2\right]$

$\mathbb{E}_{\tau_h\sim\pi}[\Pr[s_h|\tau_h]]$

$$\sum_{h=1}^{H} \sum_{s_h} g(s_h) d^{\pi}(s_h)$$

cover

$$\sum_{h=1}^{H} \mathbb{E}_{\tau_h\sim\pi_b}\left[\left(\sum_{s_h} g(s_h) \cdot \Pr[s_h|\tau_h]\right)^2\right]$$

$\mathbb{E}_{a_h\sim\pi, a_{h+1:H}\sim\pi_b}[\Delta_h V_{\mathcal{F}}|s_h]$

linear measure

belief state $\mathbf{b}(\tau_h)$

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

$\mathbb{E}_{\tau_h \sim \pi}[\Pr[s_h|\tau_h]]$

$$\sum_{h=1}^{H} \sum_{s_h} g(s_h) d^{\pi}(s_h)$$

cover

$$\sum_{h=1}^{H} \mathbb{E}_{\tau_h \sim \pi_b}\left[\left(\sum_{s_h} g(s_h) \cdot \Pr[s_h|\tau_h]\right)^2\right]$$

$\mathbb{E}_{a_h \sim \cdot, a_{h+1:H} \sim \pi_b}[\Delta_h V_{\mathcal{F}} | s_h]$

linear measure

belief state $\mathbf{b}(\tau_h)$

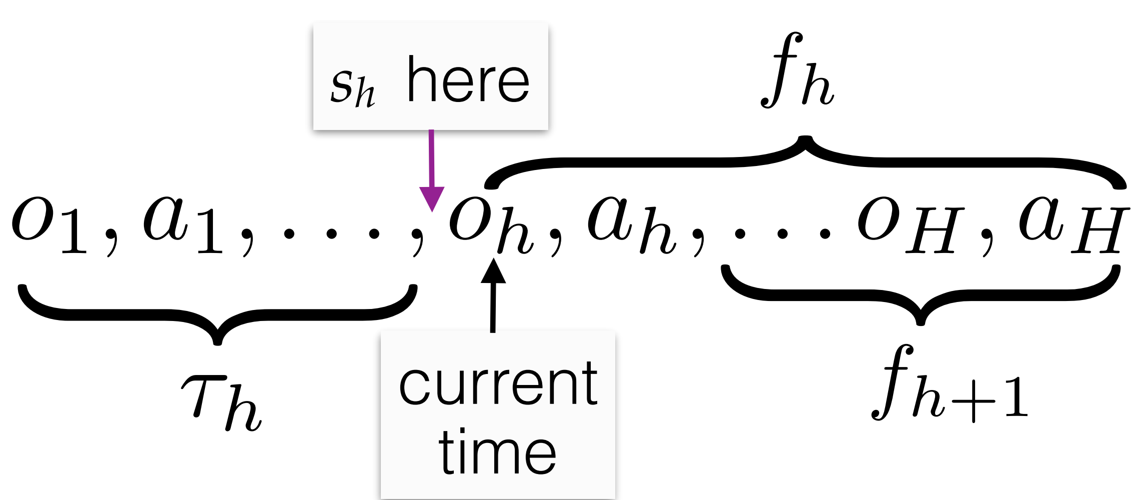
Future-Dependent Value Function

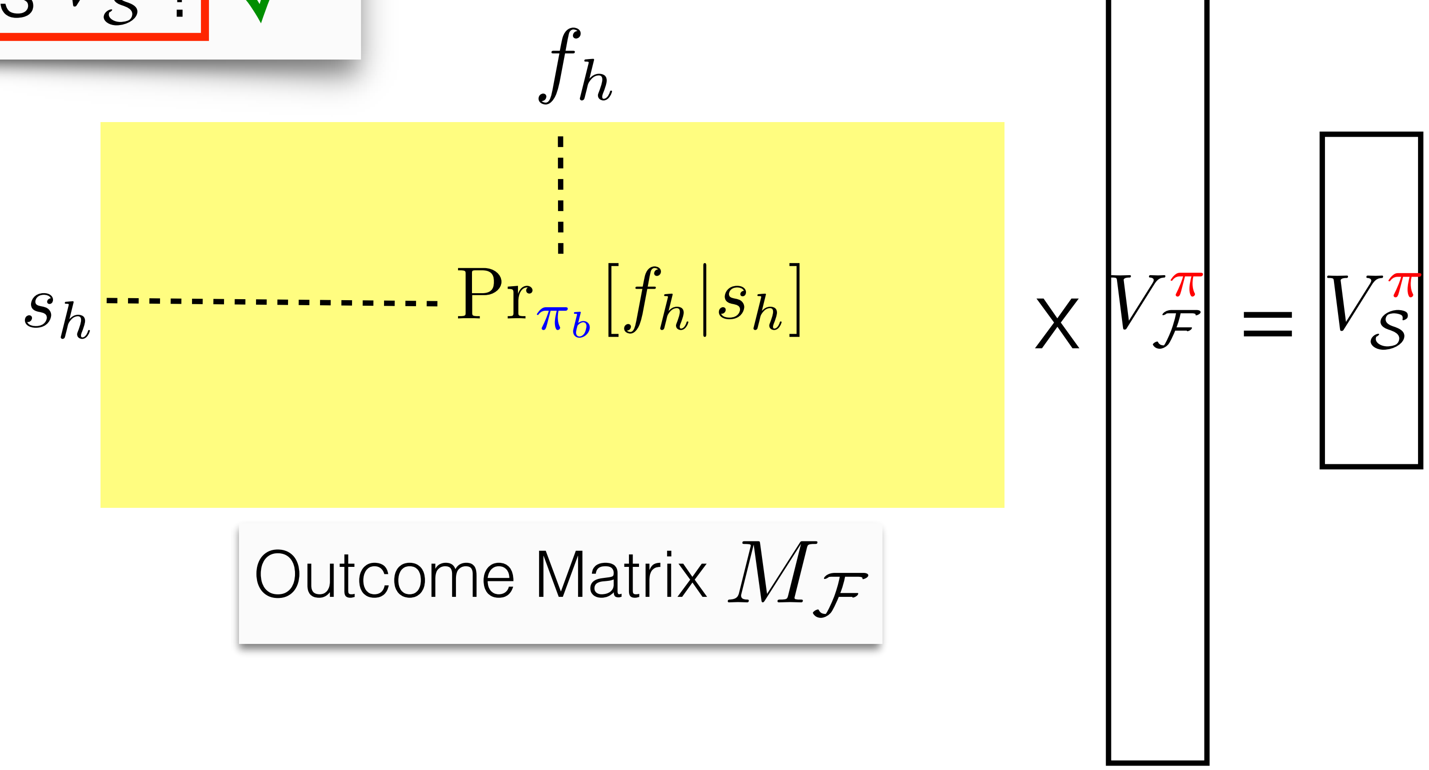
- Define: value function of latent state

