Causal Graphical Models

Professor Ajoodha

Causal Graphical Models

Causality

Professor Ajoodha

Lecture 9

School of Computer Science and Applied Mathematics
The University of the Witwatersrand, Johannesburg

ExplainableAI Lab

— MODELLING. DECISION MAKING. CAUSALITY —

Problem Statement

- As long as we can represent P^* (standard probabilistic queries) causal structure does not matter.
- Is there any value in **imposing causal semantics**?
- Bayesian networks are insufficient for causality:
  - Bayesian networks do not always capture underlying causal mechanisms
  - Bayesian networks therefore can not capture how interventions will affect the system
- E.g. reasoning about situations where we **intervene**.
  - E.g. Will **preventing** smoking cause the frequency of lung cancer to decrease?
  - E.g. Would you have passed the PGM exam **if you skipped** all the lectures?

Conditioning and Intervention

- Causal models are needed for interventions.
- An intervention is when you **take a dedicated action** that manipulates the value.
- We can perform interventions of the form $do(Z := z)$, which forces the variable Z to take the value z.
  - E.g. Intervention queries: $P(\mathbf{Y} \mid do(z), \mathbf{X} = \mathbf{x})$.

Medical Example

- "If we get a patient to take this medication, what are her chances of getting well?"
- $P(H \mid do(M := m^1))$, where H is patient health, and $M = m^1$ is her taking meds.
- **Not the same as** $P(H \mid m^1)$, which is probability of patient health given the medication.
- $P(H \mid m^1) \geq P(H \mid do(M := m^1))$, if patients who take the medication on their own are more likely to be health-conscious, the therefore more healthy in general.
- $P(H \mid do(M := m^1))$ is the **specific effect of the medication** on the individual patient's health.

More Examples of Ideal Intervention Queries

- If we lower the price of hamburgers, will people buy more ketchup?
- If we lower the interest rates, will that give rise to inflation?
- If we were to force someone to smoke, would they be more likely to get cancer?
- If you study PGMs every night, would you be more likely to get a A on the exam?
- If you drink alcohol everyday, would you be brain dead by 50?
- If we increase student fees by 10%, would this cause student protests?

Counterfactual Queries

- Another query is the counterfactual query.
- Here we have some information about the true state of the world ...
- ... and we want to inquire about the state the world would be in **had we intervened**.
- For example:
  - "Would the Rand value **have dropped** if we **rather elected Mmusi Maimane** as president instead of Cyril Ramaphosa in the 2019 election?"
  - "Would the accident occurred if the car was driven by a human instead of machine?"
  - "If a car does not start, but where the lights work; will replacing the battery make the car start?"

Intervention VS Probabilistic Query Example

- Analyse how studying affects test scores of 100 students
- Data has been collected on the study hours of each student and their corresponding test scores.
- The dataset includes variables for study hours, test scores, and tutoring intervention status.
  1. Hours studied (X): $X \in (0, 100)$
  2. Test score (Y): $Y \in (0, 100)$
  3. Tutoring intervention

$$Z = \begin{cases} 1, & \text{if the student received tutoring} \\ 0, & \text{otherwise} \end{cases}$$

(1) **Probabilistic Query**: What is the probability of a student studying for 5 hours and scoring above 80?

(2) **Intervention Query**: What is the causal effect of tutoring intervention on average test scores for students studying 8 hours?

Intervention VS Probabilistic Query Example

(1) **Probabilistic Query**: What is the probability of a student studying for 5 hours and scoring above 80?

- Assume that 20 students studied for 5 hours, and 10 of them scored above 80.
- Then $P(Y > 80 \mid X = 5) = 10/20 = 0.5$
- Thus, the chances of a student who studies 5 hours and gets above 80% is about 50%.

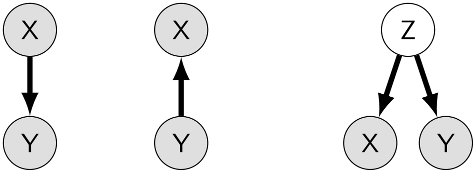
Intervention VS Probabilistic Query Example

(2) **Intervention Query**: What is the causal effect of tutoring intervention on average test scores for students studying 8 hours?

