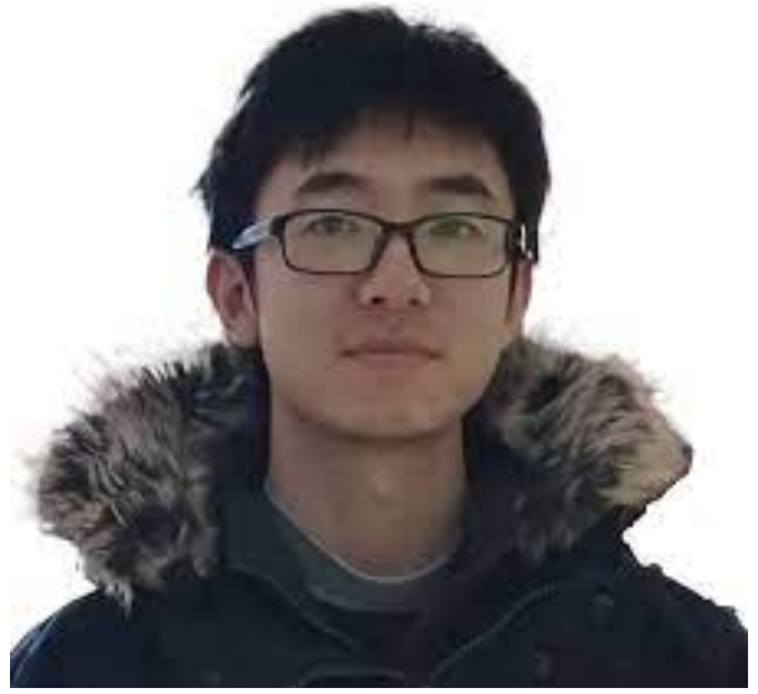
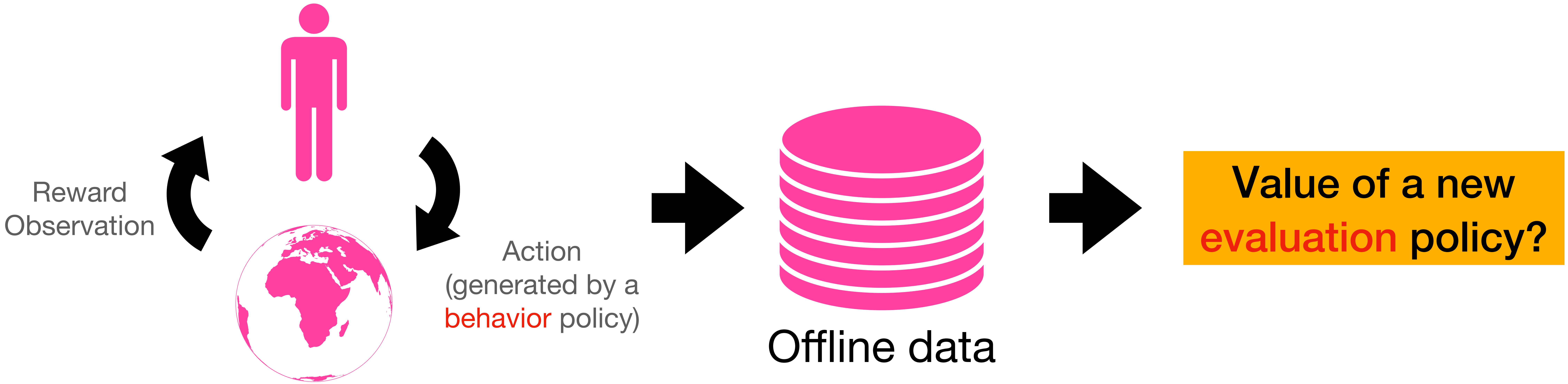
A Minimax Learning Approach to Off-Policy Evaluation in Confounded Partially Observable Markov Decision Processes

Changchun Shi (*) Masatoshi Uehara (*) Jiawei Huang Nan Jiang



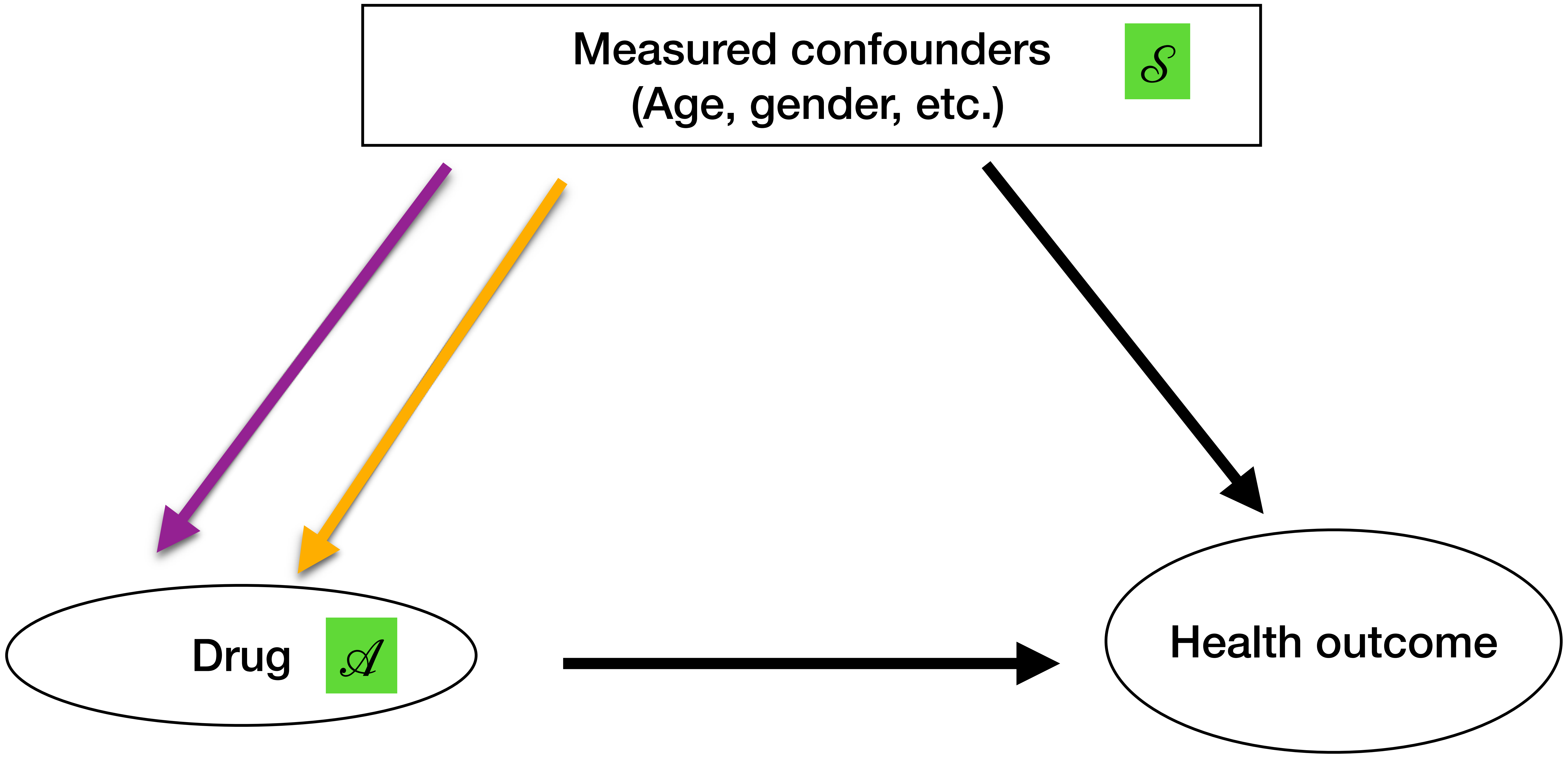
Pitfalls in OPE?

