

Course Lecture Notes

# **Introduction to Causal Inference**

**from a Machine Learning Perspective**

Brady Neal

September 22, 2020

# Preface

**Prerequisites** There is one main prerequisite: **basic probability**. This course assumes you’ve taken an introduction to probability course or have had equivalent experience. Topics from statistics and machine learning will pop up in the course from time to time, so some familiarity with those will be helpful but is not necessary. For example, if cross-validation is a new concept to you, you can learn it relatively quickly at the point in the book that it pops up. And we give a primer on some statistics terminology that we’ll use in Section 2.4.

**Active Reading Exercises** Research shows that one of the best techniques to remember material is to actively try to recall information that you recently learned. You will see “active reading exercises” throughout the book to help you do this. They’ll be marked by the [Active reading exercise](#): heading.

**Many Figures in This Book** As you will see, there are a ridiculous amount of figures in this book. This is on purpose. This is to help give you as much visual intuition as possible. We will sometimes copy the same figures, equations, etc. that you might have seen in preceding chapters so that we can make sure the figures are always right next to the text that references them.

**Sending Me Feedback** This is a book *draft*, so I greatly appreciate any feedback you’re willing to send my way. If you’re unsure whether I’ll be receptive to it or not, don’t be. Please send any feedback to me at [bradyneal11@gmail.com](mailto:bradyneal11@gmail.com) with “[Causal Book]” in the beginning of your email subject. Feedback can be at the word level, sentence level, section level, chapter level, etc. Here’s a non-exhaustive list of useful kinds of feedback:

- ▶ Typoz.
- ▶ Some part is confusing.
- ▶ You notice your mind starts to wander, or you don’t feel motivated to read some part.
- ▶ Some part seems like it can be cut.
- ▶ You feel strongly that some part absolutely should not be cut.
- ▶ Some parts are not connected well. Moving from one part to the next, you notice that there isn’t a natural flow.
- ▶ A new active reading exercise you thought of.

**Bibliographic Notes** Although we do our best to cite relevant results, we don’t want to disrupt the flow of the material by digging into exactly where each concept came from. There will be complete sections of bibliographic notes in the final version of this book, but they won’t come until after the course has finished.

# Contents

<b>Preface</b>	<b>ii</b>
<b>Contents</b>	<b>iii</b>
<b>1 Motivation: Why You Might Care</b>	<b>1</b>
1.1 Simpson's Paradox . . . . .	1
1.2 Applications of Causal Inference . . . . .	2
1.3 Correlation Does Not Imply Causation . . . . .	3
1.3.1 Nicolas Cage and Pool Drownings . . . . .	3
1.3.2 Why is Association Not Causation? . . . . .	4
1.4 Main Themes . . . . .	5
<b>2 Potential Outcomes</b>	<b>6</b>
2.1 Potential Outcomes and Individual Treatment Effects . . . . .	6
2.2 The Fundamental Problem of Causal Inference . . . . .	7
2.3 Getting Around the Fundamental Problem . . . . .	8
2.3.1 Average Treatment Effects and Missing Data Interpretation . . . . .	8
2.3.2 Ignorability and Exchangeability . . . . .	9
2.3.3 Conditional Exchangeability and Unconfoundedness . . . . .	10
2.3.4 Positivity/Overlap and Extrapolation . . . . .	12
2.3.5 No interference, Consistency, and SUTVA . . . . .	13
2.3.6 Tying It All Together . . . . .	14
2.4 Fancy Statistics Terminology Defancified . . . . .	15
2.5 A Complete Example with Estimation . . . . .	16
<b>3 The Flow of Association and Causation in Graphs</b>	<b>19</b>
3.1 Graph Terminology . . . . .	19
3.2 Bayesian Networks . . . . .	20
3.3 Causal Graphs . . . . .	22
3.4 Two-Node Graphs and Graphical Building Blocks . . . . .	23
3.5 Chains and Forks . . . . .	24
3.6 Colliders and their Descendants . . . . .	26
3.7 d-separation . . . . .	28
3.8 Flow of Association and Causation . . . . .	30
<b>4 Causal Models</b>	<b>32</b>
4.1 The <i>do</i> -operator and Interventional Distributions . . . . .	32
4.2 The Main Assumption: Modularity . . . . .	34
4.3 Truncated Factorization . . . . .	35
4.3.1 Example Application and Revisiting "Association is Not Causation" . . . . .	36
4.4 The Backdoor Adjustment . . . . .	37
4.4.1 Relation to Potential Outcomes . . . . .	39
4.5 Structural Causal Models (SCMs) . . . . .	40
4.5.1 Structural Equations . . . . .	40
4.5.2 Interventions . . . . .	42
4.5.3 Collider Bias and Why to Not Condition on Descendants of Treatment . . . . .	43
4.6 Example Applications of the Backdoor Adjustment . . . . .	44
4.6.1 Association vs. Causation in a Toy Example . . . . .	44

4.6.2	A Complete Example with Estimation . . . . .	45
4.7	Assumptions Revisited . . . . .	47
<b>5</b>	<b>Randomized Experiments</b>	<b>49</b>
5.1	Comparability and Covariate Balance . . . . .	49
5.2	Exchangeability . . . . .	50
5.3	No Backdoor Paths . . . . .	51
<b>6</b>	<b>Nonparametric Identification</b>	<b>52</b>
6.1	Frontdoor Adjustment . . . . .	52
6.2	<i>do</i> -calculus . . . . .	55
6.2.1	Application: Frontdoor Adjustment . . . . .	57
6.3	Determining Identifiability from the Graph . . . . .	58
<b>7</b>	<b>Estimation</b>	<b>61</b>
7.1	Preliminaries . . . . .	61
7.2	Conditional Outcome Modeling (COM) . . . . .	62
7.3	Grouped Conditional Outcome Modeling (GCOM) . . . . .	63
7.4	Increasing Data Efficiency . . . . .	64
7.4.1	TARNet . . . . .	64
7.4.2	X-Learner . . . . .	65
7.5	Propensity Scores . . . . .	66
7.6	Inverse Probability Weighting (IPW) . . . . .	67
7.7	Doubly Robust Methods . . . . .	69
7.8	Other Methods . . . . .	69
7.9	Concluding Remarks . . . . .	70
7.9.1	Confidence Intervals . . . . .	70
7.9.2	Comparison to Randomized Experiments . . . . .	70
<b>8</b>	<b>Counterfactuals</b>	<b>72</b>
8.1	Coming Soon . . . . .	72
<b>9</b>	<b>More Chapters Coming</b>	<b>73</b>
	<b>APPENDIX</b>	<b>74</b>
<b>A</b>	<b>Proofs</b>	<b>75</b>
A.1	Proof of Equation 6.1 from Section 6.1 . . . . .	75
A.2	Proof of Propensity Score Theorem (7.1) . . . . .	75
A.3	Proof of IPW Estimand (7.18) . . . . .	76
	<b>Bibliography</b>	<b>78</b>
	<b>Alphabetical Index</b>	<b>81</b>

# List of Figures

1.1	Causal structure for when to prefer treatment B for COVID-27	2
1.2	Causal structure for when to prefer treatment A for COVID-27	2
1.3	Number of Nicolas Cage movies correlates with number of pool drownings	3
1.4	Causal structure with getting lit as a confounder	4
2.2	Causal structure for ignorable treatment assignment mechanism	9
2.1	Causal structure of $X$ confounding the effect of $T$ on $Y$	9
2.3	Causal structure of confounding through $X$	11
2.4	Causal structure for conditional exchangeability given $X$	11
2.5	The Identification-Estimation Flowchart	16
3.3	Directed graph	19
3.1	Terminology machine gun	19
3.2	Undirected graph	19
3.4	Directed graph with cycle	19
3.5	Directed graph with immorality	20
3.6	Four node DAG where $X_4$ locally depends on only $X_3$	20
3.7	Four node DAG with many independencies	21
3.8	Two connected node DAG	22
3.9	Basic graph building blocks	24
3.11	Two connected node DAG	24
3.12	Chain with association	24
3.10	Two unconnected node DAG	24
3.13	Fork with association	25
3.14	Chain with blocked association	25
3.15	Fork with blocked association	25
3.16	Immorality with association blocked by collider	26
3.17	Immorality with association unblocked	26
3.18	Good-looking men are jerks example	27
3.19	Graphs for d-separation exercise	30
3.20	Causal association and confounding association	30
3.21	Assumptions flowchart from statistical independencies to causal dependencies	31
4.1	The Identification-Estimation Flowchart (extended)	32
4.2	Illustration of the difference between conditioning and intervening	33
4.3	Causal mechanism	34
4.4	Intervention as edge deletion in causal graphs	35
4.5	Causal structure for application of truncated factorization	36
4.6	Manipulated graph for three nodes	37
4.7	Graph for structural equation	40
4.8	Causal graph for several structural equations	41
4.9	Causal structure before simple intervention	42
4.10	Causal structure after simple intervention	42
4.11	Causal graph for completely blocking causal flow	43
4.12	Causal graph for partially blocking causal flow	43
4.13	Causal graph where a conditioned collider induces bias	43
4.14	Causal graph where child of a mediator is conditioned on	44

4.15 Magnified causal graph where child of a mediator is conditioned on . . . . .	44
4.16 Causal graph for M-bias . . . . .	44
4.17 Causal graph for toy example . . . . .	45
4.18 Causal graph for blood pressure example with collider . . . . .	46
4.19 Causal graph for M-bias with unobserved variables . . . . .	47
5.1 Causal structure of confounding through $X$ . . . . .	51
5.2 Causal structure when we randomize treatment . . . . .	51
6.1 Causal graph for frontdoor criterion . . . . .	52
6.2 Illustration of focusing analysis to a mediator . . . . .	52
6.3 Illustration of steps of frontdoor adjustment . . . . .	52
6.5 Equationtown . . . . .	53
6.4 Causal graph for frontdoor criterion . . . . .	53
6.6 Causal graph for frontdoor criterion . . . . .	54
6.7 Causal graph for frontdoor criterion . . . . .	57
6.10 Causal graph for frontdoor criterion . . . . .	58
6.8 Causal graph for frontdoor with $W - T$ edge removed . . . . .	58
6.9 Causal graph for frontdoor with $T - M$ edge removed . . . . .	58
6.11 Example graph that satisfies the unconfounded children criterion. . . . .	60
7.1 The Identification-Estimation Flowchart . . . . .	62
7.2 Different neural networks for different kinds of estimators . . . . .	65
7.3 Simple graph where $W$ satisfies the backdoor criterion . . . . .	66
7.5 Simple graph where $W$ confounds the effect of $T$ on $Y$ . . . . .	67
7.6 Effective graph for pseudo-population that we get by reweighting the data generated according to the graph in Figure 7.5 using inverse probability weighting. . . . .	67
7.4 Graphical proof of propensity score theorem . . . . .	67
A.1 Causal graph for frontdoor criterion . . . . .	75

## List of Tables

1.1 Simpson's paradox in COVID-27 data . . . . .	1
2.1 Causal Inference as Missing Data Problem . . . . .	9
3.1 Exponential number of parameters for modeling factors . . . . .	20

## Listings

2.1 Python code for estimating the ATE . . . . .	17
2.2 Python code for estimating the ATE using the coefficient of linear regression . . . . .	17
4.1 Python code for estimating the ATE, without adjusting for the collider . . . . .	46

# Motivation: Why You Might Care

# 1

## 1.1 Simpson's Paradox

Consider a purely hypothetical future where there is a new disease known as COVID-27 that is prevalent in the human population. In this purely hypothetical future, there are two treatments that have been developed: treatment A and treatment B. Treatment B is more scarce than treatment A, so the split of those currently receiving treatment A vs. treatment B is roughly 73%/27%. You are in charge of choosing which treatment your country will exclusively use, in a country that only cares about minimizing loss of life.

You have data on the percentage of people who die from COVID-27, given the treatment they were assigned and given their condition at the time treatment was decided. Their condition is a binary variable: either mild or severe. In this data, 16% of those who receive A die, whereas 19% of those who receive B die. However, when we examine the people with mild condition separately from the people with severe condition, the numbers reverse order. In the mild subpopulation, 15% of those who receive A die, whereas 10% of those who receive B die. In the severe subpopulation, 30% of those who receive A die, whereas 20% of those who receive B die. We depict these percentages and the corresponding counts in Table 1.1.

		Condition		
		Mild	Severe	Total
Treatment	A	15% (210/1400)	30% (30/100)	<b>16%</b> (240/1500)
	B	<b>10%</b> (5/50)	<b>20%</b> (100/500)	19% (105/550)

The apparent paradox stems from the fact that, in Table 1.1, the “Total” column could be interpreted to mean that we should prefer treatment A, whereas the “Mild” and “Severe” columns could both be interpreted to mean that we should prefer treatment B.<sup>1</sup> In fact, the answer is that if we know someone’s condition, we should give them treatment B, and if we do *not* know their condition, we should give them treatment A. Just kidding... that doesn’t make any sense. So really, what treatment should you choose for your country?

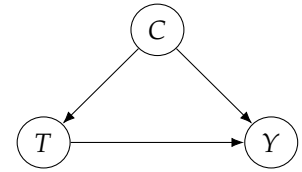
Either treatment A or treatment B could be the right answer, depending on the causal structure of the data. In other words, causality is essential to solve Simpson’s paradox. For now, we will just give the intuition for when you should prefer treatment A vs. when you should prefer treatment B, but it will be made more formal in Chapter 4.

1.1 Simpson’s Paradox . . . . .	1
1.2 Applications of Causal Inference . . . . .	2
1.3 Correlation Does Not Imply Causation . . . . .	3
Nicolas Cage and Pool Drownings . . . . .	3
Why is Association Not Causation? . . . . .	4
1.4 Main Themes . . . . .	5

**Table 1.1:** Simpson’s paradox in COVID-27 data. The percentages denote the mortality rates in each of the groups. Lower is better. The numbers in parentheses are the corresponding counts. This apparent paradox stems from the interpretation that treatment A looks better when examining the whole population, but treatment B looks better in all subpopulations.

<sup>1</sup> A key ingredient necessary to find Simpson’s paradox is the **non-uniformity of allocation of people to the groups**. 1400 of the 1500 people who received treatment A had mild condition, whereas 500 of the 550 people who received treatment B had severe condition. Because people with mild condition are less likely to die, this means that the total mortality rate for those with treatment A is lower than what it would have been if mild and severe conditions were equally split among them. The opposite bias is true for treatment B.

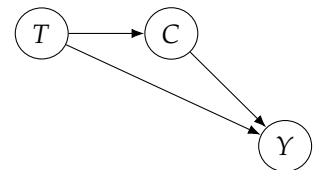
**Scenario 1** If the condition  $C$  is a cause of the treatment  $T$  (Figure 1.1), treatment B is more effective at reducing mortality  $Y$ . An example scenario is where doctors decide to give treatment A to most people who have mild conditions. And they save the more expensive and more limited treatment B for people with severe conditions. Because having severe condition causes one to be more likely to die ( $C \rightarrow Y$  in Figure 1.1) and causes one to be more likely to receive treatment B ( $C \rightarrow T$  in Figure 1.1), treatment B will be associated with higher mortality in the total population. In other words, treatment B is associated with a higher mortality rate simply because condition is a common cause of both treatment and mortality. Here, condition confounds the effect of treatment on mortality. To correct for this confounding, we must examine the relationship of  $T$  and  $Y$  among patients with the same conditions. This means that the better treatment is the one that yields lower mortality in each of the subpopulations (the “Mild” and “Severe” columns in Table 1.1): treatment B.



**Figure 1.1:** Causal structure of scenario 1, where condition  $C$  is a common cause of treatment  $T$  and mortality  $Y$ . Given this causal structure, treatment B is preferable.

**Scenario 2** If the prescription<sup>2</sup> of treatment  $T$  is a cause of the condition  $C$  (Figure 1.2), treatment A is more effective. An example scenario is where treatment B is so scarce that it requires patients to wait a long time after they were prescribed the treatment before they can receive the treatment. Treatment A does not have this problem. Because the condition of a patient with COVID-27 worsens over time, the prescription of treatment B actually causes patients with mild conditions to develop severe conditions, causing a higher mortality rate. Therefore, even if treatment B is more effective than treatment A once *administered* (positive effect along  $T \rightarrow Y$  in Figure 1.2), because *prescription* of treatment B causes worse conditions (negative effect along  $T \rightarrow C \rightarrow Y$  in Figure 1.2), treatment B is less effective in total. Note: Because treatment B is more expensive, treatment B is prescribed with 0.27 probability, while treatment A is prescribed with 0.73 probability; importantly, treatment prescription is independent of condition in this scenario.

<sup>2</sup>  $T$  refers to the prescription of the treatment, rather than the subsequent reception of the treatment.



**Figure 1.2:** Causal structure of scenario 2, where treatment  $T$  is a cause of condition  $C$ . Given this causal structure, treatment A is preferable.

**In sum**, the more effective treatment is completely dependent on the causal structure of the problem. In Scenario 1, where  $C$  was a cause of  $T$  (Figure 1.1), treatment B was more effective. In Scenario 2, where  $T$  was a cause of  $C$  (Figure 1.2), treatment A was more effective. Without causality, Simpson’s paradox cannot be resolved. With causality, it is not a paradox at all.

## 1.2 Applications of Causal Inference

Causal inference is essential to science, as we often want to make causal claims, rather than merely associational claims. For example, if we are choosing between treatments for a disease, we want to choose the treatment that causes the most people to be cured, without causing too many bad side effects. If we want a reinforcement learning algorithm to maximize reward, we want it to take actions that cause it to achieve the maximum reward. If we are studying the effect of social media on mental health, we are trying to understand what the main causes of a given mental health outcome are and order these causes by the percentage of the outcome that can be attributed to each cause.



Causal inference is essential for rigorous decision-making. For example, say we are considering several different policies to implement to reduce greenhouse gas emissions, and we must choose just one due to budget constraints. If we want to be maximally effective, we should carry out causal analysis to determine which policy will cause the largest reduction in emissions. As another example, say we are considering several interventions to reduce global poverty. We want to know which policies will cause the largest reductions in poverty.

Now that we've gone through the general example of Simpson's paradox and a few specific examples in science and decision-making, we'll move to how causal inference is so different from prediction.

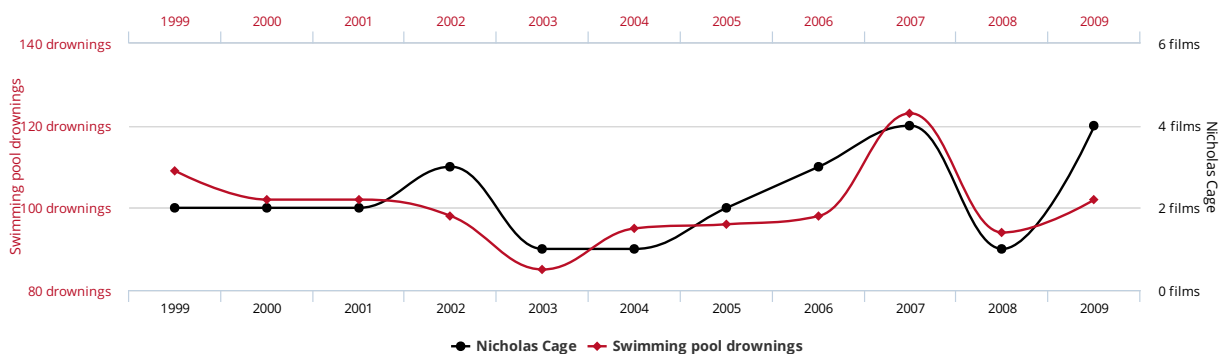
## 1.3 Correlation Does Not Imply Causation

Many of you will have heard the mantra "correlation does not imply causation." In this section, we will quickly review that and provide you with a bit more intuition about why this is the case.

### 1.3.1 Nicolas Cage and Pool Drownings

It turns out that the yearly number of people who drown by falling into swimming pools has a high degree of correlation with the yearly number of films that Nicolas Cage appears in [1]. See Figure 1.3 for a graph of this data. Does this mean that Nicolas Cage encourages bad swimmers to hop in the pool in his films? Or does Nicolas Cage feel more motivated to act in more films when he sees how many drownings are happening that year, perhaps to try to prevent more drownings? Or is there some other explanation? For example, maybe Nicolas Cage is interested in increasing his popularity among causal inference practitioners, so he travels back in time to convince his past self to do just the right number of movies for us to see this correlation, but not too close of a match as that would arouse suspicion and potentially cause someone to prevent him from rigging the data this way. We may never know for sure.

[1]: Vigen (2015), *Spurious correlations*



**Figure 1.3:** The yearly number of movies Nicolas Cage appears in correlates with the yearly number of pool drownings [1].

Of course, all of the possible explanations in the preceding paragraph seem quite unlikely. Rather, it is likely that this is a *spurious correlation*, where there is no causal relationship. We'll soon move on to a more

illustrative example that will help clarify how spurious correlations can arise.

### 1.3.2 Why is Association Not Causation?

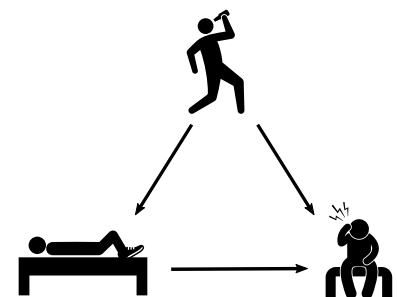
Before moving to the next example, let's be a bit more precise about terminology. "Correlation" is often colloquially used as a synonym for statistical dependence. However, "correlation" is technically only a measure of *linear* statistical dependence. We will largely be using the term *association* to refer to statistical dependence from now on.