- Assume out of 100 students, 30 studied for 8 hours without tutoring intervention (average test score: 70), while 15 students studied for 8 hours with tutoring intervention (average test score: 85).
- Assuming that there are **no confounding factors**:
- Causal effect = Average test score with intervention - Average test score without intervention
- Causal effect = 85 - 70 = 15
- Thus, the tutoring intervention increased the average test score by 15 points for students studying 8 hours.

Confounding Factors

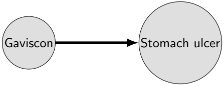
- What can we say about the **causal relationship** between two **probabilistically correlated** variables?
- Multiple causal scenarios can produce variable correlations:

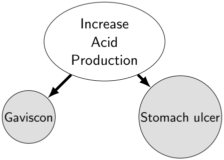


- Knowing Z can disentangle causal relationships between X and Y.
- Latent variables can create correlations that mimic causality
- Some correlations do not indicate direct causation:
  1. Confounding Factors
  2. Selection Bias

Confounding Factors

- Latent variables can model a sparse representation.
- Even if we **do not know** all the latent variables, if we have the correct marginal distribution, then **we can answer** probabilistic queries.
- However, for causal queries, it is essential to **disentangle causal relationships** due to confounding factors.
- In complex real-world settings it is impossible to identify all relevant latent variables and quantify their effects.





Selection Bias

Selection bias is when the population that the distribution represents is a segment of the population that exhibits atypical behaviour.

1. For example, suppose WITS sends a survey to its alumni about academic and sports history.
2. The survey result shows a negative correlation between athletic activities and high student marks.

   What does this mean?

   - Participating in athletic activities reduces ones marks; or
   - Students with high marks tend not to participate in athletic activities?

Selection Bias cont.

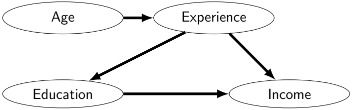
- Another explanation could be that respondents are not a representative segment population.
- Respondents only included alumni who got good marks, or were good at athletic activities.
- Students who did good/bad in either, tended not to respond.
- Without accounting for selection bias, we may falsely explain the correlation using a causal relationship.

Other factors

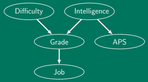
- Other factors that can contribute to non-causal correlations:
  - **Common Cause**: Correlation due to shared influence, not implying direct causation.
  - **Noisy Data**: Non-causal correlations arising by chance, especially in large datasets.
  - **Data Mining**: Testing multiple relationships can lead to spurious correlations without causal basis.
  - **Measurement Errors**: Inaccuracies or biases in measurements introducing false associations.
  - **Time-related Trends**: Similar patterns without causal link, influenced by shared temporal or external factors. (E.g. pressure and rain activity)

Causal Model

- We can view a Bayesian network as a causal model:
  1. Each variable is governed by a causal mechanism that (stochastically) determines its value based on the values of its parents.
  2. A causal mechanism takes the same form as a standard CPD.
  3. The edges represent direct causes between variables.
  4. Causality flows in the direction of the edges.
- A causal model, $\mathcal{C}$, over $\mathcal{X}$ is a Bayesian network that can answer both conditional queries and intervention queries.



Intervention VS Probabilistic Query



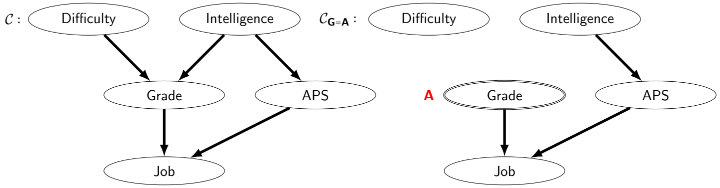
- $P(\mathbf{Y} \mid do(z), \mathbf{X} = \mathbf{x})$ verses $P(\mathbf{Y} \mid \mathbf{Z} = \mathbf{z}, \mathbf{X} = \mathbf{x})$
- Suppose Sipho gets an A in class. Then
  1. Probability of intelligence increases
  2. Probability of high APS increases
  3. Probability of good job increases
- Now suppose Sipho bribes the professor to get an A. The manipulation of the grade should not influence the intelligence or APS. Then
  1. $P(i^1 \mid do(g^1)) = P(i^1)$
  2. $P(a^1 \mid do(g^1)) = P(a^1)$
  3. In the case of getting a job, the transcript is only observed so $P(J \mid do(g^1)) = P(J \mid g^1)$.
- **Intuition:** The intervention query on grade should only affect the descendants of grade.

