CSC2541: Introduction to Causality

Lecture 3 - Identification

Instructor: Rahul G. Krishnan

TA & slides: Vahid Balazadeh-Meresht

September 26, 2022

Recap - Lecture 2

- Bayesian networks - Compact representations of joint probability distributions.
- Conditional Independencies - Blocked and Unblocked paths characterize the flow of association.
- D-separation and (global/local) Markov properties - Characterize conditional independence in a graph.
- Observational equivalence - We cannot distinguish graphs that have the same skeleton and same v-structures from data.

Recap - Lecture 2

- do-operator - Operator that corresponds to an intervention on a random variable.
- Independent mechanisms (or Modularity) - Intervention on a node only changes the mechanism associated with that node.
- Causal Bayesian Networks - $\mathcal{G}$ is causal BN if the interventional distribution is Markov compatible with it and it satisfies modularity.
- Analyzing (directed) paths in a Causal Bayesian Network lets us assess the flow of causation.
- Structural Causal Models - Functional representation of causal process that generates the data (more flexibility than Bayesian network).
- Good predictors need not be causal!

Counterfactuals - Imagination

Suppose we know (e.g., from randomized trials) that a treatment T has no causal effect on mortality Y. The corresponding causal Bayesian network will be

saying $P(Y|do(T=1)) = P(Y|do(T=0)) = P(Y)$.

Counterfactuals - Imagination

Suppose we know (e.g., from randomized trials) that a treatment T has no causal effect on mortality Y. The corresponding causal Bayesian network will be

saying $P(Y|do(T = 1)) = P(Y|do(T = 0)) = P(Y)$.

1. Should we prescribe the treatment for a new patient?

Counterfactuals - Imagination

Suppose we know (e.g., from randomized trials) that a treatment T has no causal effect on mortality Y. The corresponding causal Bayesian network will be

saying $P(Y|do(T=1)) = P(Y|do(T=0)) = P(Y)$.

1. Should we prescribe the treatment for a new patient? It has no causal effect!

Counterfactuals - Imagination

Suppose we know (e.g., from randomized trials) that a treatment T has no causal effect on mortality Y. The corresponding causal Bayesian network will be

saying $P(Y|do(T=1)) = P(Y|do(T=0)) = P(Y)$.

1. Should we prescribe the treatment for a new patient? It has no causal effect!

2. Suppose we did prescribe the treatment ($T=1$) for a patient and he died. *What would have happened had he not been treated?*

Counterfactuals - Imagination

Suppose we know (e.g., from randomized trials) that a treatment T has no causal effect on mortality Y. The corresponding causal Bayesian network will be

saying $P(Y|do(T=1)) = P(Y|do(T=0)) = P(Y)$.

1. Should we prescribe the treatment for a new patient? It has no causal effect!

2. Suppose we did prescribe the treatment ($T=1$) for a patient and he died. *What would have happened had he not been treated?*

   This is a counterfactual question. We can never observe/test counterfactuals even with RCTs

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1 : \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

$$\mathcal{M}_2 : \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1 - T) \cdot (1 - U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1 : \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned} \qquad \mathcal{M}_2 : \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1-T) \cdot (1-U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

1. What is the observational distribution of $\mathcal{M}_1$ and $\mathcal{M}_2$?

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1: \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1-T) \cdot (1-U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

1. What is the observational distribution of $\mathcal{M}_1$ and $\mathcal{M}_2$?

$$P^{\mathcal{M}_1}(T=t, Y=y) = P(U_1=t)P(U_2=y) = 0.25$$

$$P^{\mathcal{M}_2}(T=t, Y=y) = P(U_1=t)[\mathbb{I}_{t=y} \cdot P(U_2=1) + \mathbb{I}_{t\neq y} \cdot P(U_2=0)] = 0.25$$

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1: \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1-T) \cdot (1-U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

1. What is the observational distribution of $\mathcal{M}_1$ and $\mathcal{M}_2$?

   $P^{\mathcal{M}_1}(T = t, Y = y) = P(U_1 = t)P(U_2 = y) = 0.25$

   $P^{\mathcal{M}_2}(T = t, Y = y) = P(U_1 = t)[\mathbb{I}_{t=y} \cdot P(U_2 = 1) + \mathbb{I}_{t \neq y} \cdot P(U_2 = 0)] = 0.25$

2. What is the interventional distribution $P(Y|do(T))$ for $\mathcal{M}_1$ and $\mathcal{M}_2$?

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1: \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1-T) \cdot (1-U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

1. What is the observational distribution of $\mathcal{M}_1$ and $\mathcal{M}_2$?

   $P^{\mathcal{M}_1}(T=t, Y=y) = P(U_1=t)P(U_2=y) = 0.25$

   $P^{\mathcal{M}_2}(T=t, Y=y) = P(U_1=t)[\mathbb{I}_{t=y} \cdot P(U_2=1) + \mathbb{I}_{t\neq y} \cdot P(U_2=0)] = 0.25$

2. What is the interventional distribution $P(Y|do(T))$ for $\mathcal{M}_1$ and $\mathcal{M}_2$?

   $P^{\mathcal{M}_1}(Y=y|do(T=t)) = P(U_2=y) = 0.5$

   $P^{\mathcal{M}_2}(Y=y|do(T=t)) = \mathbb{I}_{t=1} \cdot P(U_2=y) + \mathbb{I}_{t=0} \cdot P(U_2=1-y) = 0.5$

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1 : \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

$$\mathcal{M}_2 : \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1 - T) \cdot (1 - U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1 : \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned} \qquad \mathcal{M}_2 : \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1-T) \cdot (1-U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

What would have happened had the sick patient not been treated?

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1: \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1-T) \cdot (1-U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

What would have happened had the sick patient not been treated? We can infer U_1 and U_2 based on the observation $T = 1$ and $Y = 1$:

$$\mathcal{M}_1: \begin{aligned} T = 1 &\implies U_1 = 1 \\ Y = 1 &\implies U_2 = 1 \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T = 1 &\implies U_1 = 1 \\ Y = 1, T = 1 &\implies U_2 = 1 \end{aligned}$$

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1: \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1-T) \cdot (1-U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

What would have happened had the sick patient not been treated? We can infer U_1 and U_2 based on the observation $T = 1$ and $Y = 1$:

$$\mathcal{M}_1: \begin{aligned} T = 1 &\implies U_1 = 1 \\ Y = 1 &\implies U_2 = 1 \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T = 1 &\implies U_1 = 1 \\ Y = 1, T = 1 &\implies U_2 = 1 \end{aligned}$$

We can answer the counterfactual question after inferring U_1 and U_2

$$P^{\mathcal{M}_{1|T=1,Y=1}}(Y = 1 | do(T = 0)) = P(U_2 = 1) = 1$$

$$P^{\mathcal{M}_{2|T=1,Y=1}}(Y = 1 | do(T = 0)) = P(1 - U_2 = 1) = P(U_2 = 0) = 0$$

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1: \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1-T) \cdot (1-U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

What would have happened had the sick patient not been treated? We can infer U_1 and U_2 based on the observation $T = 1$ and $Y = 1$:

$$\mathcal{M}_1: \begin{aligned} T = 1 &\implies U_1 = 1 \\ Y = 1 &\implies U_2 = 1 \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T = 1 &\implies U_1 = 1 \\ Y = 1, T = 1 &\implies U_2 = 1 \end{aligned}$$

We can answer the counterfactual question after inferring U_1 and U_2

$$P^{\mathcal{M}_{1|T=1,Y=1}}(Y = 1 | do(T = 0)) = P(U_2 = 1) = 1$$

$$P^{\mathcal{M}_{2|T=1,Y=1}}(Y = 1 | do(T = 0)) = P(1 - U_2 = 1) = P(U_2 = 0) = 0$$

Are counterfactuals the same as interventions?

Consider the following two SCMs that generate T and Y

$$\mathcal{M}_1: \begin{aligned} T &= U_1 \\ Y &= U_2 \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T &= U_1 \\ Y &= T \cdot U_2 + (1-T) \cdot (1-U_2) \\ U_1, U_2 &\sim \texttt{Ber}(0.5) \end{aligned}$$

What would have happened had the sick patient not been treated? We can infer U_1 and U_2 based on the observation $T = 1$ and $Y = 1$:

$$\mathcal{M}_1: \begin{aligned} T = 1 &\implies U_1 = 1 \\ Y = 1 &\implies U_2 = 1 \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} T = 1 &\implies U_1 = 1 \\ Y = 1, T = 1 &\implies U_2 = 1 \end{aligned}$$

We can answer the counterfactual question after inferring U_1 and U_2

$$P^{\mathcal{M}_{1|T=1,Y=1}}(Y = 1 | do(T = 0)) = P(U_2 = 1) = 1$$

$$P^{\mathcal{M}_{2|T=1,Y=1}}(Y = 1 | do(T = 0)) = P(1 - U_2 = 1) = P(U_2 = 0) = 0$$

For interventional questions, we can run RCTs and estimate quantities like ATE. But, for counterfactual questions, we can never go back in time and change what we did.

Pearl's three layer causal hierarchy

Association ($P(y|x)$) Seeing. How would seeing X change the belief in Y?

Example What does a symptom tell us about a disease?

Adapted from `https://web.cs.ucla.edu/~kaoru/3-layer-causal-hierarchy.pdf`

Pearl's three layer causal hierarchy

Association ($P(y|x)$) Seeing. How would seeing X change the belief in Y?

Example What does a symptom tell us about a disease?

Intervention ($P(y|do(x))$) Doing. What if I do X?

Example If I take aspirin, will my headache be cured?

Adapted from `https://web.cs.ucla.edu/~kaoru/3-layer-causal-hierarchy.pdf`

Pearl's three layer causal hierarchy

Association ($P(y|x)$) Seeing. How would seeing X change the belief in Y?

Example What does a symptom tell us about a disease?

Intervention ($P(y|do(x))$) Doing. What if I do X?

Example If I take aspirin, will my headache be cured?

Counterfactuals ($P^{\mathcal{M}_{|X=x',Y=y'}}(y|do(x))$) Imagining. Was it X that caused Y? What if I had acted differently?

Example What if I had not be smoking the past 2 years?

Adapted from `https://web.cs.ucla.edu/~kaoru/3-layer-causal-hierarchy.pdf`

Pearl's three layer causal hierarchy

Association ($P(y|x)$) Seeing. How would seeing X change the belief in Y?

Example What does a symptom tell us about a disease?

Intervention ($P(y|do(x))$) Doing. What if I do X?

Example If I take aspirin, will my headache be cured?

Counterfactuals ($P^{\mathcal{M}_{|X=x',Y=y'}}(y|do(x))$) Imagining. Was it X that caused Y? What if I had acted differently?

Example What if I had not be smoking the past 2 years?

The hierarchy is directional: Association $<$ Intervention $<$ Counterfactuals. Using counterfactuals (intervention), we can answer questions about intervention (association).

Adapted from `https://web.cs.ucla.edu/~kaoru/3-layer-causal-hierarchy.pdf`

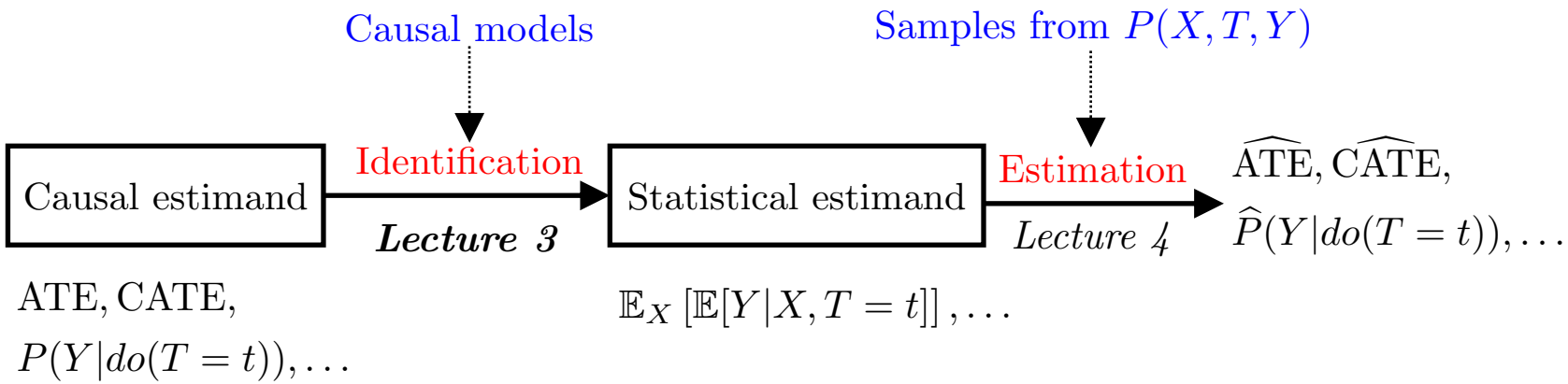
Questions?