- Offline policy evaluation (OPE) is a fundamental task in offline RL. We want to estimate the value of evaluation policies from offline data.



- Most of papers assume behavior policies depend on observable quantities. But is it really true 😶 ?

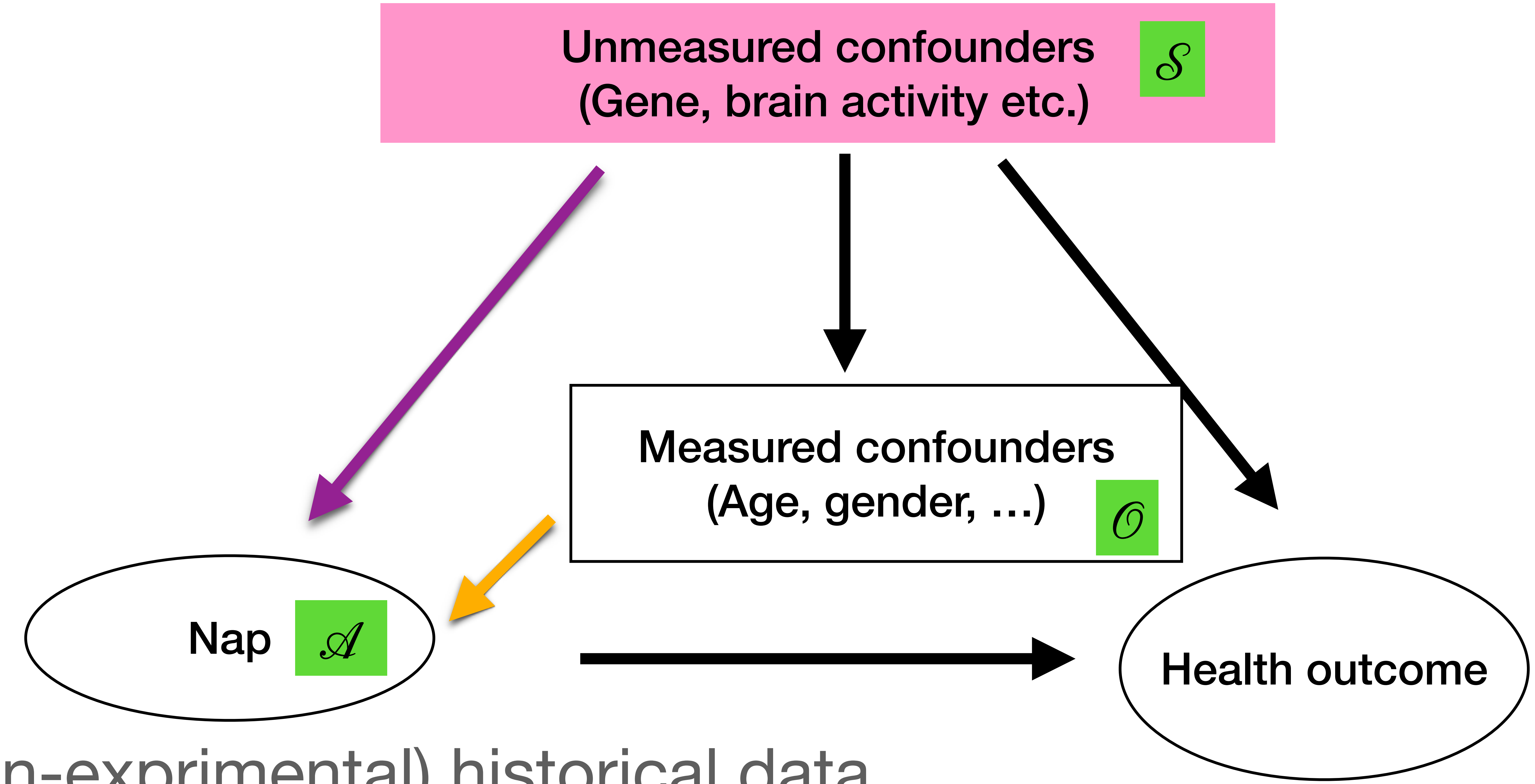
Standard OPE (without unmeasured confounders)



Consider clinical trials.

- (Behavior policies): Specified by only measured cofounders.
- (Evaluation policies): Depends on only measured cofounders.