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

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

1. Does (well-behaved) $V_{\mathcal{F}}^{\pi}$ even exist?
2. Does it work in the same way as $V_{\mathcal{S}}^{\pi}$? ✓





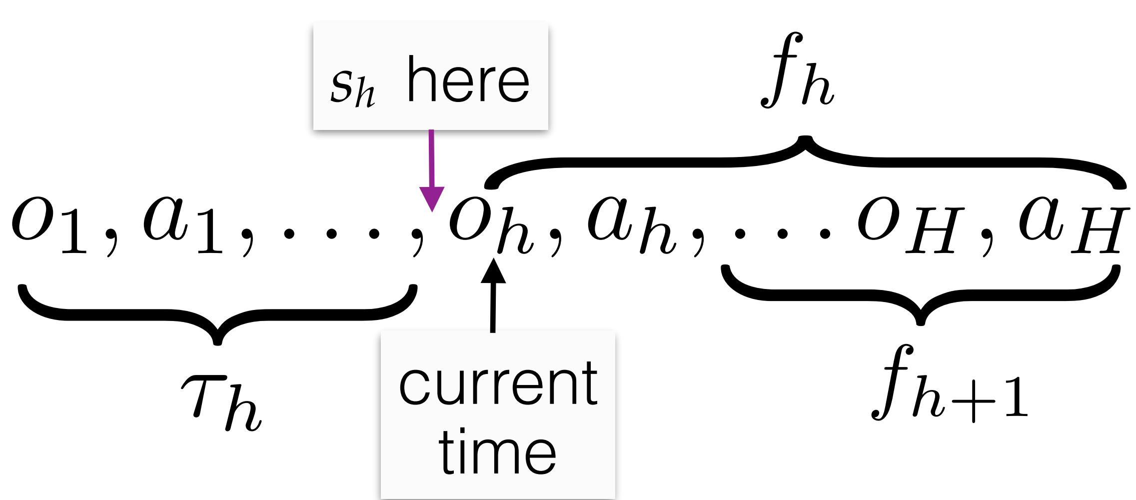
Future-Dependent Value Function

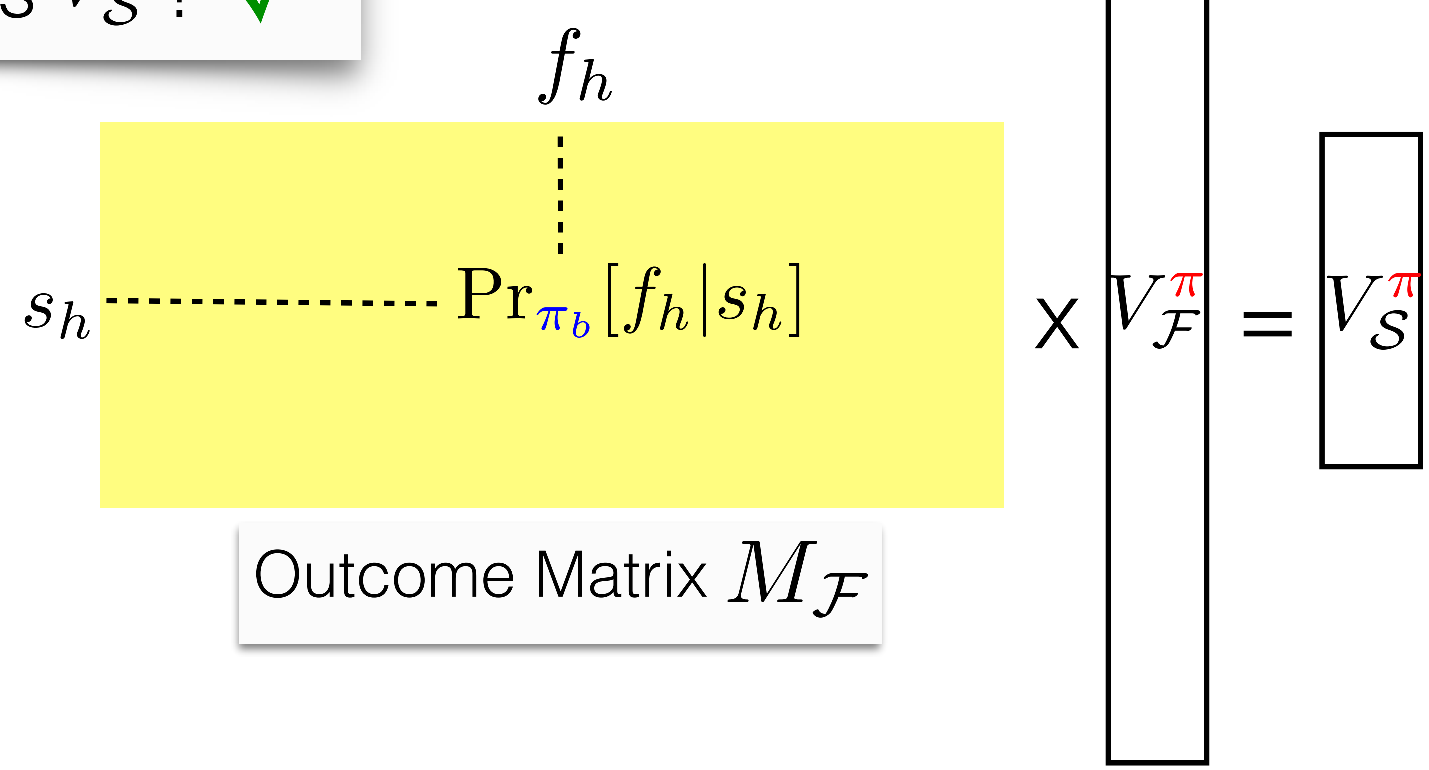
- Define: value function of latent state

