Causal Inference via Directed Acyclic Graphs and Linear/Nonlinear Structural Equation Models

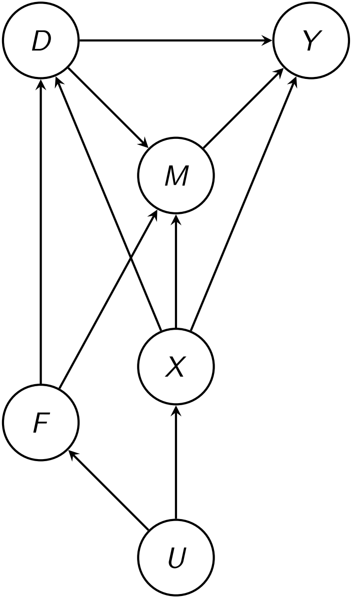
Gauranga Kr. Baishya, Chennai Mathematical Institute (CMI)

May 14, 2025

Introduction

1. Here, we explore a fully linear & nonlinear, nonparametric formulation of causal diagrams and their associated structural equation models (SEMs).
2. These models offer a powerful and flexible tool for understanding the structures that underpin causal identification, allowing us to move beyond restrictive linear assumptions.
3. Using these structures, we define counterfactuals—following what Judea Pearl terms the "First Law of Causal Inference," which establishes that SEMs naturally induce these outcomes.

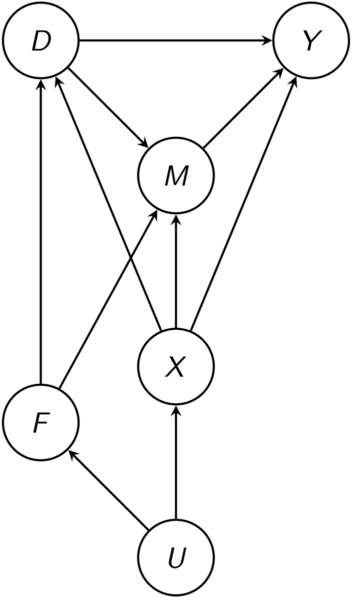
General DAG & SEM via an example



Causal Diagram for "The Impact of 401(k) Eligibility on Financial Wealth:"

Y represents net financial assets; D denotes eligibility for a 401(k) program; X includes observed worker-level covariates (e.g., income); F represents unobserved firm-level covariates; M denotes the employer's matching contribution; and U captures general latent factors.

General DAG & SEM via an example

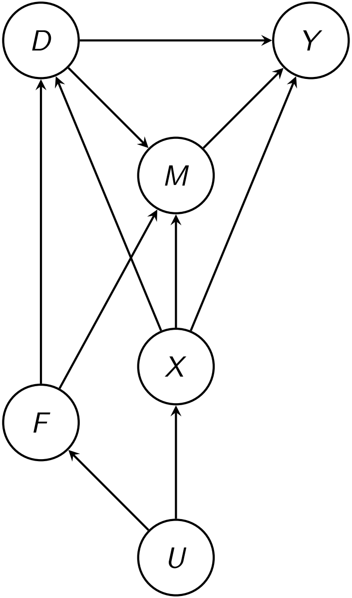


This diagram represents how 401(k) eligibility (D) might affect an individual's net financial assets (Y) both directly and indirectly through the employer's matching contribution (M). It includes observed worker-level characteristics (X), unobserved firm-level characteristics (F), and latent factors (U) that may influence the pathway from eligibility to financial outcomes.

The DAG as a Markovian Model

1. In these graphs, each node represents a random variable (or vector), and an arrow from one node (a "parent") to another (a "child") indicates that the parent directly influences the child, establishing statistical dependency.
2. The Markov property states that each variable is conditionally independent of all non-parents (and non-descendants) given its parents. Consequently, the joint probability distribution can be written as the product of each variable's conditional distribution given its parents.