Causation is not all or none. For any given amount of association, it does not need to be "all of the association is causal" or "none of the association is causal." Rather, it is possible to have a large amount of association with only *some* of it being causal. The phrase "association is not causation" simply means that the amount of association and the amount of causation can be different. Some amount of association and zero causation is a special case of "association is not causation."

Say you happen upon some data that relates wearing shoes to bed and waking up with a headache, as one does. It turns out that most times that someone wears shoes to bed, that person wakes up with a headache. And most times someone doesn't wear shoes to bed, that person doesn't wake up with a headache. It is not uncommon for people to interpret data like this (with associations) as meaning that wearing shoes to bed causes people to wake up with headaches, especially if they are looking for a reason to justify not wearing shoes to bed. A careful journalist might make claims like "wearing shoes to bed is associated with headaches" or "people who wear shoes to bed are at higher risk of waking up with headaches." However, the main reason to make claims like that is that most people will internalize claims like that as "if I wear shoes to bed, I'll probably wake up with a headache."

We can explain how wearing shoes to bed and headaches are associated without either being a cause of the other. It turns out that they are both caused by a *common cause*: drinking the night before. We depict this in Figure 1.4. You might also hear this kind of variable referred to as a "confounder" or a "lurking variable." We will call this kind of association *confounding association* since the association is facilitated by a confounder.

The total association observed can be made up of both confounding association and causal association. It could be the case that wearing shoes to bed does have some small causal effect on waking up with a headache. Then, the total association would not be solely confounding association nor solely causal association. It would be a mixture of both. For example, in Figure 1.4, causal association flows along the arrow from shoe-sleeping to waking up with a headache. And confounding association flows along the path from shoe-sleeping to drinking to headachening (waking up with a headache). We will make the graphical interpretation of these different kinds of association clear in Chapter 3.



**Figure 1.4:** Causal structure, where drinking the night before is a common cause of sleeping with shoes on and of waking up with a headache.

**The Main Problem** The main problem motivating causal inference is that association is not causation.<sup>3</sup> If the two were the same, then causal inference would be easy. Traditional statistics and machine learning would already have causal inference solved, as measuring causation would be as simple as just looking at measures such as correlation and predictive performance in data. A large portion of this book will be about better understanding and solving this problem.

<sup>3</sup> As we'll see in Chapter 5, if we randomly assign the treatment in a controlled experiment, association actually *is* causation.

## 1.4 Main Themes

There are several overarching themes that will keep coming up throughout this book. These themes will largely be comparisons of two different categories. As you are reading, it is important that you understand which categories different sections of the book fit into and which categories they do not fit into.

**Statistical vs. Causal** Even with an infinite amount of data, we sometimes cannot compute some causal quantities. In contrast, much of statistics is about addressing uncertainty in finite samples. When given infinite data, there is no uncertainty. However, association, a statistical concept, is not causation. There is more work to be done in causal inference, even after starting with infinite data. This is the main distinction motivating causal inference. We have already made this distinction in this chapter and will continue to make this distinction throughout the book.

**Identification vs. Estimation** Identification of causal effects is unique to causal inference. It is the problem that remains to solve, even when we have infinite data. However, causal inference also shares estimation with traditional statistics and machine learning. We will largely begin with identification of causal effects (in Chapters 2, 4 and 6) before moving to estimation of causal effects (in Chapter 7). The exceptions are Section 2.5 and Section 4.6.2, where we carry out complete examples with estimation to give you an idea of what the whole process looks like early on.

**Interventional vs. Observational** If we can intervene/experiment, identification of causal effects is relatively easy. This is simply because we can actually take the action that we want to measure the causal effect of and simply measure the effect after we take that action. Observational data is where it gets more complicated because confounding is almost always introduced into the data.

**Assumptions** There will be a large focus on what assumptions we are using to get the results that we get. Each assumption will have its own box to help make it difficult to not notice. Clear assumptions should make it easy to see where critiques of a given causal analysis or causal model will be. The hope is that presenting assumptions clearly will lead to more lucid discussions about causality.

In this chapter, we will ease into the world of causality. We will see that new concepts and corresponding notations need to be introduced to clearly describe causal concepts. These concepts are “new” in the sense that they may not exist in traditional statistics or math, but they should be familiar in that we use them in our thinking and describe them with natural language all the time.

**Familiar statistical notation** We will use  $T$  to denote the random variable for treatment,  $Y$  to denote the random variable for the outcome of interest and  $X$  to denote covariates. In general, we will use uppercase letters to denote random variables (except in maybe one case) and lowercase letters to denote values that random variables take on. Much of what we consider will be settings where  $T$  is binary. Know that, in general, we can extend things to work in settings where  $T$  can take on more than two values or where  $T$  is continuous.

## 2.1 Potential Outcomes and Individual Treatment Effects

We will now introduce the first causal concept to appear in this book. These concepts are sometimes characterized as being unique to the Neyman-Rubin [2–4] causal model (or potential outcomes framework), but they are not. For example, these same concepts are still present (just under different notation) in the framework that uses causal graphs (Chapters 3 and 4). It is important that you spend some time ensuring that you understand these initial causal concepts. If you have not studied causal inference before, they will be unfamiliar to see in mathematical contexts, though they may be quite familiar intuitively because we commonly think and communicate in causal language.

**Scenario 1** Consider the scenario where you are unhappy. And you are considering whether or not to get a dog to help make you happy. If you become happy after you get the dog, does this mean the dog caused you to be happy? Well, what if you would have also become happy had you *not* gotten the dog? In that case, the dog was not necessary to make you happy, so its claim to a causal effect on your happiness is weak.

**Scenario 2** Let’s switch things up a bit. Consider that you will still be happy if you get a dog, but now, if you don’t get a dog, you will remain unhappy. In this scenario, the dog has a pretty strong claim to a causal effect on your happiness.

In both the above scenarios, we have used the causal concept known as potential outcomes. Your outcome  $Y$  is happiness:  $Y = 1$  corresponds to happy while  $Y = 0$  corresponds to unhappy. Your treatment  $T$  is whether or not you get a dog:  $T = 1$  corresponds to you getting a dog while  $T = 0$

2.1 Potential Outcomes and Individual Treatment Effects . . .	6
2.2 The Fundamental Problem of Causal Inference . . . .	7
2.3 Getting Around the Fundamental Problem . . . . .	8
Average Treatment Effects and Missing Data Interpretation . . . . .	8
Ignorability and Exchangeability . . . . .	9
Conditional Exchangeability and Unconfoundedness . . .	10
Positivity/Overlap and Extrapolation . . . . .	12
No interference, Consistency, and SUTVA . . . . .	13
Tying It All Together . . . .	14
2.4 Fancy Statistics Terminology Defancified . . . . .	15
2.5 A Complete Example with Estimation . . . . .	16

[2]: Splawa-Neyman (1923 [1990]), ‘On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.’

[3]: Rubin (1974), ‘Estimating causal effects of treatments in randomized and nonrandomized studies.’

[4]: Sekhon (2008), ‘The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods’

corresponds to you not getting a dog. We denote by  $Y(1)$  the *potential* outcome of happiness you would observe if you were to get a dog ( $T = 1$ ). Similarly, we denote by  $Y(0)$  the potential outcome of happiness you would observe if you were to not get a dog ( $T = 0$ ). In scenario 1,  $Y(1) = 1$  and  $Y(0) = 1$ . In contrast, in scenario 2,  $Y(1) = 1$  and  $Y(0) = 0$ .

More generally, the *potential outcome*  $Y(t)$  denotes what your outcome would be, if you were to take treatment  $t$ . A potential outcome  $Y(t)$  is distinct from the observed outcome  $Y$  in that not all potential outcomes are observed. Rather all potential outcomes can *potentially* be observed. The one that is actually observed depends on the value that the treatment  $T$  takes on.

In the previous scenarios, there was only a single individual in the whole population: you. However, generally, there are many individuals<sup>1</sup> in the population of interest. We will denote the treatment, covariates, and outcome of the  $i_{\text{th}}$  individual using  $T_i$ ,  $X_i$ , and  $Y_i$ . Then, we can define the *individual treatment effect* (ITE)<sup>2</sup> for individual  $i$ :

$$\tau_i \triangleq Y_i(1) - Y_i(0) \quad (2.1)$$

Whenever there is more than one individual in a population,  $Y(t)$  is a random variable because different individuals will have different potential outcomes. In contrast,  $Y_i(t)$  is usually treated as non-random<sup>3</sup> because the subscript  $i$  means that we are conditioning on so much individualized (and context-specific) information, that we restrict our focus to a single individual (in a specific context) whose potential outcomes are deterministic.

ITEs are some of the main quantities that we care about in causal inference. For example, in scenario 2 above, you would choose to get a dog because the causal effect of getting a dog on your happiness is positive:  $Y(1) - Y(0) = 1 - 0 = 1$ . In contrast, in scenario 1, you might choose to not get a dog because there is no causal effect of getting a dog on your happiness:  $Y(1) - Y(0) = 1 - 1 = 0$ .

Now that we've introduced potential outcomes and ITEs, we can introduce the main problems that pop up in causal inference that are not present in fields where the main focus is on association or prediction.

## 2.2 The Fundamental Problem of Causal Inference

It is impossible to observe all potential outcomes for a given individual [3]. Consider the dog example. You could observe  $Y(1)$  by getting a dog and observing your happiness after getting a dog. Alternatively, you could observe  $Y(0)$  by not getting a dog and observing your happiness. However, you cannot observe both  $Y(1)$  and  $Y(0)$ , unless you have a time machine that would allow you to go back in time and choose the version of treatment that you didn't take the first time. You cannot simply get a dog, observe  $Y(1)$ , give the dog away, and then observe  $Y(0)$  because the second observation will be influenced by all the actions you took between the two observations and anything else that changed since the first observation.

<sup>1</sup> "Unit" is often used in the place of "individual" as the units of the population are not always people.

<sup>2</sup> The ITE is also known as the *individual causal effect*, *unit-level causal effect*, or *unit-level treatment effect*.

<sup>3</sup> Though,  $Y_i(t)$  can be treated as random.

[3]: Rubin (1974), 'Estimating causal effects of treatments in randomized and nonrandomized studies.'

This is known as the *fundamental problem of causal inference* [5]. It is fundamental because if we cannot observe both  $Y_i(1)$  and  $Y_i(0)$ , then we cannot observe the causal effect  $Y_i(1) - Y_i(0)$ . This problem is unique to causal inference because, in causal inference, we care about making causal claims, which are defined in terms of potential outcomes. For contrast, consider machine learning. In machine learning, we often only care about predicting the observed outcome  $Y$ , so there is no need for potential outcomes, which means machine learning does not have to deal with this fundamental problem that we must deal with in causal inference.

[5]: Holland (1986), ‘Statistics and Causal Inference’

The potential outcomes that you do not (and cannot) observe are known as *counterfactuals* because they are counter to fact (reality). “Potential outcomes” are sometimes referred to as “counterfactual outcomes,” but we will never do that in this book because a potential outcome  $Y(t)$  does not become counter to fact until another potential outcome  $Y(t')$  is observed. The potential outcome that is observed is sometimes referred to as a *factual*. Note that there are no counterfactuals or factuals until the outcome is observed. Before that, there are only *potential* outcomes.

## 2.3 Getting Around the Fundamental Problem

I suspect this section is where this chapter might start to get a bit unclear. If that is the case for you, don’t worry too much, and just continue to the next chapter, as it will build up parallel concepts in a hopefully more intuitive way.

### 2.3.1 Average Treatment Effects and Missing Data Interpretation

We know that we can’t access individual treatment effects, but what about *average* treatment effects? We get the *average treatment effect* (ATE)<sup>4</sup> by taking an average over the ITEs:

<sup>4</sup> The ATE is also known as the “average causal effect (ACE).”

$$\tau \triangleq \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[Y(1) - Y(0)], \quad (2.2)$$

where the average is over the individuals  $i$  if  $Y_i(t)$  is deterministic. If  $Y_i(t)$  is random, the average is also over any other randomness.

Okay, but how would we actually compute the ATE? Let’s look at some made-up data in Table 2.1 for this. If you like examples, feel free to substitute in the COVID-27 example from Section 1.1 or the dog-happiness example from Section 2.1. We will take this table as the whole population of interest. Because of the fundamental problem of causal inference, this is fundamentally a missing data problem. All of the question marks in the table indicate that we do not observe that cell.

A natural quantity that comes to mind is the *associational difference*:  $\mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$ . By linearity of expectation, we have that the ATE  $\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ . Then, maybe  $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$  equals  $\mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$ . Unfortunately, this is not true in general. If it were, that would mean that causation is simply association.  $\mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$  is an associational quantity, whereas  $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$



$i$	$T$	$Y$	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$
1	0	0	?	0	?
2	1	1	1	?	?
3	1	0	0	?	?
4	0	0	?	0	?
5	0	1	?	1	?
6	1	1	1	?	?

is a causal quantity. They are not equal due to confounding, which we discussed in Section 1.3. The graphical interpretation of this, depicted in Figure 2.1, is that  $X$  confounds the effect of  $T$  on  $Y$  because there is this  $T \leftarrow X \rightarrow Y$  path that non-causal association flows along.<sup>5</sup>

### 2.3.2 Ignorability and Exchangeability

Well, what assumption(s) would make it so that the ATE is simply the associational difference? This is equivalent to saying “what makes it valid to calculate the ATE by taking the average of the  $Y(0)$  column, ignoring the question marks, and subtracting that from the average of the  $Y(1)$  column, ignoring the question marks?”<sup>6</sup> This ignoring of the question marks (missing data) is known as *ignorability*. Assuming ignorability is like ignoring how people ended up selecting the treatment they selected and just assuming they were randomly assigned their treatment; we depict this graphically in Figure 2.2 by the lack of a causal arrow from  $X$  to  $T$ . We will now state this assumption formally.

#### Assumption 2.1 (Ignorability / Exchangeability)

$$(Y(1), Y(0)) \perp\!\!\!\perp T$$

This assumption is key to causal inference because it allows us to reduce the ATE to the associational difference:

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1) | T = 1] - \mathbb{E}[Y(0) | T = 0] \quad (2.3)$$

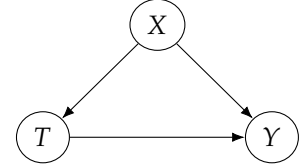
$$= \mathbb{E}[Y | T = 1] - \mathbb{E}[Y | T = 0] \quad (2.4)$$

The ignorability assumption is used in Equation 2.3. We will talk more about Equation 2.4 when we get to Section 2.3.5.

Another perspective on this assumption is that of *exchangeability*. Exchangeability means that the treatment groups are exchangeable in the sense that if they were swapped, the new treatment group would observe the same outcomes as the old treatment group, and the new control group would observe the same outcomes as the old control group. Formally, this assumption means  $\mathbb{E}[Y(1)|T = 0] = \mathbb{E}[Y(1)|T = 1]$  and  $\mathbb{E}[Y(0)|T = 1] = \mathbb{E}[Y(0)|T = 0]$ , respectively. Then, this implies  $\mathbb{E}[Y(1)|T = t] = \mathbb{E}[Y(1)]$  and  $\mathbb{E}[Y(0)|T = t] = \mathbb{E}[Y(0)]$ , for all  $t$ , which is nearly equivalent<sup>7</sup> to Assumption 2.1.

An important intuition to have about exchangeability is that it guarantees that the treatment groups are comparable. In other words, the treatment groups are the same in all relevant aspects other than the treatment. This intuition is what underlies the concept of “controlling for” or “adjusting

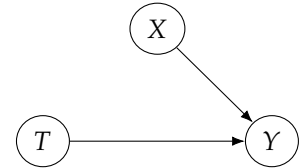
**Table 2.1:** Example data to illustrate that the fundamental problem of causal inference can be interpreted as a missing data problem.



**Figure 2.1:** Causal structure of  $X$  confounding the effect of  $T$  on  $Y$ .

<sup>5</sup> Keep reading to Chapter 3, where we will flesh out and formalize this graphical interpretation.

<sup>6</sup> **Active reading exercise:** verify that this procedure is equivalent to  $\mathbb{E}[Y|T = 1] - \mathbb{E}[Y|T = 0]$  in the data in Table 2.1.



**Figure 2.2:** Causal structure when the treatment assignment mechanism is ignorable. Notably, this means there’s no arrow from  $X$  to  $T$ , which means there is no confounding.

<sup>7</sup> Technically, this is *mean* exchangeability, which is a weaker assumption than the full exchangeability that we describe in Assumption 2.1 because it only constrains the first moment of the distribution. Generally, we only need mean ignorability/exchangeability for average treatment effects, but it is common to assume complete independence, as in Assumption 2.1.

for” variables, which we will discuss shortly when we get to conditional exchangeability.

We have leveraged Assumption 2.1 to identify causal effects. To *identify* a causal effect is to reduce a causal expression to a purely statistical expression. In this chapter, that means to reduce an expression from one that uses potential outcome notation to one that uses only statistical notation such as  $T$ ,  $X$ ,  $Y$ , expectations, and conditioning. This means that we can calculate the causal effect from just the observational distribution  $P(X, T, Y)$ .

**Definition 2.1** (Identifiability) *A causal quantity (e.g.  $\mathbb{E}[Y(t)]$ ) is identifiable if we can compute it from a purely statistical quantity (e.g.  $\mathbb{E}[Y | t]$ ).*

We have seen that ignorability is extremely important (Equation 2.3), but how realistic of an assumption is it? In general, it is completely unrealistic because there is likely to be confounding in most data we observe (causal structure shown in Figure 2.1). However, we can make this assumption realistic by running randomized experiments, which force the treatment to not be caused by anything but a coin toss, so then we have the causal structure shown in Figure 2.2. We cover randomized experiments in greater depth in Chapter 5.

We have covered two prominent perspectives on this main assumption (2.1): ignorability and exchangeability. Mathematically, these mean the same thing, but their names correspond to different ways of thinking about the same assumption. Exchangeability and ignorability are only two names for this assumption. We will see more aliases after we cover the more realistic, conditional version of this assumption.

### 2.3.3 Conditional Exchangeability and Unconfoundedness

In observational data, it is unrealistic to assume that the treatment groups are exchangeable. In other words, there is no reason to expect that the groups are the same in all relevant variables other than the treatment. However, if we control for relevant variables by conditioning, then maybe the subgroups will be exchangeable. We will clarify what the “relevant variables” are in Chapter 3, but for now, let’s just say they are all of the covariates  $X$ . Then, we can state conditional exchangeability formally.

**Assumption 2.2** (Conditional Exchangeability / Unconfoundedness)

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$$

The idea is that although the treatment and potential outcomes may be unconditionally associated (due to confounding), within levels of  $X$ , they are not associated. In other words, there is no confounding within levels of  $X$  because controlling for  $X$  has made the treatment groups comparable. We’ll now give a bit of graphical intuition for the above. We will not draw the rigorous connection between the graphical intuition and Assumption 2.2 until Chapter 3; for now, it is just meant to aid intuition.



We do not have exchangeability in the data because  $X$  is a common cause of  $T$  and  $Y$ . We illustrate this in Figure 2.3. Because  $X$  is a common cause of  $T$  and  $Y$ , there is non-causal association between  $T$  and  $Y$ . This non-causal association flows along the  $T \leftarrow X \rightarrow Y$  path; we depict this with a red dashed arc.

However, we *do* have *conditional* exchangeability in the data. This is because, when we condition on  $X$ , there is no longer any non-causal association between  $T$  and  $Y$ . The non-causal association is now “blocked” at  $X$  by conditioning on  $X$ . We illustrate this blocking in Figure 2.4 by shading  $X$  to indicate it is conditioned on and by showing the red dashed arc being blocked there.

Conditional exchangeability is the main assumption necessary for causal inference. Armed with this assumption, we can identify the causal effect within levels of  $X$ , just like we did with (unconditional) exchangeability:

$$\mathbb{E}[Y(1) - Y(0) | X] = \mathbb{E}[Y(1) | X] - \mathbb{E}[Y(0) | X] \quad (2.5)$$

$$= \mathbb{E}[Y(1) | T = 1, X] - \mathbb{E}[Y(0) | T = 0, X] \quad (2.6)$$

$$= \mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X] \quad (2.7)$$

In parallel to before, we get Equation 2.5 by linearity of expectation. And we now get Equation 2.6 by conditional exchangeability. If we want the marginal effect that we had before when assuming (unconditional) exchangeability, we can get that by simply marginalizing out  $X$ :

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X \mathbb{E}[Y(1) - Y(0) | X] \quad (2.8)$$

$$= \mathbb{E}_X [\mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X]] \quad (2.9)$$

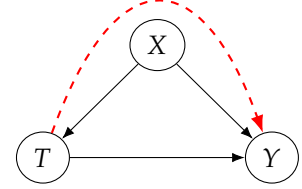
This marks an important result for causal inference, so we’ll give it its own proposition box. The proof we give above leaves out some details. Read through to Section 2.3.6 (where we redo the proof with all details specified) to get the rest of the details. We will call this result the *adjustment formula*.

**Theorem 2.1** (Adjustment Formula) *Given the assumptions of unconfoundedness, positivity, consistency, and no interference, we can identify the average treatment effect:*

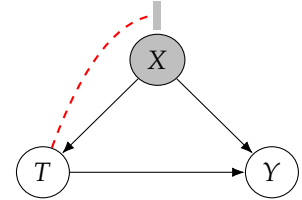
$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X [\mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X]]$$

Conditional exchangeability (Assumption 2.2) is a core assumption for causal inference and goes by many names. For example, the following are reasonably commonly used to refer to the same assumption: unconfoundedness, conditional ignorability, no unobserved confounding, selection on observables, no omitted variable bias, etc. We will use the name “unconfoundedness” a fair amount throughout this book.