Intervention Queries

- Intervention queries can be answered by a causal model.

$$P_{\mathcal{C}}(\mathbf{Y} \mid do(z), \mathbf{x}) = P_{\mathcal{C}_{\mathbf{Z}=\mathbf{z}}}(\mathbf{Y} \mid \mathbf{x})$$

- When we intervene in the original network, $\mathcal{C}$, by making Siphos grade an A, we use a mutilated network ($\mathcal{C}_{\mathbf{Z}=\mathbf{z}}$).



- Knowing the difference between a observational query and an intervention query has helped us solve some interesting problems.

Simpson's Paradox

Suppose that we are trying to test a drug to cure a disease.

1. 60% **who took** the drug are cured. $P(c^1 \mid d^1)$
2. 40% **who did not** take drug are cured. $P(c^1 \mid d^0)$

We may now believe that the drug is beneficial. However,

Female patients:

3. 75% **who took** are cured. $P(c^1 \mid d^1, f)$
4. 100% **who did not** take the drug are cured. $P(c^1 \mid d^0, f)$

Male patients:

5. 0% **who took** the drug are cured. $P(c^1 \mid d^1, m)$
6. 25% **who did not** take the drug are cured. $P(c^1 \mid d^0, m)$

$$P(c^1 \mid d^1) > P(c^1 \mid d^0)$$
$$P(c^1 \mid d^1, m) < P(c^1 \mid d^0, m)$$
$$P(c^1 \mid d^1, f) < P(c^1 \mid d^0, f)$$

A Closer Look (at those cured)

	Female or Male Together
Take Drug	3 out of 5 (were cured) (60%)
Don't Take	2 out of 5 (were cured) (40%)

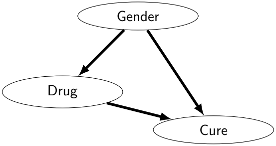
- It looks like the treatment is working!

	Female	**Male**
Take Drug	3 out of 4 (75%)	0 out of 1 (0%)
Don't Take	1 out of 1 (100%)	1 out of 4 (25%)

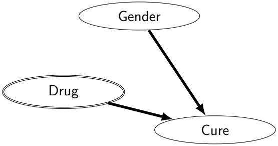
- Wait? The treatment is now not working!?

Understanding Simpson's Paradox

- Taking the drug is correlated with gender.
- Women are much more likely to take the drug than men.
- That is 75% percent of females take the drug and 0% of men take the drug.
- So which statistic do we use: the likelihood $P(c^1 \mid d^1)$ or the causal $P(c^1 \mid do(d^1))$?
- We must use the causal query which uses the mutilated network



$P(c^1 \mid d^1)$



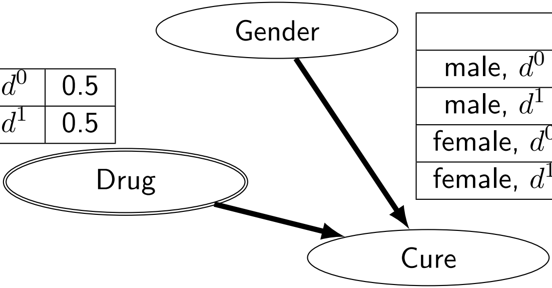
$P(c^1 \mid do(d^1))$

Visualising the Mutilated network

male	0.5
female	0.5

d^0	0.5
d^1	0.5

	c^0	c^1
male, d^0	0.75	0.25
male, d^1	1	0
female, d^0	0	1
female, d^1	0.25	0.75



$$P(c^1 \mid do(d^1))$$

Structural Causal Identifiability

- We can now answer intervention queries **given a fully specified causal model**.
- However, fully specifying a causal model is **often impossible**.
- It seems that causal inquiries may be a hopeless task since latent variables are **unavoidable**.
- ✏ Fortunately, sometimes we can answer causal queries in models which **involve latent variables** using observed correlations alone.
- We now attempt to address the question of **which intervention queries are identifiable**.

