CSC2541: Introduction to Causality

Lecture 1 - Introduction and Motivation

Instructor: Rahul G. Krishnan

TA & slides: Vahid Balazadeh-Meresht

September 12, 2022

Why am I interested in causality?

- Assistant Professor in Computer Science and Medicine, CIFAR AI Chair at the Vector Institute
- **Research goal:** Machine learning for healthcare
- **Vision:** Autonomous agents for clinical decision support
- A lot of healthcare is asking the question "So what should I do?"
- Need to understand the effect of interventions and how to build systems integrate ideas from causal inference will be an important part of realizing that vision.

Course logistics

- All course related material and announcements will be found at: `https://csc2541-2022.github.io/`
- Office hours: M11-12 in Pratt 286
- Mark breakdown:
  - Individual: Problem set (15%) and Paper summary (15%)
  - Group: Paper Presentation (15%) and Project (55%)
- Preqrequisite: Strong background in linear algebra, statistics, Bayesian networks and latent variable modeling
- Lot to cover and very little time – will post slides before class starts.

Success in the course project

- Worth more than half the grade in the course.
- Some courses start the project mid-way through the semester. Start thinking about the class project in the second week.
- Project proposal due October 10 (less than a month). See instructions here:
  `https://csc2541-2022.github.io/assignments/projectproposal`
- Talk to the people around you and start figuring out joint themes in your research/interests.
- Start taking a look at the Project Resources page `https://csc2541-2022.github.io/projectresources` to brainstorm among your colleagues.

Feedback welcome

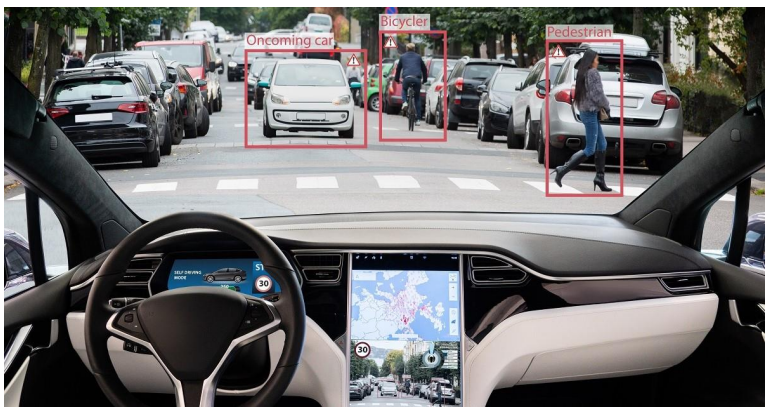
- This is the first iteration of the class, your feedback will shape it for the generations to come! We'll have a midterm survey for the course.
- Vahid and I believe the material here is fundamental enough to eventually become an undergraduate class.
- Causal inference has been studied and developed in a variety of fields ranging from statistics, biostatistics, machine learning, economics, biology. Literature is vast and notation varies across disciplines.
- The goal of this course: help you read, understand and incorporate ideas from causal inference in your own work.

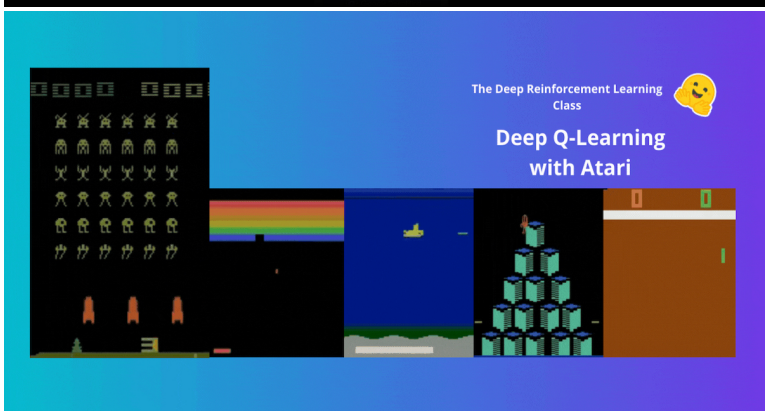
Questions?

Question

Any questions on logistics?

Deep Reinforcement learning and scientific discovery





NLP and vision

Natural language processing

- Large language models: BERT, GPT-3, PaLM
- Language generation from images
- Sentiment analysis

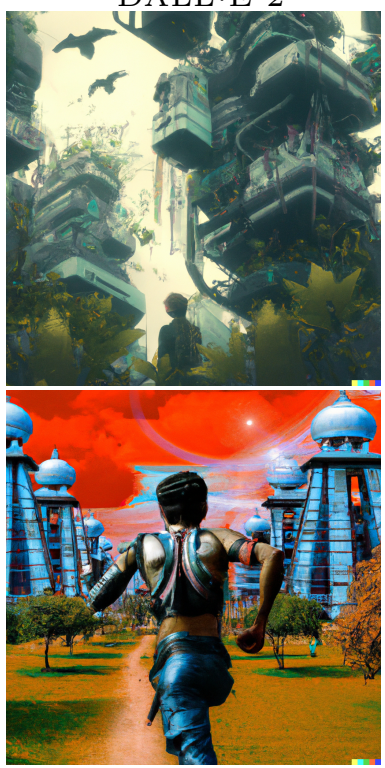
Computer vision

- Image classification
- Image generation (from text)
- Segmentation

Benefits

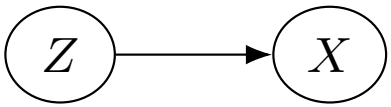
- Superhuman performance on some tasks
- Ability to learn from large datasets
- Model complex functions
- Rich representations with continuous optimization

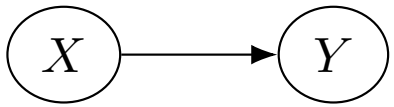
DALL·E 2



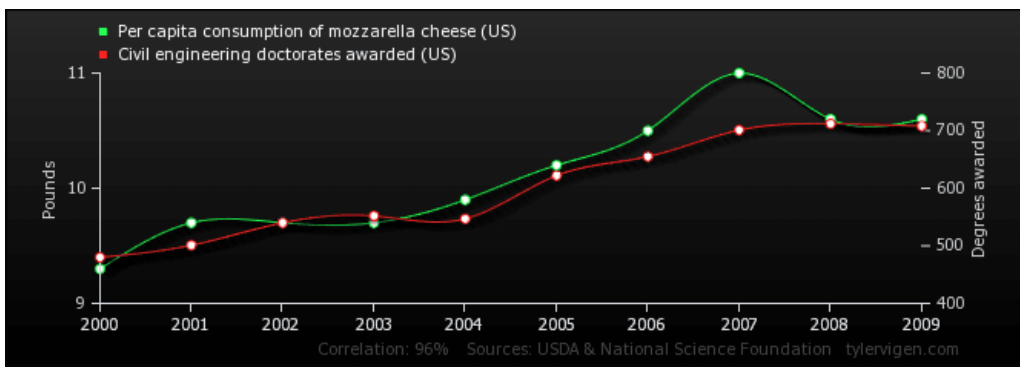
Adapted from "Causality and Deep Learning: Synergies, Challenges, and the Future," 2022.

Successes driven by advances in deep learning





- Building predictive models of labels given data X ([*]Nets, [*]formers etc.),
- Using latent variable models to extract latent structure Z from data X (GANs, VAEs),
- We've gotten very good at the art of developing new architectures and learning algorithms that can capture complex correlations between high-dimensional random variables,
- But correlation is not causation.



Source: `https://www.fastcompany.com/3030529/hilarious-graphs-prove-that-correlation-isnt-causation`

Deep learning can have poor out-of-distribution generalization

Deep learning models are excellent at picking up on latent statistical relationships. E.g., Grass and cow appears with a higher chance

(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

"Recognition in Terra Incognita," ECCV, 2018.

Why is it hard to generalize to a new environment with a new data distribution?

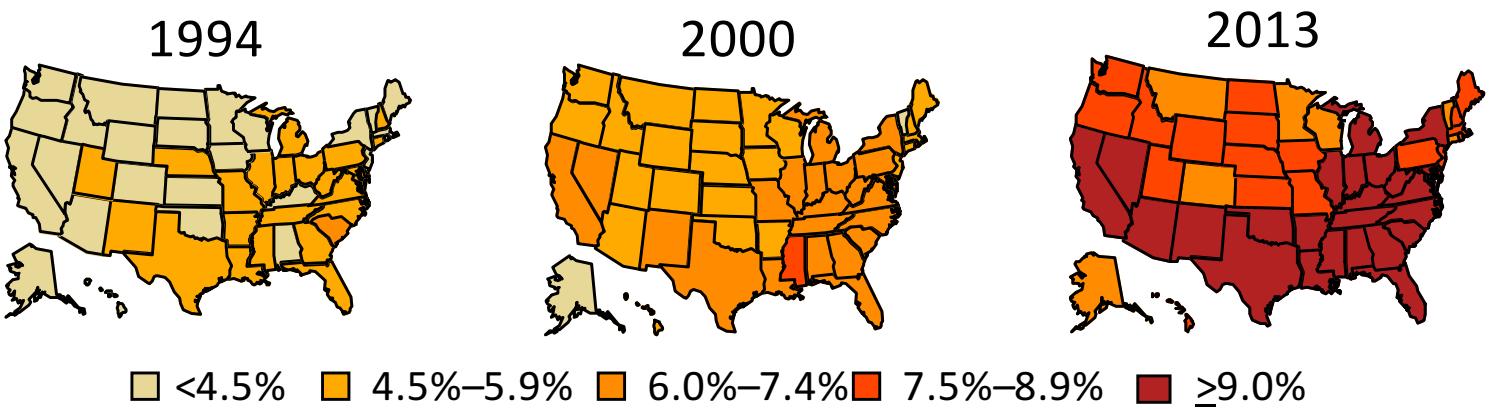
Catastrophic forgetting and continual learning

- One possibility is to retrain models in the new environment. However, this often results in degradation of performance in the original environment, a phenomena called **catastrophic forgetting**.
- A branch of ML known as **continual learning** seeks to build models that can continued to be trained in new environments.
- Human's have a remarkable ability to capture cause and effect relationships even when we move to new environments! How can we translate this ability to models that learn?

What this class is, and is not

- An overview of the foundations and the assumptions that underlie when causal inference is feasible.
- Give you knowledge of when one can tease apart the effect of an intervention from data alone and when it is not.
- Understand some of the algorithms that underlie classical work over the past decades in the field across disciplines.
- Not sufficient to start making original research contributions in causal inference, but we hope you will appreciate the hardness that underlies these problems and inspire you to think of creative projects that leverage these ideas.

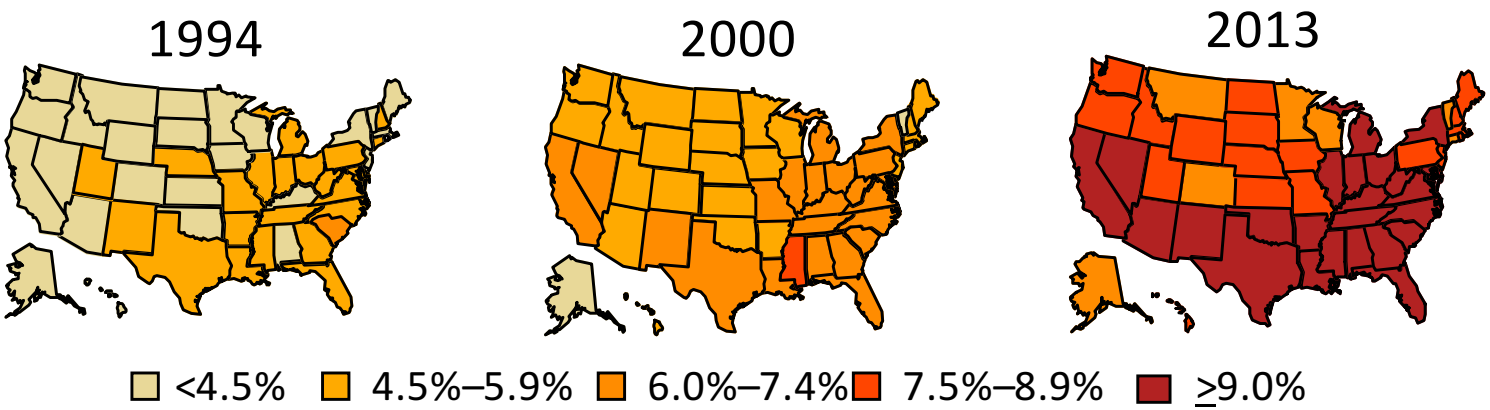
Example 1 - Risk stratification



- We can use machine learning for early detection of Type 2 diabetes
- Health system doesn't want to know how to predict diabetes - They want to know how to prevent it

Slide credits to David Sontag at MIT

Example 1 - Risk stratification



- We can use machine learning for early detection of Type 2 diabetes
- Health system doesn't want to know how to predict diabetes - They want to know how to prevent it
- Gastric bypass surgery is the highest negative weight (9^{th} most predictive feature)
  - Does this mean it would be a good intervention?