Question

Any questions on counterfactuals?

We can only estimate "statistical" quantities

- In lectures 1 & 2, we studied the (1) potential outcomes framework, (2) causal Bayesian networks, and (3) SCMs to model causal quantities
- We made assumptions like (conditional) ignorability, positivity, and modularity and used G-formula or truncated factorization to estimate the causal quantities
- We can only observe data from the observational distribution $P(X, T, Y)$ and not interventional distribution $P(X, Y|do(T))$ [2]



[2] Except in RCTs, where we do observe data from the interventional distribution

A non-identifiable example

Consider the following two SCM, where we only observe T and Y:

$$\mathcal{M}_1: \quad \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= X + \mathcal{N}(0,1) \\ Y &:= X + T + \mathcal{N}(0,1) \end{aligned}$$

$$\mathcal{M}_2: \quad \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= 1.4X + 0.2\ \mathcal{N}(0,1) \\ Y &:= 5X - 2T + \mathcal{N}(0,1) \end{aligned}$$

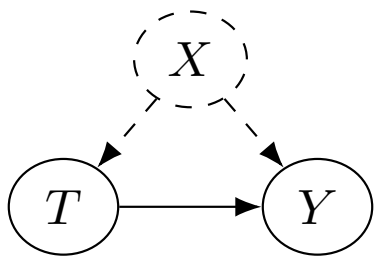
A non-identifiable example

Consider the following two SCM, where we only observe T and Y:

$$\mathcal{M}_1: \quad \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= X + \mathcal{N}(0,1) \\ Y &:= X + T + \mathcal{N}(0,1) \end{aligned}$$

$$\mathcal{M}_2: \quad \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= 1.4X + 0.2\ \mathcal{N}(0,1) \\ Y &:= 5X - 2T + \mathcal{N}(0,1) \end{aligned}$$

Nature only gives us the (observed) generated data $P(T,Y)$, and not the data generating rules. We may also know the causal graph.



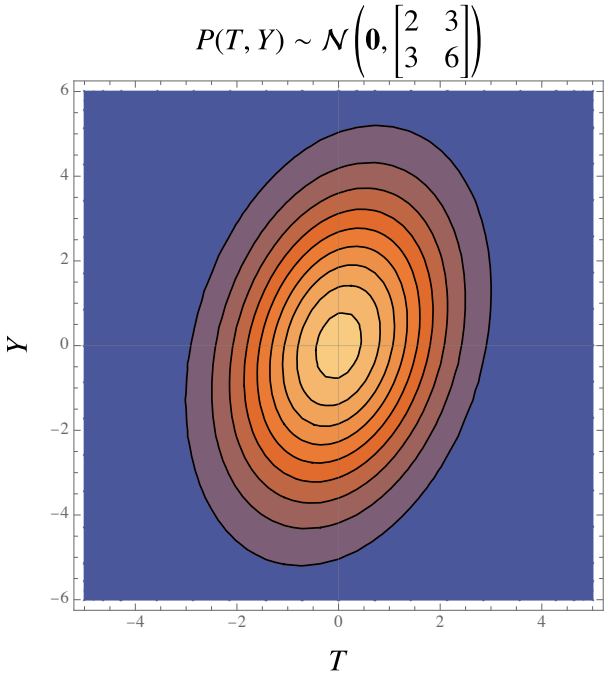
What is the observed data distributions for SCMs $\mathcal{M}_1, \mathcal{M}_2$?

A non-identifiable example

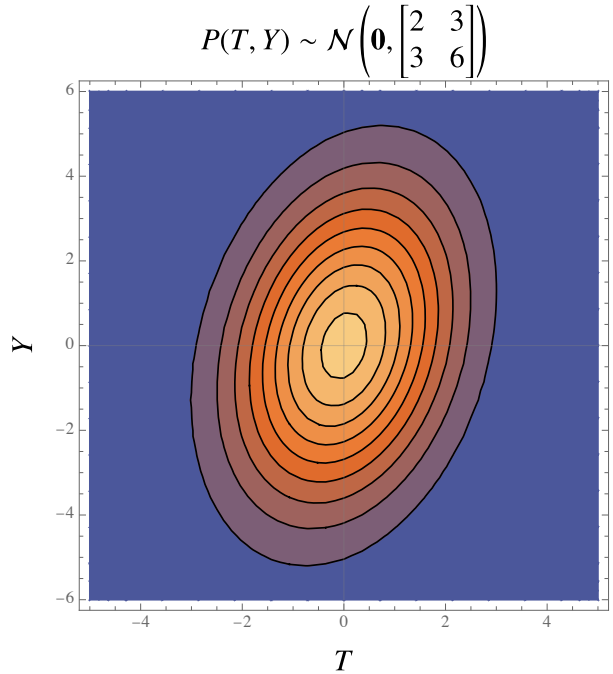
Consider the following two SCM, where we only observe T and Y:

$$\mathcal{M}_1: \quad \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= X + \mathcal{N}(0,1) \\ Y &:= X + T + \mathcal{N}(0,1) \end{aligned}$$

$$\mathcal{M}_2: \quad \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= 1.4X + 0.2\,\mathcal{N}(0,1) \\ Y &:= 5X - 2T + \mathcal{N}(0,1) \end{aligned}$$



Show it as HW



What about interventional distributions $P^{\mathcal{M}_1}(Y|do(T=1))$, $P^{\mathcal{M}_2}(Y|do(T=1))$?

A non-identifiable example - calculating the interventions

What about $P^{\mathcal{M}_1}(Y|do(T=1))$, $P^{\mathcal{M}_2}(Y|do(T=1))$?

$$\mathcal{M}_1: \quad \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= X + \mathcal{N}(0,1) \\ Y &:= X + T + \mathcal{N}(0,1) \end{aligned}$$

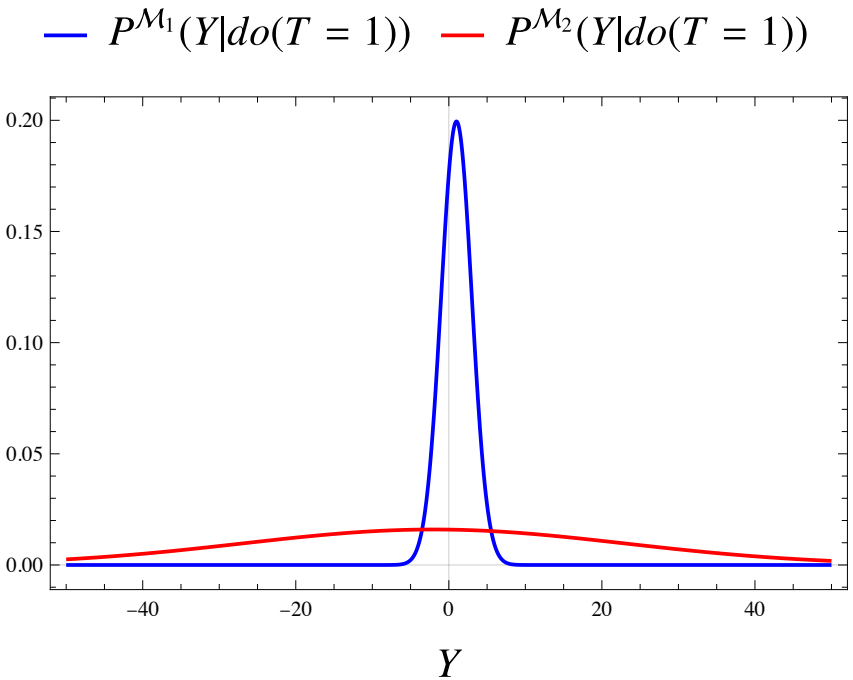
$$\mathcal{M}_2: \quad \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= 1.4X + 0.2\,\mathcal{N}(0,1) \\ Y &:= 5X - 2T + \mathcal{N}(0,1) \end{aligned}$$

A non-identifiable example - calculating the interventions

What about $P^{\mathcal{M}_1}(Y|do(T=1))$, $P^{\mathcal{M}_2}(Y|do(T=1))$?

$$\mathcal{M}_1: \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= 1 \\ Y &:= X + T + \mathcal{N}(0,1) \end{aligned} \qquad \mathcal{M}_2: \begin{aligned} X &:= \mathcal{N}(0,1) \\ T &:= 1 \\ Y &:= 5X - 2T + \mathcal{N}(0,1) \end{aligned}$$

$$P^{\mathcal{M}_1}(Y|do(T=1)) \sim \mathcal{N}(1,2) \qquad P^{\mathcal{M}_2}(Y|do(T=1)) \sim \mathcal{N}(-2,26)$$



Identifiability - Definition

- In the previous example, both the observed distributions and causal graphs were exactly the same for $\mathcal{M}_1$ and $\mathcal{M}_2$ but $P^{\mathcal{M}_1}(Y|do(T=1)) \neq P^{\mathcal{M}_2}(Y|do(T=1))$
- We cannot distinguish between $\mathcal{M}_1$ and $\mathcal{M}_2$. Hence, $P(Y|do(T=1))$ is **not identifiable**

Identifiability - Definition

- In the previous example, both the observed distributions and causal graphs were exactly the same for $\mathcal{M}_1$ and $\mathcal{M}_2$ but $P^{\mathcal{M}_1}(Y|do(T=1)) \neq P^{\mathcal{M}_2}(Y|do(T=1))$
- We cannot distinguish between $\mathcal{M}_1$ and $\mathcal{M}_2$. Hence, $P(Y|do(T=1))$ is **not identifiable**

Identifiability

Let $Q(\mathcal{M})$ be any computable quantity of a SCM $\mathcal{M}$. We say that Q is identifiable in a class $\mathfrak{M}$ of models if, for any pairs of SCMs $\mathcal{M}_1, \mathcal{M}_2 \in \mathfrak{M}$ that have the same observed distribution $P^{\mathcal{M}_1}(\boldsymbol{O}) = P^{\mathcal{M}_2}(\boldsymbol{O})$ and causal graph $\mathcal{G}(\mathcal{M}_1) = \mathcal{G}(\mathcal{M}_2)$, we have $Q(\mathcal{M}_1) = Q(\mathcal{M}_2)$.[a]

[a] We implicitly assume the positivity holds, i.e., $P^{\mathcal{M}}(\boldsymbol{O}) > 0$ for all models $\mathcal{M} \in \mathfrak{M}$

Questions?

Question

Any questions on identifiability?

Identification of interventional distributions

- Here, we will focus on quantities like $P(Y|do(T))$, where T and Y are (sets of disjoint) variables in $\boldsymbol{O}$
- We assume that we can observe a subset of all related variables $\boldsymbol{O} \subseteq \boldsymbol{V}$ and have access to the joint distribution $P(\boldsymbol{O})$[1]
- We'll also assume the causal graph $\mathcal{G}$ is given (from the expert knowledge or structure learning algorithms)

[1] In other words, we assume having infinite samples. Lecture 4 will discuss the finite-sample case.