The main reason for moving from exchangeability (Assumption 2.1) to conditional exchangeability (Assumption 2.2) was that it seemed like a more realistic assumption. However, we often cannot know for certain if conditional exchangeability holds. There may be some unobserved confounders that are not part of  $X$ , meaning conditional exchangeability is violated. Fortunately, that is not a problem in randomized experiments



**Figure 2.3:** Causal structure of  $X$  **confounding** the effect of  $T$  on  $Y$ . We depict the confounding with a red dashed line.



**Figure 2.4:** Illustration of conditioning on  $X$  leading to no confounding.

(Chapter 5). Unfortunately, it is something that we must always be conscious of in observational data. Intuitively, the best thing we can do is to observe and fit as many covariates into  $X$  as possible to try to ensure unconfoundedness.<sup>8</sup>

<sup>8</sup> As we will see in Chapters 3 and 4, it is not necessarily true that conditioning on more covariates always helps our causal estimates be less biased.

### 2.3.4 Positivity/Overlap and Extrapolation

While conditioning on many covariates is attractive for achieving unconfoundedness, it can actually be detrimental for another reason that has to do with another important assumption that we have yet to discuss: *positivity*. We will get to why at the end of this section. Positivity is the condition that all subgroups of the data with different covariates have some probability of receiving any value of treatment. Formally, we define positivity for binary treatment as follows.

**Assumption 2.3** (Positivity / Overlap / Common Support) *For all values of covariates  $x$  present in the population of interest (i.e.  $x$  such that  $P(X = x) > 0$ ),*

$$0 < P(T = 1 \mid X = x) < 1$$

To see why positivity is important, let's take a closer look at Equation 2.9:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X [\mathbb{E}[Y \mid T = 1, X] - \mathbb{E}[Y \mid T = 0, X]]$$

(2.9 revisited)

In short, if we have a positivity violation, then we will be conditioning on a zero probability event. This is because there will be some value of  $x$  with non-zero probability for which  $P(T = 1 \mid X = x) = 0$  or  $P(T = 0 \mid X = x) = 0$ . This means that for some value of  $x$  that we are marginalizing out in the above equation,  $P(T = 1, X = x) = 0$  or  $P(T = 0, X = x) = 0$ , and these are the two events that we condition on in Equation 2.9.

To clearly see how a positivity violation translates to division by zero, let's rewrite the right-hand side of Equation 2.9. For discrete covariates and outcome, it can be rewritten as follows:

$$\sum_x P(X = x) \left( \sum_y y P(Y = y \mid T = 1, X = x) - \sum_y y P(Y = y \mid T = 0, X = x) \right)$$

(2.10)

Then, applying Bayes' rule, this can be further rewritten:

$$\sum_x P(X = x) \left( \sum_y y \frac{P(Y = y, T = 1, X = x)}{P(T = 1 \mid X = x)P(X = x)} - \sum_y y \frac{P(Y = y, T = 0, X = x)}{P(T = 0 \mid X = x)P(X = x)} \right)$$

(2.11)

In Equation 2.11, we can clearly see why positivity is essential. If  $P(T = 1 \mid X = x) = 0$  for *any* level of covariates  $x$  with non-zero probability, then there is division by zero in the first term in the equation, so  $\mathbb{E}_X \mathbb{E}[Y \mid T = 1, X]$  is undefined. Similarly, if  $P(T = 1 \mid X = x) = 1$  for any level of  $x$ , then  $P(T = 0 \mid X = x) = 0$ , so there is division by zero in the second term and  $\mathbb{E}_X \mathbb{E}[Y \mid T = 0, X]$  is undefined. With either of these violations of the positivity assumption, the causal effect is undefined.

**Intuition** That’s the math for why we need the positivity assumption, but what’s the intuition? Well, if we have a positivity violation, that means that within some subgroup of the data, everyone always receives treatment or everyone always receives the control. It wouldn’t make sense to be able to estimate a causal effect of treatment vs. control in that subgroup since we see only treatment or only control. We never see the alternative in that subgroup.

Another name for positivity is *overlap*. The intuition for this name is that we want the covariate distribution of the treatment group to overlap with the covariate distribution of the control group. More specifically, we want  $P(X | T = 1)$ <sup>9</sup> to have the same support as  $P(X | T = 0)$ .<sup>10</sup> This is why another common alias for positivity is *common support*.

**The Positivity-Unconfoundedness Tradeoff** Although conditioning on more covariates could lead to a higher chance of satisfying unconfoundedness, it can lead to a higher chance of violating positivity. As we increase the dimension of the covariates, we make the subgroups for any level  $x$  of the covariates smaller.<sup>11</sup> As each subgroup gets smaller, there is a higher and higher chance that either the whole subgroup will have treatment or the whole subgroup will have control. For example, once the size of any subgroup has decreased to one, positivity is guaranteed to not hold. See [6] for a rigorous argument of high-dimensional covariates leading to positivity violations.

**Extrapolation** Violations of the positivity assumption can actually lead to demanding too much from models and getting very bad behavior in return. Many causal effect estimators<sup>12</sup> fit a model to  $\mathbb{E}[Y|t, x]$  using the  $(t, x, y)$  tuples as data. The inputs to these models are  $(t, x)$  pairs and the outputs are the corresponding outcomes. These models will be forced to extrapolate in regions (using their parametric assumptions) where  $P(T = 1, X = x) = 0$  and regions where  $P(T = 0, X = x) = 0$  when they are used in the adjustment formula (Theorem 2.1) in place of the corresponding conditional expectations.

<sup>9</sup> Whenever we use a random variable (denoted by a capital letter) as the argument for  $P$ , we are referring to the whole distribution, rather than just the scalar that something like  $P(x | T = 1)$  refers to.

<sup>10</sup> **Active reading exercise:** convince yourself that this formulation of overlap/positivity is equivalent to the formulation in Assumption 2.3.

<sup>11</sup> This is related to the *curse of dimensionality*.

[6]: D’Amour et al. (2017), *Overlap in Observational Studies with High-Dimensional Covariates*

<sup>12</sup> An “estimator” is a function that takes a dataset as input and outputs an estimate. We discuss this statistics terminology more in Section 2.4.

### 2.3.5 No interference, Consistency, and SUTVA

There are a few additional assumptions we’ve been smuggling in throughout this chapter. We will specify all the rest of these assumptions in this section. The first assumption in this section is that of *no interference*. No interference means that my outcome is unaffected by anyone else’s treatment. Rather, my outcome is only a function of my own treatment. We’ve been using this assumption implicitly throughout this chapter. We’ll now formalize it.

#### Assumption 2.4 (No Interference)

$$Y_i(t_1, \dots, t_{i-1}, t_i, t_{i+1}, \dots, t_n) = Y_i(t_i)$$

Of course, this assumption could be violated. For example, if the treatment is “get a dog” and the outcome is my happiness, it could easily be that my happiness is influenced by whether or not my friends get dogs because we could end up hanging out more to have our dogs play together. As you

might expect, violations of the no interference assumption are rampant in network data.

The last assumption is *consistency*. Consistency is the assumption that the outcome we observe  $Y$  is actually the potential outcome under the observed treatment  $T$ .

**Assumption 2.5 (Consistency)** *If the treatment is  $T$ , then the observed outcome  $Y$  is the potential outcome under treatment  $T$ . Formally,*

$$T = t \implies Y = Y(t) \quad (2.12)$$

*We could write this equivalently as follow:*

$$Y = Y(T) \quad (2.13)$$

Note that  $T$  is different from  $t$ , and  $Y(T)$  is different from  $Y(t)$ .  $T$  is a random variable that corresponds to the observed treatment, whereas  $t$  is a specific value of treatment. Similarly,  $Y(t)$  is the potential outcome for some specific value of treatment, whereas  $Y(T)$  is the potential outcome for the actual value of treatment that we observe.

When we were using exchangeability to prove identifiability, we actually assumed consistency in Equation 2.4 to get the follow equality:

$$\mathbb{E}[Y(1) | T = 1] - \mathbb{E}[Y(0) | T = 0] = \mathbb{E}[Y | T = 1] - \mathbb{E}[Y | T = 0]$$

Similarly, when we were using conditional exchangeability to prove identifiability, we assumed consistency in Equation 2.7.

It might seem like consistency is obviously true, but that is not always the case. For example, if the treatment specification is simply “get a dog” or “don’t get a dog,” this can be too coarse to yield consistency. It might be that if I were to get a puppy, I would observe  $Y = 1$  (happiness) because I needed an energetic friend, but if I were to get an old, low-energy dog, I would observe  $Y = 0$  (unhappiness). However, both of these treatments fall under the category of “get a dog,” so both correspond to  $T = 1$ . This means that  $Y(1)$  is not well defined, since it will be 1 or 0, depending on something that is not captured by the treatment specification. In this sense, consistency encompasses the assumption that is sometimes referred to as “no multiple versions of treatment.” See Sections 3.4 and 3.5 of Hernán and Robins [7] and references therein for more discussion on this topic.

**SUTVA** You will also commonly see the *stable unit-treatment value assumption* (SUTVA) in the literature. SUTVA is satisfied if unit (individual)  $i$ ’s outcome is simply a function of unit  $i$ ’s treatment. Therefore, SUTVA is a combination of consistency and no interference (and also deterministic potential outcomes).<sup>13</sup>

[7]: Hernán and Robins (2020), *Causal Inference: What If*

<sup>13</sup> **Active reading exercise:** convince yourself that SUTVA is a combination of consistency and no inference

### 2.3.6 Tying It All Together

We introduced unconfoundedness (conditional exchangeability) first because it is the main causal assumption. However, all of the assumptions are necessary:

1. Unconfoundedness (Assumption 2.2)
2. Positivity (Assumption 2.3)
3. No interference (Assumption 2.4)
4. Consistency (Assumption 2.5)

We'll now review the proof of the adjustment formula (Theorem 2.1) that was done in Equation 2.5 through Equation 2.9 and list which assumptions are used for each step. Even before we get to these equations, we use the no interference assumption to justify that the quantity we should be looking at for causal inference is  $\mathbb{E}[Y(1) - Y(0)]$ , rather than something more complex like the left-hand side of Assumption 2.4. In the proof below, the first two equalities follow from mathematical facts, whereas the last two follow from these key assumptions.

*Proof of Theorem 2.1.*

$$\begin{aligned}
 \mathbb{E}[Y(1) - Y(0)] &= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] && \text{(linearity of expectation)} \\
 &= \mathbb{E}_X [\mathbb{E}[Y(1) | X] - \mathbb{E}[Y(0) | X]] \\
 &&& \text{(law of iterated expectations)} \\
 &= \mathbb{E}_X [\mathbb{E}[Y(1) | T = 1, X] - \mathbb{E}[Y(0) | T = 0, X]] \\
 &&& \text{(unconfoundedness and positivity)} \\
 &= \mathbb{E}_X [\mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X]] \\
 &&& \text{(consistency)}
 \end{aligned}$$

□

That's how all of these assumptions tie together to give us identifiability of the ATE. We'll soon see how to use this result to get an actual estimated number for the ATE.

## 2.4 Fancy Statistics Terminology Defancified

Before we start computing concrete numbers for the ATE, we must quickly introduce some terminology from statistics that will help clarify the discussion. An *estimand* is the quantity that we want to estimate. For example,  $\mathbb{E}_X [\mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X]]$  is the estimand we care about for estimating the ATE. An *estimate* (noun) is an approximation of some estimand, which we get using data. We will see concrete numbers in the next section; these are estimates. Given some estimand  $\alpha$ , we write an estimate of that estimand by simply putting a hat on it:  $\hat{\alpha}$ . And an *estimator* is a function that maps a dataset to an estimate of the estimand. The process that we will use to go from data + estimand to a concrete number is known as *estimation*. To *estimate* (verb) is to feed data into an estimator to get an estimate.

In this book, we will use even more specific language that allows us to make the distinction between causal quantities and statistical quantities. We will use the phrase *causal estimand* to refer to any estimand that contains a potential outcome in it. We will use the phrase *statistical estimand* to denote the complement: any estimand that does not contain a potential outcome.<sup>14</sup> For an example, recall the adjustment formula

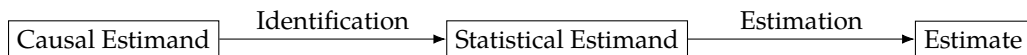
<sup>14</sup> As we will see in Chapter 4, we will equivalently refer to a causal estimand as any estimand that contains a *do*-operator, and we will refer to a statistical estimand as any estimand that does not contain a *do*-operator.

(Theorem 2.1):

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X [\mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X]] \quad (2.14)$$

$\mathbb{E}[Y(1) - Y(0)]$  is the causal estimand that we are interested in. In order to actually estimate this causal estimand, we must translate it into a statistical estimand:  $\mathbb{E}_X [\mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X]]$ .<sup>15</sup>

When we say “identification” in this book, we are referring to the process of moving from a *causal* estimand to an equivalent *statistical* estimand. When we say “estimation,” we are referring to the process of moving from a statistical estimand to an estimate. We illustrate this in the flowchart in Figure 2.5.



**Figure 2.5:** The Identification-Estimation Flowchart – a flowchart that illustrates the process of moving from a target causal estimand to a corresponding estimate, through identification and estimation.

What do we do when we go to actually estimate quantities such as  $\mathbb{E}_X [\mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X]]$ ? We will often use a model (e.g. linear regression or some more fancy predictor from machine learning) in place of the conditional expectations  $\mathbb{E}[Y | T = t, X = x]$ . We will refer to estimators that use models like this as *model-assisted estimators*. Now that we’ve gotten some of this terminology out of the way, we can proceed to an example of estimating the ATE.

## 2.5 A Complete Example with Estimation

Theorem 2.1 and the corresponding recent copy in Equation 2.14 give us identification. However, we haven’t discussed estimation at all. In this section, we will give a short example complete with estimation. We will cover the topic of estimation of causal effects more completely in Chapter 7.

We use Luque-Fernandez et al. [8]’s example from epidemiology. The outcome  $Y$  of interest is (systolic) blood pressure. This is an important outcome because roughly 46% of Americans have high blood pressure, and high blood pressure is associated with increased risk of mortality [9]. The “treatment”  $T$  of interest is sodium intake. Sodium intake is a continuous variable; in order to easily apply Equation 2.14, which is specified for binary treatment, we will binarize  $T$  by letting  $T = 1$  denote daily sodium intake above 3.5 grams and letting  $T = 0$  denote daily sodium intake below 3.5 grams.<sup>16</sup> We will be estimating the causal effect of sodium intake on blood pressure. In our data, we also have the age of the individuals and amount of protein in their urine as covariates  $X$ . Luque-Fernandez et al. [8] run a simulation, taking care to be sure that the range of values is “biologically plausible and as close to reality as possible.”

Because we are using data from a simulation, we know that the true ATE of sodium on blood pressure is 1.05. More concretely, the line of code that generates blood pressure  $Y$  looks as follows:

```
1 | blood_pressure = 1.05 * sodium + ...
```

<sup>15</sup> **Active reading exercise:** Why can’t we directly estimate a causal estimand without first translating it to a statistical estimand?

[8]: Luque-Fernandez et al. (2018), ‘Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application’

[9]: Virani et al. (2020), ‘Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association’

<sup>16</sup> As we will see, this binarization is purely pedagogical and does not reflect any limitations of adjusting for confounders.

Now, how do we actually estimate the ATE? First, we assume consistency, positivity, and unconfoundedness given  $X$ . As we recently recalled in Equation 2.14, this means that we've identified the ATE as

$$\mathbb{E}_X [\mathbb{E}[Y | T = 1, X] - \mathbb{E}[Y | T = 0, X]] .$$

We then take that outer expectation over  $X$  and replace it with an empirical mean over the data, giving us the following:

$$\frac{1}{n} \sum_i [\mathbb{E}[Y | T = 1, X = x_i] - \mathbb{E}[Y | T = 0, X = x_i]] \quad (2.15)$$

To complete our estimator, we then fit some machine learning model to the conditional expectation  $\mathbb{E}[Y | t, x]$ . Minimizing the mean-squared error (MSE) of predicting  $Y$  from  $(T, X)$  pairs is equivalent to modeling this conditional expectation [see, e.g., 10, Section 2.4]. Therefore, we can plug in any machine learning model for  $\mathbb{E}[Y | t, x]$ , which gives us a model-assisted estimator. We'll use linear regression here, which works out nicely since blood pressure is generated as a linear combination of other variables, in this simulation. We give Python code for this below, where our data are in a Pandas DataFrame called `df`. We fit the model for  $\mathbb{E}[Y | t, x]$  in line 8, and we take the empirical mean over  $X$  in lines 10-14.

[10]: Hastie et al. (2001), *The Elements of Statistical Learning*

```
1 import numpy as np
2 import pandas as pd
3 from sklearn.linear_model import LinearRegression
4
5 Xt = df[['sodium', 'age', 'proteinuria']]
6 y = df['blood_pressure']
7 model = LinearRegression()
8 model.fit(Xt, y)
9
10 Xt1 = pd.DataFrame.copy(Xt)
11 Xt1['sodium'] = 1
12 Xt0 = pd.DataFrame.copy(Xt)
13 Xt0['sodium'] = 0
14 ate_est = np.mean(model.predict(Xt1) - model.predict(Xt0))
15 print('ATE estimate:', ate_est)
```

**Listing 2.1:** Python code for estimating the ATE

Full code, complete with simulation, is available at [https://github.com/bradyneal/causal-book-code/blob/master/sodium\\_example.py](https://github.com/bradyneal/causal-book-code/blob/master/sodium_example.py).

This yields an ATE estimate of 0.85. If we were to naively regress  $Y$  on only  $T$ , which corresponds to replacing line 5 in Listing 2.1 with `Xt = df[['sodium']]`,<sup>17</sup> we would get an ATE estimate of 5.33. That's a  $\frac{5.33-1.05}{1.05} \times 100\% = 407\%$  error! In contrast, when we control for  $X$  (as in Listing 2.1), our percent error is only  $\frac{0.85-1.05}{1.05} \times 100\% = 19\%$ .

<sup>17</sup> **Active reading exercise:** This naive version is equivalent to just taking the associational difference:  $\mathbb{E}[Y | T = 1] - \mathbb{E}[Y | T = 0]$ . Why?

All of the above is done using the adjustment formula with model-assisted estimation, where we first fit a model for the conditional expectation  $\mathbb{E}[Y | t, x]$ , and then we take an empirical mean over  $X$ , using that model. However, because we are using a linear model, this is equivalent to just taking the coefficient in front of  $T$  in the linear regression as the ATE estimate. This is what we do in the following code (which gives the exact same ATE estimate):

```
1 Xt = df[['sodium', 'age', 'proteinuria']]
2 y = df['blood_pressure']
3 model = LinearRegression()
```

**Listing 2.2:** Python code for estimating the ATE using the coefficient of linear regression



```

4 | model.fit(Xt, y)
5 | ate_est = model.coef_[0]
6 | print('ATE estimate:', ate_est)

```

**Continuous Treatment** What if we allow the treatment, daily sodium intake, to remain continuous, instead of binarizing it? The cool thing about just taking the regression coefficient as the ATE estimate is that it doesn't require taking a difference between two values of treatment (e.g.  $T = 1$  and  $T = 0$ ), so it trivially generalizes to when  $T$  is continuous. When  $T$  is continuous, we care about how  $\mathbb{E}[Y(t)]$  changes with  $t$ . Since we are assuming  $\mathbb{E}[Y(t)]$  is linear, this change is completely captured by  $\frac{d}{dt}\mathbb{E}[Y(t)]$ .<sup>18</sup> When  $\mathbb{E}[Y(t)]$  is linear, it turns out that this quantity is exactly what taking the coefficient from linear regression estimates. Seemingly magically, we have compressed all of  $\mathbb{E}[Y(t)] = \mathbb{E}[Y | t]$ , which is a *function* of  $t$ , into a single value.

However, this effortless compression of all of  $\mathbb{E}[Y | t]$  for continuous  $t$  comes as a cost: the linear parametric form we assumed. If this model were misspecified,<sup>19</sup> our ATE estimate would be biased. And because linear models are so simple, they will likely be misspecified. For example, the following assumption is implicit in assuming that a linear model is well-specified: the treatment effect is the same for all individuals. See Morgan and Winship [12, Sections 6.2 and 6.3] for a more complete critique of using the coefficient in front of treatment as the ATE estimate.

<sup>18</sup> Concisely summarizing *nonlinear* functions  $\mathbb{E}[Y(t)]$  is an open problem. See, e.g., Janzing et al. [11].

[11]: Janzing et al. (2013), 'Quantifying causal influences'

<sup>19</sup> By "misspecified," we mean that the functional form of the model does not match the functional form of the data generating process.

[12]: Morgan and Winship (2014), *Counterfactuals and Causal Inference: Methods and Principles for Social Research*



# The Flow of Association and Causation in Graphs

# 3

We've been using causal graphs in the previous chapters to aid intuition. In this chapter, we will introduce the formalisms that underlie this intuition. Hopefully, we have sufficiently motivated this chapter and made the utility of graphical models clear with all of the graphical interpretations of concepts in previous chapters.

## 3.1 Graph Terminology

In this section, we will use the *terminology machine gun* (see Figure 3.1). To be able to use nice convenient graph language in the following sections, rapid-firing a lot of graph terminology is a necessary evil, unfortunately.

The term “graph” is often used to describe a variety of visualizations. For example, “graph” might refer to a visualization of a single variable function  $f(x)$ , where  $x$  is plotted on the  $x$ -axis and  $f(x)$  is plotted on the  $y$ -axis. Or “bar graph” might be used as a synonym for a bar chart. However, in graph theory, the term “graph” refers to a specific mathematical object.