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

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

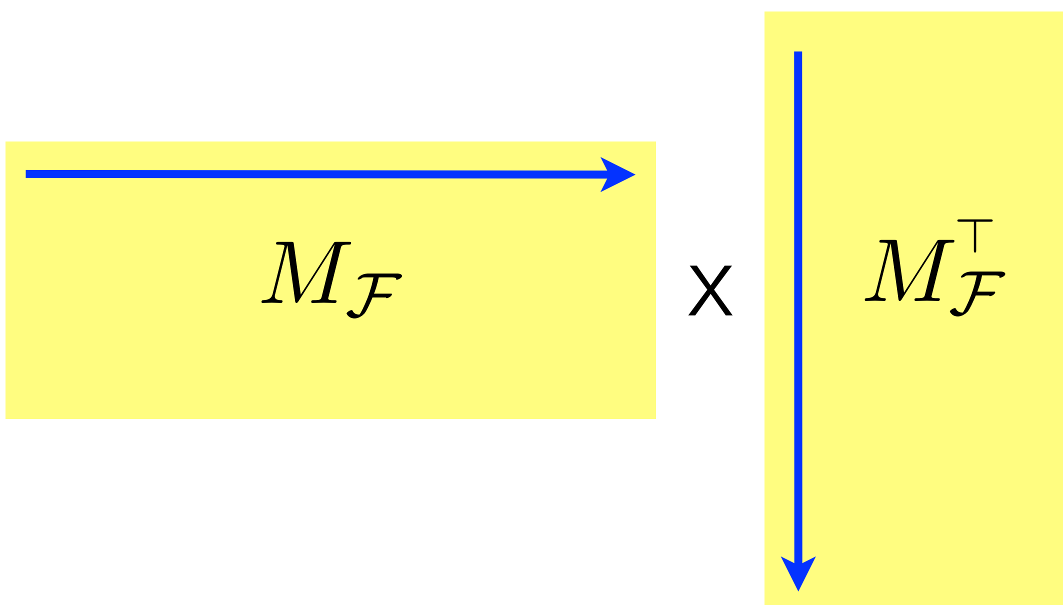
1. Does (well-behaved) $V_{\mathcal{F}}^{\pi}$ even exist?
2. Does it work in the same way as $V_{\mathcal{S}}^{\pi}$? ✓



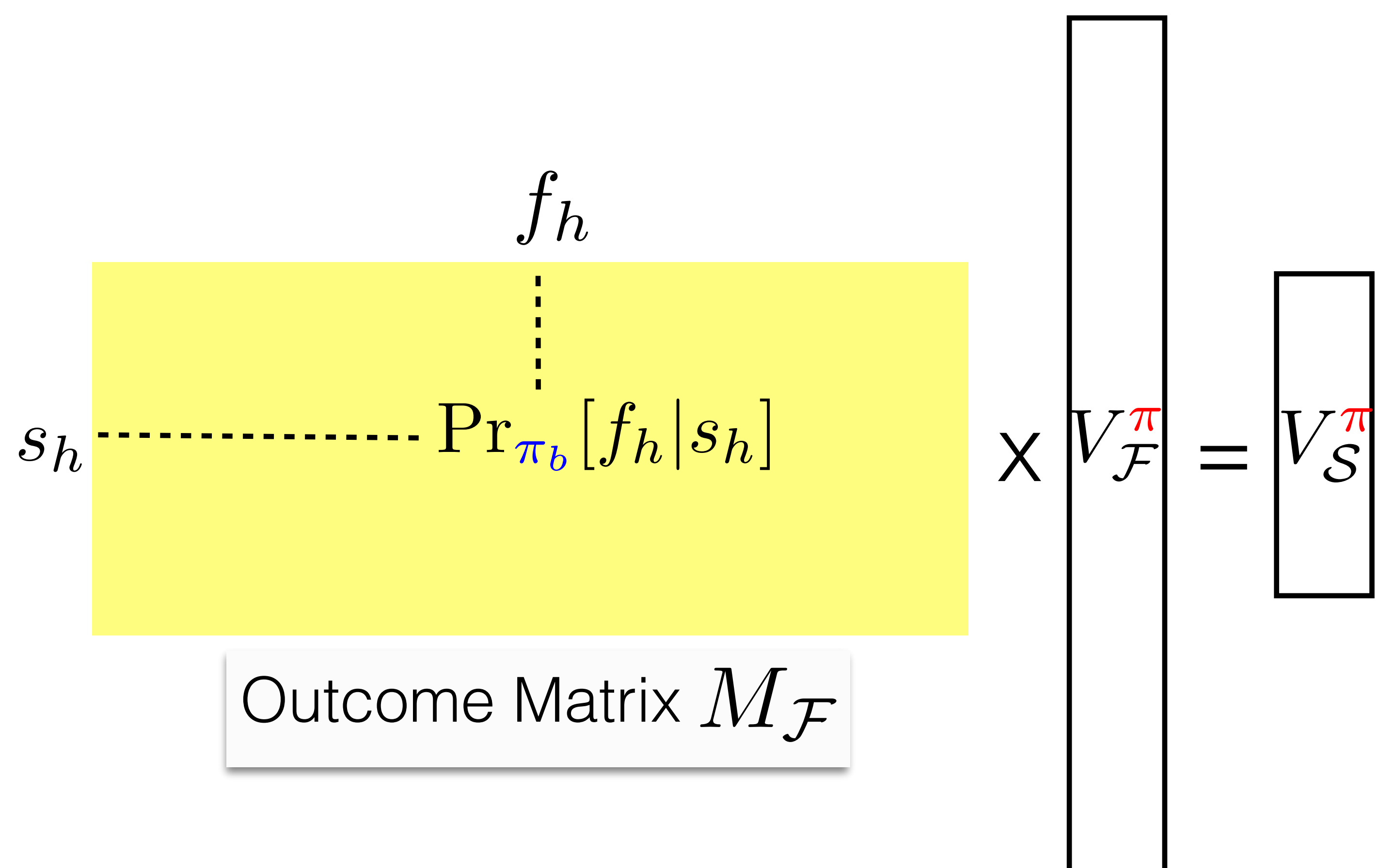


Constructing $V_{\mathcal{F}}^{\pi}$: pseudo-inverse

- Pay $1/\sigma_{\min}$ of outcome matrix, which is 1/min-eigenvalue of
  - Exponentially small when system is stochastic!
  - Problem: “linear regression" with exp. small covariates
  - Similar quantity appears in online RL (Liu et al’22)