Identification of interventional distributions

- Here, we will focus on quantities like $P(Y|do(T))$, where T and Y are (sets of disjoint) variables in $\boldsymbol{O}$
- We assume that we can observe a subset of all related variables $\boldsymbol{O} \subseteq \boldsymbol{V}$ and have access to the joint distribution $P(\boldsymbol{O})$[1]
- We'll also assume the causal graph $\mathcal{G}$ is given (from the expert knowledge or structure learning algorithms)

Question

Can we identify $P(Y|do(T))$ if we observe **all** the related variables, i.e., $\boldsymbol{O} = \boldsymbol{V}$?

[1] In other words, we assume having infinite samples. Lecture 4 will discuss the finite-sample case.

Identification with no unobserved variables

Question

Can we identify $P(Y = y|do(T = t))$ if we observe **all** the related variables, i.e., $\boldsymbol{O} = \boldsymbol{V}$?

Identification with no unobserved variables

Question

Can we identify $P(Y = y|do(T = t))$ if we observe **all** the related variables, i.e., $\boldsymbol{O} = \boldsymbol{V}$?

Yes! we can use truncated factorization
$P(y|do(T = t))$

Identification with no unobserved variables

Question

Can we identify $P(Y = y|do(T = t))$ if we observe **all** the related variables, i.e., $\boldsymbol{O} = \boldsymbol{V}$?

Yes! we can use truncated factorization

$$P(y|do(T = t))$$

$$= \sum_{v_1, v_2, \ldots, v_k \in \boldsymbol{V} \setminus (Y \cup T)} P(y, v_1, v_2, \ldots, x_k|do(T = t)) \qquad \text{(marginalization)}$$

Identification with no unobserved variables

Question

Can we identify $P(Y = y|do(T = t))$ if we observe **all** the related variables, i.e., $\boldsymbol{O} = \boldsymbol{V}$?

Yes! we can use truncated factorization

$$
\begin{aligned}
&P(y|do(T=t)) \\
&= \sum_{v_1, v_2, \ldots, v_k \in \boldsymbol{V} \setminus (Y \cup T)} P(y, v_1, v_2, \ldots, x_k | do(T=t)) && \text{(marginalization)} \\
&= \sum_{v_1, v_2, \ldots, v_k \in \boldsymbol{V} \setminus (Y \cup T)} P(y|pa_Y) \prod_{v_i \notin T} P(v_i|pa_i) && \text{(truncated factorization)}
\end{aligned}
$$

Where PA_Y (PA_i) is the set of parent nodes of Y (V_i) in causal graph $\mathcal{G}$. Note that the RHS only depends on the observational distribution P.

Identification with no unobserved variables

Question

Can we identify $P(Y = y|do(T = t))$ if we observe **all** the related variables, i.e., $\boldsymbol{O} = \boldsymbol{V}$?

Yes! we can use truncated factorization

$$
\begin{aligned}
&P(y|do(T=t)) \\
&= \sum_{v_1, v_2, \ldots, v_k \in \boldsymbol{V} \setminus (Y \cup T)} P(y, v_1, v_2, \ldots, x_k | do(T=t)) && \text{(marginalization)} \\
&= \sum_{v_1, v_2, \ldots, v_k \in \boldsymbol{V} \setminus (Y \cup T)} P(y|pa_Y) \prod_{v_i \notin T} P(v_i|pa_i) && \text{(truncated factorization)}
\end{aligned}
$$

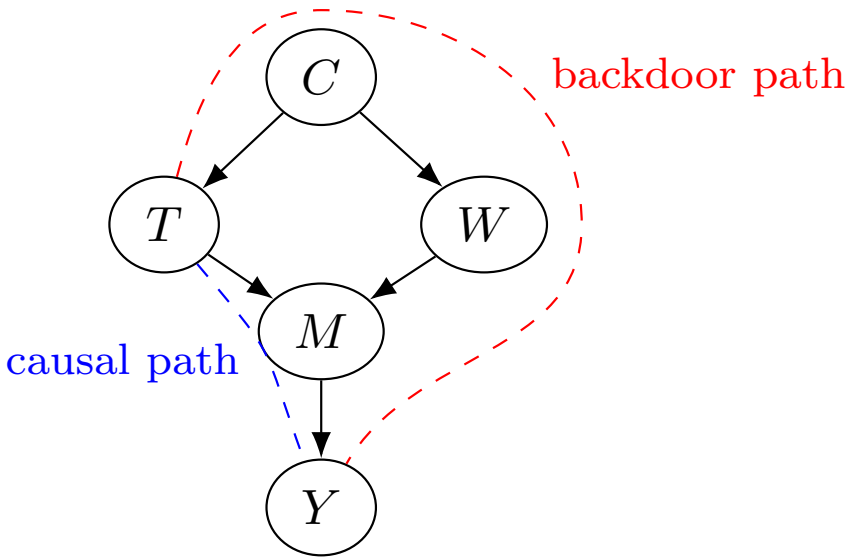
Where PA_Y (PA_i) is the set of parent nodes of Y (V_i) in causal graph $\mathcal{G}$. Note that the RHS only depends on the observational distribution P.

To calculate $P(y|do(T = t))$, we need to sum (integrate) over all variables V_i, which can be intractable. Can we simplify the formula?

Backdoor adjustment - Blocking backdoor paths

Unblocked paths show the information flow (dependencies)

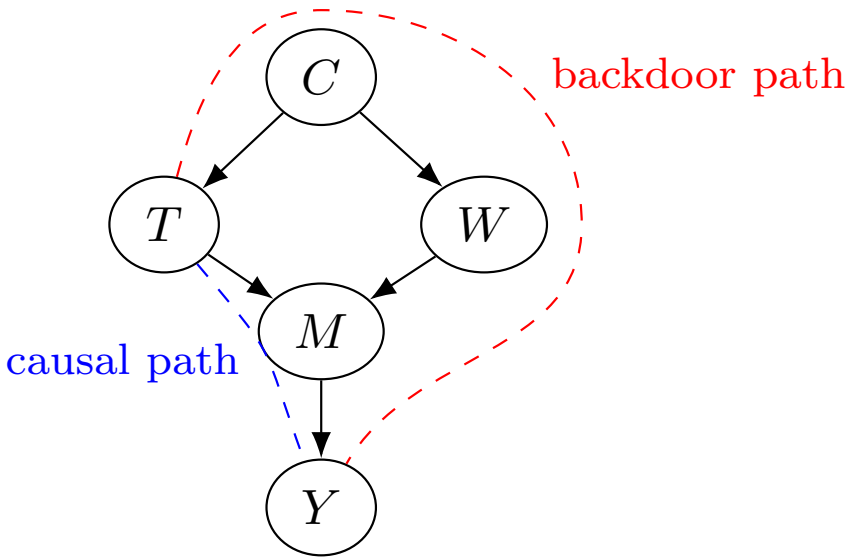
$$P(Y|T)$$



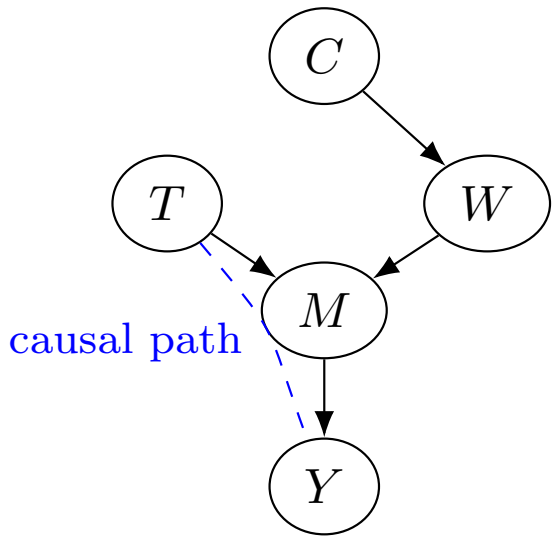
Backdoor adjustment - Blocking backdoor paths

Unblocked paths show the information flow (dependencies)

$P(Y|T)$



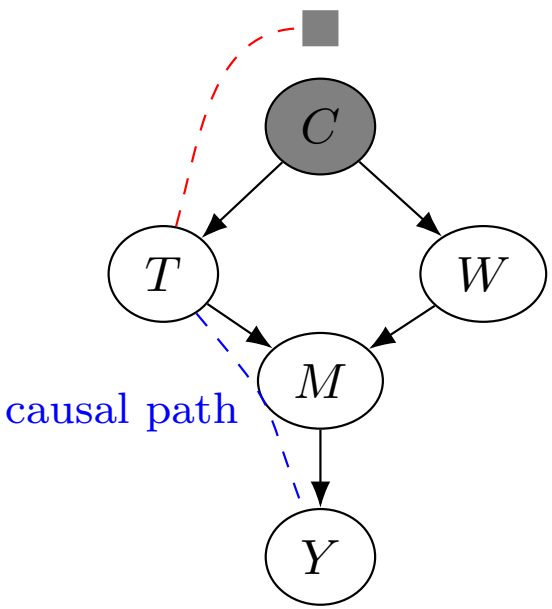
$P(Y|do(T))$



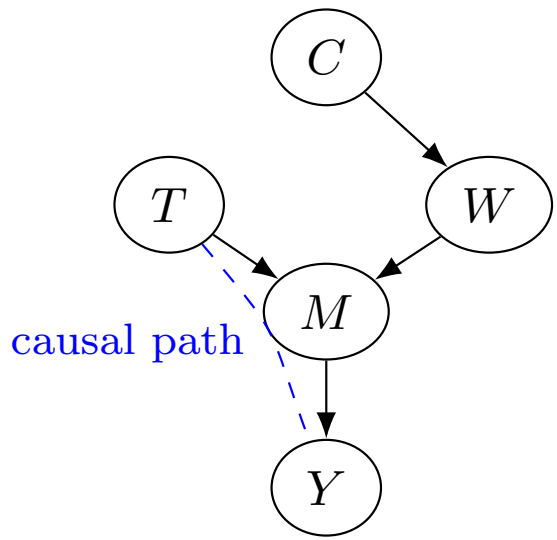
Backdoor adjustment - Blocking backdoor paths

Unblocked paths show the information flow (dependencies)

$P(Y|T, C)$



$P(Y|do(T))$

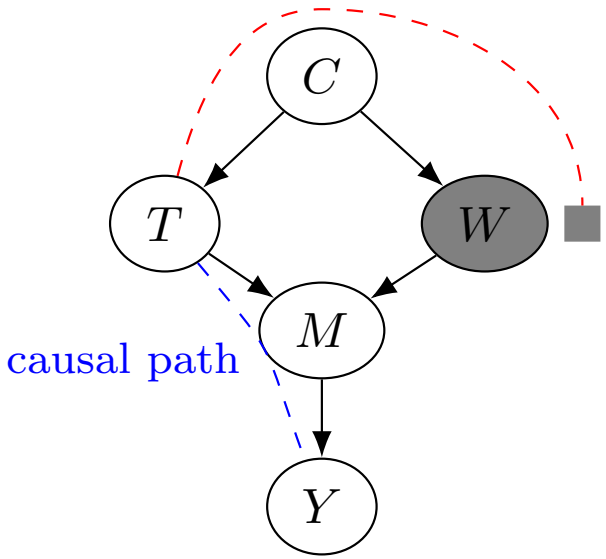


Conditioning on variables within backdoor paths blocks the non-causal associations

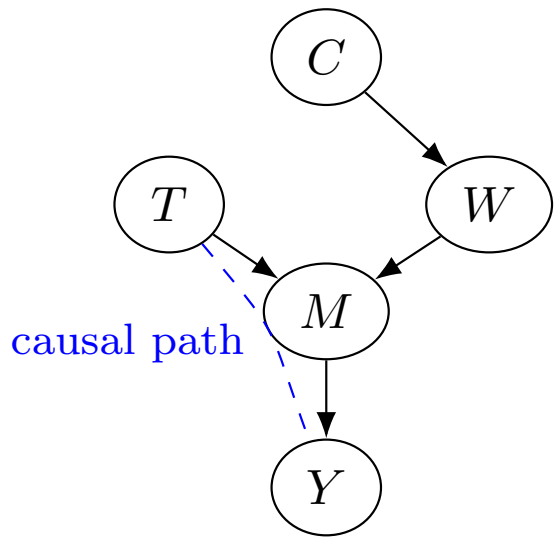
Backdoor adjustment - Blocking backdoor paths

Unblocked paths show the information flow (dependencies)

$P(Y|T,W)$



$P(Y|do(T))$



Conditioning on variables within backdoor paths blocks the non-causal associations

Backdoor criterion and the adjustment formula

Backdoor criterion

A set of variables X satisfies the backdoor criterion relative to sets of variables T and Y in a DAG $\mathcal{G}$ if

1. no node in X is a descendant of a node in T, and
2. X blocks/d-separates **every** path between T and Y that contains an arrow to T (backdoor paths)

Backdoor criterion and the adjustment formula

Backdoor criterion

A set of variables X satisfies the backdoor criterion relative to sets of variables T and Y in a DAG $\mathcal{G}$ if

1. no node in X is a descendant of a node in T, and
2. X blocks/d-separates **every** path between T and Y that contains an arrow to T (backdoor paths)

In the previous example, sets $\{C\}$ or $\{W\}$ or $\{C, W\}$ all satisfy the backdoor criterion relative to T, Y (but not $\{M\}$).