OPE with unmeaseud confounders

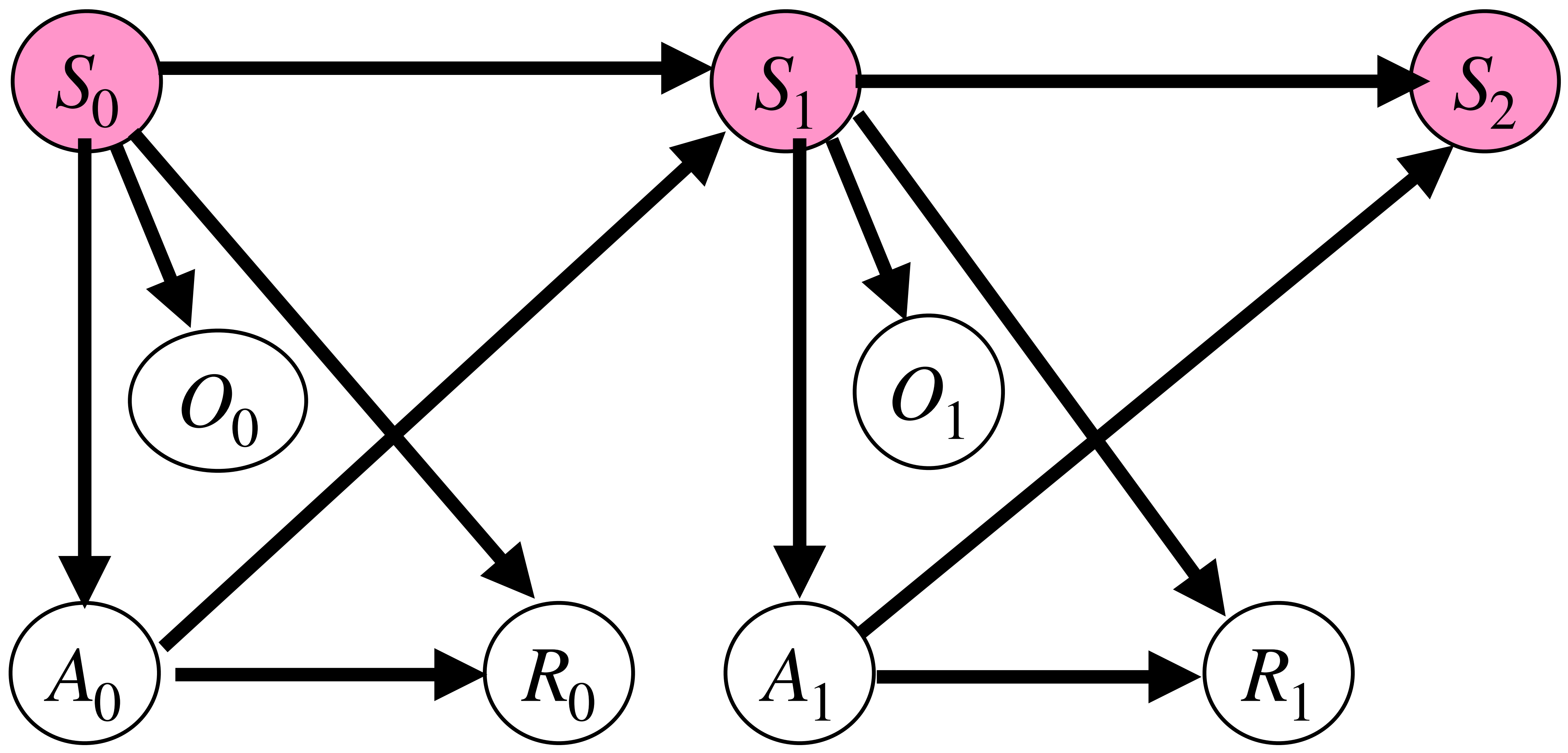


Consider (non-exprimental) historical data.

- (Behavior policies): “Nap” is affected by unmeasured variables.
- (Evaluation policies): Depend on observable variables.

Our contribution

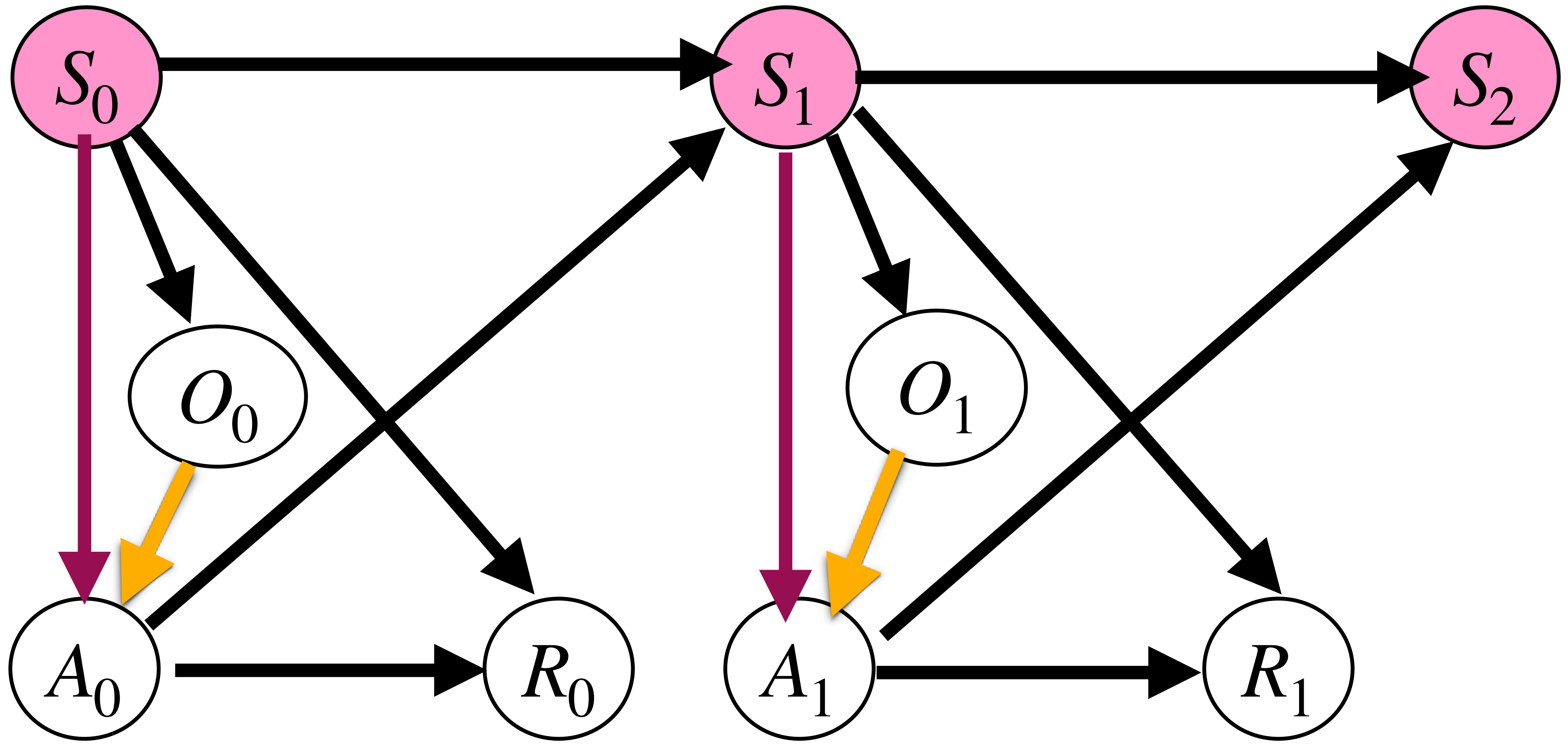
We consider OPE with unmeasured confounders in RL. (In confounded POMDPs)



1. We introduce novel value bridge functions.

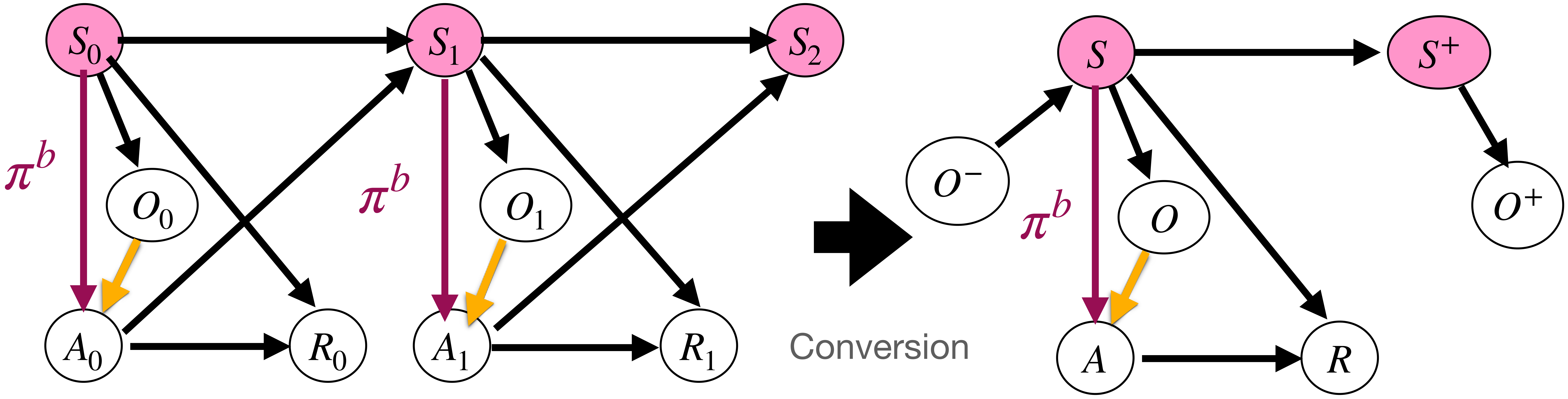
2. We propose OPE methods by estimating value bridge functions. Our proposal allows for any function approximation.

