Chapter 1

Introduction to causality and structural causal models

The notion of *causation* has long been a subject of great interest to natural philosophers and scientists. As often happens in the development of an idea, the notion of causation has been adapted from its usage in natural language into a mathematically rigorous concept which captures much of its meaning in natural language. One of the goals of this course is to expose students to one of the main frameworks using for formalizing causality, the *Structural Causal Model (SCM)* framework.

To ground the abstract mathematical definitions that will follow, we will begin with an example, adapted from the pioneering work of Sewall Wright on *path diagrams*, a forebear to structural causal models. Pay attention to this example, as we will use it throughout this lecture as a way to ground the abstract definitions that are essential for the rest of the course.

Example 1.1. *You are a biologist studying the heredity of weight in mice. You are focused on a single gene which you call the* WEIGHT *gene. You have two pools of mice: Pool 0, which contains only mice who have a non-mutated* WEIGHT *gene, and Pool 1, which contains only mice who have a mutated* WEIGHT *gene. You've obtained a grant to run the following experimental procedure 100 times. Here are the instructions for the i-th experiment:*

1. *Generate a number uniformly at random between* 0 *and* 1*. Construct a binary variable* $x_1^{(i)}$ *as follows:* $x_1^{(i)} = 1$ *if the number is less than 0.5, and* $x_1^{(i)} = 0$ *otherwise. Record the value of* $x_1^{(i)}$*.*
2. *Randomly select one male mouse from Pool* $x_1^{(i)}$*. For easy reference, call the male mouse "Mickey-i". Record the weight of Mickey-i as* $x_2^{(i)}$*.*
3. *Randomly select one female mouse from Pool* $x_1^{(i)}$*. For easy reference, call the male mouse "Minnie-i". Record the weight of Minnie-i as* $x_3^{(i)}$*.*
4. *Next, breed Mickey-i and Minnie-i, and call their child "Cheddar-i". Record the weight of Cheddar-i as* $x_4^{(i)}$*.*
5. *Now, randomly select a mouse (of the opposite gender of Cheddar-i) with equal probability from the two pools. Breed Cheddar-i with this mouse, and call their offspring "Gouda-i". Record the weight of Gouda-i as* $x_5^{(i)}$*.*

Throughout the course, we will take a probabilistic perspective on such experiments. In particular, we think of the recorded quantities $\left(x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, x_4^{(i)}, x_5^{(i)}\right)$ as independent samples coming from some joint distribution $\mathbb{P}_{\mathcal{X}}$ over **random variables** $\mathcal{X} = \{X_1, X_2, X_3, X_4, X_5\}$. The random variables associated to this example are:

- X_1, the value of the binary variable selected in Step 1.
- X_2, the weight of the male mouse "Mickey" selected in Step 2.
- X_3, the weight of the female mouse "Minnie" selected in Step 3.
- X_4, the weight of the child mouse "Cheddar" bred in Step 4.
- X_5, the weight of the grandchild mouse "Gouda" bred in Step 5.

1.1 The ladder of causation

Let us now consider some questions that we can ask about the experiment. For now, these questions will be expressed in plain English. In the next section, we will see how the questions can be formalized into mathematical expressions using different mathematical *frameworks* for causal reasoning.

Now, let us consider what questions you might ask.

Question 1. Say you ran experiment i, but only recorded the weights of Mickey-i and Cheddar-i.

If Mickey-i weighs 20 grams, Cheddar-i weighs 23 grams, what is the expected weight of Minnie-i?

This is a question about the **conditional expectation** $\mathbb{E}_{\mathcal{X}}[X_3 \mid X_2 = 20, X_4 = 23]$, where $\mathbb{E}_{\mathcal{X}}$ denotes that the conditional expectation is in the distribution $\mathbb{P}_{\mathcal{X}}$. Question 1 is a **statistical** (or *associational*) question, and can be answered without appealing to any notions of causality. The question asks about our beliefs based only on some information that we have *seen*. Let us now consider a question that ask about our beliefs for the future, based not on what we have *seen*, but on what we could *do*.

Question 2. Say we are thinking of changing the experimental procedure for a future set of experiments.

If we change Step 2 so that Mickey-i always weighs 25 grams, what is the expected weight of Gouda-i?

By changing the experimental procedure, we are asking a question about a *different* joint distribution $\mathbb{P}^I_{\mathcal{X}}$. As we will see, Question 2 about **interventions** (or *causality-in-mean*), which cannot be answered by standard statistical reasoning using only the distribution $\mathbb{P}_{\mathcal{X}}$. The question asks about how a future action would affect future observations. Finally, we may ask questions about how taking a different action *in the past* would have affected our past observations.