Slide credits to David Sontag at MIT

Example 2 - Simpson's paradox

Consider the following dataset on the recovery rate of two treatment procedures for kidney stones[1]

	Overall	Feature A	Feature B
Treatment a Open surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Treatment b Percutaneous nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

Question

Which treatment should we choose for a new patient?

[1] Table 6.1. Peters, Janzing, and Schlkopf, 2017

Example 2 - Simpson's paradox

Consider the following dataset on the recovery rate of two treatment procedures for kidney stones[1]

	Overall	Feature A	Feature B
Treatment a Open surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Treatment b Percutaneous nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

Question

Which treatment should we choose for a new patient?

Paradox: choose treatment a if the patient's feature is known, otherwise choose b!

[1] Table 6.1. Peters, Janzing, and Schlkopf, 2017

Simpson's paradox - Case 1

Case 1 Assume the groups represent the kidney stone size

	Overall	Small Stone	Large Stone
Treatment a Open surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Treatment b Percutaneous nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

Simpson's paradox - Case 1

Case 1 Assume the groups represent the kidney stone size

	Overall	Small Stone	Large Stone
Treatment a Open surgery	78%(273/350)	**93**%(**81**/**87**)	**73**%(**192**/**263**)
Treatment b Percutaneous nephrolithotomy	**83**%(**289**/**350**)	87%(234/270)	69%(55/80)

- Patients with larger stone sizes received treatment a more than the other group

Simpson's paradox - Case 1

Case 1 Assume the groups represent the kidney stone size

	Overall	Small Stone	Large Stone
Treatment a Open surgery	78%(273/350)	**93**%(**81**/**87**)	**73**%(**192**/**263**)
Treatment b Percutaneous nephrolithotomy	**83**%(**289**/**350**)	87%(234/270)	69%(55/80)

- Patients with larger stone sizes received treatment a more than the other group
- Patients with larger stones are less likely to recover (73%, 69% v.s. 93%, 87%)

Simpson's paradox - Case 1

Case 1 Assume the groups represent the kidney stone size

	Overall	Small Stone	Large Stone
Treatment a Open surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Treatment b Percutaneous nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

- Patients with larger stone sizes received treatment a more than the other group
- Patients with larger stones are less likely to recover $(73\%, 69\%$ v.s. $93\%, 87\%)$
- Hence, even though the overall data supports treatment b, treatment a has better recovery rate

Simpson's paradox - Case 2

Case 2 Assume the groups represent the blood pressure (BP) during the treatment

	Overall	Normal BP	High/low BP
Treatment a Open surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Treatment b Percutaneous nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

Simpson's paradox - Case 2

Case 2 Assume the groups represent the blood pressure (BP) during the treatment

	Overall	Normal BP	High/low BP
Treatment a Open surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Treatment b Percutaneous nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

- Patients after receiving treatment a are more likely to experience high/low BP

Simpson's paradox - Case 2

Case 2 Assume the groups represent the blood pressure (BP) during the treatment

	Overall	Normal BP	High/low BP
Treatment a Open surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Treatment b Percutaneous nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

- Patients after receiving treatment a are more likely to experience high/low BP
- Patients with high/low BP are less likely to recover

Simpson's paradox - Case 2

Case 2 Assume the groups represent the blood pressure (BP) during the treatment

	Overall	Normal BP	High/low BP
Treatment a Open surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Treatment b Percutaneous nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

- Patients after receiving treatment a are more likely to experience high/low BP
- Patients with high/low BP are less likely to recover
- Treatment a does better after stratifying by BP but high/low BP is a consequence of treatment a so it doesn't make sense to stratify by BP.

Simpson's paradox - Case 2

Case 2 Assume the groups represent the blood pressure (BP) during the treatment

	Overall	Normal BP	High/low BP
Treatment a Open surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Treatment b Percutaneous nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

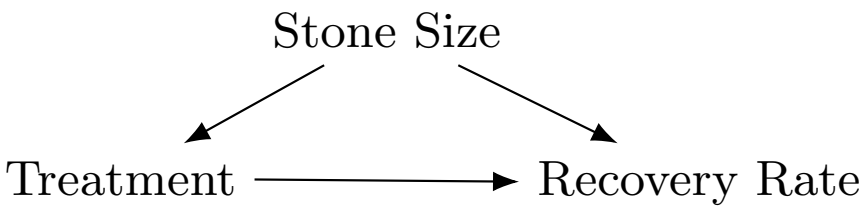
- Patients after receiving treatment a are more likely to experience high/low BP
- Patients with high/low BP are less likely to recover
- Treatment a does better after stratifying by BP but high/low BP is a consequence of treatment a so it doesn't make sense to stratify by BP.
- Choose treatment b based on the overall recovery rate

Simpson's paradox - assumptions and data

- Lets start drawing some graphs to represent these different cases.
- The data, i.e., (conditional) distributions, are the same in both cases.

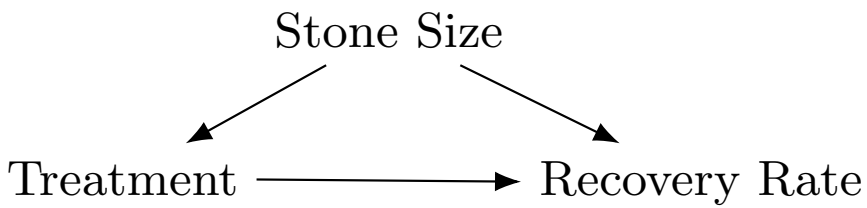
Simpson's paradox - assumptions and data

- Lets start drawing some graphs to represent these different cases.
- The data, i.e., (conditional) distributions, are the same in both cases.
- In case 1, we **assumed** the choice of treatment is influenced by the stone size, i.e.,

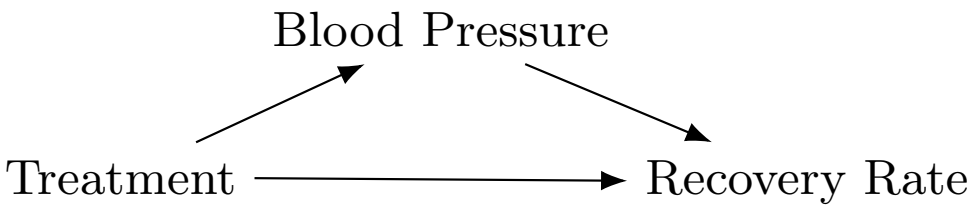


Simpson's paradox - assumptions and data

- Lets start drawing some graphs to represent these different cases.
- The data, i.e., (conditional) distributions, are the same in both cases.
- In case 1, we **assumed** the choice of treatment is influenced by the stone size, i.e.,

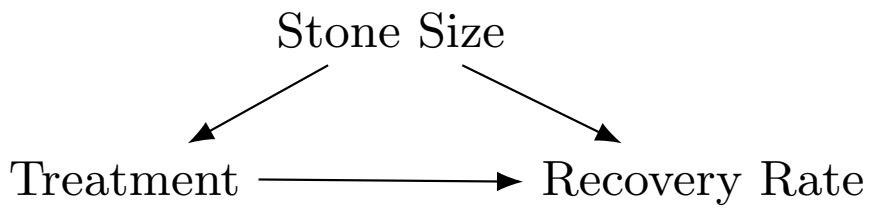


- In case 2, we **assumed** the treatment has influence on the blood pressure, i.e.,

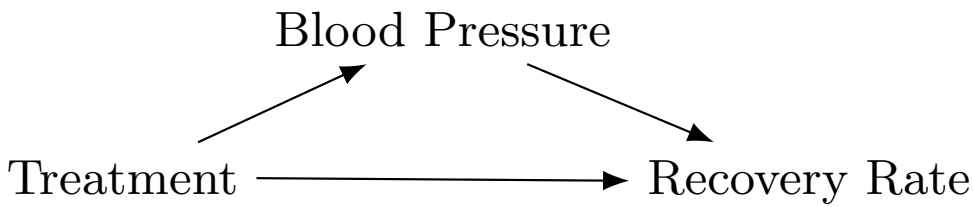


Simpson's paradox - assumptions and data

- Lets start drawing some graphs to represent these different cases.
- The data, i.e., (conditional) distributions, are the same in both cases.
- In case 1, we **assumed** the choice of treatment is influenced by the stone size, i.e.,



- In case 2, we **assumed** the treatment has influence on the blood pressure, i.e.,



- Data is not enough. We need to infer or make assumptions on how data is generated, i.e., we need to figure out what causes what
- To find good interventions/treatments, we need to define the causal effect of a treatment on the outcome of interest

Questions?

Question

Any questions on the motivating examples?

Potential outcomes and causal effects

Question

How to define the causal effect of a treatment T on the outcome of interest Y?

Potential outcomes and causal effects

Question

How to define the causal effect of a treatment T on the outcome of interest Y?

For each unit (patient) u, let

- $Y_0(u)$ be the "potential" outcome had the unit not been treated (control outcome)
- $Y_1(u)$ be the potential outcome had the unit been treated (treated outcome)

Potential outcomes and causal effects

Question

How to define the causal effect of a treatment T on the outcome of interest Y?

For each unit (patient) u, let

- $Y_0(u)$ be the "potential" outcome had the unit not been treated (control outcome)
- $Y_1(u)$ be the potential outcome had the unit been treated (treated outcome)

Individual treatment effect:

$$\text{ITE}(u) := Y_1(u) - Y_0(u)$$

For patient u, T has a causal effect on Y if $\text{ITE}(u) \neq 0$

Potential outcomes and causal effects

Question

How to define the causal effect of a treatment T on the outcome of interest Y?

For each unit (patient) u, let

- $Y_0(u)$ be the "potential" outcome had the unit not been treated (control outcome)
- $Y_1(u)$ be the potential outcome had the unit been treated (treated outcome)

Average treatment effect:

$$\text{ATE} := \mathbb{E}_{u \sim P(u)} \left[Y_1(u) - Y_0(u) \right]$$

Potential outcomes and causal effects

Question

How to define the causal effect of a treatment T on the outcome of interest Y?

For each unit (patient) u, let

- $Y_0(u)$ be the "potential" outcome had the unit not been treated (control outcome)
- $Y_1(u)$ be the potential outcome had the unit been treated (treated outcome)

The fundamental problem of causal inference

We can only ever observe one of the potential outcomes.

If the individual is treated, $T = 1$, we observe $Y_1(u)$ (factual) but $Y_0(u)$ is unknown (counterfactual)

Example - Estimants of interest

Consider the following data table, where X is a patient feature (e.g., severity of the disease) and $Y = 1$ indicates mortality. We'll pretend an oracle gave us the potential outcomes.

id	X	T	Y	Y_0	Y_1
0	0	0	0	0	1
1	0	1	1	0	1
2	0	0	1	1	0
3	0	0	0	0	0
4	0	1	0	0	0
5	1	1	0	1	0
6	1	1	1	1	1
7	1	0	0	0	1
8	1	1	0	1	0
9	1	1	0	0	0

Example - Estimants of interest

Consider the following data table, where X is a patient feature (e.g., severity of the disease) and $Y = 1$ indicates mortality. We'll pretend an oracle gave us the potential outcomes.

id	X	T	Y	Y_0	Y_1	ITE
0	0	0	0	0	1	1
1	0	1	1	0	1	1
2	0	0	1	1	0	-1
3	0	0	0	0	0	0
4	0	1	0	0	0	0
5	1	1	0	1	0	-1
6	1	1	1	1	1	0
7	1	0	0	0	1	1
8	1	1	0	1	0	-1
9	1	1	0	0	0	0

Example - Estimants of interest

Consider the following data table, where X is a patient feature (e.g., severity of the disease) and $Y = 1$ indicates mortality. We'll pretend an oracle gave us the potential outcomes.

id	X	T	Y	Y_0	Y_1	ITE
0	0	0	0	0	1	1
1	0	1	1	0	1	1
2	0	0	1	1	0	-1
3	0	0	0	0	0	0
4	0	1	0	0	0	0
5	1	1	0	1	0	-1
6	1	1	1	1	1	0
7	1	0	0	0	1	1
8	1	1	0	1	0	-1
9	1	1	0	0	0	0

$$\text{ATE} = \frac{4}{10} - \frac{4}{10} = 0$$

Example - Estimants of interest

Consider the following data table, where X is a patient feature (e.g., severity of the disease) and $Y = 1$ indicates mortality. We'll pretend an oracle gave us the potential outcomes.