Confounded POMDPs



- Behavior policies $\pi^b : \mathcal{S} \to \Delta(\mathcal{A})$, evaluation policies $\pi^e : \mathcal{O} \to \Delta(\mathcal{A})$.
- Our goal is to estimate $J(\pi^e) = \mathbb{E}_{\pi^e}[\sum_{t=0}^{\infty} \gamma^t r_t]$.
- We have data $\mathcal{D} = \{(S_i, O_i, A_i, R_i, O_{i+1}, S_{i+1})\}$ following π^b. (S_i, S_{i+1} are unobservable)

Confounded POMDPs



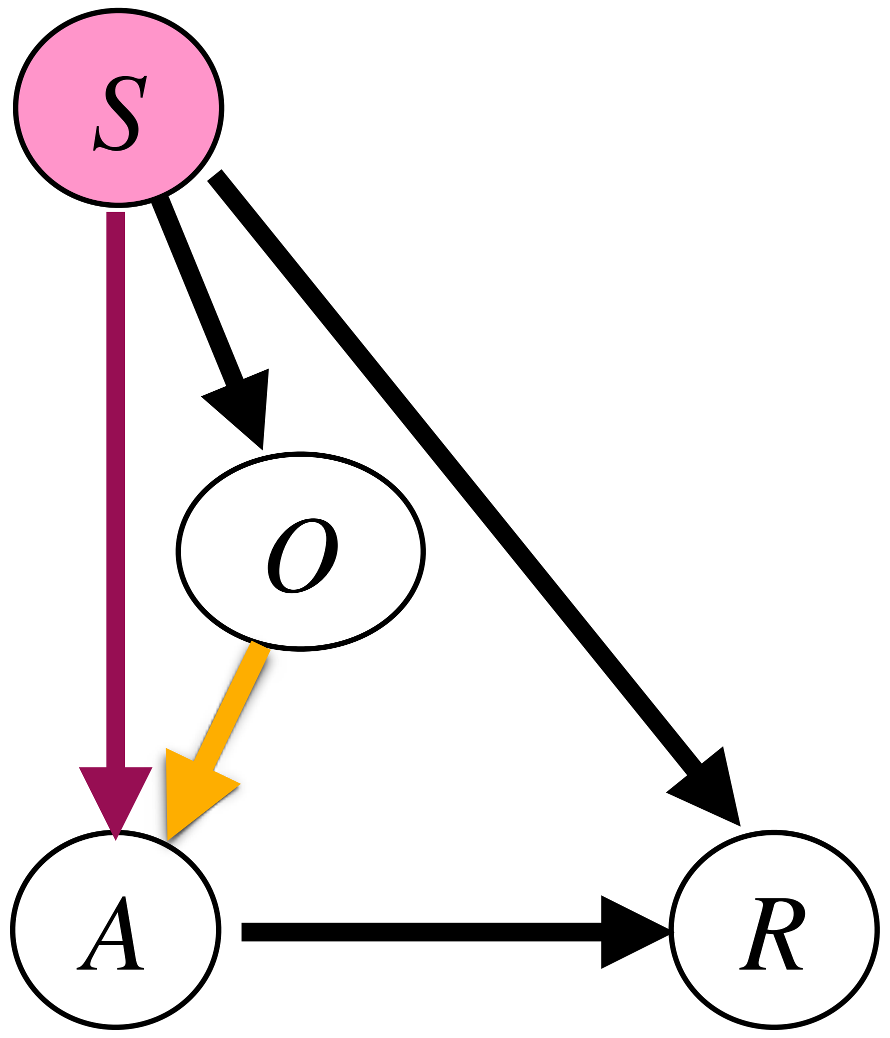
We have data

$$\mathscr{D} = \{(S_i, O_i, A_i, R_i, S_{i+1}, O_{i+1})\}\ .$$

(S_i, S_{i+1} are unobservable)

Equivalently, we have many tuples consisting of $(O^-, S, O, A, R, S^+, O^+)$.

Why diffcult?



Consider the contextual bandit setting.

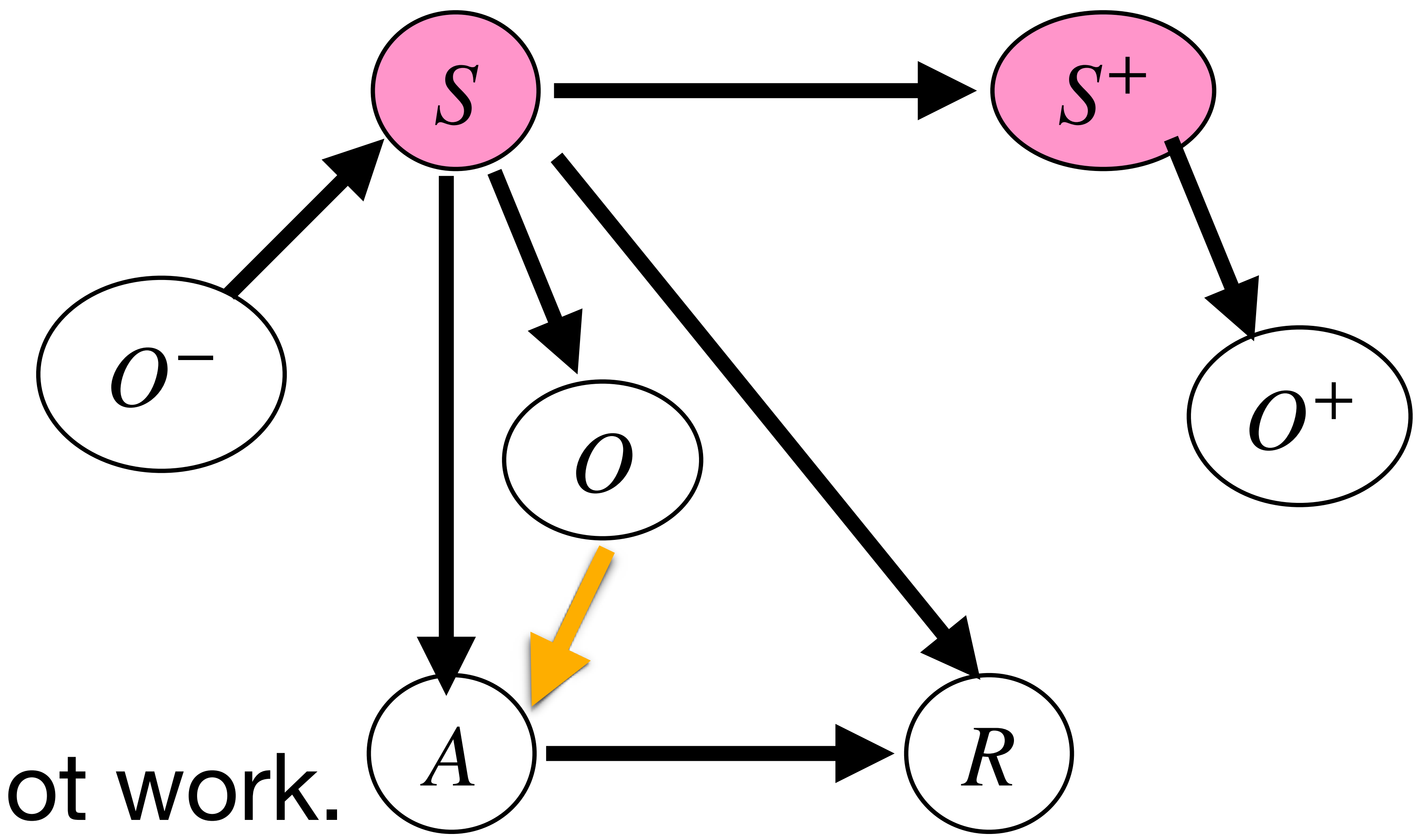
- IS estimator $\mathbb{E}_{\pi_b}\left[\frac{\pi^e(a \mid o)}{\pi^b(a \mid s)} \times r\right]$ does not work.

