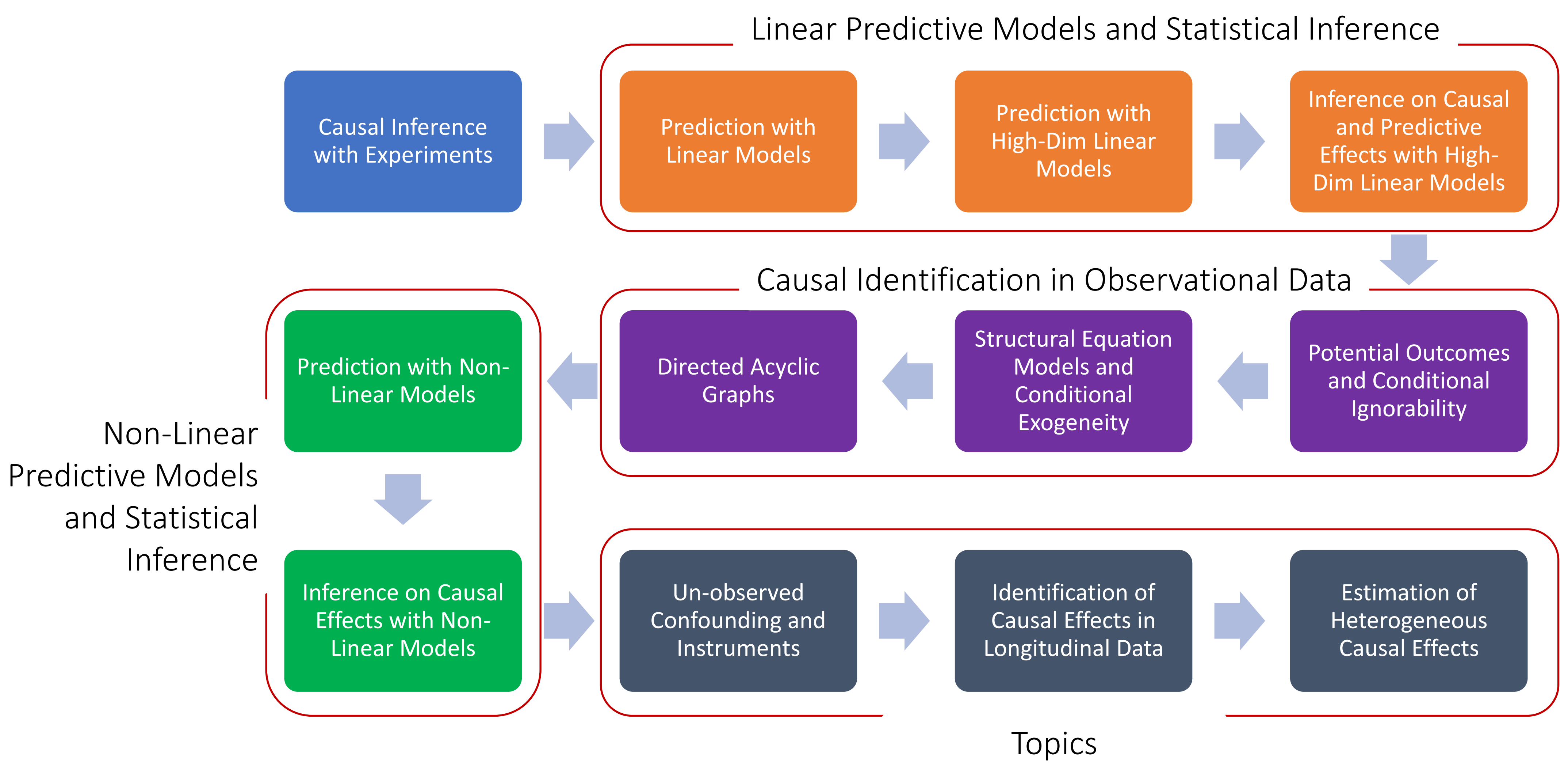
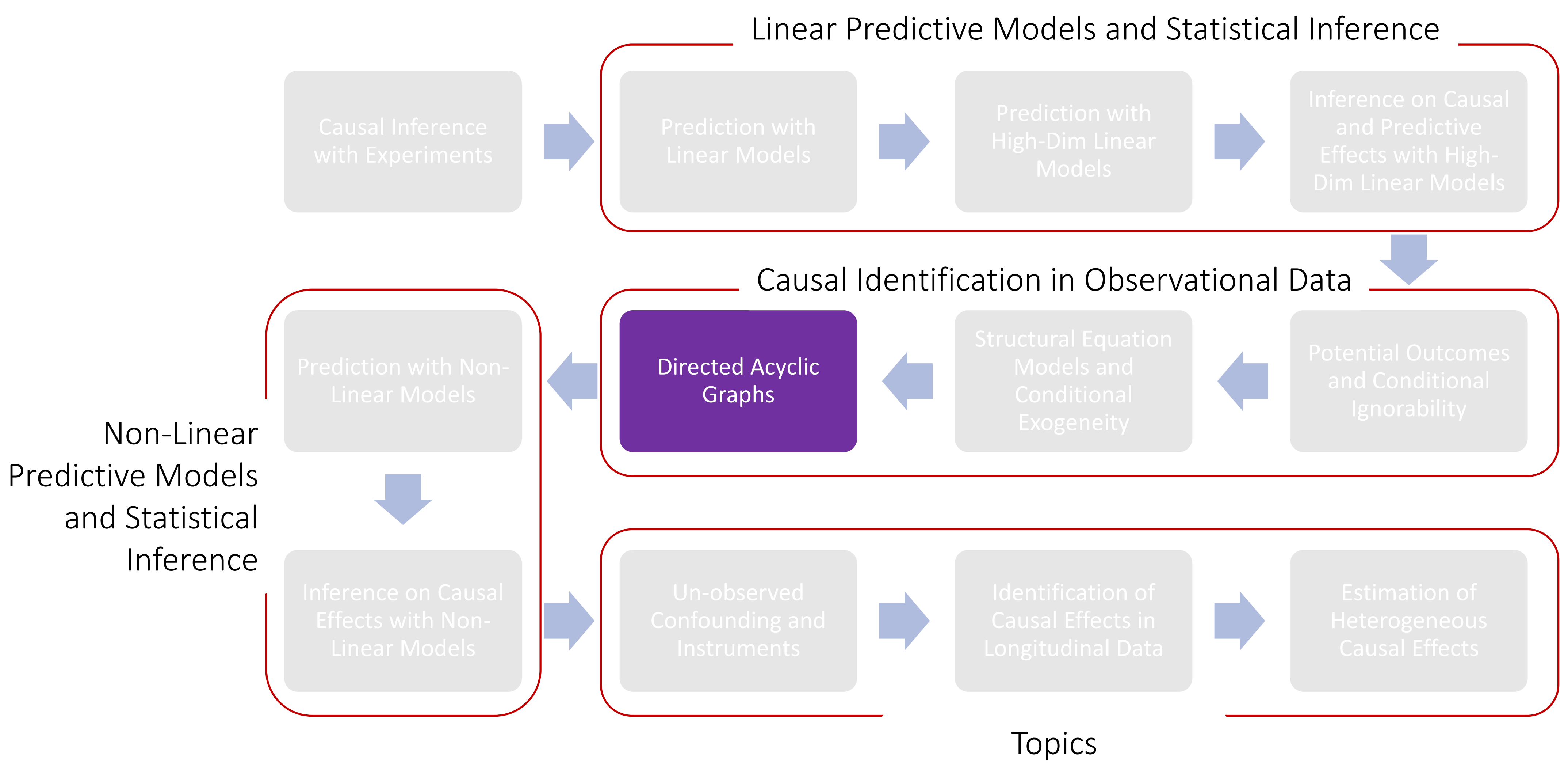
MS&E 228: Directed Acyclic Graphs and Non-Linear SEMs