Backdoor criterion and the adjustment formula

Backdoor criterion

A set of variables X satisfies the backdoor criterion relative to sets of variables T and Y in a DAG $\mathcal{G}$ if

1. no node in X is a descendant of a node in T, and
2. X blocks/d-separates **every** path between T and Y that contains an arrow to T (backdoor paths)

In the previous example, sets $\{C\}$ or $\{W\}$ or $\{C, W\}$ all satisfy the backdoor criterion relative to T, Y (but not $\{M\}$).

Theorem - Backdoor adjustment formula

If X satisfies the backdoor criterion relative to T, Y, then the interventional distribution $P(Y|do(T))$ is identifiable and is given by

$$P(Y = y|do(T = t)) = \sum_x P(Y = y|T = t, X = x)P(X = x)$$

Backdoor adjustment formula $\stackrel{?}{\equiv}$ G-formula

G-formula:

$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}_X\left[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\right] \quad \text{if } Y_1, Y_0 \perp\!\!\!\perp T|X$$

Backdoor adjustment formula:

$$\begin{aligned} P(Y|do(T=t)) &= \sum_x P(Y|T=t,X=x)P(X=x) \\ &= \mathbb{E}_X[P(Y|T=t,X)] \end{aligned}$$

Backdoor adjustment formula $\stackrel{?}{\equiv}$ G-formula

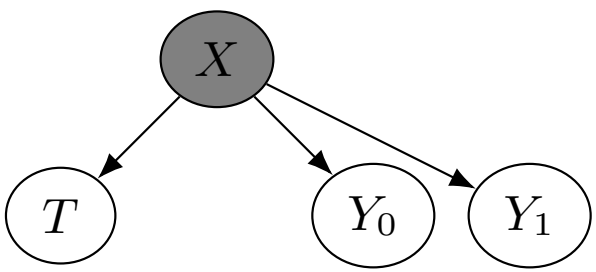
G-formula:

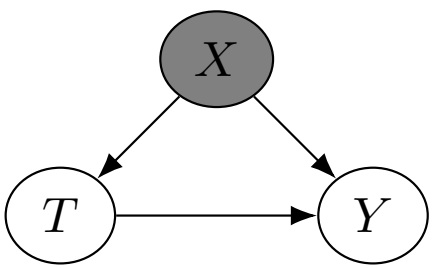
$$\mathbb{E}[Y_1 - Y_0] = \mathbb{E}_X\left[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\right] \quad \text{if } Y_1, Y_0 \perp\!\!\!\perp T|X$$

Backdoor adjustment formula:

$$\begin{aligned} P(Y|do(T=t)) &= \sum_x P(Y|T=t,X=x)P(X=x) \\ &= \mathbb{E}_X[P(Y|T=t,X)] \end{aligned}$$

X satisfies conditional ignorability $\equiv$ X satisfies the backdoor criterion

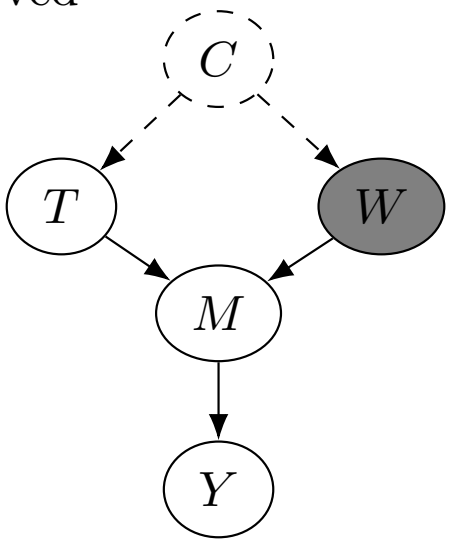




Backdoor adjustment formula for unobserved variables

- Backdoor adjustment formula works when all the variables are observed
- It can also be used when some variables are unobserved

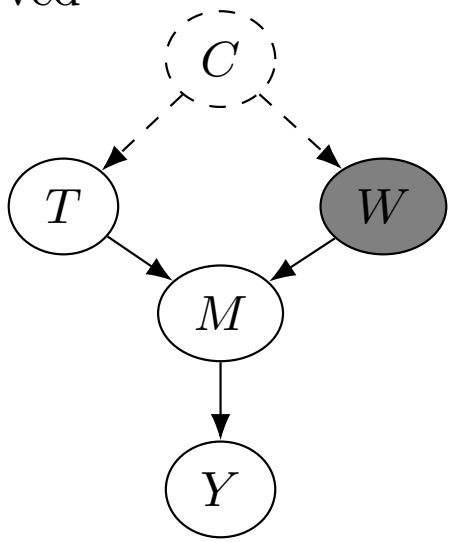
$$P(Y = y|do(T = t))$$
$$= \sum_w \underbrace{P(Y = y|T = t, W = w)}_{\text{observed}} \underbrace{P(W = w)}_{\text{observed}}$$



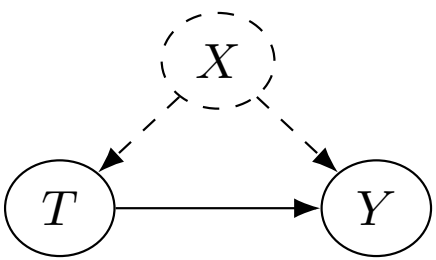
Backdoor adjustment formula for unobserved variables

- Backdoor adjustment formula works when all the variables are observed
- It can also be used when some variables are unobserved

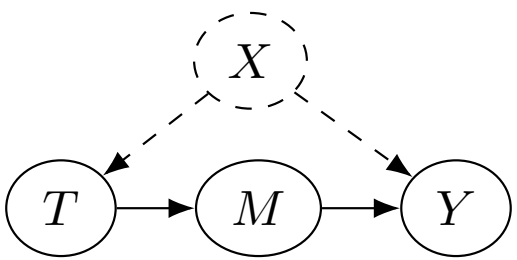
$$P(Y = y|do(T = t))$$
$$= \sum_w \underbrace{P(Y = y|T = t, W = w)}_{\text{observed}} \underbrace{P(W = w)}_{\text{observed}}$$



- What if all the variables that satisfy the backdoor criterion are unobserved?



or



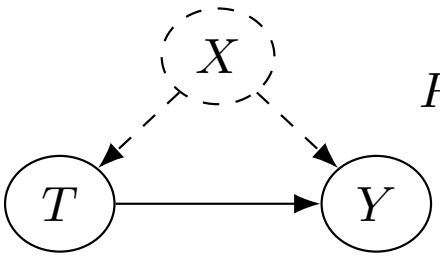
...

Is backdoor criterion necessary for identification?

- We saw that backdoor criterion is a sufficient condition for identification (using the adjustment formula)
- Is it also necessary? i.e., Is the causal effect non-identifiable if no observed variables satisfy the backdoor criterion?

Is backdoor criterion necessary for identification?

- We saw that backdoor criterion is a sufficient condition for identification (using the adjustment formula)
- Is it also necessary? i.e., Is the causal effect non-identifiable if no observed variables satisfy the backdoor criterion?

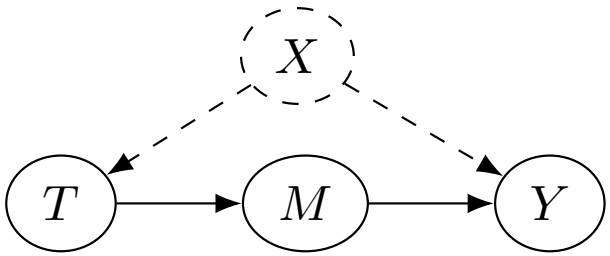


$P(Y|do(T))$ is non-identifiable for this graph

(we saw an example earlier)

Is backdoor criterion necessary for identification?

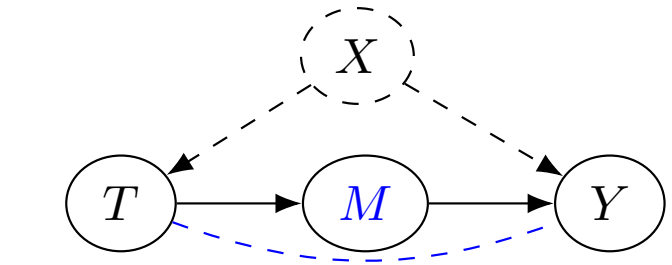
- We saw that backdoor criterion is a sufficient condition for identification (using the adjustment formula)
- Is it also necessary? i.e., Is the causal effect non-identifiable if no observed variables satisfy the backdoor criterion?



What about this graph?

Is backdoor criterion necessary for identification?

- We saw that backdoor criterion is a sufficient condition for identification (using the adjustment formula)
- Is it also necessary? i.e., Is the causal effect non-identifiable if no observed variables satisfy the backdoor criterion?

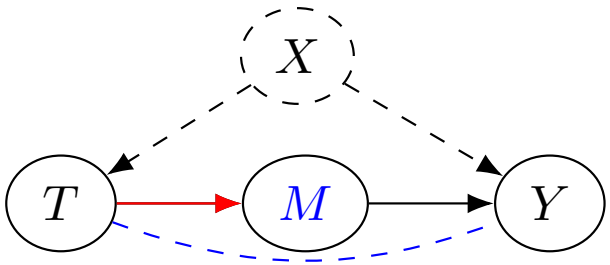


Mediator captures the causal association

What about this graph?

Is backdoor criterion necessary for identification?

- We saw that backdoor criterion is a sufficient condition for identification (using the adjustment formula)
- Is it also necessary? i.e., Is the causal effect non-identifiable if no observed variables satisfy the backdoor criterion?



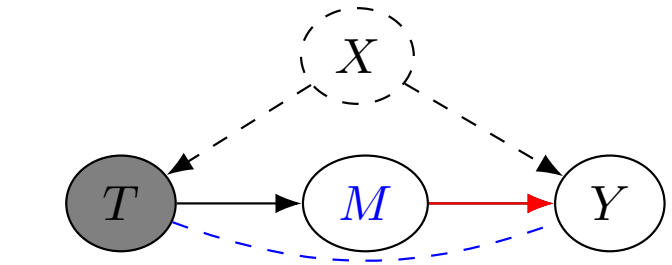
Mediator captures the causal association

What about this graph?

Step 1 Identify the causal effect of T on M: $P(m|do(T=t)) = P(m|t)$

Is backdoor criterion necessary for identification?

- We saw that backdoor criterion is a sufficient condition for identification (using the adjustment formula)
- Is it also necessary? i.e., Is the causal effect non-identifiable if no observed variables satisfy the backdoor criterion?



Mediator captures the causal association

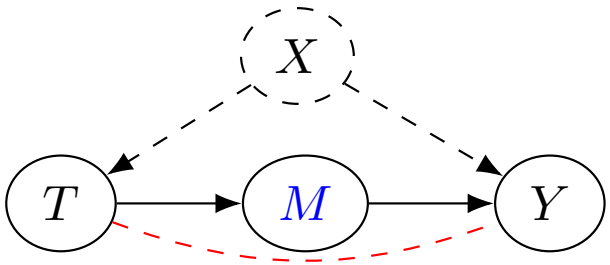
What about this graph?