General DAG & SEM via an example



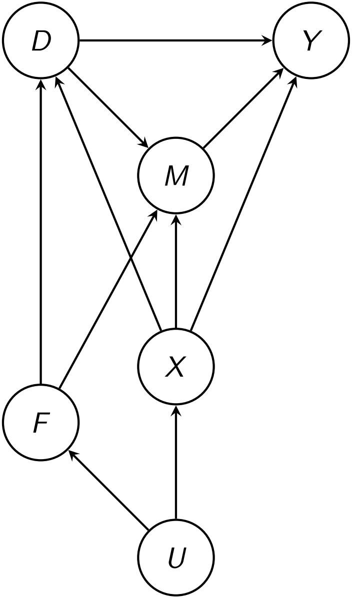
We have:

- U has no parents (a "root" node),
- F, X has parent U,
- D has parents F and X,
- M has parents D, F, and X,
- Y has parents D, M, and X.

According to the Markovian property:

$$
\begin{aligned}
p(u, f, x, d, m, y) = p(u) \\
&\times p(f \mid u) \times p(x \mid u) \\
&\times p(d \mid f, x) \\
&\times p(m \mid d, f, x) \\
&\times p(y \mid d, m, x).
\end{aligned}
$$

The DAG as a Structural Equations Model

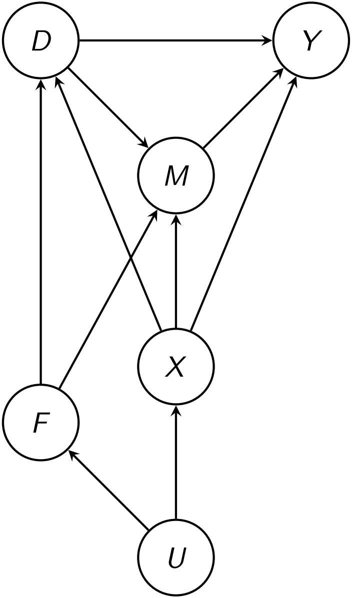


Here, each variable could be defined as a function of its parent variables (as determined by the DAG) and its own exogenous noise ϵ. For instance,

$$Y := f_Y(D, M, X, \epsilon_Y)$$

indicates that net financial assets, Y are determined by eligibility D, the matching contribution M, observed covariates X, and an unobserved shock ϵ_Y. The operator ($:=$) signifies that variables are generated recursively, starting from the root and proceeding through subsequent layers.

The DAG as a Structural Equations Model



We interpret the DAG as implying that (or being implied by) the following system of SEM:

$$
\begin{aligned}
Y &:= f_Y(D, M, X, \epsilon_Y), \\
M &:= f_M(D, F, X, \epsilon_M), \\
D &:= f_D(F, X, \epsilon_D), \\
X &:= f_X(U, \epsilon_X), \\
F &:= f_F(U, \epsilon_F), \\
U &:= \epsilon_U,
\end{aligned}
$$

where $\epsilon_Y, \epsilon_M, \epsilon_D, \epsilon_X, \epsilon_F, \epsilon_U$ are mutually independent stochastic shocks, and f_Y, f_M, f_D, f_X, f_F are structural functions.

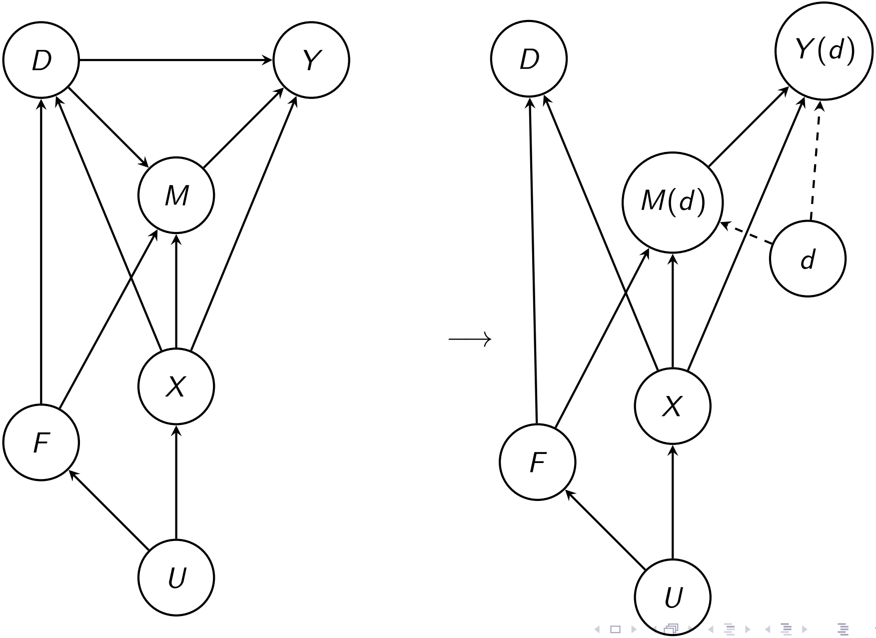
Intervention and Counterfactual DAG and SEM