id	X	T	Y	Y_0	Y_1	ITE
0	0	0	0	0	1	1
1	0	1	1	0	1	1
2	0	0	1	1	0	-1
3	0	0	0	0	0	0
4	0	1	0	0	0	0
5	1	1	0	1	0	-1
6	1	1	1	1	1	0
7	1	0	0	0	1	1
8	1	1	0	1	0	-1
9	1	1	0	0	0	0

$$\text{ATE} = \frac{4}{10} - \frac{4}{10} = 0$$

Conditional average treatment effect $\mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X]$

$$\text{CATE}(X) = \begin{cases} \frac{2}{5} - \frac{1}{5} = \frac{1}{5} & X = 0 \\ \frac{2}{5} - \frac{3}{5} = -\frac{1}{5} & X = 1 \end{cases}$$

Example - Estimants of interest

Consider the following data table, where X is a patient feature (e.g., severity of the disease) and $Y = 1$ indicates mortality. We'll pretend an oracle gave us the potential outcomes.

id	X	T	Y	Y_0	Y_1	ITE
0	0	0	0	0	1	1
1	0	1	1	0	1	1
2	0	0	1	1	0	-1
3	0	0	0	0	0	0
4	0	1	0	0	0	0
5	1	1	0	1	0	-1
6	1	1	1	1	1	0
7	1	0	0	0	1	1
8	1	1	0	1	0	-1
9	1	1	0	0	0	0

$$\text{ATE} = \frac{4}{10} - \frac{4}{10} = 0$$

Conditional average treatment effect
$\mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X]$

$$\text{CATE}(X) = \begin{cases} \frac{2}{5} - \frac{1}{5} = \frac{1}{5} & X = 0 \\ \frac{2}{5} - \frac{3}{5} = -\frac{1}{5} & X = 1 \end{cases}$$

factuals/counterfactuals

Example - Estimants of interest

Consider the following data table, where X is a patient feature (e.g., severity of the disease) and $Y = 1$ indicates mortality. We'll pretend an oracle gave us the potential outcomes.

id	X	T	Y	Y_0	Y_1	ITE
0	0	0	0	0	?	?
1	0	1	1	?	1	?
2	0	0	1	1	?	?
3	0	0	0	0	?	?
4	0	1	0	?	0	?
5	1	1	0	?	0	?
6	1	1	1	?	1	?
7	1	0	0	0	?	?
8	1	1	0	?	0	?
9	1	1	0	?	0	?

$$\text{ATE} = \frac{?}{10} - \frac{?}{10} = ?$$

Conditional average treatment effect
$\mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X]$

$$\text{CATE}(X) = \begin{cases} \frac{?}{5} - \frac{?}{5} = ? & X = 0 \\ \frac{?}{5} - \frac{?}{5} = ? & X = 1 \end{cases}$$

factuals/counterfactuals

Assumptions for causal inference

In our analysis we implicitly used the following two assumptions:

Stable unit treatment value assumption (SUTVA)

- Units do not interfere, i.e., the potential outcome of a unit does not depend on the other patients.
- The factual matches the observed outcome, i.e., $Y_T(u) = Y$ (Consistency)

- Aside: There is a rich literature on **causal inference in network data** that we will not cover in this class.

Association v.s. causation

id	X	T	Y	Y_0	Y_1
0	0	0	0	0	1
1	0	1	1	0	1
2	0	0	1	1	0
3	0	0	0	0	0
4	0	1	0	0	0
5	1	1	0	1	0
6	1	1	1	1	1
7	1	0	0	0	1
8	1	1	0	1	0
9	1	1	0	0	0

Association v.s. causation

id	X	T	Y	Y_0	Y_1
0	0	0	0	0	1
1	0	1	1	0	1
2	0	0	1	1	0
3	0	0	0	0	0
4	0	1	0	0	0
5	1	1	0	1	0
6	1	1	1	1	1
7	1	0	0	0	1
8	1	1	0	1	0
9	1	1	0	0	0

- is T associated to Y?

Association v.s. causation

id	X	T	Y	Y_0	Y_1
0	0	0	0	0	1
1	0	1	1	0	1
2	0	0	1	1	0
3	0	0	0	0	0
4	0	1	0	0	0
5	1	1	0	1	0
6	1	1	1	1	1
7	1	0	0	0	1
8	1	1	0	1	0
9	1	1	0	0	0

- is T associated to Y?

- In population $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] = \frac{2}{6} - \frac{1}{4} = \frac{1}{2}$
- In sub-populations

$$\begin{cases} \mathbb{E}[Y|T=1, X=0] - \mathbb{E}[Y|T=0, X=0] = \frac{1}{2} - \frac{1}{3} = \frac{1}{6} \\ \\ \mathbb{E}[Y|T=1, X=1] - \mathbb{E}[Y|T=0, X=1] = \frac{1}{4} - \frac{0}{1} = \frac{1}{4} \end{cases}$$

- Treatment is associated with more deaths!

Association v.s. causation

id	X	T	Y	Y_0	Y_1
0	0	0	0	0	1
1	0	1	1	0	1
2	0	0	1	1	0
3	0	0	0	0	0
4	0	1	0	0	0
5	1	1	0	1	0
6	1	1	1	1	1
7	1	0	0	0	1
8	1	1	0	1	0
9	1	1	0	0	0

- Does T causes more deaths?

- is T associated to Y?

- In population $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] = \frac{2}{6} - \frac{1}{4} = \frac{1}{2}$
- In sub-populations

$$\begin{cases} \mathbb{E}[Y|T=1, X=0] - \mathbb{E}[Y|T=0, X=0] = \frac{1}{2} - \frac{1}{3} = \frac{1}{6} \\ \\ \mathbb{E}[Y|T=1, X=1] - \mathbb{E}[Y|T=0, X=1] = \frac{1}{4} - \frac{0}{1} = \frac{1}{4} \end{cases}$$

- Treatment is associated with more deaths!

Association v.s. causation

id	X	T	Y	Y_0	Y_1
0	0	0	0	0	1
1	0	1	1	0	1
2	0	0	1	1	0
3	0	0	0	0	0
4	0	1	0	0	0
5	1	1	0	1	0
6	1	1	1	1	1
7	1	0	0	0	1
8	1	1	0	1	0
9	1	1	0	0	0

- Does T causes more deaths?
  - ATE $= 0$
  - CATE$(0) = \frac{1}{5}$, CATE$(1) = -\frac{1}{5}$
  - Treatment helps severe patients

- is T associated to Y?
  - In population $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0] = \frac{2}{6} - \frac{1}{4} = \frac{1}{2}$
  - In sub-populations

$$\begin{cases} \mathbb{E}[Y|T=1, X=0] - \mathbb{E}[Y|T=0, X=0] = \frac{1}{2} - \frac{1}{3} = \frac{1}{6} \\ \\ \mathbb{E}[Y|T=1, X=1] - \mathbb{E}[Y|T=0, X=1] = \frac{1}{4} - \frac{0}{1} = \frac{1}{4} \end{cases}$$

  - Treatment is associated with more deaths!

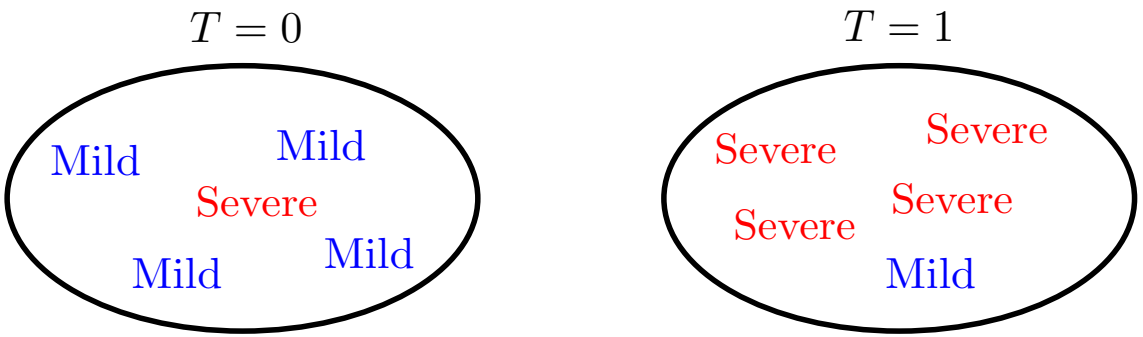
Association v.s. causation

id	X	T	Y	Y_0	Y_1
0	0	0	0	0	1
1	0	1	1	0	1
2	0	0	1	1	0
3	0	0	0	0	0
4	0	1	0	0	0
5	1	1	0	1	0
6	1	1	1	1	1
7	1	0	0	0	1
8	1	1	0	1	0
9	1	1	0	0	0

Couldn't we condition on treatment and use machine learning to predict outcomes? $\mathbb{E}[Y_1 - Y_0] \neq \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$. Why?

Association v.s. causation

id	X	T	Y	Y_0	Y_1
0	0	0	0	0	1
1	0	1	1	0	1
2	0	0	1	1	0
3	0	0	0	0	0
4	0	1	0	0	0
5	1	1	0	1	0
6	1	1	1	1	1
7	1	0	0	0	1
8	1	1	0	1	0
9	1	1	0	0	0

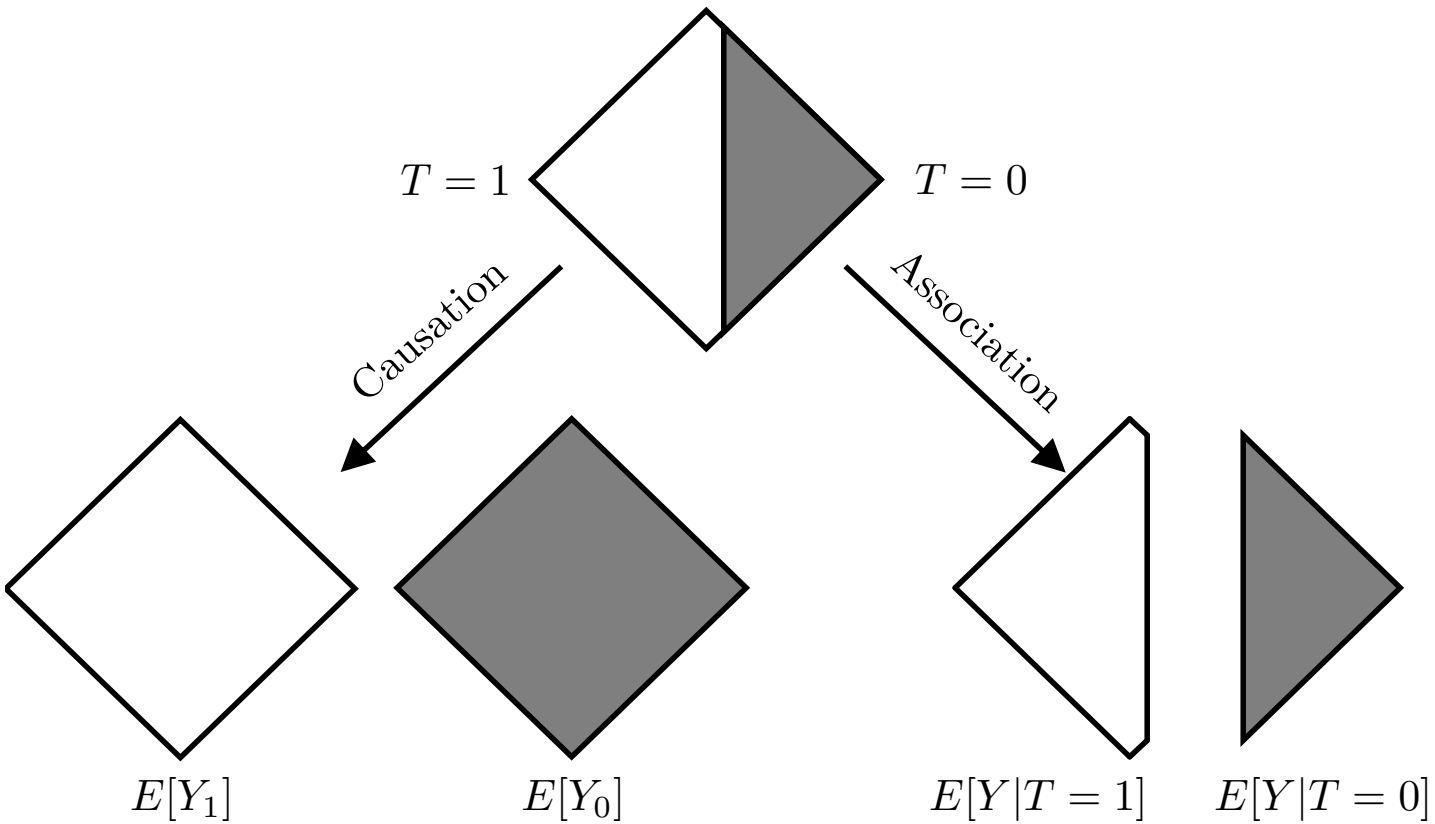


Couldn't we condition on treatment and use machine learning to predict outcomes? $\mathbb{E}[Y_1 - Y_0] \neq \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$. Why?

