$(RL)^2$ TECHNION Israel Institute of Technology

Online Learning with A Lot of Batch Data

Shie Mannor

shie@technion.ac.il
With G. Tennenholtz, U. Shalit and Y. Efroni

Technion - Israel Institute of Technology & NVIDIA Research

Motivation

- Large amounts of offline data are readily available
  - Healthcare
  - Autonomous Driving / Smart Cities
  - Education
  - Robotics
- The problem: offline data is often partially observable.
- May result in biased estimates that are confounded by spurious correlation.

Motivation

Use offline data for reinforcement learning (RL)

- Off-policy evaluation.
- Batch-mode reinforcement learning (offline RL).
- Let's start with bandits

Part I: Linear Bandits + Confouned Data

- Mixed setting: online + offline
- Linear contextual bandit (online)
  - T trials, $|\mathcal{A}|$ discrete actions, $x_t \in \mathcal{X}$ i.i.d. contexts
  - Context dimension: d
  - Reward given by $r_t = \left\langle x_t, w^*_{a_t} \right\rangle + \eta_t$
  - $\left\{w^*_a \in \mathbb{R}^d\right\}_{a \in \mathcal{A}}$ are unknown parameter vectors
  - η_t is some conditionally σ-subgaussian random noise
  - Minimize regret:

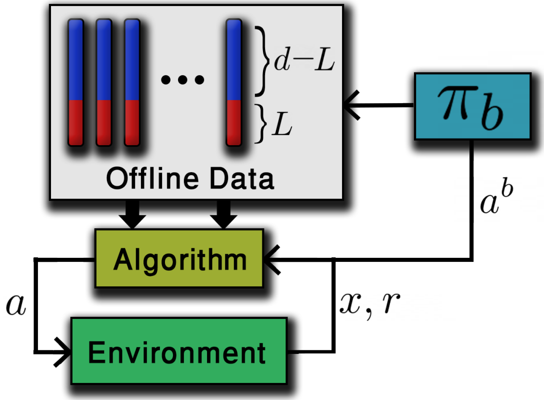
$$Regret\,(T) = \sum_{t=1}^{T} \left\langle x_t, w^*_{\pi^*(x_t)} \right\rangle - \sum_{t=1}^{T} \langle x_t, w^*_{a_t} \rangle.$$

Setup: Linear Bandits + Confouned Data

Additional access to partially observable offline data

- Data was generated by an unknown, fixed behavior policy π_b
- Only L features of the context are visible in the data
- Let $x^{\mathrm{o}}, x^{\mathrm{h}}$ denote the observed and unobserved features of the context x, respectively.

Setup: Linear Bandits + Confouned Data



Partially Observable Data = Linear Constraints

- Suppose we ignore that the data is partially observable.
- We find a least square solution to

$$\min_{b\in\mathbb{R}^L} \sum_{i=1}^{N_a} (\langle x_i^{\mathrm{o}}, b\rangle - r_i)^2 \quad \forall a \in \mathcal{A}.$$

- Denote by b_a^{LS} its solution.
- Can b_a^{LS} provide useful information for the bandit problem?

Partially Observable Data = Linear Constraints

Proposition

Let $R_{11}(a) = \mathbb{E}^{\pi_b}\left(x^{\mathrm{o}}\left(x^{\mathrm{o}}\right)^T|a\right)$, $R_{12}(a) = \mathbb{E}^{\pi_b}\left(x^{\mathrm{o}}\left(x^{\mathrm{h}}\right)^T|a\right)$. *The following holds almost surely for all* $a \in \mathcal{A}$*.*

$$\lim_{N\to\infty} b_a^{LS} = \Big(I_{L\times L}, \quad R_{11}^{-1}(a)R_{12}(a)\Big)w_a^*,$$

- b_a^{LS} provides us L independent linear relations.
- We only need to learn a lower dimensional subspace.