Step 1 Identify the causal effect of T on M: $P(m|do(T=t)) = P(m|t)$

Step 2 Identify the causal effect of M on Y: T satisfies the backdoor criterion, $P(y|do(M=m)) = \sum_{t'} P(y|m,t')P(t')$

Is backdoor criterion necessary for identification?

- We saw that backdoor criterion is a sufficient condition for identification (using the adjustment formula)
- Is it also necessary? i.e., Is the causal effect non-identifiable if no observed variables satisfy the backdoor criterion?



Mediator captures the causal association

What about this graph?

Step 1 Identify the causal effect of T on M: $P(m|do(T=t)) = P(m|t)$

Step 2 Identify the causal effect of M on Y: T satisfies the backdoor criterion, $P(y|do(M=m)) = \sum_{t'} P(y|m,t')P(t')$

Step 3 Combine steps 1 and 2:

$$P(y|do(T=t)) = \sum_m P(m|do(T=t))P(y|do(M=m)$$

$$= \sum_m P(m|t) \sum_{t'} P(y|m,t')P(t')$$

Frontdoor criterion and adjustment formula

We were able to identify the causal effect even when the backdoor criterion was not satisfied

Frontdoor criterion

A set of variables M satisfies the frontdoor criterion relative to sets of variables T and Y in a DAG $\mathcal{G}$ if

1. M blocks all directed paths from T to Y;
2. no unblocked backdoor path from T to M; and
3. all backdoor paths from M to Y are blocked by T.

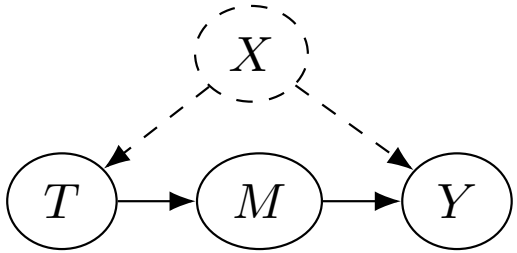
Theorem - Frontdoor adjustment formula

If M satisfies the frontdoor criterion relative to T, Y, then the interventional distribution $P(Y|do(T))$ is identifiable and is given by

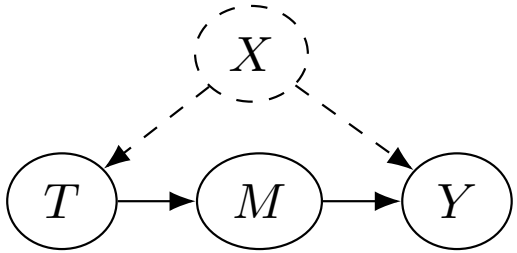
$$P(Y = y|do(T = t)) = \sum_m P(m|t) \sum_{t'} P(y|t', m)P(t')$$

Frontdoor adjustment - Proof with truncated factorization

$$
\begin{aligned}
&P(Y = y|do(T = t)) \\
&= \sum_{m,x} P(y, m, x|do(T = t)) \\
&= \sum_{m,x} P(m|t)P(y|m, x)P(x) \qquad \text{(truncated factorization)}
\end{aligned}
$$

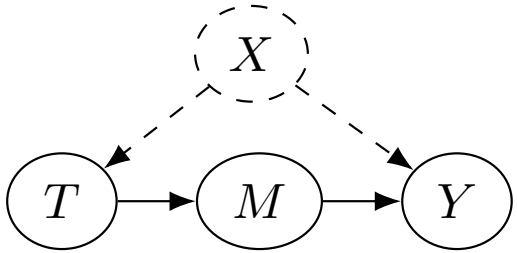


Frontdoor adjustment - Proof with truncated factorization



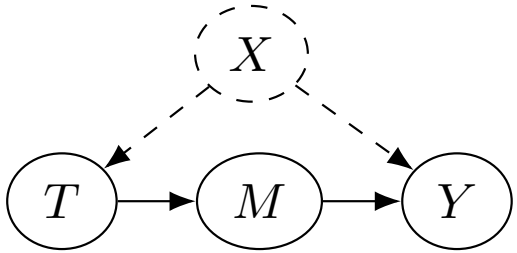
$$
\begin{aligned}
&P(Y = y|do(T = t)) \\
&= \sum_{m,x} P(y, m, x|do(T = t)) \\
&= \sum_{m,x} P(m|t)P(y|m, x)P(x) \qquad \text{(truncated factorization)} \\
&= \sum_{m} P(m|t) \sum_{x} P(y|m, x)P(x) \\
&= \sum_{m} P(m|t) \sum_{x} P(y|m, x) \sum_{t'} P(x, t') \qquad \text{(marginalize over } T\text{)}
\end{aligned}
$$

Frontdoor adjustment - Proof with truncated factorization



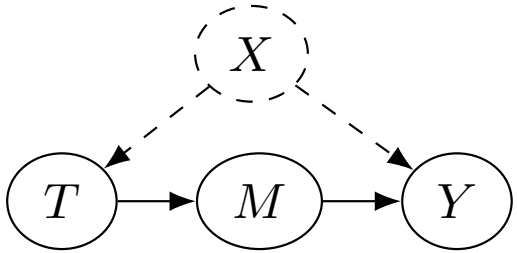
$$
\begin{aligned}
&P(Y = y|do(T = t)) \\
&= \sum_{m,x} P(y, m, x|do(T = t)) \\
&= \sum_{m,x} P(m|t)P(y|m, x)P(x) \qquad \text{(truncated factorization)} \\
&= \sum_{m} P(m|t) \sum_{x} P(y|m, x)P(x) \\
&= \sum_{m} P(m|t) \sum_{x} P(y|m, x) \sum_{t'} P(x, t') \qquad \text{(marginalize over } T\text{)} \\
&= \sum_{m} P(m|t) \sum_{x} \sum_{t'} P(y|m, x, t')P(x|t')P(t') \qquad (Y \perp\!\!\!\perp T|M, X)
\end{aligned}
$$

Frontdoor adjustment - Proof with truncated factorization



$$
\begin{aligned}
&P(Y = y|do(T = t)) \\
&= \sum_{m,x} P(y, m, x|do(T = t)) \\
&= \sum_{m,x} P(m|t)P(y|m, x)P(x) \qquad \text{(truncated factorization)} \\
&= \sum_{m} P(m|t) \sum_{x} P(y|m, x)P(x) \\
&= \sum_{m} P(m|t) \sum_{x} P(y|m, x) \sum_{t'} P(x, t') \qquad \text{(marginalize over } T\text{)} \\
&= \sum_{m} P(m|t) \sum_{x} \sum_{t'} P(y|m, x, t')P(x|t')P(t') \qquad (Y \perp\!\!\!\perp T|M, X) \\
&= \sum_{m} P(m|t) \sum_{x} \sum_{t'} P(y|m, x, t')P(x|t', m)P(t') \qquad (X \perp\!\!\!\perp M|T) \\
&= \sum_{m} P(m|t) \sum_{x} \sum_{t'} P(y, x|m, t')P(t')
\end{aligned}
$$

Frontdoor adjustment - Proof with truncated factorization



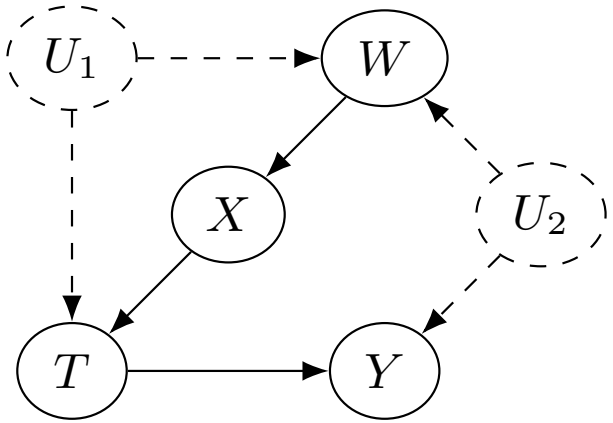
$$
\begin{aligned}
&P(Y = y|do(T = t)) \\
&= \sum_{m,x} P(y, m, x|do(T = t)) \\
&= \sum_{m,x} P(m|t)P(y|m, x)P(x) \qquad \text{(truncated factorization)} \\
&= \sum_{m} P(m|t) \sum_{x} P(y|m, x)P(x) \\
&= \sum_{m} P(m|t) \sum_{x} P(y|m, x) \sum_{t'} P(x, t') \qquad \text{(marginalize over } T\text{)} \\
&= \sum_{m} P(m|t) \sum_{x} \sum_{t'} P(y|m, x, t')P(x|t')P(t') \qquad (Y \perp\!\!\!\perp T|M, X) \\
&= \sum_{m} P(m|t) \sum_{x} \sum_{t'} P(y|m, x, t')P(x|t', m)P(t') \qquad (X \perp\!\!\!\perp M|T) \\
&= \sum_{m} P(m|t) \sum_{x} \sum_{t'} P(y, x|m, t')P(t') \\
&= \sum_{m} P(m|t) \sum_{t'} P(y|m, t')P(t') \qquad \text{(marginalize over } X\text{)}
\end{aligned}
$$

Questions?

Question

Any questions on frontdoor adjustment?

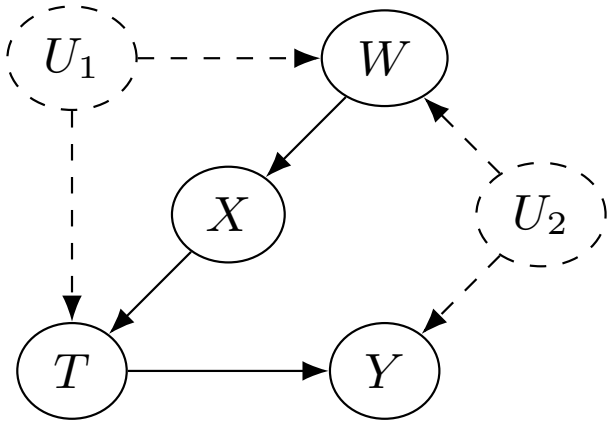
What if backdoor and frontdoor criteria don't work?



We are interested in the causal effect of cardiac output (T) on the blood pressure (Y). X is the heart rate and W is catecholamine (a stress hormone). The levels of total peripheral resistance (U_1) and analgesia (U_2) are unobserved. [1]

[1] Figure 1.a in Jung, Tian, and Bareinboim, 2021.

What if backdoor and frontdoor criteria don't work?

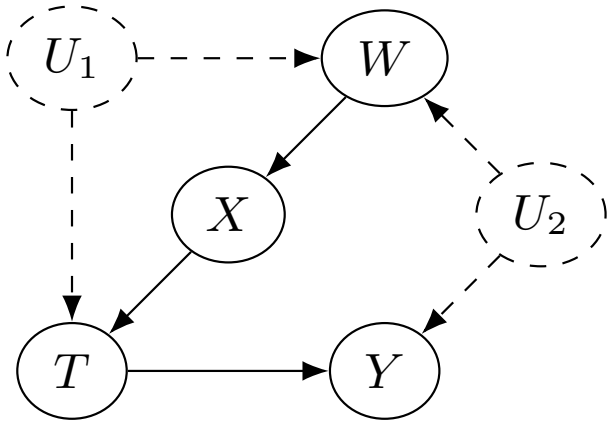


We are interested in the causal effect of cardiac output (T) on the blood pressure (Y). X is the heart rate and W is catecholamine (a stress hormone). The levels of total peripheral resistance (U_1) and analgesia (U_2) are unobserved. [1]

- There is an unobserved backdoor path between T and Y, T, U_1, W, U_2, Y: ~~Backdoor criterion~~
- There is no mediator between T and Y: ~~Frontdoor criterion~~

[1] Figure 1.a in Jung, Tian, and Bareinboim, 2021.