Vasilis Syrgkanis

MS&E, Stanford


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Prediction with Non-Linear Models
Directed Acyclic Graphs
Structural Equation Models and Conditional Exogeneity
Potential Outcomes and Conditional Ignorability
Non-Linear Predictive Models and Statistical Inference
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Identification of Causal Effects in Longitudinal Data
Estimation of Heterogeneous Causal Effects
Topics


Linear Predictive Models and Statistical Inference
Causal Inference with Experiments
Prediction with Linear Models
Prediction with High-Dim Linear Models
Inference on Causal and Predictive Effects with High-Dim Linear Models
Causal Identification in Observational Data
Prediction with Non-Linear Models
Directed Acyclic Graphs
Structural Equation Models and Conditional Exogeneity
Potential Outcomes and Conditional Ignorability
Non-Linear Predictive Models and Statistical Inference
Inference on Causal Effects with Non-Linear Models
Un-observed Confounding and Instruments
Identification of Causal Effects in Longitudinal Data
Estimation of Heterogeneous Causal Effects
Topics

Goals for Today

- Learn the “language” of Directed Acyclic Graphs (DAGs) and their associated non-linear structural equation models (SEMs)
- Introduce “intervention” concepts “do” and “fix”
- Introduce d-separation and conditional independence in DAGs
- Proof sketch of fundamental theorem d-separation $\Rightarrow$ conditional ind.

Next lecture

- Graphical criteria for selection of adjustment set
- Crash course on good and bad “controls”

DAGs

Judea Pearl. ‘Causal diagrams for empirical research’. In: *Biometrika* 82.4 (1995), pp. 669–688 (cited on page 30).

Trygve Haavelmo. ‘The probability approach in econometrics’. In: *Econometrica: Journal of the Econometric Society* 12 (1944), pp. iii–vi+1–115 (cited on pages 30, 32).

James Heckman and Rodrigo Pinto. ‘Causal analysis after Haavelmo’. In: *Econometric Theory* 31.1 (2015 (NBER 2013)), pp. 115–151 (cited on pages 30, 35).

James Robins. ‘A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect’. In: *Mathematical modelling* 7.9-12 (1986), pp. 1393–1512 (cited on page 54).

Thomas S. Richardson and James M. Robins. *Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality*. Working Paper

Non-Linear SEMs

Non-Linear SEMs

- Last time we looked at linear SEMs
- The language of SEMs does not really rely on the linearity assumption

- For example, the Triangular Structural Equation (TSEM)

$$Y \coloneqq \delta P + \beta' X + \epsilon_Y$$
$$P \coloneqq \nu' X + \epsilon_P$$
$$X \coloneqq \epsilon_X$$

Non-Linear SEMs

- Last time we looked at linear SEMs
- The language of SEMs does not really rely on the linearity assumption

- For example, the Triangular Structural Equation (TSEM)
- Can be made non-linear

$$
\begin{aligned}
Y &\coloneqq f_Y(P, X, \epsilon_Y) \\
P &\coloneqq f_P(X, \epsilon_P) \\
X &\coloneqq f_X(\epsilon_X)
\end{aligned}
$$

- While we still maintain that $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent “exogenous” shocks

Non-Linear SEMs

- Non-Linear Triangular Structural Equation (TSEM)

Endogenously determined by the structural model

$$\boxed{Y} := f_Y(P, X, \boxed{\epsilon_Y})$$
$$\boxed{P} := f_P(X, \boxed{\epsilon_P})$$
$$\boxed{X} := f_X(\boxed{\epsilon_X})$$

Exogenously determined "outside" of the model

- $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent "exogenous" shocks
- The functions f_Y, f_P, f_X are deterministic "structural functions"
- Instead of "structural parameters" we now have "structural functions"
- Moreover, the dimension of exogenous shocks is un-restricted
- Note that the TSEM implies: $\epsilon_Y \perp\!\!\!\perp P, X$ and $\epsilon_P \perp\!\!\!\perp X$

Non-Linear SEMs

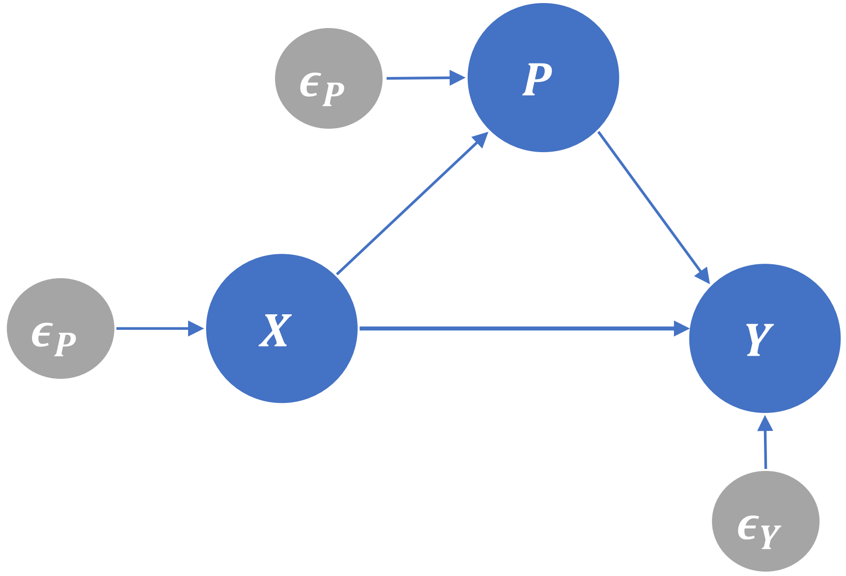
- Non-Linear Triangular Structural Equation (TSEM)

Endogenously determined by the structural model

$$\boxed{Y} := f_Y(P, X, \boxed{\epsilon_Y})$$
$$\boxed{P} := f_P(X, \boxed{\epsilon_P})$$
$$\boxed{X} := f_X(\boxed{\epsilon_X})$$

Exogenously determined “outside” of the model

- $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent “exogenous” shocks



Non-Linear SEMs

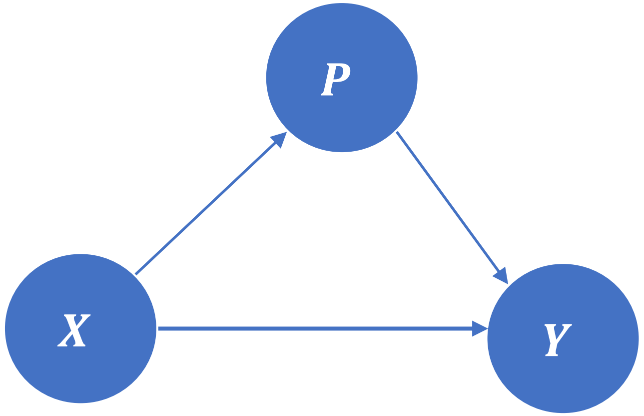
- Non-Linear Triangular Structural Equation (TSEM)

Endogenously determined by the structural model

$$\boxed{Y} := f_Y(P, X, \boxed{\epsilon_Y})$$
$$\boxed{P} := f_P(X, \boxed{\epsilon_P})$$
$$\boxed{X} := f_X(\boxed{\epsilon_X})$$