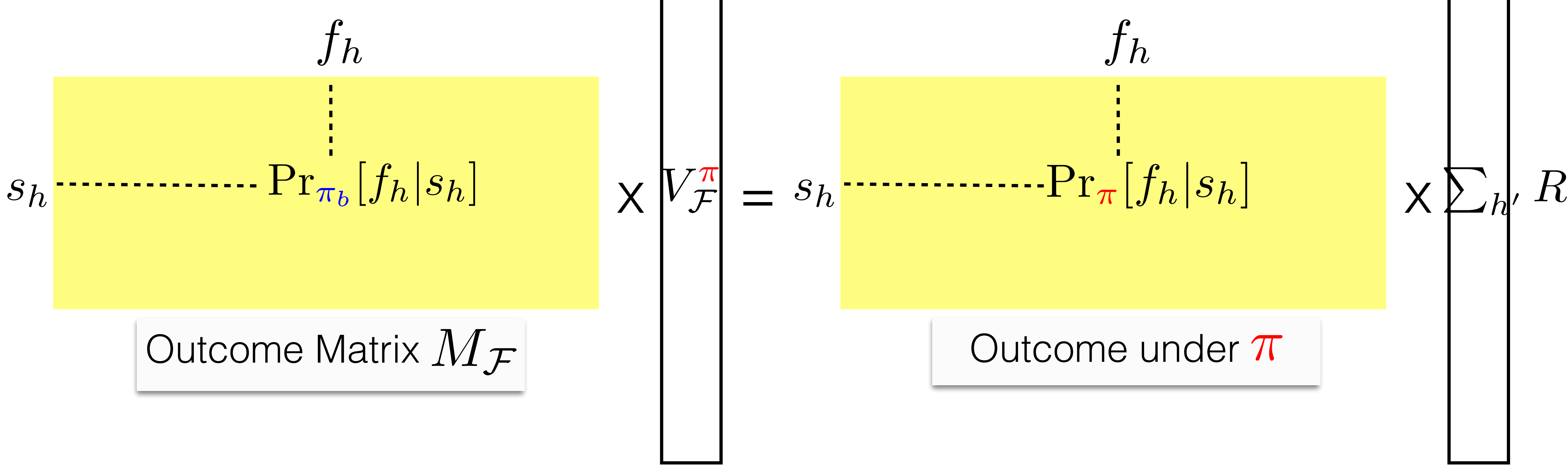


Diag entries: Simpson index
$\sum_{f_h} \Pr_{\pi_b}(f_h|s_h)^2$



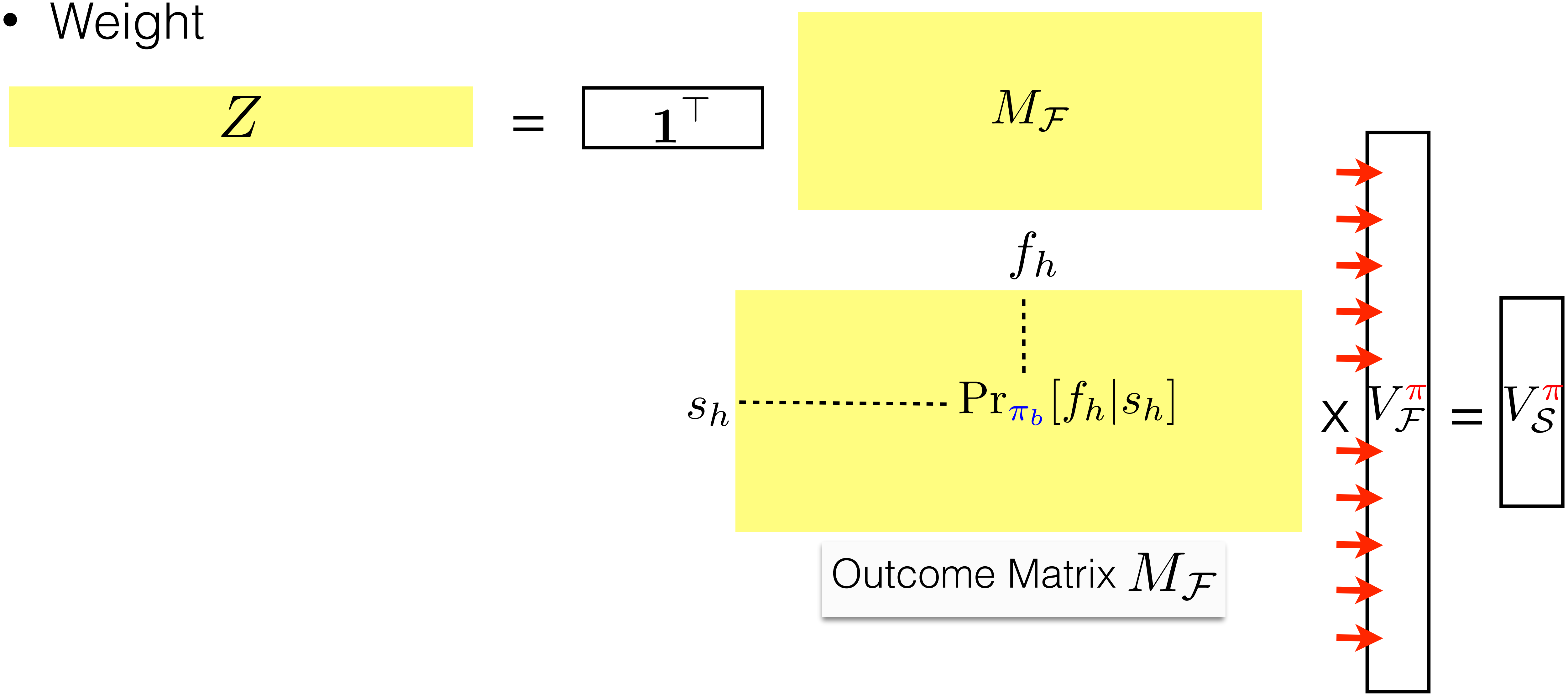
Constructing $V_{\mathcal{F}}^{\pi}$: reweighting

- General case
  - $V_{\mathcal{F}}^{\pi}(f_h) = \prod_{h'=h}^{H} \frac{\pi(a_{h'}|o_{h'})}{\pi_b(a_{h'}|o_{h'})} \left(\sum_{h'=h}^{H} R(o_{h'}, a_{h'}) \right)$!!