1. Causality emerges when we introduce the concept of an intervention. In particular, consider intervening by replacing D with a fixed value d in the equations for all descendant.
2. By assumption, the structural equations remain invariant under such an intervention - the essence of their "structural" nature.
3. The complete counterfactual system is:

$$
\begin{aligned}
Y(d) &:= f_Y(d, M(d), X, \epsilon_Y), \\
M(d) &:= f_M(d, F, X, \epsilon_M), \\
D &:= f_D(F, X, \epsilon_D), \\
X &:= f_X(U, \epsilon_X), \\
F &:= f_F(U, \epsilon_F), \\
U &:= \epsilon_U.
\end{aligned}
$$

The original equation for D remains in the model; the intervention fixes $D = d$ to determine $Y(d)$ and $M(d)$.

Intervention and Counterfactual DAG and SEM



Conditional Ignorability/Exogeneity

1. The fact that the SEM implies potential outcomes is known as the First Law of Causal Inference. Moreover, because SEMs/DAGs encapsulate the contextual knowledge of a problem, we can derive the conditional ignorability/exogeneity condition from the model rather than merely postulating it.
2. For example, in our case we deduce that:

$$Y(d) \perp\!\!\!\perp D \mid F, X,$$

which implies that:

$$\mathbb{E}[Y(d) \mid F, X] = \mathbb{E}[Y \mid D = d, F, X],$$

allowing us to identify average causal (or treatment) effects by adjusting (or conditioning on F, X). There are two ways to verify that (F, X) satisfy this condition:

Conditional Ignorability/Exogeneity: Functional (Structural) Argument.

In the counterfactual setting where we fix $D = d$, the relevant structural equations are:

$$Y(d) = f_Y(d, M(d), X, \epsilon_Y)$$

and

$$M(d) = f_M(d, F, X, \epsilon_M).$$

The random variable D is still generated by:

$$D = f_D(F, X, U, \epsilon_D).$$

Once we condition on F and X, the distribution of $Y(d)$ is determined solely by d, $M(d)$, X, and their associated noise terms, and is not influenced by the realized value of D. Formally:

$$Y(d) \perp\!\!\!\perp D \mid F, X.$$

In other words, once F and X are given, knowing D adds no additional information about $Y(d)$.

Conditional Ignorability/Exogeneity: D-Separation (Graphical) Argument

In the counterfactual DAG, $Y(d)$ receives inputs from $M(d)$, X, and the fixed node d. Although D remains in the graph (generated by its usual parents F, X, and U), there is no arrow from D to $Y(d)$. Any path from D to $Y(d)$ must traverse F or X. For example, the paths are:

1. $D \leftarrow X \rightarrow Y(d)$,
2. $D \leftarrow F \rightarrow M(d) \rightarrow Y(d)$,
3. $D \leftarrow F \leftarrow U \rightarrow X \rightarrow Y(d)$.

Conditioning on F and X is said to block these paths (conditioning on a node severs information flow), which then makes D to be d-separated from $Y(d)$ given $\{F, X\}$. By the equivalence between d-separation and conditional independence (called Global Markov property), we conclude that:

$$Y(d) \perp\!\!\!\perp D \mid F, X$$

Conditional Ignorability/Exogeneity: Backdoor Blocking (Graphical) Argument

The goal is to identify a set Z that blocks all backdoor paths between D and Y. A set Z satisfies the backdoor criterion if:

1. No variable in Z is a descendant of D, and
2. Z blocks every backdoor path from D to Y (a backdoor path starts with an arrow into D).

The first rule prevents blocking causal paths from D to Y, such as $D \to M \to Y$. The second rule ensures that conditioning on Z eliminates all non-causal paths that could confound the relationship between D and Y. Thus, conditioned on Z, the statistical association between Y and D only reflects the causal channels.