Linear Bandits with Linear Constraints

- Given side information to the bandit problem

$$M_a w_a^* = b_a \qquad , a \in \mathcal{A}.$$

- $M_a \in \mathbb{R}^{L \times d}$, $b_a \in \mathbb{R}^L$ are *known*.
- Let P_a denote the orthogonal projection onto the kernel of M_a
- Effectively dimension of problem: $d - L$
- We can thus achieve regret $\widetilde{\mathcal{O}}\left((d-L)\sqrt{KT}\right)$

Linear Bandits with Linear Constraints

Algorithm 1 OFUL with Linear Side Information

input: $\alpha > 0, M_a \in \mathbb{R}^{L \times d}, b_a \in \mathbb{R}^L, \delta > 0$
init: $V_a = \lambda I_d, Y_a = 0, \forall a \in \mathcal{A}$
for $t = 1, \ldots$ **do**
  Receive context x_t
  $\hat{w}_{t,a}^{P_a} = (P_a V_a P_a)^\dagger \left(Y_a - (V_a - \lambda I_d) M_a^\dagger b_a\right)$
  $\hat{y}_{t,a} = \left\langle x_t, M_a^\dagger b_a \right\rangle + \left\langle x_t, \hat{w}_{t,a}^{P_a} \right\rangle$
  $\mathrm{UCB}_{t,a} = \sqrt{\beta_t(\delta)} \|x_t\|_{(P_a V_a P_a)^\dagger}$
  $a_t \in \arg\max_{a \in \mathcal{A}} \{\hat{y}_{t,a} + \alpha \mathrm{UCB}_{t,a}\}$
  Play action a_t and receive reward r_t
  $V_{a_t} = V_{a_t} + x_t x_t^T, Y_{a_t} = Y_{a_t} + x_t r_t$
end for

Deconfounding Partially Obserable Data

- In our case, for partially observable offline data, we get

$$M_a = \left(I_L, \quad R_{11}^{-1}(a) R_{12}(a) \right),$$

  and b_a is the solution to $\min_{b \in \mathbb{R}^L} \sum_{i=1}^{N_a} (\langle x_i^{\mathrm{o}}, b \rangle - r_i)^2$.
- Problem: $R_{12}(a)$ is unknown (not identifiable)

(RL)²
TECHNION
Israel Institute of Technology

Deconfounding Partially Obserable Data

- Solution: Assume that at each round $t > 0$, we can query π_b.
- This lets us get an estimate for R_{12} as

$$\hat{R}_{12}(a,t) = \frac{1}{t}\sum_{i=1}^{t} \frac{1_{a_i=a}}{P^{\pi_b}(a)} (x_i^{\mathrm{o}})(x_i^{\mathrm{h}})^T$$

- Our final estimate for M_a is then

$$\hat{M}_{t,a} = \begin{pmatrix} I_L, & R_{11}^{-1}(a)\hat{R}_{12}(a,t) \end{pmatrix}$$

Deconfounding Partially Obserable Data

Algorithm 2 OFUL with Partially Observable Offline Data

1: **input:** $\alpha > 0, \delta > 0, T, b_a \in \mathbb{R}^L$ (from dataset)
2: **for** $n = 0, \ldots, \log T - 1$ **do**
3: Use 2^n previous samples from π_b to update the estimate of $\hat{M}_{2^n,a}, \forall a \in \mathcal{A}$
4: Calculate $\hat{M}^{\dagger}_{2^n,a}, \hat{P}_{2^n,a}, \forall a \in \mathcal{A}$
5: Run Algorithm 1 for 2^n time steps with bonus $\sqrt{\beta_{n,t}(\delta)}$ and $\hat{M}_{2^n,a}, b_a$
6: **end for**

Deconfounding Partially Obserable Data

Theorem (Main Result)

For any $T > 0$, with probability at least $1 - \delta$, the regret of Algorithm 2 is bounded by

$$Regret(T) \leq \widetilde{\mathcal{O}}\left((1 + f_{B_1})(d - L)\sqrt{KT}\right).$$

- f_{B_1} is a factor indicating how hard it is to estimate the linear constraints
- Worst case dependence: $f_{B_1} \leq \widetilde{\mathcal{O}}\left(\max_a \frac{(L(d-L))^{1/4}}{P^{\pi_b}(a)}\right)$

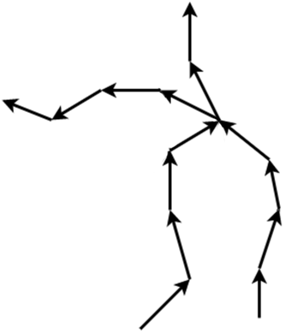
Part II: Off-Policy Evaluation in Reinforcement Learning

