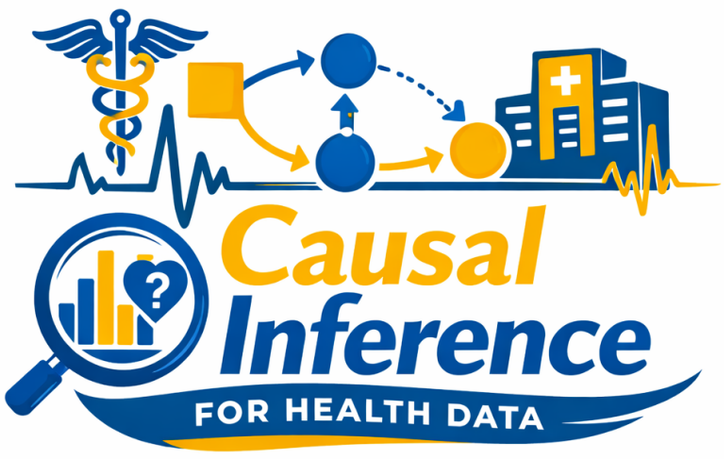
Causal Inference for Health Data

(STATS C160/C260 — Winter 2026)

Lecture 10: Counterfactuals II

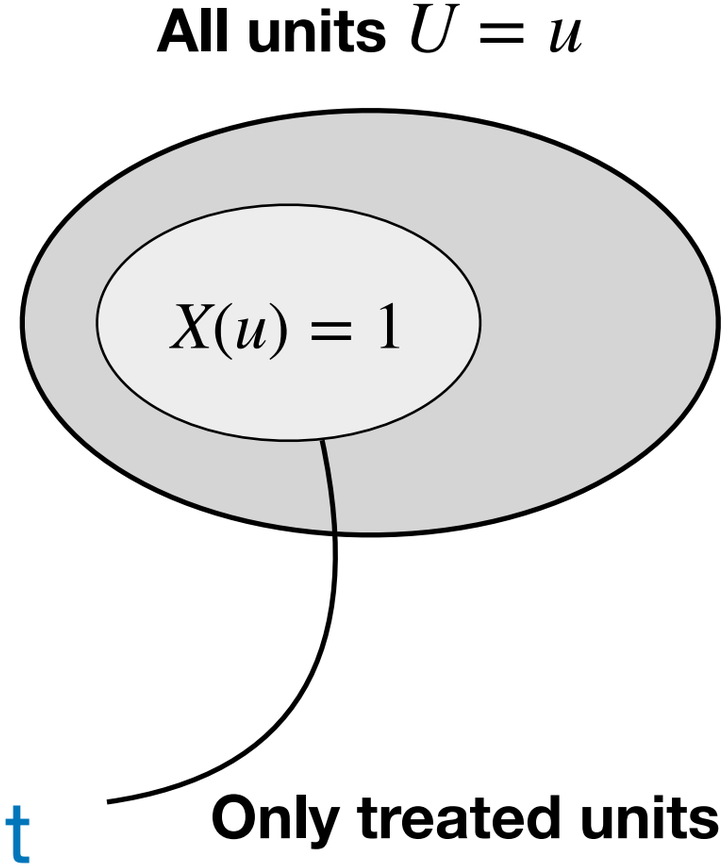
Drago Plečko

UCLA
College | Physical Sciences
Statistics & Data Science



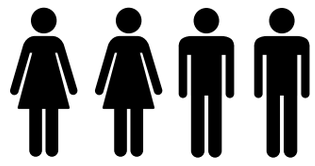
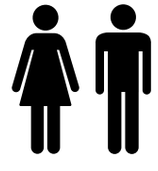
Example — Takeaways

- Causal effects may exhibit *heterogeneity*, meaning that different parts of the population are affected differently,
- The average causal effect
  $E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)]$
  is an average over the entire, possibly heterogenous, population of individuals,
- To better understand heterogeneity, we may use conditional effects, which sometimes require Layer 3 of PCH,
- This is the case for the Effect of Treatment on the Treated (ETT).



(1) Use of Counterfactuals: Granularity

Population Axis

population $P(u)$		entire population
x-subset $P(u \mid x)$		treated population
z, x-subset $P(u \mid x, z)$		treated, Stage I-II
z, x, w-subset $P(u \mid x, z, w)$		treated, Stage I-II, with side-effects
v-subset $P(u \mid v)$		treated, Stage I-II, side-effects, survived
unit-level u		Catherine, treated, Stage I-II, side-effects, survived

Example. Semaglutide Effects on Diabetes

- A diabetologist is treating his patients with semaglutide, to stabilize their HbA1c values,
- However, he realized that there are different causal pathways through which semaglutide may affect HbA1c levels: (1) through weight-loss and associated insulin sensitivity, and (2) direct GLP-1 receptor effects,