What if backdoor and frontdoor criteria don't work?



We are interested in the causal effect of cardiac output (T) on the blood pressure (Y). X is the heart rate and W is catecholamine (a stress hormone). The levels of total peripheral resistance (U_1) and analgesia (U_2) are unobserved. [1]

- There is an unobserved backdoor path between T and Y, T, U_1, W, U_2, Y: ~~Backdoor criterion~~
- There is no mediator between T and Y: ~~Frontdoor criterion~~
- We can use *do*-calculus to decide if $P(Y|do(T))$ is identifiable

[1] Figure 1.a in Jung, Tian, and Bareinboim, 2021.

Pearl's *do*-calculus

- *do*-calculus is a set of three inference rules that allows us to convert an interventional quantity into a probability expression involving observed quantities
- We'll consider general quantities $P(Y|do(T = t), X = x)$ for arbitrary (sets of) variables T, X, Y

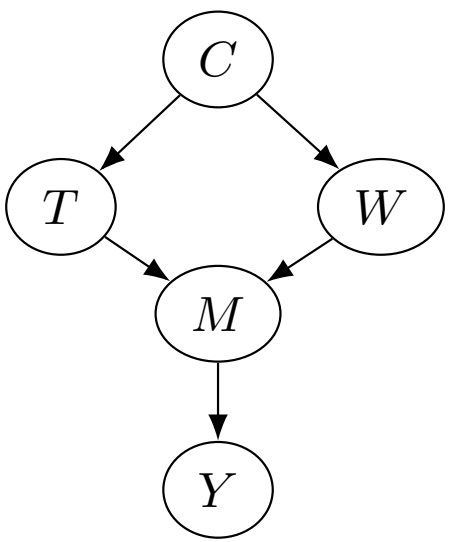
$$P(Y|do(T = t), X = x) := \frac{P(Y, X = x|do(T = t))}{P(X = x|do(T = t))}$$

Pearl's *do*-calculus

- *do*-calculus is a set of three inference rules that allows us to convert an interventional quantity into a probability expression involving observed quantities
- We'll consider general quantities $P(Y|do(T=t), X=x)$ for arbitrary (sets of) variables T, X, Y

$$P(Y|do(T=t), X=x) := \frac{P(Y, X=x|do(T=t))}{P(X=x|do(T=t))}$$

- Notation. Graph $\mathcal{G}$

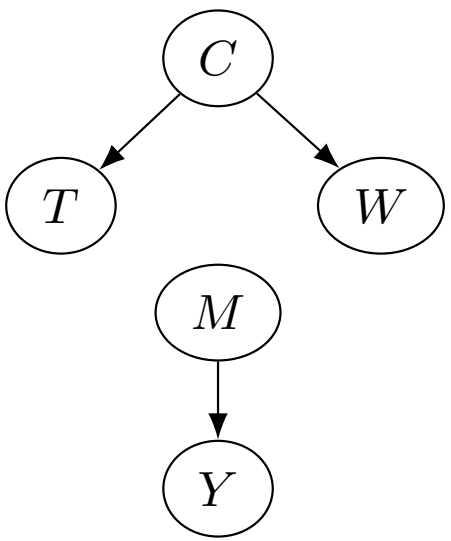


Pearl's *do*-calculus

- *do*-calculus is a set of three inference rules that allows us to convert an interventional quantity into a probability expression involving observed quantities
- We'll consider general quantities $P(Y|do(T=t), X=x)$ for arbitrary (sets of) variables T, X, Y

$$P(Y|do(T=t), X=x) := \frac{P(Y, X=x|do(T=t))}{P(X=x|do(T=t))}$$

- Notation. Graph $\mathcal{G}_{\overline{M}}$

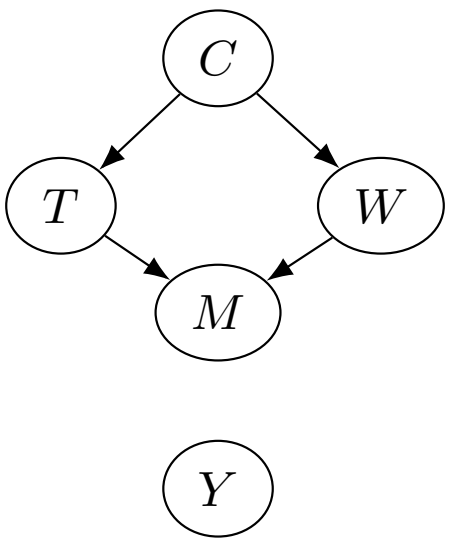


Pearl's *do*-calculus

- *do*-calculus is a set of three inference rules that allows us to convert an interventional quantity into a probability expression involving observed quantities
- We'll consider general quantities $P(Y|do(T=t), X=x)$ for arbitrary (sets of) variables T, X, Y

$$P(Y|do(T=t), X=x) := \frac{P(Y, X=x|do(T=t))}{P(X=x|do(T=t))}$$

- Notation. Graph $\mathcal{G}_{\underline{M}}$

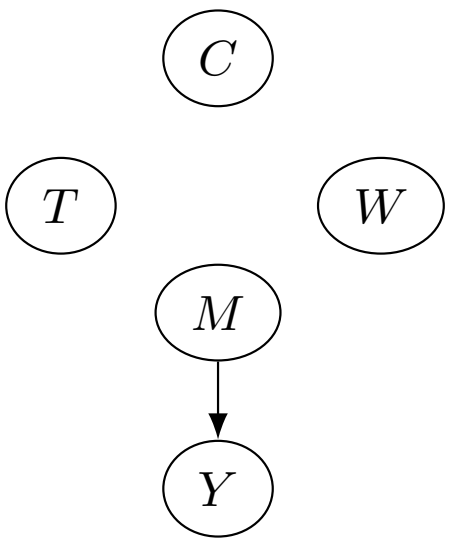


Pearl's *do*-calculus

- *do*-calculus is a set of three inference rules that allows us to convert an interventional quantity into a probability expression involving observed quantities
- We'll consider general quantities $P(Y|do(T=t), X=x)$ for arbitrary (sets of) variables T, X, Y

$$P(Y|do(T=t), X=x) := \frac{P(Y, X=x|do(T=t))}{P(X=x|do(T=t))}$$

- Notation. Graph $\mathcal{G}_{\underline{C},\overline{M}}$



Rule 1 of *do*-calculus - Insertion/deletion of observations

$$P(Y|do(T=t), X, W) = P(Y|do(T=t), W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T}}} X|T, W$$

Rule 1 of *do*-calculus - Insertion/deletion of observations

$$P(Y|do(T=t), X, W) = P(Y|do(T=t), W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T}}} X|T, W$$

Intuition:

- Remember that in the mutilated graph $\mathcal{G}_{\overline{T}}$, every path from T is causal. It can be seen as:

$$P(Y|T=t, X, W) = P(Y|T=t, W) \text{ if } Y \perp\!\!\!\perp_{\mathcal{G}} X|T, W$$

Generalization of d-separation

Rule 1 of *do*-calculus - Insertion/deletion of observations

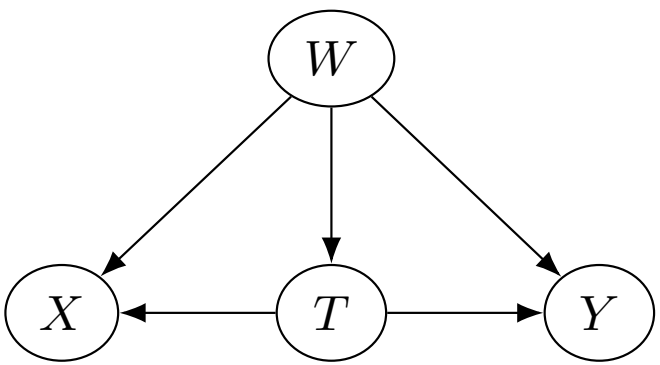
$$P(Y|do(T=t), X, W) = P(Y|do(T=t), W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T}}} X|T, W$$

Intuition:

- Remember that in the mutilated graph $\mathcal{G}_{\overline{T}}$, every path from T is causal. It can be seen as:

$$P(Y|T=t, X, W) = P(Y|T=t, W) \text{ if } Y \perp\!\!\!\perp_{\mathcal{G}} X|T, W$$

Generalization of d-separation



Rule 1 of *do*-calculus - Insertion/deletion of observations

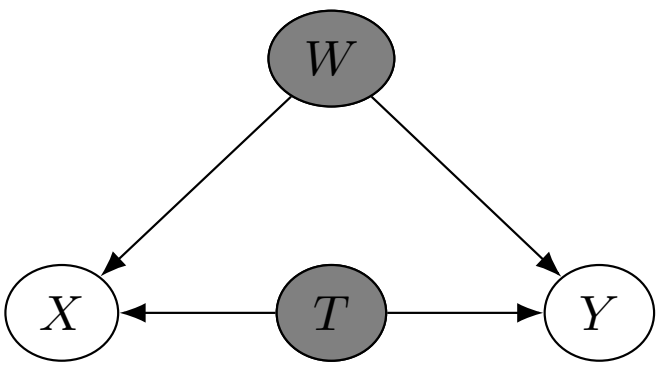
$$P(Y|do(T=t), X, W) = P(Y|do(T=t), W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T}}} X|T, W$$

Intuition:

- Remember that in the mutilated graph $\mathcal{G}_{\overline{T}}$, every path from T is causal. It can be seen as:

$$P(Y|T=t, X, W) = P(Y|T=t, W) \text{ if } Y \perp\!\!\!\perp_{\mathcal{G}} X|T, W$$

Generalization of d-separation



Rule 2 of *do*-calculus - Action/observation exchange

$$P(Y|do(T=t), do(X=x), W) = P(Y|do(T=t), X=x, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\underline{X}}} X|T, W$$

Rule 2 of *do*-calculus - Action/observation exchange

$$P(Y|do(T=t), do(X=x), W) = P(Y|do(T=t), X=x, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\underline{X}}} X|T, W$$

Intuition:

- Again, removing all edges to T can be seen as:

  $$P(Y|T=t, do(X=x)), W) = P(Y|T=t, X=x, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\underline{X}}} X|T, W$$

- All the backdoor paths from X to Y are blocked by T and W, i.e., conditioning on X = intervention on X

Generalization of backdoor criterion

Rule 2 of *do*-calculus - Action/observation exchange

$$P(Y|do(T=t), do(X=x), W) = P(Y|do(T=t), X=x, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\underline{X}}} X|T, W$$

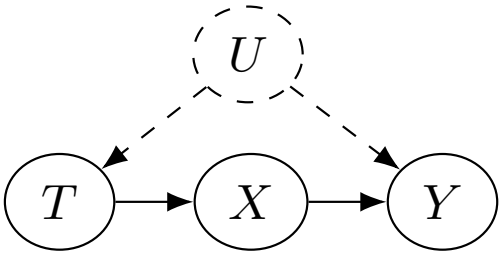
Intuition:

- Again, removing all edges to T can be seen as:

  $$P(Y|T=t, do(X=x)), W) = P(Y|T=t, X=x, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\underline{X}}} X|T, W$$

- All the backdoor paths from X to Y are blocked by T and W, i.e., conditioning on X = intervention on X

Generalization of backdoor criterion



$$P(Y|do(T=t), do(X=x)) =$$

Rule 2 of *do*-calculus - Action/observation exchange

$$P(Y|do(T=t), do(X=x), W) = P(Y|do(T=t), X=x, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\underline{X}}} X|T, W$$

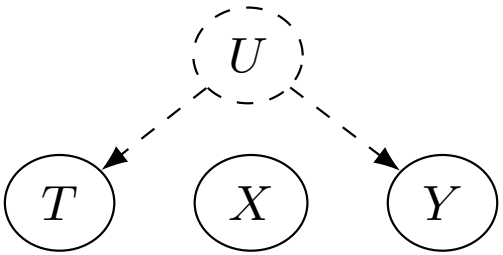
Intuition:

- Again, removing all edges to T can be seen as:

  $$P(Y|T=t, do(X=x)), W) = P(Y|T=t, X=x, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\underline{X}}} X|T, W$$

- All the backdoor paths from X to Y are blocked by T and W, i.e., conditioning on X = intervention on X

Generalization of backdoor criterion



$$P(Y|do(T=t), do(X=x)) = P(Y|do(T=t), X=x) \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\underline{X}}} X$$

Rule 3 of *do*-calculus - Insertion/deletion of actions

$$P(Y|do(T=t), do(X=x), W) = P(Y|do(T=t), W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T}, \overline{X(W)}}} X|T, W$$

where $X(W)$ is the set of nodes in X that are not ancestors of any node in W in $\mathcal{G}_{\overline{T}}$. i.e. W blocks the effect of interventions on X onto Y.