Question 3. Say that the WEIGHT gene was not mutated in experiment i.

Suppose Cheddar-i weighed 24 grams. How much would Cheddar-i have weighed, if the WEIGHT gene had been mutated in experiment i?

This is a question about a **counterfactual**. It does not only involve reasoning about what we have *seen* or what we could *do*, but mixes the two in a non-trivial way.

These three types of questions make up the "three rungs" of Judea Pearl's "ladder of causation". To answer questions at "higher" rungs of the ladder, more sophisticated mathematical frameworks are required: a mathematical framework can formally express questions at Rungs 1 and 2, without being able to express questions at Rung 3. Let us now investigate a few different frameworks.

1.2 Frameworks for causal reasoning

1.2.1 Why multiple frameworks?

First, let us justify why we are spending time to introduce different frameworks. In the course, we will only use one of these frameworks. Why not jump into that one and get started? To understand this pedagogical choice, it helps to understand why there are multiple frameworks in the first place. There are at least three reasons to cite.

The historical reason. Many fields of science and engineering are concerned with questions which are intrinsically causal. In biostatistics and medicine, researchers investigate the effect of treatments, such as chemotherapy, on the health outcomes of patients. In econometrics, researchers investigate the effect of public policies, such as compulsory high school education, on the economic success of individuals. In computer science - in particular within the fields of artificial intelligence and automated reasoning - researchers try to build autonomous systems which can predict the effects of their own actions on their environment.

Each of these fields, and others, found the need to develop a mathematical formalism with which to express their causally-centered questions. Since different fields emphasize different questions and provide different mathematical training, it is natural that they would develop different frameworks.

The philosophical reason. Although the reasons for initial existence of different formalisms may be contingent, the reasons for their persistence are longer-lasting. Different formalisms suggest different ways of thinking about the world. Over time, these ways of thinking about the world often transform into deep-rooted conceptual intuitions, and even metaphysical commitments. Two of the main divisions between frameworks are about the fundamentality of counterfactuals and about the precise definition of an intervention.

The potential outcomes framework treats counterfactuals as first-order objects, upon which the rest of the framework is built. In contrast, the structural causal model framework *constructs* counterfactuals, building from more primitive objects, such as a structural causal model and an intervention upon it.

The structural causal model framework, in its simplest form, makes the implicit assumption that the notion of intervening on a variable is well-defined. This assumption may raise issues when contemplating "interventions" on variables such as an individual's gender, which has complex interactions with other traits of the individual, such as social perceptions and sexual identity.

The practical reason. Depending on your intuitions and metaphysical beliefs, you may find it easier to think in one framework than another. However, you should not conclude that it is only worthwhile to think in terms of a single framework. Having different ways to think about a problem is not an unfortunate annoyance, it is an important tool for problem-solving. The ability to formalize a theory in different ways has several practical advantages and can be used to understand connections between problems that initially appear quite distinct. A prime example comes from physics, where different formalizations such as the Lagrangian and Hamiltonian formulations of classical mechanics are used to solve a range of problems.

For the sake of this course, we will focus on the structural causal model framework for causal reasoning. The choice to focus on a single framework will allow us to cover a wider range of topics, but it is not meant to suggest that this framework is the only one worth learning. You are encouraged to learn the other frameworks, and use the one most well-suited to your own purposes.

1.2.2 A brief summary of other frameworks

To compare the frameworks, our example from Section 1.1 is too complicated. We will consider an example which is both simpler, and more representative of a real-world task.

Example 1.2. *Suppose that you are a doctor specializing in a rare disease, for which there is only one treatment. The treatment has some unpleasant side effects, and might not be effective for all people. Thus, you must decide which patients you will give the treatment, and which ones you will not. Your workflow for each patient i looks like this:*

- *Look at the patient's info sheet, which provides the same set of covariates* $\mathbf{c}^{(i)}$ *for each patient.*
- *Decide whether to give the treatment to the patient* ($t^{(i)} = 1$) *or not* ($t^{(i)} = 0$).
- *In a six-month follow-up, record whether the treatment was "successful", i.e., whether the patient is free of symptoms* ($y^{(i)} = 1$) *or not* ($y^{(i)} = 0$).

