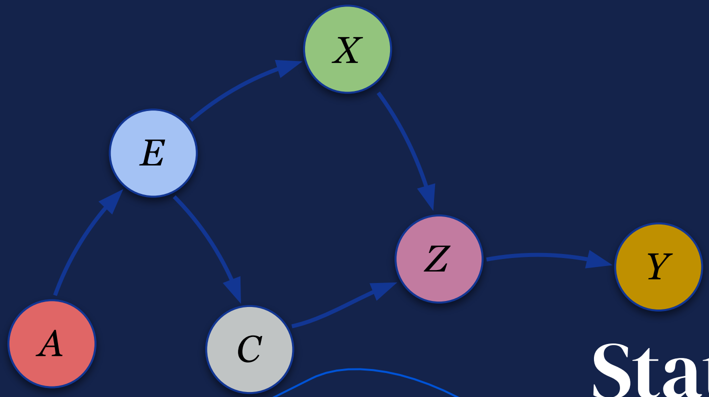
DeepMind



Statistical Causality

Causal ML Fairness, Estimation of Causal Effects, Causal Decision Making

Silvia Chiappa
csilvia@deepmind.com

Greek Stochastics
23-26/08/2022

Outline

Part I: Causal Inference (CI) via Causal Bayesian Networks (CBNs)

Part II: CBNs and CI for ML Fairness

Machine Learning

ML Prerequisites

Linear Algebra Optimization

Calculus Probability Statistics

ML Approaches and Methods

Linear Models

Regression, Classification,
Maximum Likelihood, Bayesian Inference

Neural Networks

Backpropagation, Automatic Differentiation,
Gradient Optimization, Deep Learning

Graphical Models

Directed, Undirected, Factor Graphs,
Conditional Independence, Message Passing

Sequential Decision Making

Reinforcement Learning

Approximate Inference

Variational Inference,
Monte-Carlo Methods

...

Statistical Causality

Machine Learning

ML Prerequisites

Basic Understanding of Statistical Causality,

...

Core Ingredients of ML Development/Evaluation/Deployment

Causally-based Methods/Considerations,

...

Obstacles

1. Knowledge barrier
2. Difficulties in identification (black-box)
3. Difficulties in estimation
4. Not a bottleneck

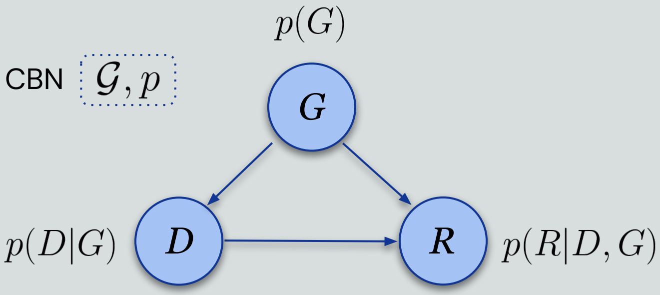
Part I
Causal Inference (CI)
via
Causal Bayesian
Networks (CBNs)

Causal Inference

Separation of Statistical Dependence into Causal Influence and Spurious Correlation

Drug Example

- Doctor gives drug more often to male individuals $G \rightarrow D$
- Recovery is influenced by drug $D \rightarrow R$ and gender $G \rightarrow R$ (male individuals recover more often)



→ [*Causality: Models, Reasoning, and Inference.*]() [J. Pearl.]()

Goal: Learn from data whether giving the drug helps recovery

Statistical Dependence (ML Approach)

Conditional **observational distribution** $p(R \mid D)$
Measures statistical dependence between D and R containing both causal influence ($D \rightarrow R$) and spurious correlation ($D \leftarrow G \rightarrow R$)

Causal Influence (Causal Inference Approach)

Conditional **interventional distribution**

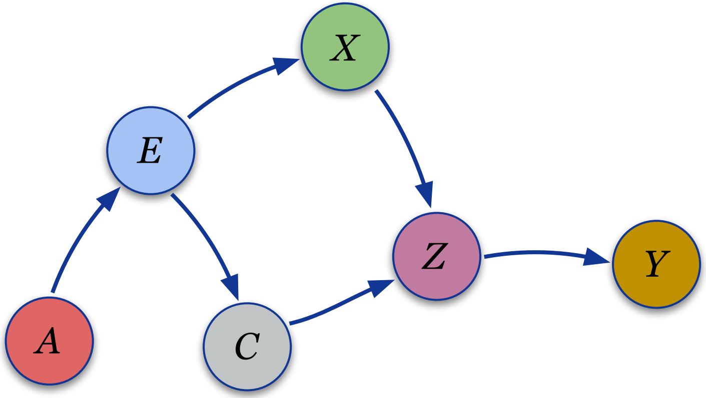
$$p_{\rightarrow D}(R|D) = p(R|\mathrm{do}(D))$$

Measures only causal influence ($D \rightarrow R$)

Causal Inference via CBNs

Identification Problem: Answer whether it is possible / how to express a causal quantity as a **functional of the observational distribution *p***, using knowledge of the causal graph

Estimation Problem: How to estimate the functional



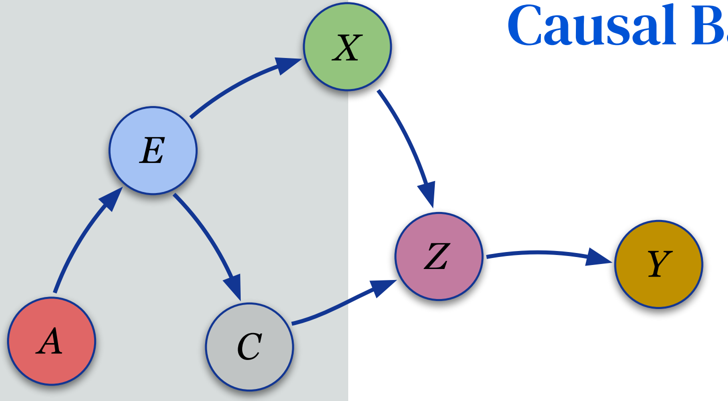
(Causal) Bayesian Networks

- Probabilistic Graphical Models: Principles and Techniques. D. Koller, N. Friedman.
- Bayesian Reasoning and Machine Learning. D. Barber.
- Pattern Recognition and Machine Learning. C. Bishop.

- Causal Inference in Statistics: A Primer. J. Pearl, M. Glymour, N. P. Jewell.
- Elements of Causal Inference Foundations and Learning Algorithms. J. Peters, D. Janzing, B. Schölkopf.
- Causality: Models, Reasoning, and Inference. J. Pearl.

Bayesian Networks

Causal Bayesian Networks



Visually express
statistical (in)dependence

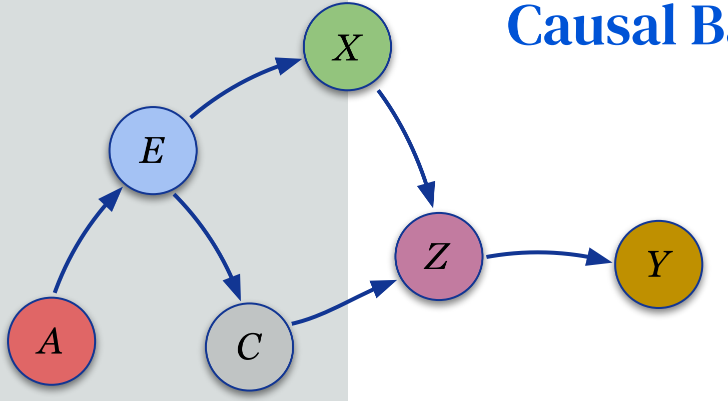
Visually express
causal dependence

Direct link between two nodes represents statistical relationship

Direct link between two nodes represents causal relationship

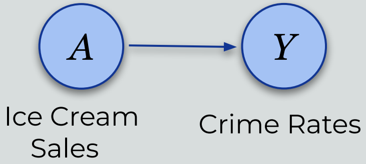
Bayesian Networks

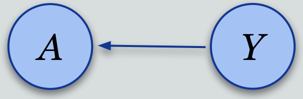
Causal Bayesian Networks

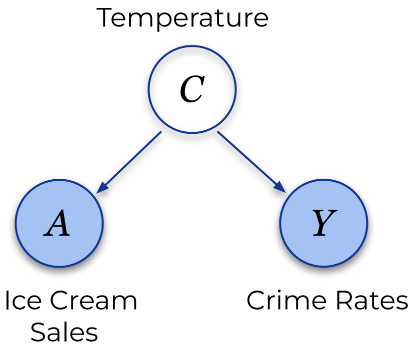


Ice Cream Example

Study shows that increase in ice cream sales A is correlated with increase in crimes Y
Not because ice cream causes crime, but because both ice cream sales and crimes are more common in hot weather







Bayesian Networks

Directed acyclic graph $\mathcal{G}$ in which each node is associated with conditional distribution given its parents $p(X_i|\text{pa}(X_i))$

Joint distribution of all nodes is given by the product of all conditional distributions

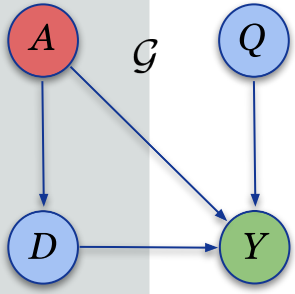
$$p(X_1, \ldots, X_I) = \prod_{i=1}^{I} p(X_i|\text{pa}(X_i))$$

Causal Bayesian Networks

Bayesian network with special semantic:
Each conditional distribution represents a causal mechanism

X is a cause of (influences) Y if there exists a causal path from X to Y

Causal path := directed path
$\rightarrow$ X is a cause of Y if X is an ancestor of Y
$\rightarrow$ X is a cause of Y if Y is a descendant of X



- A is a cause of Y
- A is not a cause of Q

$$p(A, Q, D, Y) = p(Y|A, Q, D)p(D|A)p(A)p(Q)$$

Statistical Independence in (Causal) Bayesian Networks (d-separation)

d-separation

X is d-separated from Y by Z ($X \perp\!\!\!\perp_{\mathcal{G}} Y \mid Z$) if all paths from any element of X to any element of Y are closed (or blocked) given Z

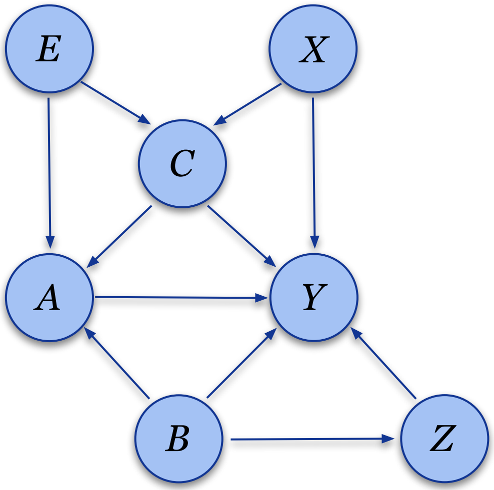
A path is closed if at least one of the following conditions is satisfied

1. There is a collider (node with 2 parents) on the path such that neither the collider nor any of its descendants belong to the conditioning set Z
2. There is a non-collider on the path which belongs to the conditioning set Z

d-separation implies statistical independence

If X is d-separated from Y by Z then X and Y are statistically independent given Z ($X \perp\!\!\!\perp_p Y \mid Z$)

$$X \perp\!\!\!\perp_{\mathcal{G}} Y \mid Z \Rightarrow X \perp\!\!\!\perp_p Y|Z$$



Statistical Independence in (Causal) Bayesian Networks (d-separation)

d-separation

X is d-separated from Y by Z ($X \perp\!\!\!\perp_{\mathcal{G}} Y \mid Z$) if all paths from any element of X to any element of Y are closed (or blocked) given Z

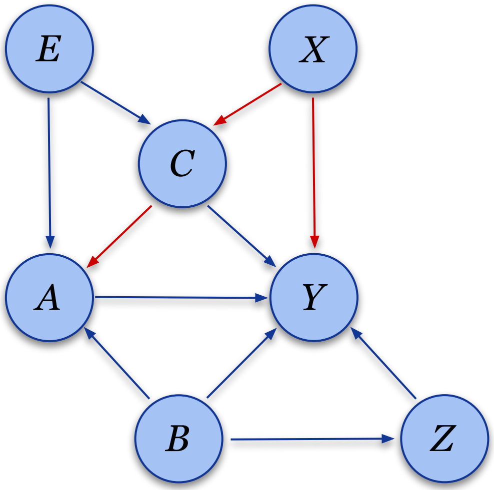
A path is closed if at least one of the following conditions is satisfied

1. There is a collider (node with 2 parents) on the path such that neither the collider nor any of its descendants belong to the conditioning set Z
2. There is a non-collider on the path which belongs to the conditioning set Z

d-separation implies statistical independence

If X is d-separated from Y by Z then X and Y are statistically independent given Z ($X \perp\!\!\!\perp_p Y \mid Z$)

$$X \perp\!\!\!\perp_{\mathcal{G}} Y \mid Z \Rightarrow X \perp\!\!\!\perp_p Y|Z$$



$A \leftarrow C \leftarrow X \rightarrow Y$ is open

Statistical Independence in (Causal) Bayesian Networks (d-separation)

d-separation

X is d-separated from Y by Z ($X \perp\!\!\!\perp_{\mathcal{G}} Y \mid Z$) if all paths from any element of X to any element of Y are closed (or blocked) given Z

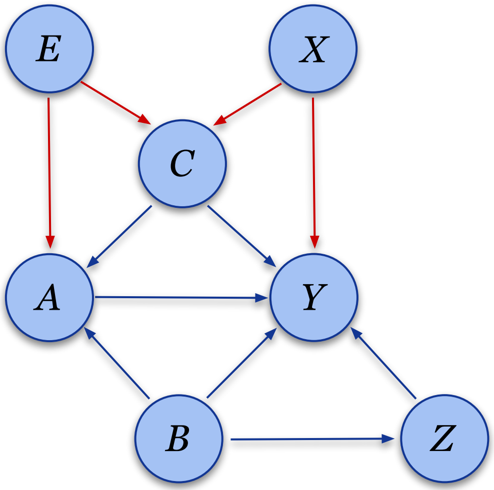
A path is closed if at least one of the following conditions is satisfied