In the 401(k) diagram, the backdoor paths from D to Y run through F and X:

1. $D \leftarrow X \to Y$,
2. $D \leftarrow F \to M \to Y$,
3. $D \leftarrow F \leftarrow U \to X \to Y$.

By conditioning on both F and X, we block all such paths

Definitions: DAG, Parents, Ancestors, and Descendants

Definition (DAG):
The graph $G = (V, E)$ is a directed acyclic graph (DAG) if it contains no cycles; equivalently, if V is partially ordered by the edge structure E.

Definition (Parents, Ancestors, and Descendants):

1. The parents of X_j are defined as:
$$Pa_j := \{X_k : X_k \to X_j\}.$$
2. The children of X_j are:
$$Ch_j := \{X_k : X_j \to X_k\}.$$
3. The ancestors of X_j are:
$$An_j := \{X_k : X_k < X_j\} \cup \{X_j\}.$$
4. The descendants of X_j are:
$$Ds_j := \{X_k : X_k > X_j\}.$$

Definitions: ASEM, Linear ASEM, and Consistency

1. The **Acyclic SEM** corresponding to the DAG $G = (V, E)$ is the collection of random variables $\{X_j\}_{j \in V}$ satisfying:

$$X_j := f_j(Pa_j, \epsilon_j), \quad j \in V,$$

where the disturbances $(\epsilon_j)_{j \in V}$ are jointly independent.

2. A **linear ASEM** is an ASEM in which the equations are linear:

$$f_j(Pa_j, \epsilon_j) := f_j' Pa_j + \epsilon_j.$$

Here, the functions $\{f_j\}$ are identified with their coefficient vectors. In linear ASEMs, the requirement of independent errors may be relaxed to uncorrelated errors.

3. The **structural potential response processes** for the ASEM corresponding to $G = (V, E)$ are given by:

$$X_j(pa_j) := f_j(pa_j, \epsilon_j), \quad j \in V,$$

viewed as stochastic processes indexed by the potential parental values pa_j.

Definitions: Consistency, Exogenous and Endogenous variables, Markovian Factorisation

1. The observable variables are generated by drawing $\{\epsilon_j\}_{j \in V}$ and then solving the system of equations for $\{X_j\}_{j \in V}$ (Consistency).
2. The stochastic shocks $\{\epsilon_j\}_{j \in V}$ are called **exogenous variables**, while the variables $\{X_j\}_{j \in V}$ are **endogenous**. The latter are determined by the model equations, whereas the former are not.
3. The ASEM $(X_j)_{j \in V}$ associated with a DAG $G = (V, E)$ satisfies the following equivalent properties:
   3.1 $$p(\{x_\ell\}_{\ell \in V}) = \prod_{\ell \in V} p(x_\ell \mid pa_\ell).$$
   3.2 Each variable is independent of its non-descendants given its parents (Local Markov Property).

Definitions: Paths, Blocked Paths, and d-Separation

Paths and Backdoor Paths on DAGs:

1. A directed path is a sequence:

$$X_{v_1} \rightarrow X_{v_2} \rightarrow \cdots \rightarrow X_{v_m}.$$

2. A non-directed path is a path in which some, but not all, arrows are replaced by $\leftarrow$.
3. A node X_j is a collider on a path if the path includes a subpath of the form:

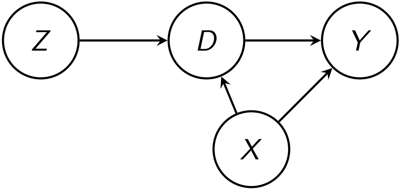
$$\rightarrow X_j \leftarrow .$$

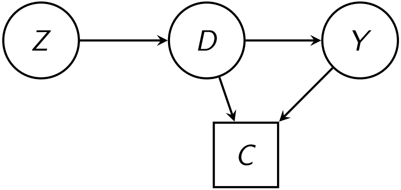
4. A backdoor path from X_l to X_k is a non-directed path that starts at X_l and ends with an arrow into X_k.

