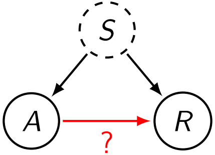
Using Confounded Data in Offline RL

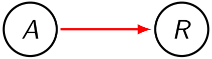
Maxime Gasse, Damien Grasset, Guillaume Gaudron, Pierre-Yves Oudeyer

NeurIPS 2022 - Offline RL workshop
Dec 2, 2022



Confounding in Offline RL?

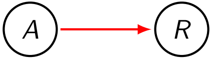
Online data = interventions, measures $p(r|do(a))$



$$a^\star = \arg\max_a \mathop{\mathbb{E}}_{p(r|do(a))}[r]$$

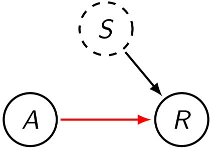
Confounding in Offline RL?

Online data = interventions, measures $p(r|do(a))$



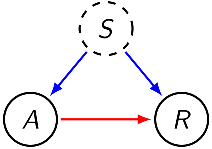
$$a^\star = \arg\max_a \mathop{\mathbb{E}}_{p(r|do(a))}[r]$$

Offline data = observations, measures $p(r|a)$



$p(r|a) = p(r|do(a))$

No confounding



$p(r|a) \neq p(r|do(a))$

Confounding, self-delusion [Ortega et al., 2021]

Model-based RL in POMDPs

Causal transition model: $p(o_{t+1}|o_{0\to t}, do(a_{0\to t}))$

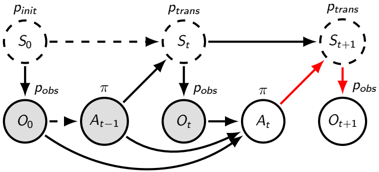
Observed transition model: $p(o_{t+1}|o_{0\to t}, a_{0\to t})$

Model-based RL in POMDPs

Causal transition model: $p(o_{t+1} | o_{0 \to t}, do(a_{0 \to t}))$

Observed transition model: $p(o_{t+1} | o_{0 \to t}, a_{0 \to t})$

Standard regime



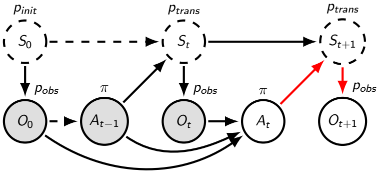
Observed = causal

Model-based RL in POMDPs

Causal transition model: $p(o_{t+1}|o_{0\to t}, do(a_{0\to t}))$

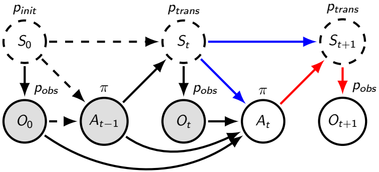
Observed transition model: $p(o_{t+1}|o_{0\to t}, a_{0\to t})$

Standard regime



Observed = causal

Privileged regime



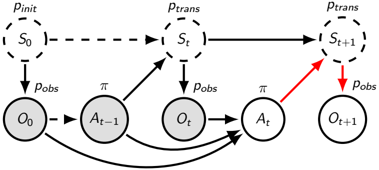
Observed $\neq$ causal !!

Model-based RL in POMDPs

Causal transition model: $p(o_{t+1}|o_{0\rightarrow t}, do(a_{0\rightarrow t}))$

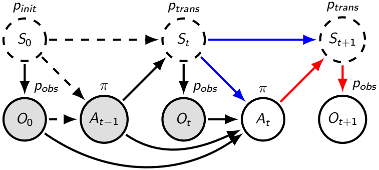
Observed transition model: $p(o_{t+1}|o_{0\rightarrow t}, a_{0\rightarrow t})$

Standard regime



Observed = causal

Privileged regime



Observed $\neq$ causal !!