1. There is a collider (node with 2 parents) on the path such that neither the collider nor any of its descendants belong to the conditioning set Z
2. There is a non-collider on the path which belongs to the conditioning set Z

d-separation implies statistical independence

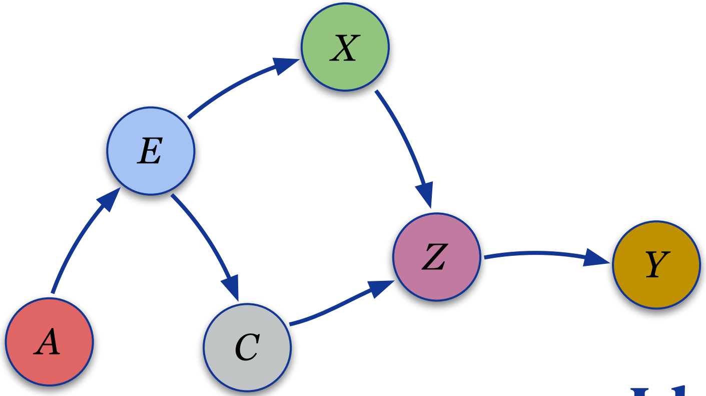
If X is d-separated from Y by Z then X and Y are statistically independent given Z ($X \perp\!\!\!\perp_p Y \mid Z$)

$$X \perp\!\!\!\perp_{\mathcal{G}} Y \mid Z \Rightarrow X \perp\!\!\!\perp_p Y|Z$$



$A \leftarrow E \rightarrow C \leftarrow X \rightarrow Y$ is closed

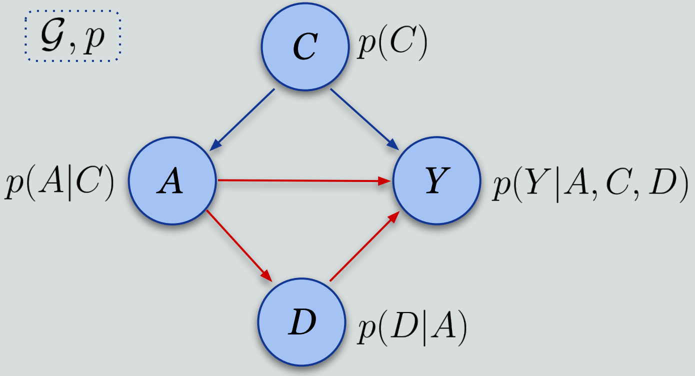
Conditioning on C opens the path



Identification of Causal Quantities

- *Probabilistic Graphical Models: Principles and Techniques.* D. Koller, N. Friedman.
- *Bayesian Reasoning and Machine Learning.* D. Barber.
- *Pattern Recognition and Machine Learning.* C. Bishop.
- *Causal Inference in Statistics: A Primer.* J. Pearl, M. Glymour, N. P. Jewell.
- *Elements of Causal Inference Foundations and Learning Algorithms.* J. Peters, D. Janzing, B. Schölkopf.
- *Causality: Models, Reasoning, and Inference.* J. Pearl.

Statistical Dependence between A and Y versus Causal Influence of A on Y



- Causal paths $A \rightarrow Y$
  $A \rightarrow D \rightarrow Y$
- Non-causal path (back-door path)

  $A \leftarrow C \rightarrow Y$

$A \leftarrow C \rightarrow Y$ Statistical Dependence

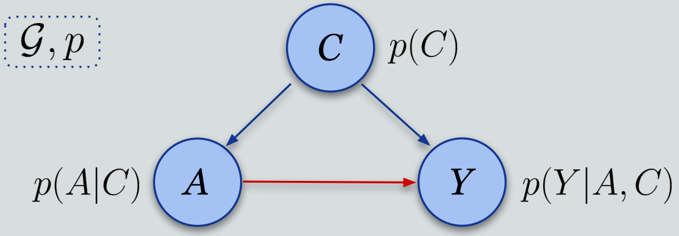
$\approx$

$A \rightarrow Y$
$A \rightarrow D \rightarrow Y$ Causal Influence

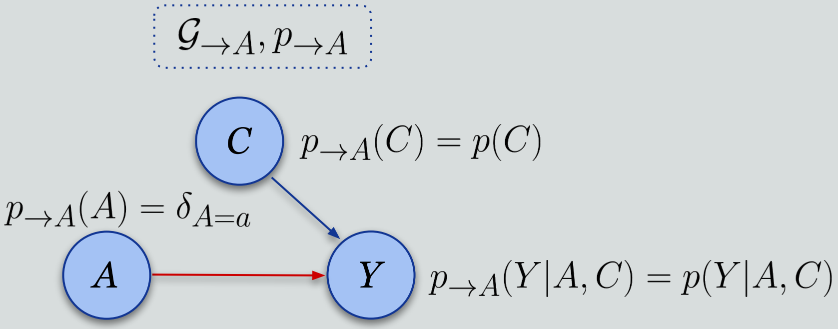
Causal influence $p_{\rightarrow A}(Y|A)$

Conditional distribution of Y given A without the information from A traveling through $A \leftarrow C \rightarrow Y$

Actual Hard Intervention



1. Get new data from interventional distribution $p_{\rightarrow A}$ where A i set to value a (does not depend on C) $p_{\rightarrow A}(A) = \delta_{A=a}$



Causal Influence

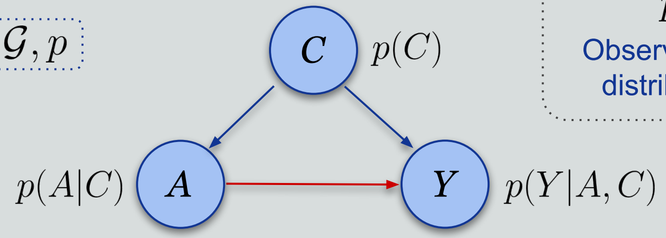
2. Compute conditional interventional distribution $p_{\rightarrow A}(Y|A)$ using the interventional data

$$
\begin{aligned}
p_{\rightarrow A}(Y|A) &= \sum_C p_{\rightarrow A}(Y|A, C) p_{\rightarrow A}(C|A) \\
&= \sum_C p_{\rightarrow A}(Y|A, C) p_{\rightarrow A}(C) \\
&= \sum_C p(Y|A, C) p(C)
\end{aligned}
$$

Hypothetical Hard Intervention

$\mathcal{G}, p$

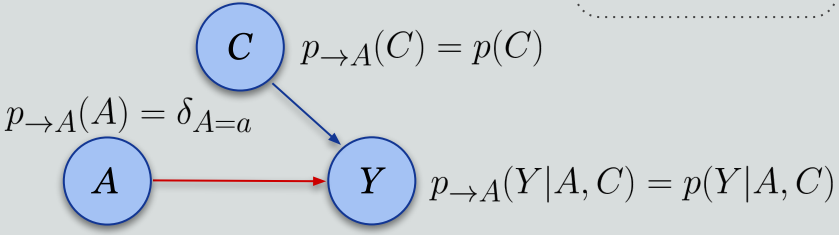
p
Observational distribution



1. Change $p(A|C)$ into $p_{\to A}(A) = \delta_{A=a}$
2. Leave other conditional distributions as in p

$\mathcal{G}_{\to A}, p_{\to A}$

$p_{\to A}$
Interventional distribution

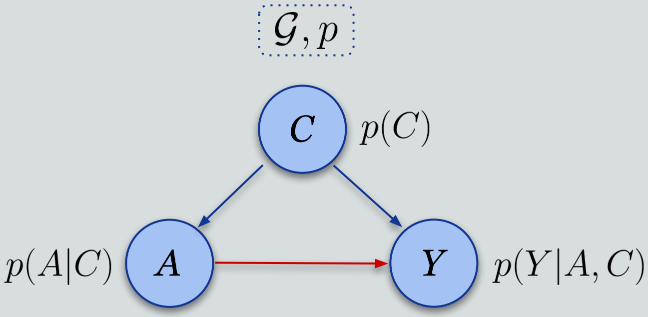


Causal Influence

3. Express conditional interventional distribution as a functional of the observational distribution

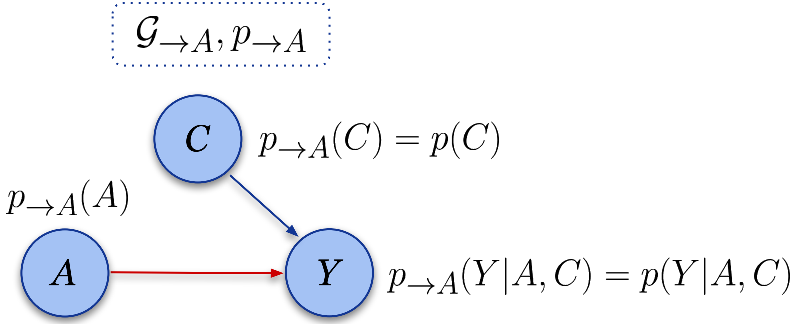
$$
\begin{aligned}
p_{\to A}(Y|A) &= \sum_C p_{\to A}(Y|A,C)\,p_{\to A}(C|A) \\
&= \sum_C p_{\to A}(Y|A,C)\,p_{\to A}(C) \\
&= \sum_C p(Y|A,C)\,p(C)
\end{aligned}
$$

Statistical Dependence



$$p(Y|A) = \sum_{C} p(Y|A, C)p(C|A)$$

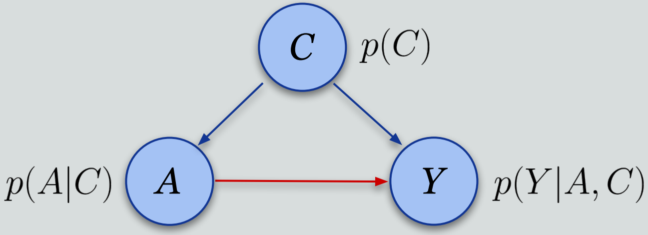
Causal Influence



$$p_{\to A}(Y|A) = \sum_{C} p(Y|A, C)p(C)$$

Statistical Dependence

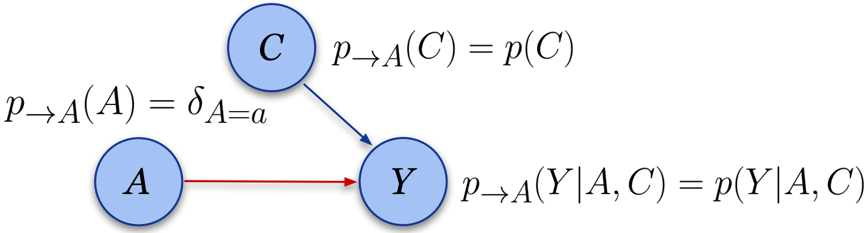
$\mathcal{G}, p$



$$\boxed{p(Y|A) = \sum_C p(Y|A,C)\,p(C|A)}$$

Causal Influence

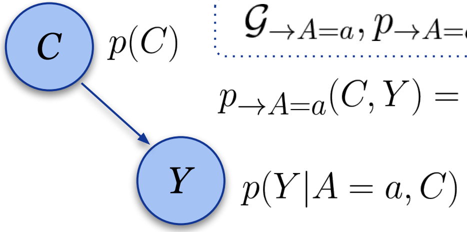
$\mathcal{G}_{\to A}, p_{\to A}$



NOTE $p_{\to A}(Y) = p_{\to A}(Y|A=a)$

$$= \sum_C p(Y|A=a,C)\,p(C)$$

$\mathcal{G}_{\to A=a}, p_{\to A=a}$



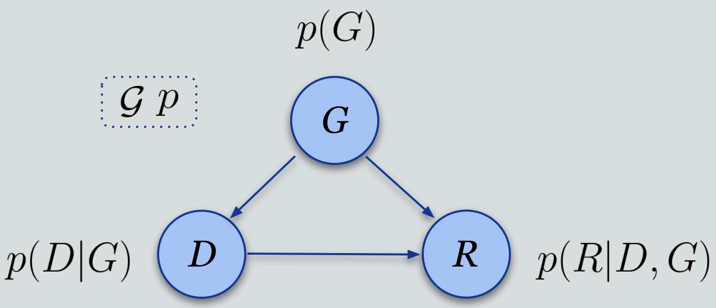
$$p_{\to A=a}(C,Y) = p(Y|A=a,C)\,p(C)$$

Drug Example

Separation of Causal Influence from Spurious Correlation

Data Generation Mechanism

- Doctor gives drug more often to male individuals $G \to D$
- Recovery is influenced by drug $D \to R$ and gender $G \to R$



→ [*Causality: Models, Reasoning, and Inference.* J. Pearl.]

Females	$R=0$	$R=1$		Recovery Rate
No Drug $(D=0)$	21	9	30	30%
Drug $(D=1)$	8	2	10	20%
	29	11	40	

Males	$R=0$	$R=1$		Recovery Rate
No Drug $(D=0)$	3	7	10	70%
Drug $(D=1)$	12	18	30	60%
	15	25	40	

	$R=0$	$R=1$		Recovery Rate
No Drug $(D=0)$	24	16	40	40%
Drug $(D=1)$	20	20	40	50%
	44	36	80	

Drug Example

Separation of Causal Influence from Spurious Correlation

Example of Simpson's Paradox

Statistical association that holds in several different groups of data is reversed when groups are combined

- According to the female data and to the male data the drug decreases recovery

- According to the combined male and female data the drug increases recovery

We believe this is a paradox as we wrongly believe that conditional distribution gives causal effect

Females	$R=0$	$R=1$		Recovery Rate
No Drug $(D=0)$	21	9	30	30%
Drug $(D=1)$	8	2	10	20%
	29	11	40	

Males	$R=0$	$R=1$		Recovery Rate
No Drug $(D=0)$	3	7	10	70%
Drug $(D=1)$	12	18	30	60%
	15	25	40	

	$R=0$	$R=1$		Recovery Rate
No Drug $(D=0)$	24	16	40	40%
Drug $(D=1)$	20	20	40	50%
	44	36	80	

Simpson's Paradox

Females	$R=0$	$R=1$		Recovery Rate
No Drug ($D=0$)	21	9	30	30%
Drug ($D=1$)	8	2	10	20%
	29	11	40	

Males	$R=0$	$R=1$		Recovery Rate
No Drug ($D=0$)	3	7	10	70%
Drug ($D=1$)	12	18	30	60%
	15	25	40	

	$R=0$	$R=1$		Recovery Rate
No Drug ($D=0$)	24	16	40	40%
Drug ($D=1$)	20	20	40	50%
	44	36	80	

Conditional Distribution

$$p(G=F|D=0)=\frac{30}{40}, p(G=M|D=0)=\frac{10}{40}$$
$$p(G=F|D=1)=\frac{10}{40}, p(G=M|D=1)=\frac{30}{40}$$

$$\begin{aligned} p(R=1|D=0) &= \sum_G p(R=1|D=0,G)p(G|D=0) \\ &= \frac{9}{\cancel{30}}\cdot\frac{\cancel{30}}{40}+\frac{7}{\cancel{10}}\cdot\frac{\cancel{10}}{40}=\frac{16}{40}=0.4 \\ p(R=1|D=1) &= \sum_G p(R=0|D=1,G)p(G|D=1) \\ &= \frac{2}{\cancel{10}}\cdot\frac{\cancel{10}}{40}+\frac{18}{\cancel{30}}\cdot\frac{\cancel{30}}{40}=\frac{20}{40}=0.5 \end{aligned}$$