Definitions: Paths, Blocked Paths, and d-Separation

1. **Blocked Paths:** A path that is not blocked is open. A path π is blocked by a set of nodes S if either:
   - π contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ with $m \in S$, or
   - π contains a collider $i \rightarrow m \leftarrow j$ with $\{m\} \bigcup D_m \notin S$.

   The backdoor path $Y \leftarrow X \rightarrow D$ is blocked by setting $S = X$.
2. **Opening a Path by Conditioning:** A path containing a collider is opened by conditioning on that collider or one of its descendants. The path $Y \rightarrow C \leftarrow D$ is blocked (by the empty set) but becomes open when conditioned on the collider C.





Definitions: Paths, Blocked Paths, and d-Separation

d-Separation: Given a DAG G, a set of nodes S d-separates nodes X and Y if S blocks all paths between X and Y. We denote this as:

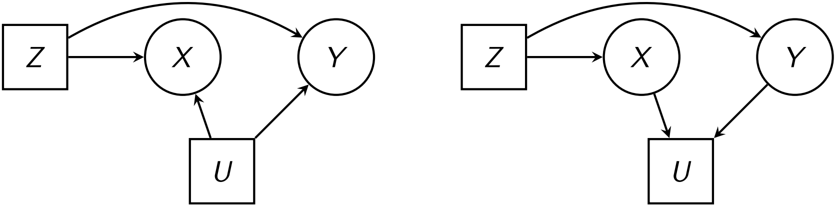
$$(X \perp\!\!\!\perp_d Y|S)_G$$

In which case,

$$X \perp\!\!\!\perp Y|S \quad \textit{(Pearl and Verma)}$$

Intuitively, conditioning on S interrupts the information flow between X and Y, impossible to predict each other given S. The formal proof is nontrivial. The converse does not generally.

Definitions: Paths, Blocked Paths, and d-Separation



We illustrate how d-separation implies conditional independence:

1. In Figure left, X and Y are d-separated by $S = \{Z, U\}$ since S blocks all paths between them. By Markov factorization:

   $$p(y, x \mid u, z) = p(y \mid x, z, u)\, p(x \mid z, u) = p(y \mid u, z)\, p(x \mid z, u),$$

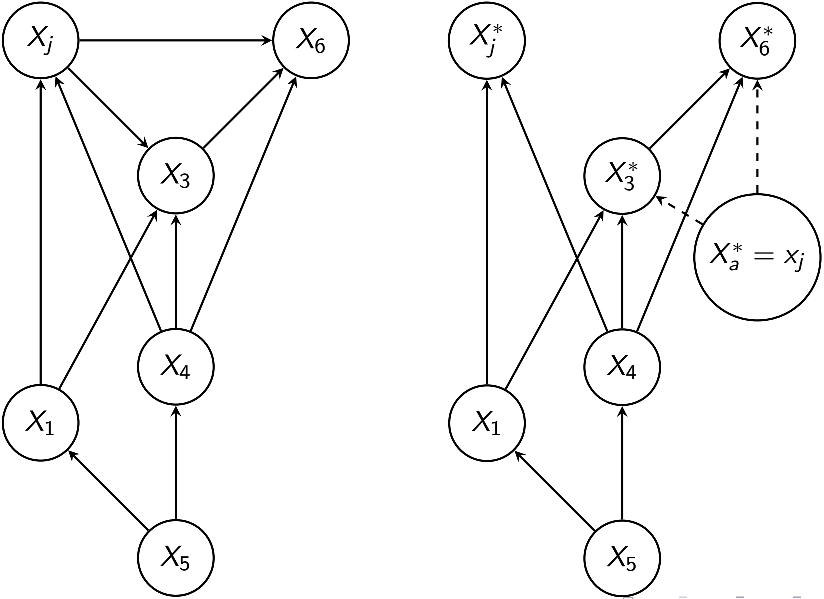
   implying $X \perp\!\!\!\perp Y \mid Z, U$.