A *graph* is a collection of *nodes* (also called “vertices”) and *edges* that connect the nodes. For example, in Figure 3.2,  $A$ ,  $B$ , and  $C$  are the nodes of the graph, and the lines connecting them are the edges. Figure 3.2 is called an *undirected graph* because the edges do not have any direction. In contrast, Figure 3.3 is a *directed graph*. A directed graph's edges go out of a *parent* node and into a *child* node, with the arrows signifying which direction the edges are going. We will denote the parents of a node  $X$  with  $\text{pa}(X)$ . We'll use an even simpler shorthand when the nodes are ordered so that we can denote the  $i^{\text{th}}$  node by  $X_i$ ; in that case, we will also denote the parents of  $X_i$  by  $\text{pa}_i$ . Two nodes are said to be *adjacent* if they are connected by an edge. For example, in both Figure 3.2 and Figure 3.3,  $A$  and  $C$  are adjacent, but  $A$  and  $D$  are not.

A *path* in a graph is any sequence of adjacent nodes, regardless of the direction of the edges that join them. For example,  $A - C - B$  is a path in Figure 3.2, and  $A \rightarrow C \leftarrow B$  is a path in Figure 3.3. A *directed path* is a path that consists of directed edges that are all directed in the same direction (no two edges along the path both point into or both point out of the same node). For example,  $A \rightarrow C \rightarrow D$  is a directed path in Figure 3.3, but  $A \rightarrow C \leftarrow B$  and  $C \leftarrow A \rightarrow B$  are not.

If there is a directed path that starts at node  $X$  and ends at node  $Y$ , then  $X$  is an *ancestor* of  $Y$ , and  $Y$  is a *descendant* of  $X$ . We will denote descendants of  $X$  by  $\text{de}(X)$ . For example, in Figure 3.3,  $A$  is an ancestor of  $B$  and  $D$ , and  $B$  and  $D$  are both descendants of  $A$  ( $\text{de}(A)$ ). If  $X$  is an ancestor of itself, then some funky time travel has taken place. In seriousness, a directed path from some node  $X$  back to itself is known as a *cycle* (see Figure 3.4). If there are no cycles in a directed graph, the graph is known

3.1 Graph Terminology . . . . .	19
3.2 Bayesian Networks . . . . .	20
3.3 Causal Graphs . . . . .	22
3.4 Two-Node Graphs and Graphical Building Blocks . . . . .	23
3.5 Chains and Forks . . . . .	24
3.6 Colliders and their Descendants . . . . .	26
3.7 d-separation . . . . .	28
3.8 Flow of Association and Causation . . . . .	30

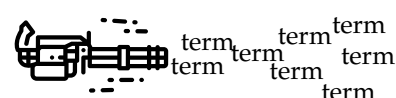


Figure 3.1: Terminology machine gun

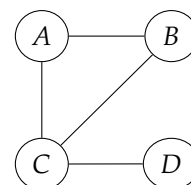


Figure 3.2: Undirected graph

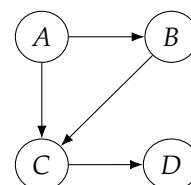


Figure 3.3: Directed graph

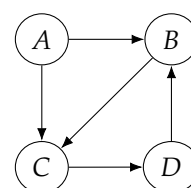
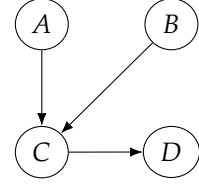


Figure 3.4: Directed graph with cycle

as a *directed acyclic graph* (DAG). The graphs we focus on in this book will mostly be DAGs.

If two parents  $X$  and  $Y$  share some child  $Z$ , but there is no edge connecting  $X$  and  $Y$ , then  $X \rightarrow Z \leftarrow Y$  is known as an *immorality*. Seriously; that's a real term in graphical models. For example, if we remove the  $A \rightarrow B$  from Figure 3.3 to get Figure 3.5, then  $A \rightarrow C \leftarrow B$  is an immorality.



**Figure 3.5:** Directed graph with immorality

## 3.2 Bayesian Networks

It turns out that much of the work for causal graphical models was done in the field of probabilistic graphical models. Probabilistic graphical models are statistical models while causal graphical models are causal models. Bayesian networks are the main probabilistic graphical model that causal graphical models (causal Bayesian networks) inherit most of their properties from.

Imagine that we only cared about modeling association, without any causal modeling. We would want to model the data distribution  $P(x_1, x_2, \dots, x_n)$ . In general, we can use the chain rule of probability to factorize any distribution:

$$P(x_1, x_2, \dots, x_n) = P(x_1) \prod_i P(x_i | x_{i-1}, \dots, x_1) \quad (3.1)$$

However, if we were to model these factors with tables, it would take an exponential number of parameters. To see this, take each  $x_i$  to be binary and consider how we would model the factor  $P(x_n | x_{n-1}, \dots, x_1)$ . Since  $x_n$  is binary, we only need to model  $P(X_n = 1 | x_{n-1}, \dots, x_1)$  because  $P(X_n = 0 | x_{n-1}, \dots, x_1)$  is simply  $1 - P(X_n = 1 | x_{n-1}, \dots, x_1)$ . Well, we would need  $2^{n-1}$  parameters to model this. As a specific example, let  $n = 4$ . As we can see in Table 3.1, this would require  $2^{4-1} = 8$  parameters:  $\alpha_1, \dots, \alpha_8$ . This brute-force parametrization quickly becomes intractable as  $n$  increases.

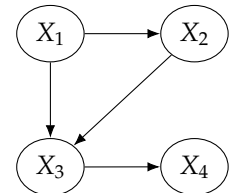
An intuitive way to more efficiently model many variables together in a joint distribution is to only model local dependencies. For example, rather than modeling the  $X_4$  factor as  $P(x_4 | x_3, x_2, x_1)$ , we could model it as  $P(x_4 | x_3)$  if we have reason to believe that  $X_4$  only locally depends on  $X_3$ . In fact, in the corresponding graph in Figure 3.6, the only node that feeds into  $X_4$  is  $X_3$ . This is meant to signify that  $X_4$  only locally depends on  $X_3$ . Whenever we use a graph  $G$  in relation to a probability distribution  $P$ , there will always be a one-to-one mapping between the nodes in  $G$  and the random variables in  $P$ , so when we talk about nodes being independent, we mean the corresponding random variables are independent.

Given a probability distribution and a corresponding directed acyclic graph (DAG), we can formalize the specification of independencies with the *local Markov assumption*:

**Assumption 3.1** (Local Markov Assumption) *Given its parents in the DAG, a node  $X$  is independent of all its non-descendants.*

**Table 3.1:** Table required to model the single factor  $P(x_n | x_{n-1}, \dots, x_1)$  where  $n = 4$  and the variables are binary. The number of parameters to necessary is exponential in  $n$ .

$x_1$	$x_2$	$x_3$	$P(x_4   x_3, x_2, x_1)$
0	0	0	$\alpha_1$
0	0	1	$\alpha_2$
0	1	0	$\alpha_3$
0	1	1	$\alpha_4$
1	0	0	$\alpha_5$
1	0	1	$\alpha_6$
1	1	0	$\alpha_7$
1	1	1	$\alpha_8$



**Figure 3.6:** Four node DAG where  $X_4$  locally depends on only  $X_3$ .

This assumption (along with specific DAGs) gives us a lot. We will demonstrate this in the next few equations. In our four variable example, the chain rule of probability tells us that we can factorize any  $P$  such that

$$P(x_1, x_2, x_3, x_4) = P(x_1) P(x_2 \mid x_1) P(x_3 \mid x_2, x_1) P(x_4 \mid x_3, x_2, x_1). \quad (3.2)$$

If  $P$  is Markov with respect to the graph<sup>1</sup> in Figure 3.6, then we can simplify the last factor:

$$P(x_1, x_2, x_3, x_4) = P(x_1) P(x_2 \mid x_1) P(x_3 \mid x_2, x_1) P(x_4 \mid x_3). \quad (3.3)$$

If we further remove edges, removing  $X_1 \rightarrow X_2$  and  $X_2 \rightarrow X_3$  as in Figure 3.7, we can further simplify the factorization of  $P$ :

$$P(x_1, x_2, x_3, x_4) = P(x_1) P(x_2) P(x_3 \mid x_1) P(x_4 \mid x_3). \quad (3.4)$$

With the understanding that we have hopefully built up from a few examples,<sup>2</sup> we will now state one of the main consequences of the local Markov assumption:

**Definition 3.1** (Bayesian Network Factorization) *Given a probability distribution  $P$  and a DAG  $G$ ,  $P$  factorizes according to  $G$  if*

$$P(x_1, \dots, x_n) = \prod_i P(x_i \mid \text{pa}_i)$$

Hopefully you see the resemblance between the move from Equation 3.2 to Equation 3.3 or the move to Equation 3.4 and the generalization of this that is presented in Definition 3.1.

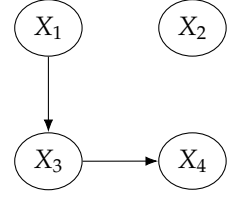
The Bayesian network factorization is also known as the *chain rule for Bayesian networks* or *Markov compatibility*. For example, if  $P$  factorizes according to  $G$ , then  $P$  and  $G$  are Markov compatible.

We have given the intuition of how the local Markov assumption implies the Bayesian network factorization, and it turns out that the two are actually equivalent. In other words, we could have started with the Bayesian network factorization as the main assumption (and labeled it as an assumption) and shown that it implies the local Markov assumption. See Koller and Friedman [13, Chapter 3] for these proofs and more information on this topic.

As important as the local Markov assumption is, it only gives us information about the *independencies* in  $P$  that a DAG implies. It does not even tell us that if  $X$  and  $Y$  are adjacent in the DAG, then  $X$  and  $Y$  are dependent. And this additional information is very commonly assumed in causal DAGs. To get this guaranteed dependence between adjacent nodes, we will generally assume a slightly stronger assumption than the local Markov assumption: *minimality*.

**Assumption 3.2** (Minimality Assumption) 1. *Given its parents in the DAG, a node  $X$  is independent of all its non-descendants (Assumption 3.1).*  
2. *Adjacent nodes in the DAG are dependent.*<sup>3</sup>

<sup>1</sup> A probability distribution is said to be (locally) Markov with respect to a DAG if they satisfy the local Markov assumption.



**Figure 3.7:** Four node DAG with even more independencies.

<sup>2</sup> **Active reading exercise:** ensure that you know how we get from Equation 3.2 to Equation 3.3 and to Equation 3.4 using the local Markov assumption.

[13]: Koller and Friedman (2009), *Probabilistic Graphical Models: Principles and Techniques*

<sup>3</sup> This is often equivalently stated in the following way: if we were to remove any edges from the DAG,  $P$  would not be Markov with respect to the graph with the removed edges [see, e.g., 14, Section 6.5.3].

[14]: Peters et al. (2017), *Elements of Causal Inference: Foundations and Learning Algorithms*

To see why this assumption is named “minimality” consider, what we know when we know that  $P$  is Markov with respect to a DAG  $G$ . We know that  $P$  satisfies a set of independencies that are specific to the structure of  $G$ . If  $P$  and  $G$  also satisfy minimality, then this set of independencies is *minimal* in the sense the  $P$  does not satisfy any additional independencies. This is equivalent to saying that adjacent nodes are dependent.

For example, if the DAG were simply two connected nodes  $X$  and  $Y$  as in Figure 3.8, the local Markov assumption would tell us that we can factorize  $P(x, y)$  as  $P(x)P(y|x)$ , but it would also allow us to factorize  $P(x, y)$  as  $P(x)P(y)$ , meaning it allows distributions where  $X$  and  $Y$  are independent. In contrast, the minimality assumption does not allow this additional independence. Minimality would tell us to factorize  $P(x, y)$  as  $P(x)P(y|x)$ , and it would tell us that no additional independencies ( $X \perp\!\!\!\perp Y$ ) exist in  $P$  that are minimal with respect to Figure 3.8.

Because removing edges in a Bayesian network is equivalent to adding independencies,<sup>4</sup> the minimality assumption is equivalent to saying that we can’t remove any more edges from the graph. In a sense, every edge is “active.” More concretely, consider that  $P$  and  $G$  are Markov compatible and that  $G'$  is what we get when we remove some edge from  $G$ . If  $P$  is also Markov with respect to  $G'$ , then  $P$  is not minimal with respect to  $G$ .

Armed with the minimality assumption and what it implies about how distributions factorize when they are Markov with respect to some DAG (Definition 3.1), we are now ready to discuss the flow of association in DAGs. However, because everything in this section is purely statistical, we are not ready to discuss the flow of *causation* in DAGs. To do that, we must make causal assumptions. Pedagogically, this will also allow us to use intuitive causal language when we explain the flow of association.

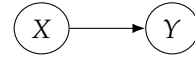


Figure 3.8: Two connected nodes

<sup>4</sup> **Active reading exercise:** why is removing edges in a Bayesian network equivalent to adding independencies?

### 3.3 Causal Graphs

The previous section was all about statistical models and modeling association. In this section, we will augment these models with causal assumptions, turning them into causal models and allowing us to study causation. In order to introduce causal assumptions, we must first have an understanding of what it means for  $X$  to be a cause of  $Y$ .

**Definition 3.2** (What is a cause?) *A variable  $X$  is said to be a cause of a variable  $Y$  if  $Y$  can change in response to changes in  $X$ .*<sup>5</sup>

Another phrase commonly used to describe this primitive is that  $Y$  “listens” to  $X$ . With this, we can now specify the main causal assumption that we will use throughout this book.

**Assumption 3.3** ((Strict) Causal Edges Assumption) *In a directed graph, every parent is a direct cause of all its children.*

Here, the set of *direct causes* of  $Y$  is everything that  $Y$  directly responds to; if we fix all of the direct causes of  $Y$ , then changing any other cause of  $Y$  won’t induce any changes in  $Y$ . This assumption is “strict” in the sense

<sup>5</sup> See Section 4.5.1 for a definition using mathematical notation.

that every edge is “active,” just like in DAGs that satisfy minimality. In other words, because the definition of a cause (Definition 3.2) implies that a cause and its effect are dependent and because we are assuming all parents are causes of their children, we are assuming that parents and their children are dependent. So the second part of minimality (Assumption 3.2) is baked into the strict causal edges assumption.

In contrast, the *non-strict* causal edges assumption would allow for some parents to not be causes of their children. It would just assume that children are not causes of their parents. This allows us to draw graphs with extra edges to make fewer assumptions, just like we would in Bayesian networks, where more edges means fewer independence assumptions. Causal graphs are sometimes drawn with this kind of non-minimal meaning, but the vast majority of the time, when someone draws a causal graph, they mean that parents are causes of their children. Therefore, unless we specify otherwise, throughout this book, we will use “causal graph” to refer to a DAG that satisfies the strict causal edges assumption. And we will often omit the word “strict” when we refer to this assumption.

When we add the causal edges assumption, directed paths in the DAG take on a very special meaning; they correspond to causation. This is in contrast to other paths in the graph, which association may flow along, but causation certainly may not. This will become more clear when we go into detail on these other kinds of paths in Sections 3.5 and 3.6.

Moving forward, we will now think of the edges of graphs as causal, in order to describe concepts intuitively with causal language. However, all of the associational claims about statistical independence will still hold, even when the edges do not have causal meaning like in the vanilla Bayesian networks of Section 3.2.

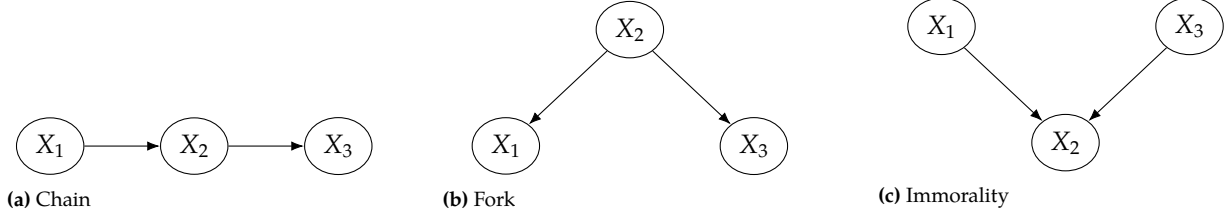
As we will see in the next few sections, the main assumptions that we need for our causal graphical models to tell us how association and causation flow between variables are the following two:

1. Local Markov Assumption (Assumption 3.1)
2. Causal Edges Assumption (Assumption 3.3)

We will discuss these assumptions throughout the next few sections and come back to discuss them more fully again in Section 3.8 after we’ve established the necessary preliminaries.

## 3.4 Two-Node Graphs and Graphical Building Blocks

Now that we’ve gotten the basic assumptions and definitions out of the way, we can get to the core of this chapter: the flow of association and causation in DAGs. We can understand this flow in general DAGs by understanding the flow in the minimal building blocks of graphs. These minimal building blocks consist of chains (Figure 3.9a), forks (Figure 3.9b), immoralities (Figure 3.9c), two unconnected nodes (Figure 3.10), and two connected nodes (Figure 3.11).



**Figure 3.9:** Basic graph building blocks

By “flow of association,” we mean whether any two nodes in a graph are associated or not associated. Another way of saying this is whether two nodes are (statistically) dependent or (statistically) independent. Additionally, we will study whether two nodes are *conditionally* independent or not.

For each building block, we will give the intuition for why two nodes are (conditionally) independent or not, and we will give a proof as well. We can prove that two nodes  $A$  and  $B$  are conditionally independent given some set of nodes  $C$  by simply showing that  $P(a, b|c)$  factorizes as  $P(a|c)P(b|c)$ . We will now do this in the case of the simplest basic building block: two unconnected nodes.

Given a graph that is just two unconnected nodes, as depicted in Figure 3.10, these nodes are not associated simply because there is no edge between them. To show this, consider the factorization of  $P(x_1, x_2)$  that the Bayesian network factorization (Definition 3.1) gives us:

$$P(x_1, x_2) = P(x_1)P(x_2) \quad (3.5)$$

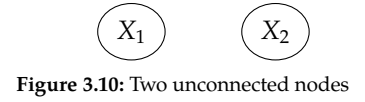
That’s it; applying the Bayesian network factorization immediately gives us a proof that the two nodes  $X_1$  and  $X_2$  are unassociated (independent) in this building block. And what is the assumption that allows us to prove this? That  $P$  is Markov with respect to the graph in Figure 3.10.

In contrast, if there is an edge between the two nodes (as in Figure 3.11), then the two nodes are associated. The assumption we leverage here is the causal edges assumption (Assumption 3.3), which means that  $X_1$  is a cause of  $X_2$ . Since  $X_1$  is a cause of  $X_2$ ,  $X_2$  must be able to change in response to changes in  $X_1$ , so  $X_2$  and  $X_1$  are associated. In general, any time two nodes are adjacent in a causal graph, they are associated.<sup>6</sup> We will see this same concept several more times in Section 3.5 and Section 3.6.

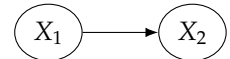
Now that we’ve covered the relevant two-node graphs, we’ll cover the flow of association in the remaining graphical building blocks (three-node graphs in Figure 3.9), starting with chain graphs.

### 3.5 Chains and Forks

Chains (Figure 3.12) and forks (Figure 3.13) share the same set of dependencies. In both structures,  $X_1$  and  $X_2$  are dependent, and  $X_2$  and  $X_3$  are dependent for the same reason that we discussed toward the end of Section 3.4. Adjacent nodes are always dependent when we make the causal edges assumption (Assumption 3.3). What about  $X_1$  and  $X_3$ ,

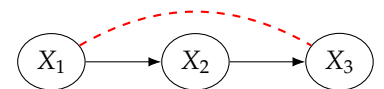


**Figure 3.10:** Two unconnected nodes



**Figure 3.11:** Two connected nodes

<sup>6</sup> Two adjacent nodes in a *non-strict* causal graph can be unassociated.



**Figure 3.12:** Chain with flow of association drawn as a dashed red arc.



though? Does association flow from  $X_1$  to  $X_3$  through  $X_2$  in chains and forks?

Usually, yes,  $X_1$  and  $X_3$  are associated in both chains and forks. In chain graphs,  $X_1$  and  $X_3$  are usually dependent simply because  $X_1$  causes changes in  $X_2$  which then causes changes in  $X_3$ . In a fork graph,  $X_1$  and  $X_3$  are also usually dependent. This is because the same value that  $X_2$  takes on is used to determine both the value that  $X_1$  takes on and the value that  $X_3$  takes on. In other words,  $X_1$  and  $X_3$  are associated through their (shared) common cause. We use the word “usually” throughout this paragraph because there exist pathological cases where the conditional distributions  $P(x_2|x_1)$  and  $P(x_3|x_2)$  are misaligned in such a specific way that makes  $X_1$  and  $X_3$  not actually associated [see, e.g., 15, Section 2.2].

An intuitive graphical way of thinking about  $X_1$  and  $X_3$  being associated in chains and forks is to visualize the flow of association. We visualize this with a dashed red line in Figure 3.12 and Figure 3.13. In the chain graph (Figure 3.12), association flows from  $X_1$  to  $X_3$  along the path  $X_1 \rightarrow X_2 \rightarrow X_3$ . Symmetrically, association flows from  $X_3$  to  $X_1$  along that same path, just running opposite the arrows. In the fork graph (Figure 3.13), association flows from  $X_1$  to  $X_3$  along the path  $X_1 \leftarrow X_2 \rightarrow X_3$ . And similarly, we can think of association flowing from  $X_3$  to  $X_1$  along that same path, just as was the case with chains. In general, the flow of association is symmetric.