Simpson's Paradox

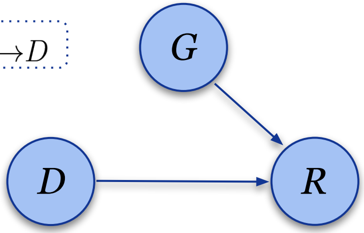
Females	$R=0$	$R=1$		Recovery Rate
No Drug ($D=0$)	21	9	30	30%
Drug ($D=1$)	8	2	10	20%
	29	11	40	

Males	$R=0$	$R=1$		Recovery Rate
No Drug ($D=0$)	3	7	10	70%
Drug ($D=1$)	12	18	30	60%
	15	25	40	

	$R=0$	$R=1$		Recovery Rate
No Drug ($D=0$)	24	16	40	40%
Drug ($D=1$)	20	20	40	50%
	44	36	80	

Causal Influence

$\mathcal{G}_{\to D}, p_{\to D}$



$$p(G=F) = \frac{40}{80}, p(G=M) = \frac{40}{80}$$

$$p_{\to D=0}(R=1|D=0) = \sum_G p(R=1|D=0,G)p(G)$$

$$= \frac{9}{30}\cdot\frac{40}{80} + \frac{7}{10}\cdot\frac{40}{80} = \frac{9+21}{30}\cdot\frac{1}{2} = 0.5$$

$$p_{\to D=1}(R=1|D=1) = \sum_G p(R=1|D=1,G)p(G)$$

$$= \frac{2}{10}\cdot\frac{40}{80} + \frac{18}{30}\cdot\frac{40}{80} = \frac{6+18}{30}\cdot\frac{1}{2} = 0.4$$

Identification via Do-calculus

Three graphical rules that, combined, might allow to express a causal quantity as a functional of the observational distribution

Might: In the presence of latent variables, for certain graph structures identification is not possible

conditioning

Rule 1 – Insertion/Deletion of Observations

Rule 2 – Action/Observation Exchange

Rule 3 – Insertion/Deletion of Actions

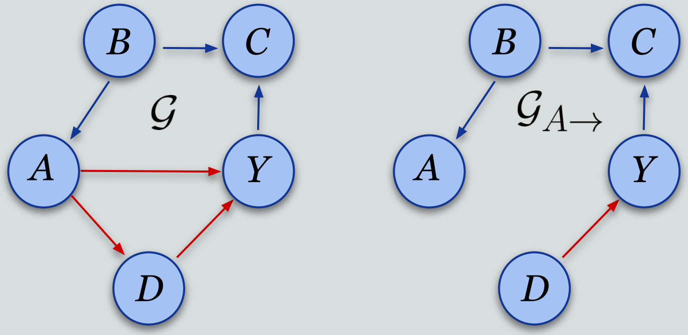
intervening

Rules of Do-calculus (Simpler Version)

Rule 2 - Action/Observation Exchange

$$p_{\rightarrow A}(Y|A) = p(Y|A)$$

$$\text{if } Y \perp\!\!\!\perp_{\mathcal{G}_{A\rightarrow}} A$$



Intuition

The graph $\mathcal{G}_{A\rightarrow}$ has all links emerging from A removed.
If Y is independent of A in this graph it means that all backdoor paths (non-causal paths) are closed, therefore the interventional and conditional distributions are equivalent.
The only open paths are causal paths!

1. Backdoor path from A to Y: $A \leftarrow$ ... Y
Backdoor paths must be non-causal as otherwise we would create a cycle (A is a cause of Y which is a causal of A)

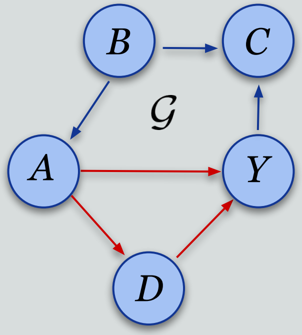
2. Frontdoor path from A to Y: $A \rightarrow$ ... Y
Only causal frontdoor paths are open, as to be non-causal a frontdoor path must contain a collider $A \rightarrow$ ...$\rightarrow Z \leftarrow$... Y

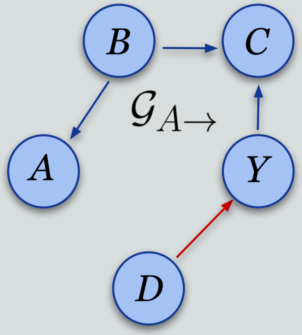
Rules of Do-calculus (Simpler Version)

Rule 2 - Action/Observation Exchange

$$p_{\to A}(Y|A) = p(Y|A)$$

$$\text{if } Y \perp\!\!\!\perp_{\mathcal{G}_{A\to}} A$$



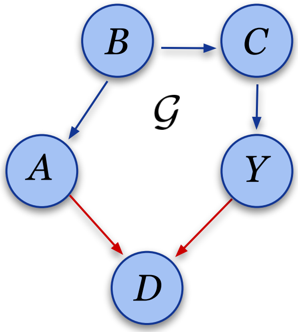


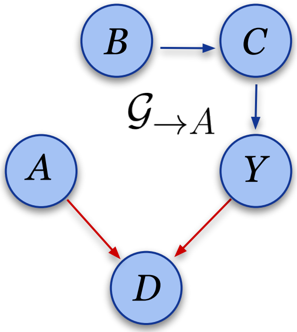
Intuition
The graph $\mathcal{G}_{A\to}$ has all links emerging from A removed.
If Y is independent of A in this graph it means that all backdoor paths (non-causal paths) are closed, therefore the interventional and conditional distributions are equivalent.
The only open paths are causal paths!

Rule 3 - Insertion/Deletion of Action

$$p_{\to A}(Y|A) = p(Y)$$

$$\text{if } Y \perp\!\!\!\perp_{\mathcal{G}_{\to A}} A$$





Intuition
The graph $\mathcal{G}_{\to A}$ has all links pointing to A removed.
If Y is independent of A in this graph it means that there are no causal paths from A to Y, i.e. A is not a cause of Y, therefore intervening on A or ignoring A is the same. There is no influence reaching Y from A.
The only possibly open paths are non-causal paths!

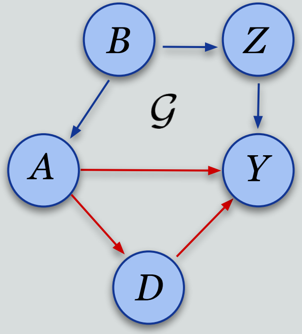
Rules of Do-calculus

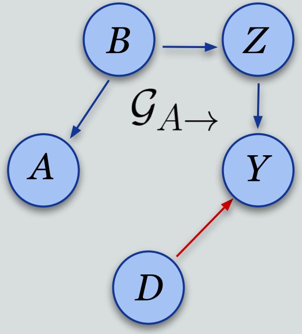
Called Adjustment Criterion for a set Z satisfying different conditions (called adjustment set)

Backdoor Criterion

$$p_{\to A}(Y|A) = \sum_Z p(Y|A,Z)p(Z)$$

If Z blocks all backdoor paths from A to Y ($Y \perp\!\!\!\perp_{\mathcal{G}_{A\to}} A \,|\, Z$)
No node in Z is a descendant of A





$$\begin{aligned} p_{\to A}(Y|A) &= \sum_Z p_{\to A}(Y|A,Z)p_{\to A}(Z|A) \\ &= \sum_Z p(Y|A,Z)p(Z) \end{aligned}$$

$Y \perp\!\!\!\perp_{\mathcal{G}_{A\to}} A \,|\, Z$

Rule 2
Action / Observation exchange

$Z \perp\!\!\!\perp_{\mathcal{G}_{\to A}} A$

Rule 3
Insertion / Deletion of Action

1. Z is in a backdoor path from A to Y: Removing links pointing to A blocks the subpaths $A\leftarrow$ ... Z
A frontdoor path from A to Z must contain a collider otherwise Y would be a cause of A.

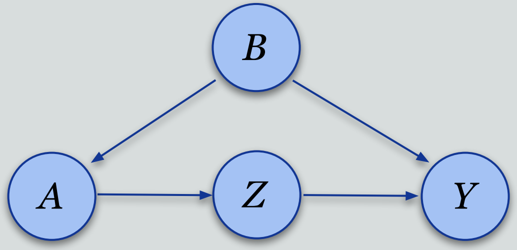
2. Z is in a frontdoor path from A to Y: The subpath $A\to$ ... Z cannot be causal, as otherwise Z would be a descendant of A, and therefore is closed (must contain a collider).

Rules of Do-calculus

Frontdoor Criterion

$$p_{\to A=a}(Y|A=a) = \sum_Z p(Z|A=a) \sum_A p(Y|Z,A)p(A)$$

1. If all backdoor paths from Z to Y are blocked by A ($Y \perp\!\!\!\perp_{\mathcal{G}_{Z\to}} Z \mid A$)
2. There is no open backdoor path from A to Z
3. Z intercepts all causal paths from A to Y



$Z \perp\!\!\!\perp_{\mathcal{G}_{A\to}} A$

Rule 2

Action / Observation exchange

Backdoor paths from A to Z: are closed. Causal paths from from A to Z: are cut in the graph with emerging links from A removed.

$Y \perp\!\!\!\perp_{\mathcal{G}_{\to A, Z\to}} Z$

Rule 2 Action / Observation exchange

$$\begin{aligned} p_{\to A=a}(Y|A=a) &= \sum_Z p_{\to A=a}(Y|A=a,Z)p_{\to A=a}(Z|A=a) \\ &= \sum_Z p_{\to A=a,\to Z}(Y|A=a,Z)p(Z|A=a) \\ &= \sum_Z p_{\to Z}(Y|Z)p(Z|A=a) \\ &= \sum_Z p(Z|A=a)\sum_A p(Y|Z,A)p(A) \end{aligned}$$

$Y \perp\!\!\!\perp_{\mathcal{G}_{\to A, Z\to}} A$

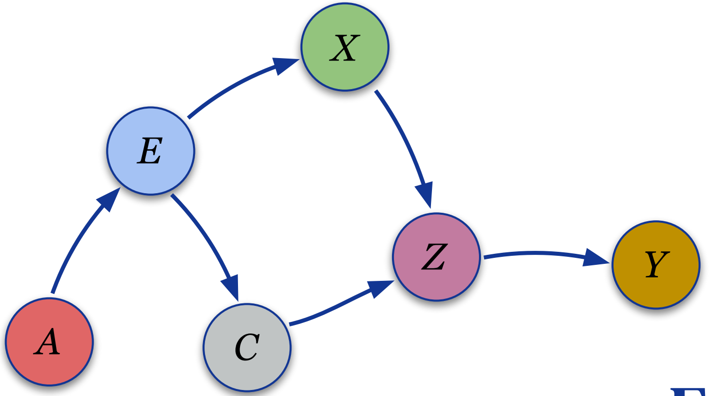
Rule 3

Insertion / Deletion of Action

Backdoor paths from A to Y: are cut in the graph with incoming links into A removed.
Causal paths from from A to Y: Z intercepts these paths which are cut in the graph with emerging links from Z removed.

$Y \perp\!\!\!\perp_{\mathcal{G}_{Z\to}} Z \mid A$

Backdoor Criterion



Estimation of Causal Quantities

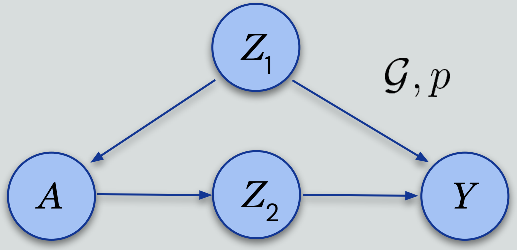
Estimation of Causal Quantities

Using knowledge of the CBN, do-calculus enables us (when possible) to obtain identification formulas, i.e. formulas that express a causal quantity as a functional of the observational distribution

1. Obtain an estimate of an identification formula with desirable properties
2. Choose between several identification formulas

→ *Asymptotically Best Causal Effect Identification with Multi-Armed Bandits. A. Malek, S. Chiappa, 2021.*

Selection of Identification Formulas



Identification Formulas for $\tau = \mathbb{E}_{p_{\rightarrow A}(Y|A)}[Y]$

- Adjustment criterion using Z_1
- Frontdoor criterion using Z_2

- **Statistical Considerations:** Asymptotic variance (each formula associated with asymptotically linear estimator of $\tau \rightarrow$ convergence rate $O\left(\sqrt{n}\right)$)
- **Practical Considerations:** Costs in observing covariates

Asymptotic Variance

Graphical Criteria for Selection of Adjustment Sets

[1] **Graphical Criteria for Efficient Total Effect Estimation via Adjustment in Causal Linear Models.** L. Henckel, E. Perković, M. H. Maathuis, 2019.

[2] **Efficient Adjustment Sets for Population Average Causal Treatment Effect Estimation in Graphical Models.** A. Rotnitzky and E. Smucler, 2020.

[3] **On efficient adjustment in causal graphs.** J. Witte, L. Henckel, M. H. Maathuis, V. Didelez, 2020.

[4] **Efficient Adjustment Sets in Causal Graphical Models with Hidden Variables**. E. Smucler, F. Sapienza, A. Rotnitzky, 2021.