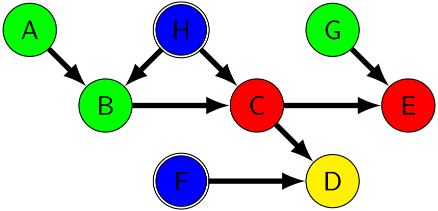
Query Simplification Rules

- The **causal model structure** provides equivalence rules over interventional queries
- These rules can **simplify queries** by replacing them with equivalent ones.
- This can convert a causal query to a **non-interventional one** that can be answered using observational data.
- The rules are based on an **augmented causal model** that incorporates interventions explicitly within the graph structure.
- Causal intervention uses a **decision variable, $\hat{Z}$, to intervene** at Z, where $\hat{Z} \in \{\epsilon\} \bigcup Val(Z)$
  1. If $\hat{Z} = \epsilon$, then Z behaves normally.
  2. If $\hat{Z} = z$, then it deterministically sets $Z = z$.

Rule 1: Graph-Dependent Intervention Rule

$$P(\mathbf{Y} \mid do(\mathbf{Z} := \mathbf{z}), \mathbf{X} = \mathbf{x}, \mathbf{W} = \mathbf{w}) = P(\mathbf{Y} \mid do(\mathbf{Z} := \mathbf{z}), \mathbf{X} = \mathbf{x})$$

if **Y** is d-sep from **W** given **Z**, **X**, in $\mathcal{G}^{\dagger}_{\bar{Z}}$.



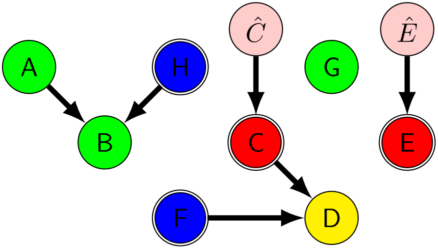
Intuition: Intervention query probabilities depend on the graph's independence and d-separation.

Rule 2: Information Conservation Rule

$$P(\mathbf{Y} \mid do(\mathbf{Z} := \mathbf{z}), \boxed{do(\mathbf{X} = \mathbf{x})}, \mathbf{W} = \mathbf{w})$$

$$= P(\mathbf{Y} \mid do(\mathbf{Z} := \mathbf{z}), \boxed{\mathbf{X} = \mathbf{x}}, \mathbf{W} = \mathbf{w})$$

if **Y** is d-sep from $\hat{\mathbf{X}}$ given **X**, **Z**, **W** in $\mathcal{G}^{\dagger}_{\bar{Z}}$.

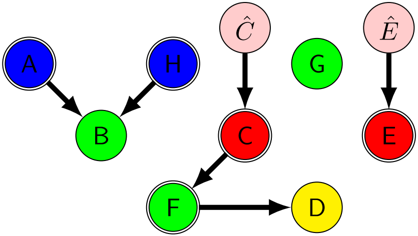


Intuition: An intervention at variable X provides no additional information about variable Y beyond the specific values of X.

Rule 3: Irrelevant Intervention Rule

$$P(\mathbf{Y} \mid do(\mathbf{Z} := \mathbf{z}), \boxed{do(\mathbf{X} = \mathbf{x})}, \mathbf{W} = \mathbf{w})$$
$$= P(\mathbf{Y} \mid do(\mathbf{Z} := \mathbf{z}), \mathbf{W} = \mathbf{w})$$

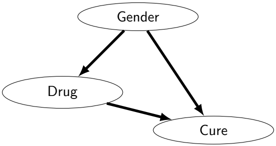
if $\mathbf{Y}$ is d-sep from $\hat{\mathbf{X}}$ given $\mathbf{Z}$, $\mathbf{W}$ in $\mathcal{G}^{\dagger}_{\bar{Z}}$.



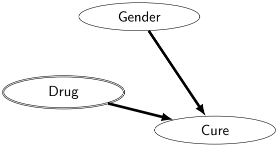
Intuition: Since the values of X are not requisite, we can disregard both the knowledge of the intervention and the knowledge of the actual values taken.

Iterated Query Simplification

- We can use this to answer Simpson's paradox.