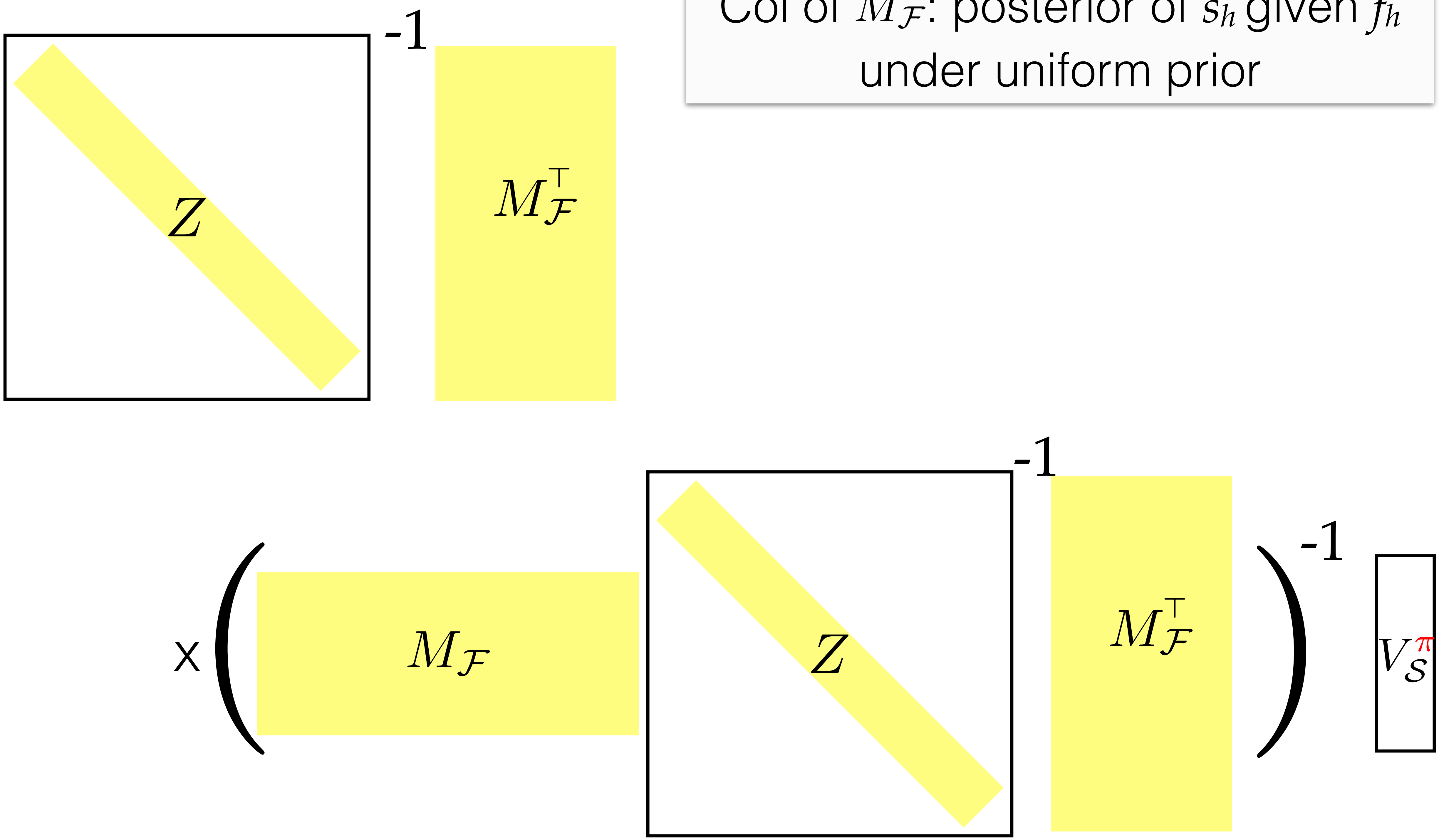
Constructing $V^{\pi}_{\mathcal{F}}$: *weighted* pseudo-inverse

- Pseudo-inverse minimizes L_2 norm (we want L_∞)
- L_2 norm treats all exponentially many coordinates equally — not informative
- Find solution that minimizes *weighted* L_2 norm
- Weight



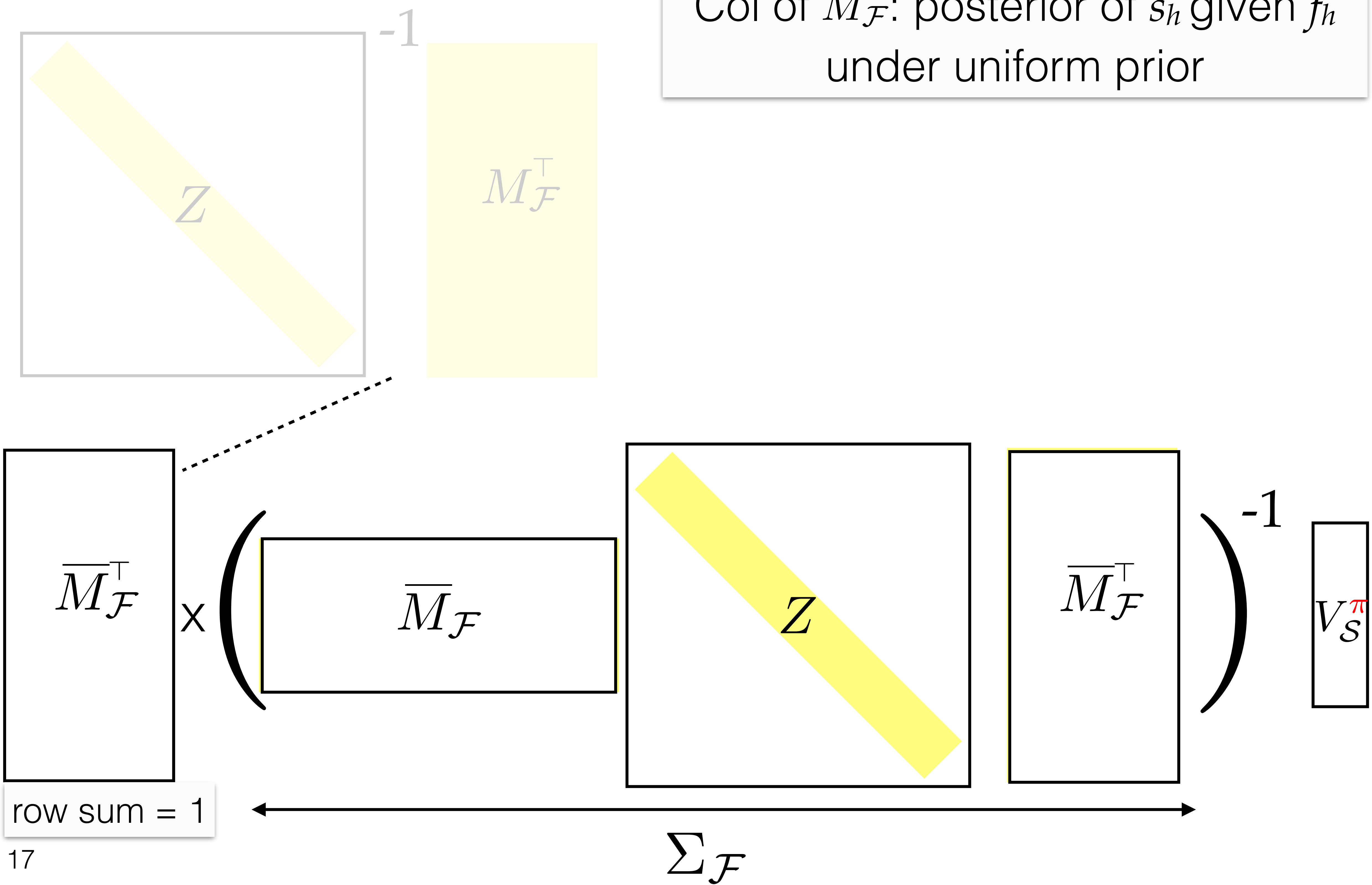
Constructing $V_{\mathcal{F}}^{\pi}$: *weighted* pseudo-inverse

