Causality in Real-World RL

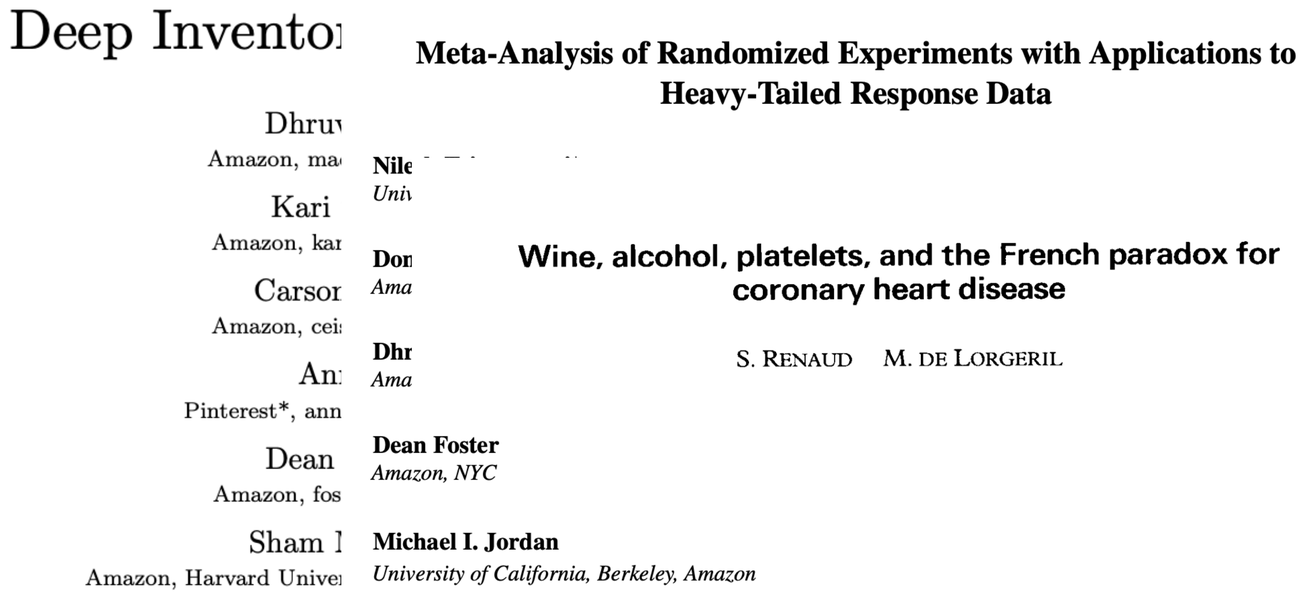
A Practitioners Perspective

Dhruv Madeka*

***Amazon, maded@amazon.com**

Outline

- Part I: Randomization (Easy)
- Part II: Utilizing Observational Data (Medium)
- Part III: Multivariate Regression (Hard)


Deep Inventor

Dhruv
Amazon, ma

Kari
Amazon, kar

Carson
Amazon, cei

An
Pinterest*, ann

Dean
Amazon, fos

Sham
Amazon, Harvard Univer

Meta-Analysis of Randomized Experiments with Applications to Heavy-Tailed Response Data

Nile
Univ

Don
Ama

Dhr
Ama

Dean Foster
Amazon, NYC

Michael I. Jordan
University of California, Berkeley, Amazon

Wine, alcohol, platelets, and the French paradox for coronary heart disease

S. RENAUD M. DE LORGERIL

Largely based on these papers: arxiv/2210.03137
arxiv/2112.07602
10.1016/0140-6736(92)91277-f

I: Randomization

Potential Outcomes Framework

- The potential outcomes framework phrases causality in the following way:

$$Y = \begin{cases} Y(1) & \text{if } T = 1 \\ Y(0) & \text{if } T = 0 \end{cases}$$

- The treatment effect for a unit i becomes:

$$Y_i = Y_i(1) - Y_i(0)$$

- Formally, we augment the random variable space of (T, Y) with $(Y(1), Y(0))$
- Of course we never really observe the “counterfactual” - so what can we do?

Formal Definitions

- Denote by $\mathcal{T}$, the (randomly) assigned set of units for which the treatment is applied
- We are aiming to estimate the Average Treatment Effect Δ:

$$\Delta = \mathbb{E}[Y_i(1) - Y_i(0)]$$

- Then we define the Difference in Means Estimator as:

$$\hat{\Delta}_{DM} = \frac{1}{|\mathcal{T}|} \sum_{i=1}^{N} T_i Y_i(1) - \frac{1}{|\mathcal{C}|} \sum_{i=1}^{N} (1 - T_i) Y_i(0)$$

Unbiasedness of the DM Estimator implies:

$$\mathbb{E}[\hat{\Delta}_{DM}] = \Delta$$

Regression Interpretation

$$Y \equiv Y(0) + T\dot{(}Y(1) - Y(0)) \implies \mathbb{E}[Y \,|\, T] = \alpha + \beta\Delta$$

- Can be interpreted as regressing Y_i on $(1, T_i)$
- Of course we need not restrict ourselves to such simple regressions -
- Define a covariate X as a random variable which is independent of the treatment T
  - Note this allows us to add any covariate observed *before* the randomization
- We can define the “Conditional Average Treatment Effect” (CATE) as:

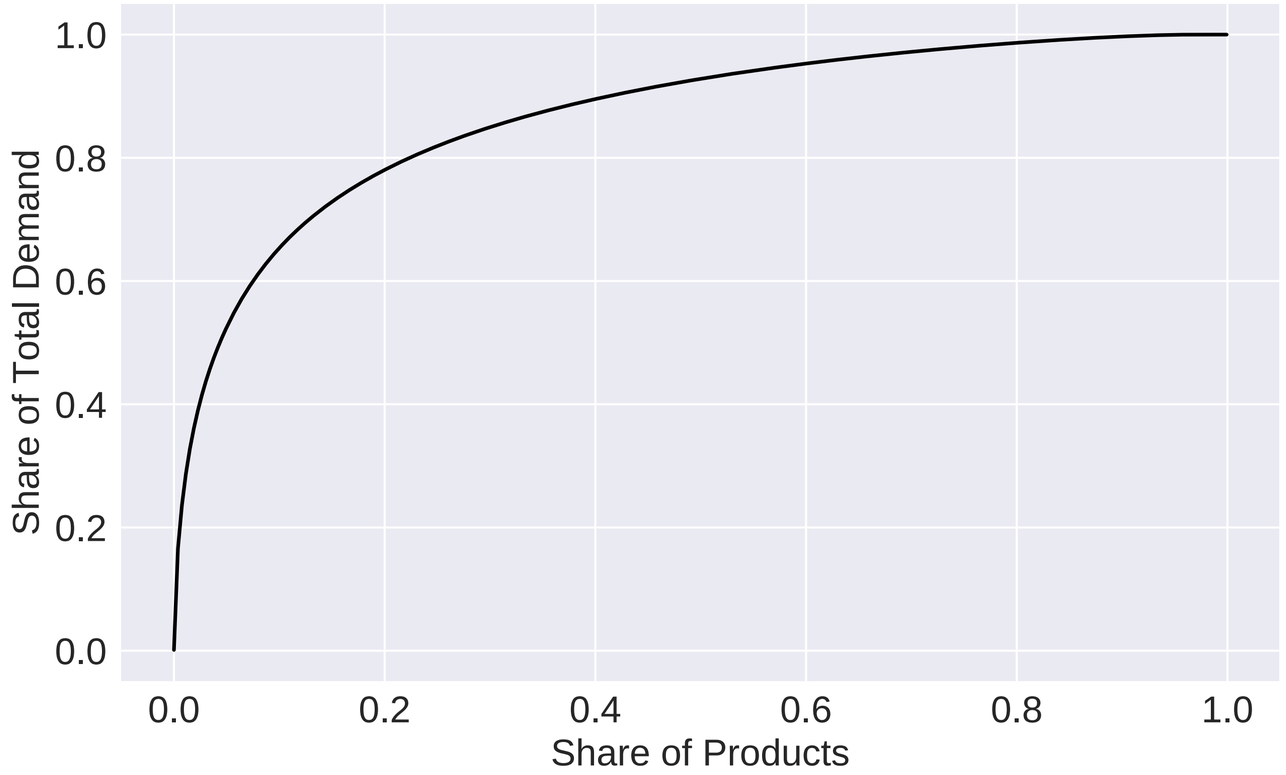
$$\Delta(X) = \mathbb{E}[Y_i(T_i) \,|\, X_i = x, T_i = 1] - \mathbb{E}[Y_i(T_i) \,|\, X_i = x, T_i = 0]$$

- Then define the generalized Difference in Difference Estimator:

$$Y_i = \alpha + T_i\Delta + X_i^T\beta + \epsilon_i$$

Problem with these estimators

- Amazon has an incredibly power-lawed set of data
- A few products contribute to most of the revenue
- Makes classical OLS a really bad tool!



What about in the real world?

- Traditionally the solution in supervised learning is to use Weighted Least Squares
- Key idea is to downweight large residuals in OLS to deal with the heteroskedasticity that the power law induces

$$\text{argmin}_{\beta} ||Y - X^T\beta||_2^2 \quad \longleftarrow \quad \text{OLS}$$

$$\text{argmin}_{\beta_{WLS}} ||W^{\frac{1}{2}}(Y - X^T\beta_{WLS})||_2^2 \quad \longleftarrow \quad \text{WLS}$$

What's the issue with this?

$$Y_i = \alpha + X_i^T \beta_{WLS} + \frac{\epsilon_i}{W_i}$$

$$\implies \hat{\beta}_{WLS} = (X^T W X)^{-1}(X^T W Y)$$

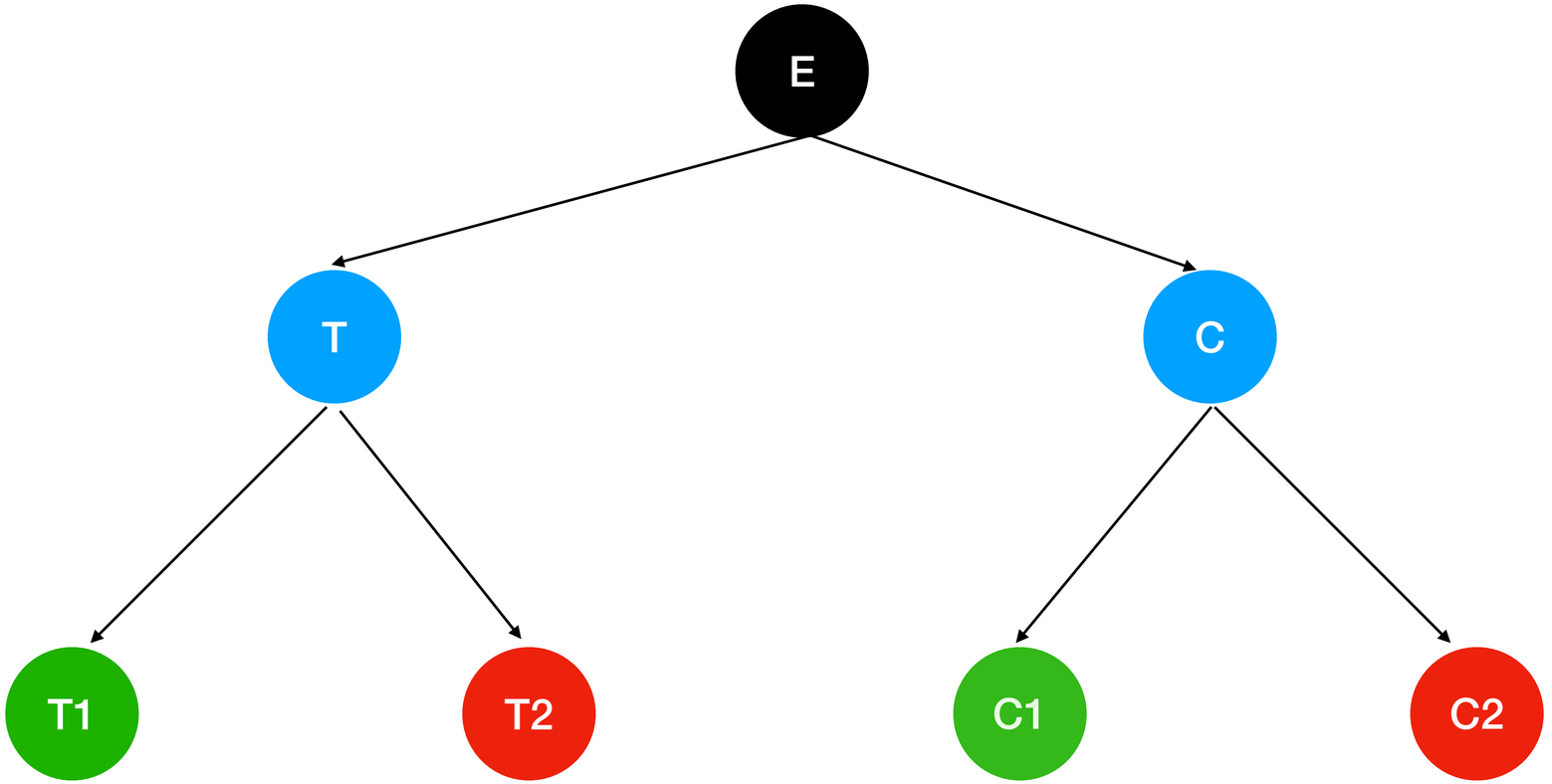
- WLS produces biased estimates if the weights depend on the covariates!