Chains and forks also share the same set of *independencies*. When we condition on  $X_2$  in both graphs, it blocks the flow of association from  $X_1$  to  $X_3$ . This is because of the local Markov assumption; each variable can locally depend on only its parents. So when we condition on  $X_2$  ( $X_3$ ’s parent in both graphs),  $X_3$  becomes independent of  $X_1$  (and vice versa).

We will refer to this independence as an instance of a *blocked path*. We illustrate these blocked paths in Figure 3.14 and Figure 3.15. Conditioning blocks the flow of association in chains and forks. Without conditioning, association is free to flow in chains and forks; we will refer to this as an *unblocked path*. However, the situation is completely different with immoralities, as we will see in the next section.

That’s all nice intuition, but what about the proof? We can prove that  $X_1 \perp\!\!\!\perp X_3 \mid X_2$  using just the local Markov assumption. We will do this by showing that  $P(x_1, x_3 \mid x_2) = P(x_1 \mid x_2) P(x_3 \mid x_2)$ . We’ll show the proof for chain graphs. It is usually useful to start with the Bayesian network factorization. For chains, we can factorize  $P(x_1, x_2, x_3)$  as follows:

$$P(x_1, x_2, x_3) = P(x_1) P(x_2|x_1) P(x_3|x_2) \quad (3.6)$$

Bayes’ rule tells us that  $P(x_1, x_3 \mid x_2) = \frac{P(x_1, x_2, x_3)}{P(x_2)}$ , so we have

$$P(x_1, x_3 \mid x_2) = \frac{P(x_1) P(x_2|x_1) P(x_3|x_2)}{P(x_2)} \quad (3.7)$$

Since we’re looking to end up with  $P(x_1 \mid x_2) P(x_3 \mid x_2)$  and we already have  $P(x_3|x_2)$ , we must turn the rest into  $P(x_1 \mid x_2)$ . We can do this by

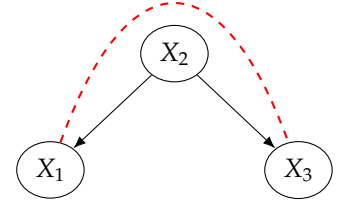


Figure 3.13: Fork with flow of association drawn as a dashed red arc.

[15]: Pearl et al. (2016), *Causal inference in statistics: A primer*

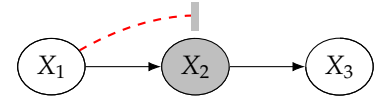


Figure 3.14: Chain with association blocked by conditioning on  $X_2$ .

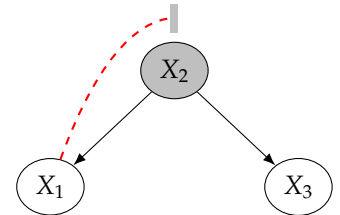


Figure 3.15: Fork with association blocked by conditioning on  $X_2$ .

another application of Bayes rule:

$$P(x_1, x_3 | x_2) = \frac{P(x_1, x_2)}{P(x_2)} P(x_3 | x_2) \quad (3.8)$$

$$= P(x_1 | x_2) P(x_3 | x_2) \quad (3.9)$$

With that, we've shown that  $X_1 \perp\!\!\!\perp X_3 | X_2$ . Try it yourself; prove the analog in forks.<sup>7</sup>

**Flow of Causation** The flow of association is symmetric, whereas the flow of causation is not. Under the causal edges assumption (Assumption 3.3), causation only flows in a single direction. Causation only flows along *directed* paths. Association flows along any path that does not contain an immorality.

<sup>7</sup> **Active reading exercise:** prove that  $X_1 \perp\!\!\!\perp X_3 | X_2$  for forks (Figure 3.15).

### 3.6 Colliders and their Descendants

Recall from Section 3.1 that we have an immorality when we have a child whose two parents do not have an edge connecting them (Figure 3.16). And in this graph structure, the child is known as a bastard. No, just kidding; it's called a *collider*.

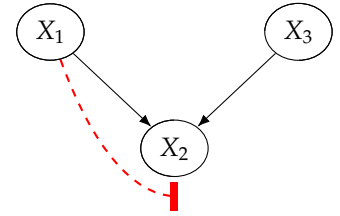
In contrast to chains and forks, in an immorality,  $X_1 \perp\!\!\!\perp X_3$ . Look at the graph structure and think about it a bit. Why would  $X_1$  and  $X_3$  be associated? One isn't the descendant of the other like in chains, and they don't share a common cause like in forks. Rather, we can think of  $X_1$  and  $X_3$  simply as unrelated events that happen, which happen to both contribute to some common effect ( $X_2$ ). To show this, we apply the Bayesian network factorization and marginalize out  $x_2$ :

$$P(x_1, x_3) = \sum_{x_2} P(x_1, x_2, x_3) \quad (3.10)$$

$$= \sum_{x_2} P(x_1) P(x_3) P(x_2 | x_1, x_3) \quad (3.11)$$

$$= P(x_1) P(x_3) \sum_{x_2} P(x_2 | x_1, x_3) \quad (3.12)$$

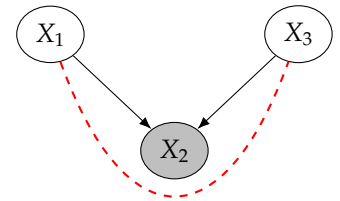
$$= P(x_1) P(x_3) \quad (3.13)$$



**Figure 3.16:** Immorality with **association** blocked by a collider.

We illustrate the independence of  $X_1$  and  $X_3$  in Figure 3.16 by showing that the association that we could have imagined as flowing along the path  $X_1 \rightarrow X_2 \leftarrow X_3$  is actually blocked at  $X_2$ . Because we have a collider on the path connecting  $X_1$  and  $X_3$ , association does not flow through that path. This is another example of a *blocked path*, but this time the path is not blocked by conditioning; the path is blocked by a collider.

**Good-Looking Men are Jerks** Oddly enough, when we condition on the collider  $X_2$ , its parents  $X_1$  and  $X_3$  become dependent (depicted in Figure 3.17). An example is the easiest way to see why this is the case. Imagine that you're out dating men, and you notice that most of the nice men you meet are not very good-looking, and most of the good-looking men you meet are jerks. It seems that you have to choose between looks and kindness. In other words, it seems like kindness and looks are negatively associated. However, what if I also told you that there is an important third variable here: availability (whether men are already in



**Figure 3.17:** Immorality with **association** unblocked by conditioning on the collider.



a relationship or not)? And what if told you that a man's availability is largely determined by their looks and kindness; if they are both good-looking and kind, then they are in a relationship. The available men are the remaining ones, the ones who are either not good-looking or not kind. You see an association between looks and kindness because you've conditioned on a collider (availability). You're only looking at men who are not in a relationship. You can see the causal structure of this example by taking Figure 3.17 and replacing  $X_1$  with "looks,"  $X_3$  with "kindness," and  $X_2$  with "availability."

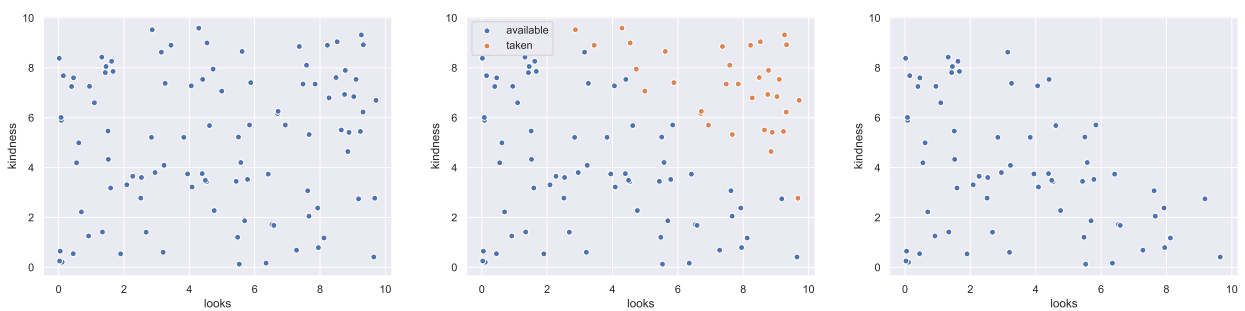
The above example naturally suggests that, when dating men, maybe you should consider not conditioning on  $X_2$  = "not in a relationship" and, instead, condition on  $X_2$  = "in a relationship." However, you could run into other variables  $X_4$  that introduce new immoralities there. Such moral questions are outside the scope of this book.

Returning to inside the scope of this book, we have that conditioning on a collider can turn a blocked path into an *unblocked path*. The parents  $X_1$  and  $X_3$  are not associated in the general population, but when we condition on their shared child  $X_2$  taking on a specific value, they become associated. Conditioning on the collider  $X_2$  allows associated to flow along the path  $X_1 \rightarrow X_2 \leftarrow X_3$ , despite the fact that it does not when we don't condition on  $X_2$ . We illustrate this in the move from Figure 3.16 to Figure 3.17.

We also illustrate this with a scatter plot in Figure 3.18. In Figure 3.18a, we plot the whole population, with kindness on the x-axis and looks on the y-axis. As you can see, the variables are not associated in the general population. However, if we remove the ones who are already in a relationship (the orange ones in Figure 3.18b), we are left with the clear negative association that we see in Figure 3.18c. This phenomenon is known as *Berkson's paradox*. The fact that we see this negative association simply because we are selecting a biased subset of the general population to look at is why this is sometimes referred to as *selection bias* [see, e.g., 7, Chapter 8].

**Active reading exercise:** Come up with your own example of an immorality and how conditioning on the collider induces association between its parents. Hint: think of rare events for  $X_1$  and  $X_3$  where, if either of them happens, some outcome  $X_2$  will happen.

[7]: Hernán and Robins (2020), *Causal Inference: What If*



(a) Looks and kindness data for the whole population. Looks and kindness are independent.

(b) Looks and kindness data grouped by whether the person is available or not. Within each group, there is a negative correlation.

(c) Looks and kindness data for only the available people. Now, there is a negative correlation.

**Figure 3.18:** Example data for the "good-looking men are jerks" example. Both looks and kindness are continuous values on a scale from 0 to 10.

**Numerical Example** All of the above has been to give you intuition about why conditioning on a collider induces association between its parents, but we have yet to give a concrete numerical example of this. We will give a simple one here. Consider the following *data generating*

process (DGP), where  $X_1$  and  $X_3$  are drawn independently from standard normal distributions and then used to compute  $X_2$ :

$$X_1 \sim N(0, 1), \quad X_3 \sim N(0, 1) \quad (3.14)$$

$$X_2 = X_1 + X_3 \quad (3.15)$$

We've already stated that  $X_1$  and  $X_3$  are independent, but to juxtapose the two calculations, let's compute their covariance:

$$\begin{aligned} \text{Cov}(X_1, X_3) &= \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_3 - \mathbb{E}[X_3])] \\ &= \mathbb{E}[X_1 X_3] && \text{(zero mean)} \\ &= \mathbb{E}[X_1] \mathbb{E}[X_3] && \text{(independent)} \\ &= 0 \end{aligned}$$

Now, let's compute their covariance, conditional on  $X_2$ :

$$\text{Cov}(X_1, X_3 \mid X_2 = x) = \mathbb{E}[X_1 X_3 \mid X_2 = x] \quad (3.16)$$

$$= \mathbb{E}[X_1(x - X_1)] \quad (3.17)$$

$$= x\mathbb{E}[X_1] - \mathbb{E}[X_1^2] \quad (3.18)$$

$$= -1 \quad (3.19)$$

Crucially, in Equation 3.17, we used Equation 3.15 to plug in for  $X_3$  in terms of  $X_1$  and  $X_2$  (conditioned to  $x$ ). This led to a second-order term, which led to the calculation giving a nonzero number, which means  $X_1$  and  $X_3$  are associated, conditional on  $X_2$ .

**Descendants of Colliders** Conditioning on descendants of a collider also induces association in between the parents of the collider. The intuition is that if we learn something about a collider's descendant, we usually also learn something about the collider itself because there is a direct causal path from the collider to its descendants, and we know that nodes in a chain are usually associated (see Section 3.5), assuming minimality (Assumption 3.2). In other words, a descendant of a collider can be thought of as a proxy for that collider, so conditioning on the descendant is similar to conditioning on the collider itself.

**Active reading exercise:** We have provided several techniques for how to think about colliders: high-level examples, numerical examples, and abstract reasoning. Use at least one of them to convince yourself that conditioning on a descendant of a collider can induce association between the collider's parents.

### 3.7 d-separation

Before we define d-separation, we'll codify what we mean by the concept of a "blocked path," which we've been discussing in the previous sections:

**Definition 3.3** (blocked path) *A path between nodes  $X$  and  $Y$  is blocked by a (potentially empty) conditioning set  $Z$  if either of the following is true:*

1. *Along the path, there is a chain  $\cdots \rightarrow W \rightarrow \cdots$  or a fork  $\cdots \leftarrow W \rightarrow \cdots$ , where  $W$  is conditioned on ( $W \in Z$ ).*
2. *There is a collider  $W$  on the path that is not conditioned on ( $W \notin Z$ ) and none of its descendants are conditioned on ( $\text{de}(W) \not\subseteq Z$ ).*

Then, an *unblocked path* is simply the complement; an unblocked path is a

path that is not blocked. The graphical intuition to have in mind is that association flows along unblocked paths, and association does not flow along blocked paths. If you don't have this intuition in mind, then it is probably worth it to reread the previous two sections, with the goal of gaining this intuition. Now, we are ready to introduce a very important concept: *d-separation*.

**Definition 3.4** (d-separation) *Two (sets of) nodes  $X$  and  $Y$  are d-separated by a set of nodes  $Z$  if all of the paths between (any node in)  $X$  and (any node in)  $Y$  are blocked by  $Z$  [16].*

[16]: Pearl (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*

If all the paths between two nodes  $X$  and  $Y$  are blocked, then we say that  $X$  and  $Y$  are *d-separated*. Similarly, if there exists at least one path between  $X$  and  $Y$  that is unblocked, then we say that  $X$  and  $Y$  are *d-connected*.

As we will see in Theorem 3.1, d-separation is such an important concept because it implies conditional independence. We will use the notation  $X \perp\!\!\!\perp_G Y \mid Z$  to denote that  $X$  and  $Y$  are d-separated in the graph  $G$  when conditioning on  $Z$ . Similarly, we will use the notation  $X \perp\!\!\!\perp_P Y \mid Z$  to denote that  $X$  and  $Y$  are independent in the distribution  $P$  when conditioning on  $Z$ .

**Theorem 3.1** *Given that  $P$  is Markov with respect to  $G$  (satisfies the local Markov assumption, Assumption 3.1), if  $X$  and  $Y$  are d-separated in  $G$  conditioned on  $Z$ , then  $X$  and  $Y$  are independent in  $P$  conditioned on  $Z$ . We can write this succinctly as follows:*

$$X \perp\!\!\!\perp_G Y \mid Z \implies X \perp\!\!\!\perp_P Y \mid Z \quad (3.20)$$

Because this is so important, we will give Equation 3.20 a name: the *global Markov assumption*. Theorem 3.1 tells us that the local Markov assumption implies the global Markov assumption.

**Markov assumption** Just as we built up the intuition that suggested that the local Markov assumption (Assumption 3.1) implies the Bayesian network factorization (Definition 3.1) and alerted you to the fact that the Bayesian network factorization also implies the local Markov assumption (the two are equivalent), it turns out that the global Markov assumption also implies the local Markov assumption. In other words, the local Markov assumption, global Markov assumption, and the Bayesian network factorization are all equivalent [see, e.g., 13, Chapter 3]. Therefore, we will use the slightly shortened phrase **Markov assumption** to refer to these concepts as a group, or we will simply write “ $P$  is Markov with respect to  $G$ ” to convey the same meaning.

[13]: Koller and Friedman (2009), *Probabilistic Graphical Models: Principles and Techniques*

**Active reading exercise:** To get some practice with d-separation, here are some questions about d-separation in Figure 3.19.

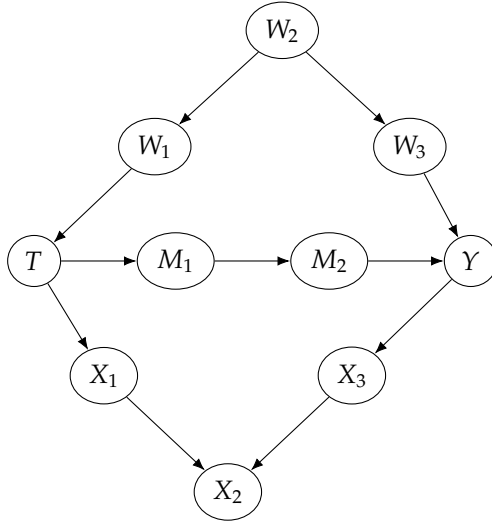
Questions about Figure 3.19a:

1. Are  $T$  and  $Y$  d-separated by the empty set?
2. Are  $T$  and  $Y$  d-separated by  $W_2$ ?
3. Are  $T$  and  $Y$  d-separated by  $\{W_2, M_1\}$ ?
4. Are  $T$  and  $Y$  d-separated by  $\{W_1, M_2\}$ ?
5. Are  $T$  and  $Y$  d-separated by  $\{W_1, M_2, X_2\}$ ?

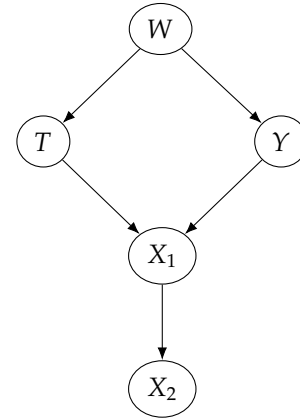
6. Are  $T$  and  $Y$  d-separated by  $\{W_1, M_2, X_2, X_3\}$ ?

Questions about Figure 3.19b:

1. Are  $T$  and  $Y$  d-separated by the empty set?
2. Are  $T$  and  $Y$  d-separated by  $W$ ?
3. Are  $T$  and  $Y$  d-separated by  $\{W, X_2\}$ ?



(a)



(b)

Figure 3.19: Graphs for d-separation exercise

### 3.8 Flow of Association and Causation

Now that we have covered the necessary preliminaries (chains, forks, colliders, and d-separation), it is worth emphasizing how association and causation flow in directed graphs. Association flows along all unblocked paths. In causal graphs, causation flows along directed paths. Recall from Section 1.3.2 that not only is association not causation, but causation is a sub-category of association. That's why association and causation both flow along directed paths.

We refer to the flow of association along directed paths as *causal association*. A common type of non-causal association that makes total association not causation is *confounding association*. In the graph in Figure 3.20, we depict the confounding association in red and the causal association in blue.

Regular Bayesian networks are purely statistical models, so we can only talk about the flow of association in Bayesian networks. Association still flows in exactly the same way in Bayesian networks as it does in causal graphs, though. In both, association flows along chains and forks, unless a node is conditioned on. And in both, a collider blocks the flow of association, unless it is conditioned on. Combining these building blocks, we get how association flows in general DAGs. We can tell if two nodes are not associated (no association flows between them) by whether or not they are d-separated.

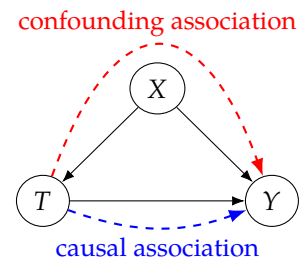


Figure 3.20: Causal graph depicting an example of how confounding association and causal association flow.

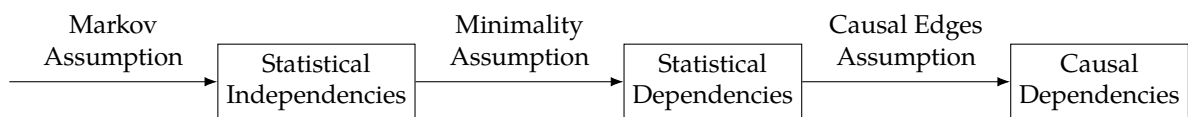
Causal graphs are special in that we additionally assume that the edges have causal meaning (causal edges assumption, Assumption 3.3). This assumption is what introduces causality into our models, and it makes one type of path take on a whole new meaning: directed paths. This assumption endows directed paths with the unique role of carrying causation along them. Additionally, this assumption is asymmetric; “ $X$  is a cause of  $Y$ ” is not the same as saying “ $Y$  is a cause of  $X$ .” This means that there is an important difference between association and causation: association is symmetric, whereas causation is asymmetric.

**d-separation Implies Association is Causation** Given that we have tools to measure association, how can we isolate causation? In other words, how can we ensure that the association we measure is causation, say, for measuring the causal effect of  $X$  on  $Y$ ? Well, we can do that by ensuring that there is no non-causal association flowing between  $X$  and  $Y$ . This is true if  $X$  and  $Y$  are d-separated in the augmented graph where we remove outgoing edges from  $X$ . This is because all of  $X$ ’s causal effect on  $Y$  would flow through its outgoing edges, so once those are removed, the only association that remains is purely non-causal association.