Offline data from human demonstrations

- autonomous driving → privileged
- medical treatment recommendation → privileged
- atari → standard

Combining online and offline data

Online (standard) data

- $\mathcal{D}_{std} \sim p(o_{t+1} | o_{0 \rightarrow t}, do(a_{0 \rightarrow t}))$

Offline (privileged) data

- $\mathcal{D}_{prv} \sim p(o_{t+1} | o_{0 \rightarrow t}, a_{0 \rightarrow t})$

Combining online and offline data

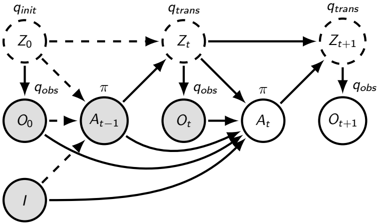
Online (standard) data

- $\mathcal{D}_{std} \sim p(o_{t+1} | o_{0\to t}, do(a_{0\to t}))$

Offline (privileged) data

- $\mathcal{D}_{prv} \sim p(o_{t+1} | o_{0\to t}, a_{0\to t})$

Augmented POMDP model



Combining online and offline data

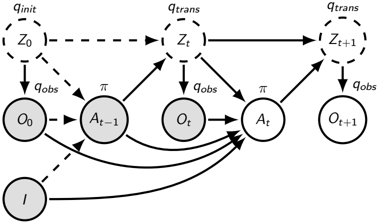
Online (standard) data

- $\mathcal{D}_{std} \sim p(o_{t+1} | o_{0\to t}, do(a_{0\to t}))$

Offline (privileged) data

- $\mathcal{D}_{prv} \sim p(o_{t+1} | o_{0\to t}, a_{0\to t})$

Augmented POMDP model



$$\hat{q} = \arg\max_{q\in\mathcal{Q}} \sum_{(\tau)}^{\mathcal{D}_{prv}} \log q(\tau | i = 0) + \sum_{(\tau)}^{\mathcal{D}_{std}} \log q(\tau | i = 1)$$

Correct and sample-efficient (guarantees in the paper).

Model learning baselines

No obs

- online data only (correct)

$$\hat{q} = \arg\max_{q \in \mathcal{Q}} \sum_{(\tau)}^{\mathcal{D}_{std}} \log q(\tau | i = 1)$$

Naive

- offline + online (incorrect)

$$\hat{q} = \arg\max_{q \in \mathcal{Q}} \sum_{(\tau)}^{\mathcal{D}_{prv} \cup \mathcal{D}_{std}} \log q(\tau | i = 1)$$

Augmented

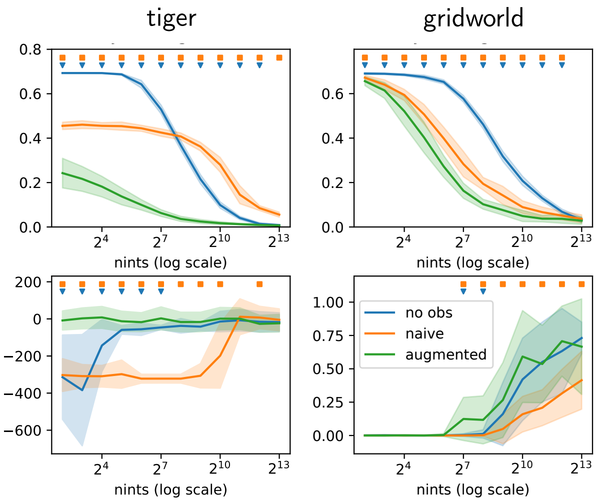
- offline + online (correct)

$$\hat{q} = \arg\max_{q \in \mathcal{Q}} \sum_{(\tau)}^{\mathcal{D}_{prv}} \log q(\tau | i = 0) + \sum_{(\tau)}^{\mathcal{D}_{std}} \log q(\tau | i = 1)$$

Experiments on synthetic POMDPs

model quality
(JS divergence)

agent performance
(cumulated reward)

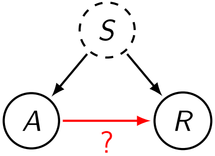


Take-home message: RL is causal!

Causality provides useful tools to reason about offline data

Beware **privileged agents** and **confounding**

- Using offline data naively can **degrade** the performance of an online RL agent
- Using offline data safely can **improve** the performance of an online RL agent



Come to our poster!
"Using Confounded Data in Offline RL"

Using Confounded Data in Offline RL

Maxime Gasse, Damien Grasset, Guillaume Gaudron, Pierre-Yves Oudeyer

NeurIPS 2022 - Offline RL workshop
Dec 2, 2022