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_{WLS}] - \beta &= \mathbb{E}[(X^T W X)^{-1}(X^T W Y)] - \beta \\
&= \mathbb{E}[(X^T W X)^{-1}(X^T W \epsilon)] \\
&= Cov((X^T W X)^{-1} X^T W, \epsilon)
\end{aligned}$$

What do we do?

- We can’t really estimate any causal effects with such power lawed data
- And we can’t magically believe (ala Pearl/Rubin etc) that we can wish away the bias
- But, we do work at a tech company
- Which means we have a LOT of randomized trials. Can we use those in anyway to fit more complex “causal” models?

So how do we perform Model Selection?



Treat one split as “in-sample”, treat the other as out of sample

Train a model on “in-sample”, and test it “out of sample”

Theorem: Sample Splitting is a valid procedure

Theorem [Tripuraneni, Joncas, M., Foster, Jordan '22]:

Consider two estimators A and B of the ATE Δ. If T, C are independent and T_1, T_2, C_1, C_2 are independent then we have that:

$$\mathbb{E}[(\hat{\Delta}_A(T_1, C_1) - \hat{\Delta}_{DM}(T_2, C_2))^2] - \mathbb{E}[(\hat{\Delta}_B(T_1, C_1) - \hat{\Delta}_{DM}(T_2, C_2))^2] =$$
$$\mathbb{E}[(\hat{\Delta}_A(T_1, C_1) - \hat{\Delta}_{DM}(T_2, C_2))^2] - \mathbb{E}[(\Delta_B(T_1, C_1) - \Delta_{DM}(T_2, C_2))^2]$$

- Implications:
  - We can rank *causal* estimators based on their out of sample performance
  - We can train complex (possibly biased) estimators as our causal models of the world

Win Table for Different Estimators

On 800 Amazon Supply Chain Trials

- We see how often one estimator wins against the other (Borda counts)

Method	dm	mom1000	gen_dd	gen_dd_w1	dm_wins.001
dm	x	(-3.58, 0.000363)	(-12.68, 2.38e-33)	(-22.36, 3.6e-84)	(-28.19, 7.99e-118)
mom1000	(3.58, 0.000363)	x	(-2.12, 0.0342)	(-11.89, 7.32e-30)	(-13.51, 3.78e-37)
gen_dd	(12.68, 2.38e-33)	(2.12, 0.0342)	x	(-21.1, 4.73e-77)	(-19.01, 2e-65)
gen_dd_w1	(22.36, 3.6e-84)	(11.89, 7.32e-30)	(21.1, 4.73e-77)	x	(-0.26, 0.794)
dm_wins.001	(28.19, 7.99e-118)	(13.51, 3.78e-37)	(19.01, 2e-65)	(0.26, 0.794)	x

- Implying:

$$\mathbf{gen\_dd\_w1\_wins.001} > \mathbf{gen\_dd\_wins.001} > \mathbf{dm\_wins.001} \approx \mathbf{gen\_dd\_w1} > \mathbf{gen\_dd} > \mathbf{mom1000} > \mathbf{dm}$$

Decision Making for Randomizations

- When we're running a Supply Chain (or search engine, or social media site), mere estimation is not enough
- What we actually want to optimize is the decision to launch a new policy or not. This is a "meta" policy
- Consider a Decision Policy for a new product (I) D_I. We wish to optimize across a series of product decisions

$$\text{argmax}_D \quad S(\Delta, \hat{\Delta}) = \text{argmax}_D \sum_{I \in \text{Products}} \Delta_I D(\hat{\Delta}_I)$$

Can we use sample splitting for decision making?

- Simple procedure: Estimate $\hat{D}_I(\hat{\Delta}_A(T_1, C_1))$ for some estimator A
- Evaluate the "reward" from the sample splits $\hat{\Delta}_{I,DM}(T_2, C_2)$
- Giving us the following:

$$\hat{S}(\hat{\Delta}_A) = \sum_{I \in \text{Products}} \hat{\Delta}_{I,DM}(T_2, C_2)\hat{D}(\hat{\Delta}_{I,A}(T_1, C_1))$$

Theorem: Sample Splitting for launch decisions

Theorem [Tripuraneni, Joncas, M., Foster, Jordan '22]:

Consider two estimators A and B of the ATE Δ. If T, C are independent and T_1, T_2, C_1, C_2 are independent then we have that:

$$\mathbb{E}[\hat{S}(\hat{\Delta}_A)] - \mathbb{E}[\hat{S}(\hat{\Delta}_B)] = \sum_{I \in \text{Products}} \Delta_I(\mathbb{E}[D_I(\hat{\Delta}_A)] - \mathbb{E}[D_I(\hat{\Delta}_B)])$$