We cannot observe S.

- Direct method $\mathbb{E}_{\pi^b}\left[\sum_{a'} \pi^e(a' \mid o)\mathbb{E}[r \mid s, a']\right]$ does not work.

Why diffcult?

Consider the RL setting.



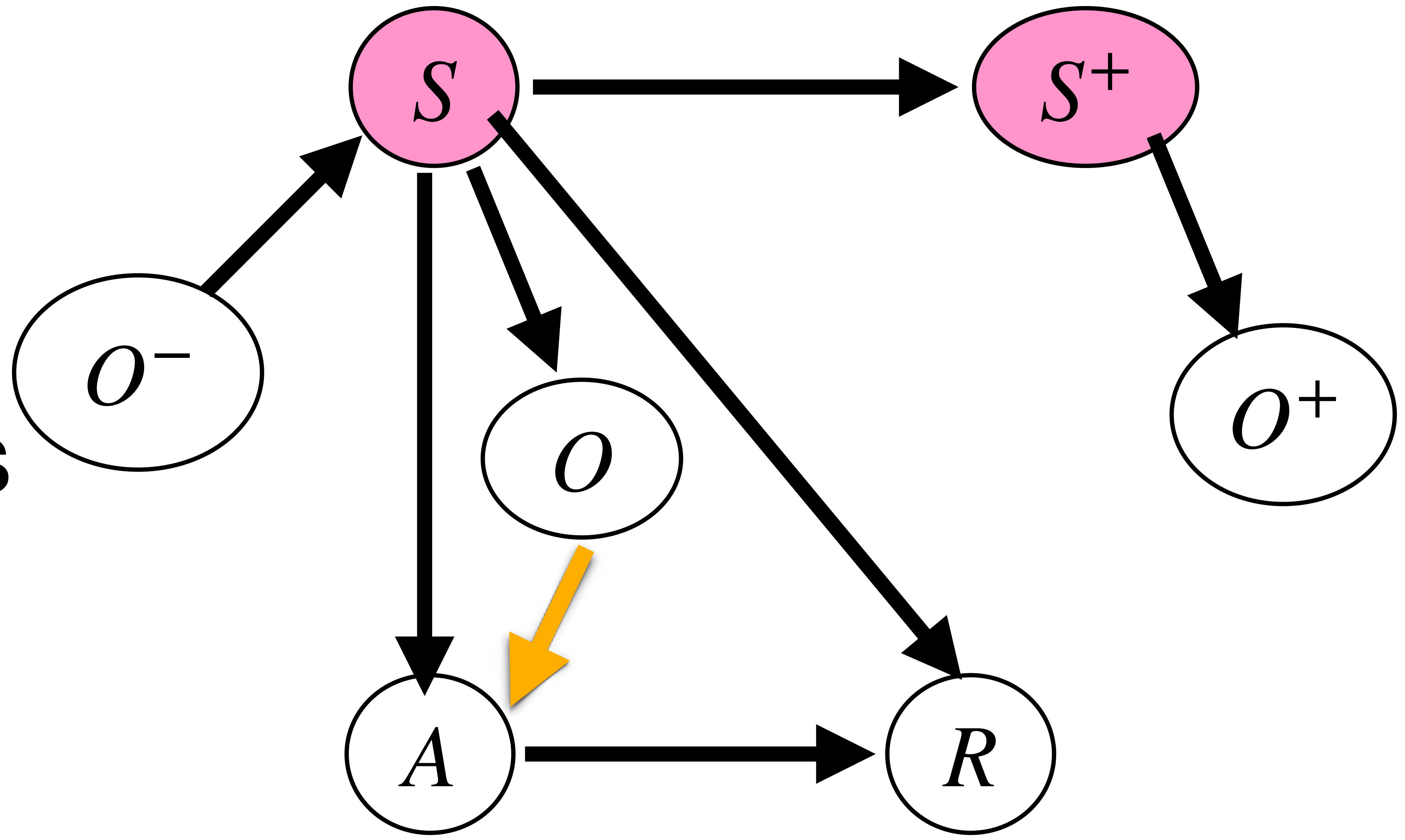
- IS estimator $(1-\gamma)^{-1}\mathbb{E}_{\pi_b}\left[w_{\pi^e/\pi^b}(s)\times\frac{\pi^e(a\mid o)}{\pi^b(a\mid s)}\times r\right]$ does not work.

Weight functions. Ratio of occupancy distributions $P_{\pi^e}(s)/P_{\pi^b}(s)$

Q-functions $\mathbb{E}_{\pi^e}[\sum_{t=0}^{\infty}\gamma^t r_t \mid s_0=s, a_0=a]$

- Direct method estimator $\mathbb{E}_{\pi^b}\left[\sum_a \pi^e(a\mid o)q^{\pi^e}(s,a)\right]$ does not work.

Value bridge functions



Can we consider the analog of weight functions and Q-functions in confounded POMDPs?