Graphical Criteria for Selection of Adjustment Sets

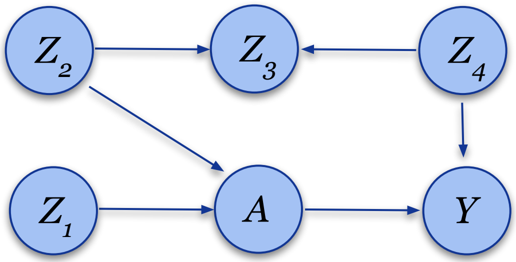
$$B \cap G = \emptyset$$

1.
Addition of Precision Set

Adding to the adjustment set B a *precision* set G, i.e. variables that d-separated from A by B ($A \perp\!\!\!\perp_{\mathcal{G}} G \mid B$) gives an adjustment set $B \cup G$ with smaller asymptotic variance

2.
Removal of Overadjustment Set

Removing from the adjustment set $B \cup G$ the *overadjustment* set B, i.e. variables that are d-separated from Y by G and A ($Y \perp\!\!\!\perp_{\mathcal{G}} B \mid G \cup A$) gives an adjustment set G with smaller asymptotic variance



Adjustment sets

$\emptyset, \{Z_1\}, \{Z_2\}, \{Z_4\}$
$\{Z_1, Z_2\}, \{Z_1, Z_4\}$
$\{Z_2, Z_3\}, \{Z_2, Z_4\}, \{Z_3, Z_4\}$
$\{Z_1, Z_2, Z_3\}, \{Z_1, Z_2, Z_4\}$
$\{Z_1, Z_3, Z_4\}, \{Z_2, Z_3, Z_4\}$
$\{Z_1, Z_2, Z_3, Z_4\}$

2. Adjustment set $\{Z_1, Z_2, Z_3, Z_4\}$

$Y \perp\!\!\!\perp_{\mathcal{G}} \{Z_1, Z_2, Z_3\} \mid Z_4 \cup A$ ➡ $\{Z_1, Z_2, Z_3\}$ overadjustment set

2. Adjustment set Z_4

$Y \not\!\perp\!\!\!\perp_{\mathcal{G}} Z_4 \mid A$ ➡ Z_4 not an overadjustment set

1. Adjustment set $\emptyset$

$A \perp\!\!\!\perp_{\mathcal{G}} Z_4$ ➡ Z_4 precision set

2. Adjustment set $\{Z_1, Z_2\}$

$Y \perp\!\!\!\perp_{\mathcal{G}} \{Z_1, Z_2\} \mid A$ ➡ $\{Z_1, Z_2\}, Z_1, Z_2$ overadjustment sets

2. Adjustment set $\{Z_1, Z_2, Z_3\}$

$Y \perp\!\!\!\perp_{\mathcal{G}} Z_1 \mid \{Z_2, Z_3\} \cup A$ ➡ Z_1 overadjustment set

Graphical Criteria for Selection of Adjustment Sets

Causal nodes $\mathrm{cn}(A, Y, \mathcal{G})$: Nodes on causal paths from A to Y, excluding A

Forbidden nodes $\mathrm{forb}(A, Y, \mathcal{G}) = \mathrm{de}(\mathrm{cn}(A, Y, \mathcal{G}), \mathcal{G}) \cup A$

Comparison of Adjustment Sets

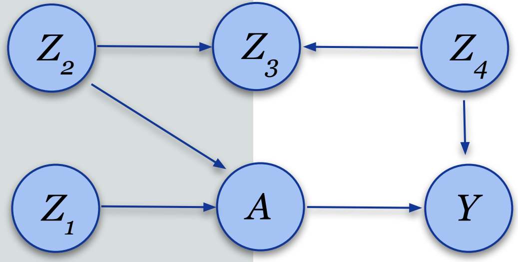
If B and G are adjustment sets with $A \perp\!\!\!\perp_{\mathcal{G}} G \setminus B \mid B$ and $Y \perp\!\!\!\perp_{\mathcal{G}} B \setminus G \mid G \cup A$ then the variance of G is smaller than the variance of B

Optimal Adjustment Set

The variance of the adjustment set

$$O(A, Y, \mathcal{G}) = \mathrm{pa}(\mathrm{cn}(A, Y, \mathcal{G}), \mathcal{G}) \setminus \mathrm{forb}(A, Y, \mathcal{G})$$

is smaller than for any other adjustment set



$$\mathrm{cn}(A, Y, \mathcal{G}) = \{Y\}$$
$$\mathrm{forb}(A, Y, \mathcal{G}) = \{A, Y\}$$
$$\mathrm{pa}(\mathrm{cn}(A, Y, \mathcal{G}), \mathcal{G}) = \{Z_4\}$$
$$O(A, Y, \mathcal{G}) = \{Z_4\}$$

Variance of Z_4 smaller than variance of empty set

$$B = \emptyset, G = \{Z_4\} \quad A \perp\!\!\!\perp_{\mathcal{G}} G \setminus B \mid B$$
$$B = \{Z_4\}, G = \emptyset \quad Y \not\perp\!\!\!\perp_{\mathcal{G}} B \setminus G \mid G \cup A$$

Cannot compare $\{Z_2, Z_3\}$ and empty set

$$B = \emptyset, G = \{Z_2, Z_3\} \quad A \not\perp\!\!\!\perp_{\mathcal{G}} G \setminus B \mid B$$
$$B = \{Z_2, Z_3\}, G = \emptyset \quad Y \not\perp\!\!\!\perp_{\mathcal{G}} B \setminus G \mid G \cup A$$

Z_4 latent variable

$\emptyset, \cancel{\{Z_1\}}, \cancel{\{Z_2\}}, \cancel{\{Z_4\}}$
$\cancel{\{Z_1, Z_2\}}, \cancel{\{Z_1, Z_4\}}$
$\{Z_2, Z_3\}, \cancel{\{Z_2, Z_4\}}, \cancel{\{Z_3, Z_4\}}$
$\cancel{\{Z_1, Z_2, Z_3\}}, \cancel{\{Z_1, Z_2, Z_4\}}$
$\cancel{\{Z_1, Z_3, Z_4\}}, \cancel{\{Z_2, Z_3, Z_4\}}$
$\cancel{\{Z_1, Z_2, Z_3, Z_4\}}$

→ cannot select

Asymptotically Best Causal Effect Identification with Multi-Armed Bandit

A selection method that is applicable to arbitrary identification formulas and accounts for both statistical performance and practical considerations must make use of data

Setting

- The investigator collects observational data
- Sequential strategy in which the investigator decides, observation-by-observation (rounds), which subset of variables to observe with the goal of identifying the best formula with the fewest observations

Multi-Armed Bandits Formalism

Cast this setting into the best-arm-identification bandit framework, by considering each formula as one arm and by replacing the goal of learning the arm with the best mean with the goal of learning the formula with the best cost-adjusted asymptotic variance

Each arm k corresponds to an estimator $\hat{\tau}_k$ of τ with asymptotic variance σ_k^2 and cost c_k

Goal: $k^* = \arg\min_k c_k \sigma_k^2$

Leverage algorithms from the bandit literature

Adapt LUCB and Successive Elimination (SE) by introducing finite-sample confidence sequences on $\hat{\sigma}_k^2(\mathcal{D}_n) - \sigma_k^2$ (confidence bounds that hold for all formulas and all rounds simultaneously)

Estimators of τ

Asymptotically linear estimator $\hat{\tau}_k$ with nuisance function η_k (e.g. propensity score for IPW)

Influence function ϕ_k satisfying $\mathbb{E}_p[\phi_k(W_k, \eta_k, \tau)] = 0 \quad \mathbb{E}_p[\phi_k^2(W_k, \eta_k, \tau)] < \infty$ $\quad W_k$ covariates for formula k, A, Y

$$\sqrt{n}(\hat{\tau}_k(\mathcal{D}_n) - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_k(w_k^i, \eta_k, \tau) + o_p(1)$$

Asymptotically normal with variance $\sigma_k^2 = \mathbb{E}_p[\phi_k^2(W_k, \eta_k, \tau)]$

Can assume an uncentered influence function $\phi_k(W_k, \eta_k, \tau) = \psi_k(W_k, \eta_k) - \tau$

$$\sigma_k^2 = \mathrm{Var}_p[\psi_k(W_k, \eta_k)]$$

When need to estimate η_k, converge of $\hat{\eta}_k$ at rate slower than $O(n^{-1/2})$ could cause loss of asymptotically linearity (can happen if e.g. $\hat{\eta}_k$ is modelled with with neural networks)

Asymptotically linear estimators for any identification formula [Jung, Tian, Bareinboim 2021] using double machine learning

Estimating the Asymptotic Variance

- Our goal: estimate $\sigma_k^2 = \mathbb{E}_p[\psi_k(W_k, \eta_k) - \tau)^2]$ and derive finite-sample confidence sequences
- Our estimator $\hat{\sigma}_k^2$ for data $\mathcal{D}_n$ (inspired by [Chernozhukov et al., 2016])
  - Randomly split $\mathcal{D}_n$ into two folds $\mathcal{D}_n^\eta, \mathcal{D}_n^\sigma$
  - Fit the nuisance function $\hat{\eta}_k(\mathcal{D}_n^\eta)$
  - Fit $\hat{\sigma}_k^2$ with an empirical variance: $\hat{\sigma}_k^2(\mathcal{D}_n) = \mathrm{Var}_{\mathcal{D}_n^\sigma}\left[\psi_k(W_k, \hat{\eta}_k(\mathcal{D}_n^\eta))\right]$

Main Theorem

$O\left(\sqrt{n}\right)$ estimation of the asymptotic variance is possible whenever $O\left(\sqrt{n}\right)$ estimation of τ is possible

Confidence Sequence LUCB (CS-LUCB)

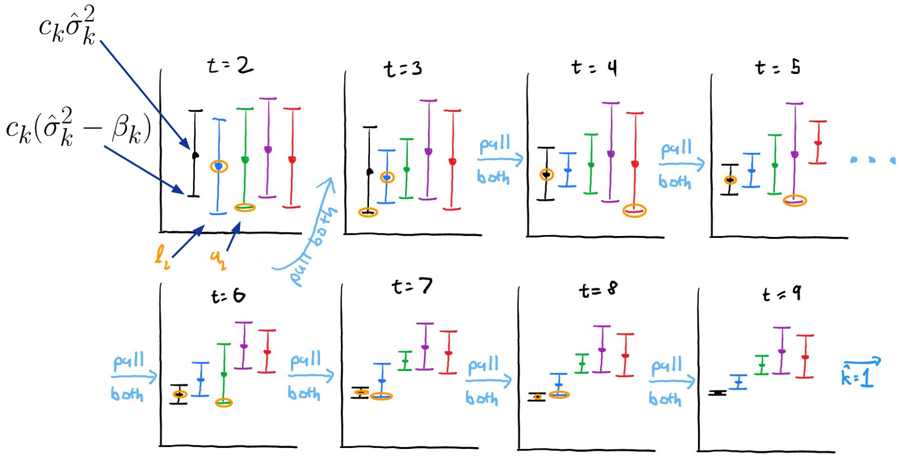
At each round *t*

lower bound

- Sample arms $l_t \leftarrow \arg\min_{k \in [K]} c_k \hat{\sigma}_k^2$
- Sample arms $u_t \leftarrow \arg\min_{k \neq l_t} c_k(\hat{\sigma}_k^2 - \beta_k)$
- If $c_{l_t}\left(\hat{\sigma}_{l_t}^2 + \beta_{l_t}\right) \leq c_{u_t}\left(\hat{\sigma}_{u_t}^2 - \beta_{u_t}\right) - \epsilon$ return $\hat{k} = l_t$ otherwise collect new data and update l_t and u_t

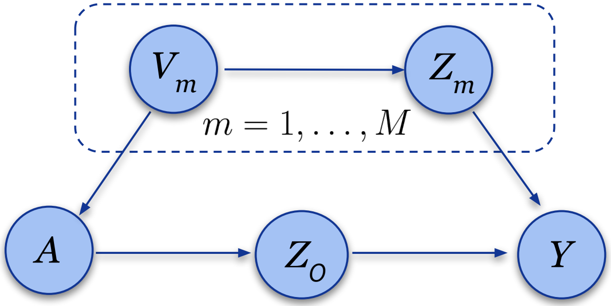
$\hat{k}$ is at most ϵ-suboptimal with probability at least $1 - \delta$

$$\mathbb{P}\left(c_{\hat{k}}\sigma_{\hat{k}}^2 \geq \min_k c_k \sigma_k^2 + \epsilon\right) \leq \delta$$



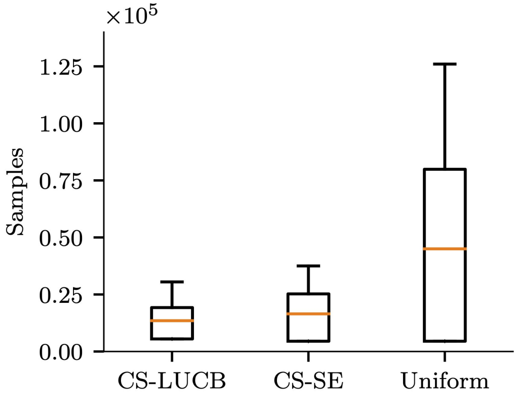
➡ Create gap between upper bounds of best arms and lower bounds of remaining arms

Linear Experiments



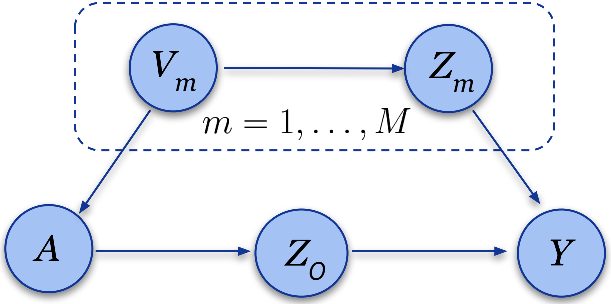
$$V_m \sim \mathcal{N}(0, I_2) \quad A \sim \text{Bernoulli}\left(1 \Big/ \left(1 + e^{-\sum_{m=1}^{M} B_m^\top V_m}\right)\right)$$

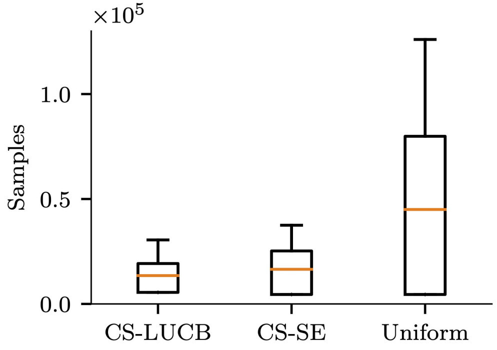
$$Z_m = A_m V_m + \epsilon_{z,m} \quad Z_0 \,|\, A \sim \text{Cat} \quad Y = \sum_{m=0}^{M} C_m^\top V_m + \epsilon_y$$

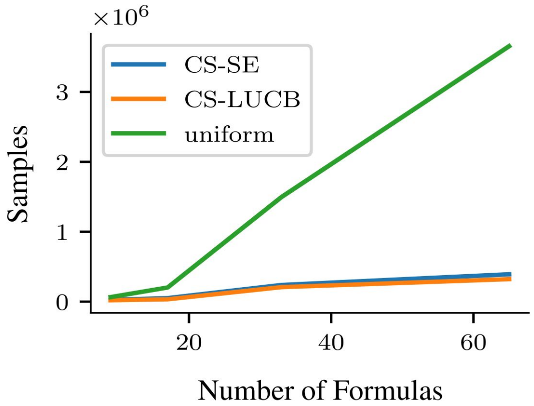


- $2^{M=3}$ =8 formulas from adjustment criterion
- One formula from frontdoor criterion Z_M

Non-Linear Experiments



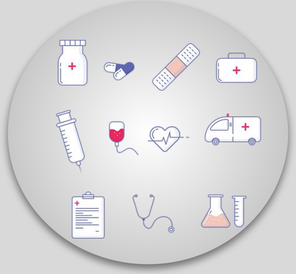




Part II
CBNs and CI for ML Fairness

ML Fairness

Machine Learning deployed to make decisions that affect people lives can treat individuals unfairly on the basis of sensitive attributes such as race, gender, disabilities, etc.