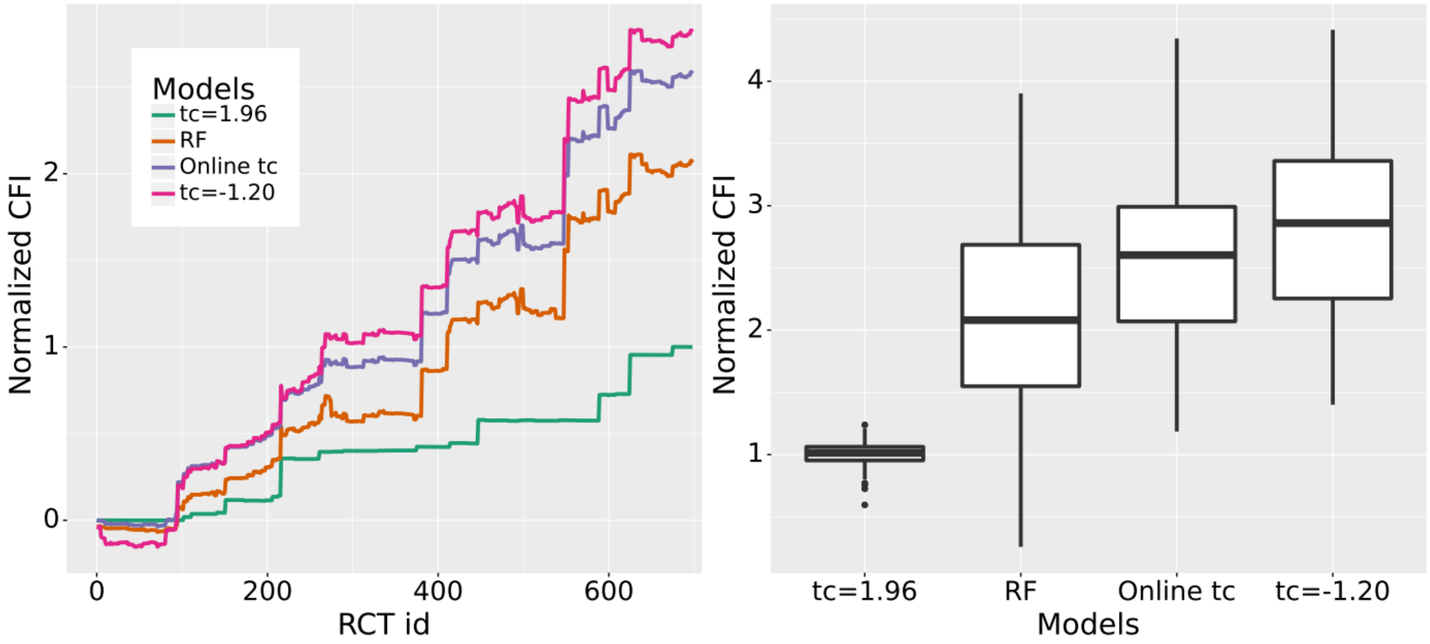
- Implications:
  - We can rank "launch" policies directly on their out of sample performance
  - By passes the need to worry about "when" to launch a product

Making optimal launch decisions

- Using an online bandit to make the decision works as well as anything basically

- Surprisingly, for Supply Chain Data using a t-value threshold of -1.2 is good!

- Indicates the human effort put in is worth every launch!

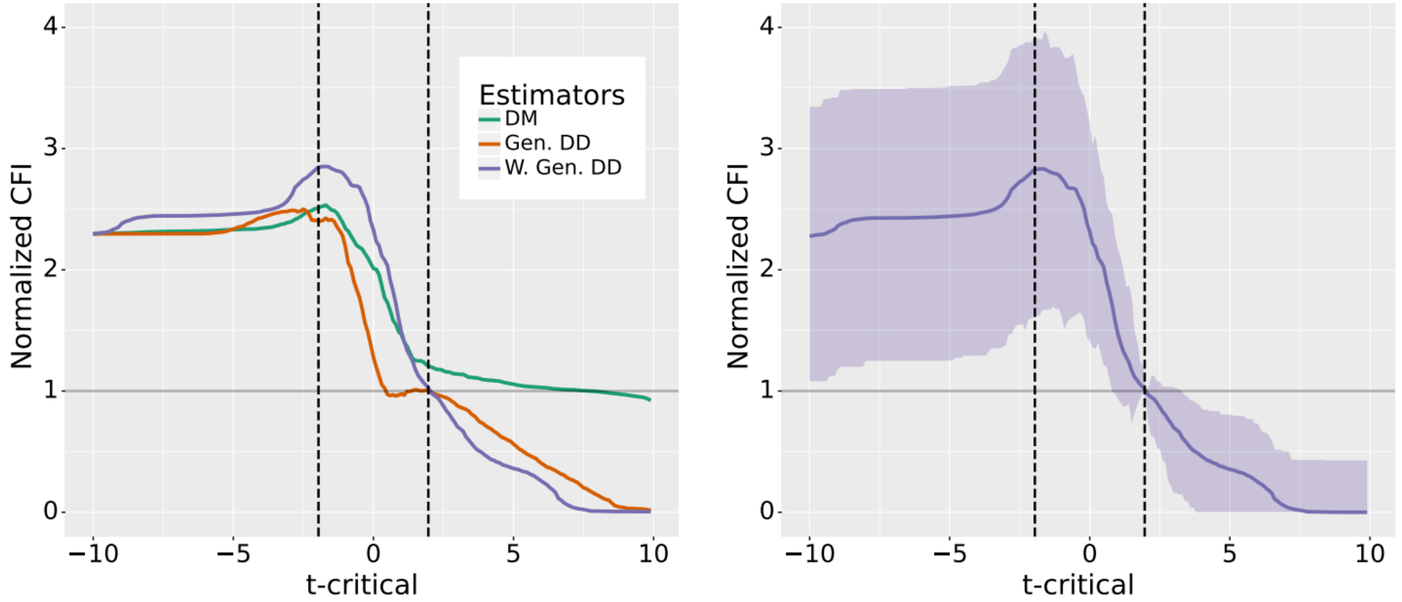


Making optimal launch decisions

- Using an online bandit to make the decision works as well as anything basically

- Surprisingly, for Supply Chain Data using a t-value threshold of -1.2 is good!

- Indicates the human effort put in is worth every launch!



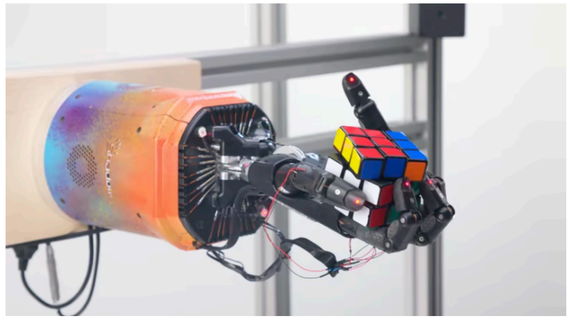
II: Utilizing observational data

RL is hard!