In Figure 3.21, we illustrate what each of the important assumptions gives us in terms of interpreting this flow of association. First, we have the (local/global) Markov assumption (Assumption 3.1). As we saw in Section 3.7, this assumption allows us to know which nodes are unassociated. In other words, the Markov assumption tells along which paths the association does *not* flow. When we slightly strengthen the Markov assumption to the minimality assumption (Assumption 3.2), we get which paths association *does* flow along (except in intransitive edges cases). When we further add in the causal edges assumption (Assumption 3.3), we get that causation flows along directed paths. Therefore, the following two<sup>8</sup> assumptions are essential for graphical causal models:

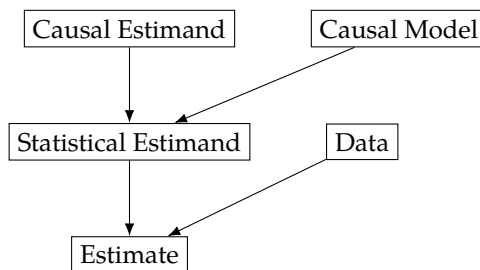
1. Markov Assumption (Assumption 3.1)
2. Causal Edges Assumption (Assumption 3.3)

<sup>8</sup> Recall that the first part of the minimality assumption is just the local Markov assumption and that the second part is contained in the causal edges assumption.



**Figure 3.21:** A flowchart that illustrates what kind of claims we can make about our data as we add each additional important assumption.

Causal models are essential for identification of causal quantities. When we presented the Identification-Estimation Flowchart (Figure 2.5) back in Section 2.4, we described identification as the process of moving from a causal estimand to a statistical estimand. However, to do that, we must have a causal model. We depict this more full version of the Identification-Estimation Flowchart in Figure 4.1.



**Figure 4.1:** The Identification-Estimation Flowchart – a flowchart that illustrates the process of moving from a target causal estimand to a corresponding estimate, through identification and estimation. In contrast to Figure 2.5, this version is augmented with a causal model and data.

The previous chapter gives graphical intuition for causal models, but it doesn’t explain how to identify causal quantities and formalize causal models. We will do that in this chapter.

## 4.1 The *do*-operator and Interventional Distributions

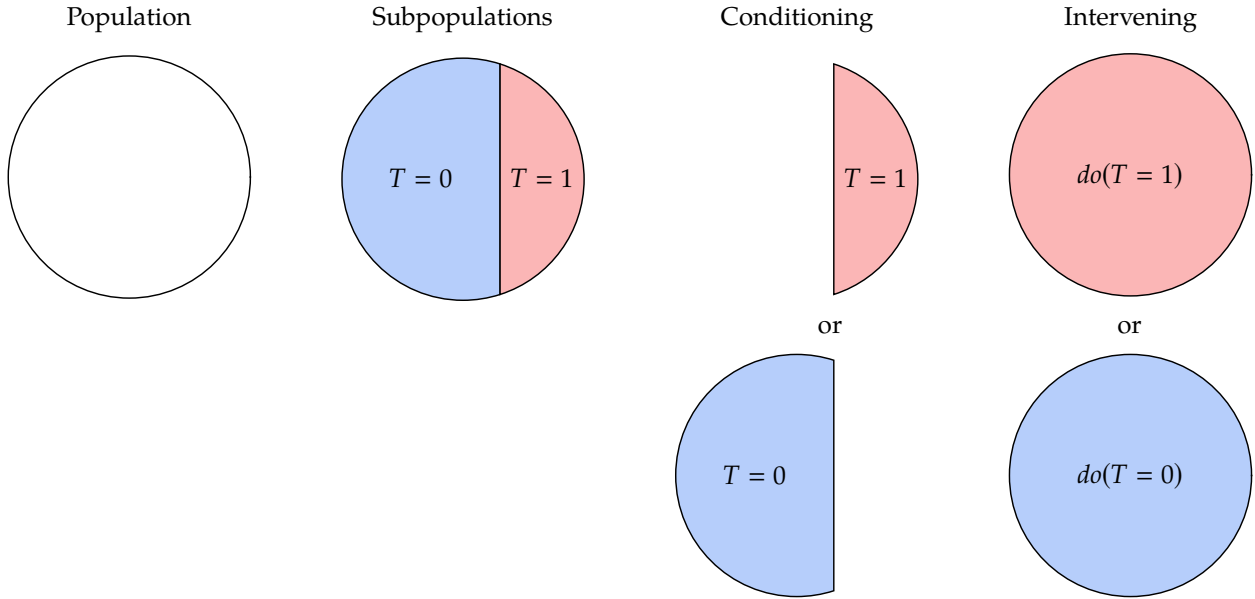
The first thing that we will introduce is a mathematical operator for intervention. In the regular notation for probability, we have conditioning, but that isn’t the same as intervening. Conditioning on  $T = t$  just means that we are restricting our focus to the subset of the population to those who received treatment  $t$ . In contrast, an intervention would be to take the whole population and give everyone treatment  $t$ . We illustrate this in Figure 4.2. We will denote intervention with the *do*-operator:  $do(T = t)$ . This is the notation commonly used in graphical causal models, and it has equivalents in potential outcomes notation. For example, we can write the distribution of the potential outcome  $Y(t)$  that we saw in Chapter 2 as follows:

$$P(Y(t) = y) \triangleq P(Y = y \mid do(T = t)) \triangleq P(y \mid do(t)) \quad (4.1)$$

Note that we shorten  $do(T = t)$  to just  $do(t)$  in the last option in Equation 4.1. We will use this shorthand throughout the book. We can similarly write the ATE (average treatment effect) when the treatment is binary as follows:

$$\mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)] \quad (4.2)$$

4.1 The <i>do</i> -operator and Interventional Distributions . .	32
4.2 The Main Assumption: Modularity . . . . .	34
4.3 Truncated Factorization . .	35
Example Application and Revisiting “Association is Not Causation” . . . . .	36
4.4 The Backdoor Adjustment Relation to Potential Outcomes . . . . .	37
4.5 Structural Causal Models (SCMs) . . . . .	39
Structural Equations . . . .	40
Interventions . . . . .	42
Collider Bias and Why to Not Condition on Descendants of Treatment . . . . .	43
4.6 Example Applications of the Backdoor Adjustment . . .	44
Association vs. Causation in a Toy Example . . . . .	44
A Complete Example with Estimation . . . . .	45
4.7 Assumptions Revisited . . .	47



**Figure 4.2:** Illustration of the difference between conditioning and intervening

We will often work with full distributions like  $P(Y \mid do(t))$ , rather than their means, as this is more general; if we characterize  $P(Y \mid do(t))$ , then we've characterized  $\mathbb{E}[Y \mid do(t)]$ . We will commonly refer to  $P(Y \mid do(T = t))$  and other expressions with the *do*-operator in them as *interventional distributions*.

Interventional distributions such as  $P(Y \mid do(T = t))$  are conceptually quite different from the *observational distribution*  $P(Y)$ . Observational distributions such as  $P(Y)$  or  $P(Y, T, X)$  do not have the *do*-operator in them. Because they don't have the *do*-operator, we can observe data from them without needing to carry out any experiment. This is why we call data from  $P(Y, T, X)$  *observational data*. If we can reduce an expression  $Q$  with *do* in it (an interventional expression) to one without *do* in it (an observational expression), then  $Q$  is said to be *identifiable*. An expression with a *do* in it is fundamentally different from an expression without a *do* in it, despite the fact that in *do*-notation, *do* appears after a regular conditioning bar. As we discussed in Section 2.4, we will refer to an estimand as a *causal estimand* when it contains a *do*-operator, and we refer to an estimand as a *statistical estimand* when it doesn't contain a *do*-operator.

Whenever,  $do(t)$  appears after the conditioning bar, it means that everything in that expression is in the *post-intervention* world where the intervention  $do(t)$  occurs. For example,  $\mathbb{E}[Y \mid do(t), Z = z]$  refers to the expected outcome in the subpopulation where  $Z = z$  after the whole subpopulation has taken treatment  $t$ . In contrast,  $\mathbb{E}[Y \mid Z = z]$  simply refers to the expected value in the (*pre-intervention*) population where individuals take whatever treatment they would normally take ( $T$ ). This distinction will become important when we get to counterfactuals in Chapter 8.



## 4.2 The Main Assumption: Modularity

Before we can describe a very important assumption, we must specify what a *causal mechanism* is. There are a few different ways to think about causal mechanisms. In this section, we will refer to the causal mechanism that generates  $X_i$  as the conditional distribution of  $X_i$  given all of its causes:  $P(x_i \mid \text{pa}_i)$ . As we show graphically in Figure 4.3, the causal mechanism that generates  $X_i$  is all of  $X_i$ 's parents and their edges that go into  $X_i$ . We will give a slightly more specific description of what a causal mechanism is in Section 4.5.1, but these suffice for now.

In order to get many causal identification results, the main assumption we will make is that interventions are local. More specifically, we will assume that intervening on a variable  $X_i$  only changes the causal mechanism for  $X_i$ ; it does not change the causal mechanisms that generate any other variables. In this sense, the causal mechanisms are *modular*. Other names that are used for the modularity property are *independent mechanisms*, *autonomy*, and *invariance*. We will now state this assumption more formally.

**Assumption 4.1** (Modularity / Independent Mechanisms / Invariance)

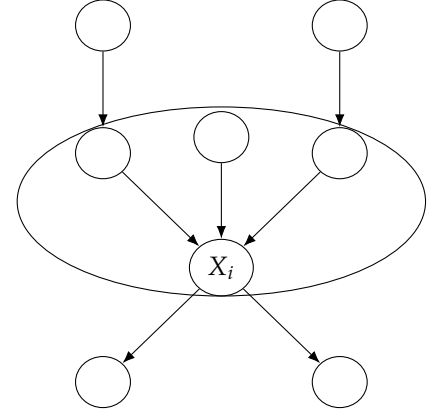
If we intervene on a set of nodes  $S \subseteq [n]$ ,<sup>1</sup> setting them to constants, then for all  $i$ , we have the following:

1. If  $i \notin S$ , then  $P(x_i \mid \text{pa}_i)$  remains unchanged.
2. If  $i \in S$ , then  $P(x_i \mid \text{pa}_i) = 1$  if  $x_i$  is the value that  $X_i$  was set to by the intervention; otherwise,  $P(x_i \mid \text{pa}_i) = 0$ .

In the second part of the above assumption, we could have alternatively said  $P(x_i \mid \text{pa}_i) = 1$  if  $x_i$  is *consistent with the intervention*<sup>2</sup> and 0 otherwise. More explicitly, we will say (in the future) that if  $i \in S$ , a value  $x_i$  is consistent with the intervention if  $x_i$  equals the value that  $X_i$  was set to in the intervention.

The modularity assumption is what allows us to encode many different interventional distributions all in a single graph. For example, it could be the case that  $P(Y)$ ,  $P(Y \mid \text{do}(T = t))$ ,  $P(Y \mid \text{do}(T = t'))$ , and  $P(Y \mid \text{do}(T_2 = t_2))$  are all completely different distributions that share almost nothing. If this were the case, then each of these distributions would need their own graph. However, by assuming modularity, we can encode them all with the same graph that we use to encode the joint  $P(Y, T, T_2, \dots)$ , and we can know that all of the factors (except ones that are intervened on) are shared across these graphs.

The causal graph for interventional distributions is simply the same graph that was used for the observational joint distribution, but with all of the edges to the intervened node(s) removed. This is because the probability for the intervened factor has been set to 1, so we can just ignore that factor (this is the focus of the next section). Another way to see that the intervened node has no causal parents is that the intervened node is set to a constant value, so it no longer depends on any of the variables it depends on in the observational setting (its parents). The graph with edges removed is known as the *manipulated graph*.



**Figure 4.3:** A causal graph with the causal mechanism that generates  $X_i$  depicted inside an ellipse.

<sup>1</sup> We use  $[n]$  to refer to the set  $\{1, 2, \dots, n\}$ .

<sup>2</sup> Yes, the word “consistent” is extremely overloaded.



For example, consider the causal graph for an observational distribution in Figure 4.4a. Both  $P(Y \mid \text{do}(T = t))$  and  $P(Y \mid \text{do}(T = t'))$  correspond to the causal graph in Figure 4.4b, where the incoming edge to  $T$  has been removed. Similarly,  $P(Y \mid \text{do}(T_2 = t_2))$  corresponds to the graph in Figure 4.4c, where the incoming edges to  $T_2$  have been removed. Although it is not expressed in the graphs (which only express conditional independencies and causal relations), under the modularity assumption,  $P(Y)$ ,  $P(Y \mid T = t')$ , and  $P(Y \mid \text{do}(T_2 = t_2))$  all shared the exact same factors (that are not intervened on).

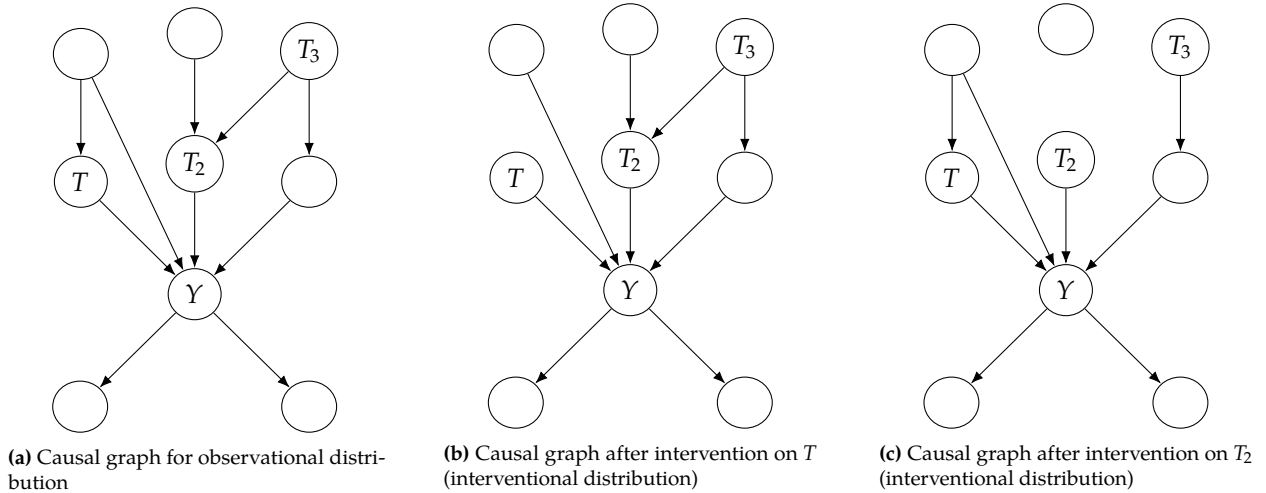


Figure 4.4: Intervention as edge deletion in causal graphs

What would it mean for the modularity assumption to be violated? Imagine that you intervene on  $X_i$ , and this causes the mechanism that generates a different node  $X_j$  to change; an intervention on  $X_i$  changes  $P(x_j \mid \text{pa}_j)$ , where  $j \neq i$ . In other words, the intervention is not local to the node you intervene on; causal mechanisms are not invariant to when you change other causal mechanisms; the causal mechanisms are not modular.

This assumption is so important that Judea Pearl refers to a closely related version (which we will see in Section 4.5.2) as **The Law of Counterfactuals (and Interventions)**, one of two key principles from which all other causal results follow.<sup>3</sup> Incidentally, taking the modularity assumption (Assumption 4.1) and the Markov assumption (the other key principle) together gives us *causal Bayesian networks*. We'll now move to one of the important results that follow from these assumptions.

<sup>3</sup> The other key principle is the global Markov assumption (Theorem 3.1), which is the assumption that d-separation implies conditional independence.

### 4.3 Truncated Factorization

Recall the Bayesian network factorization (Definition 3.1), which tells us that if  $P$  is Markov with respect to a graph  $G$ , then  $P$  factorizes as follows:

$$P(x_1, \dots, x_n) = \prod_i P(x_i \mid \text{pa}_i) \quad (4.3)$$

where  $\text{pa}_i$  denotes the parents of  $X_i$  in  $G$ . Now, if we intervene on some set of nodes  $S$  and assume modularity (Assumption 4.1), then all of the factors should remain the same except the factors for  $X_i \in S$ ; those factors

should change to 1 (for values consistent with the intervention) because those variables have been intervened on. This is how we get the *truncated factorization*.

**Proposition 4.1** (Truncated Factorization) *We assume that  $P$  and  $G$  satisfy the Markov assumption and modularity. Given, a set of intervention nodes  $S$ , if  $x$  is consistent with the intervention, then*

$$P(x_1, \dots, x_n \mid do(S = s)) = \prod_{i \notin S} P(x_i \mid pa_i). \quad (4.4)$$

*Otherwise,  $P(x_1, \dots, x_n \mid do(S = s)) = 0$ .*

The key thing that changed when we moved from the regular factorization in Equation 4.3 to the truncated factorization in Equation 4.4 is that the latter's product is only over  $i \notin S$  rather than all  $i$ . In other words, the factors for  $i \in S$  have been truncated.

### 4.3.1 Example Application and Revisiting “Association is Not Causation”

To see the power that the truncated factorization gives us, let's apply it to identify the causal effect of treatment on outcome in a simple graph. Specifically, we will identify the causal quantity  $P(y \mid do(t))$ . In this example, the distribution  $P$  is Markov with respect to the graph in Figure 4.5. The Bayesian network factorization (from the Markov assumption), gives us the following:

$$P(y, t, x) = P(x) P(t \mid x) P(y \mid t, x) \quad (4.5)$$

When we intervene on the treatment, the truncated factorization (from adding the modularity assumption) gives us the following:

$$P(y, x \mid do(t)) = P(x) P(y \mid t, x) \quad (4.6)$$

Then, we simply need to marginalize out  $x$  to get what we want:

$$P(y \mid do(t)) = \sum_x P(y \mid t, x) P(x) \quad (4.7)$$

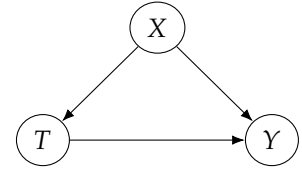
We assumed  $X$  is discrete when we summed over its values, but we can simply replace the sum with an integral if  $X$  is continuous. Throughout this book, that will be the case, so we usually won't point it out.

If we massage Equation 4.7 a bit, we can clearly see how association is not causation. The purely associational counterpart of  $P(y \mid do(t))$  is  $P(y \mid t)$ . If the  $P(x)$  in Equation 4.7 were  $P(x \mid t)$ , then we would actually recover  $P(y \mid t)$ . We briefly show this:

$$\sum_x P(y \mid t, x) P(x \mid t) = \sum_x P(y, x \mid t) \quad (4.8)$$

$$= P(y \mid t) \quad (4.9)$$

This gives some concreteness to the difference between association and causation. In this example (which is representative of a broader



**Figure 4.5:** Simple causal structure where  $X$  confounds the effect of  $T$  on  $Y$  and where  $X$  is the only confounder.

phenomenon), the difference between  $P(y \mid do(t))$  and  $P(y \mid t)$  is the difference between  $P(x)$  and  $P(x \mid t)$ .

To round this example out, say  $T$  is a binary random variable, and we want to compute the ATE.  $P(y \mid do(T = 1))$  is the distribution for  $Y(1)$ , so we can just take the expectation to get  $\mathbb{E}[Y(1)]$ . Similarly, we can do the same thing with  $Y(0)$ . Then, we can write the ATE as follows:

$$\mathbb{E}[Y(1) - Y(0)] = \sum_y y P(y \mid do(T = 1)) - \sum_y y P(y \mid do(T = 0)) \quad (4.10)$$

If we then plug in Equation 4.7 for  $P(y \mid do(T = 1))$  and  $P(y \mid do(T = 0))$ , we have a fully identified ATE. Given the simple graph in Figure 4.5, we have shown how we can use the truncated factorization to identify causal effects in Equations 4.5 to 4.7. We will now generalize this identification process to a more general formula.

## 4.4 The Backdoor Adjustment

Recall from Chapter 3 that causal association flows from  $T$  to  $Y$  along directed paths and that non-causal association flows along any other paths from  $T$  to  $Y$  that aren't blocked by either 1) a non-collider that is conditioned on or 2) a collider that isn't conditioned on. These non-directed unblocked paths from  $T$  to  $Y$  are known as *backdoor paths* because they have an edge that goes in the "backdoor" of the  $T$  node. And it turns out that if we can block these paths by conditioning, we can identify causal quantities like  $P(Y \mid do(t))$ .<sup>4</sup>

This is precisely what we did in the previous section. We blocked the backdoor path  $T \leftarrow X \rightarrow Y$  in Figure 4.5 simply by conditioning on  $X$  and marginalizing it out (Equation 4.7). In this section, we will generalize Equation 4.7 to arbitrary DAGs. But before we do that, let's graphically consider why the quantity  $P(y \mid do(t))$  is purely causal.

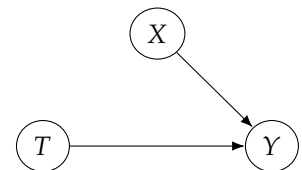
As we discussed in Section 4.2, the graph for the interventional distribution  $P(Y \mid do(t))$  is the same as the graph for the observational distribution  $P(Y, T, X)$ , but with the incoming edges to  $T$  removed. For example, if we take the graph from Figure 4.5 and intervene on  $T$ , then we get the manipulated graph in Figure 4.6. In this manipulated graph, there cannot be any backdoor paths because no edges are going into the backdoor of  $T$ . Therefore, all of the association that flows from  $T$  to  $Y$  in the manipulated graph is purely causal.

With that digression aside, let's prove that we can identify  $P(y \mid do(t))$ . We want to turn the causal estimand  $P(y \mid do(t))$  into a statistical estimand (only relies on the observational distribution). We'll start with assuming we have a set of variables  $W$  that satisfy the backdoor criterion:

**Definition 4.1** (Backdoor Criterion) *A set of variables  $W$  satisfies the backdoor criterion relative to  $T$  and  $Y$  if the following are true:*

1.  $W$  blocks all backdoor paths from  $T$  to  $Y$ .
2.  $W$  does not contain any descendants of  $T$ .