Treated and untreated populations are not always comparable

For instance, $\mathbb{E}[Y|T = 1]$ is biased towards the outcome of patients with more severe disease

Association v.s. causation



Estimating treatment effects

- We do not observe both Y_0 and Y_1. How to estimate ITE, ATE, or CATE?

Estimating treatment effects

- We do not observe both Y_0 and Y_1. How to estimate ITE, ATE, or CATE?
- ITEs are generally impossible as counterfactuals are unknown

id	X	T	Y	Y_0	Y_1	ITE
0	0	0	0	0	1	1
1	0	1	1	0	1	1
2	0	0	1	1	0	-1
3	0	0	0	0	0	0
4	0	1	0	0	0	0
5	1	1	0	1	0	-1
6	1	1	1	1	1	0
7	1	0	0	0	1	1
8	1	1	0	1	0	-1
9	1	1	0	0	0	0

Estimating treatment effects

- We do not observe both Y_0 and Y_1. How to estimate ITE, ATE, or CATE?
- ITEs are generally impossible as counterfactuals are unknown

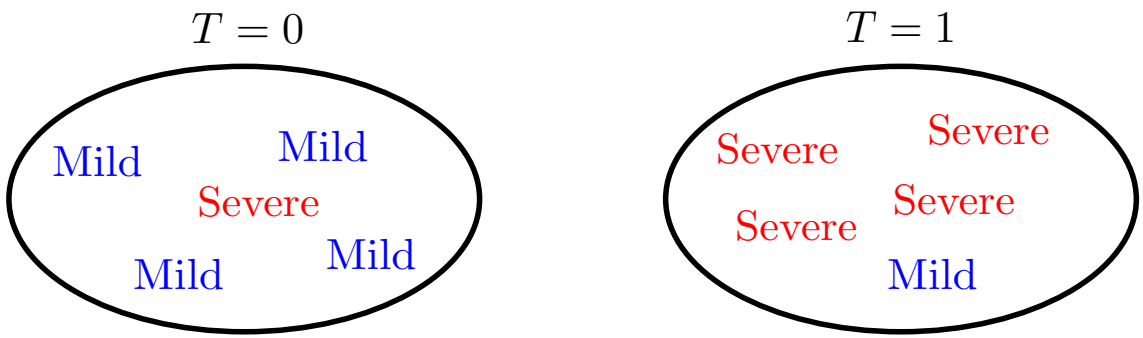
id	X	T	Y	Y_0	Y_1	ITE
0	0	0	0	0	0	0
1	0	1	1	1	1	0
2	0	0	1	1	1	0
3	0	0	0	0	1	1
4	0	1	0	1	0	-1
5	1	1	0	0	0	0
6	1	1	1	0	1	1
7	1	0	0	0	0	0
8	1	1	0	0	0	0
9	1	1	0	1	0	-1

Estimating treatment effects

- Let's focus on the simplest quantity defined on the population, i.e., $\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$

Estimating treatment effects

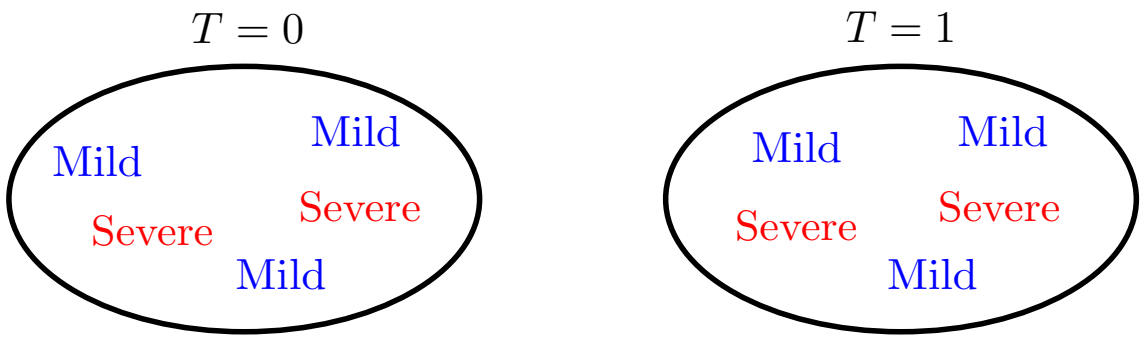
- Let's focus on the simplest quantity defined on the population, i.e., $\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$



- We saw that generally $\mathbb{E}[Y_1] - \mathbb{E}[Y_0] \neq \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$
- But, when is association causation?

Estimating treatment effects

- Let's focus on the simplest quantity defined on the population, i.e., $\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$

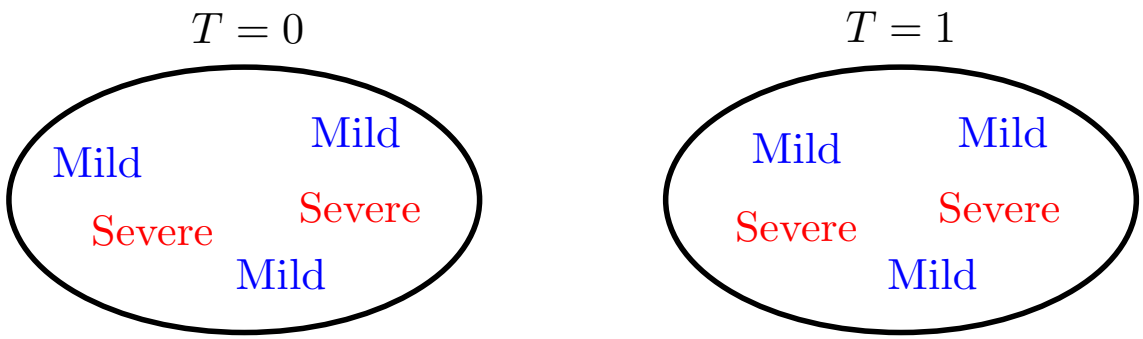


- We saw that generally $\mathbb{E}[Y_1] - \mathbb{E}[Y_0] \neq \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$
- But, when is association causation?
- When the treated and untreated populations are similar, i.e., they have **similar potential outcomes**

$$P(Y_1|T = 1) = P(Y_1|T = 0) \text{ and } P(Y_0|T = 0) = P(Y_0|T = 0)$$

Estimating treatment effects

- Let's focus on the simplest quantity defined on the population, i.e., $\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0]$



- We saw that generally $\mathbb{E}[Y_1] - \mathbb{E}[Y_0] \neq \mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$
- But, when is association causation?
- When the treated and untreated populations are similar, i.e., they have **similar potential outcomes**

$$Y_1, Y_0 \perp\!\!\!\perp T$$

Estimating treatment effects - Ignorability

Ignorability/Exchangeability assumption

$Y_1, Y_0 \perp\!\!\!\perp T$ i.e. the potential outcomes are independent of treatment assignment. **Intuitively:** Knowing the treatment assigned to the patient gives us no information about what the outcome looks like.

Estimating treatment effects - Ignorability

Ignorability/Exchangeability assumption

$Y_1, Y_0 \perp\!\!\!\perp T$ i.e. the potential outcomes are independent of treatment assignment. **Intuitively:** Knowing the treatment assigned to the patient gives us no information about what the outcome looks like.

$$\mathbb{E}[Y_0] = P(T = 1) \cdot \mathbb{E}[Y_0|T = 1] + P(T = 0) \cdot \mathbb{E}[Y_0|T = 0]$$

Estimating treatment effects - Ignorability

Ignorability/Exchangeability assumption

$Y_1, Y_0 \perp\!\!\!\perp T$ i.e. the potential outcomes are independent of treatment assignment. **Intuitively:** Knowing the treatment assigned to the patient gives us no information about what the outcome looks like.

$$\begin{aligned}\mathbb{E}[Y_0] =& P(T=1)\cdot\mathbb{E}[Y_0|T=1] + P(T=0)\cdot\mathbb{E}[Y_0|T=0] \\ =& P(T=1)\cdot\mathbb{E}[Y_0|T=0] + P(T=0)\cdot\mathbb{E}[Y_0|T=0] \qquad \text{(ignorability)}\end{aligned}$$

Estimating treatment effects - Ignorability

Ignorability/Exchangeability assumption

$Y_1, Y_0 \perp\!\!\!\perp T$ i.e. the potential outcomes are independent of treatment assignment. **Intuitively:** Knowing the treatment assigned to the patient gives us no information about what the outcome looks like.

$$
\begin{aligned}
\mathbb{E}[Y_0] =& P(T=1)\cdot\mathbb{E}[Y_0|T=1] + P(T=0)\cdot\mathbb{E}[Y_0|T=0] \\
=& P(T=1)\cdot\mathbb{E}[Y_0|T=0] + P(T=0)\cdot\mathbb{E}[Y_0|T=0] && \text{(ignorability)} \\
=& P(T=1)\cdot\mathbb{E}[Y|T=0] + P(T=0)\cdot\mathbb{E}[Y|T=0] && \text{(consistency)} \\
=& \mathbb{E}[Y|T=0]
\end{aligned}
$$

Estimating treatment effects - Ignorability

Ignorability/Exchangeability assumption

$Y_1, Y_0 \perp\!\!\!\perp T$ i.e. the potential outcomes are independent of treatment assignment. **Intuitively:** Knowing the treatment assigned to the patient gives us no information about what the outcome looks like.

$$\begin{aligned}
\mathbb{E}[Y_0] =& P(T=1)\cdot\mathbb{E}[Y_0|T=1] + P(T=0)\cdot\mathbb{E}[Y_0|T=0] \\
=& P(T=1)\cdot\mathbb{E}[Y_0|T=0] + P(T=0)\cdot\mathbb{E}[Y_0|T=0] && \text{(ignorability)} \\
=& P(T=1)\cdot\mathbb{E}[Y|T=0] + P(T=0)\cdot\mathbb{E}[Y|T=0] && \text{(consistency)} \\
=& \mathbb{E}[Y|T=0]
\end{aligned}$$

Hence, we can estimate ATE under the ignorability and consistency assumptions

$$\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$$

Estimating treatment effects - Ignorability

Ignorability/Exchangeability assumption

$Y_1, Y_0 \perp\!\!\!\perp T$ i.e. the potential outcomes are independent of treatment assignment. **Intuitively:** Knowing the treatment assigned to the patient gives us no information about what the outcome looks like.

$$\begin{aligned}
\mathbb{E}[Y_0] =& P(T=1)\cdot\mathbb{E}[Y_0|T=1] + P(T=0)\cdot\mathbb{E}[Y_0|T=0] \\
=& P(T=1)\cdot\mathbb{E}[Y_0|T=0] + P(T=0)\cdot\mathbb{E}[Y_0|T=0] && \text{(ignorability)} \\
=& P(T=1)\cdot\mathbb{E}[Y|T=0] + P(T=0)\cdot\mathbb{E}[Y|T=0] && \text{(consistency)} \\
=& \mathbb{E}[Y|T=0]
\end{aligned}$$

Hence, we can estimate ATE under the ignorability and consistency assumptions

$$\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$$

Ignorability is also called *exchangeability*. Since we can exchange the treated and untreated population:

$$Y_0 \perp\!\!\!\perp T \implies \mathbb{E}[Y_0|T=1] = \mathbb{E}[Y_0] = \mathbb{E}[Y_0|T=0]$$

Randomized controlled trials

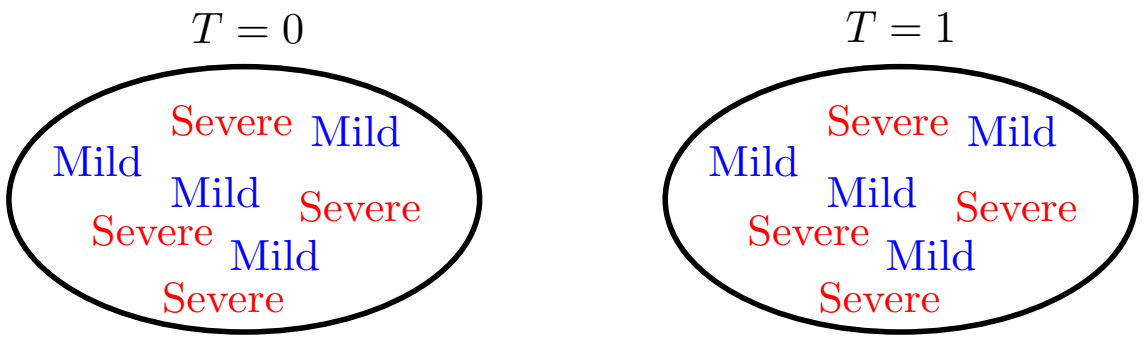
Where can we make the ignorability assumption? i.e., $Y_1, Y_0 \perp\!\!\!\perp T$

- We have no control on potential outcomes Y_1, Y_0. But we can control the treatment assignment

Randomized controlled trials

Where can we make the ignorability assumption? i.e., $Y_1, Y_0 \perp\!\!\!\perp T$

- We have no control on potential outcomes Y_1, Y_0. But we can control the treatment assignment
- *Randomized controlled trials (RCTs)*: Flip a coin to put participants in treated or untreated groups



$$\forall y_0, y_1 : \quad P(T = 1 | Y_0 = y_0, Y_1 = y_1) = c \implies Y_0, Y_1 \perp\!\!\!\perp T$$

Observational data

RCTs are gold-standard to study causal effects but not always feasible

- They can be unethical, e.g., causal effect of smoking on lung cancer
- They are costly with a small number of participants. So, they often cannot capture the heterogeneity of the population
- Participants are not necessarily representative of the whole population
- ...