(Definition) Value bridge functions $b_V : \mathscr{A} \times \mathscr{O} \to \mathbb{R}$ are defined as solutions to

$$\mathbb{E}_{\pi^b}[b_V(a,o) \mid a,s] = q^{\pi^e}(s,a)\mathbb{E}_{\pi^b}[\pi^e(a \mid o) \mid s]$$

(Definition) Weight bridge functions $b_W : \mathscr{A} \times \mathscr{O} \to \mathbb{R}$ are defined as solutions to

$$\mathbb{E}_{\pi^b}[b_W(a,o^-) \mid a,s] = w_{\pi^e/\pi^b}(s)/\pi^b(a \mid s)\,.$$

When do they exist?

Existence of value bridge functions

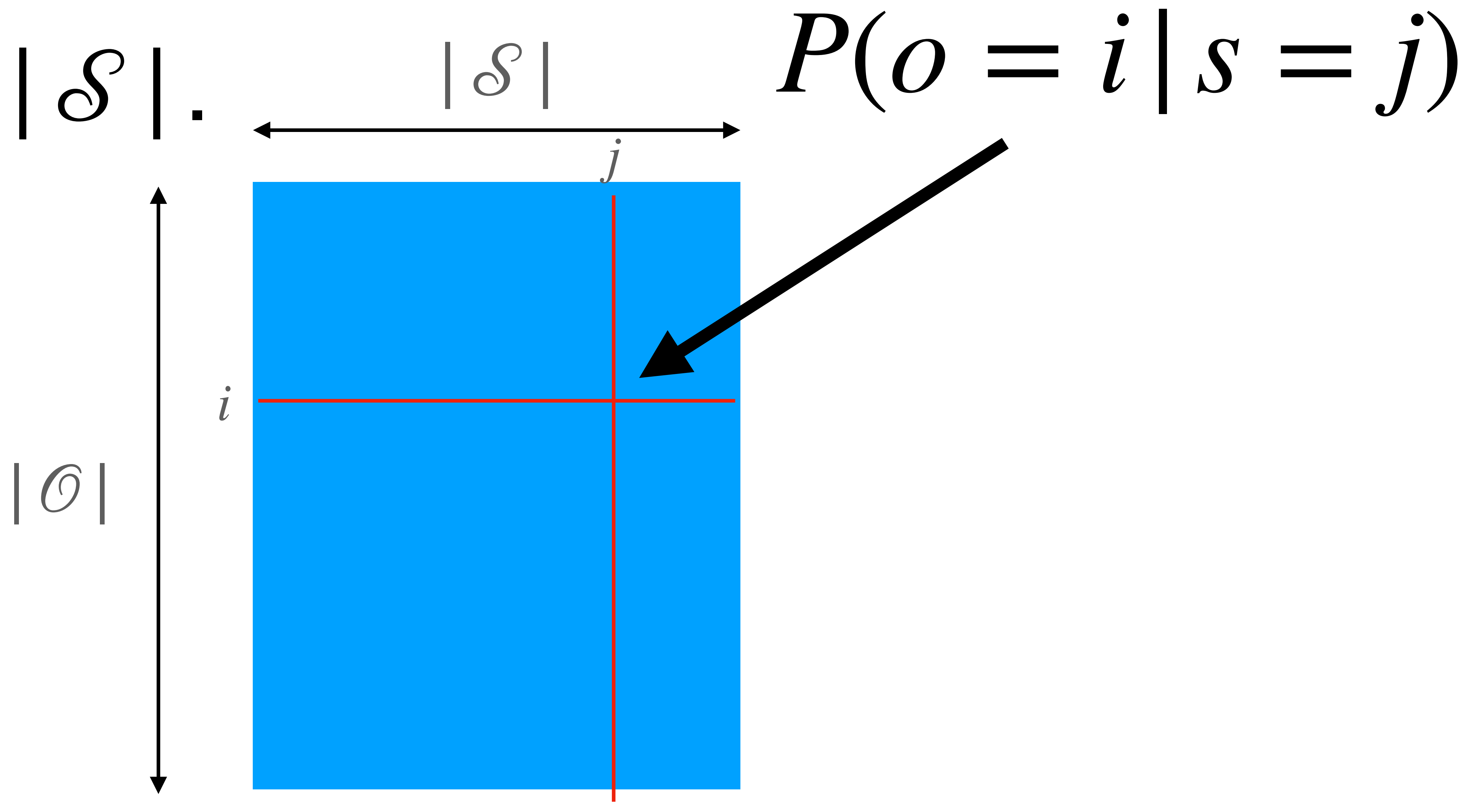
- We need the existence of value bridge functions b_V s.t.

$$\mathbb{E}_{\pi^b}[b_V(a,o) \mid a,s] = q^{\pi^e}(s,a)\mathbb{E}_{\pi^b}[\pi^e(a \mid o) \mid s].$$

- Roughly, it is satisfied O retains enough information about S.

- In the tabular case, $\mathrm{rank}(G) = |\mathcal{S}|$.

  * Assumed in many HMM/POMDP works.