<sup>4</sup> As we mentioned in Section 3.8, blocking all backdoor paths is equivalent to having d-separation in the graph where edges going out of  $T$  are removed. This is because these are the only edges that causation flows along, so once they are removed, all that remains is non-causation association.



**Figure 4.6:** Manipulated graph that results from intervening on  $T$ , when the original graph is

<sup>5</sup> **Active reading exercise:** In a general DAG, which set of nodes related to  $T$  will always be a sufficient adjustment set? Which set of nodes related to  $Y$  will always be a sufficient adjustment set?

Satisfying the backdoor criterion makes  $W$  a *sufficient adjustment set*.<sup>5</sup> We saw an example of  $X$  as a sufficient adjustment set in Section 4.3.1. Because there was only a single backdoor path in Section 4.3.1, a single node ( $X$ ) was enough to block all backdoor paths, but, in general, there can be multiple backdoor paths.

To introduce  $W$  into the proof, we'll use the usual trick of conditioning on variables and marginalizing them out:

$$P(y \mid do(t)) = \sum_w P(y \mid do(t), w) P(w \mid do(t)) \quad (4.11)$$

Given that  $W$  satisfies the backdoor criterion, we can write the following:

$$\sum_w P(y \mid do(t), w) P(w \mid do(t)) = \sum_w P(y \mid t, w) P(w \mid do(t)) \quad (4.12)$$

This follows from the modularity assumption (Assumption 4.1). If  $W$  is all of the parents for  $Y$  (other than  $T$ ), it should be clear that the modularity assumption immediately implies  $P(y \mid do(t), w) = P(y \mid t, w)$ . If  $W$  isn't the parents of  $Y$  but still blocks all backdoor paths another way, then this equality is still true but requires using the graphical knowledge we built up in Chapter 3.

In the manipulated graph (for  $P(y \mid do(t), w)$ ), all of the  $T$ - $Y$  association flows along the directed path(s) from  $T$  to  $Y$ , since there cannot be any backdoor paths because  $T$  has no incoming edges. Similarly, in the regular graph (for  $P(y \mid t, w)$ ), all of the  $T$ - $Y$  association flows along the directed path(s) from  $T$  to  $Y$ . This is because, even though there exist backdoor paths, the association that would flow along them is blocked by  $W$ , leaving association to only flow along directed paths. In both cases, association flows along the exact same directed paths, which correspond to the exact same conditional distributions (by the modularity assumption).

Although we've justified Equation 4.12, there is still a *do* in the expression:  $P(w \mid do(t))$ . However,  $P(w \mid do(t)) = P(w)$ . To see this, consider how  $T$  might influence  $W$  in the manipulated graph. It can't be through any path that has an edge into  $T$  because  $T$  doesn't have any incoming edges in the manipulated graph. It can't be through any path that has an edge going out of  $T$  because such a path would have to have a collider, that isn't conditioned on, on the path. We know any such colliders are not conditioned on because we have assumed that  $W$  does not contain descendants of  $T$  (second part of the backdoor criterion).<sup>6</sup> Therefore, we can write the final step:

$$\sum_w P(y \mid t, w) P(w \mid do(t)) = \sum_w P(y \mid t, w) P(w) \quad (4.13)$$

This is known as the *backdoor adjustment*.

**Theorem 4.2** (Backdoor Adjustment) *Given the modularity assumption (Assumption 4.1), that  $W$  satisfies the backdoor criterion (Definition 4.1), and*

<sup>6</sup> We will come back to what goes wrong if we condition on descendants of  $T$  in Section 4.5.3, after we cover some important concepts that we need before we can fully explain that.

positivity (Assumption 2.3), we can identify the causal effect of  $T$  on  $Y$ :

$$P(y \mid do(t)) = \sum_w P(y \mid t, w) P(w)$$

Here's a concise recap of the proof (Equations 4.11 to 4.13) without all of the explanation/justification:

*Proof.*

$$P(y \mid do(t)) = \sum_w P(y \mid do(t), w) P(w \mid do(t)) \quad (4.14)$$

$$= \sum_w P(y \mid t, w) P(w \mid do(t)) \quad (4.15)$$

$$= \sum_w P(y \mid t, w) P(w) \quad (4.16)$$

□

**Relation to d-separation** We can use the backdoor adjustment if  $W$  d-separates  $T$  from  $Y$  in the manipulated graph. Recall from Section 3.8 that we mentioned that we would be able to isolate the causal association if  $T$  is d-separated from  $Y$  in the manipulated graph. "Isolation of the causal association" is identification. We can also isolate the causal association if  $Y$  is d-separated from  $T$  in the manipulated graph, *conditional on*  $W$ . This is what the first part of the backdoor criterion is about and what we've codified in the backdoor adjustment.

#### 4.4.1 Relation to Potential Outcomes

Hmm, the backdoor adjustment (Theorem 4.2) looks quite similar to the adjustment formula (Theorem 2.1) that we saw back in the potential outcomes chapter:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_W [\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W]] \quad (4.17)$$

We can derive this from the more general backdoor adjustment in a few steps. First, we take an expectation over  $Y$ :

$$\mathbb{E}[Y \mid do(t)] = \sum_w \mathbb{E}[Y \mid t, w] P(w) \quad (4.18)$$

Then, we notice that the sum over  $w$  and  $P(w)$  is an expectation (for discrete  $x$ , but just replace with an integral if not):

$$\mathbb{E}[Y \mid do(t)] = \mathbb{E}_W \mathbb{E}[Y \mid t, W] \quad (4.19)$$

And finally, we look at the difference between  $T = 1$  and  $T = 0$ :

$$\mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)] = \mathbb{E}_W [\mathbb{E}[Y \mid T = 1, W] - \mathbb{E}[Y \mid T = 0, W]] \quad (4.20)$$

Since the *do*-notation  $\mathbb{E}[Y \mid do(t)]$  is just another notation for the potential outcomes  $\mathbb{E}[Y(t)]$ , we are done! If you remember, one of the main assumptions we needed to get Equation 4.17 (Theorem 2.1) was conditional

exchangeability (Assumption 2.2), which we repeat below:

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \quad (4.21)$$

However, we had no way of knowing how to choose  $W$  or knowing that that  $W$  actually gives us conditional exchangeability. Well, using graphical causal models, we know how to choose a valid  $W$ : we simply choose  $W$  so that it satisfies the backdoor criterion. Then, under the assumptions encoded in the causal graph, conditional exchangeability provably holds; the causal effect is provably identifiable.

## 4.5 Structural Causal Models (SCMs)

Graphical causal models such as causal Bayesian networks give us powerful ways to encode statistical and causal assumptions, but we have yet to explain exactly what an intervention is or exactly what a causal mechanism is. Moving from causal Bayesian networks to full structural causal models will give us this additional clarity along with the power to compute counterfactuals.

### 4.5.1 Structural Equations

As Judea Pearl often says, the equals sign in mathematics does not convey any causal information. Saying  $A = B$  is the same as saying  $B = A$ . Equality is symmetric. However, in order to talk about causation, we must have something asymmetric. We need to be able to write that  $A$  is a cause of  $B$ , meaning that changing  $A$  results in changes in  $B$ , but changing  $B$  does not result in changes in  $A$ . This is what we get when we write the following *structural equation*:

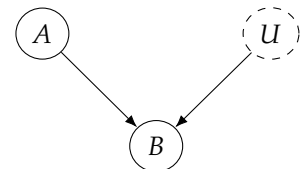
$$B := f(A), \quad (4.22)$$

where  $f$  is some function that maps  $A$  to  $B$ . While the usual “=” symbol does not give us causal information, this new “:=” symbol does. This is a major difference that we see when moving from statistical models to causal models. Now, we have the asymmetry we need to describe causal relations. However, the mapping between  $A$  and  $B$  is deterministic. Ideally, we’d like to allow it to be probabilistic, which allows room for some unknown causes of  $B$  that factor into this mapping. Then, we can write the following:

$$B := f(A, U), \quad (4.23)$$

where  $U$  is some unobserved random variable. We depict this in Figure 4.7, where  $U$  is drawn inside a dashed node to indicate that it is unobserved. The unobserved  $U$  is analogous to the randomness that we would see by sampling units (individuals); it denotes all the relevant (noisy) background conditions that determine  $B$ . More concretely, there are analogs to every part of the potential outcome  $Y_i(t)$ :  $B$  is the analog of  $Y$ ,  $A = a$  is the analog of  $T = t$ , and  $U$  is the analog of  $i$ .

The functional form of  $f$  does not need to be specified, and when left unspecified, we are in the *nonparametric* regime because we aren’t making any assumptions about parametric form. Although the mapping



**Figure 4.7:** Graph for simple structural equation. The dashed node  $U$  means that  $U$  is unobserved.

is deterministic, because it takes a random variable  $U$  (a “noise” or “background conditions” variable) as input, it can represent any stochastic mapping, so structural equations generalize the probabilistic factors  $P(x_i \mid \text{pa}_i)$  that we’ve been using throughout this chapter. Therefore, all the results that we’ve seen such as the truncated factorization and the backdoor adjustment still hold when we introduce structural equations.

**Cause and Causal Mechanism Revisited** We have now come to the more precise definitions of what a cause is (Definition 3.2) and what a causal mechanism is (introduced in Section 4.2). A causal mechanism that generates a variable is the structural equation that corresponds to that variable. For example, the causal mechanism for  $B$  is Equation 4.23. Similarly,  $X$  is a *direct cause* of  $Y$  if  $X$  appears on the right-hand side of the structural equation for  $Y$ . We say that  $X$  is a *cause* of  $Y$  if  $X$  is a direct cause of  $Y$ <sup>7</sup> or if  $X$  is a direct cause of  $Y$ .

We only showed a single structural equation in Equation 4.23, but there can be a large collection of structural equations in a single model, which we will commonly label  $M$ . For example, we write structural equations for Figure 4.8 below:

$$\begin{aligned} B &:= f_B(A, U_B) \\ M : \quad C &:= f_C(A, B, U_C) \\ D &:= f_D(A, C, U_D) \end{aligned} \quad (4.24)$$

In causal graphs, the noise variables are often implicit, rather than explicitly drawn. The variables that we write structural equations for are known as *endogenous* variables. These are the variables whose causal mechanisms we are modeling – the variables that have parents in the causal graph. In contrast, *exogenous* variables are variables who do not have any parents in the causal graph; these variables are external to our causal model in the sense that we choose not to model their causes. For example, in the causal model described by Figure 4.8 and Equation 4.24, the endogenous variables are  $\{B, C, D\}$ . And the exogenous variables are  $\{A, U_B, U_C, U_D\}$ .

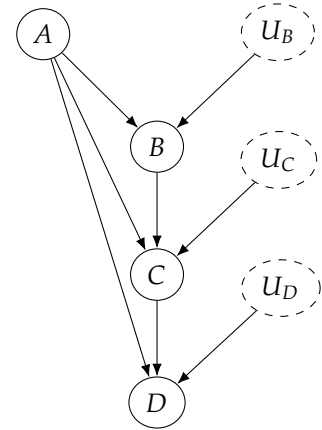
**Definition 4.2** (Structural Causal Model (SCM)) *A structural causal model is a tuple of the following sets:*

1. A set of endogenous variables  $V$
2. A set of exogenous variables  $U$
3. A set of functions  $f$ , one to generate each endogenous variable as a function of other variables

For example,  $M$ , the set of three equations above in Equation 4.24 constitutes an SCM with corresponding causal graph in Figure 4.8. Every SCM implies an associated causal graph: for each structural equation, draw an edge from every variable on the right-hand side to the variable on the left-hand side.

If the causal graph contains no cycles (is a DAG) and the noise variables  $U$  are independent, then the causal model is *Markovian*; the distribution  $P$  is Markov with respect to the causal graph. If the causal graph doesn’t contain cycles but the noise terms are dependent, then the model is *semi-Markovian*. For example, if there is unobserved confounding, the model

<sup>7</sup> Trust me; the recursion ends. The base case was specified.



**Figure 4.8:** Graph for the structural equations in Equation 4.24.



is semi-Markovian. Finally, the graphs of *non-Markovian* models contain cycles. We will largely be considering Markovian and semi-Markovian models in this book.

### 4.5.2 Interventions

Interventions in SCMs are remarkably simple. The intervention  $do(T = t)$  simply corresponds to replacing the structural equation for  $T$  with  $T := t$ . For example, consider the following causal model  $M$  with corresponding causal graph in Figure 4.9:

$$M : \begin{aligned} T &:= f_T(X, U_T) \\ Y &:= f_Y(X, T, U_Y) \end{aligned} \quad (4.25)$$

If we then intervene on  $T$  to set it to  $t$ , we get the *interventional* SCM  $M_t$  below and corresponding manipulated graph in Figure 4.10.

$$M_t : \begin{aligned} T &:= t \\ Y &:= f_Y(X, T, U_Y) \end{aligned} \quad (4.26)$$

The fact that  $do(T = t)$  only changes the equation for  $T$  and no other variables is a consequence of the modularity assumption; these causal mechanisms (structural equations) are modular. Assumption 4.1 states the modularity assumption in the context of causal Bayesian networks, but we need a slightly different translation of this assumption for SCMs.

**Assumption 4.2** (Modularity Assumption for SCMs) *Consider an SCM  $M$  and an interventional SCM  $M_t$  that we get by performing the intervention  $do(T = t)$ . The modularity assumption states that  $M$  and  $M_t$  share all of their structural equations except the structural equation for  $T$ , which is  $T := t$  in  $M_t$ .*

In other words, the intervention  $do(T = t)$  is localized to  $T$ . None of the other structural equations change because they are modular; the causal mechanisms are independent. The modularity assumption for SCMs is what gives us what Pearl calls the **The Law of Counterfactuals**, which we briefly mentioned at the end of Section 4.2, after we defined the modularity assumption for causal Bayesian networks. But before we can get to that, we must first introduce a bit more notation.

In the causal inference literature, there are many different ways of writing the unit-level potential outcome. In Chapter 2, we used  $Y_i(t)$ . However, there are other ways such as  $Y_i^t$  or even  $Y_t(u)$ . For example, in his prominent potential outcomes paper, Holland [5] uses the  $Y_t(u)$  notation. In this notation,  $u$  is the analog of  $i$ , just as we mentioned is the case for the  $U$  in Equation 4.23 and the paragraph that followed it. This is the notation that Pearl uses for SCMs as well [see, e.g., 17, Definition 4]. So  $Y_t(u)$  denotes the outcome that unit  $u$  would observe if they take treatment  $t$ , given that the SCM is  $M$ . Similarly, we define  $Y_{M_t}(u)$  as the outcome that unit  $u$  would observe if they take treatment  $t$ , given that the SCM is  $M_t$  (remember that  $M_t$  is the same SCM as  $M$  but with the structural equation for  $T$  changed to  $T := t$ ). Now, we are ready to

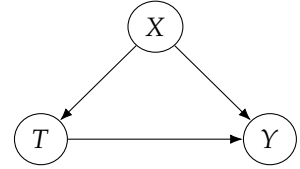


Figure 4.9: Basic causal graph

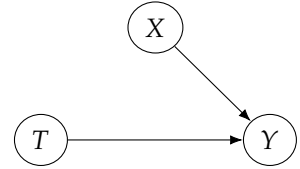


Figure 4.10: Basic causal with the incoming edges to  $T$  removed, due to the intervention  $do(T = t)$ .

[5]: Holland (1986), ‘Statistics and Causal Inference’

[17]: Pearl (2009), ‘Causal inference in statistics: An overview’

present one of Pearl’s two key principles from which all other causal results follow:<sup>8</sup>

**Definition 4.3** (The Law of Counterfactuals (and Interventions))

$$Y_t(u) = Y_{M_t}(u) \quad (4.27)$$

This is called “The Law of Counterfactuals” because it gives us information about counterfactuals. Given an SCM with enough details about it specified, we can actually compute counterfactuals. This is a big deal because this is exactly what the fundamental problem of causal inference (Section 2.2) told us we cannot do. We won’t say more about how to do this until we get to the dedicated chapter for counterfactuals: Chapter 8.

### 4.5.3 Collider Bias and Why to Not Condition on Descendants of Treatment

In defining the backdoor criterion (Definition 4.1) for the backdoor adjustment (Theorem 4.2), not only did we specify that the adjustment set  $W$  blocks all backdoor paths, but we also specified that  $W$  does not contain any descendants of  $T$ . Why? There are two categories of things that could go wrong if we condition on descendants of  $T$ :

1. We block the flow of causation from  $T$  to  $Y$ .
2. We induce non-causal association between  $T$  and  $Y$ .

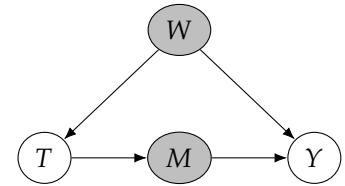
As we’ll see, it is fairly intuitive why we want to avoid the first category. The second category is a bit more complex, and we’ll break it up into two different parts, each with their own paragraph. This more complex part is actually why we delayed this explanation to after we introduced SCMs, rather than back when we introduced the backdoor criterion/adjustment in Section 4.4.

If we condition on a node that is on a directed path from  $T$  to  $Y$ , then we block the flow of causation along that causal path. We will refer to a node on a directed path from  $T$  to  $Y$  as a *mediator*, as it mediates the effect of  $T$  on  $Y$ . For example, in Figure 4.11, all of the causal flow is blocked by  $M$ . This means that we will measure zero association between  $T$  and  $Y$  (given that  $W$  blocks all backdoor paths). In Figure 4.12, only a portion of the causal flow is blocked by  $M$ . This is because causation can still flow along the  $T \rightarrow Y$  edge. In this case, we will get a non-zero estimate of the causal effect, but it will still be biased, due to the causal flow that  $M$  blocks.

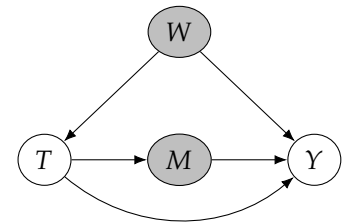
If we condition on a descendant of  $T$  that isn’t a mediator, it could unblock a path from  $T$  to  $Y$  that was blocked by a collider. For example, this is the case with conditioning on  $Z$  in Figure 4.13. This induces non-causal association between  $T$  and  $Y$ , which biases the estimate of the causal effect. Consider the following general kind of path, where  $\rightarrow \dots \rightarrow$  denotes a directed path:  $T \rightarrow \dots \rightarrow Z \leftarrow \dots \leftarrow Y$ . Conditioning on  $Z$ , or any descendant of  $Z$  in a path like this, will induce *collider bias*. That is, the causal effect estimate will be biased by the non-causal association that we induce when we condition on  $Z$  or any of its descendants (see Section 3.6).

<sup>8</sup> **Active reading exercise:** Can you recall which was the other key principle/assumption?

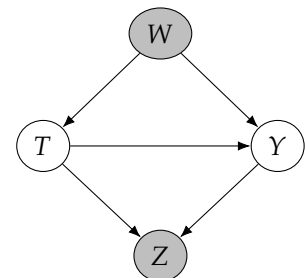
**Active reading exercise:** Take what you now know about structural equations, and relate it to other parts of this chapter. For example, how do interventions in structural equations relate to the modularity assumption? How does the modularity assumption for SCMs (Assumption 4.2) relate to the modularity assumption in causal Bayesian networks (Assumption 4.1)? Does this modularity assumption for SCMs still give us the backdoor adjustment?



**Figure 4.11:** Causal graph where all causation is blocked by conditioning on  $M$ .



**Figure 4.12:** Causal graph where part of the causation is blocked by conditioning on  $M$ .



**Figure 4.13:** Causal graph where conditioning on the collider  $Z$  induces bias.

What about conditioning on  $Z$  in Figure 4.14? Would that induce bias? Recall that graphs are frequently drawn without explicitly drawing the noise variables. If we *magnify* part of the graph, making  $M$ 's noise variable explicit, we get Figure 4.15. Now, we see that  $T \rightarrow M \leftarrow U_M$  forms an immorality. Therefore, conditioning on  $Z$  induces an association between  $T$  and  $U_M$ . This induced non-causal association is another form of collider bias. You might find this unsatisfying because  $Y$  is not one of the immoral parents here; rather  $T$  and  $U_M$  are the ones living the immoral lifestyle. So why would this change the association between  $T$  and  $Y$ ? One way to get the intuition for this is that there is now induced association flowing between  $T$  and  $U_M$  through the edge  $T \rightarrow M$ , which is also an edge that causal association is flowing along. You can think of these two types of association getting tangled up along the  $T \rightarrow M$  edge, making the observed association between  $T$  and  $Y$  not purely causal. See Pearl [18, Section 11.3.1 and 11.3.3] for more information on this topic.

Note that we actually can condition on some descendants of  $T$  without inducing non-causal associations between  $T$  and  $Y$ . For example, conditioning on descendants of  $T$  that aren't on any causal paths to  $Y$  won't induce bias. However, as you can see from the above paragraph, this can get a bit tricky, so it is safest to just not condition on any descendants of  $T$ , as the backdoor criterion prescribes. Even outside of graphical causal models (e.g. in potential outcomes literature), this rule is often applied; it is usually described as not conditioning on any *post-treatment covariates*.

**M-Bias** Unfortunately, even if we only condition on pretreatment covariates, we can still induce collider bias. Consider what would happen if we condition on the collider  $Z_2$  in Figure 4.16. Doing this opens up a backdoor path, along which non-causal association can flow. This is known as *M-bias* due to the M shape that this non-causal association flows along when the graph is drawn with children below their parents. For many examples of collider bias, see Elwert and Winship [19].