Exogenously determined “outside” of the model

- $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent “exogenous” shocks; typically omitted from DAG visualization



Non-Linear SEMs

- Non-Linear Triangular Structural Equation (TSEM)

Endogenously determined by the structural model

$$\boxed{Y} := f_Y(P, X, \boxed{\epsilon_Y})$$
$$\boxed{P} := f_P(X, \boxed{\epsilon_P})$$
$$\boxed{X} := f_X(\boxed{\epsilon_X})$$

Exogenously determined "outside" of the model

- $\epsilon_Y, \epsilon_P, \epsilon_X$ are independent "exogenous" shocks; typically omitted from DAG visualization
- A TSEM is simply a statistical "generative" model that determines a distribution over observed random variables (*c.f. Neural-Causal Models in further reading)

Structural Form

- TSEM is “structural” in that it is endowed with the following properties
- Made up of a collection of stochastic potential outcome processes indexed by (p, x)

$$Y(p, x) \coloneqq f_Y(p, x, \epsilon_Y)$$
$$P(x) \coloneqq f_P(x, \epsilon_P)$$
$$X \coloneqq f_X(\epsilon_X)$$

- **Exogeneity:** $\epsilon_P, \epsilon_Y, \epsilon_X$ are independent “shock” variables generated outside of the model
- **Consistency:** endogenous variables (Y, P, X) generated by recursive substitutions

$$Y \coloneqq Y(P, X), \qquad P \coloneqq P(X), \qquad X \coloneqq \epsilon_X$$

- **Invariance:** structure remains invariant to changes of distributions of shocks

Link to Potential Outcomes

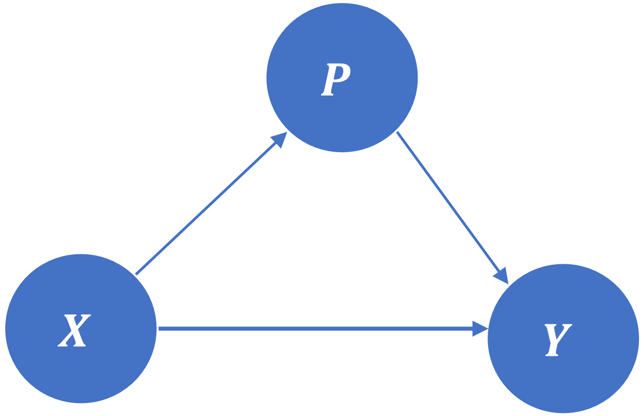
- Define arbitrary random potential outcomes
$$Y(p,x), P(x), X(.)$$
- Which are independent random processes
$$\{Y(p,x)\}_{p,x} \perp \{P(x)\}_x \perp X(.)$$
- Each potential outcome process can be represented as a SEM
- We can define $\epsilon_Y \coloneqq \epsilon_Y(p,x) = Y(p,x)$ and
$$f_Y(p,x,\epsilon_Y) = \epsilon_Y(p,x) = Y(p,x)$$

The Language of Interventions and Intervention Counterfactuals

Do Interventions: $do(P = p)$

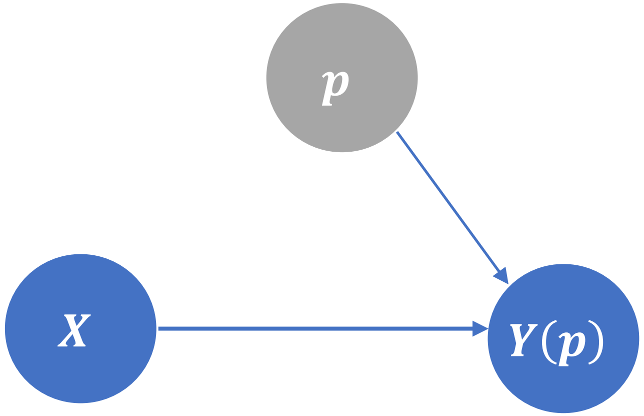
Original Data Generative Model

$$
\begin{aligned}
Y &:= f_Y(P, X, \epsilon_Y) \\
P &:= f_P(X, \epsilon_P) \\
X &:= \epsilon_X
\end{aligned}
$$



Data Generative Model under $do(P = p)$

$$
\left.\begin{array}{l} Y \\ P \\ X \end{array}\right| do(P = p) \begin{array}{l} := f_Y(p, X, \epsilon_Y) \\ := p \\ := \epsilon_X \end{array}
$$



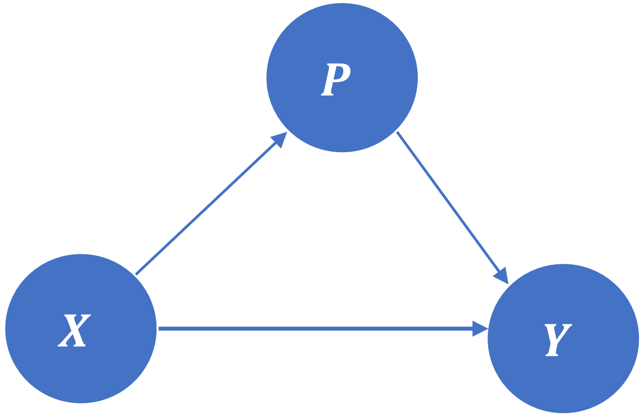
Interventions

- Do-interventions is only one way of defining counterfactuals
- We can define any type of counterfactual by simply changing one of the equations to something else
- Wright in his seminal work in ‘28 defined an intervention where the demand equation was replaced by another one that reflects a tax hike
- We can also define “soft-interventions”: increase price by 10% of its current value
- Another useful variant of do-interventions does not replace the treatment equation are “fix” interventions

Fix Interventions: $\mathrm{fix}(P = p)$

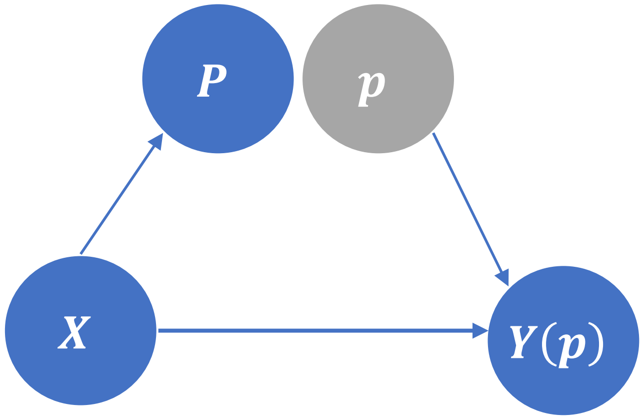
Original Data Generative Model

$$\begin{aligned} Y &:= f_Y(P, X, \epsilon_Y) \\ P &:= f_P(X, \epsilon_P) \\ X &:= \epsilon_X \end{aligned}$$



Data Generative Model under $\mathrm{fix}(P = p)$

$$\left. \begin{array}{l} Y \\ P \\ X \end{array} \right| \mathrm{fix}(P = p) \begin{array}{l} := f_Y(p, X, \epsilon_Y) \\ := f_P(X, \epsilon_P) \\ := \epsilon_X \end{array}$$



Fix Interventions

- A fix intervention is a form of “localized” do intervention
- We are only fixing the value of P in the structural equation for Y
- The random variables generated by the fix intervention are the triplets
$$(Y(p), P, X)$$
- The intervention does not affect the P, X equations nor the distribution of the exogenous shock ϵ_Y in the outcome equation