Rule 3 of *do*-calculus - Insertion/deletion of actions

$$P(Y|do(T=t), do(X=x), W) = P(Y|do(T=t), W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\overline{X(W)}}} X|T,W$$

where $X(W)$ is the set of nodes in X that are not ancestors of any node in W in $\mathcal{G}_{\overline{T}}$. i.e. W blocks the effect of interventions on X onto Y.

Intuition:

- Again, removing all edges to T the rule becomes:

$$P(Y|T=t, do(X=x), W) = P(Y|T=t, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{X(W)}}} X|T,W$$

Rule 3 of *do*-calculus - Insertion/deletion of actions

$$P(Y|do(T=t), do(X=x), W) = P(Y|do(T=t), W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T}, \overline{X(W)}}} X|T, W$$

where $X(W)$ is the set of nodes in X that are not ancestors of any node in W in $\mathcal{G}_{\overline{T}}$. i.e. W blocks the effect of interventions on X onto Y.

Intuition:

- Again, removing all edges to T the rule becomes:

$$P(Y|T=t, do(X=x), W) = P(Y|T=t, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{X}}} X|T, W$$

- Let first look at $\overline{X}$. It says there is no directed path between X and W, so intervention has no effect (we can delete it)

Rule 3 of *do*-calculus - Insertion/deletion of actions

$$P(Y|do(T=t), do(X=x), W) = P(Y|do(T=t), W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T}, \overline{X(W)}}} X|T, W$$

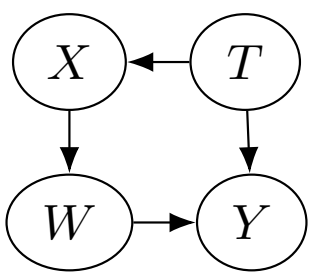
where $X(W)$ is the set of nodes in X that are not ancestors of any node in W in $\mathcal{G}_{\overline{T}}$. i.e. W blocks the effect of interventions on X onto Y.

Intuition:

- Again, removing all edges to T the rule becomes:

$$P(Y|T=t, do(X=x), W) = P(Y|T=t, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{X(W)}}} X|T, W$$

- Let first look at $\overline{X}$. It says there is no directed path between X and W, so intervention has no effect (we can delete it)
- Why $\overline{X(W)}$?



Rule 3 of *do*-calculus - Insertion/deletion of actions

$$P(Y|do(T=t), do(X=x), W) = P(Y|do(T=t), W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T}, \overline{X(W)}}} X|T, W$$

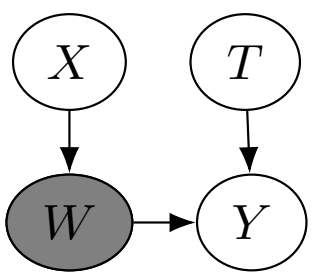
where $X(W)$ is the set of nodes in X that are not ancestors of any node in W in $\mathcal{G}_{\overline{T}}$. i.e. W blocks the effect of interventions on X onto Y.

Intuition:

- Again, removing all edges to T the rule becomes:

$$P(Y|T=t, do(X=x), W) = P(Y|T=t, W) \quad \text{if} \quad Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{X(W)}}} X|T, W$$

- Let first look at $\overline{X}$. It says there is no directed path between X and W, so intervention has no effect (we can delete it)
- Why $\overline{X(W)}$? X is independent of Y given W in $\mathcal{G}_{\overline{X}}$ but still has causal effect on Y. That's why ancestors of W are excluded



do-calculus is complete[1]

Theorem - Completeness of *do*-calculus

A causal effect $P(Y = y|do(T = t))$ is identifiable if and only if there exists a finite sequence of transformations, each conforming to one of the following inference rules that reduce $P(Y = y|do(T = t))$ into an expression involving observed quantities

1. Rule 1:

   $$P(Y|do(T = t), X, W) = P(Y|do(T = t), W) \quad \text{if } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T}}} X|T, W$$

2. Rule 2:

   $$P(Y|do(T = t), do(X = x), W) = P(Y|do(T = t), X = x, W)$$
   $$\text{if } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\underline{X}}} X|T, W$$

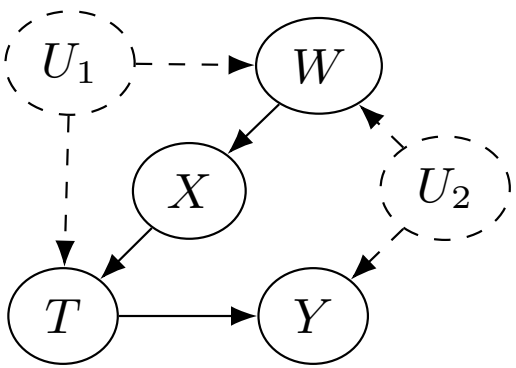
3. Rule 3:

   $$P(Y|do(T = t), do(X = x), W) = P(Y|do(T = t), W)$$
   $$\text{if } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\overline{X(W)}}} X|T, W$$

[1] Proof in Huang and Valtorta, 2012 and Shpitser and Pearl, 2012

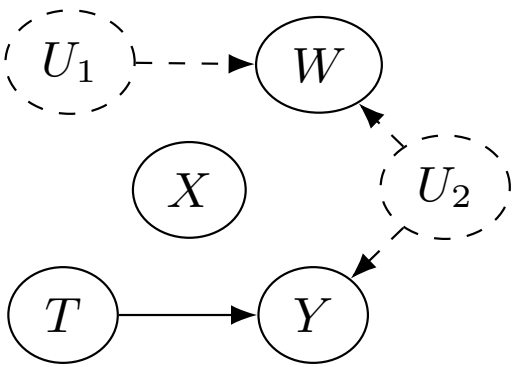
Example - Identification with *do*-calculus

$P(y|do(T = t))$



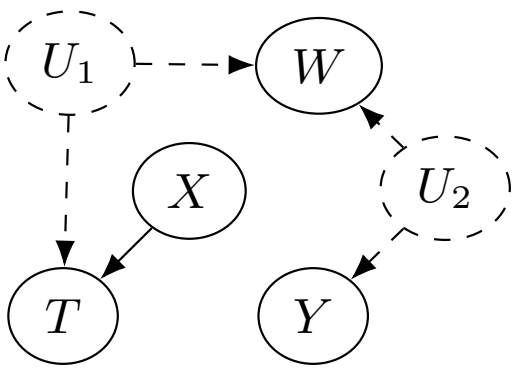
Example - Identification with *do*-calculus

$$P(y|do(T=t))$$
$$= P(y|do(T=t), do(X=x)) \quad \text{(Rule 3: insertion of actions - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\overline{X}}} X|T)$$



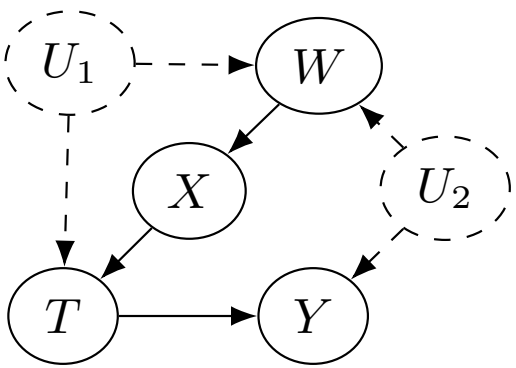
Example - Identification with *do*-calculus

$$
\begin{aligned}
&P(y|do(T=t)) \\
&= P(y|do(T=t), do(X=x)) \quad \text{(Rule 3: insertion of actions - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\overline{X}}} X|T) \\
&= P(y|T=t, do(X=x)) \quad \text{(Rule 2: action/observation exchange - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{X},\underline{T}}} T|X)
\end{aligned}
$$



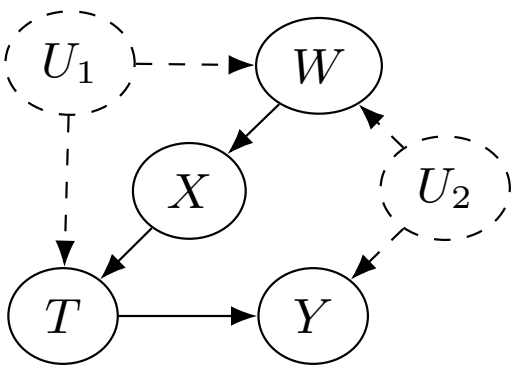
Example - Identification with *do*-calculus

$$
\begin{aligned}
&P(y|do(T=t)) \\
&= P(y|do(T=t), do(X=x)) \quad \text{(Rule 3: insertion of actions - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\overline{X}}} X|T) \\
&= P(y|T=t, do(X=x)) \quad \text{(Rule 2: action/observation exchange - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{X},\underline{T}}} T|X) \\
&= \frac{P(y,t|do(X=x))}{P(t|do(X=x))}
\end{aligned}
$$



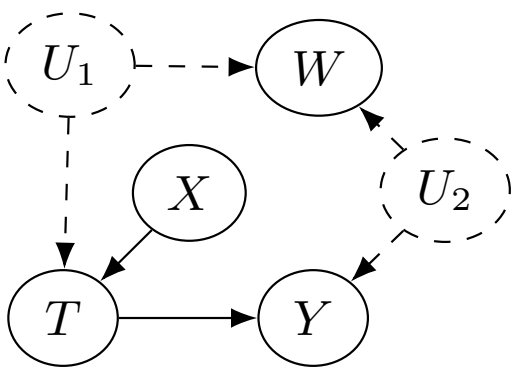
Example - Identification with *do*-calculus

$$
\begin{aligned}
&P(y|do(T=t)) \\
&= P(y|do(T=t), do(X=x)) \quad \text{(Rule 3: insertion of actions - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\overline{X}}} X|T) \\
&= P(y|T=t, do(X=x)) \quad \text{(Rule 2: action/observation exchange - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{X},\underline{T}}} T|X) \\
&= \frac{P(y,t|do(X=x))}{P(t|do(X=x))} \\
&= \frac{\sum_w P(y,t|W=w, do(X=x))P(w|do(X=x))}{\sum_w P(t|W=w, do(X=x)P(w|do(X=x))} \quad \text{(Marginalization over } W)
\end{aligned}
$$



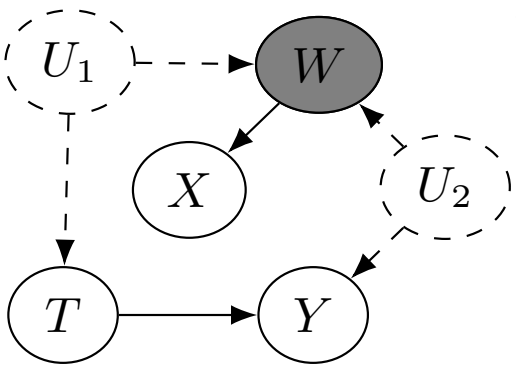
Example - Identification with *do*-calculus

$$
\begin{aligned}
&P(y|do(T=t)) \\
&= P(y|do(T=t), do(X=x)) \quad \text{(Rule 3: insertion of actions - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\overline{X}}} X|T) \\
&= P(y|T=t, do(X=x)) \quad \text{(Rule 2: action/observation exchange - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{X},\underline{T}}} T|X) \\
&= \frac{P(y,t|do(X=x))}{P(t|do(X=x))} \\
&= \frac{\sum_w P(y,t|W=w, do(X=x))P(w|do(X=x))}{\sum_w P(t|W=w, do(X=x)P(w|do(X=x))} \quad \text{(Marginalization over } W) \\
&= \frac{\sum_w P(y,t|W=w, do(X=x))P(w)}{\sum_w P(t|W=w, do(X=x)P(w)} \quad \text{(Rule 3: deletion of actions - } W \perp\!\!\!\perp_{\mathcal{G}_{\overline{X}}} X)
\end{aligned}
$$