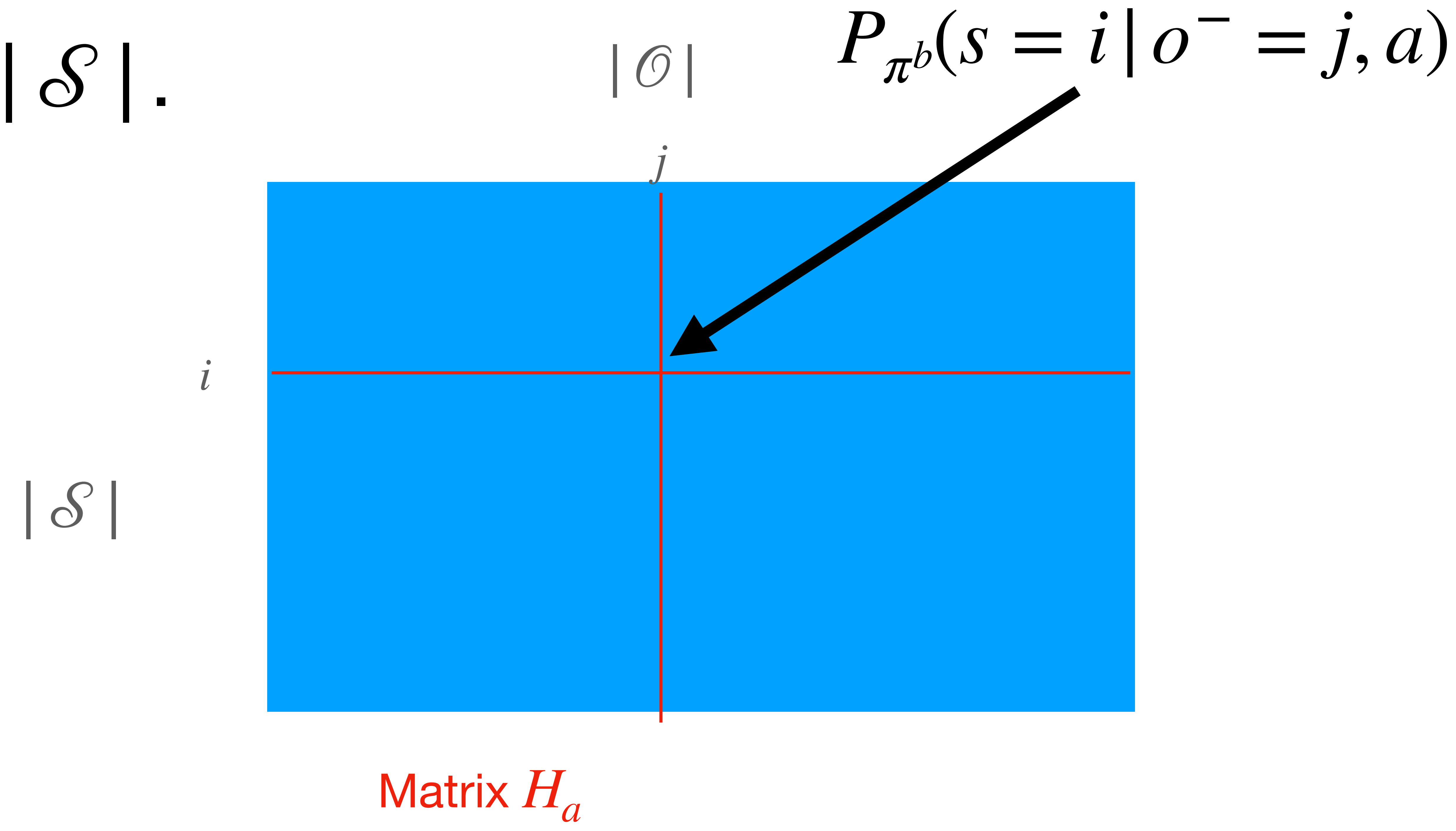
Matrix G

Existence of weight bridge functions

- We need the existence of value bridge functions:

$$\mathbb{E}_{\pi^b}[b_W(a, o^-) \mid a, s] = w_{\pi^e/\pi^b}(s)/\pi^b(a \mid s)\,.$$

- Roughly, it is satisfied O^- retains enough information about S.
- In the tabular case, $\mathrm{rank}(H_a) = |\mathcal{S}|$.



How to use bridge functions for OPE?

When bridge functions exist, we can ensure

Direct method: $J = \mathbb{E}_{o \sim \nu}[\sum_{a'} b_V(a', o)]$

IS method: $J = \mathbb{E}[b_W(a, o^-)\pi^e(a \mid o)r]$

Learnable Value bridge functions

- Definition of value bridge functions

  $\mathbb{E}_{\pi^b}[b_V(a,o) \mid a,s] = q^{\pi^e}(s,a)\mathbb{E}_{\pi^b}[\pi^e(a \mid o) \mid s]$ is not still useful for learning 😂
- We can use the analog of Bellman equations for value bridge functions:

  $$\mathbb{E}_{\pi^b}[\gamma \sum_{a'} b_V(a',o^+) + r\pi^e(a \mid o) - b_V(a,o) \mid a, o^-] = 0.$$
- This is equivalent to

$\mathbb{E}_{\pi^b}[\{\gamma \sum_{a'} b_V(a',o^+) + r\pi^e(a \mid o) - b_V(a,o)\}f(a,o^-)] = 0$ for any $f \in [\mathcal{A} \times \mathcal{O} \to \mathbb{R}]$

- We can similarly define Bellman flow equations for weight bridge functions b_W.

This forms a basis for learning b_V, b_W 😀

IS/Direct method with minimax estimators

PO-MQL
(Partially Observable Minimax Q-function learning)

Function classes: $\mathscr{V} \subset [\mathscr{A} \times \mathscr{O} \to \mathbb{R}], \mathscr{V}^{\dagger} \subset [\mathscr{A} \times \mathscr{O} \to \mathbb{R}]$.

(1) Construct $\hat{b}_V := \mathrm{argmin}_{g \in \mathscr{V}} \max_{f \in \mathscr{V}^{\dagger}} \mathbb{E}_{\mathscr{D}} \left[\{\gamma \sum_{a'} g(a', o^+) + r\pi^e(a \mid o) - g(a, o)\} f(a, o^-) \right]$

(2) Direct method $\hat{J}_{VM} = \mathbb{E}_{o \sim \nu}[\sum_{a'} \hat{b}_V(a', o)]$

Empirical approximation

PO-MWL
(Partially Observable Minimax Weight learning)

Function classes: $\mathscr{W} \subset [\mathscr{A} \times \mathscr{O} \to \mathbb{R}], \mathscr{W}^{\dagger} \subset [\mathscr{A} \times \mathscr{O} \to \mathbb{R}]$.

(1) Construct $\hat{b}_W := \mathrm{argmin}_{g \in \mathscr{W}} \max_{f \in \mathscr{W}^{\dagger}} \mathbb{E}_{\mathscr{D}}[L_W(g, f)]$ for some loss L_w.

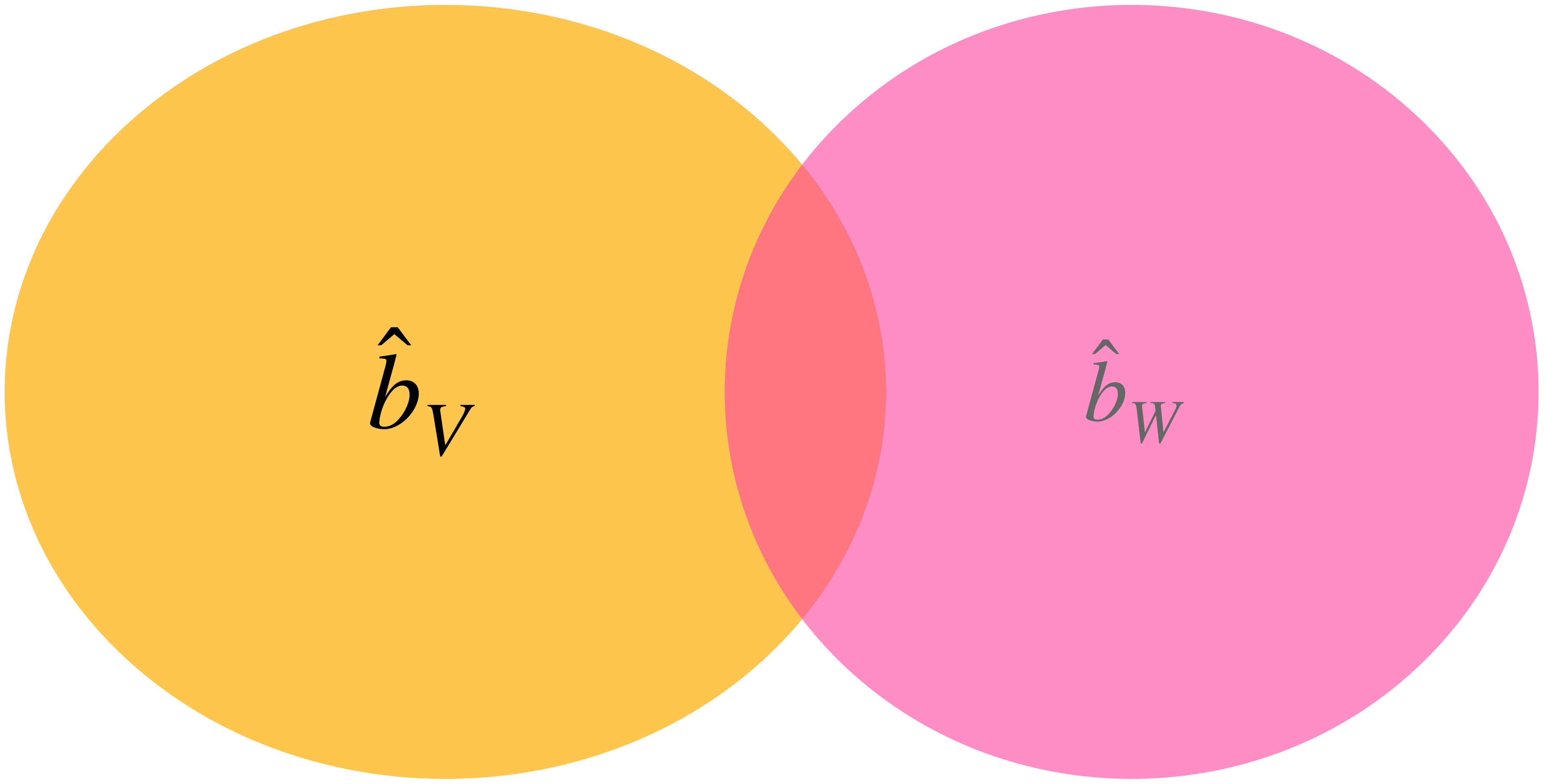
(2) IS method $\hat{J}_{IS} = \mathbb{E}_{\mathscr{D}}[\hat{b}_W(a, o^-) r\pi^e(a \mid o)]$