Conditional Ignorability and Mean Intervention Counterfactuals

- Identification by conditioning and conditional ignorability extends to average intervention counterfactuals
- Recall to do identification by conditioning we need for a set S
$$Y(p) \perp\!\!\!\perp P \mid S$$

- Then predictive response equals interventional response
$$E[Y(p) \mid S] = E[\,Y(p) \mid P = p, S\,] = E[\,Y(P) \mid P = p, S\,] = E[\,Y \mid P = p, S\,]$$
- Average predictive response equals average interventional response
$$E[Y(p)] = E\big[E[Y|P = p, S]\big]$$

Single World Intervention Graphs

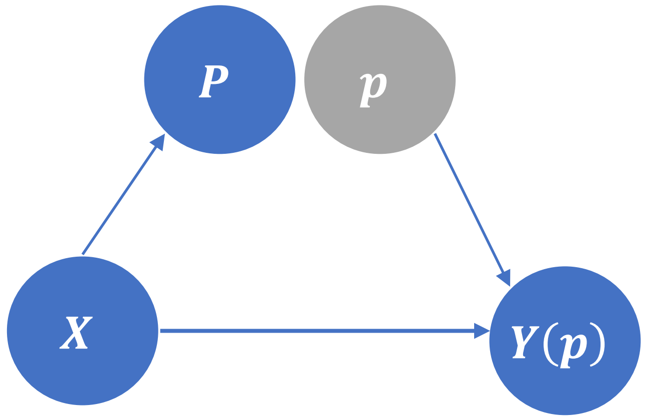
- The graphs that represent the generative model under a fix intervention

- Easy to verify visually that
$$Y(p) \perp\!\!\!\perp P \mid X$$

- Then we can do identification based on conditional ignorability

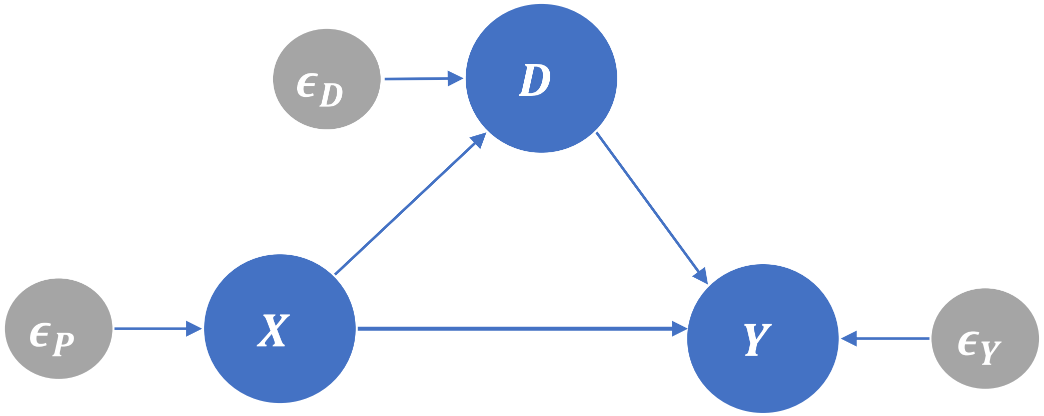
Data Generative Model under $\mathrm{fix}(P = p)$

$$\left.\begin{array}{l} Y \\ P \\ X \end{array}\right| \mathrm{fix}(P = p) \quad \begin{array}{l} := f_Y(p, X, \epsilon_Y) \\ \ f_P(X, \epsilon_P) \\ := \epsilon_X \end{array}$$



Single World Intervention Graph

Non-Linear versions of structural equation models are equivalent to Directed Acyclic Graphs



Exogenously determined "outside" of the model

Endogenously determined by the structural model

For any DAG, we can write ASEM

$$\boxed{X_j} := f_j(\text{Parents}_j, \epsilon_j) = f_j(\text{Pa}_j, \boxed{\epsilon_j})$$

Shocks ϵ_j are jointly independent and independent of $\{X_j\}$

Corresponding structural response functions

$$\boxed{X_j(pa_j)} := \boxed{f_j}(\boxed{pa_j}, \boxed{\epsilon_j})$$

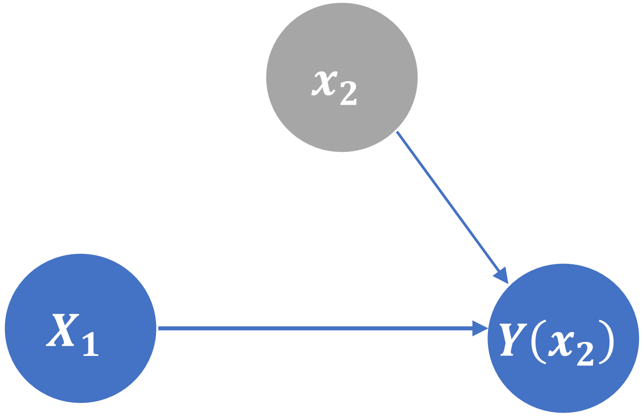
Shocks can be multi-dimensional e.g. separate shock variable per parental value

Potential/Counterfactual Outcome Processes

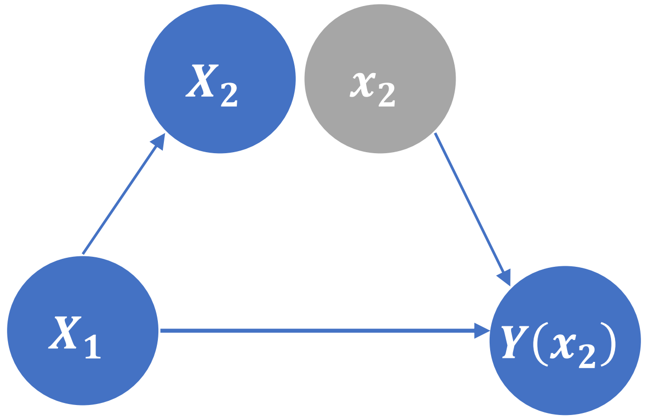
Structural Response Function

Potential values of parents

Fix Interventions $\mathrm{fix}(X_j = x_j)$
Locally replace X_j in every RHS of a structural equation with x_j. Leave as-is structural response of X_j. Also measures potential outcome $Y(x_j)$



Fix intervention visually represented as **SWIG** $\tilde{G}(x_j)$. Depicts potential outcome $Y(x_j)$ and original variable X_j on the same graph



If we can check $Y(x_j) \perp X_j \mid S$ based on the SWIG, we can identify $E[Y(x_j)]$ via conditioning

Graphical Criteria for Conditional Independence

DAGs Encode Factorization of Probability

- Graph implies factorization of the probability law

$$p(y, d, x, z) = p(y|x, d)p(d|x, z)p(x)p(z)$$

Proof.

$$\begin{tikzcd}
Z \arrow[rr] & & D \arrow[rr] & & Y \\
& & & X \arrow[ul] \arrow[ur] &
\end{tikzcd}$$

- By Bayes rule: $p(y, d, x, z) = p(y|d, x, z)p(d, x, z)$
- From graph: $p(y|d, x, z) = p(y|d, x)$
- By Bayes rule: $p(d, x, z) = p(d|x, z)p(x, z)$
- By Bayes rule: $p(x, z) = p(x|z)p(z)$
- From graph: $p(x \mid z) = p(x)$