In order to make better treatment decisions, you and other doctors specializing in this disease, who all follow the same workflow, have compiled a dataset $\{(\mathbf{c}^{(i)}, t^{(i)}, y^{(i)})\}_{i=1}^{n}$ *of* n *previous patients.*

In each framework, we will assume that each sample $(\mathbf{c}^{(i)}, t^{(i)}, y^{(i)})$ is drawn from some distribution $\mathbb{P}_{\mathcal{X}}$ over random variables $\mathcal{X} = \{\mathbf{C}, T, Y\}$. In the potential outcomes framework, $\mathbb{P}_{\mathcal{X}}$ is viewed as the marginal of a distribution which contains other variables. In the structural causal model framework, $\mathbb{P}_{\mathcal{X}}$ is not viewed as a marginal distribution, and is used as a starting point from which to construct other distributions.

The potential outcomes (PO) framework. In this framework, patient i is associated with two "potential" outcomes: $y_0^{(i)}$, their outcome in the world where they were not given the treatment, and $y_1^{(i)}$, their outcome in the world where they were given the treatment. Then, their "factual" outcome $y^{(i)}$ depends on the value of $t^{(i)}$: if $t^{(i)} = 0$, then $y^{(i)} = y_0^{(i)}$, and if $t^{(i)} = 1$, then $y^{(i)} = 0$.

Thus, patient i is associated to five values $(\mathbf{c}^{(i)}, t^{(i)}, y_0^{(i)}, y_1^{(i)}, y^{(i)})$, sampled from a distribution over random variables $\mathbf{C}, T, Y_0, Y_1$, and Y. However, we only observe their values $(\mathbf{c}^{(i)}, t^{(i)}, y^{(i)})$, and their "counterfactual" outcome is viewed as missing data. The expected value of Y, if treatment $t \in \{0, 1\}$ was *always* assigned for every patient, is thus $\mathbb{E}[Y_t]$.

The structural causal model (SCM) framework. In this framework, patient i is only associated with a single outcome $y^{(i)}$. To express the effect of the intervention I that *always* assigns the treatment $t \in \{0, 1\}$ to every patient, we construct a new distribution $\mathbb{P}_{\mathcal{X}}^{I}(\mathbf{C}, T, Y)$. $\mathbb{P}_{\mathcal{X}}^{I}$ is closely related to the original distribution $\mathbb{P}_{\mathcal{X}}$, in a way that depends on a *causal graph* between the variables $\mathbf{C}, T$, and Y.

The expected value of Y under this intervention is $\mathbb{E}^{I}[Y]$, where $\mathbb{E}^{I}$ denotes that the expectation is taken with respect to the distribution $\mathbb{P}_{\mathcal{X}}^{I}$. This value can be expressed without appealing to counterfactual outcomes, which are derived using a combination of interventions and conditioning.

Other frameworks. There are many frameworks for causal reasoning, with subtle variations between themselves. Two are worth mentioning here. The first is the Single World Intervention Graph (SWIG) framework, which defines a "graphical" version of the PO framework and allows one to translate between the PO and SCM frameworks. The second is the Decision-theoretic (DT) framework, which does not define counterfactuals at all, only aiming to formalize up to Rung 2 on the ladder of causation. Pedagogically, it will actually be helpful for us to start with a framework that is quite similar to the DT framework, and see why counterfactuals are not defined within that framework.

1.3 Causal Bayesian networks

We start with an example of a Bayesian network for Example 1.1.

Example 1.3. *Suppose that:*

1. *The distribution of a Bernoulli random variable with probability* p *of equaling 1 is denoted* $\mathsf{Ber}(p)$.
2. *The weight of a male mouse with a non-mutated weight gene follows the distribution* $\mathcal{N}(25, 1)$, *and the weight of a male mouse with a mutation weight gene follows the distribution* $\mathcal{N}(27, 1)$.
3. *Similarly, suppose that the weight of a female mouse with a non-mutated weight gene follows the distribution* $\mathcal{N}(20, 1)$, *and the weight of a male mouse with a mutation weight gene follows the distribution* $\mathcal{N}(22, 1)$.

4. *Next, given two mice with weights a and b, suppose that the weight of their offspring follows the distribution $\mathcal{N}(\frac{1}{2}(a+b), 1)$.*

5. *Finally, for a single mouse of weight c mated with a random mouse of the opposite gender, suppose that the weight of their offspring follows the distribution $\mathcal{N}(c, 2)$.*

Then the Bayesian network that describes our running example is:

$$\begin{aligned} X_1 &\sim \mathsf{Ber}(0.5) \\ X_2 \mid X_1 &\sim \mathcal{N}(25 + 2 \cdot X_1, 1) \\ X_3 \mid X_1 &\sim \mathcal{N}(20 + 2 \cdot X_1, 1) \\ X_4 \mid X_2, X_3 &\sim \mathcal{N}\left(1/2(a+b), 1\right) \\ X_5 \mid X_4 &\sim \mathcal{N}(X_4, 1) \end{aligned} \tag{1.1}$$