Given: data generated by a behavior policy π_b

$(RL)^2$ TECHNION Israel Institute of Technology

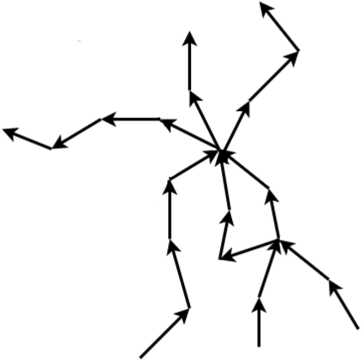
Off-Policy Evaluation in Reinforcement Learning

Given: data generated by a behavior policy π_b



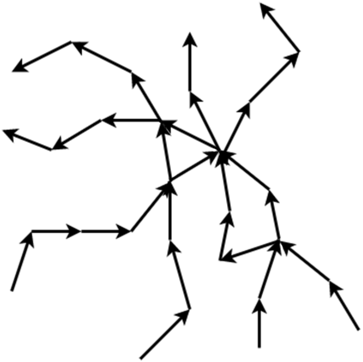
Off-Policy Evaluation in Reinforcement Learning

Given: data generated by a behavior policy π_b



Off-Policy Evaluation in Reinforcement Learning

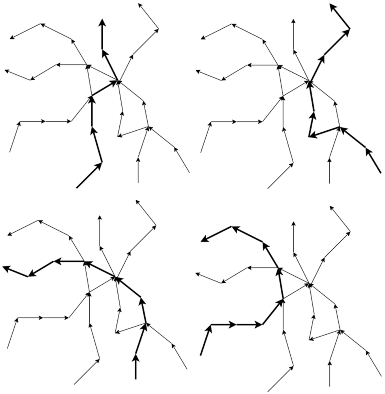
Given: data generated by a behavior policy π_b



$(RL)^2$ TECHNION Israel Institute of Technology

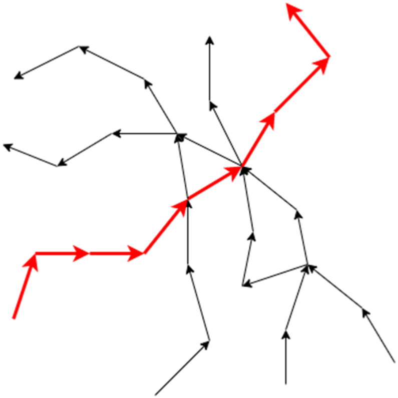
Off-Policy Evaluation in Reinforcement Learning

Given: data generated by a behavior policy π_b



Off-Policy Evaluation in Reinforcement Learning

Evaluate the value of a different policy π_e



Off-Policy Evaluation (OPE)

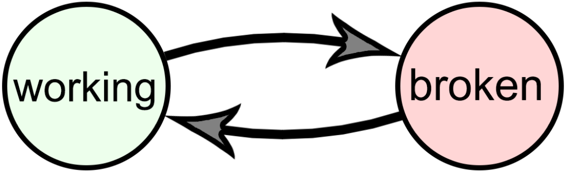
- **Given:** Data generated by a behavioral policy π_b in a Markov Decision Process (MDP)
- **Objective:** Evaluate the value of an evaluation policy π_e.
- **Methods:**
  - Direct methods (model based and model free)
  - Inverse propensity scoring (e.g., importance sampling)
  - Doubly robust methods
- How do we define OPE under partial observability?

Example: Probabilistic Maintenance

- We have an expensive machine that needs monthly maintenance.
- Every month an expert comes and checks the machine.
- The expert knows whether the machine is working properly.
- The expert can choose to fix the machine or leave it as is.