2. In Figure right, X and Y are d-separated by $S = \{Z\}$, and similarly:

   $$p(y, x \mid z) = p(y \mid z)\, p(x \mid z),$$

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

Intervention and Counterfactual DAG and SEM



Counterfactuals and Identification by Conditioning

The intervention $\text{fix}(X_j = x_j)$ on an ASEM creates a counterfactual ASEM (CF-ASEM) defined by a modified DAG, known as a Single World Intervention Graph (SWIG):

$$\widetilde{G}(x_j) := (V, \widetilde{E}),$$

along with a collection of counterfactual variables $\{X_k^*\}_{k \in V} \cup \{X_a^*\}$.
Here, the node X_j is split into two distinct entities:

- $X_j^* := X_j$, representing the natural value.
- A new deterministic node a with $X_a^* := x_j$, representing the intervened value.

Counterfactuals and Identification by Conditioning

The construction proceeds as follows:

1. The intervention node X_a^* inherits only the outgoing edges from X_j (i.e., $\widetilde{e}_{ai} = e_{ji}$ for all i) and has no incoming edges ($\widetilde{e}_{ia} = 0$ for all i), reflecting that it is fixed by the intervention.
2. The node X_j^* inherits only the incoming edges from X_j (i.e., $\widetilde{e}_{ij} = e_{ij}$ for all i) and has no outgoing edges ($\widetilde{e}_{ji} = 0$ for all i), preserving its dependence on its original causes.
3. All remaining edges are preserved: $\widetilde{e}_{ik} = e_{ik}$ for all i and for all $k \neq j, k \neq a$, ensuring the rest of the graph structure remains intact.
4. The counterfactual variables are assigned according to:

$$X_k^* := f_k(Pa_k^*, \epsilon_k), \quad \text{for } k \neq a,$$

where Pa_k^* denotes the parents of X_k^* under $\widetilde{E}$, adapting the structural equations to the new graph.

Theorem: A Counterfactual Criterion for Identification by Conditioning

Consider any ASEM with DAG, G. Re-label the treatment node X_j as D, and let Y be any descendant of D representing the outcome. Construct the SWIG $\widetilde{G}(d)$ induced by the fix$(D = d)$ intervention, and let S be any subset of nodes common to both G and $\widetilde{G}(d)$ such that $Y(d)$ is d-separated from D by S in $\widetilde{G}(d)$. Then:

1. The conditional exogeneity condition holds:

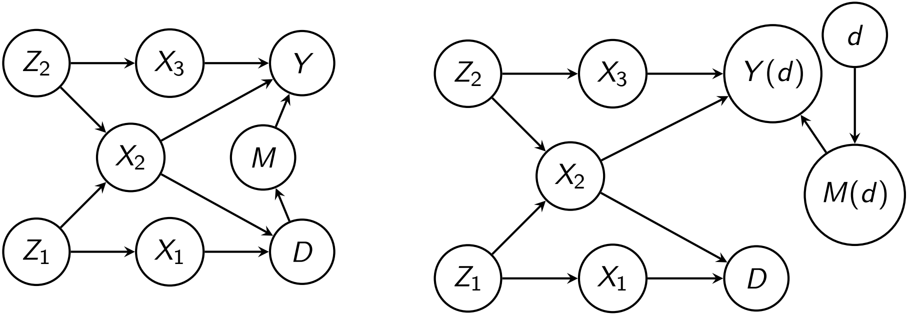
$$Y(d) \perp\!\!\!\perp D \mid S.$$

2. The conditional average potential outcome is identified by the corresponding regression:

$$E\left[g(Y(d)) \mid S = s\right] = E\left[g(Y) \mid D = d, S = s\right],$$

for all s with $p(s, d) > 0$ and for all bounded functions g.

Pearl's example



Consider the DAG above (left) and The SWIG induced by the $\mathrm{fix}(D = d)$ intervention is shown in right. We aim to estimate the causal effect of D on Y, i.e., the mapping $d \mapsto Y(d)$. Here,

$$S = \{\{X_1, X_2\}, \quad \{X_2, X_3\}, \quad \{X_2, Z_2\}, \quad \{X_2, Z_1\}\}$$

Conditioning on X_2 alone is insufficient because, although it blocks the inner backdoor paths, it opens an outer path where X_2 acts as a collider; adding X_1, X_3, Z_1, or Z_2 blocks this additional path, ensuring ignorability.