In particular, the joint distribution over $\mathcal{X}$ is

$$\begin{aligned} \mathbb{P}_{\mathcal{X}}(X_1, X_2, X_3, X_4, X_5) = \mathsf{Ber}(X_1; 0.5) &\times \mathcal{N}(X_2; 25 + 2X_1, 1) \times \mathcal{N}(X_3; 20 + 2X_1, 1) \\ &\times \mathcal{N}(X_4; 1/2(X_2 + X_3), 1) \times \mathcal{N}(X_5; X_4, 2) \end{aligned}$$

1.3.1 Directed acyclic graphs

Both causal Bayesian networks and structural causal models are defined with respect to directed graphs. For simplicity, we will consider only directed *acyclic* graphs (DAGs), which are directed graphs that contain no directed cycles. In both cases, the nodes of the directed graph are in one-to-one correspondence with the random variables $\mathcal{X}$, and a directed edge $X_i \to X_j$ for $X_i, X_j \in \mathcal{X}$ has the interpretation that the variable X_i has a direct causal influence on the variable X_j.

Graph-theoretic notation. In graph theory, DAGs have a standard notation that we will use here: given a DAG $\mathcal{G}$ and a node V_i in $\mathcal{G}$, we denote $\mathrm{pa}_{\mathcal{G}}(V_i)$ to denote its **parents**, and $\mathrm{ch}_{\mathcal{G}}(V_i)$ to denote its **children**. We use $\mathrm{an}_{\mathcal{G}}(V_i)$ to denote the **ancestors** of V_i and $\mathrm{de}_{\mathcal{G}}(V_i)$ to denote its **descendants**. We will add a bar to indicate an "inclusive" version of these sets, e.g. $\overline{\mathrm{pa}}_{\mathcal{G}}(V_i) = \mathrm{pa}_{\mathcal{G}}(V_i) \cup \{V_i\}$.

1.3.2 Bayesian networks

We will define a Bayesian network as a mathematical model consisting of three components: a **signature** which describes the variables in the model, a **causal graph** which defines the (qualitative) relationships between the variables, and a set of **causal conditionals**, which define the relationships between variables in a more quantitative way.

Often, one does not need to separately define the signature of a Bayesian network, since it is generally clear from context. To give a rigorous treatment, we will define the signature here, but we will generally let it be inferred from context throughout the rest of the course.

Definition 1.1. *A **signature** $\mathcal{S}$ consists of:*

- *A set $\mathcal{X}$ of variables, and*
- *A **range function** $\mathcal{R}$ which maps each variable $X_i \in \mathcal{X}$ to its **alphabet** $\mathcal{R}(X_i)$.*

The signature for our running example has $\mathcal{R}(X_1) = \{0, 1\}$ and $\mathcal{R}(X_i) = \mathbb{R}$ for $i = 2, 3, 4, 5$.

For convenience, we will overload the notation on the range function to take a set as input, i.e., for a set $\mathbf{V} \subseteq \mathcal{X}$, we define

$$\mathcal{R}(\mathbf{V}) := \times_{V \in \mathbf{V}} \mathcal{R}(V)$$

where $\times$ denotes the Cartesian product. The causal graph of a Bayesian network is simply a DAG with nodes $\mathcal{X}$. Finally, the causal conditionals of a Bayesian network will relate the causal graph to a distribution over random variables $\mathcal{X}$.

Definition 1.2. *Let $\mathcal{S} = (\mathcal{X}, \mathcal{R})$ be a signature and let $\mathcal{G}$ be a causal graph over variables $\mathcal{X}$. A **causal conditional** for variable $X_i \in \mathcal{X}$ is a conditional probability distribution $\mathbb{P}_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i))$.*

Now, sticking these components together, we get a Bayesian network:

Definition 1.3. *A **Bayesian network** $\mathcal{BN}$ is comprised of:*

- *a signature $\mathcal{S} = (\mathcal{X}, \mathcal{R})$,*
- *a causal graph $\mathcal{G}$ over variables $\mathcal{X}$, and*
- *an indexed set $\{\mathbb{P}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i))\}_{X_i \in \mathcal{X}}$ of causal conditionals.*

The **entailed distribution** *of a $\mathcal{BN}$ is*

$$\mathbb{P}_{\mathcal{X}}(\mathcal{X}) = \prod_{X_i \in \mathcal{X}} \mathbb{P}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i))$$

Now, it is clear that our running example is a Bayesian network with causal graph $\mathcal{G} = \{X_1 \to X_2, X_1 \to X_3, X_2 \to X_4, X_3 \to X_4, X_4 \to X_5\}$.