Doubly robust method with minimax estimators

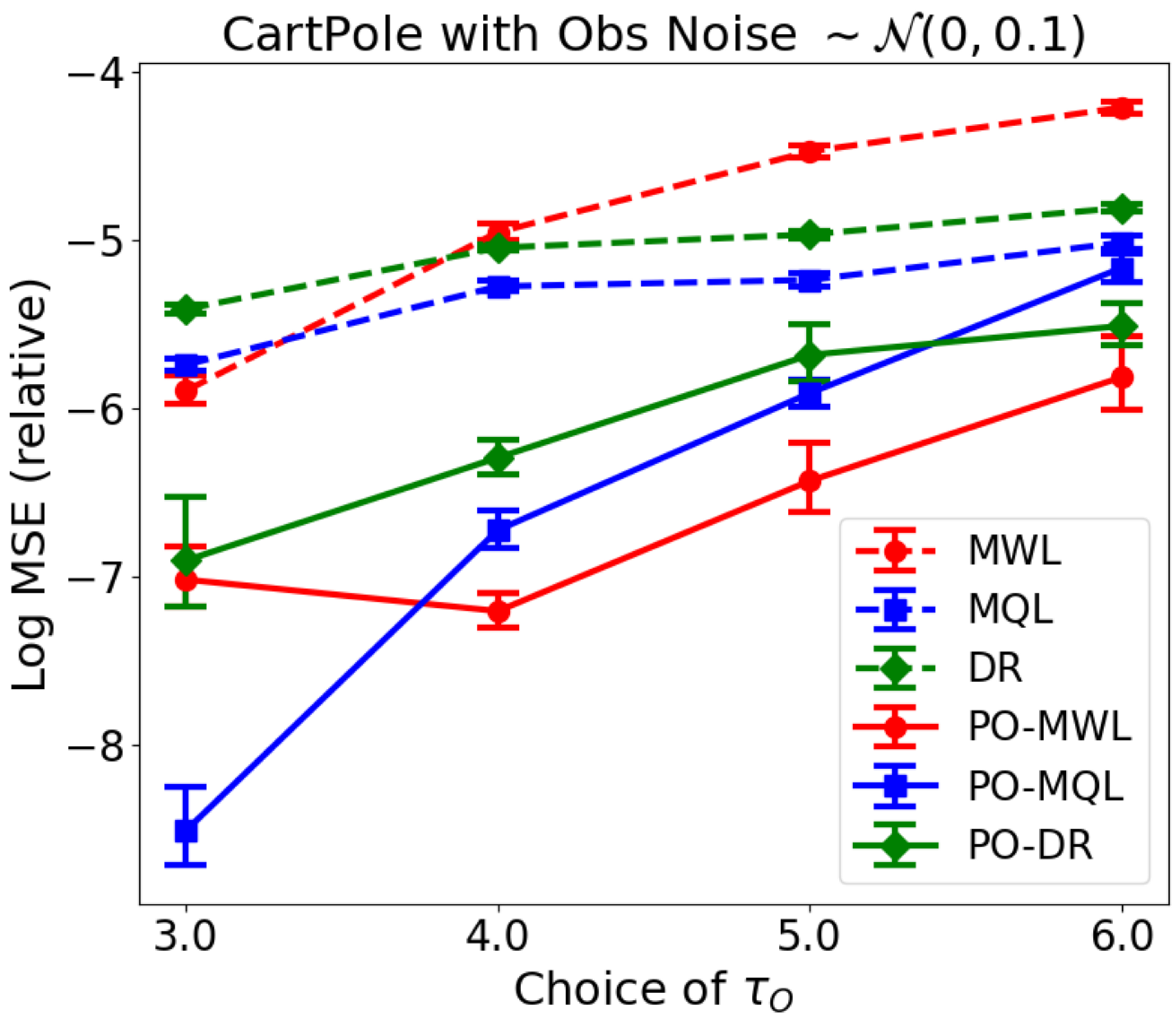
PO-DR
(Partially observable doubly robust)

$$\hat{J}_{DR} = \mathbb{E}_{o \sim \nu_o}[\sum_{a'} \hat{b}_V(a', o)] + \mathbb{E}_{\mathcal{D}}\left[(1-\gamma)^{-1}\hat{b}_W(a, o^-)\left[\{r + \gamma \sum_{a'} \hat{b}_V(a', o^+)\}\pi^e(a \mid o) - \hat{b}_V(a, o)\right]\right]$$

We can prove $\hat{J}_{DR}$ is consistent as long as either $\hat{b}_V$ or $\hat{b}_W$ is consistent.



Experiment



Setting

- We consider confounded POMDPs using Cartpole envrionments.
- We add gaussian noise to states.

Result

- MWL, MQL, DR are existing methods for MDPs.
- PO-MWL, PO-MQL, PO-DR are our proposal.

More contents

- Various finite sample results (realizability+ bellman completeness, doubly realizability, etc)
- Finite horizon case.
- Memory-based policies.

Summary

- Consider OPE methods with unmeasured cofounders.
- We can estimate the policy value via value/weight bridge functions.
  - (1) Estimate value/weight bridge functions using the minimax loss function.
  - (2) Plug them into IS (PO-MWL), direct methods (PO-MQL), and doubly robust methods (PO-DR).