Example: Probabilistic Maintenance

- State Space: Working / Broken
- Action Space: Fix / Not Fix



Example: Probabilistic Maintenance (Numeric)

F - fix
NF - not fix

	working	broken
working	F NF 1 / 0.9	F NF 0 / 0.1
broken	F NF 1 / 0	F NF 0 / 1

Example: Probabilistic Maintenance (Numeric)

F - fix
NF - not fix

	working	broken
working	F NF 1 / 0.9	F NF 0 / 0.1
broken	F NF 1 / 0	F NF 0 / 1

$$r(\cdot, \text{fix}) = -1, \quad r(\text{broken}, \cdot) = -10$$

Example: Probabilistic Maintenance (Numeric)

F - fix
NF - not fix

	working	broken
working	F NF 1 / 0.9	F NF 0 / 0.1
broken	F NF 1 / 0	F NF 0 / 1

$$\pi_b(\text{working}) = \begin{cases} F & \text{, w.p. } 0.1 \\ NF & \text{, w.p. } 0.9 \end{cases}$$

$$\pi_b(\text{broken}) = \begin{cases} F & \text{, w.p. } 0.9 \\ NF & \text{, w.p. } 0.1 \end{cases}$$

$$r(\cdot, \text{fix}) = -1, \quad r(\text{broken}, \cdot) = -10$$

Example: Probabilistic Maintenance (Numeric)

We only see a noisy observation of the state:

Temperature of machine

Example: Probabilistic Maintenance (Numeric)

We only see a noisy observation of the state:

Temperature of machine

$$O(\text{working}) = \begin{cases} \text{HOT} & \text{, w.p. } 0.1 \\ \text{NORMAL} & \text{, w.p. } 0.9 \end{cases}$$

$$O(\text{broken}) = \begin{cases} \text{HOT} & \text{, w.p. } 0.9 \\ \text{NORMAL} & \text{, w.p. } 0.1 \end{cases}$$

Example: Probabilistic Maintenance (Numeric)

$$v_{\mathrm{IS}}^{\pi_e} = \mathbb{E}\left(\left(\sum_{t=0}^{H} r_t\right) \prod_{t=0}^{H} \frac{\pi_e(a_i|s_i)}{\pi_b(a_i|s_i)} \middle| \pi_b\right)$$

Example: Probabilistic Maintenance (Numeric)

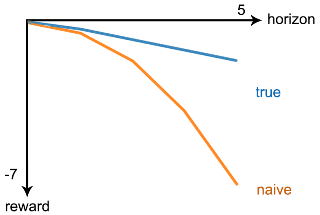
$$v_{\text{IS}}^{\pi_e} = \mathbb{E}\left(\left(\sum_{t=0}^{H} r_t\right) \prod_{t=0}^{H} \frac{\pi_e(a_i|s_i)}{\pi_b(a_i|s_i)} \middle| \pi_b\right)$$

$$v_{\text{naive}}^{\pi_e} = \mathbb{E}\left(\left(\sum_{t=0}^{H} r_t\right) \prod_{t=0}^{H} \frac{\pi_e(a_i|o_i)}{\pi_b(a_i|o_i)} \middle| \pi_b\right)$$

Example: Probabilistic Maintenance (Numeric)

$$v_{\mathrm{IS}}^{\pi_e} = \mathbb{E}\left(\left(\sum_{t=0}^{H} r_t\right)\prod_{t=0}^{H}\frac{\pi_e(a_i|s_i)}{\pi_b(a_i|s_i)}\middle|\pi_b\right)$$

$$v_{\mathrm{naive}}^{\pi_e} = \mathbb{E}\left(\left(\sum_{t=0}^{H} r_t\right)\prod_{t=0}^{H}\frac{\pi_e(a_i|o_i)}{\pi_b(a_i|o_i)}\middle|\pi_b\right)$$



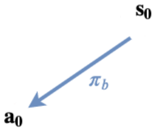
OPE in Partially Observable Environments

How do we define OPE under partial observability?