General DAGs and Factorization

- The probability law factorizes as:

$$p(\{x_\ell\}_{\ell \in V}) = \prod_{\ell \in V} p(x_\ell | pa_\ell)$$

DAGs Encode Conditional Independencies

- Any two variables X, Y are independent conditional on a set S if they are D(irected)-separated in the graph

$$\left(X \perp\!\!\!\perp_d Y \mid S \right)_G \quad \Rightarrow \quad X \perp\!\!\!\perp Y \mid S$$

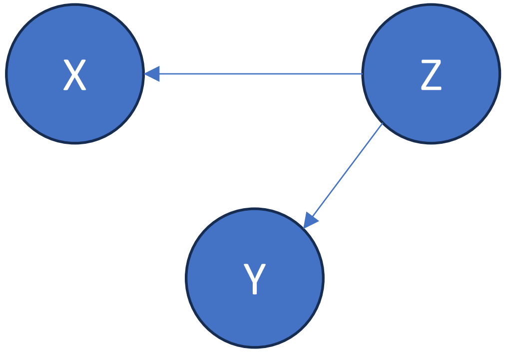
- Need to define the concept of D-separation

Graph Separation and Conditional Independence

- Looking at a graph, when can we conclude that

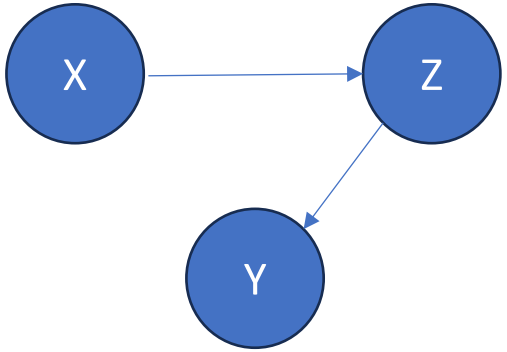
$$X \perp\!\!\!\perp Y \mid Z$$

Case 1:
Z is common cause



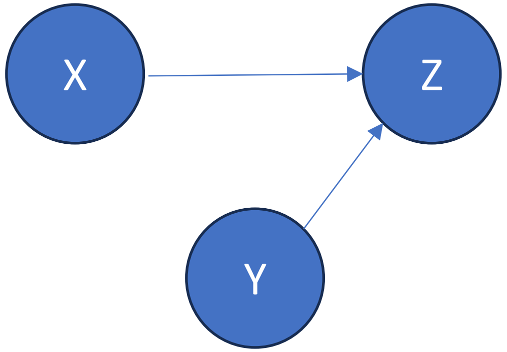
$$p(x,y,z) = p(x \mid z) \cdot p(y \mid z) \cdot p(z)$$
$$= f(x,z) \cdot g(y,z)$$

Case 2:
Z is mediator



$$p(x,y,z) = p(z \mid x) \cdot p(x) \cdot p(y \mid z)$$
$$= f(x,z) \cdot g(y,z)$$

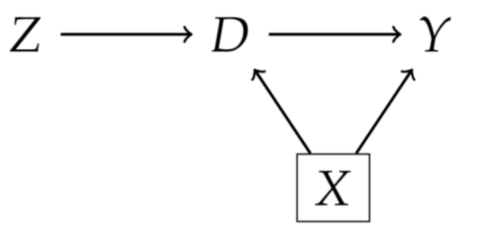
Case 3:
Z is collider

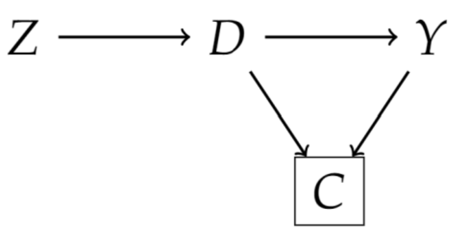


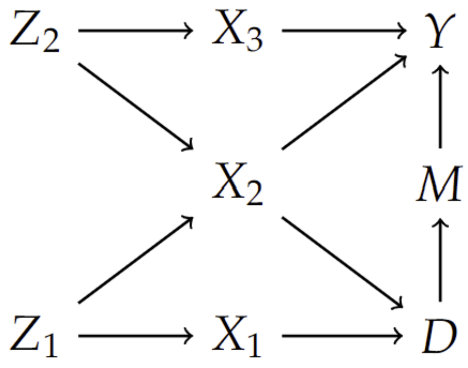
$$p(x,y,z) = p(z \mid x,y) \cdot p(x) \cdot p(y)$$
$$\neq f(x,z) \cdot g(y,z)$$

Some Graph Definitions

- A path π in a graph is blocked by a set of nodes S if
  - Either π contains a chain $i \to m \to j$ or a fork $i \leftarrow m \to j$ and $m \in S$
  - Or π contains a collider $i \to m \leftarrow j$ and neither m nor its descendants are in S







D-Separation

- In a DAG G, two nodes X, Y are D-separated by s set of nodes S if S blocks all paths between X and Y

- We denote it as:

$$\left(X \perp\!\!\!\perp_d Y \mid S \right)_G$$

D-separation implies conditional independency

$$\left(X \perp\!\!\!\perp_d Y \mid S \right)_G \quad \Rightarrow \quad X \perp\!\!\!\perp Y \mid S,$$

$$(\text{Verma, Pearl,}'\ 88)$$

DAGs encode conditional independencies: S d-separates X from Y in DAG G implies $X \perp\!\!\!\perp Y \mid S$

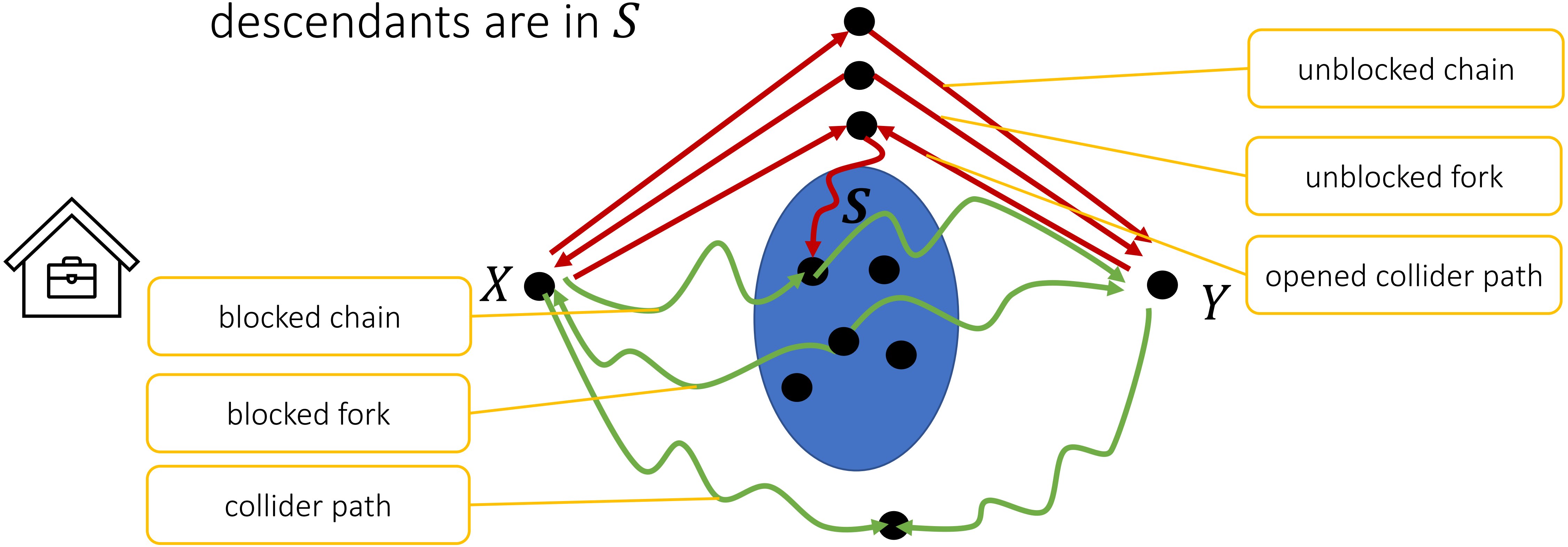
$$(X \perp\!\!\!\perp_d Y \mid S)_G \quad \Rightarrow \quad X \perp\!\!\!\perp Y \mid S$$

Implies testable restrictions we can use to refute DAG from data; e.g. for linear ASEMs, BLP of Y using X, S should have zero on X

$$Y = \boxed{\alpha} X + \beta' S + \epsilon, \qquad \epsilon \perp (X, S)$$

Test whether it is non-zero!