Col of $\overline{M}_{\mathcal{F}}$: posterior of s_h given f_h under uniform prior



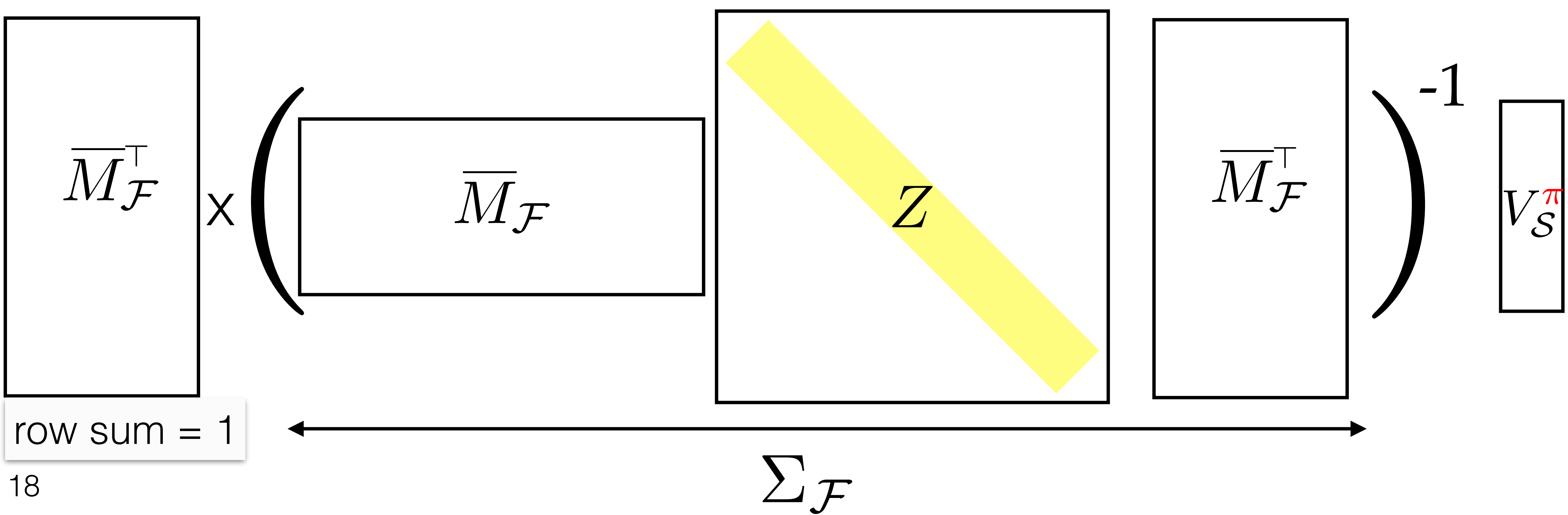
Constructing $V_{\mathcal{F}}^{\pi}$: *weighted* pseudo-inverse

Col of $\overline{M}_{\mathcal{F}}$: posterior of s_h given f_h under uniform prior