- Sample complexity can be as large as $\min(|\Theta|, 2^T)$
- Large state/action spaces
- Exploration
- Credit assignment problem

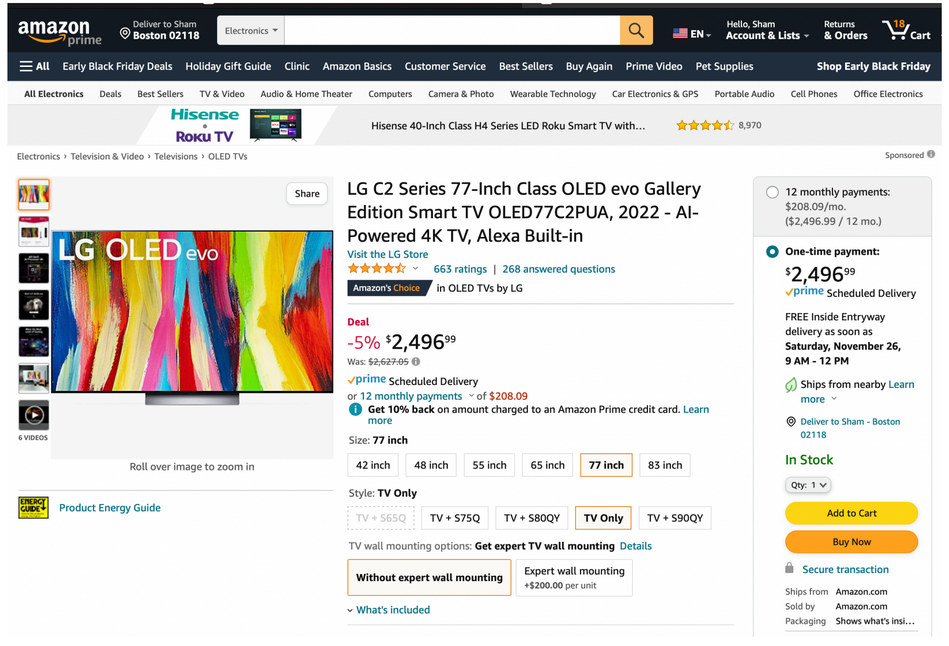
Dexterous Robotic Hand Manipulation
OpenAI, '19

Real-world RL is hard.



The core challenges Amazon faces are sequential decision making problems.

Can RL help in this space?



RL is hard!

- Exploration and Credit Assignment are trying to solve the same problem
  - Learning the “causal” structure of the world
- In RL notation: $\mathbb{P}[s_{t+1} \mid s_t, a_t], R(s_t, a_t)$
- The dependence on the action separates this from conventional supervised learning

The Supply Chain Problem

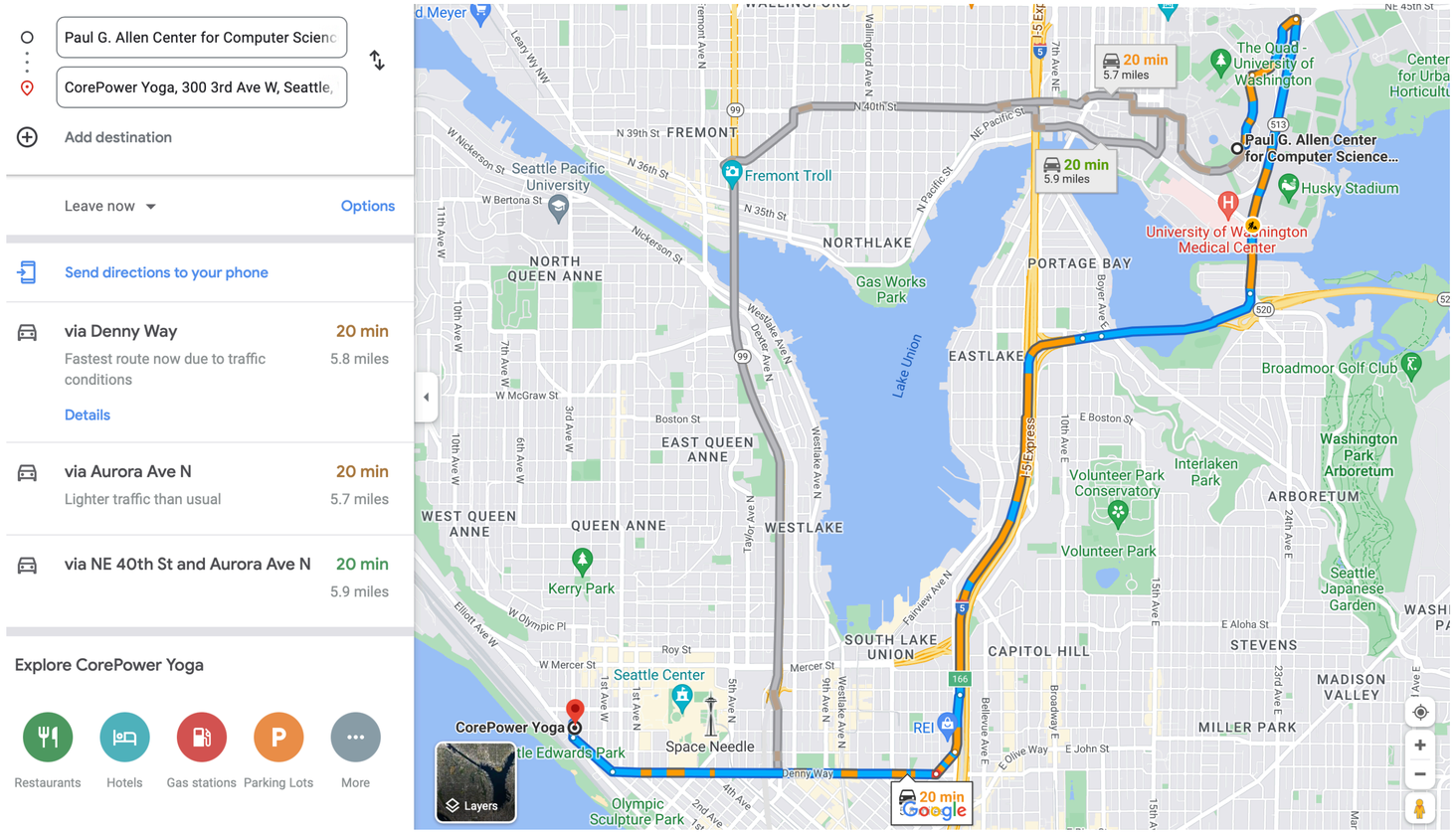
- Supply Chain is about buying, storing, pricing, and transporting goods.
- Amazon has been running it's Supply Chain for decades now
  - There is a lot of historical "off-policy" data
  - How do we use it?
  - Concern: counterfactual issue?
- This talk: how can we use this data to solve the inventory management problem?

A practical approach to Real-World RL

- Some problems inherently duck the counterfactual issue
- If our actions don’t really affect the world, we can ignore causality and frame the problem as “supervised learning”
- This is what “ExoMDP”s do

Warm up: Vehicle Routing

(when using historical data might be ok)



- We want a good policy for routing a single car.

- Policy π: features -> directions

  features:

  time of day, holiday indicators, current traffic, sports games, accidents, location, weather,

- Historical Data:

  suppose we have logged historical data of features

- Backtesting policies:
  - Key idea: a single route minimally affects traffic
  - Counterfactual: with the historical data, we can see what would have happened with another policy.