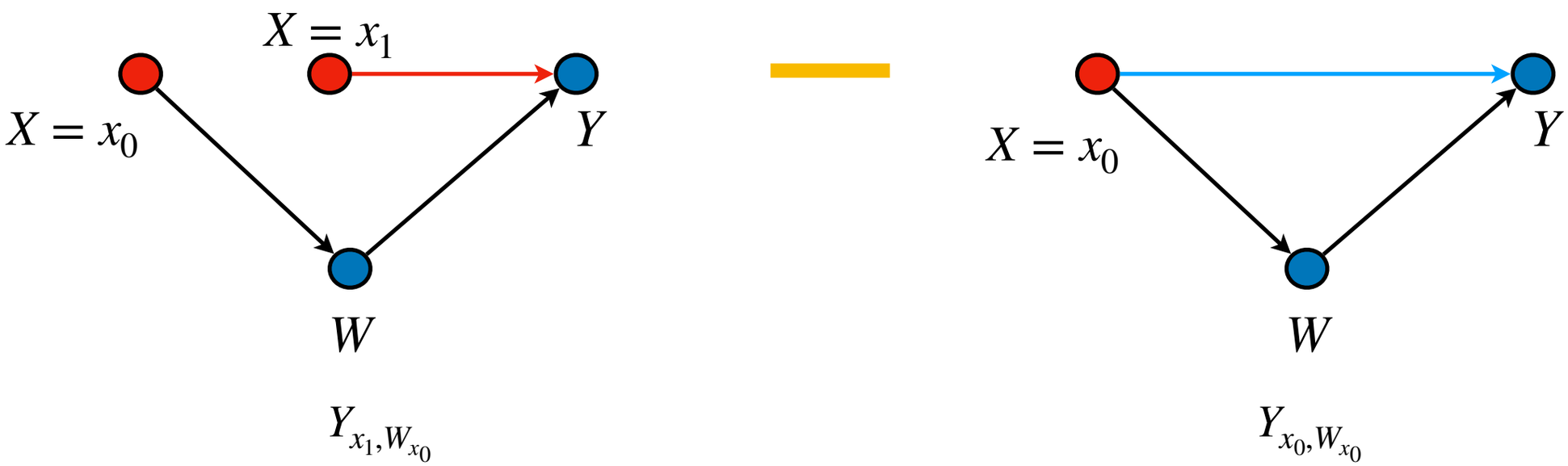
- The doctor is interested in how strong each of these pathways is, and he wonders if he can obtain an answer from data:
- $X = 1$ denotes semaglutide treatment, $X = 0$ standard treatment, $Y = 1$ denote normal HbA1c levels ($Y = 0$ otherwise); $W = 1$ for weight loss ($W = 0$ otherwise).

Gedankenexperiment (NDE)

- For investigating this, the doctor performs following thought experiments,

- For an untreated individual ($X = x_0$), how would his HbA1c level (Y) change had he been treated ($X = x_1$), while keeping the weight loss unchanged (at the natural level $X = x_0$)?

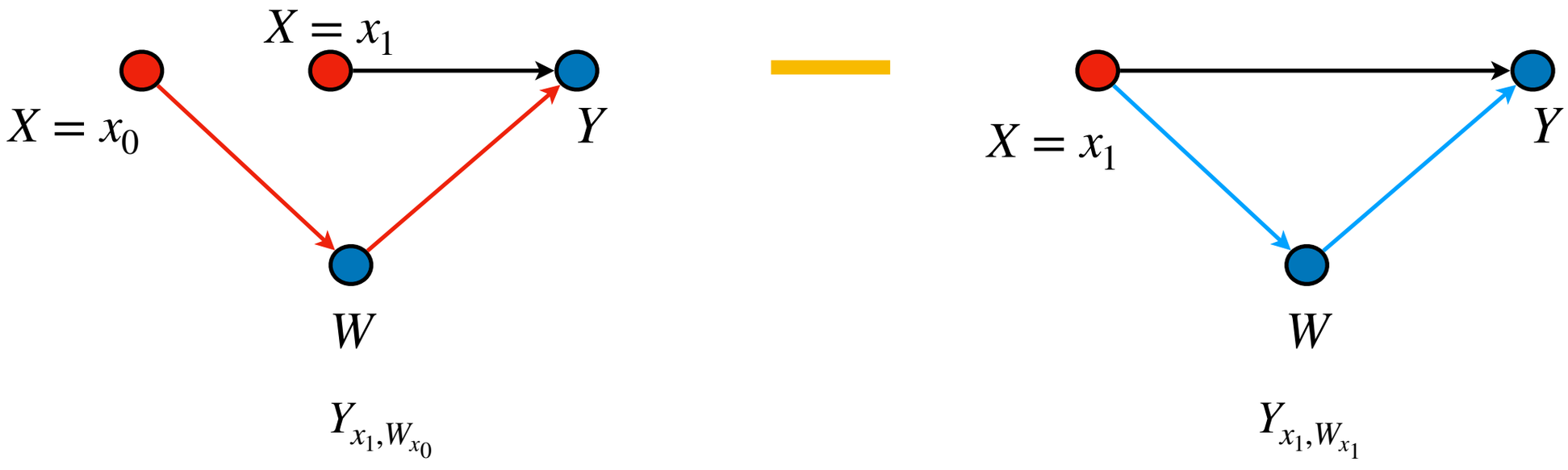
$$\mathbf{NDE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0,W_{x_0}})$$



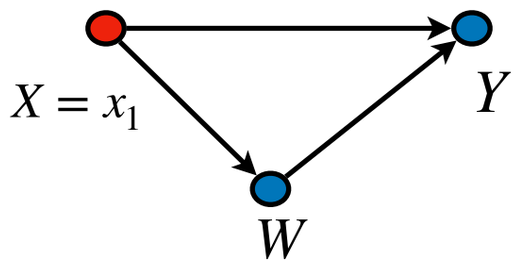
Gedankenexperiment (NIE)

- For a treated individual ($X = x_1$), how would their HbA1c level (Y) change had they not been treated ($X = x_0$), while keeping treatment unchanged along the direct causal pathway (at the natural level $X = x_1$)?

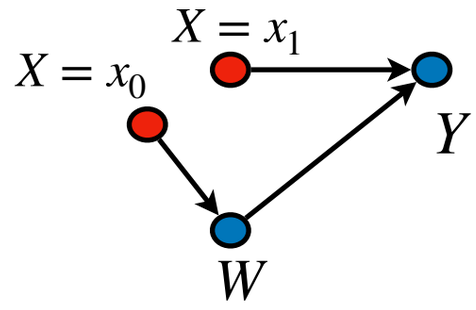
$$\textbf{NIE}_{x_1,x_0}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_1,W_{x_1}})$$



ATE Decomposition



$Y \mid do(X = x_1)$



$Y_{x_1, W_{x_0}}$

$-\text{NIE}_{x_1,x_0}(y)$

Theorem. *The average total effect can be decomposed into its direct and indirect parts:*

$$ATE_{x_0,x_1}(y) = NDE_{x_0,x_1}(y) - NIE_{x_1,x_0}(y)\,.$$

(Y)