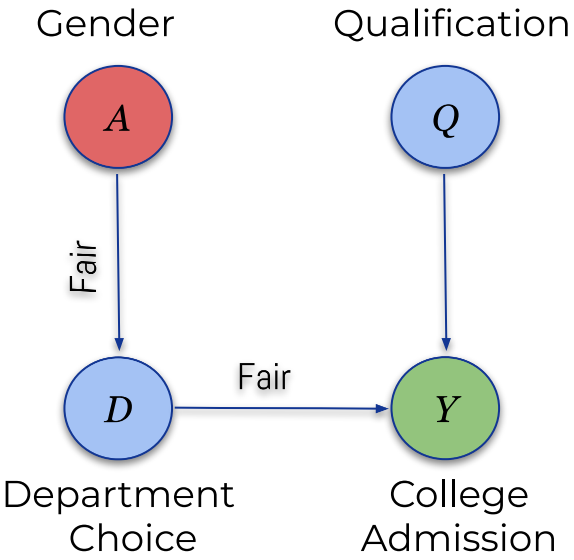
College Admission Example Characterising Patterns of Unfairness

Berkeley's alleged sex bias (1974)

Female applicants were rejected more often than male applicants

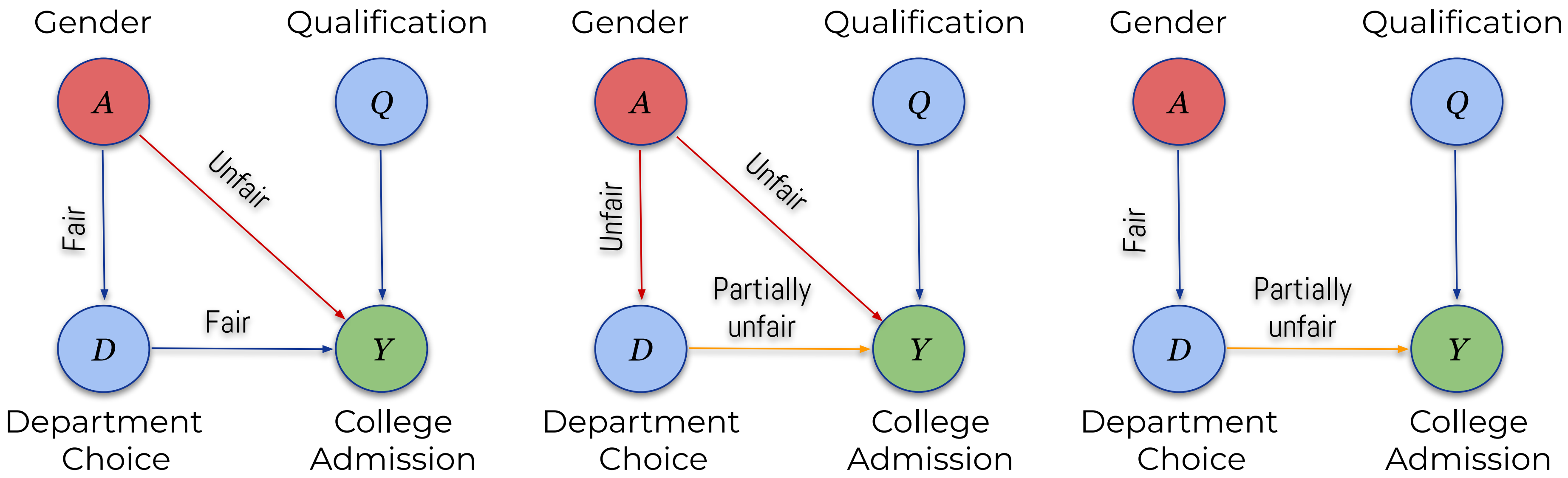
College Admission Example
Characterising Patterns of Unfairness

Berkeley's alleged sex bias (1974)



Women spontaneously apply to departements with lower admission rates

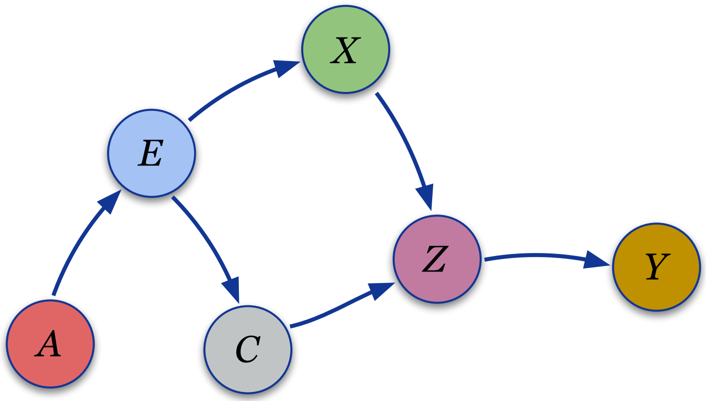
Different Possible Unfairness Scenarios



Women spontaneously apply to departements with lower admission rates

Women apply to departments with lower admission rates due to due systemic historical or cultural pressures

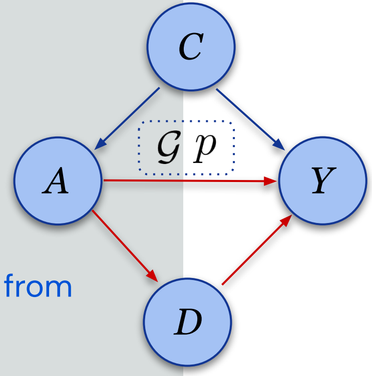
College lower admission rates of departments chosen spontaneously more often by women



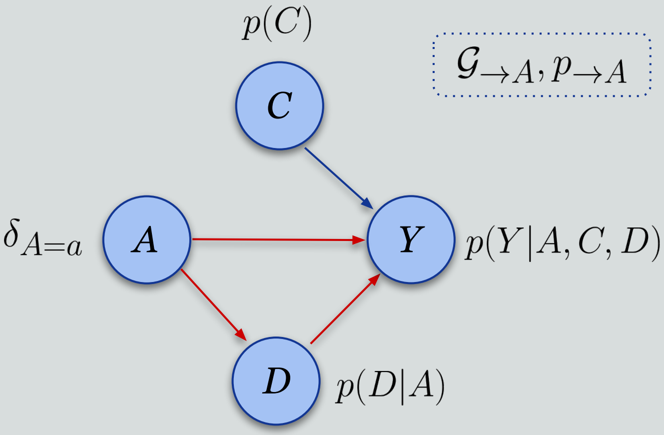
Path-Specific Influence

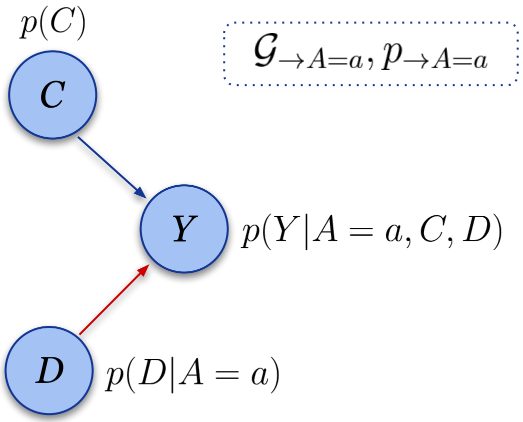
- *Probabilistic Graphical Models: Principles and Techniques.* D. Koller, N. Friedman.
- *Bayesian Reasoning and Machine Learning.* D. Barber.
- *Pattern Recognition and Machine Learning.* C. Bishop.
- *Causal Inference in Statistics: A Primer.* J. Pearl, M. Glymour, N. P. Jewell.
- *Elements of Causal Inference Foundations and Learning Algorithms.* J. Peters, D. Janzing, B. Schölkopf.
- *Causality: Models, Reasoning, and Inference.* J. Pearl.

Causal Influence



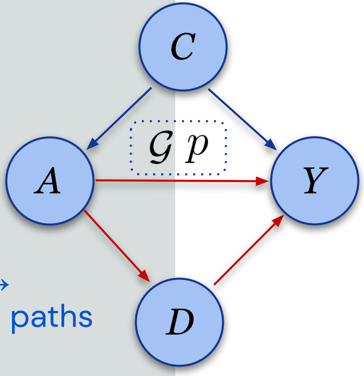
Hard intervention → separate causal influence from spurious correlation



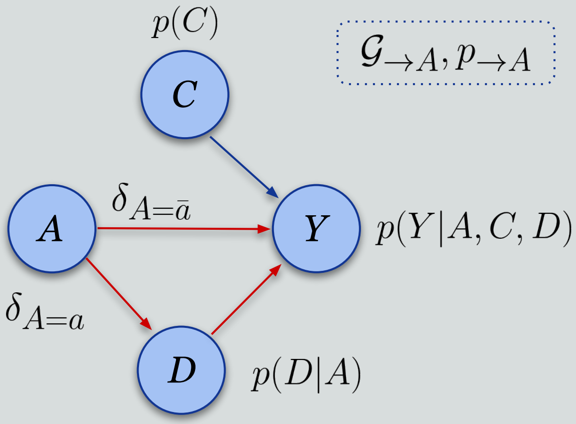


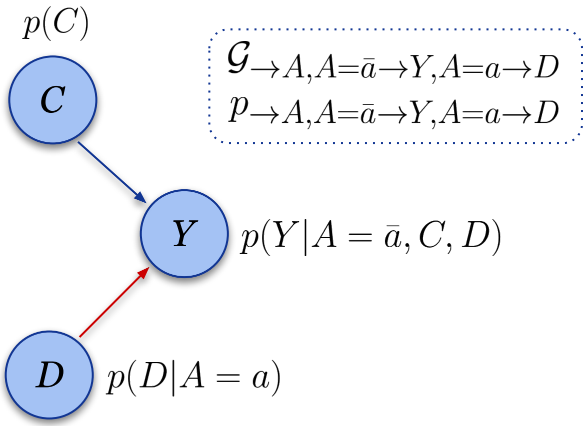
$$
\begin{aligned}
p_{\rightarrow A=a}(Y) &= \sum_{C,D} p_{\rightarrow A=a}(C, D, Y) \\
&= \sum_{C,D} p(Y|A = a, C, D)p(D|A = a)p(C)
\end{aligned}
$$

Path-Specific Influence



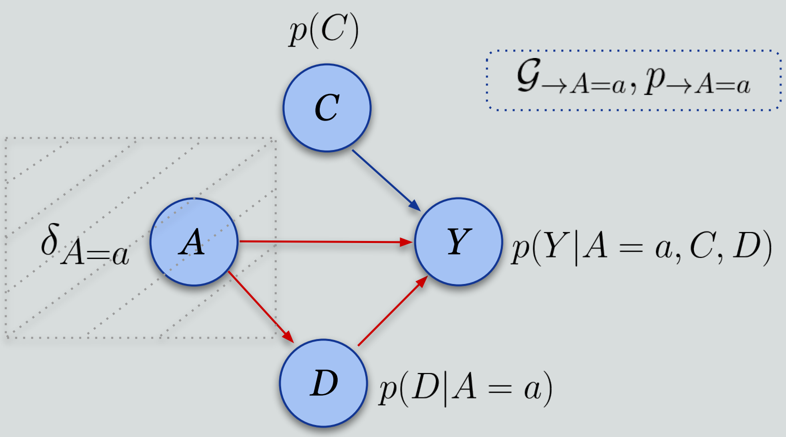
Different hard interventions on emerging links → separate causal influence along different causal paths





$$p_{\to A, A=\bar{a}\to Y, A=a\to D}(Y) = \sum_{C,D} p_{\to A, A=\bar{a}\to Y, A=a\to D}(C, D, Y)$$
$$= \sum_{C,D} (Y|A=\bar{a}, C, D)p(D|A=a)p(C)$$

Causal Influence

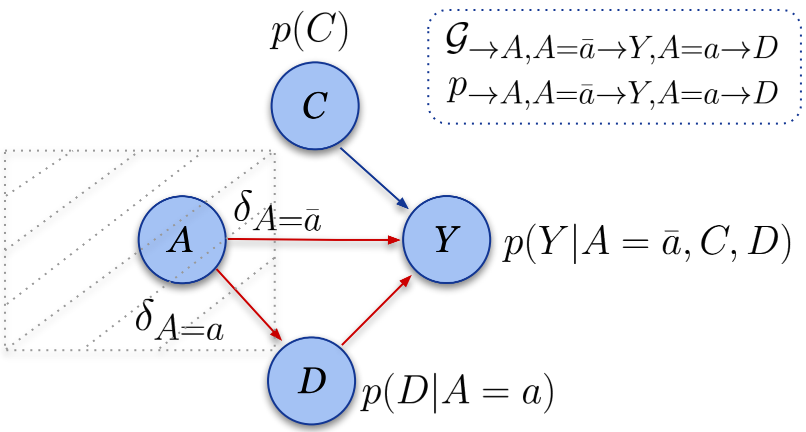


Potential Outcome Y_a
Random variable with distribution

$$p_{\mathrm{PO}}(Y_a) := p_{\to A=a}(Y)$$

$$\begin{aligned} p_{\mathrm{PO}}(Y_a) &= p_{\to A=a}(Y) \\ &= \sum_{C,D} p_{\to A=a}(C,D,Y) \\ &= \sum_{C,D} p(Y|A=a,C,D)p(D|A=a)p(C) \end{aligned}$$

Path-Specific Influence



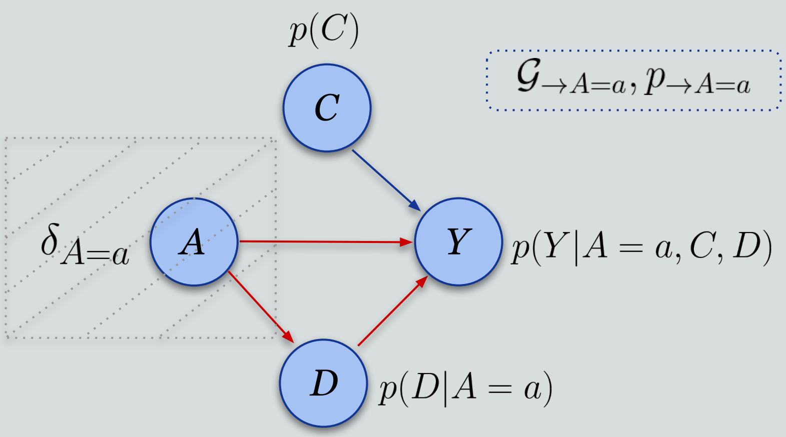
Path-Specific Potential Outcome $Y_{\bar{a}}(D_a)$