1.3.3 Interventions

Let us return to Question 2. In the new experimental procedure, we replace Step 2 of the original experimental procedure with Step 2', which always picks a male mouse with a weight of 25 grams. The new experimental procedure can be modeled as taking samples from a Bayesian network with the following causal conditionals:

$$\begin{aligned} X_1 &\sim \mathsf{Ber}(0.5) \\ X_2 &\sim \delta_{25} \\ X_3 \mid X_1 &\sim \mathcal{N}(20 + 2 \cdot X_1, 1) \\ X_4 \mid X_2, X_3 &\sim \mathcal{N}\left(\tfrac{1}{2}(a+b), 1\right) \\ X_5 \mid X_4 &\sim \mathcal{N}(X_4, 1) \end{aligned}$$

where δ_x denotes a Dirac distribution which assigns all probability mass to the value x.

Notably, all causal conditionals in the new Bayesian network are equal to the causal conditionals in the original Bayesian network, except for the causal conditional of X_2. This notion of creating a new Bayesian network from another Bayesian network is formally captured in the definition of an *intervention*.

Definition 1.4. *Let $\mathcal{BN}$ be a Bayesian network with signature $\mathcal{S} = (\mathcal{X}, \mathcal{R})$, causal graph $\mathcal{G}$, and causal conditionals $\{\mathbb{P}_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i))\}_{X_i \in \mathcal{X}}$. An **intervention** I on $\mathcal{BN}$ consists of:*

- *A set $T(I) \subseteq \mathcal{X}$ of **intervention targets**, and*
- *An indexed set $\{\mathbb{P}^I_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}})\}_{X_i \in T(I)}$ of **interventional causal conditionals** for each intervention target.*

*Then, the **intervened Bayesian network** $\mathcal{BN}^I$ is a Bayesian network with the same signature and the same causal graph as $\mathcal{BN}$, but with causal conditionals*

$$\{\mathbb{P}^I_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}})\}_{X_i \in T(I)} \cup \{\mathbb{P}_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i)\}_{X_i \in \mathcal{X} \setminus T(I)}$$

.

Finally, let us also point out some important terminology when dealing with interventions. The interventions that we have defined are known as **soft** or **imperfect** interventions in the literature, and are the most general class of interventions that are normally considered. Two more restricted classes of interventions are also widely studied.

Definition 1.5. *A* **hard** *or* **perfect** *intervention is an intervention I where each of the intervened variables does not depend on its parents, i.e., for $X_i \in T(I)$, we have $\mathbb{P}_{\mathcal{X}}(X_i \mid \text{pa}_{\mathcal{G}}(X_i)) = \mathbb{P}_{\mathcal{X}}(X_i)$.*

A **do**-*intervention is a perfect intervention I where $\mathbb{P}_{\mathcal{X}}(X_i)$ is a Dirac distribution for each $X_i \in T(I)$. In the case of a do-intervention with $T(I) = \mathbf{A}$, setting the values of $\mathbf{A}$ deterministically to $\mathbf{a}$, we denote the interventional distribution as $\mathbb{P}_{\mathcal{X}}(\mathcal{X} \mid \texttt{do}(\mathbf{A} = \mathbf{a}))$.*

1.3.4 Intervention-augmented graphs

It is natural to treat an intervention as a transformation of a Bayesian network, as done in Definition 1.12. However, from a technical standpoint, it is often convenient to treat an intervention as an expansion of a Bayesian network into a new model which encapsulates both the original Bayesian network and the intervened Bayesian network, as we now describe. First, we introduce an expansion on the signature:

Definition 1.6. *Let $\mathcal{S} = (\mathcal{X}, \mathcal{R})$ be a signature and let I be an intervention. The* **interventional signature** *$\mathcal{S}^I$ has:*

- *Variables $\mathcal{X}^I := \mathcal{X} \cup \{\zeta^I\}$, for a new variable ζ^I called the* **intervention indicator***, and*
- *a range function $\mathcal{R}^I$, which matches $\mathcal{R}$ with the additional criterion that $\mathcal{R}^I(\zeta^I) = \{0, 1\}$*

Next, we introduce an expansion on the causal graph:

Definition 1.7. *Let $\mathcal{S}$ be a signature and $\mathcal{G}$ be a causal graph and let I be an intervention. Then the* **interventional causal graph***, denoted $\mathcal{G}^I$, is a DAG with nodes $\mathcal{X}^I$, and the following edges:*

- *For $X_i, X_j \in \mathcal{X} \cup \mathcal{E}$, let $X_i \to X_j$ if and only if $V_i \to V_j$ in $\mathcal{G}$, and‘*
- *$\zeta^I \to X_i$ for all $X_i \in T(I)$.*

From these pieces, we may define a new Bayesian.