X is d-separated from Y by S if **every path** from X to Y is **blocked**.
S **blocks a path** if one of the following holds:
- path contains chain $X \rightarrow M \rightarrow Y$ or fork $X \leftarrow M \rightarrow Y$ and $M \in S$
- path contains collider $X \rightarrow M \leftarrow Y$ and neither M nor its descendants are in S



Graphical Criteria for Valid Adjustment Sets

Conditional Ignorability

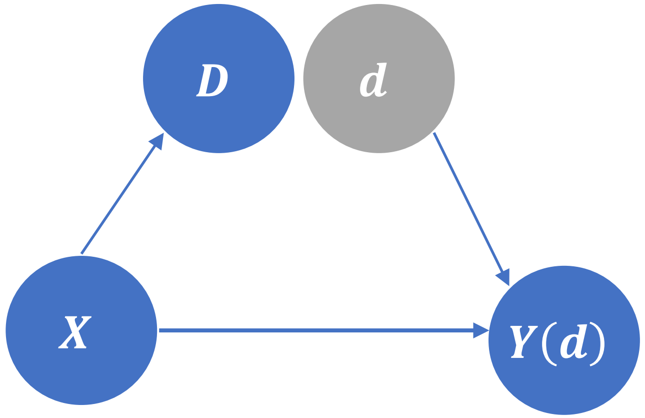
- Recall to do identification by conditioning we need for a set S

$$Y(d) \perp\!\!\!\perp D \mid S$$

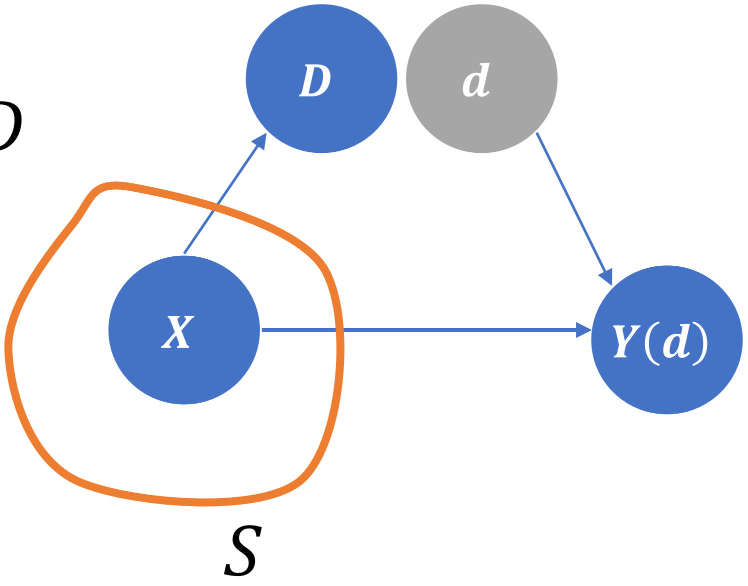
- Then predictive response equals structural response
$$E[Y(d) \mid S] = E[\, Y \mid D = d, S\,]$$
- Average predictive response equals average structural response
$$E[Y(d)] = E\big[E[Y|D = d, S]\big]$$

How can we check Conditional Ignorability

- Given a DAG, can we visually inspect if conditional ignorability holds
- Note that the SWIG graph contains both $Y(d)$ and D!
- We can simply check if $Y(d)$ is independent of D conditional on S on the SWIG graph!
- This is just a conditional independence statement on a DAG
- We can use d-separation!



Conditional ignorability between treatment D and outcome Y conditional on set S holds if **$\boldsymbol{Y}(\boldsymbol{d})$ is d-separated from $\boldsymbol{D}$ on SWIG $\widetilde{\boldsymbol{G}}(\boldsymbol{d})$** induced by $\mathrm{fix}(D = d)$ by the set S



Example

$$
\begin{tikzcd}
Z \arrow[r] & D & d \arrow[r] & Y(d) \\
 & & X \arrow[ul] \arrow[ur] &
\end{tikzcd}
$$

Example

$$
\begin{tikzcd}
Z_2 \arrow[r] \arrow[dr] & X_3 \arrow[r] & Y \\
 & X_2 \arrow[ur] \arrow[dr] & M \arrow[u] \\
Z_1 \arrow[r] \arrow[ur] & X_1 \arrow[r] & D \arrow[u]
\end{tikzcd}
$$

$$
\begin{tikzcd}
Z_2 \arrow[r] \arrow[dr] & X_3 \arrow[r] & Y(d) & \\
 & X_2 \arrow[ur] \arrow[dr] & & M(d) \arrow[ul] \\
Z_1 \arrow[r] \arrow[ur] & X_1 \arrow[r] & D & d \arrow[u]
\end{tikzcd}
$$

Proving the Main Theorem!

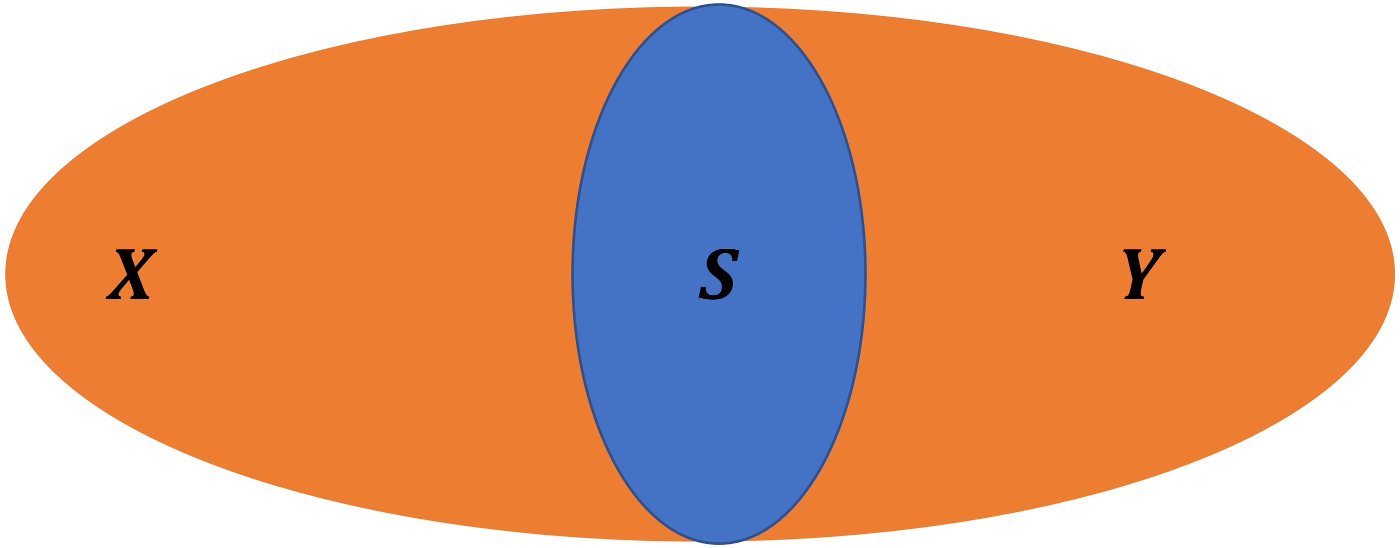
Proof Step 1

- A set of nodes $\boldsymbol{X}$ is called ancestral if all ancestors of $\boldsymbol{X}$ are in $\boldsymbol{X}$