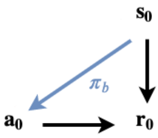
Partially Observable Markov Decision Process (POMDP)

s_0

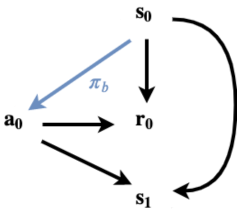
Partially Observable Markov Decision Process (POMDP)



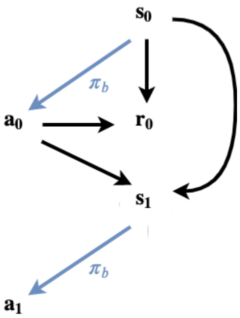
Partially Observable Markov Decision Process (POMDP)



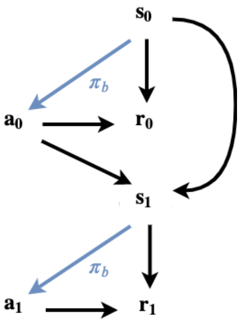
Partially Observable Markov Decision Process (POMDP)



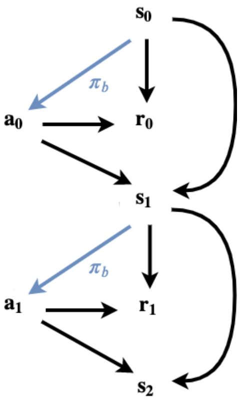
Partially Observable Markov Decision Process (POMDP)



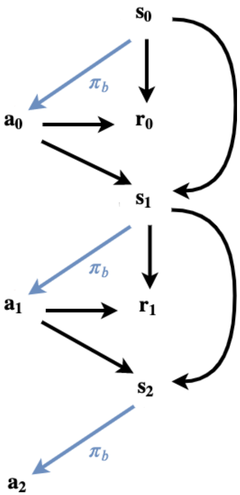
Partially Observable Markov Decision Process (POMDP)



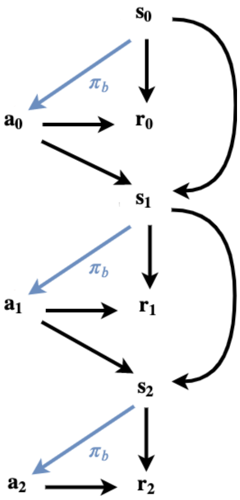
Partially Observable Markov Decision Process (POMDP)



Partially Observable Markov Decision Process (POMDP)



Partially Observable Markov Decision Process (POMDP)



Partially Observable Markov Decision Process (POMDP)

s_0

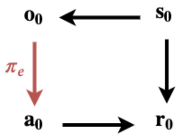
Partially Observable Markov Decision Process (POMDP)

$$o_0 \longleftarrow s_0$$

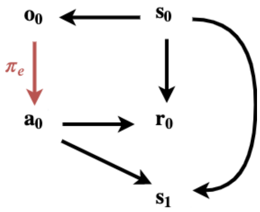
Partially Observable Markov Decision Process (POMDP)

$$
\begin{tikzcd}
\mathbf{o}_0 \arrow[d, "\pi_e"'] & \mathbf{s}_0 \arrow[l] \\
\mathbf{a}_0 &
\end{tikzcd}
$$

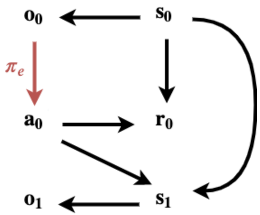
Partially Observable Markov Decision Process (POMDP)



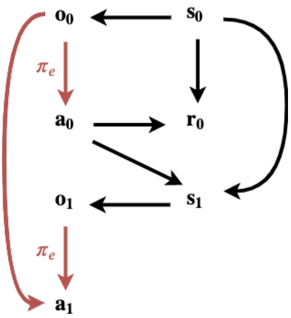
Partially Observable Markov Decision Process (POMDP)



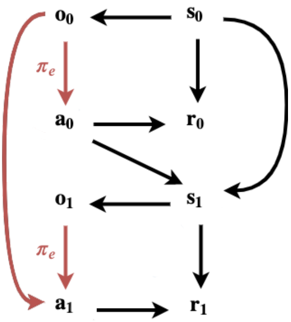
Partially Observable Markov Decision Process (POMDP)



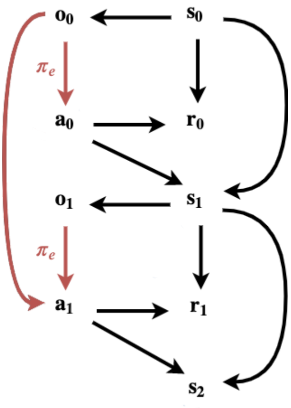
Partially Observable Markov Decision Process (POMDP)