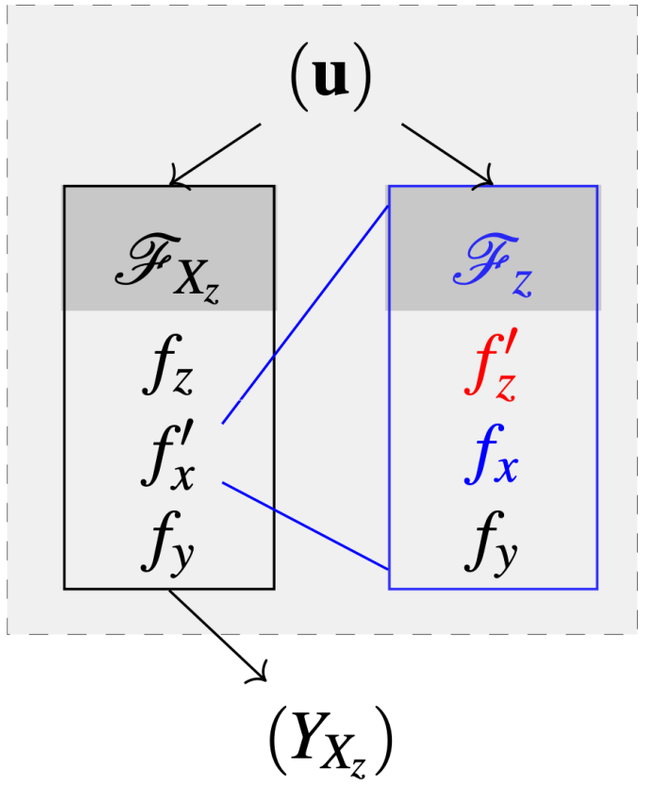
$Y_{x_1, W_{x_0}}$

$Y \mid do(X = x_0)$

$\text{NDE}_{x_0,x_1}(y)$

Nested Counterfactuals

- We used the notion of *nested counterfactuals*.
- Y_x refers to Y under intervention $do(X = x)$, that is, when X is fixed to a constant value x.
- Here, we consider an intervention $do(X \leftarrow X_z)$, where X is not fixed to a constant but is supposed to behave as another counterfactual variable X_z.

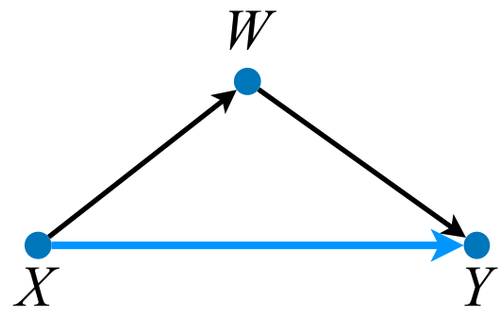


- For such intervention, we first need to evaluate X_z (from $\mathcal{M}_z$), and then consider a model $\mathcal{M}_{X_z}$ where X is given the value dictated by X_z.

Natural Direct Effects

- The Natural Direct Effect (NDE) is defined as

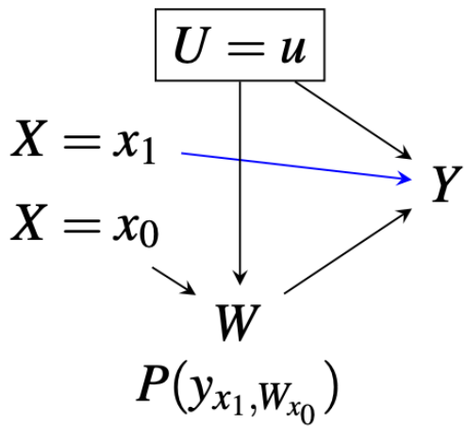
$$\boxed{\text{NDE}_{x_0,x_1}(y) = P(y_{x_1,W_{x_0}}) - P(y_{x_0,W_{x_0}})} \quad (5.101)$$



- $Y_{x_1,W_{x_0}}$ refers to the outcome Y under $X = x_1$ and $W = W_{x_0}$, the value that W would attain had X been x_0.

- Y_{x_0} represents a baseline where Y perceives $X = x_0$ in all causal paths. It is also equivalent to $Y_{x_0 W_{x_0}}$, that is, a situation where W also gets $X = x_0$.

- Taking the difference of those two quantities keeps the path $X \rightarrow W \rightarrow Y$ constant while changing the level of X from x_0 to x_1 in the path $X \rightarrow Y$, effectively measuring the direct impact of X on Y.