A useful Remark

A surprisingly useful limitation of the counterfactual DAG approach is that it avoids selecting valid yet unhelpful control variables for adjustment. Consider the simple DAG

$$Z \leftarrow D \rightarrow Y,$$

and its corresponding counterfactual DAG

$$Z(d) \leftarrow d \rightarrow Y(d).$$

Under the counterfactual approach, no adjustment is required—in other words, the empty set is a valid adjustment set:

$$Y(d) \perp\!\!\!\perp D.$$

A useful Remark

However, we know that Z is a valid control. Its validity can be deduced by considering a cross-world DAG that combines factual and counterfactual variables from the respective ASEMs:

$$\begin{tikzcd}
\epsilon_z \arrow[r] \arrow[dr] & Z & D \arrow[l] \arrow[r] & Y & \epsilon_y \arrow[l] \arrow[dl] \\
 & Z(d) & d \arrow[l] \arrow[r] & Y(d) &
\end{tikzcd}$$

In this cross-world DAG, $Y(d)$ is d-separated from D by Z, so that

$$Y(d) \perp\!\!\!\perp D \mid Z.$$

Thus, while Z is a valid control variable, it is arguably superfluous, as it does not add useful information about $Y(d)$; in fact, adjusting for Z can reduce the precision of our estimates of the average causal effect of D on Y.

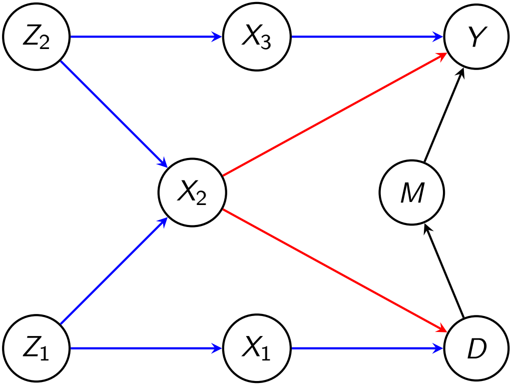
Theorem: Backdoor Criterion

A backdoor path is one that starts at D and ends with an arrow pointing into D, indicating confounding influences. Consider any ASEM and its associated DAG. Re-label a treatment node X_j as D, and let Y, an outcome of interest, be any descendant of D. An adjustment set S is valid - meaning it implies conditional ignorability, i.e.,

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

if the backdoor criterion is satisfied. That is, no element of S is a descendant of D, and all backdoor paths from Y to D are blocked by S.

Pearl's Example using Backdoor Criterion



In the DAG above, there are two backdoor paths from D to Y:

(i) $D \leftarrow X_2 \rightarrow Y$,

(ii) $D \leftarrow X_1 \leftarrow Z_1 \rightarrow X_2 \leftarrow Z_2 \rightarrow X_3 \rightarrow Y$.

Pearl's Example using Backdoor Criterion

The two backdoor paths from D to Y:

$$\text{(i) } D \leftarrow X_2 \rightarrow Y,$$

$$\text{(ii) } D \leftarrow X_1 \leftarrow Z_1 \rightarrow X_2 \leftarrow Z_2 \rightarrow X_3 \rightarrow Y.$$

Conditioning on X_2 alone blocks the inner backdoor path but opens an outer backdoor path (since X_2 acts as a collider in that sequence). To block the outer backdoor path, we must also condition on an additional variable (e.g., X_1, X_3, Z_1, or Z_2). Thus, valid adjustment sets include, for instance:

$$S_1 = \{X_1, X_2\} \quad \text{or} \quad S_2 = \{X_2, X_3\}.$$

Note that conditioning on M is invalid because M is a descendant of D (an intermediate outcome) which can bias the effect estimate.

Comments on Adjustment: Valid Yet Unhelpful Controls

The backdoor criterion systematically yields minimal adjustment sets for identification. However, it may not capture every valid set. For example, consider the simple DAG:

$$Z \leftarrow D \rightarrow Y.$$

Here, conditioning on Z does not satisfy the backdoor criterion (since Z is a descendant of D), yet Z is a valid control. In this case, D directly causes Y without confounding, so there is no need to adjust for Z. Adjusting for Z may even lower the precision of the estimated effect. This limitation is useful as it helps disregard controls that, while valid, do not add any meaningful information for identifying the causal effect. The same observation applies within the counterfactual approach.