Observational data

RCTs are gold-standard to study causal effects but not always feasible

- They can be unethical, e.g., causal effect of smoking on lung cancer
- They are costly with a small number of participants. So, they often cannot capture the heterogeneity of the population
- Participants are not necessarily representative of the whole population
- ...

What about millions of *observational data* points that are not RCT?

- In healthcare (EHR data), patients are often treated based on their symptoms
- Mild heart problem gets regular exercise while stage D heart failure gets heart transplant
- $P(Y_{\text{exercise}}|T = \text{exercise}) < P(Y_{\text{exercise}}|T = \text{heart surgery})$
- Therefore, $Y_1, Y_0 \not\perp\!\!\!\perp T$

Estimating treatment effects - Conditional ignorability

What is the effect of heart transplant in patients with heart failure?

Estimating treatment effects - Conditional ignorability

What is the effect of heart transplant in patients with heart failure?

- Ignorability does not hold. Patients with more severe symptoms are more likely to get transplant

Estimating treatment effects - Conditional ignorability

What is the effect of heart transplant in patients with heart failure?

- Ignorability does not hold. Patients with more severe symptoms are more likely to get transplant
- However, within patients with similar symptoms, hearts are assigned to the ones with compatible HLA[1] genes, which we "believe" is independent of mortality

[1] human leukocyte antigen

Estimating treatment effects - Conditional ignorability

What is the effect of heart transplant in patients with heart failure?

- Ignorability does not hold. Patients with more severe symptoms are more likely to get transplant
- However, within patients with similar symptoms, hearts are assigned to the ones with compatible HLA[1] genes, which we "believe" is independent of mortality
- In other words, $Y_1, Y_0 \perp\!\!\!\perp T|X$, where X is the severity of symptoms

[1] human leukocyte antigen

Estimating treatment effects - Conditional ignorability

What is the effect of heart transplant in patients with heart failure?

- Ignorability does not hold. Patients with more severe symptoms are more likely to get transplant
- However, within patients with similar symptoms, hearts are assigned to the ones with compatible HLA[1] genes, which we "believe" is independent of mortality
- In other words, $Y_1, Y_0 \perp\!\!\!\perp T|X$, where X is the severity of symptoms: *conditional ignorability*

[1] human leukocyte antigen

Estimating treatment effects - Conditional ignorability

Conditional Ignorability assumption

$$Y_1, Y_0 \perp\!\!\!\perp T | X$$

Estimating treatment effects - Conditional ignorability

Conditional Ignorability assumption

$$Y_1, Y_0 \perp\!\!\!\perp T|X$$

$$\begin{aligned}\mathbb{E}[Y_0] &= \mathbb{E}_X\left[\mathbb{E}[Y_0|X]\right] \\ &= \mathbb{E}_X\left[\mathbb{E}[Y_0|X, T=1] \cdot P(T=1|X) + \mathbb{E}[Y_0|X, T=0] \cdot P(T=0|X)\right]\end{aligned}$$

Estimating treatment effects - Conditional ignorability

Conditional Ignorability assumption

$$Y_1, Y_0 \perp\!\!\!\perp T|X$$

$$\begin{aligned}
\mathbb{E}[Y_0] &= \mathbb{E}_X\left[\mathbb{E}[Y_0|X]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X, T=1] \cdot P(T=1|X) + \mathbb{E}[Y_0|X, T=0] \cdot P(T=0|X)\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X, T=0] \cdot P(T=1|X) + \mathbb{E}[Y_0|X, T=0] \cdot P(T=0|X)\right] \\
&\qquad\qquad \text{(conditional ignorability)}
\end{aligned}$$

Estimating treatment effects - Conditional ignorability

Conditional Ignorability assumption

$$Y_1, Y_0 \perp\!\!\!\perp T|X$$

$$\begin{aligned}
\mathbb{E}[Y_0] &= \mathbb{E}_X\left[\mathbb{E}[Y_0|X]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X,T=1]\cdot P(T=1|X) + \mathbb{E}[Y_0|X,T=0]\cdot P(T=0|X)\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X,T=0]\cdot P(T=1|X) + \mathbb{E}[Y_0|X,T=0]\cdot P(T=0|X)\right] \\
&\qquad\qquad \text{(conditional ignorability)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\cdot P(T=1|X) + \mathbb{E}[Y|X,T=0]\cdot P(T=0|X)\right] \\
&\qquad\qquad \text{(consistency)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\right]
\end{aligned}$$

Estimating treatment effects - Conditional ignorability

Conditional Ignorability assumption

$$Y_1, Y_0 \perp\!\!\!\perp T|X$$

$$\begin{aligned}
\mathbb{E}[Y_0] &= \mathbb{E}_X\left[\mathbb{E}[Y_0|X]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X,T=1]\cdot P(T=1|X) + \mathbb{E}[Y_0|X,T=0]\cdot P(T=0|X)\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X,T=0]\cdot P(T=1|X) + \mathbb{E}[Y_0|X,T=0]\cdot P(T=0|X)\right] \\
&\qquad\qquad \text{(conditional ignorability)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\cdot P(T=1|X) + \mathbb{E}[Y|X,T=0]\cdot P(T=0|X)\right] \\
&\qquad\qquad \text{(consistency)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\right]
\end{aligned}$$

- Adjustment formula (G-formula):
  $\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}_X\left[\mathbb{E}\left[Y|X,T=1\right]\right] - \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\right]$

Estimating treatment effects - Conditional ignorability

Conditional Ignorability assumption

$$Y_1, Y_0 \perp\!\!\!\perp T|X$$

$$\begin{aligned}
\mathbb{E}[Y_0] &= \mathbb{E}_X\left[\mathbb{E}[Y_0|X]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X,T=1]\cdot P(T=1|X) + \mathbb{E}[Y_0|X,T=0]\cdot P(T=0|X)\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X,T=0]\cdot P(T=1|X) + \mathbb{E}[Y_0|X,T=0]\cdot P(T=0|X)\right] \\
&\qquad\qquad \text{(conditional ignorability)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\cdot P(T=1|X) + \mathbb{E}[Y|X,T=0]\cdot P(T=0|X)\right] \\
&\qquad\qquad \text{(consistency)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\right]
\end{aligned}$$

- Adjustment formula (G-formula):
  $\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}_X\left[\mathbb{E}\left[Y|X,T=1\right]\right] - \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\right]$
- X is called sufficient (valid) adjustment set

Estimating treatment effects - Conditional ignorability

Conditional Ignorability assumption

$$Y_1, Y_0 \perp\!\!\!\perp T|X$$

$$\begin{aligned}
\mathbb{E}[Y_0] &= \mathbb{E}_X\left[\mathbb{E}[Y_0|X]\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X,T=1]\cdot P(T=1|X) + \mathbb{E}[Y_0|X,T=0]\cdot P(T=0|X)\right] \\
&= \mathbb{E}_X\left[\mathbb{E}[Y_0|X,T=0]\cdot P(T=1|X) + \mathbb{E}[Y_0|X,T=0]\cdot P(T=0|X)\right] \\
&\qquad\qquad \text{(conditional ignorability)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\cdot P(T=1|X) + \mathbb{E}[Y|X,T=0]\cdot P(T=0|X)\right] \\
&\qquad\qquad \text{(consistency)} \\
&= \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\right]
\end{aligned}$$

- Adjustment formula (G-formula):
  $\text{ATE} = \mathbb{E}[Y_1] - \mathbb{E}[Y_0] = \mathbb{E}_X\left[\mathbb{E}\left[Y|X,T=1\right]\right] - \mathbb{E}_X\left[\mathbb{E}[Y|X,T=0]\right]$
- X is called sufficient (valid) adjustment set
- Conditional ignorability (unconfoundedness) is an **untestable** assumption. Can never guarantee $Y_0, Y_1 \perp\!\!\!\perp T|X$ for a non-random T

Estimating treatment effects - Positivity

- G-formula:

$$\text{ATE} = \mathbb{E}_X \left[\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0] \right]$$

- How to estimate ATE given a dataset $\{(x_i, t_i, y_i)_{i=1}^N\}$?

Estimating treatment effects - Positivity

- G-formula:

$$\text{ATE} = \mathbb{E}_X \left[\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0] \right]$$

- How to estimate ATE given a dataset $\{(x_i, t_i, y_i)_{i=1}^N\}$?

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{x_i} \mathbb{E}[Y|X = x_i, T = 1] - \mathbb{E}[Y|X = x_i, T = 0]$$

Estimating treatment effects - Positivity

- G-formula:

$$\text{ATE} = \mathbb{E}_X \left[\mathbb{E}[Y|X, T = 1] - \mathbb{E}[Y|X, T = 0]\right]$$

- How to estimate ATE given a dataset $\{(x_i, t_i, y_i)_{i=1}^N\}$?

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{x_i} \mathbb{E}[Y|X = x_i, T = 1] - \mathbb{E}[Y|X = x_i, T = 0]$$

- To estimate both $\mathbb{E}[Y|X = x_i, T = 0]$ and $\mathbb{E}[Y|X = x_i, T = 1]$, we need a *positive* probability of getting treatment and control

$$\mathbb{E}[Y|X = x_i, T = 0] = \sum_y y \cdot \frac{P(Y = y, X = x_i, T = 0)}{P(X = x_i)P(T = 0|X = x_i)}$$

Estimating treatment effects - Positivity

- G-formula:

$$\text{ATE} = \mathbb{E}_X \left[\mathbb{E}[Y|X, T=1] - \mathbb{E}[Y|X, T=0]\right]$$

- How to estimate ATE given a dataset $\{(x_i, t_i, y_i)_{i=1}^N\}$?

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{x_i} \mathbb{E}[Y|X = x_i, T=1] - \mathbb{E}[Y|X = x_i, T=0]$$

- To estimate both $\mathbb{E}[Y|X = x_i, T=0]$ and $\mathbb{E}[Y|X = x_i, T=1]$, we need a *positive* probability of getting treatment and control

$$\mathbb{E}[Y|X = x_i, T=1] = \sum_y y \cdot \frac{P(Y=y, X=x_i, T=1)}{P(X=x_i)P(T=1|X=x_i)}$$

Estimating treatment effects - Positivity

- G-formula:

$$\text{ATE} = \mathbb{E}_X \left[\mathbb{E}[Y|X, T=1] - \mathbb{E}[Y|X, T=0]\right]$$

- How to estimate ATE given a dataset $\{(x_i, t_i, y_i)_{i=1}^N\}$?

$$\widehat{\text{ATE}} = \frac{1}{N} \sum_{x_i} \mathbb{E}[Y|X=x_i, T=1] - \mathbb{E}[Y|X=x_i, T=0]$$

- To estimate both $\mathbb{E}[Y|X=x_i, T=0]$ and $\mathbb{E}[Y|X=x_i, T=1]$, we need a *positive* probability of getting treatment and control

$$\mathbb{E}[Y|X=x_i, T=1] = \sum_y y \cdot \frac{P(Y=y, X=x_i, T=1)}{P(X=x_i)P(T=1|X=x_i)}$$

Positivity assumption

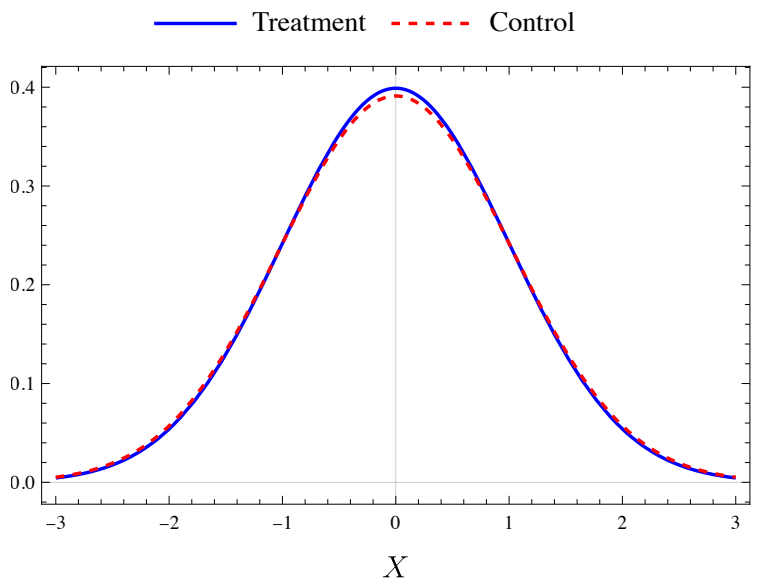
$\forall x$ with $P(x) > 0, \quad 0 < P(T=1|X=x) < 1$