Example - Identification with *do*-calculus

$$
\begin{aligned}
&P(y|do(T=t)) \\
&= P(y|do(T=t), do(X=x)) \quad \text{(Rule 3: insertion of actions - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{T},\overline{X}}} X|T) \\
&= P(y|T=t, do(X=x)) \quad \text{(Rule 2: action/observation exchange - } Y \perp\!\!\!\perp_{\mathcal{G}_{\overline{X},\underline{T}}} T|X) \\
&= \frac{P(y,t|do(X=x))}{P(t|do(X=x))} \\
&= \frac{\sum_w P(y,t|W=w, do(X=x))P(w|do(X=x))}{\sum_w P(t|W=w, do(X=x)P(w|do(X=x))} \quad \text{(Marginalization over } W) \\
&= \frac{\sum_w P(y,t|W=w, do(X=x))P(w)}{\sum_w P(t|W=w, do(X=x)P(w)} \quad \text{(Rule 3: deletion of actions - } W \perp\!\!\!\perp_{\mathcal{G}_{\overline{X}}} X) \\
&= \frac{\sum_w P(y,t|W=w, X=x)P(w)}{\sum_w P(t|W=w, X=x)P(w)} \\
&\qquad \text{(Rule 2: action/observation exchange - } T, Y \perp\!\!\!\perp_{\mathcal{G}_{\underline{X}}} X|W)
\end{aligned}
$$



Questions?

Question

Any questions on do-calculus?

Where does the graph come from?

- Prior knowledge
- Guess a graph and test whether its edges match the conditional independencies in data
- Discovery algorithms

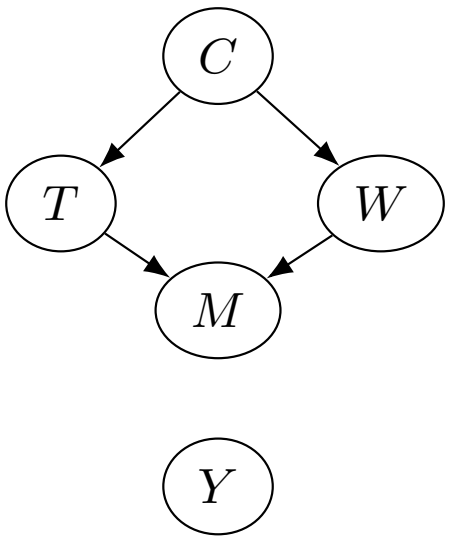
Prior knowledge

- Talk to a practitioner and ask them to create a *story* of how the random variables interact in practice,
- Convert the story into a graphical representation.

Problem: Process is prone to error and subjective biases.

Guessing and testing

- Start with a graph,
- Find all the variables that are d-separated in the graph,
- Run independence tests to assess whether the d-separation set holds.



Test for $\{Y \perp\!\!\!\perp T, M, W, C\}$, $\{T \perp\!\!\!\perp W | C\}$, $\{M \perp\!\!\!\perp C | T, W\}$

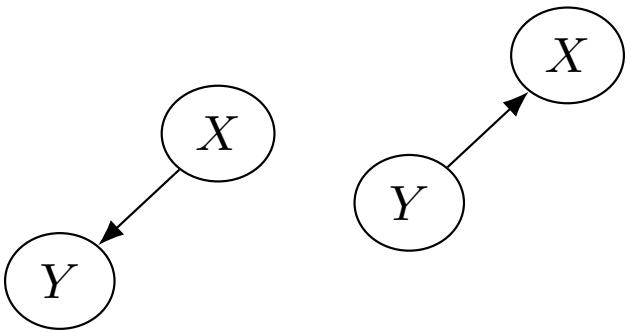
Testing conditional independence

Completely general CI testing is open and active area of research!

- $X \perp\!\!\!\perp Y|Z \iff I[X;Y|Z] = 0$; but efficient estimators for mutual information are tricky to find.
- Discrete variables: χ^2 test
- Normally distributed variables $X \perp\!\!\!\perp Y|Z$ equivalent to zero partial correlation
- $P(Y|X,Z) = P(Y|Z)$ could be checked via non-parameteric density estimation.
- By the Markov Properties we know that $X \perp\!\!\!\perp Y|Z \iff Z$ d-separates X, Y.
- **Can we use this principle in conjunction with conditional tests to find graphs?**.

Two variable case

Lets start with the simplest two-variable case:



- Both these cases imply $X \not\perp\!\!\!\perp Y$.
- Cannot distinguish between them due to Observational Equivalence but we know there is an edge.

Three variable case

Enumerate the possibilities of graphs that could have generated the data:

1. $X \longrightarrow Z \longrightarrow Y$
2. $X \longleftarrow Z \longrightarrow Y$
3. $X \longleftarrow Z \longleftarrow Y$
4. $X \longrightarrow Z \longleftarrow Y$

- (1-3) are observationally equivalent but (4) represents a collider
- If $X \perp\!\!\!\perp Y | Z$ then we can narrow down the skeleton of the graph even if we don't know the orientation.
- If $X \not\perp\!\!\!\perp Y | Z$ then we know there is a collider and can orient edges.

General case - PC Algorithm Spirtes and Glymour, 1991

1. Start with a complete undirected graph.
2. For each pair X, Y if $X \perp\!\!\!\perp Y$, remove their edge.
3. For each X, Y still connected and each third variable Z check $X \perp\!\!\!\perp Y|Z$. If yes, remove edge between X and Y.
4. For each X, Y still connected and each third/fourth variable Z_1, Z_2 check $X \perp\!\!\!\perp Y|Z_1, Z_2$. If yes remove edge between X and Y.
5. .....
6. For each X, Y still connected and all other $N-2$ variables $Z_1, Z_2, \ldots, Z_k$ check $X \perp\!\!\!\perp Y|Z_1, Z_2, \ldots, Z_k$. If yes remove edge between X and Y.

Analysis - PC Algorithm

Assumptions for the PC algorithm

- P satisfies the causal Markov property on $\mathcal{G}$.
- There are no unobserved variables.
- There is one graph $\mathcal{G}$ to which P satisfies the Markov property.

Problem: The DAG learned by the algorithm need not be acyclic!
Why is this a problem?

Learning directed acyclic graphs

- For learning DAGs, there are several score based hill-climbing algorithms for structure learning of directed acyclic graphs.

- They learn via the following optimization problem:

$$\min_{\mathcal{G}} \text{loss}(\mathcal{G}) \text{ s.t. } \mathcal{G} \in \text{DAG}$$

- What constitutes a good score function?
  - Number should be low if the model *explains* the data and high if it does not.
  - When learning $p(y|x)$ we maximize the log-likelihood of labels y given features x to learn parameters of the conditional distribution.
  - Posit a class of functions that generates the observations and use fit to data for learning *structure*.

Learning DAGs with linear structural causal models

- We can represent any d-dimensional graph of linear structural causal models in matrix notation as follows:

1. Let $W \in \mathbb{R}^{d \times d}$ be a weight matrix representing the strength of edges and $G(W)$ denote the graph,
2. $B \in \{0,1\}^{d \times d}$ where $B[i,j] = 0 \iff w_{ij} = 0$ is the (binary) adjacency matrix,
3. $X_j = w_j^T X_j + \epsilon_j$ where $X = (X_1, \ldots, X_d)$ are each dimensions of data (nodes in the graph) and $\epsilon = (\epsilon_1, \ldots, \epsilon_d)$ are noise variables,
4. For data matrix D, we can measure fit to data via a least-squares loss $l(W, D) = \frac{1}{2n}||D - DW||_F^2$.
5. We can regularize the loss function to learn a sparse DAG fits the data: $F(W, D) = l(W, D) + \lambda||W||_1$.
6. Finding DAGs then reduces to $\min_{W \in \mathbb{R}^{d \times d}} F(W, D)$ s.t. $G(W) \in$ DAGs

Searching over DAGs

- Optimization problem is NP hard. Challenging due to the constraint in the optimization problem,
- Acyclicity is a combinatorial constraint with the number of structures increasing super exponentially in d,
- DAGS with no TEARS, Zheng et al., 2018, comes up with a creative solution to this problem!

Insight 1: Binary Adjacency Matrices and cycles

- Fact 1: $\operatorname{tr} B^k$ counts the number of length k closed paths (cycles) in a directed graph,
- Fact 2: DAG has no cycle iff $\sum_{k=1}^{\infty} \sum_{i=1}^{d} (B^k)_{ii} = 0$
- Consequence, B is a DAG iff $\operatorname{tr}(\mathbb{I} - B)^{-1} = d$

$$
\begin{aligned}
\operatorname{tr}(\mathbb{I} - B)^{-1} &= \operatorname{tr} \sum_{k=0}^{\infty} B^k && \text{(Infinite geometric series)} \\
&= \operatorname{tr} \mathbb{I} + \operatorname{tr} \sum_{k=1}^{\infty} B^k \\
&= d + \sum_{k=1}^{\infty} \sum_{i=1}^{d} (B^k)_{ii} \\
&= d
\end{aligned}
$$

However B^k is difficult to compute and represent in computer memory.

Insight 2: Matrix exponents and weighted graphs

- We can use the matrix exponential $\exp X = \sum_{k=0}^{\infty} \frac{1}{k!} X^k$ which is well-defined.
- Consequence, B is a DAG iff $\operatorname{tr} \exp B = d$, and its extension to the graph with weighted edges (Linear SCM) case yields:

Theorem - Characterizing DAGs with matrix exponents Zheng et al., 2018

A matrix $W \in \mathbb{R}^{d \times d}$ is a DAG iff:

$$h(W) = \operatorname{tr} \exp(W \circ W) - d = 0$$

where $\circ$ is the Hadamard product and

$$\nabla_W h(W) = \exp(W \circ W)^T \circ 2W$$

DAGS with no TEARS

Smooth characterizations of acyclicity

- $h(W) = 0$ iff W is acyclic (i.e. G(W) represents a DAG),
- $h(W)$ quantifies the DAGness of a graph,
- h is smooth and has easy to compute derivatives.

Now, structure learning of a DAG (under a linear SCM) can be done via : $\min_{W \in \mathbb{R}^{d \times d}} F(W)$ s.t. $h(W) = 0$.

Recap - Lecture 3

- Counterfactuals: Answering them requires the recovery of the unobserved noise that generated the data,
- Identifiability: Translating interventional queries into their observational counterparts:
  - Backdoor criteria: Identical to adjustment via the G-formula,
  - Frontdoor criteria: Using mediators to identify causal effect on outcomes.
- Do-Calculus: Three rules to identify causal effects:
  1. Insertion or deletion of observations : Generalization of d-separation,
  2. Interchanging actions with observations : Generalization of the backdoor criteria,
  3. Insertion or delection of actions

Jung, Yonghan, Jin Tian, and Elias Bareinboim (2021). "Estimating identifiable causal effects through double machine learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 13, pp. 12113–12122.

Huang, Yimin and Marco Valtorta (2012). "Pearl's calculus of intervention is complete". In: *arXiv preprint arXiv:1206.6831*.

Shpitser, Ilya and Judea Pearl (2012). "Identification of conditional interventional distributions". In: *arXiv preprint arXiv:1206.6876*.

Spirtes, Peter and Clark Glymour (1991). "An algorithm for fast recovery of sparse causal graphs". In: *Social science computer review* 9.1, pp. 62–72.

Zheng, Xun et al. (2018). "Dags with no tears: Continuous optimization for structure learning". In: *Advances in Neural Information Processing Systems* 31.