$P(c^1 \mid d^1)$



$P(c^1 \mid do(d^1))$

Applying **Rule 2: Information Conservation Rule**

$$P(c^1 \mid do(d^1)) = \sum_g P(\overset{\mathsf{Y}}{c^1} \mid \overset{\mathsf{X}}{do(d^1)}, \overset{\mathsf{W}}{g})P(g)$$

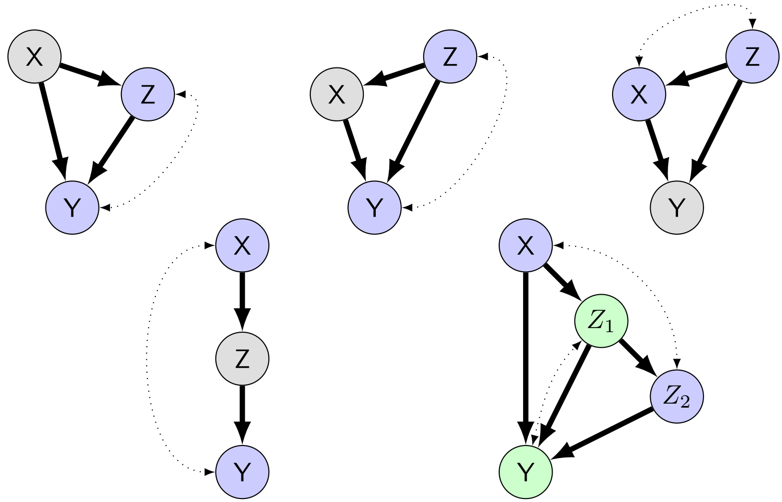
$$P(c^1 \mid do(d^1)) = \sum_g P(\overset{\mathsf{Y}}{c^1} \mid \overset{\mathsf{X}}{d^1}, \overset{\mathsf{W}}{g})P(g)$$

Solving Simpson's paradox

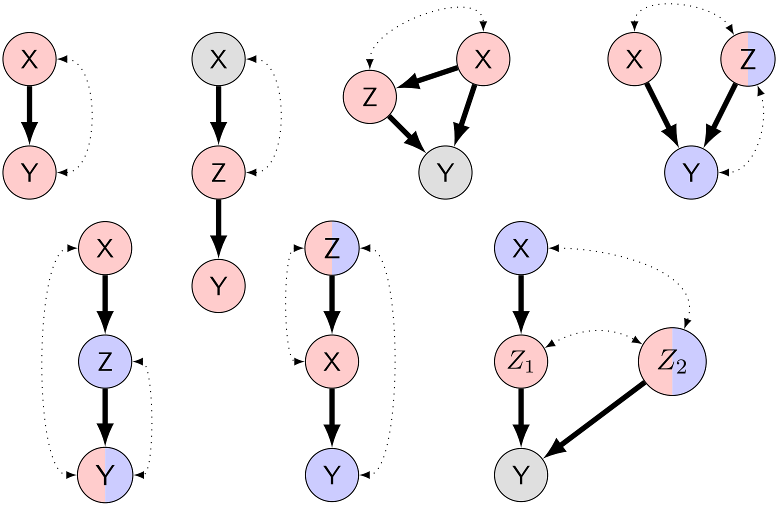
$$
\begin{aligned}
P(c^1 \mid do(d^1)) &= \sum_g P(c^1 \mid d^1, g)P(g) \\
&= P(c^1 \mid d^1, m)P(m) + P(c^1 \mid d^1, f)P(f) \\
&= (0 \times 0.5) + (0.75 \times 0.5) \\
&= 0.375 \\
P(c^1 \mid do(d^0)) &= \sum_g P(c^1 \mid d^0, g)P(g) \\
&= P(c^1 \mid d^0, m)P(m) + P(c^1 \mid d^0, f)P(f) \\
&= (0.25 \times 0.5) + (1 \times 0.5) \\
&= 0.625
\end{aligned}
$$

- And therefore, we should **not prescribe the drug.**

Models where $P(Y \mid do(X))$ is **identifiable**



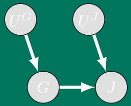
Models where $P(Y \mid do(X))$ is NOT identifiable



Mechanism and Response Variable

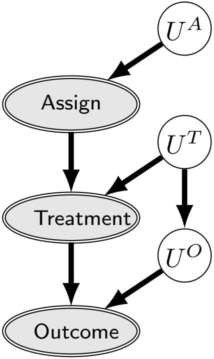
- The causal graph structure defines a **causal progression**.
- We have to explicitly reason about the **mechanism governing variable behaviour** and interaction with latent variables.
- Knowing each variable's mechanism improves reliable inferences and preserves when reasoning about **counterfactual events**.
- Answering counterfactual queries involved knowledge about:
  1. **Endogenous variables**: Included in the causal model.
  2. **Exogenous variables**: Models stochasticity in the world.
- **Functional causal models** provide more detailed specification of causal mechanisms than standard models.