Warm up 2: Fleet Routing

WAYMO

- We want to route a whole fleet of self-driving taxis.

- Policy π: features -> directions
  - features:
    customer demand, time of day,
    holiday indicators, current traffic, sports games,
    accidents, location, weather…

- Historical Data:
  suppose we have logged historical data of features

- Backtesting policies:
  - Key idea: a small fleet route may have small affects on traffic.
  - Counterfactual: with the historical data, we can see what would have happened with another policy.

Supply Chain Data

Price= $2
Cost= $1

Time	Inventory	Demand	Order	Revenue
0	100	20	-	40
0	80	-	10	-10
1	90	20	-	40
1	70	-	50	-50
2	120	60	-	120
2	60	-	10	-10

Backtesting a policy

Time	Inventory	Demand	Order	Revenue
0	100	20	-	40
0	80	-	~~10~~ 40	~~-10~~ -40
1	~~90~~ 120	20	-	40
1	~~70~~ 100	-	~~50~~ 20	~~-50~~ -20
2	120	60	-	120
2	60	-	10	-10

Price= $2

Cost= $1

- Current order doesn't impact future demand.
  - This allows us to backtest!
  - Empirically, backlog due to unmet demand does not look significant.[1]

1. See Verhoef et al (2006)

Formalization of the Supply Chain Problem

- Growing literature around a class of MDPs where a large part of the state is driven by an exogenous noise process [Efroni et al 2021, Sinclair et al 2022]

- A formalization of the model:
  - Action a_t: how much you buy
  - Exogenous random variables: evolving under $\Pr$ and not dependent on our actions
    $(\text{Demand}_t, \text{Price}_t, \text{Cost}_t, \text{Lead Time}_t, \text{Covariates}_t) := s_t$
  - Controllable part (inventory) I_t: evolution is dependent on our action.
    - $I_t = \max(I_{t-1} + a_{t-1} - D_t, 0)$ (and suppose we start at I_0).
  - Reward is just the sum of profits: $r(s_t, I_t, a_t) := \text{Price}_t \times \min(\text{Demand}_t, I_t) - \text{Cost}_t \times a_t$

- Learning setting:
  - Data collection: We observe N historical trajectories, where each sequence is sampled $s_1, \ldots, s_T \sim \Pr$
  - Goal: maximize our cumulative reward over T periods

$$V_T(\pi) = E_\pi\left[\sum_{t=1}^{T} \gamma^t r(s_t, I_t, a_t)\right]$$

Causal Freeness

- ExoMDPs have a mapping to the classical Causality language
- In the original construction, several random variables were defined

$$(\text{Demand}_t, \text{Price}_t, \text{Cost}_t, \text{Lead Time}_t, \text{Covariates}_t) := s_t$$

- And then an MDP was constructed on top of them
  - That is all the state evolution and actions were defined as functions of these variables

Causal Freeness

- We could instead use a more measure-theoretic definition of causal freeness
- The spirit is similar to the augmentation construction of causality

ASSUMPTION 1 **(Causally free)**. *Let $\mathcal{F}_t^i$ be the sequence of sigma fields generated by all the random variables up to time t, namely $\mathcal{F}_t^i = \sigma(\{D_s^i, p_s^i, c_s^i, v_s^i, a_s^i\}_{s \leq t})$. Let $\mathcal{G}_t^i = \sigma(\{D_s^i, p_s^i, c_s^i, v_s^i\}_{s \leq t})$, namely the sigma field generated by everything that isn't an action. We will say that D_t^i is causally free of θ if*

- *$(\forall d)$ and $(\forall t)$ there exists a random variable $F_t^{i,d}$ which is $\mathcal{F}_t^i$ measurable and*

$$(\forall \theta \in \Theta) \qquad \mathbb{P}_\theta^i(D_{t+1}^i \leq d | \mathcal{F}_t^i) \overset{a.s.}{=} F_t^{i,d}$$

and

- *D_{t+1}^i and the previous actions are conditionally independent, so $D_{t+1}^i \perp \{a_s^i\}_{s \leq t} | \mathcal{G}_t$ under all measures $\mathbb{P}_\theta^i$.*

Likewise we will assume that the price process p_t^i, the cost process c_t^i and the vendor lead time process v_t^i are all causally free of θ.[4]

Theorem: Backtesting in ExoMDPs

Theorem [M., Torkkola, Eisenach, Luo, Foster, Kakade '22]:

Suppose we have a set of K policies $\Pi = \{\pi_1, \ldots \pi_K\}$, and we have N sampled exogenous paths. Then we can accurately backtest up to nearly $K \approx 2^N$ policies.

Formally, for any $\delta \in (0,1)$, with probability greater than $1 - \delta$ - we have that for all $\pi \in \Pi$:

$$| V_T(\pi) - \hat{V}_T(\pi) | \leq T \sqrt{\frac{\log(K/\delta)}{N}}$$

(assuming the reward r_t is bounded by 1).

- Implications:
  - We can optimize a neural policy on the past data.
  - In the usual RL setting (not exogenous), we would have an amplification factor of (at least) $\min\{2^T, K\}$, using historical data due to the counterfactual issue.

What do ExoMDPs buy us?

- In classic (tabular) RL, we typically need the max error over states and actions to be bounded:

$$\max_{s,a} \|P(\cdot|s,a) - \widehat{P}(\cdot|s,a)\|_1 \leq (1-\gamma)^2\epsilon\,.$$

- In an ExoMDP, if we were to use a generative model, we only need a good average case prediction for the future:

$$\frac{1}{N}\sum_{i=1}^{N} \text{TotalVar}\left(\mathbb{P}^i, \widehat{\mathbb{P}}^i\right) \leq \epsilon_{\text{sup}} \implies |V_T(\pi) - \hat{V}_T(\pi)| \leq T\left(\epsilon_{\text{sup}} + \sqrt{\frac{\log(K/\delta)}{N}}\right)$$

- This allows us to avoid the counterfactual/causality issue

The Simulator

- Collection of historical trajectories:
  - 1 million products
  - 104 weeks of data per product

- Uncensoring:
  - Demand
  - Vendor Lead Times

- Policy gradient methods in a “gym”:
  - “gym” ↔ backtesting ↔ simulator
    (note the “simulator” isn’t a good world model).
  - The policy can depend on many features.
    (seasonality, holiday indicators, demand history, product details, text features)