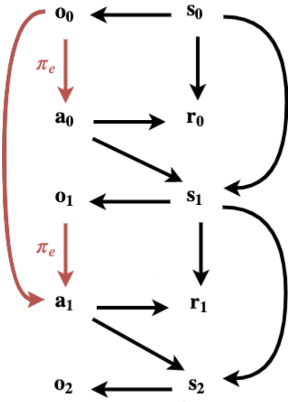
Partially Observable Markov Decision Process (POMDP)



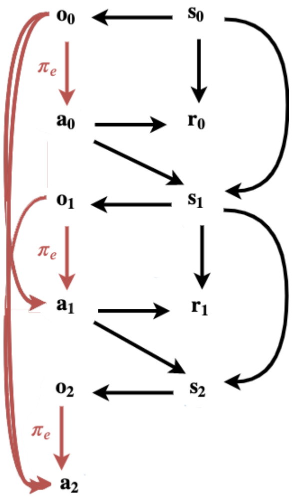
Partially Observable Markov Decision Process (POMDP)



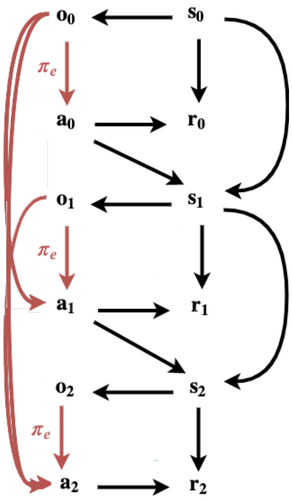
Partially Observable Markov Decision Process (POMDP)



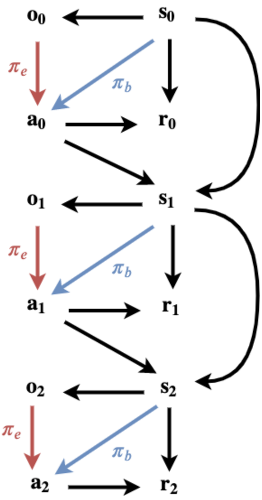
Partially Observable Markov Decision Process (POMDP)



Partially Observable Markov Decision Process (POMDP)

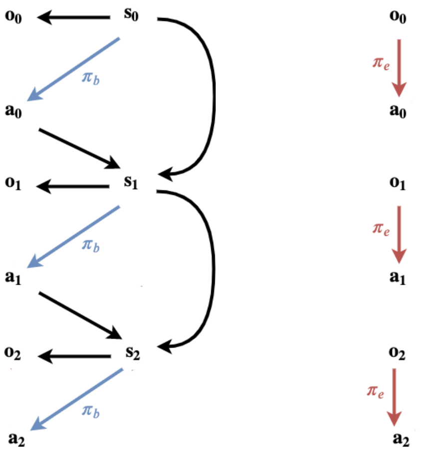


Partially Observable Markov Decision Process (POMDP)



Partially Observable Markov Decision Process (POMDP)

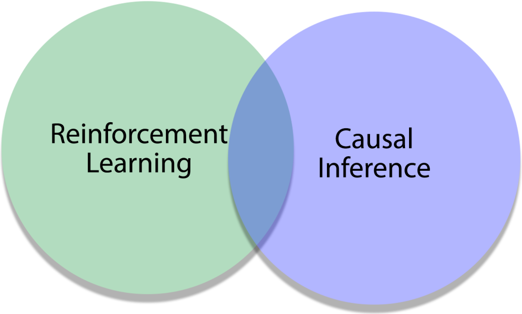
this is what created the data | this is what we see



Off-Policy Evaluation in POMDPs

The goal of off-policy evaluation in POMDPs is to evaluate $v(\pi_e)$ using the measure $P^{\pi_b}(\cdot)$ over observable trajectories $\mathcal{T}_L^o$ and the given policy π_e.

Reinforcement Learning and Causal Inference: Better Together



A Meeting Point of RL and CI

- When performing off-policy evaluation (or learning) on data where we do not have access to the same data as the agent.
- Example: physicians treating patients in an intensive care unit (ICU)
- Mistakes were made: applying RL to observational ICU data without considering hidden confounders or overlap (common support, positivity)
- In RL, hidden confounding can be described using partial observability.