Definition 1.8. *Let $\mathcal{BN}$ be a Bayesian network with signature $\mathcal{S}$, causal graph $\mathcal{G}$, causal mechanisms $\{f_{X_i}\}_{X_i \in \mathcal{X}}$, and exogenous distribution $\mathbb{P}_{\mathcal{E}}$. Let I be an intervention on this Bayesian network, with intervention targets $T(I)$ and interventional causal conditionals $\{\mathbb{P}^I_{\mathcal{X}}(X_i \mid \text{pa}_{\mathcal{G}}(X_i)\}_{X_i \in T(I)}$. Let $\mathbb{P}_{\zeta^I}$ be a Bernoulli distribution over whether or not the intervention is performed.*

The **expanded interventional Bayesian network** *is a Bayesian network $\mathcal{BN}^I_+$, where $\mathcal{BN}^I_+$ has*

- *Signature $\mathcal{S}^I$,*
- *Causal graph $\mathcal{G}^I$, and*
- *Causal conditionals as follows:*
  - *$\mathbb{P}_{\zeta_I}$ for ζ^I, and*
  - *$\mathbb{P}(X_i \mid \text{pa}_{\mathcal{G}^I}(X_i)) = \mathbb{P}(X_i \mid \text{pa}_{\mathcal{G}^I}(X_i))$ for all $X_i \notin T(I)$*
  - *$\mathbb{P}(X_i \mid \text{pa}_{\mathcal{G}^I}(X_i)) = \mathbb{P}(X_i \mid \text{pa}_{\mathcal{G}^I}(X_i))^{\mathbb{1}_{\zeta^I=0}} \mathbb{P}^I(X_i \mid \text{pa}_{\mathcal{G}^I}(X_i))^{\mathbb{1}_{\zeta^I=1}}$ for all $X_i \in T(I)$*

The next claim formalizes the fact that $\mathcal{BN}^I_+$ encapsulates the interventional SCM:

Claim 1.1. *Let $\mathbb{P}^I_{\mathcal{X}}$ be the entailed distribution of M^I, and let $\mathbb{P}_{\mathcal{X}^I}$ be the entailed distribution of M^I_+. Then*

$$\mathbb{P}^I_{\mathcal{X}}(\mathcal{X}) = \mathbb{P}_{\mathcal{X}^I}(\mathcal{X} \mid \zeta^I = 1)$$

1.4 Structural causal models

While causal Bayesian networks give us a formalism for answering Question 2, they do not provide the ability to answer *counterfactual* questions such as Question 3. For those, we will need to upgrade causal Bayesian networks into structural causal models. As in the previous section, we will start by giving a structural causal model associated with Example 1.1.

Example 1.4. *W process that gives rise to the causal conditionals given in Equation (1.1).*

1. *As initially described, the binary variable X_1 is generated by first sampling a variable ε_1 from a uniform distribution over the interval $[0, 1]$, followed by thresholding.*
2. *Mickey's weight, X_2, is generated by first sampling a variable $\varepsilon_2 \sim \mathcal{N}(25, 1)$, representing the genetic variation in weight due to genes besides the WEIGHT gene. If Mickey's WEIGHT gene is not mutated, then Mickey's weight is equal to ε_2, but if his WEIGHT gene is mutated, then Mickey's weight is equal to $\varepsilon_2 + 2$.*
3. *In a similar fashion, Minnie's weight, X_3, is generated by sampling $\varepsilon_3 \sim \mathcal{N}(20, 1)$, with $X_3 = \varepsilon_3$ if Minnie's WEIGHT gene is not mutated, and $X_3 = \varepsilon_3 + 2$ if her WEIGHT gene is mutated.*
4. *Cheddar's weight, X_4, is generated by averaging Mickey and Minnie's weights, and adding a variable $\varepsilon_4 \sim \mathcal{N}(0, 1)$ which captures the random variation in weight due to genetic and environmental factors.*
5. *Similarly, Gouda's weight, X_5, is generated by taking Cheddar's weight and adding a variable $\varepsilon_5 \sim \mathcal{N}(0, 2)$ which captures the random variation in weight from genetic and environmental factors and from the random selection of Gouda's mom.*