Random variable with distribution

$$p_{\mathrm{PO}}(Y_{\bar{a}}(D_a)) := p_{\to A, A=\bar{a}\to Y, A=a\to D}(Y)$$

$$p_{\mathrm{PO}}(Y_{\bar{a}}(D_a)) = \sum_{C,D} p(Y|A=\bar{a},C,D)p(D|A=a)p(C)$$

Average Treatment Effect

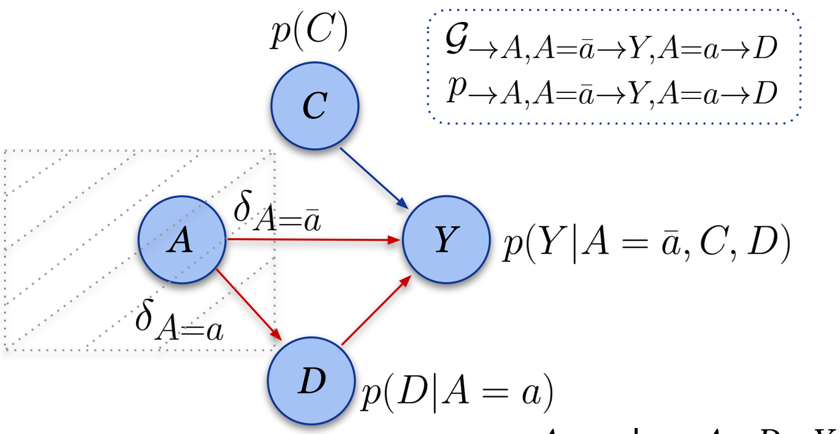


Average Treatment Effect

$$\mathrm{ATE}_{a\bar{a}} = \mathbb{E}_{p(Y_{\bar{a}})}[Y_{\bar{a}}] - \mathbb{E}_{p(Y_a)}[Y_a]$$

Measure difference of causal influence when A is set to $A=\bar{a}$ and to $A=a$

Path-Specific Effect



$A=a$ along $A\to D\to Y$

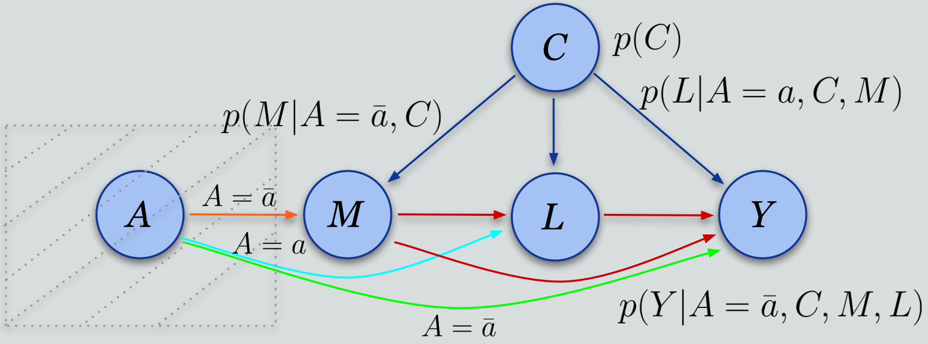
Path-Specific Effect

$$\mathrm{PSE}_{a\bar{a}} = \mathbb{E}_{p(Y_{\bar{a}}(D_a))}[Y_{\bar{a}}(D_a)] - \mathbb{E}_{p(Y_a)}[Y_a]$$

Measure difference of causal influence when A is set to $A=\bar{a}$ and to $A=a$ along $A\to Y$, whilst keeping $A=a$ along $A\to D\to Y$

Path-Specific Effect

Different hard interventions on emerging links



Path-Specific Potential Outcome $Y_{\bar{a}}(M_{\bar{a}}, L_a(M_{\bar{a}}))$

$$\mathrm{PSE}_{a\bar{a}} = \mathbb{E}_{p_{\mathrm{PO}}(Y_{\bar{a}}(M_{\bar{a}},L_a(M_{\bar{a}})))}[Y_{\bar{a}}(M_{\bar{a}}, L_a(M_{\bar{a}}))] - \mathbb{E}_{p_{\mathrm{PO}}(Y_a)}[Y_a]$$

$$p_{\mathrm{PO}}(Y_{\bar{a}}(M_{\bar{a}}, L_a(M_{\bar{a}}))) = \int_{C,M,L} p(Y|A=\bar{a},C,M,L)p(L|A=a,C,M)p(M|A=\bar{a},C)p(C)$$

Different hard interventions on different causal paths

Causal paths

$A \to M \to L \to Y$	$A=\bar{a}$	$A=\bar{a}$
$A \to M \to Y$	$A=\bar{a}$	$A=a$
$A \to L \to Y$	$A=a$	$A=a$
$A \to Y$	$A=\bar{a}$	$A=\bar{a}$

$p(M|A=a,C)$ $p(M|A=\bar{a},C)$

Path-Specific Potential Outcome $Y_{\bar{a}}(\boxed{M_a}, L_a(\boxed{M_{\bar{a}}}))$

$$\mathrm{PSE}_{a\bar{a}} = \mathbb{E}_{p_{\mathrm{PO}}(Y_{\bar{a}}(M_a,L_a(M_{\bar{a}})))}[Y_{\bar{a}}(M_a, L_a(M_{\bar{a}}))] - \mathbb{E}_{p_{\mathrm{PO}}(Y_a)}[Y_a]$$

Need Structural Causal Models to Identify PSE

Path-Specific Effect

Different hard interventions on different causal paths

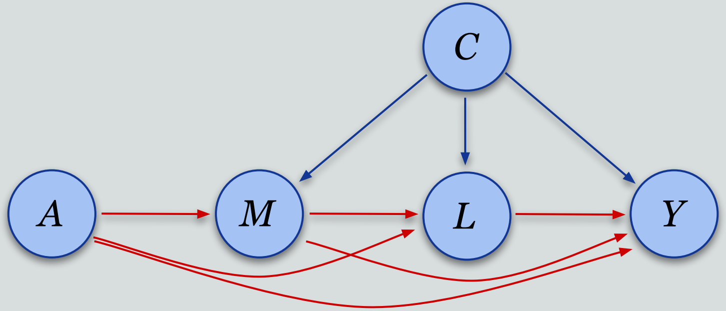
Structural Causal Models

$$\langle \mathcal{E}, V, F, p(\mathcal{E}) \rangle$$

$$\mathcal{E} = \{\mathcal{E}_{V_1}, \ldots, \mathcal{E}_{V_K}\}, p(\mathcal{E}) = \prod_{k=1}^{K} p(\mathcal{E}_{V_k})$$ Exogenous (latent)

$$F = \{f_{V_1}, \ldots, f_{V_K}\}, V = \{V_1, \ldots, V_K\}$$ Endogenous

$$V_k = f_{V_k}(\mathrm{pa}(V_k), \mathcal{E}_{V_k})$$



$$L = f_L(A, C, M, \mathcal{E}_L)$$

Intervention on L = replace the function with value l

$A \to M \to L \to Y$	$\boxed{A = \bar{a}}$
$A \to M \to Y$	$\boxed{A = a}$
$A \to L \to Y$	$A = a$
$A \to Y$	$A = \bar{a}$

$$A \sim \mathrm{Bern}(\pi), C = \epsilon_c$$

$$M = \theta^m + \theta_a^m A + \theta_c^m C + \epsilon_m \qquad \epsilon_c, \epsilon_m, \epsilon_l, \epsilon_y \sim \mathcal{N}(0,1)$$

$$L = \theta^l + \theta_a^l A + \theta_c^l C + \theta_m^l M + \epsilon_l$$

$$Y = \theta^y + \theta_a^y A + \theta_c^y C + \theta_m^y M + \theta_l^y L + \epsilon_y$$

$A \to Y$ $A = \bar{a}$

$A \to M \to Y$ $A = a$

$A \to M \to L \to Y$ $A = \bar{a}$

$A \to L \to Y$ $A = a$

$$M_a = \theta^m + \theta_a^m a + \theta_c^m C + \epsilon_m$$

$$M_{\bar{a}} = \theta^m + \theta_a^m \bar{a} + \theta_c^m C + \epsilon_m$$

$$L_a(M_{\bar{a}}) = \theta^l + \theta_a^l a + \theta_c^l C + \theta_m^l M_{\bar{a}} + \epsilon_l$$

$$Y_{\bar{a}}(M_a, L_a(M_{\bar{a}})) = \theta^y + \theta_a^y \bar{a} + \theta_c^y C + \theta_m^y M_a + \theta_l^y L_a(M_{\bar{a}}) + \epsilon_y$$

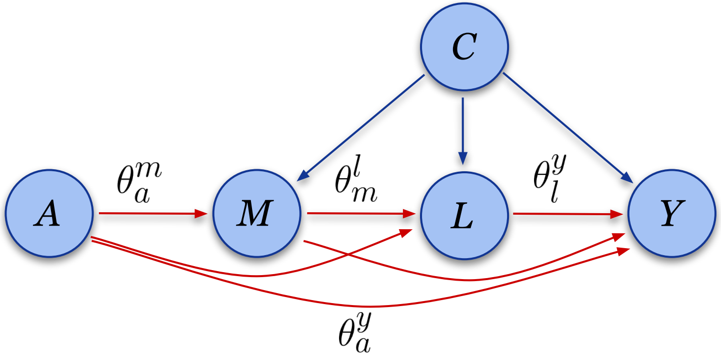
Path-Specific Effect

Different hard interventions on different causal paths

$$M_a = \theta^m + \theta^m_a a + \theta^m_c C + \epsilon_m$$
$$M_{\bar{a}} = \theta^m + \theta^m_a \bar{a} + \theta^m_c C + \epsilon_m$$
$$L_a(M_{\bar{a}}) = \theta^l + \theta^l_a a + \theta^l_c C + \theta^l_m M_{\bar{a}} + \epsilon_l$$
$$Y_{\bar{a}}(M_a, L_a(M_{\bar{a}})) = \theta^y + \theta^y_a \bar{a} + \theta^y_c C + \theta^y_m M_a + \theta^y_l L_a(M_{\bar{a}}) + \epsilon_y$$

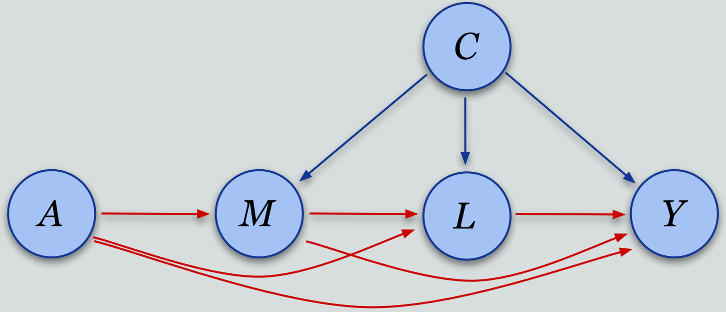
$$\mathbb{E}_{p_{\mathrm{PO}}(Y_{\bar{a}}(M_a, L_a(M_{\bar{a}})))}[Y_{\bar{a}}(M_a, L_a(M_{\bar{a}}))]$$
$$= \theta^y + \boxed{\theta^y_a \bar{a}} + \theta^y_m(\theta^m + \theta^m_a a) + \theta^y_l(\theta^l + \theta^l_a a + \theta^l_m(\theta^m + \boxed{\theta^m_a \bar{a}}))$$
$$\mathbb{E}_{p_{\mathrm{PO}}(Y_a)}[Y_a]$$
$$= \theta^y + \boxed{\theta^y_a a} + \theta^y_m(\theta^m + \theta^m_a a) + \theta^y_l(\theta^l + \theta^l_a a + \theta^l_m(\theta^m + \boxed{\theta^m_a a}))$$

$$\mathrm{PSE}_{a\bar{a}} = \mathbb{E}_{p_{\mathrm{PO}}(Y_{\bar{a}}(M_a, L_a(M_{\bar{a}})))}[Y_{\bar{a}}(M_a, L_a(M_{\bar{a}}))] - \mathbb{E}_{p_{\mathrm{PO}}(Y_a)}[Y_a]$$
$$= (\theta^y_a + \theta^y_l \theta^l_m \theta^m_a)(\bar{a} - a)$$



- **Causal effect along a path = product of coefficients**
- **Causal effect along a set of paths = sum of causal effects along all paths**

Path-Specific Counterfactual PO



$A \sim \mathrm{Bern}(\pi), C = \epsilon_c$

$M = \theta^m + \theta^m_a A + \theta^m_c C + \epsilon_m$

$L = \theta^l + \theta^l_a A + \theta^l_c C + \theta^l_m M + \epsilon_l$

$Y = \theta^y + \theta^y_a A + \theta^y_c C + \theta^y_m M + \theta^y_l L + \epsilon_y$

$A \to M \to L \to Y$	$A = \bar{a}$
$A \to M \to Y$	$A = a$
$A \to L \to Y$	$A = a$
$A \to Y$	$A = \bar{a}$

$Y_{\bar{a}}(M_a, L_a(M_{\bar{a}}))$

$p_{\mathrm{PO}}(Y_{\bar{a}}(M_a, L_a(M_{\bar{a}})))$

Observation for individual/unit *n*

$o^n = \{a^n = a, c^n, m^n, l^n, y^n\}$

$p_{\mathrm{PO}}(Y_{\bar{a}}(M_a, L_a(M_{\bar{a}})|o^n)$

$\epsilon^n = \{\epsilon^n_c, \epsilon^n_m, \epsilon^n_l, \epsilon^n_y\}$ Specific realization of exogenous variables for individual/unit *n*

$$
\begin{aligned}
c^n &= \epsilon^n_c \\
m^n &= \theta^m + \theta^m_a a + \theta^m_c c^n + \epsilon^n_m \\
l^n &= \theta^l + \theta^l_a a + \theta^l_c c^n + \theta^l_m m^n + \epsilon^n_l \\
y^m &= \theta^y + \theta^y_a a + \theta^y_c c^n + \theta^y_m m^n + \theta^y_l l^n + \epsilon^n_y
\end{aligned}
$$