The Simulator

- Policy gradient methods in a “gym”:
  - “gym” ↔ backtesting ↔ simulator
    (note the “simulator” isn’t a good world model).
  - The policy can depend on many features.
    (seasonality, holiday indicators, demand history, product details, text features)

- Note that the gym is not a true “simulator” in the usual sense
  - It does not simulate every possible starting state, only the historical ones
  - It is a collection of diracs at the historical starting states

    Different from optimal control, and traditional DeepRL

Differentiable Control Problem

- Note that each term of our state evolution is a differentiable function of previous actions

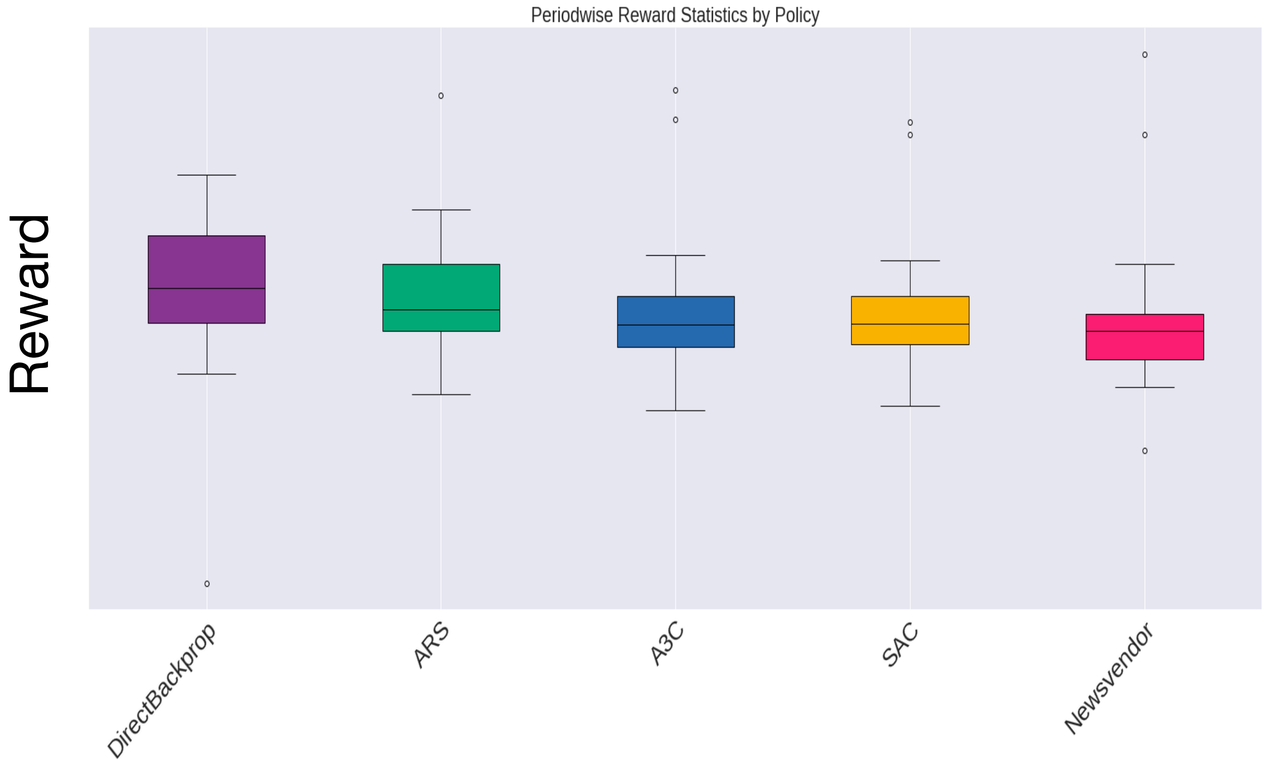
- So, we can take gradients directly from our Reward through our policy

- This is our current production policy, called *DirectBackprop*

Sim to Real Transfer

- Sim: the backtest of DirectBackprop improves on Newsvendor.
- Real: DirectBackprop significantly reduces inventory without significantly reducing total revenue.

Simulation



Real World

Metrics	% change
Inventory Level	**-12** ± 6
Revenue	**2.6%**

III: Multivariate Regression

The French Paradox

- French people eat “worse” than Americans
  - Higher fat
  - They drink wine
  - They smoke more
- Yet they live longer
- Why?

Is it red wine?

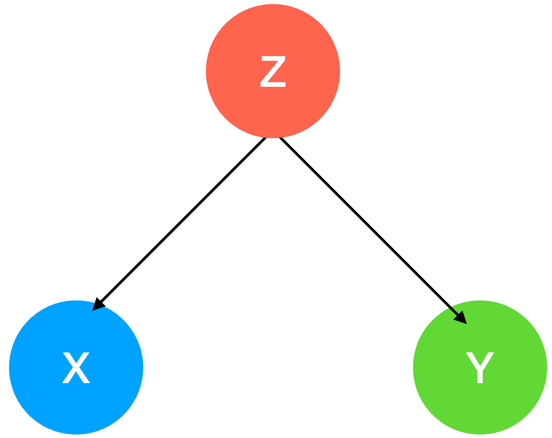
- Renaud (1991/1992) conjectured that it was the small amount of red wine they consumed that caused this

- Hard to randomize this!

- So how do we get a causal effect?

Key issue in Causality: Lurking Variables

- If Z causes X and Z causes Y, we see correlation but not causality between X and Y
- We can randomize X to break linkages
- Or if we know Z, we can control for Z



The French Paradox

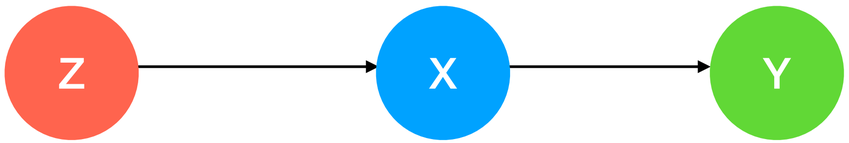
- When randomization is hard, problem becomes finding missing covariates
- Possible explanations:
  - Wealthy people drink more red wine, being wealthy makes you live longer
  - French people walk more, walking is the relevant covariate
  - People who drink are more relaxed, being relaxed is the real win
- One solution:
  - Multiple Regression

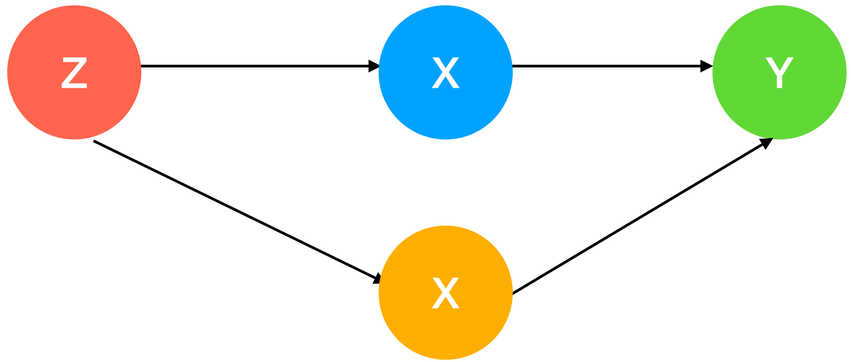
Should you drink a glass of red wine every night?