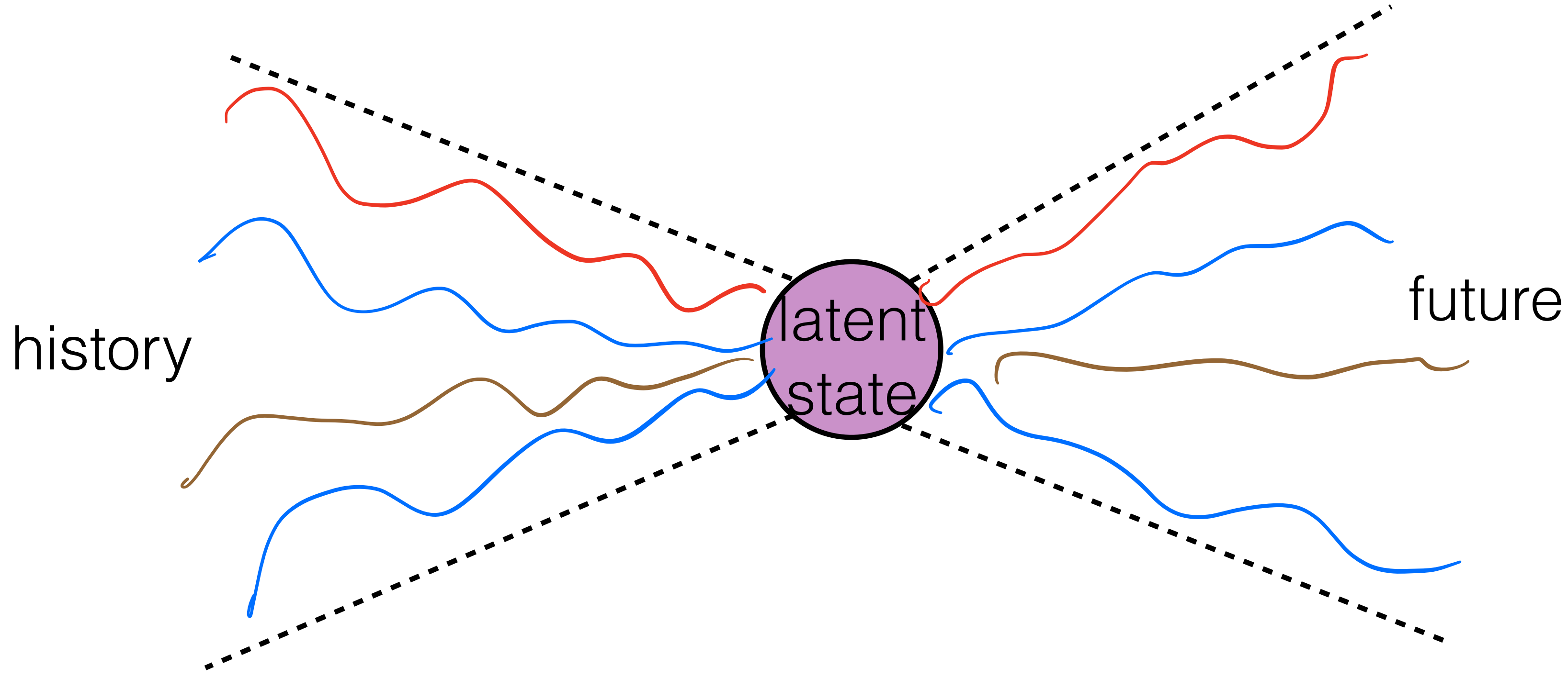
Properties of $\Sigma_{\mathcal{F}}$

- $\Sigma_{\mathcal{F}}$ is doubly stochastic
- When f_h reveals s_h, $\Sigma_{\mathcal{F}} = \mathbf{I}$
- More generally, confusion matrix of predicting s_h from f_h
- **Outcome Coverage**: $\|\Sigma_{\mathcal{F}}^{-1} V_{\mathcal{S}}^{\pi}\|_\infty \leqslant C_{\mathcal{F}}$ (not $\|\Sigma_{\mathcal{F}}^{-1/2} V_{\mathcal{S}}^{\pi}\|_2$)
    - $\Rightarrow \ \|V_{\mathcal{F}}^{\pi}\|_\infty \leqslant C_{\mathcal{F}}$