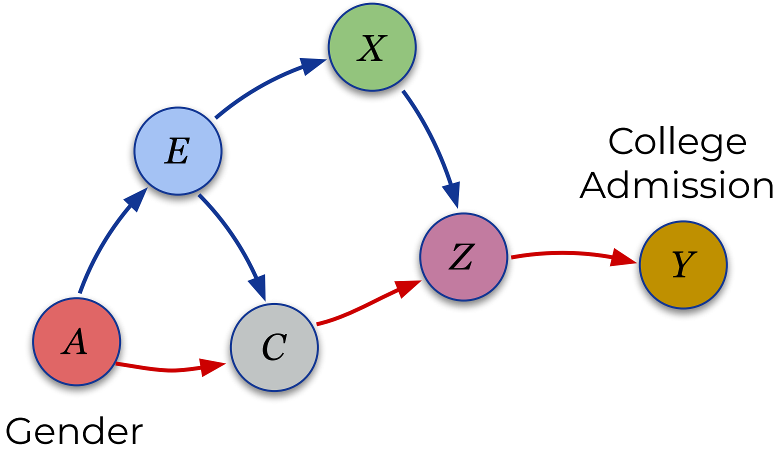
Abduction

Infer randomness from observation

$$
\begin{aligned}
\epsilon^n_c &= c^n \\
\epsilon^n_m &= m^n - (\theta^m + \theta^m_a a + \theta^m_c c^n) \\
\epsilon^n_l &= l^n - (\theta^l + \theta^l_a a + \theta^l_c c^n + \theta^l_m m^n) \\
\epsilon^n_y &= (y^m - (\theta^y + \theta^y_a a + \theta^y_c c^n + \theta^y_m m^n + \theta^y_l l^n)
\end{aligned}
$$

$$p_{\mathrm{PO}}(Y_{\bar{a}}(M_a, L_a(M_{\bar{a}})|o^n) = p_{\mathrm{PO}}(Y_{\bar{a}}(M_a, L_a(M_{\bar{a}})|a, \epsilon^n)$$

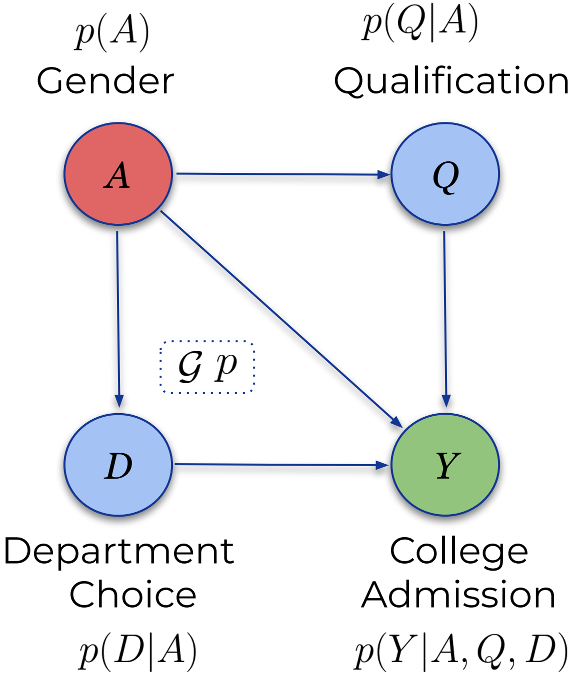
$$
\begin{aligned}
M_a|\{a, \epsilon^n\} &= \theta^m + \theta^m_a a + \theta^m_c \epsilon^n_c + \epsilon^n_m = m^n \\
M_{\bar{a}}|\{a, \epsilon^n\} &= \theta^m + \theta^m_a \bar{a} + \theta^m_c \epsilon^n_c + \epsilon^n_m = m^n + \theta^m_a(\bar{a} - a) \\
L_a(M_{\bar{a}})|\{a, \epsilon^n\} &= \theta^l + \theta^l_a a + \theta^l_c \epsilon^n_c + \theta^l_m(m^n + \theta^m_a(\bar{a} - a)) + \epsilon^n_l = l^n + \theta^l_m \theta^m_a(\bar{a} - a) \\
Y_{\bar{a}}(M_a, L_a(M_{\bar{a}}))|\{a, \epsilon^n\} &= \theta^y + \theta^y_a \bar{a} + \theta^y_c \epsilon^n_c + \theta^y_m m^n + \theta^y_l(l^n + \theta^l_m \theta^m_a(\bar{a} - a)) + \epsilon^n_y \\
&= y^n + (\theta^y_a + \theta^y_l \theta^l_m \theta^m_a)(\bar{a} - a)
\end{aligned}
$$



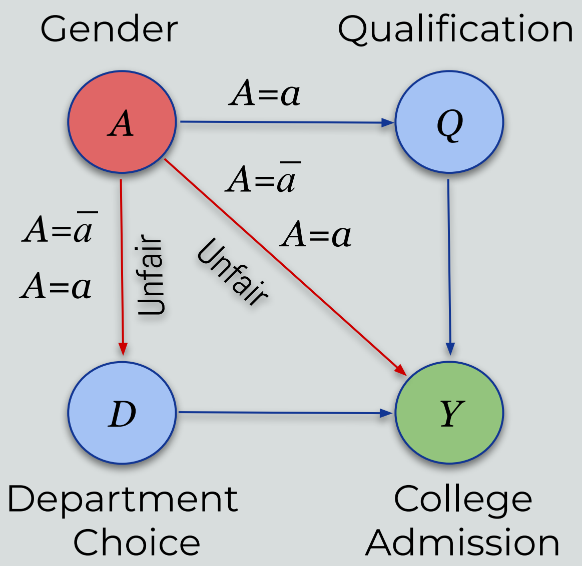
CBNs and CI for Measuring Unfairness Underlying a Dataset

- [A Causal Bayesian Networks Viewpoint on Fairness. S. Chiappa, W. Isaac, 2019.]
- [Fairness in Machine Learning. L Oneto, S. Chiappa, 2020.]

College Admission Example
Measuring Unfairness



Path-Specific Fairness: Measure Influence of A on Y along $A \rightarrow Y$ and $A \rightarrow D \rightarrow Y$ at the Population Level



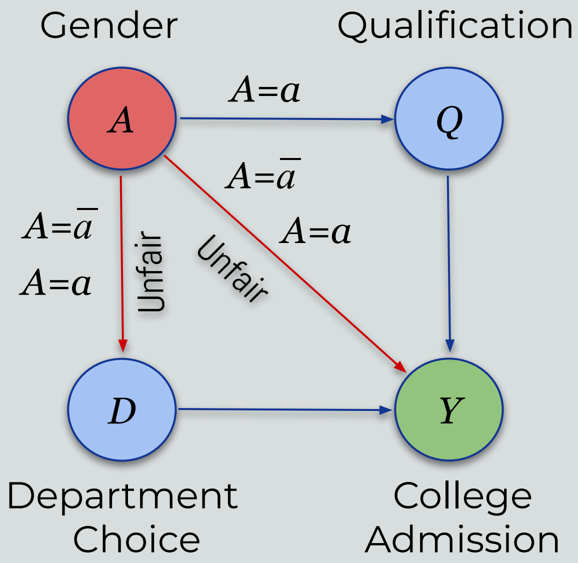
A=a indicates female applicant
A=$\bar{a}$ indicates male applicant

Path-specific potential outcome $Y_{\bar{a}}(Q_a, D_{\bar{a}})$

$$\mathrm{PSE}_{a\bar{a}} = \mathbb{E}_{p_{\mathrm{PO}}(Y_{\bar{a}}(Q_a, D_{\bar{a}}))}[Y_{\bar{a}}(Q_a, D_{\bar{a}})] - \mathbb{E}_{p_{\mathrm{PO}}(Y_a)}[Y_a]$$

Measures difference in expectation when setting the value of A to male along $A \rightarrow Y$ and $A \rightarrow D \rightarrow Y$ and to female along $A \rightarrow Q \rightarrow Y$ wrt to setting it to female along all causal paths

Path-Specific Counterfactual Fairness: Measure Influence of A on Y along $A{\rightarrow}Y$ and $A{\rightarrow}D{\rightarrow}Y$ for a Specific Individual



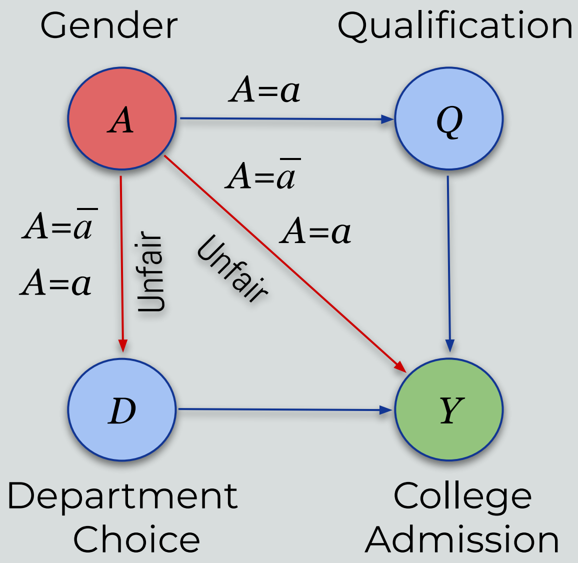
$A{=}a$ indicates female applicant
$A{=}\bar{a}$ indicates male applicant

By conditioning $Y_{\bar{a}}(Q_a, D_{\bar{a}})$ on observation

from a female individual $\{a^n = a, q^n, d^n, y^n\}$ we answer the counterfactual question of whether the individual would have been admitted had she being male only along $A{\rightarrow}Y$ and $A{\rightarrow}D{\rightarrow}Y$

$$p_{\text{PO}}(Y_{\bar{a}}(Q_a, D_{\bar{a}}) | a^n = a, q^n, d^n, y^n)$$

Path-Specific Counterfactual Fairness: Measure Influence of A on Y along $A \rightarrow Y$ and $A \rightarrow D \rightarrow Y$ for a Specific Individual

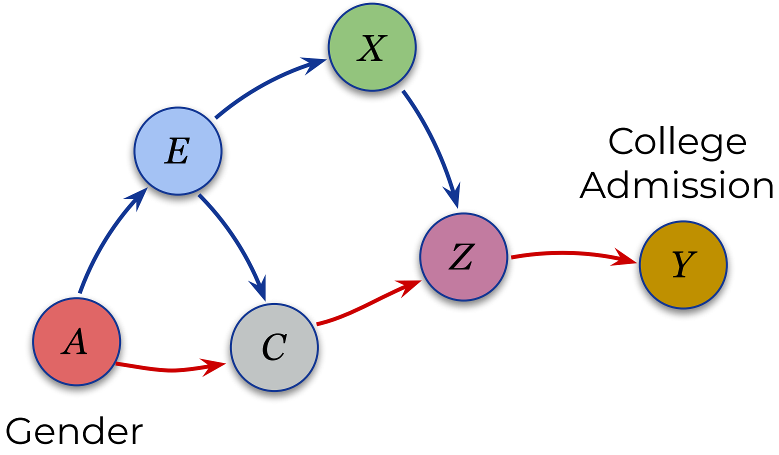


A=a indicates female applicant
A=$\bar{a}$ indicates male applicant

$$A \sim \mathrm{Bern}(\pi)$$
$$Q = \theta^q + \theta^q_a A + \epsilon_q$$
$$D = \theta^d + \theta^d_a A + \epsilon_d$$
$$Y = \theta^y + \theta^y_a A + \theta^y_q Q + \theta^y_d D + \epsilon_y$$

$$Q_a|o^n = \theta^q + \theta^q_a a + \epsilon^n_q$$
$$D_{\bar{a}}|o^n = \theta^d + \theta^d_a \bar{a} + \epsilon^n_d$$
$$Y_{\bar{a}}(Q_a, D_{\bar{a}})|o^n = \theta^y + \theta^y_a \bar{a} + \theta^y_q q^n + \theta^y_d(d^n + \theta^d_a(\bar{a} - a)) + \epsilon^n_y$$

Useful viewpoint
Modification of y^n obtained by modifying part of the observation for which the value of A is different from the one observed



CBNs and CI for Developing Fair Prediction Models

- Path-Specific Counterfactual Fairness. S. Chiappa, 2019.
- A General Approach to Fairness with Optimal Transport. S. Chiappa, R. Jiang, T. Stepleton, A. Pacchiano, H. Jiang, J. Aslanides, 2020.

- Graphical Conditions for Introduced Unfairness: Why Fair Labels Can Yield Unfair Predictions. C. Ashurst, R. Carey, S. Chiappa, T. Everitt, 2022

Difference with Measuring Unfairness in Datasets

1

Need to not absorb bias in data in the model

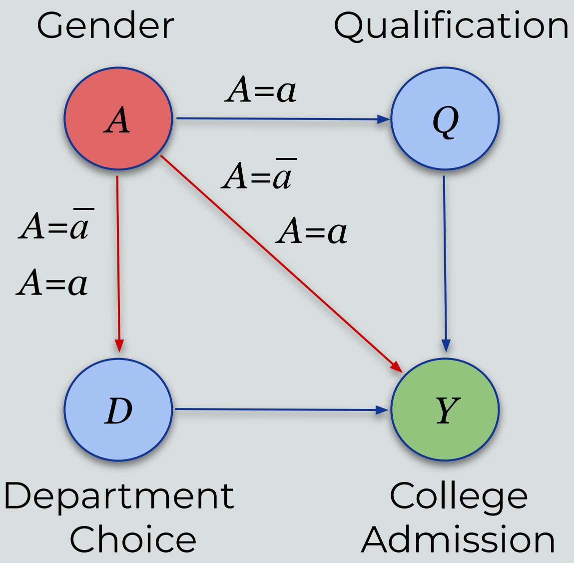
2

Need to maintain accuracy

3

Training and deployment settings are different

Path-Specific Counterfactually Fair Predictor

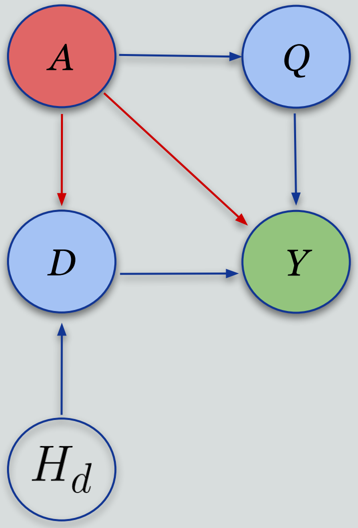


- Training: Objective function that estimates *p*
- Deployment: Observe $\{a^n = a, q^n, d^n\}$ and assign outcome as mean of

$$\hat{p}_{\mathrm{PO}}(Y_{\bar{a}}(Q_a, D_{\bar{a}})|a^n = a, q^n, d^n)$$

Need to enforce independence in the latent space

Path-Specific Counterfactually Fair Predictor

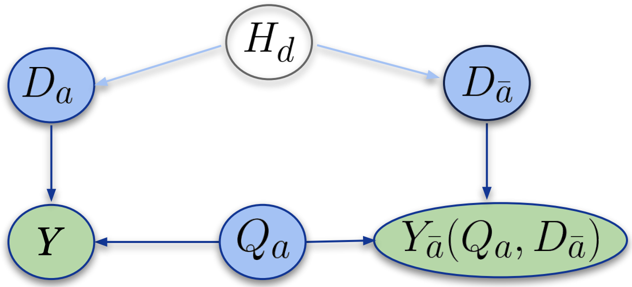


$$A \sim \mathrm{Bern}(\pi)$$
$$Q \sim p_\theta^q(\cdot|A)$$
$$H_d \sim p_\theta^h(\cdot),\ D \sim p_\theta^d(\cdot|A, H_d)$$
$$Y \sim p_\theta^y(\cdot|A, Q, D)$$