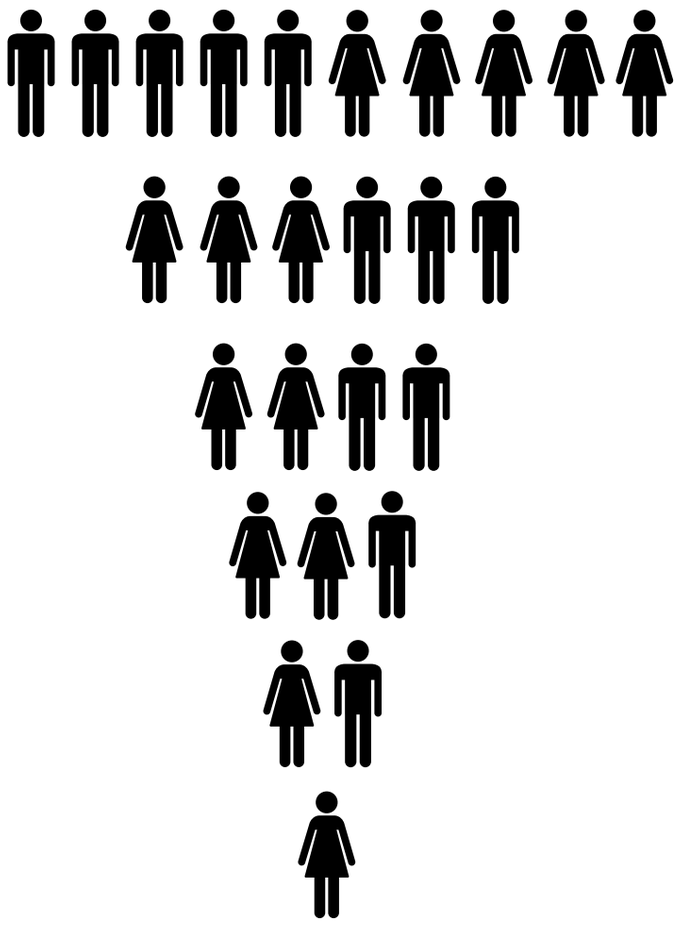
Summary: Using of Counterfactuals

Total Effect — TE

$$E[Y \mid do(x_1)] - E[Y \mid do(x_0)]$$

We spent the first 8 lectures building foundations to understand this quantity

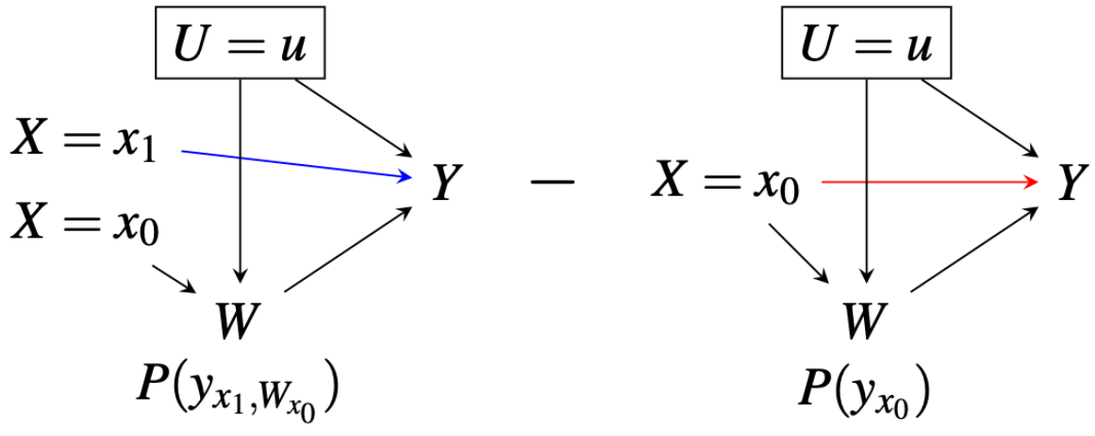
Ext. 1: Granularity



Ext. 2: Mechanisms

DE

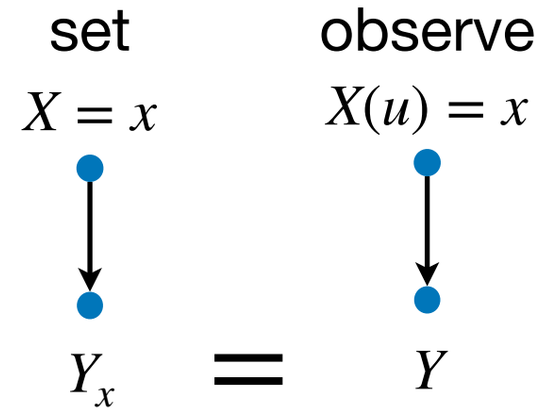
IE



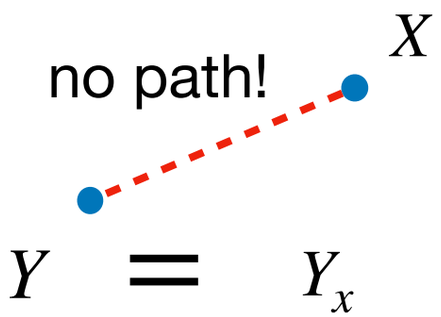
Counterfactual Constraints

We will study three types of constraints, namely:

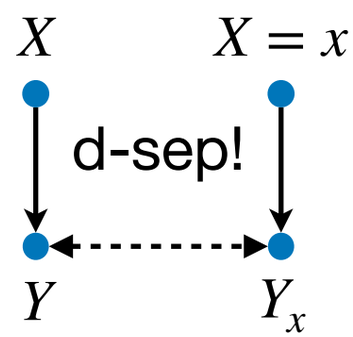
1. Consistency: variables behave the same with or without intervention under a certain observational context.



2. Exclusion: variables behave the same regardless of the exclusion of one or more interventions under a certain interventional context.



3. Independence: (counterfactual) variables are probabilistically iindependent.



Consistency

SCM M

unit $u = (u_1, \ldots, u_k) \sim P(u)$

after u is fixed, the evaluation is deterministic

$$V_1 \leftarrow f_1(u_1)$$
$$X \leftarrow f_X(v_1, u_x)$$
$$\vdots$$
$$Y \leftarrow f_Y(v_1, \ldots, v_k, u_y)$$

$X(u) = x$

same solution!

SCM M_x

for the same fixed u, the evaluation is again deterministic

$$V_1 \leftarrow f_1(u_1)$$
$$X \leftarrow x$$
$$\vdots$$
$$Y_x \leftarrow f_Y(v_1, \ldots, (v_k)_x, u_y)$$

Consistency

Lemma - Consistency. Given a structural causal model M and $X, Y \in \mathbf{V}$, $\mathbf{T} \subseteq \mathbf{V}$, let x be a value in the domain of X. Then,

$$X_{\mathbf{T}} = x \Rightarrow Y_{\mathbf{T}x} = Y_{\mathbf{T}}.$$

Whenever $\mathbf{T}$ is empty, consistency can be written as:

$$X = x \Rightarrow Y_x = Y.$$

(In some settings, consistency is used when T is empty, and is called *composition* otherwise.)