Response Variables



- Suppose we given the causal model for Sipho: $G \rightarrow J$
- We are interested in calculating Sipho's job prospects.
- If **U** is exogenous variables and G, J and endogenous, then J is a deterministic function of $G, \mathbf{U}$. I.e $f_J(\mathbf{u}, g)$.
- This leads to the following cases of recruiting Sipho:
  1. $\mu^J_{1\mapsto 1, 0\mapsto 1}$: $f_J(\mathbf{u}, g^1) = j^1$ and $f_J(\mathbf{u}, g^0) = j^1$
  2. $\mu^J_{1\mapsto 1, 0\mapsto 0}$: $f_J(\mathbf{u}, g^1) = j^1$ and $f_J(\mathbf{u}, g^0) = j^0$
  3. $\mu^J_{1\mapsto 0, 0\mapsto 1}$: $f_J(\mathbf{u}, g^1) = j^0$ and $f_J(\mathbf{u}, g^0) = j^1$
  4. $\mu^J_{1\mapsto 0, 0\mapsto 0}$: $f_J(\mathbf{u}, g^1) = j^0$ and $f_J(\mathbf{u}, g^0) = j^0$
- The **response variable**, U^J, encodes all of these functions. Since G has no parents $U^G = \{g^0, g^1\}$.

Functional Causal Models for Clinical Trails



$P(U^A)$

Medication	Placebo
0.5	0.5

$P(U^T)$: Patient's treatment compliance model.

Always Taker	Complier	Defier	Never taker
0.05	0.7	0.05	0.2

$P(U^O \mid U^T)$: Encodes patient's treatment response.

U^T	Always well	Helped	Hurt	Never well
Always Taker	0.2	0.5	0.1	0.2
Complier	0.5	0.4	0.05	0.05
Defier	0.3	0.4	0.1	0.2
Never taker	0.6	0.3	0.05	0.05

Counterfactual Queries

- Counterfactual queries involve real-world scenarios and **modified counterfactual scenarios** via causal intervention.
- So we reason in parallel about two different worlds: the **real one**, and the **counterfactual one**.
- We assume response variable values are randomly selected in the real world and **preserved in the counterfactual world**.
- Would the patient been healthy if they complied with prescribed treatment instead of not getting well due to non-compliance?

$$P(O' = o^1 \mid A = a^1, T = t^0, O = o^0, do(T' := t^1))$$

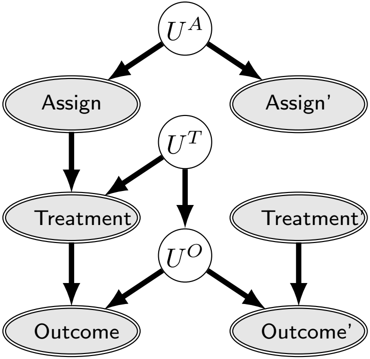
Twinned Counterfactual Network

$$P(O' = o^1 \mid A = a^1, T = t^0, O = o^0, do(T' := t^1))$$

Treatment Group

Defier/ Never taker

Never well



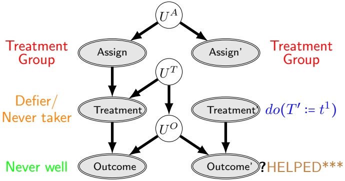
Treatment Group

$do(T' := t^1)$

?