This data-generating process is captured by the following equations and distributional statements:

$$\begin{aligned} X_1 &= \mathbb{1}_{\varepsilon_1 \leq 0.5} & \varepsilon_1 &\sim \mathsf{Unif}([0, 1]) \\ X_2 &= \varepsilon_2 + 2X_1 & \varepsilon_2 &\sim \mathcal{N}(25, 1) \\ X_3 &= \varepsilon_3 + 2X_1 & \varepsilon_2 &\sim \mathcal{N}(20, 1) \\ X_4 &= {}^1\!/_2\,(X_2 + X_3) + \varepsilon_4 & \varepsilon_4 &\sim \mathcal{N}(0, 1) \\ X_5 &= X_4 + \varepsilon_5 & \varepsilon_5 &\sim \mathcal{N}(0, 2) \end{aligned}$$

where the distributions of $\varepsilon_1, \ldots, \varepsilon_5$ are mutually independent. By looking at the pushforward of the distribution over ε_i to a distribution on X_i, we see that the the conditionals $\mathbb{P}_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i))$ match the causal conditionals from our causal Bayesian network in Example 1.3.

Note the introduction of the random variables $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_5$ compared to the causal Bayesian network. These variables are essential to defining counterfactuals with SCMs. These are called exogenous variables, and are added to the signature associated with a structural causal model.

1.4.1 Structural Causal Models

As we did with Bayesian network, we will define a structural causal model as a mathematical model consisting of *four* components: a **signature** which describes the variables in the model, a **causal graph** which defines the (qualitative) relationships between the variables, and a set of **causal mechanisms**, which are deterministic functions specifying the relationships between variable, and an **exogenous distribution** over the random noise variables.

The signature of a causal model is similar to the signature of a Bayesian network, but includes the exogenous noise variables. Formally,

Definition 1.9. *A* **signature** $\mathcal{S}$ *consists of:*

- *A set $\mathcal{X}$ of* **endogenous variables**,
- *A set $\mathcal{E}$ of* **exogenous variables**, *and*
- *A* **range function** $\mathcal{R}$ *mapping each variable $V \in \mathcal{X} \cup \mathcal{E}$ to its alphabet $\mathcal{R}(V)$.*

Here, we will take the causal graph of a structural causal model to be a DAG over the endogenous variables $\mathcal{X}$. One may also define SCMs such that the exogenous variables are included in the causal graph, and require that each exogenous variable to be a sink, i.e., a node with no parents. In this course, we will consider only **Markovian** structural causal models, where each exogenous variable ε_i only has a direct effect on a single endogenous variable X_i. [1] In that case, the causal graph with exogenous variables would just have extra edges $\varepsilon_i \to X_i$ for each $X_i \in \mathcal{X}$, so we will only work with a causal graph over $\mathcal{X}$.

Next, we define the notion of a causal mechanism:

Definition 1.10. *Let $\mathbf{S} = (\mathcal{X}, \mathcal{E}, \mathcal{R})$ be a signature and $\mathcal{G}$ be a causal graph over $\mathcal{X}$. A* **causal mechanism** *for $X_i \in \mathcal{X}$ is a function*

$$f_{X_i} : \mathcal{R}(\mathrm{pa}_{\mathcal{G}}(X_i)) \to \mathcal{R}(X_i).$$

We let $\mathrm{mech}_{\mathcal{G},\mathcal{S}}(X_i)$ denote the set of all such functions.

This leads to the fundamental formal definition:

Definition 1.11. *Let be a signature and $\mathcal{G}$ be a causal graph over $\mathcal{G}$. A* **structural causal model** *(SCM) $\mathcal{M}$ consists of:*

- *A signature $\mathcal{S} = (\mathcal{X}, \mathcal{E}, \mathcal{R})$,*
- *A causal graph $\mathcal{G}$ over $\mathcal{X}$,*
- *An indexed set $\{f_{X_i}\}_{X_i \in \mathcal{X}}$ of causal mechanisms, with $f_{X_i} \in \mathrm{mech}_{\mathcal{G},\mathcal{S}}(X_i)$, and*
- *A product distribution $\mathbb{P}_{\mathcal{E}}$ over the exogenous variables, called the* **exogenous distribution**.

1.4.2 Interventions

Now, we will define the notion of an intervention on a structural causal model, which changes some SCM $\mathcal{M}$ into a related SCM $\mathcal{M}^I$.