- Removing all nodes outside of an ancestral set and looking at the resulting graph and ASEM, the probability law is the same as the probability law of $\boldsymbol{X}$ in the original graph (exercise)

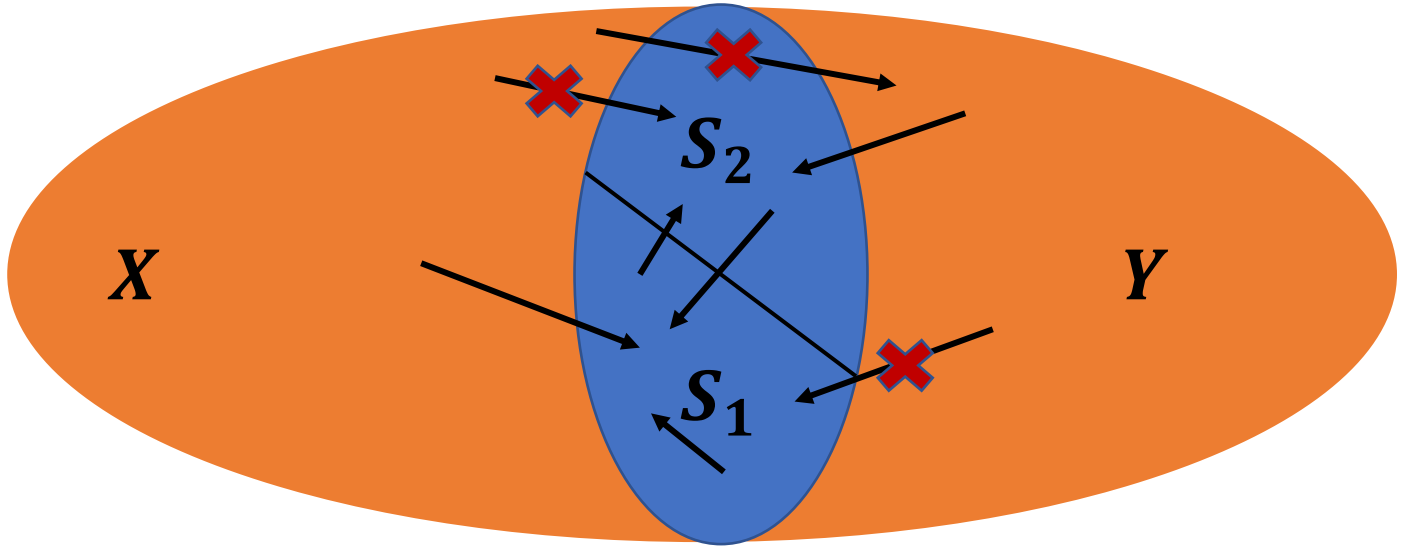
Proof Step 2

- Suppose that a set of nodes $\boldsymbol{X}$ is D-separated from a set of nodes $\boldsymbol{Y}$ by a set of nodes $\boldsymbol{S}$
- And that $\boldsymbol{X} \cup \boldsymbol{Y} \cup \boldsymbol{S}$ is the set of all nodes



Proof Step 2

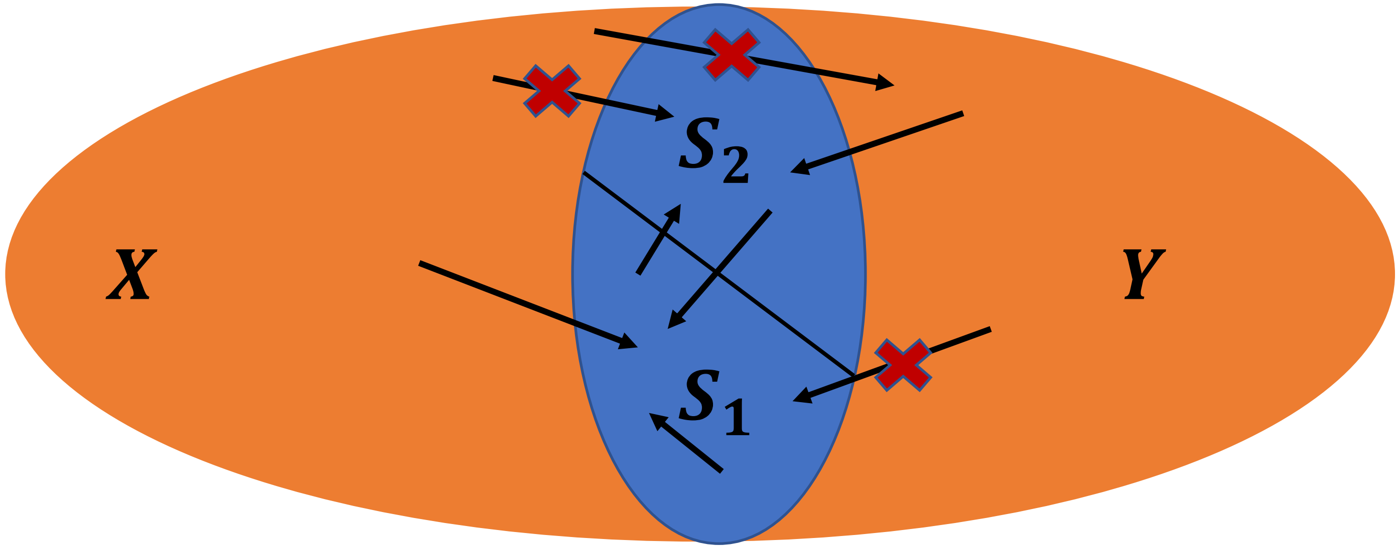
- Let S_1 the subset of S that has a parent in X. Let S_2 the remainder.
- It has to be that $Pa(X \cup S_1) \in X \cup S$
- It has to be that $Pa(Y \cup S_2) \in Y \cup S$



Proof Step 2

- We can factorize:

$$p(x,y,s) = \prod_{W \in X \cup S_1} p(w|pa_W) \prod_{W \in Y \cup S_2} p(w|pa_W) = f(x,s_1)g(y,s_2)$$



Proof Step 2

- We can factorize:

