Causal Reinforcement Learning using Observational and Interventional Data

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

Published in Transactions on Machine Learning Research (08/2023)

Presented by: Annie Raichev
Outline

• Problem
• Background
• Model based RL as a causal inference problem
  • Interventional regime
  • Observational regime
• Augmented POMDP: combining regimes
• Learning the causal transition model
• Experiments
Problem

• How to combine offline + online data?
• Agent can collect online experiences by interacting with their environment
  • Interventional Data
• Agent has access to offline experiences, from observing another agent
  • Observational data
• Can we combine to learn a latent based causal model and improve agent’s performance?
Motivating Example

• Autonomous driving
  • Offline data is collected by human drivers
  • Human drivers have wider field of vision
  • Robot drivers rely on cameras

• Human view: break when they see a person waiting to cross the street

• Agent view: Only see the person when it enters camera’s field of vision, by walking in front of the car

• Issue: Agent may wrongly infer from its own observations that breaking causes a person to appear in front of the car
Motivating Example cont...

- Agent may *wrongly infer* from its own observations that breaking causes a person to appear in front of the car.
- Agent could infer it should never brake if it’s goal is to avoid hitting someone.
- Eventually could infer the right causal effects if enough online data is collected.

**Issues:** performing many interventions to see what happens
  - Potentially dangerous
  - Not affordable

**Solution:** collect offline data by observing humans.
Background: Pearl’s Causal Hierarchy

3. Counterfactuals “What if \( X = x \) had been \( X = x' \)?”
   Ex: “Would I be home by now had I gotten off the 405?”
   Denizens: Structural Causal Models

2. Interventions “What if I do \( X = x \)?”
   Ex: “Will exercise lower my cholesterol?”
   Denizens: Causal Bayesian Networks, Reinforcement Learners

1. Associations “What if I see \( X = x \)?”
   Ex: “Is symptom \( X \) associated with disease \( Y \)?”
   Denizens: Bayesian Networks, Supervised & Unsupervised Learning
Background: Do-Calculus

Rule 1: Adding/removing Observations

\[ P(y|do(x), z, w) = P(y|do(x), w) \quad \text{if} \quad (Z \perp Y | W)_{G_X} \]

Rule 2: Action/observation exchange

\[ P(y|do(x), do(z), w) = P(y|do(x), z, w) \quad \text{if} \quad (Z \perp Y | X, W)_{G_{XZ}} \]

Rule 3: Adding/removing Actions

\[ P(y|do(x), do(z), w) = P(y|do(x), w) \quad \text{if} \quad (Z \perp Y | X, W)_{G_{XZ}(W)} \]

where \( Z(W) \) is the set of \( Z \)-nodes that are not ancestors of any \( W \)-node in \( G_X \).
A POMDP: $M = (S, O, A, p_{init}, p_{prv}, p_{trans}, r)$

- $s \in S$ hidden/partially observable states
- $a \in A$ actions, $o \in O$ observations
- $p_{init}(s_0)$, initial state distribution
- $p_{trans}(s_{t+1}|s_t, a_t)$, state transition distribution
- $p_{prv}(o_t|s_t)$, observation distribution
- $r: O \rightarrow \mathbb{R}$, reward function
Background: POMDP

- **Stochastic policy** $\pi(a_t|h_t)$
- Actions based on all info from the past
- A **trajectory** refers to a sequence of states, actions, and observations over time

**Probability of a trajectory:**

$$p_{std}(\tau) = \sum_{s_0 \rightarrow |\tau|} p_{init}(s_0)p_{prv}(o_0|s_0) \prod_{t=0}^{\tau-1} \pi_{std}(a_t|h_t)p_{trans}(s_{t+1}|s_t, a_t)p_{prv}(o_{t+1}|s_{t+1}).$$

In this setting $S_t \perp A_t|H_t$ is always true
Background: POMDP

• **Goal:** find optimal policy \( \pi^* = \arg \max_{\pi} \mathbb{E}_{\tau \sim p_{std}} \left[ \sum_{t=0}^{|\tau|} r(o_t) \right] \).