Definition 1.12. *Let $\mathcal{M}$ be a structural causal model with signature $\mathcal{S} = (\mathcal{X}, \mathcal{E}, \mathcal{R})$, causal graph $\mathcal{G}$, causal mechanisms $\{f_{X_i}\}_{X_i \in \mathcal{X}}$, and exogenous distribution $\mathbb{P}_{\mathcal{E}}$. An* **intervention** *I (also called a* **mechanism change***) consists of*

- *A set $T(I) \subseteq \mathcal{X}$ of* **intervened variables** *(also called* **intervention targets***), and*
- *An indexed set $\{g_{X_i}\}_{X_i \in \mathcal{X}(I)}$ of* **interventional causal mechanisms**, *with $g_{X_i} \in \mathrm{mech}_{\tilde{\mathcal{G}},\mathcal{S}}(X_i)$.*

Then, the ***interventional structural causal model*** *$\mathcal{M}^I$ is a Bayesian network with the same signature, the same causal graph, and the same exogenous distribution as $\mathcal{M}$, but with causal mechanisms*

$$\{g_{X_i}\}_{X_i \in T(I)} \cup \{f_{X_i}\}_{X_i \in \mathcal{X} \setminus T(I)}$$

.

Remark 1.1. *Note that the causal mechanisms for non-intervened variables are* **invariant**, *e.g., they do not change between $\mathcal{M}$ and $\mathcal{M}^I$. This is one of the defining properties of causal models: invariance of causal mechanisms unless they are explicitly postulated to change. An important consequence of invariance is* **consistency**: *if a variable X_i is not intervened, then we have*

$$\mathbb{P}^I_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i)) = \mathbb{P}_{\mathcal{X}}(X_i \mid \mathrm{pa}_{\mathcal{G}}(X_i))$$

[1] If the exogenous variables are allowed to have direct effects on multiple endogenous variables, then we obtain **semi-Markovian** structural causal models.

1.4.3 Counterfactuals

Finally, we define counterfactuals in terms of structural causal models.

Definition 1.13. *Let $\mathcal{M}$ be a structural causal model with signature $\mathcal{S} = (\mathcal{X}, \mathcal{E}, \mathcal{R})$, causal graph $\mathcal{G}$, causal mechanisms $\{f_{X_i}\}_{X_i \in \mathcal{X}}$, and exogenous distribution $\mathbb{P}_{\mathcal{E}}$. Let $\mathbf{S} \subseteq \mathcal{X}$ be a subset of endogenous variables and $\mathbf{s} \in \mathcal{R}(\mathbf{S})$ be a realization of $\mathcal{S}$.*

The **counterfactual SCM**, *denoted $\mathcal{M}_{\mathbf{S}=\mathbf{s}}$, is a SCM with the same signature, the same causal graph, and the same mechanisms as $\mathcal{M}$, but with exogenous distribution $\mathbb{P}_{\mathcal{E}|\mathbf{S}=\mathbf{s}}$.*

Then, we may transform $\mathcal{M}_{\mathbf{S}=\mathbf{s}}$ by an intervention, obtaining $\mathcal{M}^I_{\mathbf{S}=\mathbf{s}}$, to ask about counterfactual outcomes if some intervention had been performed. We close with an example.

Example 1.5. *Let $\mathcal{M}$ be the structural causal model from Example 1.4. Let $\mathbf{S} = \mathcal{X}$ with $x_1 = 0$, $x_2 = 26$, $x_3 = 20$, $x_4 = 23$, and $x_5 = 23$. Then*

$$\mathbb{P}_{\mathcal{E}|\mathbf{S}=\mathbf{s}}(\mathcal{E}) = \mathsf{Unif}(\varepsilon_1; (0.5, 1]) \times \delta_{26}(\varepsilon_2) \times \delta_{20}(\varepsilon_3) \times \delta_0(\varepsilon_4) \times \delta_0(\varepsilon_5)$$

We can use $M_{\mathbf{S}=\mathbf{s}}$ to ask about counterfactual statements, e.g. "what if we had set $X_1 = 1$?". This questions corresponds to asking about the distribution $\mathbb{P}^I_{\mathbf{S}=\mathbf{s}}$ for $\mathcal{X}(I) = \{X_1\}$ and $g_{X_1} = 1$, and we find that

$$\mathbb{P}^I_{\mathbf{S}=\mathbf{s}}(X_1, \ldots, X_5) = \delta_1(X_1) \times \delta_{28}(X_2) \times \delta_{22}(X_3) \times \delta_{25}(X_4) \times \delta_{25}(X_5)$$

1.5 Additional Reading

- **Cyclic models**: In this course, we will only consider causal models without cycles in their augmented graph. This definition can be extended to allow for cycles, see for example Bongers et al. (2021).
- **Counterfactuals**: We will not use counterfactuals again in this course, except perhaps when discussing how causality may be brought into machine learning areas such as explainability research.

Bibliography

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915.