Positivity (Overlap)

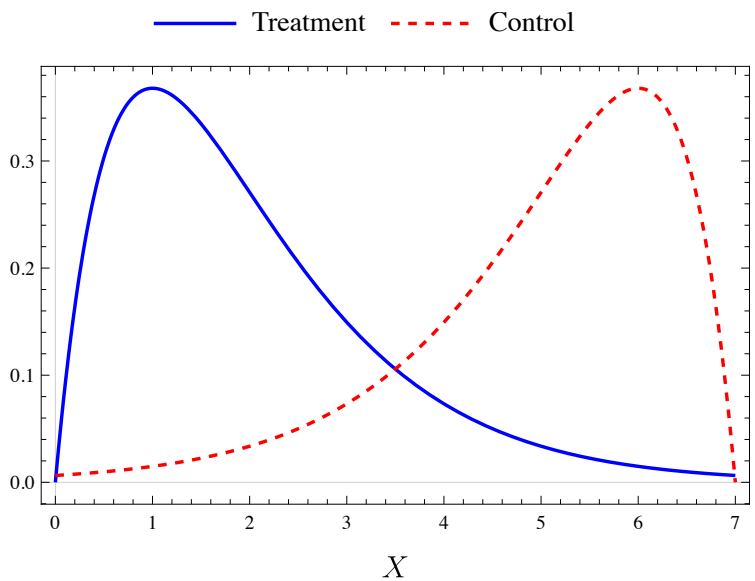
- Treatment group: $P(X|T=1)$, Control group: $P(X|T=0)$

Positivity (Overlap)

- Treatment group: $P(X|T=1)$, Control group: $P(X|T=0)$
- Positivity holds iff the support of treatment and control groups completely overlap



(a) RCT - complete overlap



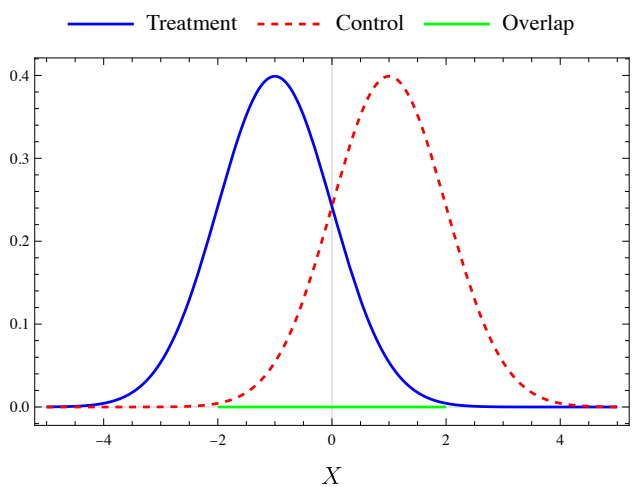
(b) Observational - complete overlap

Positivity-Unconfoundedness trade off

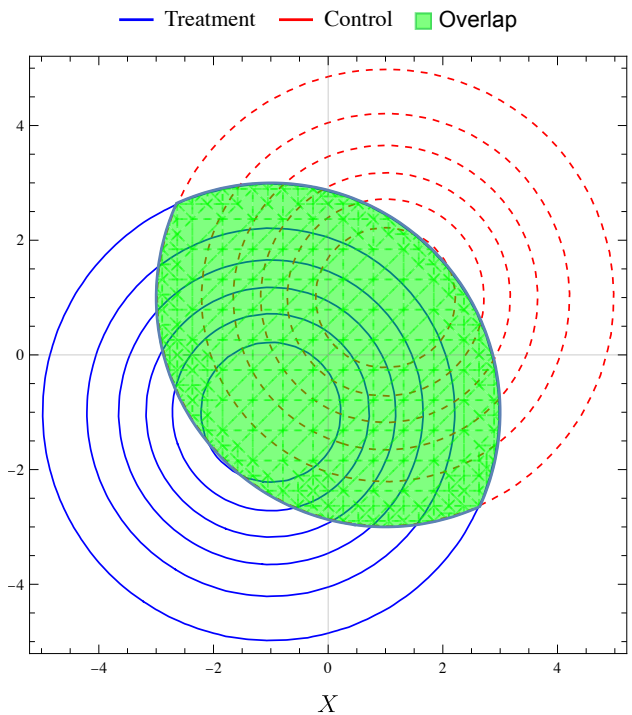
- Unconfoundedness is more plausible when more covariates are included in the analysis
- More information on treatment assignment (larger dimension d) $\rightarrow$ $Y_0, Y_1 \perp\!\!\!\perp T | X_{1:d}$

Positivity-Unconfoundedness trade off

- Unconfoundedness is more plausible when more covariates are included in the analysis
- More information on treatment assignment (larger dimension d) $\rightarrow$ $Y_0, Y_1 \perp\!\!\!\perp T | X_{1:d}$
- But, overlap condition is more difficult to satisfy



(a) $\approx \frac{2}{3}$ overlap in 1-dim



(b) $\approx (\frac{2}{3})^2$ overlap in 2-dim

Positivity-Unconfoundedness trade off

Theorem - Corollary 3 in D'Amour et al., 2021

Let $(X_k)_{k>0}$ be a sequence of covariates, and for each d, let $X_{1:d}$ be a finite subset of $(X_k)_{k>0}$. Also, let P_1 be the distribution of treatment group, i.e., $P_1(A) = P(A|T = 1)$ and P_0 denote the control group distribution. As d grows large, the (strict) positivity assumption implies

$$\frac{1}{d}\sum_{k=1}^{d}\mathbb{E}_{P_1}\left[KL(P_1(X_k|X_{1:k-1}||P_0(X_k|X_{1:k-1}))\right] = O(d^{-1})$$

With high-dimensional covariates, the positivity assumption requires the average conditional distributions of treatment and control group to be close $\approx$ RCTs

Questions?

Question

Any questions on potential outcomes?

Modeling the joint distribution

$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

Modeling the joint distribution

$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | x_{i-1}, \ldots, x_1)$$

Modeling the joint distribution

$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | x_{i-1}, \ldots, x_1)$$

pa_i: a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i | x_{i-1}, \ldots, x_1) = P(x_i | pa_i)$

Modeling the joint distribution

$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

pa_i: a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i | x_{i-1}, \ldots, x_1) = P(x_i | pa_i)$

Modeling the joint distribution

$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

pa_i: a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i | x_{i-1}, \ldots, x_1) = P(x_i | pa_i)$

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_2, x_1)P(x_4|x_3, x_2, x_1)$$

Modeling the joint distribution

$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

pa_i: a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i | x_{i-1}, \ldots, x_1) = P(x_i | pa_i)$

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_3, x_2)$$

(Bayesian network factorization)

Needs $2^0 + 2^1 + 2^1 + 2^2 = 9$ parameters $< 2^4 - 1 = 15$

Modeling the joint distribution

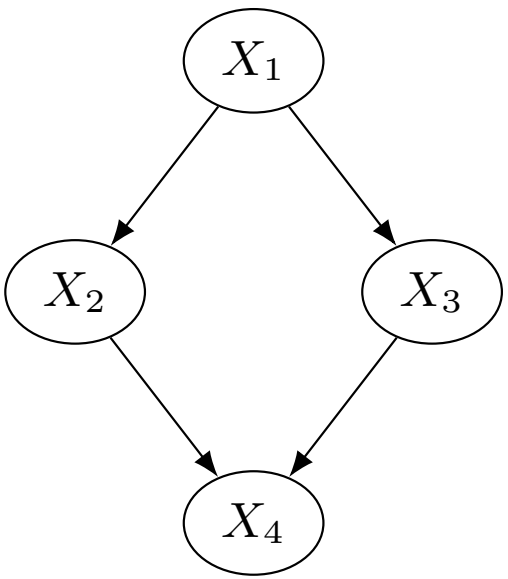
$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

pa_i: a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i | x_{i-1}, \ldots, x_1) = P(x_i | pa_i)$

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_3, x_2)$$

(Bayesian network factorization)



- Directed Acyclic Graph (DAG) $\mathcal{G}$ (Bayesian network)

Modeling the joint distribution

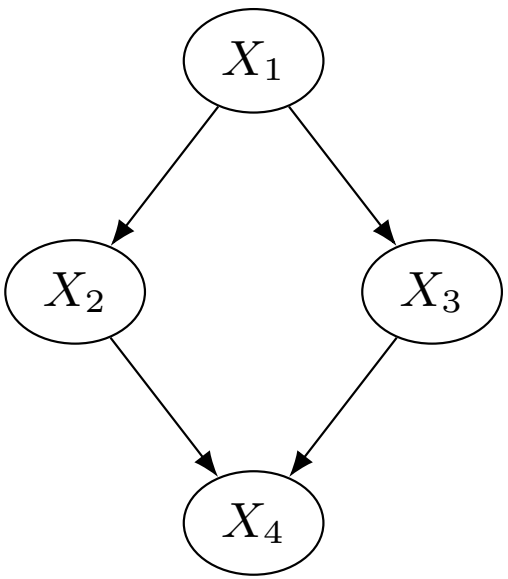
$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

pa_i: a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i | x_{i-1}, \ldots, x_1) = P(x_i | pa_i)$

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_3, x_2)$$

(Bayesian network factorization)



- Directed Acyclic Graph (DAG) $\mathcal{G}$ (Bayesian network)
- P is Markov compatible with $\mathcal{G}$

Modeling the joint distribution

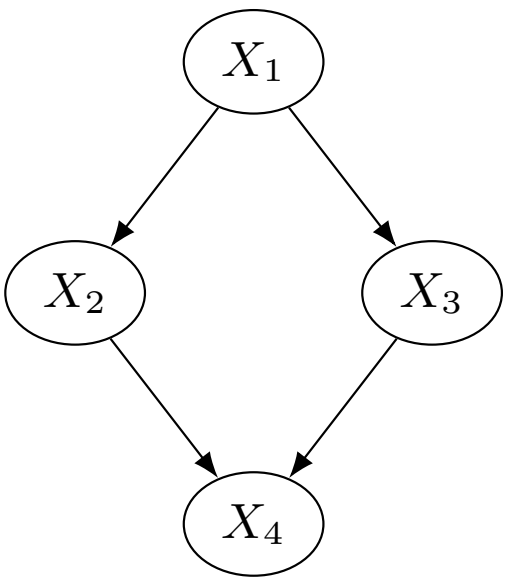
$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

pa_i: a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i | x_{i-1}, \ldots, x_1) = P(x_i | pa_i)$

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_3, x_2)$$

(Bayesian network factorization)



- Directed Acyclic Graph (DAG) $\mathcal{G}$ (Bayesian network)
- P is Markov compatible with $\mathcal{G}$
- $\mathcal{G}$ describes the conditional independence (CI) structure of P

Modeling the joint distribution

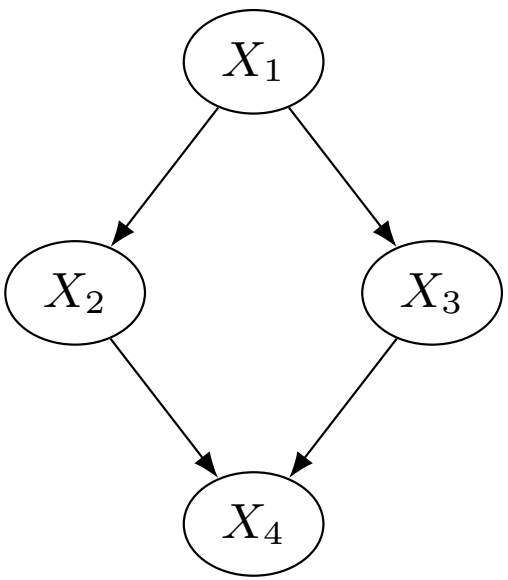
$P(x_1, x_2, \ldots, x_n)$ needs $2^n - 1$ parameters to store for binary x_i

$$P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | pa_i)$$

pa_i: a minimal subset of $\{x_1, \ldots, x_{i-1}\}$ that $P(x_i | x_{i-1}, \ldots, x_1) = P(x_i | pa_i)$

$$P(x_1, x_2, x_3, x_4) = P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_3, x_2)$$

(Bayesian network factorization)



- Directed Acyclic Graph (DAG) $\mathcal{G}$ (Bayesian network)
- P is Markov compatible with $\mathcal{G}$
- $\mathcal{G}$ describes the conditional independence (CI) structure of P
- Given $\mathcal{G}$, how to find the CI set of compatible distributions?