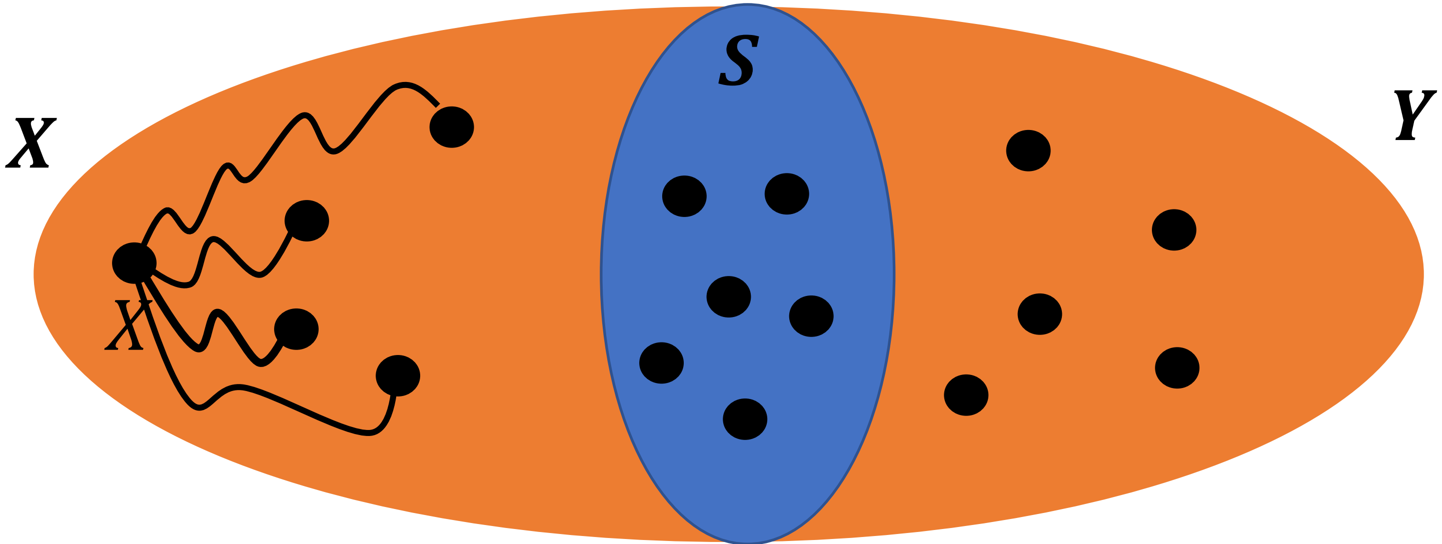
$$p(x, y, s) = \prod_{W \in X \cup S_1} p(w|pa_W) \prod_{W \in Y \cup S_2} p(w|pa_W) = f(x, s_1) g(y, s_2)$$

- Implies that:

$$\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid S$$

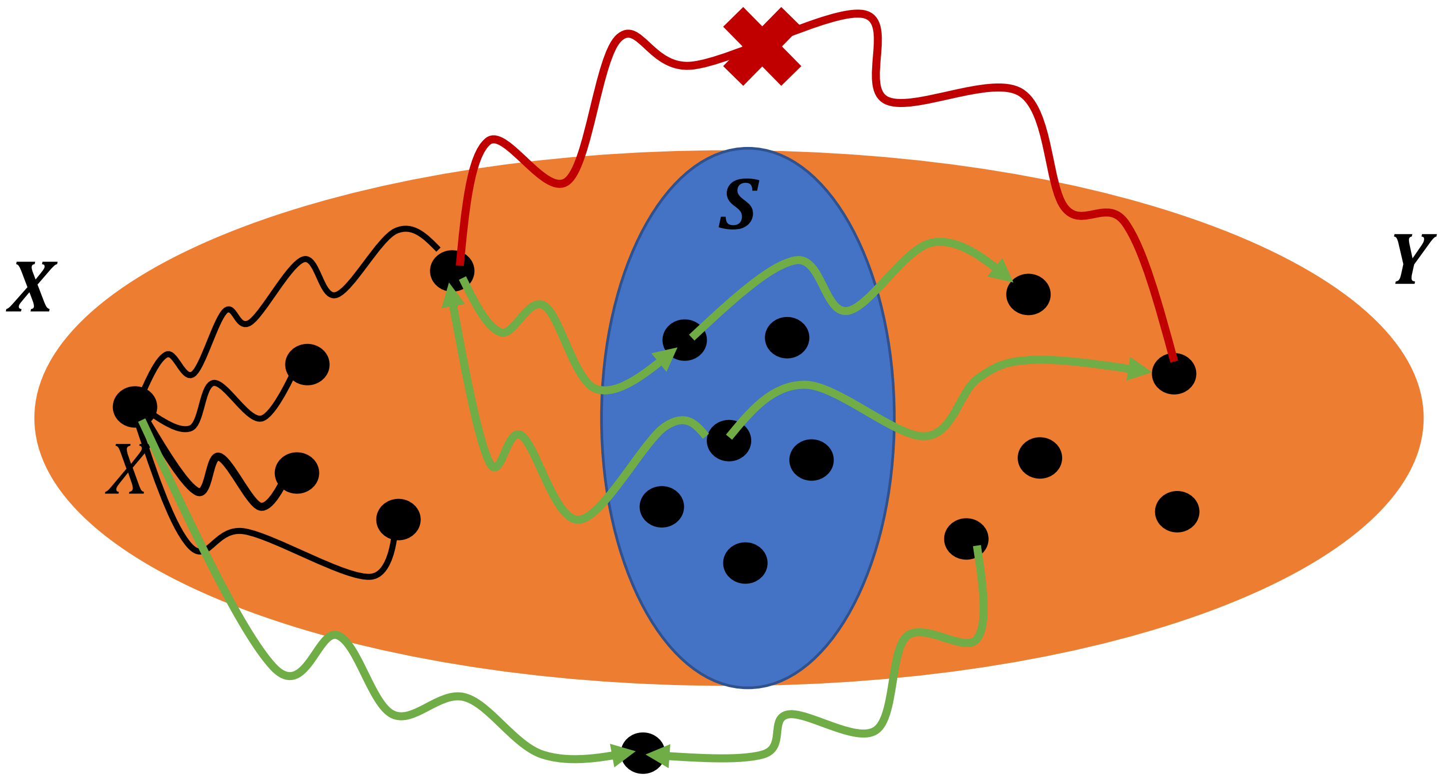
Final Step

- By first step, we can restrict to ancestral set of $X \cup Y \cup S$
- Does not change conditional independence relations (exercise)
- Does not change d-separation relations (exercise)
- Define $\boldsymbol{X}$ nodes in ancestral set of $X \cup Y \cup S$ not d-separated from X
- Define $\boldsymbol{Y}$ the remainder of nodes in ancestral set not in $\boldsymbol{X}, \boldsymbol{S}$.



Final Step

- By definition of d-separation, S must d-separate $\boldsymbol{X}$ from $\boldsymbol{Y}$ (exercise)
- We can invoke previous critical lemma



Final Step

- By marginalization

$$p(x, y, \boldsymbol{s}) = \int \int p(x, \boldsymbol{x}', y, \boldsymbol{y}', \boldsymbol{s}) d\boldsymbol{x}' \, d\boldsymbol{y}'$$

- By step 2

$$p(x, y, \boldsymbol{s}) = \int \int f(x, \boldsymbol{x}', \boldsymbol{s}) \, g(y, \boldsymbol{y}', \boldsymbol{s}) \, d\boldsymbol{x}' \, d\boldsymbol{y}'$$

- We can split integrals

$$p(x, y, \boldsymbol{s}) = \int f(x, \boldsymbol{x}', \boldsymbol{s}) d\boldsymbol{x}' \int g(y, \boldsymbol{y}', \boldsymbol{s}) \, d\boldsymbol{y}'$$

- Thus

$$p(x, y, \boldsymbol{s}) = \bar{f}(x, \boldsymbol{s}) \bar{g}(y, \boldsymbol{s}) \Rightarrow X \perp\!\!\!\perp Y \mid S$$