Consistency Example

- Recall the ETT quantity:

$$E(Y_{X=1} \mid X = 1) - E(Y_{X=0} \mid X = 1)$$

consistency? ✓

consistency? ✗

$$= E(Y \mid X = 1)$$

we compute it later!
(need other constraints)

Counterfactual Unnesting Theorem (CUT)

Theorem – (CUT) Let $Y, X \in \mathbf{V}$, $\mathbf{T}, \mathbf{Z} \subseteq \mathbf{V}$, and let $\mathbf{z}$ be a set of values for $\mathbf{Z}$. Then, the nested counterfactual $P(Y_{\mathbf{T}X_{\mathbf{z}}} = y)$ can be written as a non-nested counterfactual, as follows:

$$P(Y_{\mathbf{T}X_{\mathbf{z}}} = y) = \sum_{x} P(Y_{\mathbf{T}x} = y, X_{\mathbf{Tz}} = x).$$

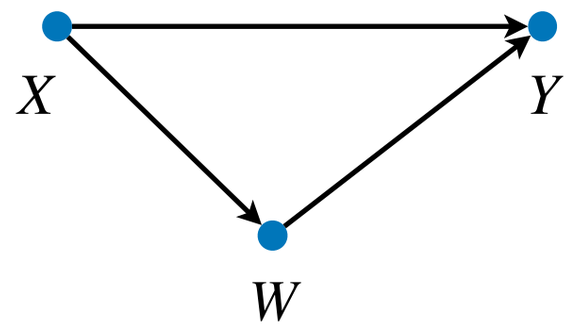
This statement's proof only requires two steps, the law of total probability and consistency itself, i.e.:

$$P(Y_{\mathbf{T}X_{\mathbf{z}}} = y) = \sum_{x} P(Y_{\mathbf{T}X_{\mathbf{z}}} = y, X_{\mathbf{Tz}} = x) \quad \text{(sum over } X_{\mathbf{Tz}}\text{)}$$

$$= \sum_{x} P(Y_{\mathbf{T}x} = y, X_{\mathbf{Tz}} = x) \quad \text{(consistency nested ctfs.)}.$$

Un-nesting Example: NDE

- Recall the notion of natural direct effect (NDE) that we defined



$$\text{NDE}_{x_0,x_1}(y) \stackrel{\delta}{=} E[Y_{x_1,W_{x_0}} - Y_{x_0}],$$

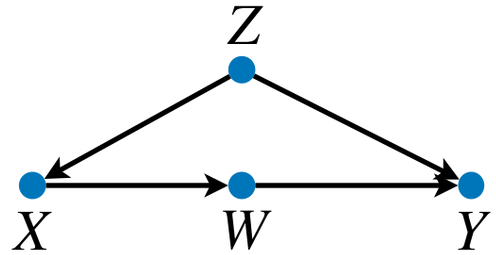
- This quantity requires the nested $Y_{x_0,W_{x_1}}$, and applying the CUT we get:

$$P(Y_{x_0,W_{x_1}} = y) = \sum_w P(Y_{x_1 w} = y, W_{x_0} = w) \quad \text{(CUT)}$$

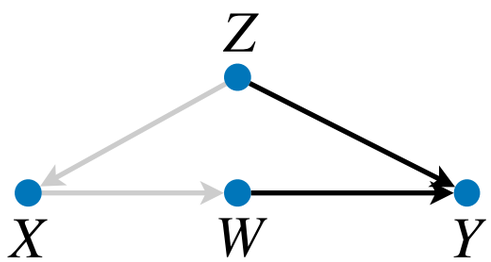
Counterfactual Ancestors

Definition – Ancestors (of a counterfactual). Let $Y_{\mathbf{x}}$ be such that $Y \in \mathbf{V}, \mathbf{X} \subseteq \mathbf{V}$. Then, the set of (counterfactual) ancestors of $Y_{\mathbf{x}}$, denoted $An(Y_{\mathbf{x}})$, consist of each $W_{\mathbf{z}}$, such that $W \in An_{\mathcal{G}_{\overline{\mathbf{X}}}}(Y) \backslash \mathbf{X}$ (which includes Y itself), and $\mathbf{z} = \mathbf{x} \cap An_{\mathcal{G}_{\overline{\mathbf{X}}}}(W)$.

Example. Consider the causal diagram:



- $An(Y_{xw}) = \{Z, Y_w\}$



$An_{\mathcal{G}_{\overline{XW}}}(Y) \backslash \{X, W\} = \{Z, Y\}$

Get ancestors

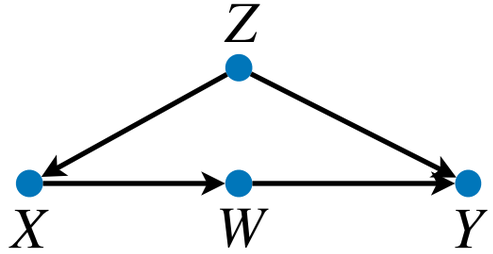
$\varnothing = \{x, w\} \cap An_{\mathcal{G}_{\overline{XW}}}(Z)$
$\{w\} = \{x, w\} \cap An_{\mathcal{G}_{\overline{XW}}}(Y)$

Compute subscripts

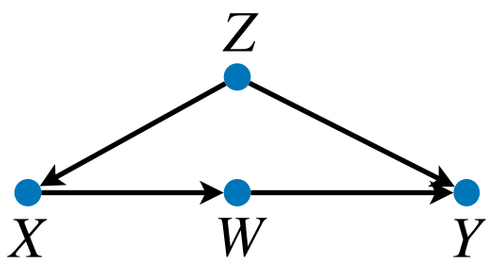
Counterfactual Ancestors (cont)

Definition – Ancestors (of a counterfactual). Let $Y_{\mathbf{x}}$ be such that $Y \in \mathbf{V}, \mathbf{X} \subseteq \mathbf{V}$. Then, the set of (counterfactual) ancestors of $Y_{\mathbf{x}}$, denoted $An(Y_{\mathbf{x}})$, consist of each $W_{\mathbf{z}}$, such that $W \in An_{\mathcal{G}_{\overline{\mathbf{X}}}}(Y) \backslash \mathbf{X}$ (which includes Y itself), and $\mathbf{z} = \mathbf{x} \cap An_{\mathcal{G}_{\overline{\mathbf{X}}}}(W)$.