CI-RL Dictionary

Causal Term	RL Term	Example
confounder (possibly hidden)	state (possibly unobserved)	information available to the doctor
action, treatment	action	medications, procedures
outcome	reward	mortality
treatment assigment process	behavior policy	the way doctors treat patients
proxy variable	observations	electronic health record

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions
1. Association $P(y\|x)$	Seeing	What is? How would seeing x change my belief in y?
2. Intervention $P(y\|\mathbf{do}(x), z)$	Doing Intervening	What if? What if I do x?
3. Counterfactual $P(y_x\|x', y')$	Imagining Retrospection	Why? Was it x that caused y? What if I had acted differently?

Pearl, Judea. "Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution." Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining. ACM, 2018.

Off-Policy Evaluation in POMDPs

Theorem 1 (POMDP Evaluation)

Assume $|\mathcal{O}| \geq |\mathcal{S}|$ and under invertibility assumptions of the dynamics we have an estimator

$$v(\pi_e) = f(\text{observable data})$$

Off-Policy Evaluation in POMDPs

Theorem 1 (POMDP Evaluation)

Assume $|\mathcal{O}| \geq |\mathcal{S}|$ and that the matrices $P^b(O_i|a_i, O_{i-1})$ are invertible for all i and all $a_i \in \mathcal{A}$. For any $\tau^o \in \mathcal{T}_t^o$ define the generalized weight matrices

$$W_i(\tau^o) = P^b(O_i|a_i, O_{i-1})^{-1} P^b(O_i, o_{i-1}|a_{i-1}, O_{i-2})$$

for $i \geq 1$, and $W_0(\tau^o) = P^b(O_0|a_0, O_{-1})^{-1} P^b(O_0)$.
Denote $\Pi_e(\tau^o) = \prod_{i=0}^t \pi_e^{(i)}(a_i|h_i^o)$, $\Omega(\tau^o) = \prod_{i=0}^t W_{t-i}(\tau^o)$.
Then

$$P^e(r_t) = \sum_{\tau^o \in \mathcal{T}_t^o} \Pi_e(\tau^o) P^b(r_t, o_t|a_t, O_{t-1}) \Omega(\tau^o).$$

(RL)^2 TECHNION Israel Institute of Technology

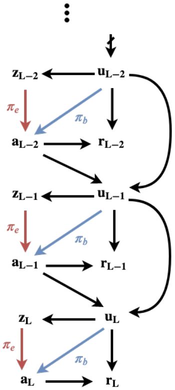
POMDP Limitation

- Causal structure of POMDPs is restricting.
- Must invert matrices of dimension $\mathcal{S}$ even when $\mathcal{O} \subset \mathcal{S}$.

- **Solution:**
  - Detach observed and unobserved variables.
  - Decoupled POMDPs (more in our AAAI 20' paper).

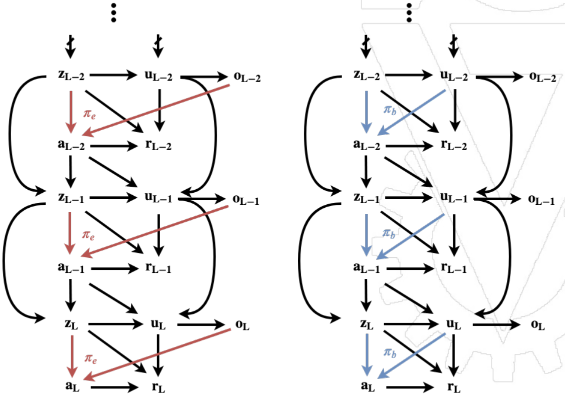
A Special POMDP (DE-POMDP)

POMDP



Z observations, U state

Decoupled PODMP



$Z \& U$ is state O is observations

$(RL)^2$ TECHNION Israel Institute of Technology

Conclusions

- Unknown states (confounders) produce bias through factors that affect both observed actions and rewards.
- This is a major problem in offline off-policy data.
- Be aware of such biases when using off-policy data that was not generated by them.

- Our work is a first step to introducing OPE for partially observable environments in RL.
- Causality and RL: Better together