- Renaud's (1991/1992) talk/paper considered
  - Income
  - Education
  - French people in the US
  - Americans in France….
- I read it and started drinking a glass of red wine every night!

Does the “Pearl” perspective add anything?

- I think what the Pearl perspective gets right is that a causal model can only be falsified (or fail to be falsified)
- What it gets wrong is that finding the variables in the graph is the hard problem
- Consider an instrumental variable approach:





A practitioners “recipe” for causality

- We wanted to answer the question - “Does more inventory lead to more demand?”
- Regress Demand on Inventory
- Then throw the kitchen sink of covariates
- If the t-stat on the OLS coefficient doesn’t change, likely the variable is “causal”

Spot the Error

- Taller people die younger (about 1 inch per year)
  - If you were tall, you were born
  - So the only tall people who di
  - Missing covariate was birth yea
- Taller people get paid more
  - Taller people are second or third generation American
  - Flynn Effect is working for them

SCIENCE

I Wish I Was a Little Bit Shorter

The research is clear: Being tall is hazardous to your health.

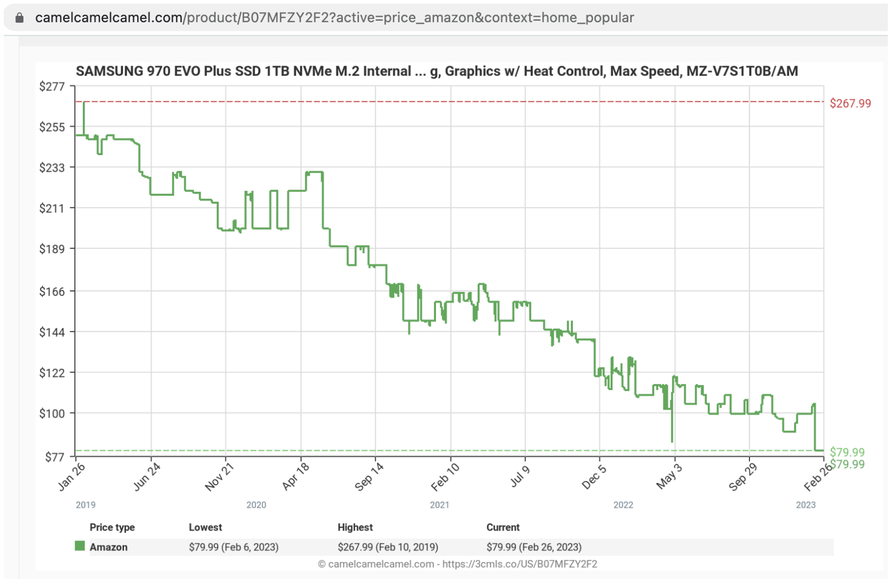
BY BRIAN PALMER JULY 30, 2013 • 12:08 PM

Marshmallow Test

- Adolescents delaying gratification have better outcomes in life
- Missing covariate:
  - Income
  - Education
- Possible explanations:
  - Parental investment
  - Circumstances

When we can not do this?

- Pricing is an **incredibly** hard problem
  - We want to understand how an x% change in price affects Amazon
  - What happens if we randomize price every week
  - People start gaming the price!
- Strategic behavior through time means the response is not truly random



Causal Confusion in Imitation Learning

- The example given:
  - Scenario A: Image with Dashboard ar
  - Scenario B: Image without Dashboar
- B does better than A (explanation is indicator light)
- Implies “causal confusion” according to the paper

Causal Confusion in Imitation Learning

Pim de Haan*[1], Dinesh Jayaraman†‡, Sergey Levine†
*Qualcomm AI Research, University of Amsterdam,
†Berkeley AI Research, ‡ Facebook AI Research

Causal Confusion in Imitation Learning

- Incorrect causal reasoning!
  - Causality isn’t the issue here, the issue is a trivially improvable error
  - Mask out the dashboard indicator since there is already a redundant signal from the break light
  - According to them, this would be a strictly better model
- This isn’t “causality” - it’s a claim on a better model

Causal Confusion in Imitation Learning

- Incorrect causal reasoning!
  - If past states and actions affect the future, adding them as features would be a trivially better model
  - If they merely confuse the model, masking them out in the first layer would be a trivially better model
- Simpler solution then graphs/ disagreement scores etc

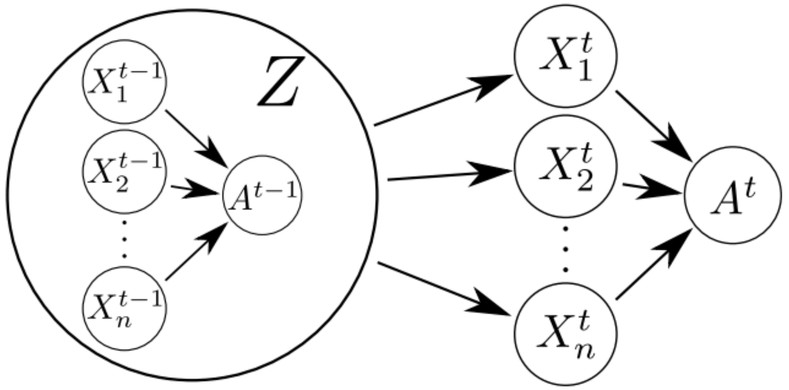
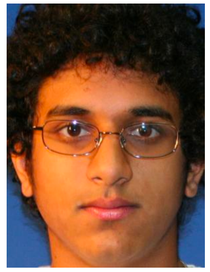


Figure 2: Causal dynamics of imitation. Parents of a node represent its causes.

Conclusion

- There are classes of RL problems that duck the issue of causality
- One can validate more complex causal models through a meta-analysis
- Causality is primarily about missing confounders, identifying them is the hard problem



Nilesh

Mike

Dominique

Dean

Carson

Kari

Anna

Dean

Sham