Example (2). Consider the causal diagram from before:



- $An(Y_z) = \{X_z, W_z, Y_z\}$



$An_{\mathcal{G}_{\overline{Z}}}(Y) \backslash \{Z\} = \{Y, W, X\}$

Get ancestors

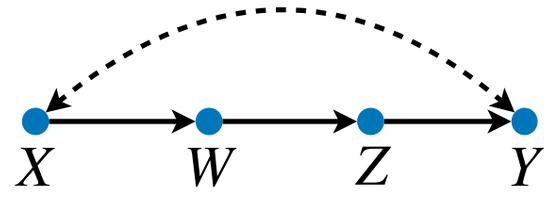
$\{z\} = \{z\} \cap An_{\mathcal{G}_{\overline{Z}}}(X)$
$\{z\} = \{z\} \cap An_{\mathcal{G}_{\overline{Z}}}(W)$
$\{z\} = \{z\} \cap An_{\mathcal{G}_{\overline{Z}}}(Y)$

Compute subscripts

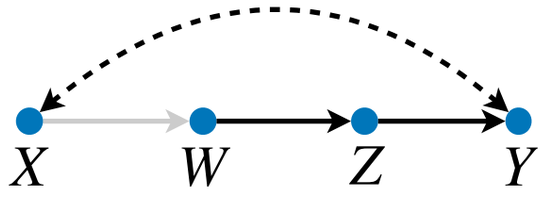
Counterfactual Ancestors (cont)

Definition – Ancestors (of a counterfactual). Let $Y_{\mathbf{x}}$ be such that $Y \in \mathbf{V}, \mathbf{X} \subseteq \mathbf{V}$. Then, the set of (counterfactual) ancestors of $Y_{\mathbf{x}}$, denoted $An(Y_{\mathbf{x}})$, consist of each $W_{\mathbf{z}}$, such that $W \in An_{\mathcal{G}_{\overline{\mathbf{X}}}}(Y) \backslash \mathbf{X}$ (which includes Y itself), and $\mathbf{z} = \mathbf{x} \cap An_{\mathcal{G}_{\overline{\mathbf{X}}}}(W)$.

Example (3). Consider the causal diagram:



- $An(Y_w) = \{Y_w, Z_w\}$



$An_{\mathcal{G}_{\overline{W}}}(Y) \backslash \{W\} = \{Y, Z\}$

Get ancestors

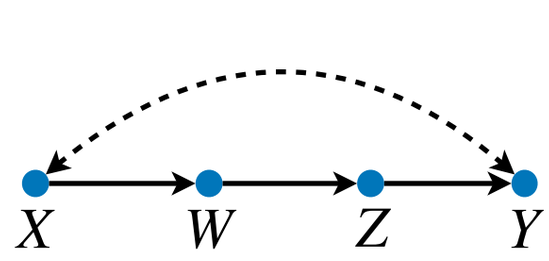
$\{w\} = \{w\} \cap An_{\mathcal{G}_{\overline{W}}}(Y)$
$\{w\} = \{w\} \cap An_{\mathcal{G}_{\overline{W}}}(Z)$

Compute subscripts

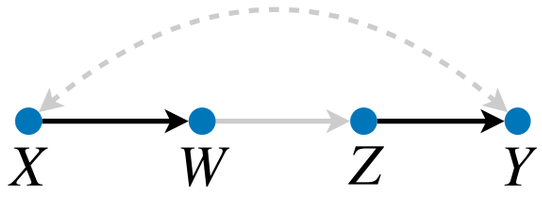
Counterfactual Ancestors (cont)

Definition – Ancestors (of a counterfactual). Let $Y_{\mathbf{x}}$ be such that $Y \in \mathbf{V}, \mathbf{X} \subseteq \mathbf{V}$. Then, the set of (counterfactual) ancestors of $Y_{\mathbf{x}}$, denoted $An(Y_{\mathbf{x}})$, consist of each $W_{\mathbf{z}}$, such that $W \in An_{\mathcal{G}_{\overline{\mathbf{X}}}}(Y) \backslash \mathbf{X}$ (which includes Y itself), and $\mathbf{z} = \mathbf{x} \cap An_{\mathcal{G}_{\overline{\mathbf{X}}}}(W)$.

Example (4). Consider the causal diagram:



- $An(W_{xz}) = \{W_x\}$



$An_{\mathcal{G}_{\overline{XZ}}}(W) \backslash \{X, Z\} = \{W\}$ — *Get ancestors*

$\{x\} = \{xz\} \cap An_{\mathcal{G}_{\overline{XZ}}}(W)$ — *Compute subscripts*

Exclusion Restrictions

- Semantically, one can consider counterfactuals $Y_{\mathbf{t}}$ for arbitrary $Y \in \mathbf{V}$ and $\mathbf{T} \subseteq \mathbf{V}$.