Conditional independence in DAGs

Question

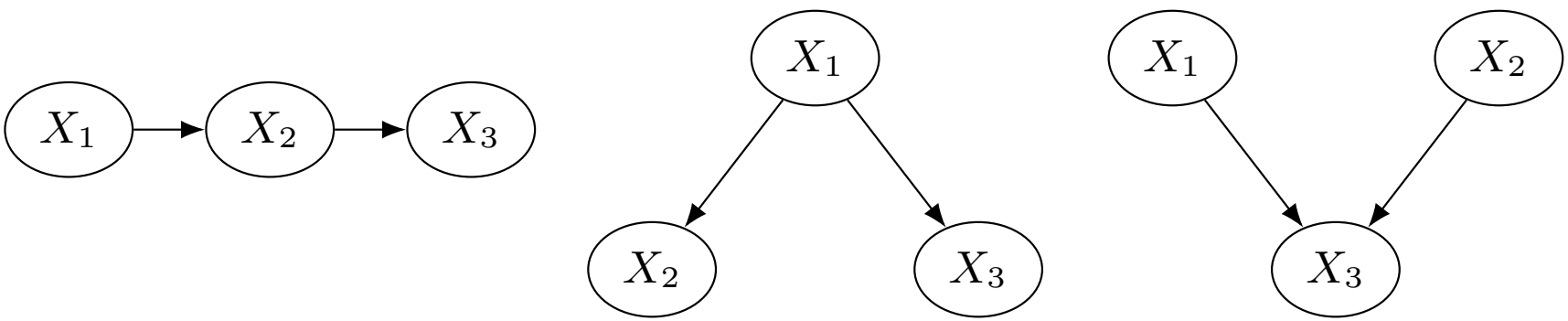
Given a DAG $\mathcal{G}$, how to find the CI set of compatible distributions?

Conditional independence in DAGs

Question

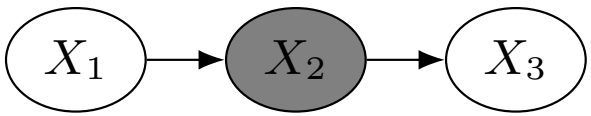
Given a DAG $\mathcal{G}$, how to find the CI set of compatible distributions?

We first consider the building blocks of DAGs



Chain Fork Collider (v-structure)

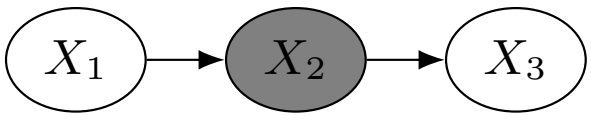
Conditional independence - Chain and v-structure



$$
\begin{aligned}
&P(x_1, x_3|x_2) \\
&= \frac{P(x_1, x_2, x_3)}{P(x_2)} \\
&= \frac{P(x_1)P(x_2|x_1)P(x_3|x_2)}{P(x_2)} \\
&= P(x_1|x_2)P(x_3|x_2) \quad \text{(Bayes rule)}
\end{aligned}
$$

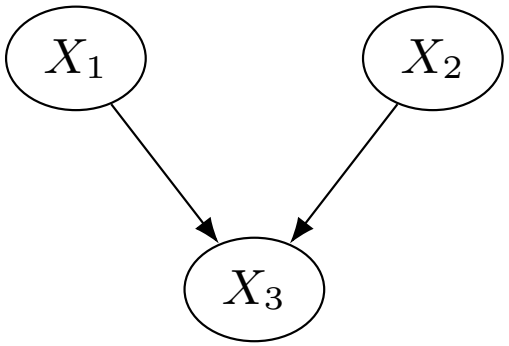
$$X_1 \perp\!\!\!\perp X_3|X_2$$

Conditional independence - Chain and v-structure



$$
\begin{aligned}
&P(x_1, x_3|x_2) \\
&= \frac{P(x_1, x_2, x_3)}{P(x_2)} \\
&= \frac{P(x_1)P(x_2|x_1)P(x_3|x_2)}{P(x_2)} \\
&= P(x_1|x_2)P(x_3|x_2) \quad \text{(Bayes rule)}
\end{aligned}
$$

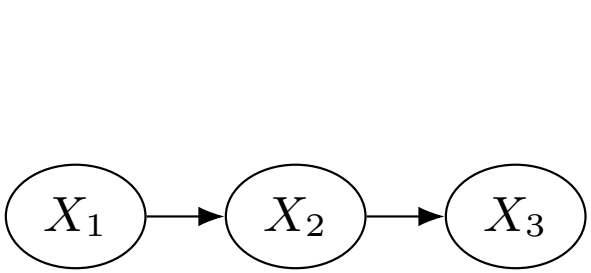
$$X_1 \perp\!\!\!\perp X_3|X_2$$

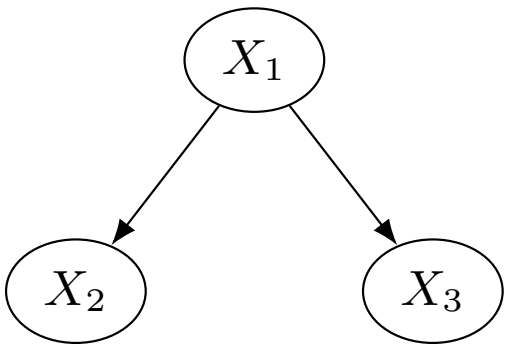


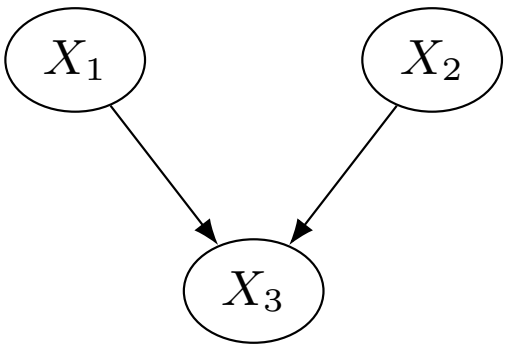
$$
\begin{aligned}
P(x_1, x_2) &= \sum_{x_3} P(x_1, x_2, x_3) \\
&= \sum_{x_3} P(x_1)P(x_2)P(x_3|x_1, x_2) \\
&= P(x_1)P(x_2)\sum_{x_3} P(x_3|x_1, x_2) \\
&= P(x_1)P(x_2)
\end{aligned}
$$

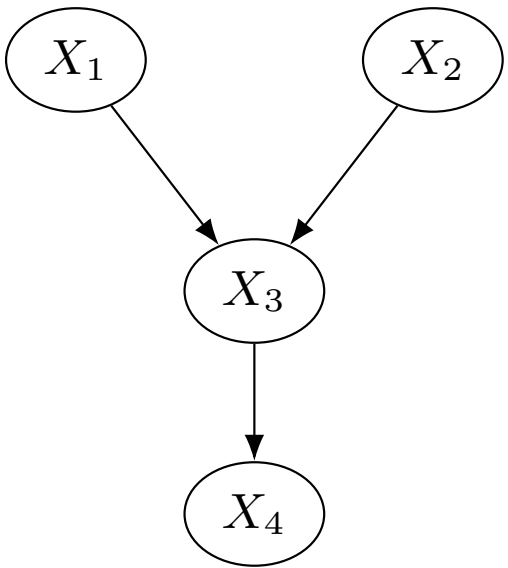
$$X_1 \perp\!\!\!\perp X_3$$

Conditional independence - Building blocks

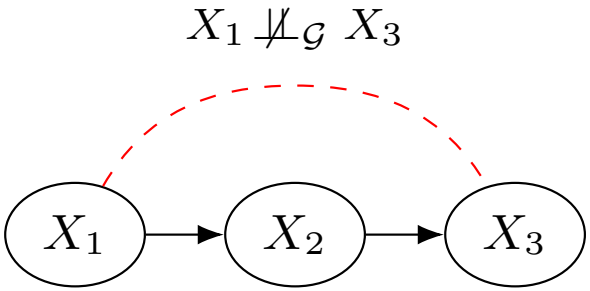


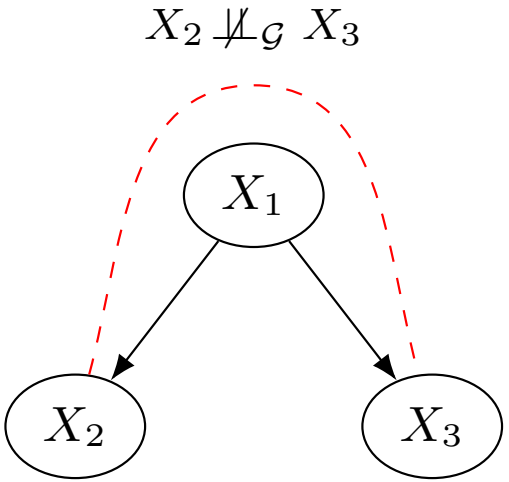


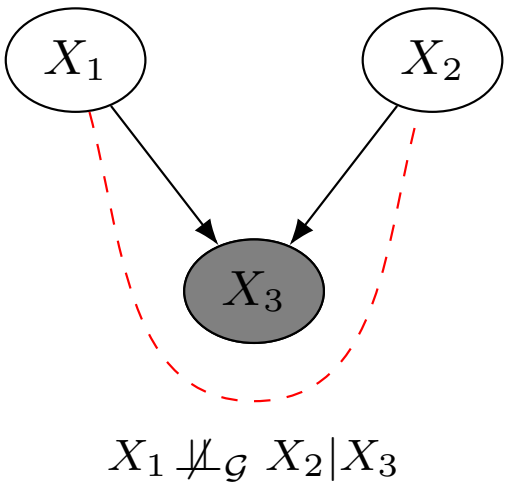


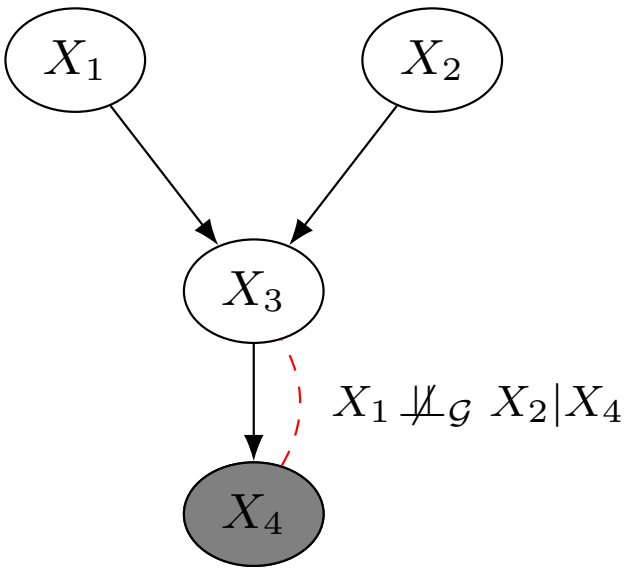


Conditional independence - Unblocked paths

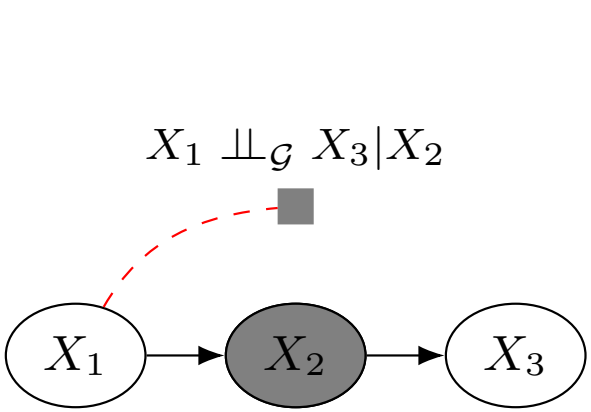


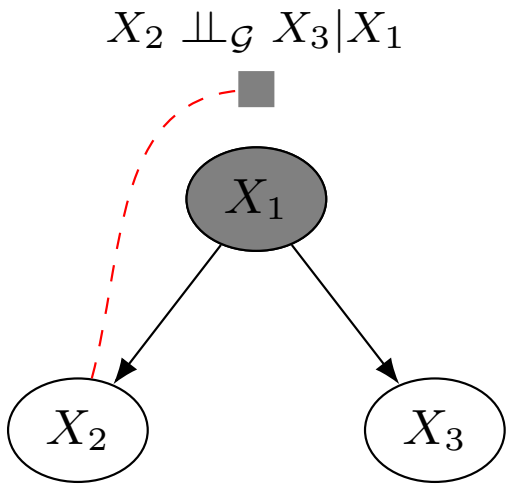


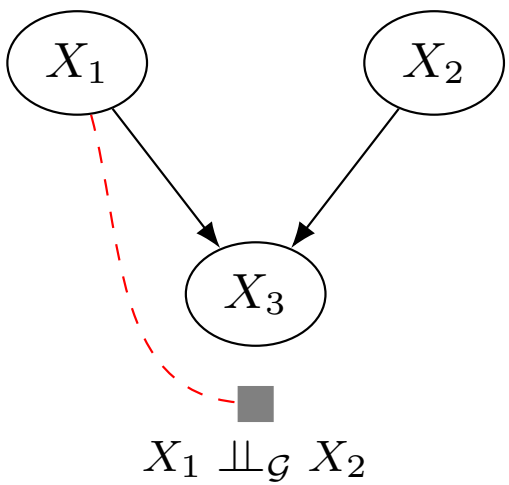


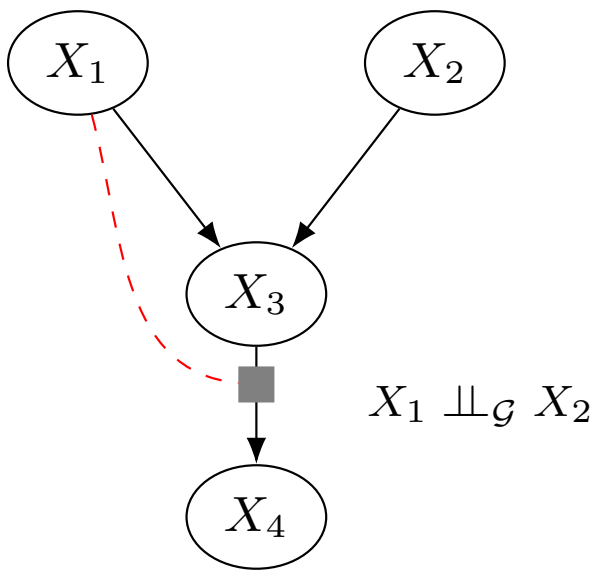


Conditional independence - Blocked paths









D-separation

Blocked path

Given a DAG $\mathcal{G}$, a (undirected) path between nodes X and Y is blocked by a set Z iff