• Need to estimate the POMDP transition model \( p_{std}(o_{t+1}|h_t, a_t) \)
  1. **learning:** given a dataset \( \mathcal{D} \), estimate a transition model \( \hat{q}(o_{t+1}|h_t, a_t) \approx p_{std}(o_{t+1}|h_t, a_t) \)
  2. **planning:** given a history \( h_t \) and a transition model \( \hat{q} \), decide on an optimal action \( a_t \).
Model Based RL as a causal inference problem

Guiding example. Consider a door, a light, and two buttons A and B. The light is red 60% of the time, and green the rest of the time. When the light is red, button A opens the door, while when the light is green, then button B opens the door. I am told that the mechanism responsible for opening the door depends on both the light color and the button pressed (light $\rightarrow$ door $\leftarrow$ button), but I am not given the mechanism itself. Suppose now that I am colorblind, and I want to open the door. Which button should I press? In the do-calculus framework, the question I am asking is

$$\arg \max_{\text{button} \in \{A,B\}} p(\text{door}=\text{open}|\text{do(button)}).$$
Interventional Regime: Causal Perspective

- Interventional data set $D_{std}$ of episodes $\tau$ from policy $\pi(a_t|h_t)$
- Want to evaluate the effect of each action on the system

$$p_{std}(o_{t+1} | o_{0\rightarrow t}, do(a_{0\rightarrow t})) = p_{std}(o_{t+1} | h_t, a_t).$$

From Rule 2 of Do-Calc:

$$O_{t+1} \perp A_t | H_t$$

In mutilated model
Interventional Regime: Estimation

• As long as $\pi$ has a non zero chance to explore every action, we can obtain an unbiased estimate of the POMDP transition model
  • Via log likelihood maximization

$$\hat{q} = \arg \max_{q \in \mathcal{Q}} \sum_{\tau} \sum_{t=0}^{\mathcal{D}_{\text{int}} |\tau|-1} \log q(o_{t+1}|h_t, a_t).$$
Interventional Regime: Example

**Guiding example.** Consider again our door example. If I am able to observe myself or another colorblind person interacting with the door, then I know that which button is pressed is unrelated to which color the light is (light $\not\rightarrow$ button). Then I can directly estimate the causal effect of the button on the door,

\[ p(\text{door}=\text{open}|\text{do(button)}) = p(\text{door}=\text{open}|\text{button}). \]

Whichever policy is used to collect (button, door) samples\(^3\), eventually I realise that button A has more chances of opening the door (60%) than button B (40%), and thus is the optimal choice.
Observational Regime: privileged POMDP

- Learning agent either doesn’t observe or only partially observes state $S_t$
- $S_t$ - unobserved confounder between $O_t$ and $A_t$
- Observational data set $D_{prv}$ of episodes $\tau$ collected from an external agent who has access to privileged information
- $S_t \perp A_t | H_t$ does not hold anymore

\[ p(o_{t+1}|o_{0\rightarrow t}, do(a_{0\rightarrow t})) \neq p_{prv}(o_{t+1}|h_t, a_t). \]
Observational Regime: Example

Guiding example. Take again the door example in figure 3, and assume that you observe someone else interacting with the door. You do not know whether that person is colorblind or not (Light $\rightarrow$ Button is possible). In this regime, without additional knowledge, you cannot recover the causal queries $p(\text{door}=\text{open}|\text{do(button)})$ from the observed distribution $p(\text{door}, \text{button})$. In the do-calculus framework, the queries are said non identifiable. However, if that person was to tell you the light color they see before they press A or B, then you could recover those queries via deconfounding,

$$p(\text{door}=\text{open}|\text{do(button)}) = \sum_{\text{light} \in \{\text{red, green}\}} p_{\text{pru}}(\text{light}) p_{\text{pru}}(\text{door}=\text{open}|\text{light}, \text{button}).$$

This formula eventually yields the correct causal transition probabilities regardless of the observed policy, given that enough (light, button, door) samples are collected\(^5\).
Combining Observational & Interventional data