- However, the variations allowed for counterfactual variables usually depend on the topology and the sparsity of the causal system.

$$\mathcal{F}_z = \begin{cases} X_z \leftarrow f_X(U_x) \\ Z_z \leftarrow z \\ Y_z \leftarrow f_Y(z, U_y) \end{cases}$$

- Consider for instance Y_z and Y_{zx} in the chain graph above. Note that for every unit $\mathbf{U} = \mathbf{u}$, the variables Y_z and Y_{zx} get always the same value. Formally:

$$\mathcal{F}_{zx} = \begin{cases} X_{zx} \leftarrow x \\ Z_{zx} \leftarrow z \\ Y_{zx} \leftarrow f_Y(z, U_y), \end{cases}$$

$$Y_{zx}(\mathbf{u}) = f_Y(z, u_y) = Y_z(\mathbf{u}).$$

Exclusion Operator

Lemma – Exclusion Operator. Consider a causal diagram G and a counterfactual variable $Y_{\mathbf{x}}$. Let the subscript restriction operator be labeled with η, defined by

$$\eta(Y_{\mathbf{x}}) := Y_{\mathbf{z}},$$

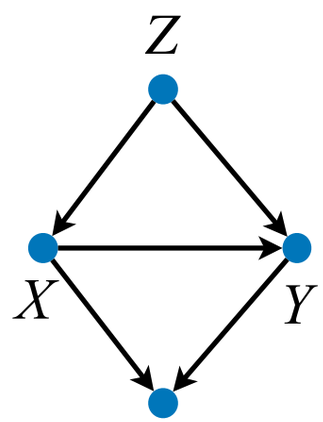
where $\mathbf{Z} = \mathbf{X} \cap \mathrm{An}_{G_{\overline{\mathbf{X}}}}(Y)$ and $\mathbf{z} = \mathbf{x} \cap \mathbf{Z}$. Then, $\eta(Y_{\mathbf{x}})$ and $Y_{\mathbf{x}}$ are the same counterfactual variable in any model compatible with G.

For a set $\mathbf{Y}_*$, define

$$\eta(\mathbf{Y}_*) = \bigcup_{Y_{\mathbf{t}} \in \mathbf{Y}_*} \eta(Y_{\mathbf{t}}).$$

Exclusion Operator Examples

Consider the causal diagram to the right. The following are some counterfactual variables that can be minimized using the exclusion operator:



$$\eta(X_{zyw}) = X_z \qquad \{Z, Y, W\} \cap An_{\mathcal{G}_{\overline{ZYW}}}(X) = \{Z\}$$

$$\eta(Y_{xz}) = Y_{xz} \qquad \{X, Z\} \cap An_{\mathcal{G}_{\overline{XZ}}}(Y) = \{X, Z\}$$

$$\eta(W_{zxy}) = W_{xy} \qquad \{Z, X, Y\} \cap An_{\mathcal{G}_{\overline{ZXY}}}(W) = \{X, Y\}$$

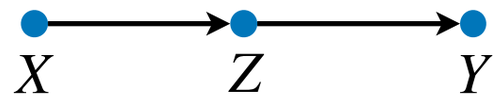
$$\eta(W_{zy}) = W_{zy} \qquad \{Z, Y\} \cap An_{\mathcal{G}_{\overline{ZY}}}(W) = \{Z, Y\}$$

Causal Diagrams for Counterfactuals: Ancestral Multi-World Networks (AMWN)

Input: graph G, ctf. variables $\mathbf{Y}_*$

Output: Ctf. ancestral graph $G_A(\mathbf{Y}_*)$ over ctf. variables $\mathbf{Y}_*$

Example:



- compute all $\mathrm{An}(\mathbf{Y}_*)$ & add them as nodes to initial G'

- add directed arrows witnessing ancestrality

- for each edge $U \to V$ in G

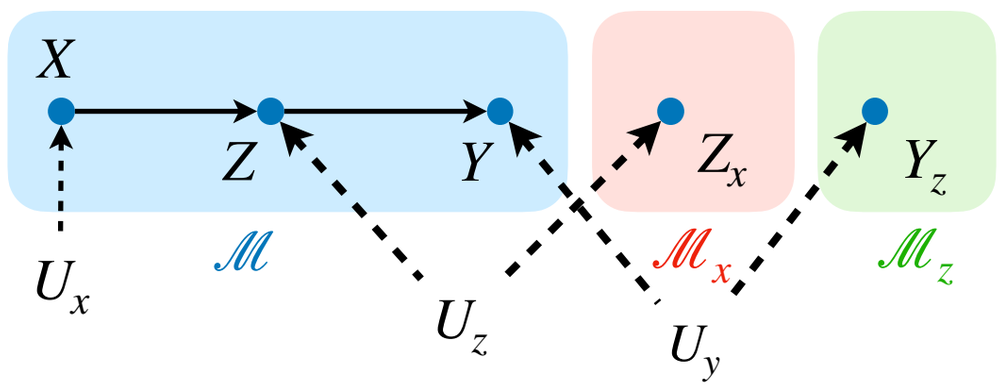
  - If more than one instance of V in G'

    add node U with edges to each copy V_x of V

- return G'

$$Y_z \perp Z_x \mid Y$$

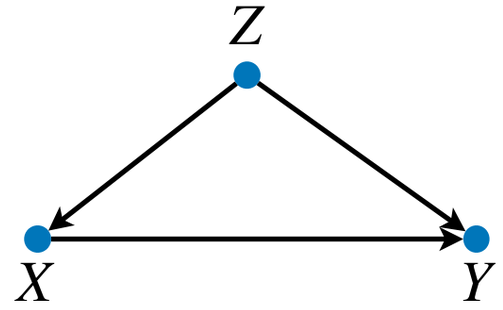
$$An(Y_z) = \{Y_z\}$$
$$An(Z_x) = \{Z_x\}$$
$$An(Y) = \{Y, Z, X\}$$



ETT Derivation

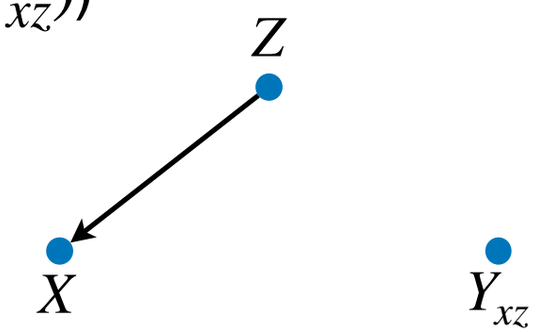
$$P(y_x \mid x') = \sum_z P(y_x \mid z, x')P(z \mid x') \text{ (Conditioning on } Z\text{)}$$

$$= \sum_z P(y_x \mid z_x, x')P(z \mid x') \quad \text{(exclusion: } \{X\} \cap An(Z) = \varnothing\text{)}$$



$$= \sum_z P(y_{xz} \mid z_x, x')P(z \mid x') \quad \text{(consistency } (Z_x = z \Rightarrow Y_x = Y_{xz})\text{)}$$

$$= \sum_z P(y_{xz} \mid z, x')P(z \mid x') \quad \text{(exclusion: } \{X\} \cap An(Z) = \varnothing\text{)}$$



$$= \sum_z P(y_{xz} \mid z, x)P(z \mid x') \quad \text{(indep. } Y_{xz} \perp X \mid Z \text{ in } \mathcal{G}_A\text{)}$$