Causal Inference

U^T	U^O	Prob
Always taker	Always well	0.01
Always taker	Helped	0.025
Always taker	Hurt	0.005
Always taker	Never well	0.01
Complier	Always well	0.35
Complier	Helped	0.28
Complier	Hurt	0.035
Complier	Never well	0.035
Defier	Always well	0.015
Defier	Helped	0.02
Defier	Hurt	0.005
Defier	Never well	0.02
Never taker	Always well	0.12
Never taker	Helped	0.06
Never taker	Hurt	0.01
Never taker	Never well	0.01



U^T	U^O	Prob
Defier	Helped	0.31
Defier	Never well	0.15
Never taker	Helped	0.46
Never taker	Never well	0.08

- Answer = 0.77.
- **The treatment would have helped!**

Learning Causal Models

- How can we learn a causal model from data?
- There are many learning problems here:
  1. Learning a Causal Model Parameters?
  2. Learning Response variables for Functional Causal Model?
  3. Learning Causal Structure?
  4. Learning from intervention data or observational data?
  5. Are there confounding variables?
- There are two learning assumptions:
  1. **Causal Markov Assumption**: in P^*, each variable is conditionally independent of its non-effects given its direct causes. (Does not hold in quantum mechanics)
  2. **Faithfulness Assumption**: Only conditional assumption independencies in P^* are those that arise from d-separation in the corresponding causal graph $\mathcal{G}^*$.
- $\mathcal{G}^*$ is therefore a **perfect map** for P^*.

Identifying a Model

- As stated previously, it is impossible to identify $\mathcal{G}^*$ from P^*, even with infinite data.
- However, we can obtain **the equivalence class** for $\mathcal{G}^*$.
- Can use constraint-based or score-based structure learning
- Bayesian model averaging is best as it provides a distribution over structures, encoding uncertainty about **edge orientation and presence/absence**.
- Hybrid approaches are also available.
- E.g. Constrained-based methods are a prior to **avoid local maxima**, with Bayesian model averaging **preventing irreversible decisions** on independencies.

Learning from Interventional Data

- It is natural to use **interventional data** to learn causal models.

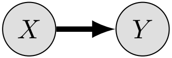
$$P(\xi \mid do(\mathbf{Z} := \mathbf{z}), \mathcal{C}) = \prod_{X_i \notin \mathbf{Z}} P(x_i \mid \mathbf{u}_i)$$

$$M[x_i; \mathbf{u}_i] = \sum_{m : X_i \notin \mathbf{Z}[m]} \mathbb{1}\{X_i[m] = x_i, Pa_{X_i}[m] = \mathbf{u}_i\}$$

- **Intuition**: counts the number of occurrences of this event, in data instances where there is no intervention at X_i.

$$L(\mathcal{C} : \mathcal{D}) = \prod_{i=1}^{n} \prod_{x_i \in Val(X_i), \mathbf{u}_i \in Val(Pa_{X_i})} \theta_{x_i \mid \mathbf{u}_i}^{M[x_i; \mathbf{u}_i]}$$

Example of Learning from Interventional Data



Intervention	x^1, y^1	x^1, y^0	x^0, y^1	x^0, y^0
$\varnothing$	4	1	1	4
$do(X := x^1)$	2	0	0	0
$do(Y := y^1)$	1	0	1	0

$$M[x^1] = M[x^1y^1 \mid \varnothing] + M[x^1y^0 \mid \varnothing] + M[x^1y^1 \mid do(y^1)] = 4 + 1 + 1$$

$$M[x^0] = M[x^0y^1 \mid \varnothing] + M[x^0y^0 \mid \varnothing] + M[x^0y^1 \mid do(y^1)] = 1 + 4 + 1$$

$$M[y^1, x^1] = M[x^1y^1 \mid \varnothing] + M[x^1y^1 \mid do(x^1)] = 4 + 2$$

$$M[y^0, x^1] = M[x^1y^0 \mid \varnothing] + M[x^1y^0 \mid do(x^1)] = 1 + 0$$

$$M[y^1, x^0] = M[x^0y^1 \mid \varnothing] + M[x^0y^1 \mid do(x^0)] = 1 + 0$$

$$M[y^0, x^0] = M[x^0y^0 \mid \varnothing] + M[x^0y^0 \mid do(x^0)] = 4 + 0$$

Therefore $\theta_{x^1} = 0.5$, $\theta_{y^1|x^1} = 0.86$, and $\theta_{y^1|x^0} = 0.2$

Summary

- Causal networks model causal relationships, not just statistical ones
- Ideal intervention and counterfactual queries
- Correlation doesn't imply causation
- Simplify models to make them identifiable
- Describe causality more fine grained using Functional Causal Models
- Learning Causal Models is challenging due to confounding, unobserved variables, and non-linear relationships.