Represent conditionals using neural networks

$$Q_a \sim p_\theta^q(\cdot|A = a)$$
$$H_d \sim p_\theta^h(\cdot),\ D_{\bar{a}} \sim p_\theta^d(\cdot|A = \bar{a}, H_d)$$
$$Y_{\bar{a}}(Q_a, D_{\bar{a}}) \sim p_\theta^y(\cdot|A = \bar{a}, Q_a, D_{\bar{a}})$$

Deployment



$$\{a^n = a, q^n, d^n\}$$

$h_d^{n,i}$ $i = 1, \ldots, I$ from $p_\theta(H_d|A, Q, D)$ Abduction

$d_{\bar{a}}^{n,i} \sim p_\theta^d(\cdot|\bar{a}, h_d^{n,i})$ Action-Prediction

$$y_{\bar{a}}^n = \frac{1}{I}\sum_{i=1}^{I} \mathbb{E}_{p_\theta^y(Y_{\bar{a}}(Q_a, D_{\bar{a}})|\bar{a}, q^n, d_{\bar{a}}^{n,i})}[Y_{\bar{a}}(Q_a, D_{\bar{a}})]$$

A-Independent Representation H_d with Variational Autoencoder

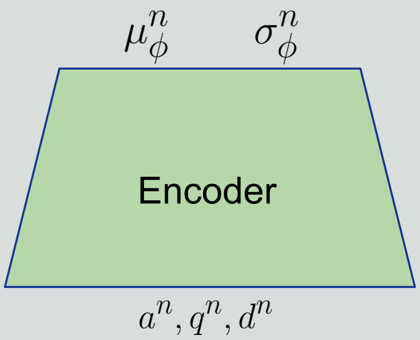
Intractable $p_\theta(H_d|A,Q,D)$

Variational approximation

$$p_\theta(H_d|A,Q,D) \approx q_\phi(H_d|A,Q,D) = \mathcal{N}(\mu_\phi, \sigma_\phi)$$

$$H_d = \mu_\phi + \sigma_\phi \epsilon, \;\; q_\epsilon = \mathcal{N}(0,1)$$

$\epsilon^i \sim \mathcal{N}(0,1)$ $\boxed{h_d^{n,i} = \mu_\phi^n + \sigma_\phi^n \epsilon^i}$

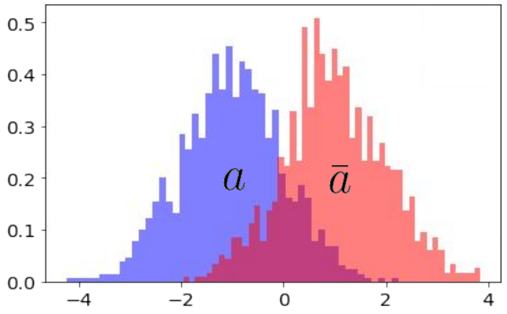


Maximise lower bound on the marginal log-likelihood with a penalty term for enforcing independence on A

$$\begin{aligned}\mathcal{F}_{\theta,\phi} = \;& \mathbb{E}_{q_\phi(H_d|A,Q,D)}[\log p_\theta(A,Q,D,Y,H_d)] \\ & - \mathbb{E}_{q_\phi(H_d|A,Q,D)}[\log q_\phi(H_d|A,Q,D)] \\ & - \beta \mathrm{MMD}(a, \bar{a})\end{aligned}$$

β weighting factor that determines the degree of independence

Empirical distributions of $h_d^{n,i}$ for male and female individuals



CBNs and CI for ML Fairness

Formalization

Unified Framework

Complex Unfairness Scenarios

Unfairness underlying Data

- CBNs for describing different possible unfairness scenarios underlying data
- CBNs for measuring unfairness underlying data

Fair Prediction Models

- CBNs for developing fair prediction models
- Causal influence diagrams for auditing fair prediction models

CBNs and CI for ML Fairness

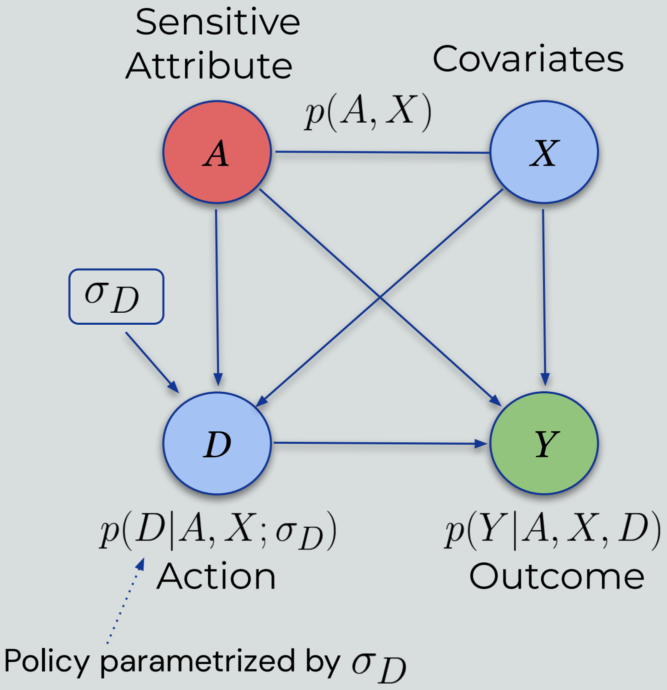
Causal Decision Making

Beyond Prediction Models

- Design optimized policies under fairness constraint

→ *Pragmatic Fairness: Developing Policies with Outcome Disparity Control. L. Gultchin, S. Guo, A. Malek, S. Chiappa, R. Silva, 2022.*

Optimized Policies under Fairness Constraints



Goal

Design a new decision making system that specifies how to select actions D that maximize downstream outcome Y subject to some fairness constraints

Extended view wrt designing fair prediction systems

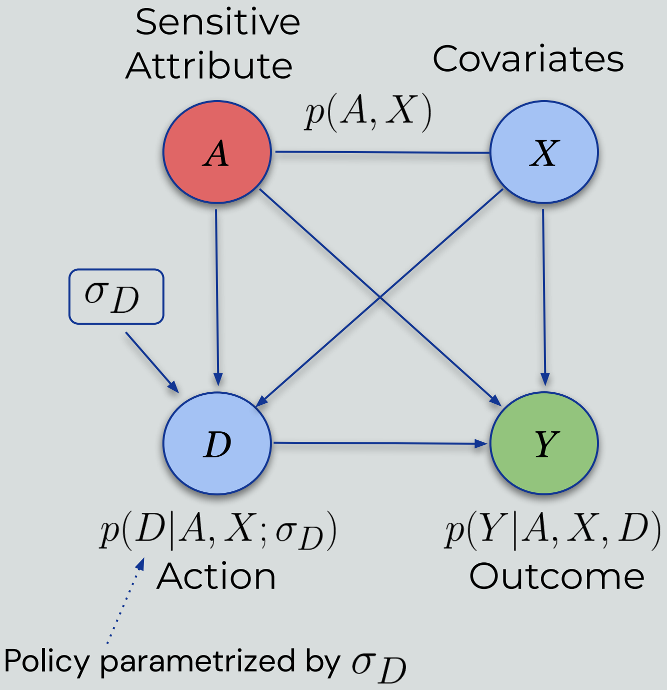
Distinguish between our choice of policy or action allocation (D), and the downstream effect we *truly* aim for (Y)

Granting Loans/Offering College Admission Examples

We would like to be fair with respect to desirable downstream outcomes, such as repayment or academic and economic success, and not just the very allocation of the action

We are not aiming to match a distribution of historical outcomes but to optimize a future expected utility

Optimized Policies under Fairness Constraints



Setting

- A and X are allowed to be associated and potentially directly influence Y
- D can only indirectly control this influence

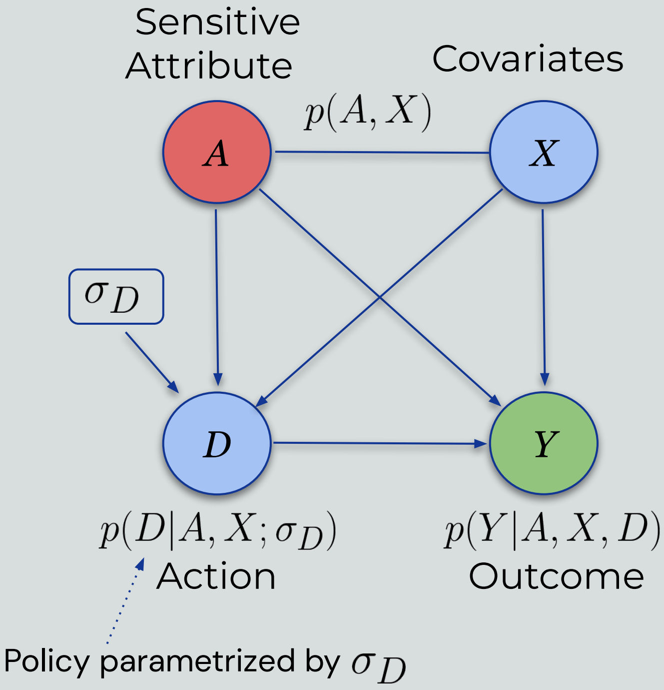
This setting formalizes a situation in which control for association of A and Y can only be achieved to some extent through a predefined set of available actions, i.e., a given action space. The level to which we can minimize such unfair impact therefore depends on the choice of the action space

Outreach Campaign Example

Company looking to change its outreach campaign to mitigate imbalances in the demographics A and potentially associated level of experience X of job applicants Y

The company cannot control factors such as cultural preference among applicants for industry sectors, but can induce modifications such as focusing recruiting efforts in events organized by minority groups in relevant conferences, thus optimizing for the things we can control

Optimized Policies under Fairness Constraints



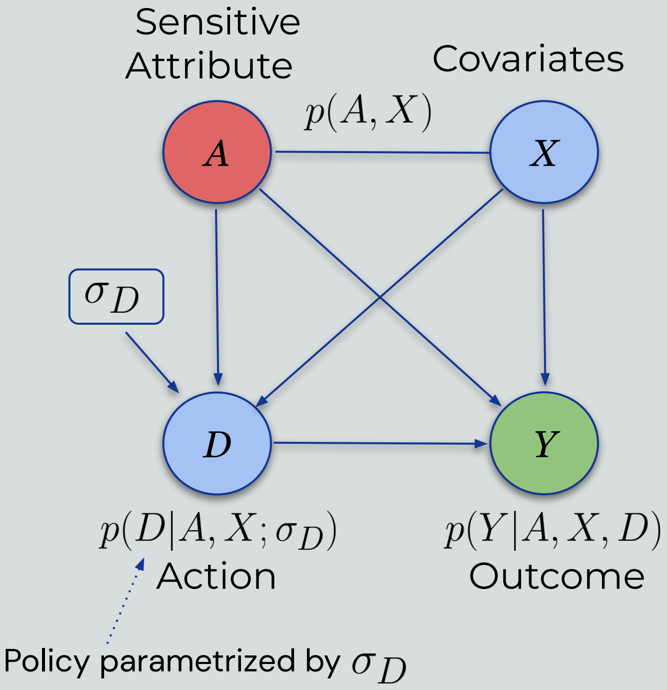
Setting

- Observational-data regime: learn optimized policy based on historical data collected using a baseline policy $\sigma_\emptyset$, representing the action allocation that was in place during the collection of the data, i.e. the "status-quo" pre-optimization.
  Reflects considerations such as the cost, ethical constraints, or difficulty of collecting interventional data.

$$\mathcal{D} = \{a^i, x^i, d^i, y^i\}_{i=1}^N$$

$$(a^i, x^i, d^i, y^i) \sim p(A, X, D, Y; \sigma_\emptyset)$$

Optimized Policies under Fairness Constraints



Formalization

Find policy $\mu^D_{\sigma_D}(a, x) = \mathbb{E}[D|a, x; \sigma_D]$ that maximizes utility $\mathbb{E}[Y_{\sigma_D}]$ while controlling for disparity

Moderation Breaking Constraint

Remove influence of A on $\mathbb{E}[Y_{\sigma_D}]$ as much as possible what we can control: the allocation of D as determined by $p(D|A, X; \sigma_D)$.

Express a possible modulation of the effect of the action on the outcome by the sensitive attribute via the decomposition

$$\mu^Y(a, x, d) := \mathbb{E}[Y|a, x, d]$$
$$= f(a, x) + g(a, x, d) + h(x, d)$$

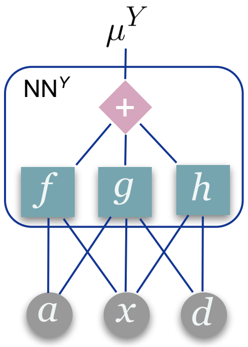
Optimized Policies under Fairness Constraints

$$\mu^Y_{\sigma_D}(a,x) := \mathbb{E}[Y_{\sigma_D}|a,x] = \int_d \mu^Y(a,x,d)p(d|a,x;\sigma_D)$$
$$= f(a,x) + g_{\sigma_D}(a,x) + h_{\sigma_D}(x)$$

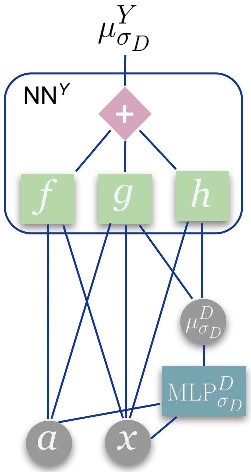
Objective

$$\arg\max_{\sigma_D} \mathbb{E}[Y_{\sigma_D}] \quad \text{s.t.}$$
$$(\mathbb{E}[g_{\sigma_D}(a,X) \mid A=a] - \mathbb{E}[g_{\sigma_D}(\bar{a},X) \mid A=\bar{a}])^2 \leq \epsilon, \forall a, \bar{a}$$

Model $\mu^Y(a,x,d)$ using a structured neural network that separates into the three components reflecting the decomposition



Model $\mu^D_{\sigma_D}(a,x) = \mathbb{E}[D|a,x;\sigma_D]$ using a MLP neural network



Teal blocks: parameter layers
Light green blocks: fixed parameters.
Purple diamonds: additive gates

Phase I

Learn parameters of NN^Y using

Phase II

Learn parameters of $\mathrm{MLP}^D_{\sigma_D}$ to learn the objective with

$$\mathbb{E}[Y_{\sigma_D}] \approx \frac{1}{N}\sum_{i=1}^{N} \mu^Y_{\sigma_D}(a^i, x^i)$$

DeepMind

The end and thank you