$\mathcal{G}_A(Y_{xz}, X, Z)$

$$= \sum_z P(y \mid z, x)P(z \mid x') \quad \text{(consistency } (Z = z, X = x \Rightarrow Y_{xz} = Y)\text{)}$$

can be computed from observational data!

NDE Derivation

$$P(y_{x'W_x}) = \sum_w P(y_{x'w}, w_x) \qquad \text{(CUT)}$$

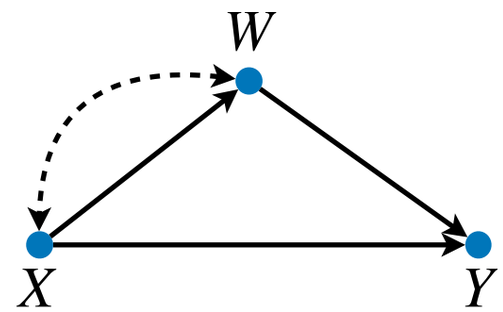
$$= \sum_w P(y_{x'w} \mid w_x) P(w_x) \qquad \text{(Chain rule)}$$

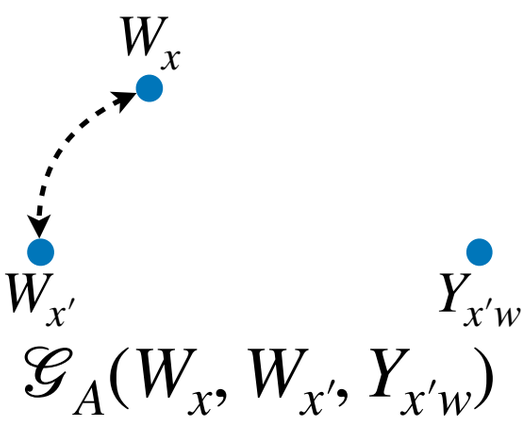
$$= \sum_w P(y_{x'w} \mid w_{x'}) P(w_x) \qquad \text{(indep.: } (Y_{x'w} \perp W_x, W_{x'}) \text{ in } \mathcal{G}_A$$

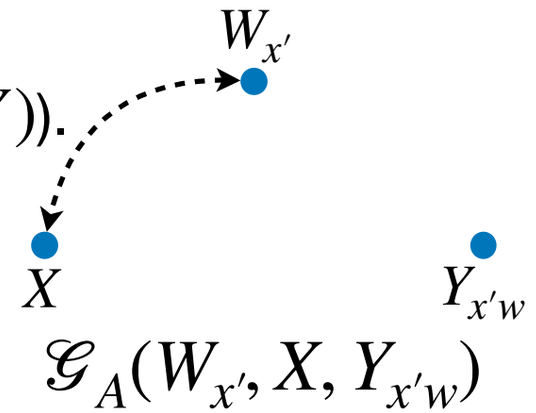
$$= \sum_w P(y_{x'w} \mid w_{x'}, x') P(w_x) \qquad \text{(indep: } (Y_{x'w} \perp X \mid W_{x'}) \text{ in } \mathcal{G}_A$$

$$= \sum_w P(y_{x'w} \mid w, x') P(w_x) \qquad \text{(consist: } (X = x' \Rightarrow W_{x'} = W))$$

$$= \sum_w P(y \mid w, x') P(w_x) \qquad \text{(consist: } (W = w, X = x' \Rightarrow Y_{x'w} = Y)).$$







Counterfactual Calculus

Theorem. Let $\mathcal{G}$ be a causal diagram, then for $\mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{W}, \mathbf{T} \subseteq \mathbf{V}$, the following rules hold for the probability distributions generated by any model compatible with $\mathcal{G}$:

Rule 1 (Consistency rule – Observation/intervention exchange)

$$P(\mathbf{y}_{\mathbf{Tx}}, \mathbf{x}_{\mathbf{T}}, \mathbf{w}_*) = P(\mathbf{y}_{\mathbf{T}}, \mathbf{x}_{\mathbf{T}}, \mathbf{w}_*)$$

Rule 2 (Independence Rule – Adding/removing counterfactual observations)

$$P(\mathbf{y}_{\mathbf{r}} \mid \mathbf{x}_{\mathbf{t}}, \mathbf{w}_*) = P(\mathbf{y}_{\mathbf{r}} \mid \mathbf{w}_*) \qquad \text{if } (\mathbf{Y}_{\mathbf{r}} \perp \mathbf{X}_{\mathbf{t}} \mid \mathbf{W}_*) \text{ in } \mathcal{G}_A,$$

Rule 3 (Exclusion Rule – Adding/removing interventions)

$$P(\mathbf{y}_{\mathbf{xz}}, \mathbf{w}_*) = P(\mathbf{y}_{\mathbf{z}}, \mathbf{w}_*) \qquad \text{if } \mathbf{X} \cap An(\mathbf{Y}) = \varnothing \text{ in } \mathcal{G}_{\overline{\mathbf{Z}}},$$

where $\mathcal{G}_A$ is the counterfactual ancestral graph $\mathcal{G}_A(\mathbf{Y}_{\mathbf{r}}, \mathbf{X}_{\mathbf{t}}, \mathbf{W}_*)$.