- There is a **chain** $U \to W \to V$ or a **fork** $U \leftarrow W \to V$ on the path, where $W \in Z$, *or*

D-separation

Blocked path

Given a DAG $\mathcal{G}$, a (undirected) path between nodes X and Y is blocked by a set Z iff

- There is a **chain** $U \to W \to V$ or a **fork** $U \leftarrow W \to V$ on the path, where $W \in Z$, *or*
- There is a **collider** $U \to W \leftarrow V$ on the path, where $W \notin Z$ and $Desc(W) \notin Z$

D-separation

Blocked path

Given a DAG $\mathcal{G}$, a (undirected) path between nodes X and Y is blocked by a set Z iff

- There is a **chain** $U \to W \to V$ or a **fork** $U \leftarrow W \to V$ on the path, where $W \in Z$, *or*
- There is a **collider** $U \to W \leftarrow V$ on the path, where $W \notin Z$ and $Desc(W) \notin Z$

D-separation ($X \perp\!\!\!\perp_{\mathcal{G}} Y|Z$)

Given a DAG $\mathcal{G}$, two sets of nodes X and Y are d-separated by a set Z iff all the paths between nodes of X and Y are blocked by Z

Global and local Markov properties

Question

Given $\mathcal{G}$, how to find the CI set of compatible distributions?

Global and local Markov properties

Question

Given $\mathcal{G}$, how to find the CI set of compatible distributions?

Global Markov property

A distribution P satisfies the *global Markov property* w.r.t. a DAG $\mathcal{G}$ if $X \perp\!\!\!\perp_{\mathcal{G}} Y|Z \implies X \perp\!\!\!\perp Y|Z$ for all disjoint sets of nodes X, Y, Z.

Global and local Markov properties

Question

Given $\mathcal{G}$, how to find the CI set of compatible distributions?

Global Markov property

A distribution P satisfies the *global Markov property* w.r.t. a DAG $\mathcal{G}$ if $X \perp\!\!\!\perp_{\mathcal{G}} Y|Z \implies X \perp\!\!\!\perp Y|Z$ for all disjoint sets of nodes X, Y, Z.

Local Markov property

A distribution P satisfies the *local Markov property* w.r.t. a DAG $\mathcal{G}$ if each variable is independent of its nondescendants (in $\mathcal{G}$) conditioned on its parents.

Global and local Markov properties

Question

Given $\mathcal{G}$, how to find the CI set of compatible distributions?

Theorem - Equivalence of Markov properties

Given a distribution P and a DAG $\mathcal{G}$, if P has a density function, then the followings are equivalent

1. P is Markov compatible w.r.t. $\mathcal{G}$
2. P satisfies the global Markov property w.r.t. $\mathcal{G}$
3. P satisfies the local Markov property w.r.t. $\mathcal{G}$

In Markov Random Fields, these properties are shown by the Hammersley-Clifford Theorem.

Observational equivalence

Question

Markov properties relate graphical separation to conditional independencies. Is it possible to have multiple graphs with the same CI structure?

Observational equivalence

Question

Markov properties relate graphical separation to conditional independencies. Is it possible to have multiple graphs with the same CI structure?

Markov equivalence of graphs

Let $\mathcal{M}(\mathcal{G}) := \{P; P \text{ is Markov compatible with } \mathcal{G}\}$. Then, $\mathcal{G}_1$ and $\mathcal{G}_2$ are called Markov equivalent if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$.

Observational equivalence

Question

Markov properties relate graphical separation to conditional independencies. Is it possible to have multiple graphs with the same CI structure?

Markov equivalence of graphs

Let $\mathcal{M}(\mathcal{G}) := \{P; P \text{ is Markov compatible with } \mathcal{G}\}$. Then, $\mathcal{G}_1$ and $\mathcal{G}_2$ are called Markov equivalent if $\mathcal{M}(\mathcal{G}_1) = \mathcal{M}(\mathcal{G}_2)$.

Theorem - Observational equivalence

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent iff they have the same skeleton and sets of v-structures

Observational equivalence

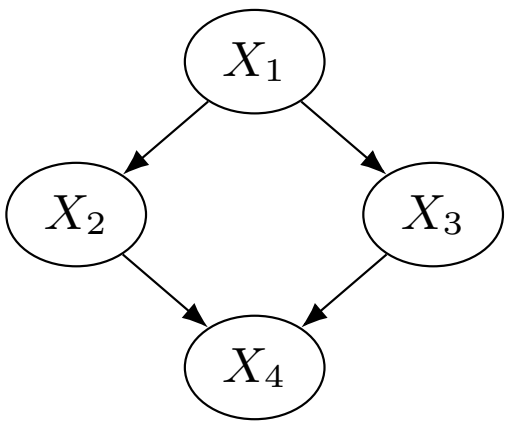
Theorem - Observational equivalence

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent iff they have the same skeleton and sets of v-structures

Observational equivalence

Theorem - Observational equivalence

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent iff they have the same skeleton and sets of v-structures

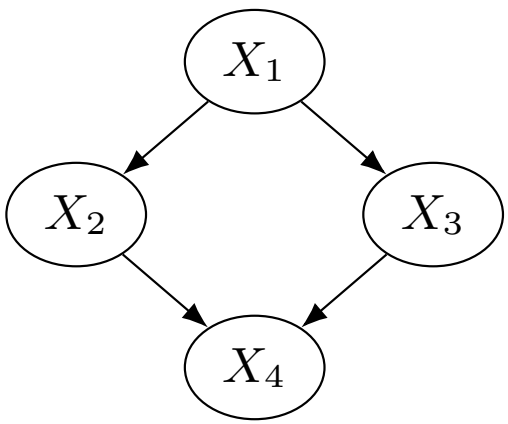


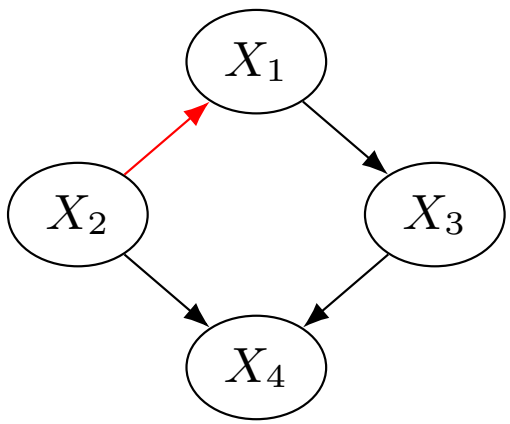
$$P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2,x_3)$$

Observational equivalence

Theorem - Observational equivalence

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent iff they have the same skeleton and sets of v-structures





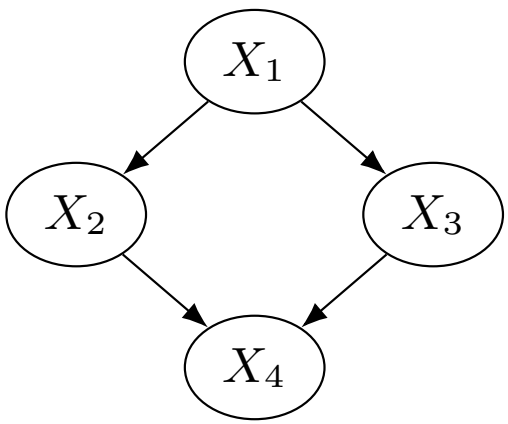
$$P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2,x_3) \quad = \quad P(x_2)P(x_1|x_2)P(x_3|x_1)P(x_4|x_2,x_3)$$

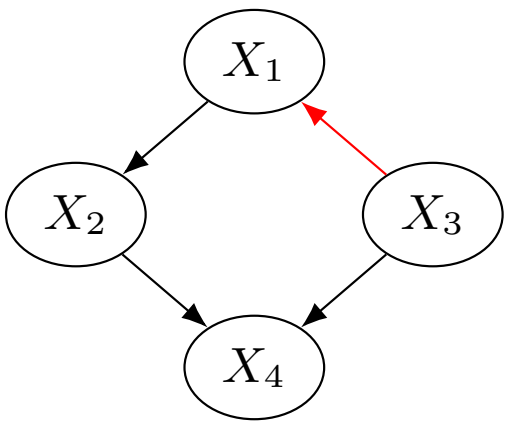
All these DAGs are observationally valid - They capture the same CI structure

Observational equivalence

Theorem - Observational equivalence

Two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent iff they have the same skeleton and sets of v-structures





$$P(x_1)P(x_2|x_1)P(x_3|x_1)P(x_4|x_2,x_3) \quad = \quad P(x_3)P(x_1|x_3)P(x_2|x_1)P(x_4|x_2,x_3)$$

All these DAGs are observationally valid - They capture the same CI structure

Lecture 1 Recap

- **What is Causal Inference:** It is the study of statistical methods to identify the effect of interventions.
- **Fundamental Problem Of Causal Inference:** We never observe both **potential outcomes** $(Y_1(u), Y_0(u))$ simultaneously.
- **Estimands of interest:**
  1. Individual Treatment Effect (ITE): What is the effect of an intervention on this individual: $\text{ITE}(u) := Y_1(u) - Y_0(u)$.
  2. Average Treatment Effect (ATE): What is the effect of an intervention on a population: $\text{ATE} := \mathbb{E}_{u \sim P(u)}\left[Y_1(u) - Y_0(u)\right]$.
  3. Conditional Average Treatement Effect: What is the effect of an intervention on a group summarized by covariates that can be conditioned on: $\mathbb{E}[Y_1|X] - \mathbb{E}[Y_0|X]$.

Lecture 1 Recap

Problem: The fundamental problem of causal inference makes it challenging to find these estimands without access to an oracle.

Strategy:

1. Write down the estimate of interest,
2. Make assumptions about the behavior of random variables in the problem,
3. Assumptions enable us to write down causal effects using quantities we can estimate from data.

We'll see this strategy arise time and again in this class.

Lecture 1 Recap

Assumptions we covered:

1. SUTVA: $Y_{0,1}(u_1) \perp\!\!\!\perp Y_{0,1}(u_k)\ \forall k \neq 1$
2. Consistency: Factual matches the observed outcome
3. Ignorability/Exchangeability: Potential outcomes are independent given treatment
4. Conditional Ignorability/Exchangeability: Potential outcomes are independent given treatment conditional on covariates [adjustment set]
5. Positivity/Overlap: The non-parameteric estimator for ATE requires us to have a positive probability of being assigned treatment or control for each configuration of patient

Lecture 1 Recap

- **Positivity Unconfoundedness tradeoff:** Including more variables means we're likely to have a valid adjustment set. Comes at the cost of satisfying overlap due to high-dimensionality
- **Overview of graphical models:** Graphical representation of relationships between random variables
- Deep connections between the conditional independences in a joint probability distribution and the graph structure

 Peters, Jonas, Dominik Janzing, and Bernhard Schlkopf (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press. ISBN: 0262037319.

 D'Amour, Alexander et al. (2021). "Overlap in observational studies with high-dimensional covariates". In: *Journal of Econometrics* 221.2, pp. 644–654.