- Augmented POMDP
  - Policy regime indicator \( I = \{0, 1\} \)
    - 1: Interventional regime
    - 0: Observational regime
- We have data sets \( D_{prv}, D_{std} \)
  \( D_{prv} \sim p(\tau | i = 0) := p_{prv}(\tau) \), and
  \( D_{std} \sim p(\tau | i = 1) := p_{std}(\tau) \).  
- \( S_t \perp A_t | H_t, I = 1 \)
- Transition Model:
  
  \[
p(o_{t+1}|o_{0:t}, do(a_{0:t})) = p_{std}(o_{t+1}|h_t, a_t) = p(o_{t+1}|h_t, a_t, i = 1).
  \]
  \[
  \pi(a_t|h_t, s_t, i = 0) := \pi_{prv}(a_t|h_t, s_t)
  \]
  \[
  \pi(a_t|h_t, s_t, i = 1) := \pi_{std}(a_t|h_t)
  \]
Augmented learning problem

• **Goal:** fit a model $\hat{q}$ that explains the interventional and observational data with discrete latent space $Z$

• Two step process:
  • Learning
  • Inference
Learning

• **Standard maximum likelihood problem** $\hat{q} = \arg \max_{q \in \mathcal{Q}} \sum_{(\tau)} \log q(\tau|i = 0) + \sum_{(\tau)} \log q(\tau|i = 1)$,

$$q(\tau|i = 0) = \sum_{z_0 \to |\tau|} q_{init}(z_0) q_{obs}(o_0|z_0) \prod_{t=0}^{\tau-1} q_{prv}(a_t|h_t, z_t) q_{trans}(z_{t+1}|a_t, z_t) q_{obs}(o_{t+1}|z_{t+1}),$$

$$q(\tau|i = 1) = \sum_{z_0 \to |\tau|} q_{init}(z_0) q_{obs}(o_0|z_0) \prod_{t=0}^{\tau-1} q_{std}(a_t|h_t) q_{trans}(z_{t+1}|a_t, z_t) q_{obs}(o_{t+1}|z_{t+1}).$$

• **four components**: $\hat{q}(z_0), \hat{q}(o_t|z_t), \hat{q}(z_{t+1}|z_t, a_t), \hat{q}(a_t|z_t, i = 0)$, each of which can be approximated using any black-box model, like a neural network.
Inference

• Initialize belief state at $t = 0$ with a prior on $q_{init}(z_0)$:
\[
\hat{q}(z_0|h_0, i = 1) = \frac{\hat{q}_{init}(z_0)\hat{q}_{obs}(o_0|z_0)}{\sum_{z_0} \hat{q}_{init}(z_0)\hat{q}_{obs}(o_0|z_0)}
\]

• Recursively compute causal transition model by propagating beliefs on latent state:
\[
\hat{q}(z_{t+1}, o_{t+1}|h_t, a_t, i = 1) = \sum_{z_t} \hat{q}(z_t|h_t, i = 1)\hat{q}_{trans}(z_{t+1}|z_t, a_t)\hat{q}_{obs}(o_{t+1}|z_{t+1})
\]

• Compute observational probability:
\[
\hat{q}(o_{t+1}|h_t, a_t, i = 1) = \sum_{z_t+1} \hat{q}(z_{t+1}, o_{t+1}|h_t, a_t, i = 1)
\]

• Update belief for next time step:
\[
\hat{q}(z_{t+1}|h_{t+1}, i = 1) = \frac{\hat{q}(z_{t+1}, o_{t+1}|h_t, a_t, i = 1)}{\sum_{z_{t+1}} \hat{q}(z_{t+1}, o_{t+1}|h_t, a_t, i = 1)}
\]
Intuition

• Model learns both distributions: \( q(\tau|i = 0) \& q(\tau|i = 1) \) together by fitting both observational and interventional data

• \( \hat{q}_{prv}(a_t|z_t) \), the privileged model, is the only component that offers limited flexibility to differentiate between regimes

• Thus \( q(\tau|i = 0) \) acts as a regularizer on \( q(\tau|i = 1) \)
  • Prevents overfitting of the interventional data
Theoretical Analysis

• Assume latent space is sufficiently large—so we have enough expressivity to learn true augmented POMDP distribution

• Assume large enough $D_{prv}$ to act as a strong regularizer

Corollary 1. The estimator $\hat{q}(o_{t+1}|h_t, a_t, i = 1)$ recovered after solving (4) with $|D_{obs}| \to \infty$ offers strictly better generalization guarantees than the one with $|D_{obs}| = 0$, for any $D_{int}$.