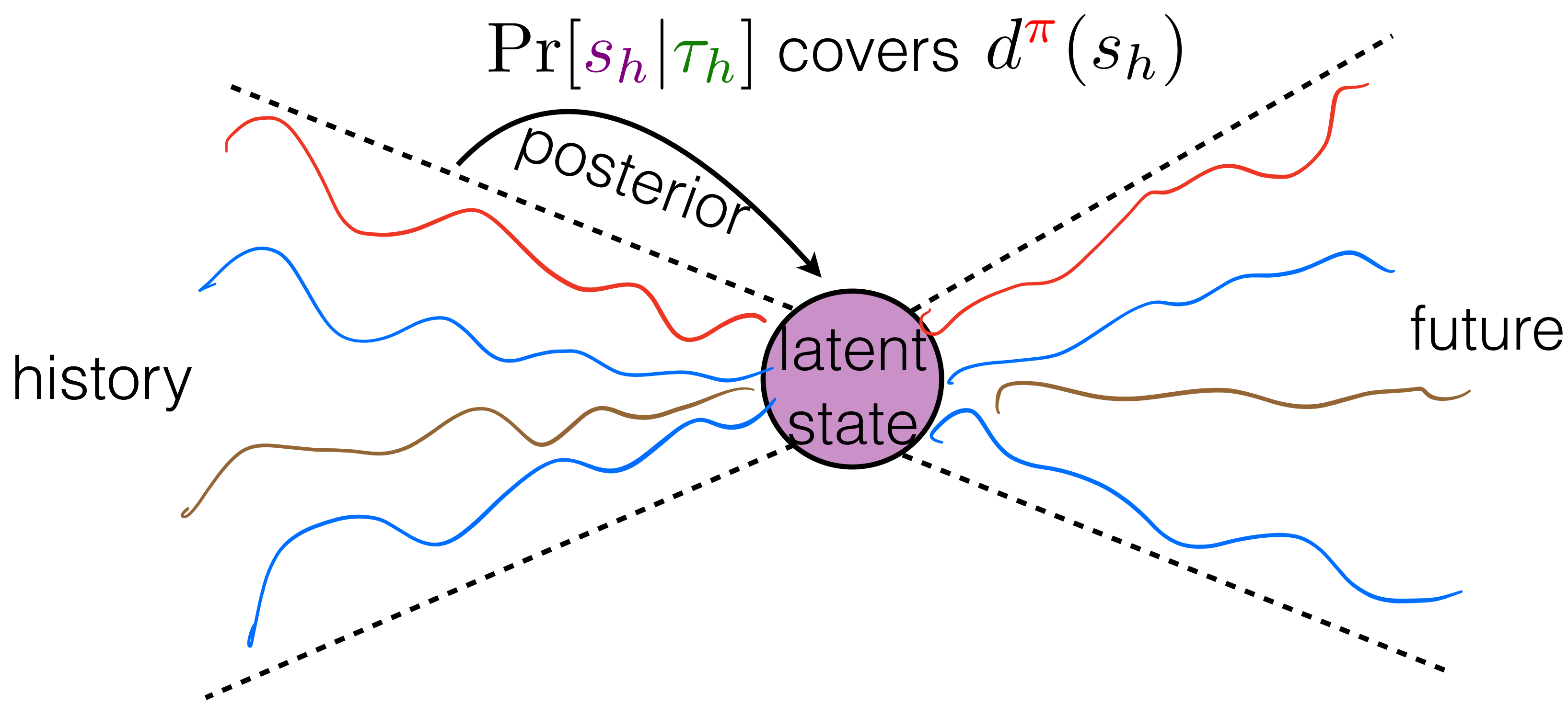
Conclusion

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



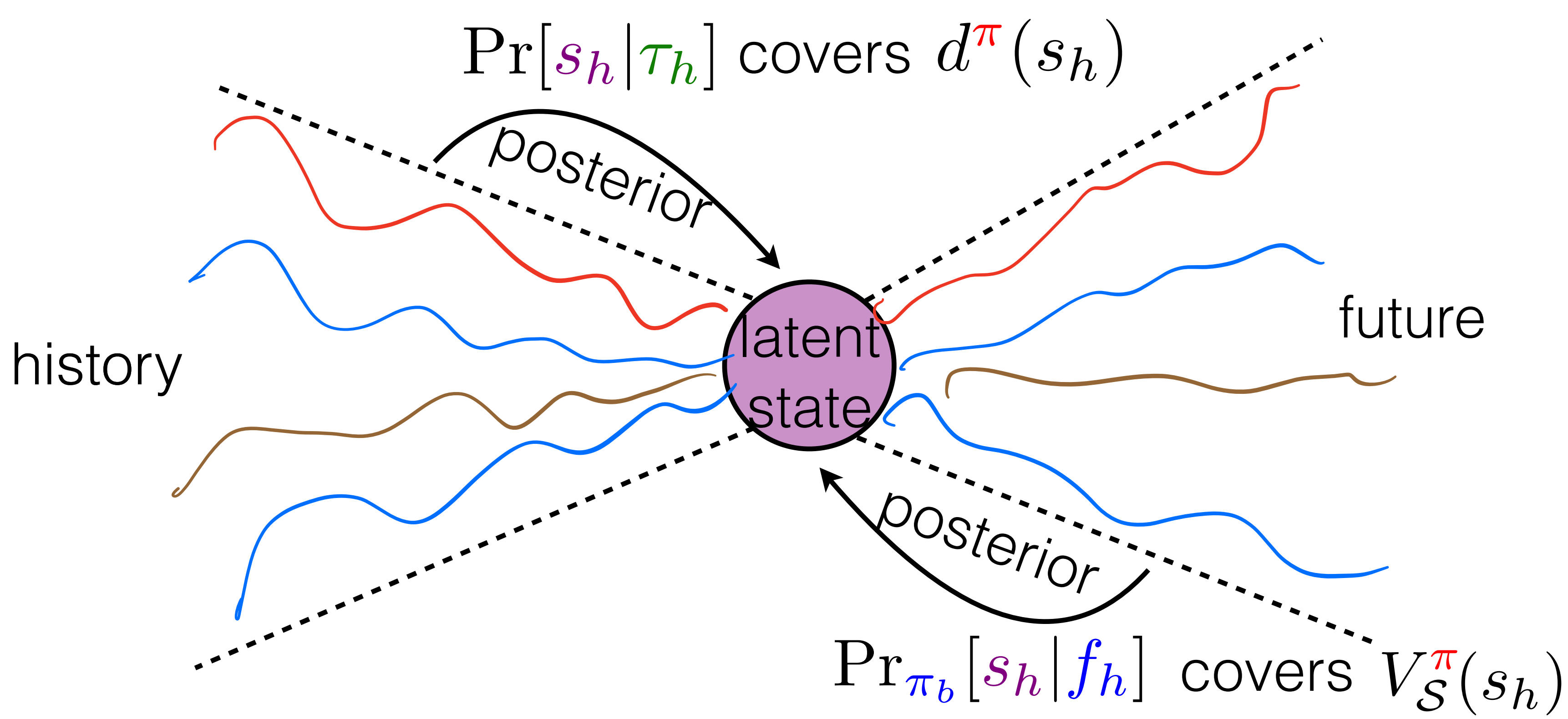
Conclusion

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



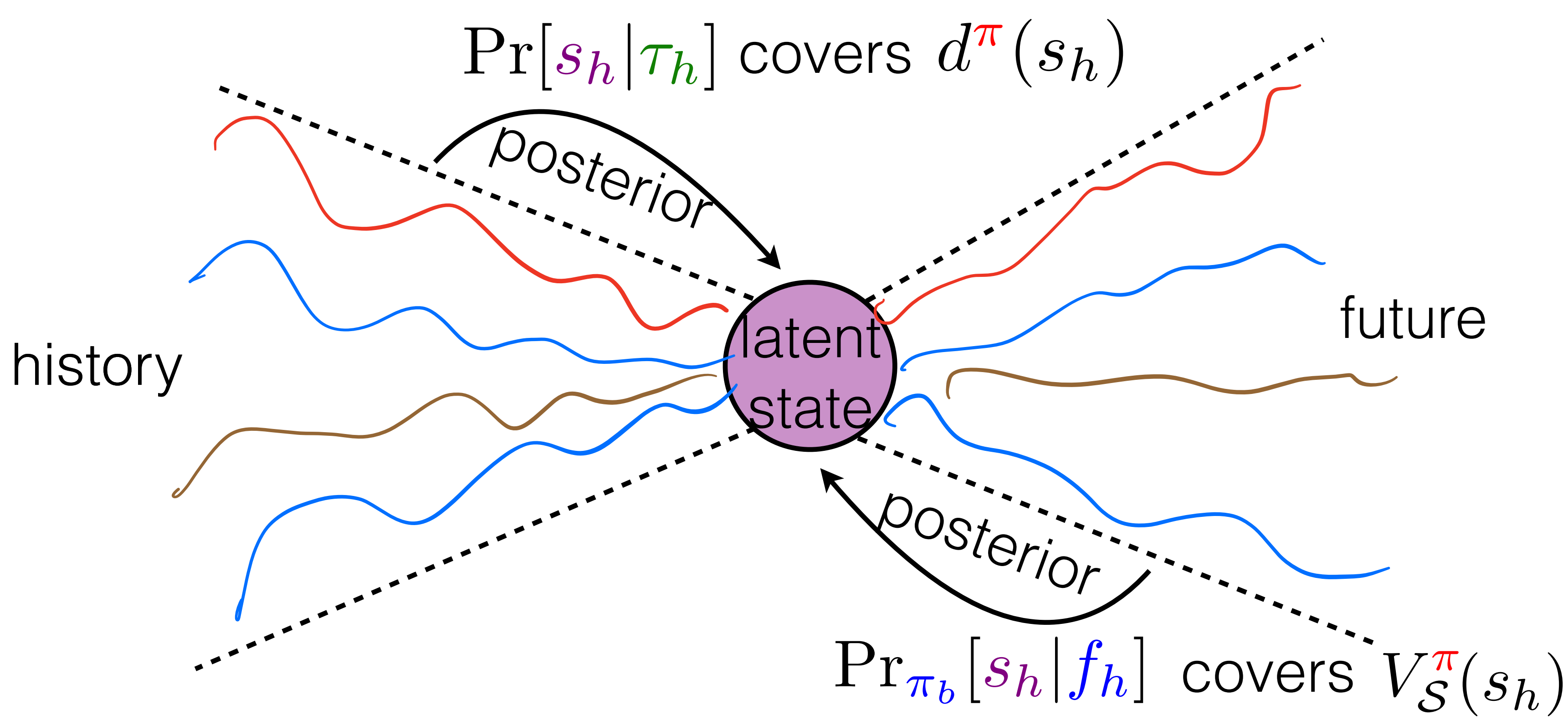
Conclusion

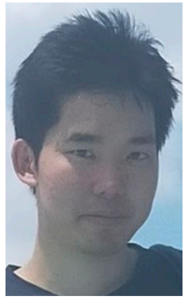
- Problem: OPE in POMDPs
- New framework: future-dependent value function $V_{\mathcal{F}}^{\pi}$
- New assumptions:
  - **Belief coverage =>** error transfer
  - **Outcome coverage =>** bounded $V_{\mathcal{F}}^{\pi}$



Conclusion

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





Masatoshi Uehara

Yuheng Zhang

Conclusion

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

Thank you! Questions?