$\prod_{t=0}^{T-1} q(o_{t+1}|o_0 \to t, a_0 \backslash (o_0 \to t)) \leq \prod_{t=0}^{T-1} p(a_t|h_t, i = 0)p(o_{t+1}|h_t, a_t, i = 0), \forall a_0 \backslash (o_0 \to t) \ \text{Bound aren’t tight}$

$\forall h_{T-1}, a_{T-1}, T \geq 1 \text{ where } p(h_{T-1}, a_{T-1}, i = 0) > 0.$

• Result is a generalization of Manski’s bounds: $P(x,y) \leq P(y|do(X=x)) \leq P(x,y) + 1 - P(x)$
Example

**Guiding example.** Let us now examine our door example in light of Theorem 1. Assume this time that you observe many (button, door) interactions from a non-colorblind person (privileged, $i = 0$), who’s policy is $\pi(\text{button}=A|\text{light}=\text{red}) = 0.9$ and $\pi(\text{button}=A|\text{light}=\text{green}) = 0.4$. Then you can already infer from Theorem 1 that $p(\text{door}=\text{open}|\text{do(button}=A)) \in [0.54, 0.84]$ and $p(\text{door}=\text{open}|\text{do(button}=B)) \in [0.24, 0.94]$. You now get a chance to interact with the door ($i = 1$), and you decide to press $A$ 10 times and $B$ 10 times. You are unlucky, and your interventional study results in the following probabilities: $q(\text{door}=\text{open}|\text{do(button}=A)) = 0.5$ and $q(\text{door}=\text{open}|\text{do(button}=B)) = 0.5$. This does not coincide with your (reliable) observational study, and therefore you adjust $q(\text{door}=\text{open}|\text{do(button}=A))$ to its lower bound 0.54. You now believe that pressing $A$ is more likely to be your optimal strategy.
Experiments: Tiger

• The learning agent receives a noisy signal of the tiger’s position (roar left or roar right).
• It can wait and listen to a new roar at cost of -1 reward, or move left or right.
• Treasure gives +10 reward
• Tiger gives -100 reward.
• The privileged agent knows the exact location of the tiger. The game stops when treasure or tiger is found, or after a maximum horizon of T = 10.
• Contains hidden states
Experiments: Tiger

- Models:
  - Augmented: learning augmented POMDP
  - No obs: discards observational distribution
  - Naïve: combines observational and interventional without accounting for confounding

- Quality of causal transition Model:
  - Likelihood on new interventional data

- Agent Performance:
  - Performance of resulting policy in terms of cumulated reward
  - Privileged agents to collect: $D_{pap}$

- Random (no confounding), noisy good, perfectly good, and perfect bad
Experiments: Hidden Treasures

- the agent must collect a treasure (+1 reward), which is randomly located in one of the four corners.
- The privileged agent knows the treasure’s position at all times, the learning agent doesn’t.
- The treasure is reset to a new location when found, and the game stops after a fixed horizon of $T = 10$. 

Experiments: Sloppy Dark Room

• the agent must reach a treasure (+1 reward) located behind a wall, and slips to a random adjacent tile instead of moving to the chosen direction 50% of the time.

• The privileged agent knows its position at all times, while the learning agent is only revealed its position with 20% chances at each time step, and is blind otherwise (a dummy position is revealed).

• The time horizon is fixed to $T = 30$. 
Experiments: Results

**model likelihood**

- hidden treasures
- sloppy dark room

**agent performance**

- density of agent trajectories
  - at specific time step *

- $|\mathcal{D}_{\text{test}}| = 1000$
  - no obs
  - naive
  - augmented

- $|\mathcal{D}_{\text{test}}| = 2000$
  - no obs
  - naive
  - augmented
Question:

• How does this method differ Bareinboim’s causal Thompson Sampling approach in “Transfer Learning in Multi-Armed Bandits: A Causal Approach”?  
  • Which one do you think is better?