## 4.6 Example Applications of the Backdoor Adjustment

### 4.6.1 Association vs. Causation in a Toy Example

In this section, we posit a toy generative process and derive the bias of the associational quantity  $\mathbb{E}[Y | t]$ . We compare this to the causal quantity  $\mathbb{E}[Y | do(t)]$ , which gives us exactly what we want. Note that both of these quantities are actually functions of  $t$ . If the treatment were binary, then we would just look at the difference between the quantities with  $T = 1$  and with  $T = 0$ . However, because our generative processes will be linear,  $\frac{d\mathbb{E}[Y|t]}{dt}$  and  $\frac{d\mathbb{E}[Y|do(t)]}{dt}$  actually gives us all the information about the treatment effect, regardless of if treatment is continuous, binary, or multi-valued. We will assume infinite data so that we can work with expectations. This means this section has nothing to do with estimation; for estimation, see the next section

The generative process that we consider has the causal graph in Figure 4.17

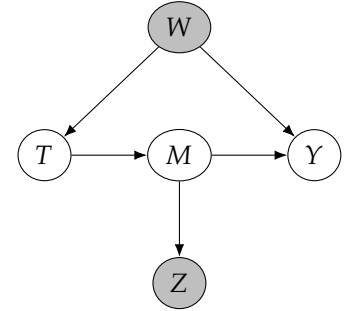


Figure 4.14: Causal graph where the child of a mediator is conditioned on.

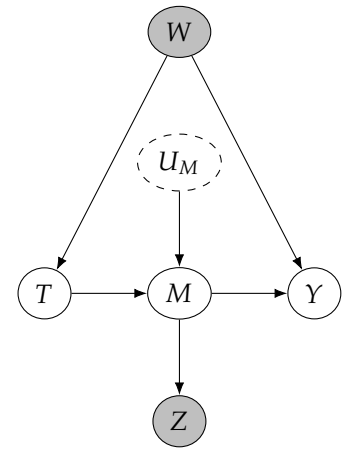


Figure 4.15: Magnified causal graph where the child of a mediator is conditioned on.

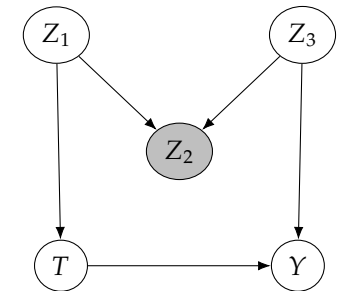


Figure 4.16: Causal graph depicting M-Bias.

and the following structural equations:

$$T := \alpha_1 X \quad (4.28)$$

$$Y := \beta T + \alpha_2 X. \quad (4.29)$$

Note that in the structural equation for  $Y$ ,  $\beta$  is the coefficient in front of  $T$ . This means that the causal effect of  $T$  on  $Y$  is  $\beta$ . Keep this in mind as we go through these calculations.

From the causal graph in Figure 4.17, we can see that  $X$  is a sufficient adjustment set. Therefore,  $\mathbb{E}[Y \mid do(t)] = \mathbb{E}_X \mathbb{E}[Y \mid t, X]$ . Let's calculate the value of this quantity in our example.

$$\mathbb{E}_X \mathbb{E}[Y \mid t, X] = \mathbb{E}_X [\mathbb{E}[\beta T + \alpha_2 X \mid T = t, X]] \quad (4.30)$$

$$= \mathbb{E}_X [\beta t + \alpha_2 X] \quad (4.31)$$

$$= \beta t + \alpha_2 \mathbb{E}[X] \quad (4.32)$$

Importantly, we made use of the equality that the structural equation for  $Y$  (Equation 4.29) gives us in Equation 4.30. Now, we just have to take the derivative to get the causal effect:

$$\frac{d \mathbb{E}_X \mathbb{E}[Y \mid t, X]}{dt} = \beta. \quad (4.33)$$

We got exactly what we were looking for. Now, let's move to the associational quantity:

$$\mathbb{E}[Y \mid T = t] = \mathbb{E}[\beta T + \alpha_2 X \mid T = t] \quad (4.34)$$

$$= \beta t + \alpha_2 \mathbb{E}[X \mid T = t] \quad (4.35)$$

$$= \beta t + \frac{\alpha_2}{\alpha_1} t \quad (4.36)$$

In Equation 4.36, we made use of the equality that the structural equation for  $T$  (Equation 4.28) gives us. If we then take the derivative, we see that there is confounding bias:

$$\frac{d \mathbb{E}[Y \mid t]}{dt} = \beta + \frac{\alpha_2}{\alpha_1}. \quad (4.37)$$

To recap,  $\mathbb{E}_X \mathbb{E}[Y \mid t, X]$  gave us the causal effect we were looking for (Equation 4.33), whereas the associational quantity  $\mathbb{E}[Y \mid t]$  did not (Equation 4.37). Now, let's go through an example that also takes into account estimation.

#### 4.6.2 A Complete Example with Estimation

Recall that we estimated a concrete value for the causal effect of sodium intake on blood pressure in Section 2.5. There, we used the potential outcomes framework. Here, we will do the same thing, but using causal graphs. The spoiler is that the 19% error that we saw in Section 2.5 was due to conditioning on a collider.

First, we need to write down our causal assumptions in terms of a causal graph. Remember that in Luque-Fernandez et al. [8]'s example from epidemiology, the treatment  $T$  is sodium intake, and the outcome  $Y$  is

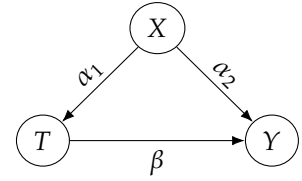


Figure 4.17: Causal graph for toy example

[8]: Luque-Fernandez et al. (2018), 'Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application'

blood pressure. The covariates are age  $W$  and amount of protein in urine (proteinuria)  $Z$ . Age is a common cause of both blood pressure and the body's ability to self-regulate sodium levels. In contrast, high amounts of urinary protein are caused by high blood pressure and high sodium intake. This means that proteinuria is a collider. We depict this causal graph in Figure 4.18.

Because  $Z$  is a collider, conditioning on it induces bias. Because  $W$  and  $Z$  were grouped together as “covariates”  $X$  in Section 2.5, we conditioned on all of them. This is why we saw that our estimate was 19% off from the true causal effect 1.05. Now that we've made the causal relationships clear with a causal graph, the backdoor criterion (Definition 4.1) tells us to only adjust for  $W$  and to not adjust for  $Z$ . More precisely, we were doing the following adjustment in Section 2.5:

$$\mathbb{E}_{W,Z} \mathbb{E}[Y | t, W, Z] \quad (4.38)$$

And now, we will use the backdoor adjustment (Theorem 4.2) to change our statistical estimand to the following:

$$\mathbb{E}_W \mathbb{E}[Y | t, W] \quad (4.39)$$

We have simply removed the collider  $Z$  from the variables we adjust for. For estimation, just as we did in Section 2.5, we use a model-assisted estimator. We replace the outer expectation over  $W$  with an empirical mean over  $W$  and replace the conditional expectation  $\mathbb{E}[Y | t, W]$  with a machine learning model (in this case, linear regression).

Just as writing down the graph has lead us to simply not condition on  $Z$  in Equation 4.39, the code for estimation also barely changes. We need to change just a single line of code in our previous program (Listing 2.1). We display the full program with the fixed line of code below:

```
1 import numpy as np
2 import pandas as pd
3 from sklearn.linear_model import LinearRegression
4
5 Xt = df[['sodium', 'age']]
6 y = df['blood_pressure']
7 model = LinearRegression()
8 model.fit(Xt, y)
9
10 Xt1 = pd.DataFrame.copy(Xt)
11 Xt1['sodium'] = 1
12 Xt0 = pd.DataFrame.copy(Xt)
13 Xt0['sodium'] = 0
14 ate_est = np.mean(model.predict(Xt1) - model.predict(Xt0))
15 print('ATE estimate:', ate_est)
```

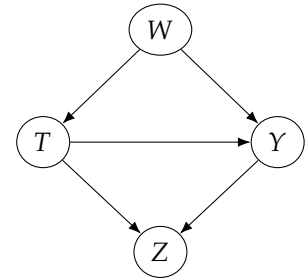
Namely, we've changed line 5 from

```
5 | Xt = df[['sodium', 'age', 'proteinuria']]

in Listing 2.1 to

5 | Xt = df[['sodium', 'age']]
```

in Listing 4.1. When we run this revised code, we get an ATE estimate of 1.0502, which corresponds to **0.02%** error (true value is 1.05) when using



**Figure 4.18:** Causal graph for the blood pressure example.  $T$  is sodium intake.  $Y$  is blood pressure.  $W$  is age. And, importantly, the amount of protein excreted in urine  $Z$  is a collider.

**Listing 4.1:** Python code for estimating the ATE, without adjusting for the collider

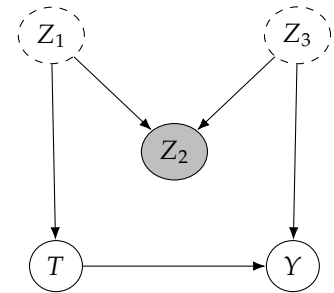
Full code, complete with simulation, is available at [https://github.com/bradyneal/causal-book-code/blob/master/sodium\\_example.py](https://github.com/bradyneal/causal-book-code/blob/master/sodium_example.py).

a fairly large sample.<sup>9</sup>

**Progression of Reducing Bias** When looking at the total association between  $T$  and  $Y$  by simply regressing  $Y$  on  $T$ , we got an estimate that was a staggering 407% off of the true causal effect, due largely to confounding bias (see Section 2.5). When we adjusted for all covariates in Section 2.5, we reduced the percent error all the way down to 19%. In this section, we saw this remaining error is due to collider bias. When we removed the collider bias, by not conditioning on the collider  $Z$ , the error became non-existent.

**Potential Outcomes and M-Bias** In fairness to the general culture around the potential outcomes framework, it is common to only condition on pretreatment covariates. This would prevent a practitioner who adheres to this rule from conditioning on the collider  $Z$  in Figure 4.18. However, there is no reason that there can't be pretreatment colliders that induce M-bias (Section 4.5.3). In Figure 4.19, we depict an example of M-bias that is created by conditioning on  $Z_2$ . We could fix this by additionally conditioning on  $Z_1$  and/or  $Z_3$ , but in this example, they are unobserved (indicated by the dashed lines). This means that the only way to avoid M-bias in Figure 4.19 is to not condition on the covariates  $Z_2$ .

<sup>9</sup> **Active reading exercise:** Given that  $Y$  is generated as a linear function of  $T$  and  $W$ , could we have just used the coefficient in front of  $T$  in the linear regression as an estimate for the causal effect?



**Figure 4.19:** Causal graph depicting M-Bias that can only be avoided by not conditioning on the collider  $Z_2$ . This is due to the fact that the dashed nodes  $Z_1$  and  $Z_3$  are unobserved.

## 4.7 Assumptions Revisited

The first main set of assumptions is encoded by the causal graph that we write down. Exactly what this causal graph means is determined by two main assumptions, each of which can take on several different forms:

### 1. The Modularity Assumption

Different forms:

- ▶ Modularity Assumption for Causal Bayesian Networks (Assumption 4.1)
- ▶ Modularity Assumption for SCMs (Assumption 4.2)
- ▶ The Law of Counterfactuals (Definition 4.3)

### 2. The Markov Assumption

Different equivalent forms:

- ▶ Local Markov assumption (Assumption 3.1)
- ▶ Bayesian network factorization (Definition 3.1)
- ▶ Global Markov assumption (Theorem 3.1)

Given, these two assumptions (and positivity), if the backdoor criterion (Definition 4.1) is satisfied in our assumed causal graph, then we have identification. Note that although the backdoor criterion is a sufficient condition for identification, it is not a necessary condition. We will see this more in Chapter 6.

**More Formal** If you're really into fancy formalism, there are some relevant sources to check out. You can see the fundamental axioms that underlie The Law of Counterfactuals in [20, 21], or if you want a textbook, you can find them in [18, Chapter 7.3]. To see proofs of the equivalence of all three forms of the Markov assumption, see, for example, [13, Chapter 3].

Now that you're familiar with causal graphical models and SCMs, it may be worth going back and rereading Chapter 2 while trying to make connections to what you've learned about graphical causal models in these past two chapters.

[20]: Galles and Pearl (1998), 'An Axiomatic Characterization of Causal Counterfactuals'

[21]: Halpern (1998), 'Axiomatizing Causal Reasoning'

[18]: Pearl (2009), *Causality*

[13]: Koller and Friedman (2009), *Probabilistic Graphical Models: Principles and Techniques*

**Connections to No Interference, Consistency, and Positivity** The no interference assumption (Assumption 2.4) is commonly implicit in causal graphs, since the outcome  $Y$  (think  $Y_i$ ) usually only has a single node  $T$  (think  $T_i$ ) for treatment as a parent, rather than having multiple treatment nodes  $T_i, T_{i-1}, T_{i+1}$ , etc. as parents. However, causal DAGs can be extended to settings where there is interference [22]. Consistency (Assumption 2.5) follows from the axioms of SCMs (see [18, Corollary 7.3.2] and [23]). Positivity (Assumption 2.3) is still a very important assumption that we must make, though it is sometimes neglected in the graphical models literature.

[22]: Ogburn and VanderWeele (2014), ‘Causal Diagrams for Interference’

[18]: Pearl (2009), *Causality*

[23]: Pearl (2010), ‘On the consistency rule in causal inference: axiom, definition, assumption, or theorem?’



# Randomized Experiments

# 5

Randomized experiments are noticeably different from observational studies. In randomized experiments, the experimenter has complete control over the *treatment assignment mechanism* (how treatment is assigned). For example, in the most simple kind of randomized experiment, the experimenter randomly assigns (e.g. via coin toss) each participant to either the treatment group or the control group. This complete control over how treatment is chosen is what distinguishes randomized experiments from observational studies. In this simple experimental setup, the treatment isn't a function of covariates at all! In contrast, in observational studies, the treatment is almost always a function of some covariate(s). As we will see, this difference is key to whether or not confounding is present in our data.

In randomized experiments, association *is* causation. This is because randomized experiments are special in that they guarantee that there is no confounding. As a consequence, this allows us to measure the causal effect  $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$  via the associational difference  $\mathbb{E}[Y | T = 1] - \mathbb{E}[Y | T = 0]$ . In the following sections, we explain why this is the case from a variety of different perspectives. If any one of these explanations clicks with you, that might be good enough. Definitely stick through to the most visually appealing explanation in Section 5.3.

5.1 Comparability and Covariate Balance . . . . .	49
5.2 Exchangeability . . . . .	50
5.3 No Backdoor Paths . . . . .	51

## 5.1 Comparability and Covariate Balance

Ideally, the treatment and control groups would be the same, in all aspects, except for treatment. This would mean they only differ in the treatment they receive (i.e. they are *comparable*). This would allow us to attribute any difference in the outcomes of the treatment and control groups to the treatment. Saying that these treatment groups are the same in everything other than their treatment and outcomes is the same as saying they have the same distribution of confounders. Because people often check for this property on observed variables (often what people mean by “covariates”), this concept is known as *covariate balance*.

**Definition 5.1** (Covariate Balance) *We have covariate balance if the distribution of covariates  $X$  is the same across treatment groups. More formally,*

$$P(X | T = 1) \stackrel{d}{=} P(X | T = 0) \quad (5.1)$$

Randomization implies covariate balance, across all covariates, even unobserved ones. Intuitively, this is because the treatment is chosen at random, regardless of  $X$ , so the treatment and control groups should look very similar. The proof is simple. Because  $T$  is not at all determined by  $X$  (solely by a coin flip),  $T$  is independent of  $X$ . This means that

The symbol  $\stackrel{d}{=}$  means “equal in distribution.”

$P(X | T = 1) \stackrel{d}{=} P(X)$ . Similarly, it means  $P(X | T = 0) \stackrel{d}{=} P(X)$ . Therefore, we have  $P(X | T = 1) \stackrel{d}{=} P(X | T = 0)$ .

Although we have proven that randomization implies covariate balance, we have not proven that that covariate balance implies that association is causation.<sup>1</sup> We'll now prove that by showing that  $P(y | do(t)) = P(y | t)$ . For the proof, the main property we utilize is that covariate balance implies  $X$  and  $T$  are independent.

<sup>1</sup> Recall that the intuition is that covariate balance means that everything is the same between the treatment groups, except for the treatment, so the treatment must be the explanation for the change in  $Y$ .

*Proof.* First, let  $X$  be a sufficient adjustment set that potentially contains *unobserved* variables (randomization also balances unobserved covariates). Such an adjustment set must exist because we allow it to contain any variables, observed or unobserved. Then, we have the following from the backdoor adjustment (Theorem 4.2):

$$P(y | do(t)) = \sum_x P(y | t, x)P(x) \quad (5.2)$$

By multiplying by  $\frac{P(t|x)}{P(t|x)}$ , we get the joint distribution in the numerator:

$$= \sum_x \frac{P(y | t, x)P(t | x)P(x)}{P(t | x)} \quad (5.3)$$

$$= \sum_x \frac{P(y, t, x)}{P(t | x)} \quad (5.4)$$

Now, we use the important property that  $X \perp\!\!\!\perp T$ :

$$= \sum_x \frac{P(y, t, x)}{P(t)} \quad (5.5)$$

An application of Bayes rule and marginalization gives us the rest:

$$= \sum_x P(y, x | t) \quad (5.6)$$

$$= P(y | t) \quad (5.7)$$

□

## 5.2 Exchangeability

Exchangeability (Assumption 2.1) gives us another perspective on why randomization makes causation equal to association. To see why, consider the following thought experiment. We decide an individual's treatment group using a random coin flip as follows: if the coin is heads, we assign the individual to the treatment group ( $T = 1$ ), and if the coin is tails, we assign the individual to the control group ( $T = 0$ ). If the groups are exchangeable, we could exchange these groups, and the average outcomes would remain the same. This is intuitively true if we chose the groups with a coin flip. Imagine simply swapping the meaning of "heads" and "tails" in this experiment. Would you expect that to change the results at all? No. This is why randomized experiments give us exchangeability.

Recall from Section 2.3.2 that mean exchangeability is formally the following:

$$\mathbb{E}[Y(1) \mid T = 1] = \mathbb{E}[Y(1) \mid T = 0] \quad (5.8)$$

$$\mathbb{E}[Y(0) \mid T = 0] = \mathbb{E}[Y(0) \mid T = 1] \quad (5.9)$$

The “exchange” is when we go from  $Y(1)$  in the treatment group to  $Y(1)$  in the control group (Equation 5.8) and from  $Y(0)$  in the control group to  $Y(0)$  in the treatment group (Equation 5.8).

To see the proof of why association is causation in randomized experiments through the lens of exchangeability, recall the proof from Section 2.3.2. First, recall that Equation 5.8 means that both quantities in it are equal to the marginal expected outcome  $\mathbb{E}[Y(1)]$  and, similarly, that Equation 5.9 means that both quantities in it are equal to the marginal expected outcome  $\mathbb{E}[Y(0)]$ . Then, we have the following proof:

$$\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1) \mid T = 1] - \mathbb{E}[Y(0) \mid T = 0] \quad (2.3 \text{ revisited})$$

$$= \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] \quad (2.4 \text{ revisited})$$

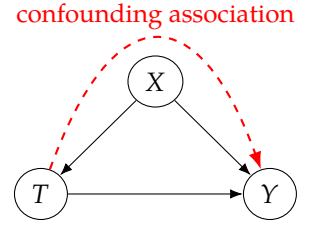
### 5.3 No Backdoor Paths

The final perspective that we’ll look at to see why association is causation in randomized experiments is that of graphical causal models. In regular observational data, there is almost always confounding. For example, in Figure 5.1, we see that  $X$  is a confounder of the effect of  $T$  on  $Y$ . Non-causal association flows along the backdoor path  $T \leftarrow X \rightarrow Y$ .

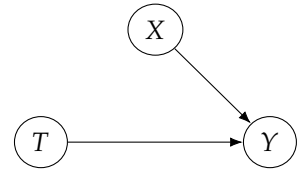
However, if we randomize  $T$ , something magical happens:  $T$  no longer has any causal parents, as we depict in Figure 5.2. This is because  $T$  is purely random. It doesn’t depend on anything other than the output of a coin toss (or a quantum random number generator, if you’re into the kind of stuff). Because  $T$  has no incoming edges, under randomization, there are no backdoor paths. So the empty set is a sufficient adjustment set. This means that all of the association that flows from  $T$  to  $Y$  is causal. We can identify  $P(Y \mid \text{do}(T = t))$  by simply applying the backdoor adjustment (Theorem 4.2), adjusting for the empty set:

$$P(Y \mid \text{do}(T = t)) = P(Y \mid T = t)$$

With that, we conclude our discussion of why association is causation in randomized experiments. Hopefully, at least one of these three explanations is intuitive to you and easy to store in long-term memory.



**Figure 5.1:** Causal structure of  $X$  confounding the effect of  $T$  on  $Y$ .



**Figure 5.2:** Causal structure when we randomize treatment.

# Nonparametric Identification

# 6

In Section 4.4, we saw that satisfying the backdoor criterion is sufficient to give us identifiability, but is the backdoor criterion also necessary? In other words, is it possible to get identifiability without being able to block all backdoor paths?

As an example, consider that we have data generated according to the graph in Figure 6.1. We don't observe  $W$  in this data, so we can't block the backdoor path through  $W$  and the confounding association that flows along it. But we still need to identify the causal effect. It turns out that it is possible to identify the causal effect in this graph, using the frontdoor criterion. We'll see the frontdoor criterion and corresponding adjustment in Section 6.1. Then, we'll consider even more general identification in Section 6.2 when we introduce *do*-calculus. We'll conclude with graphical conditions for identifiability in Section 6.3.

## 6.1 Frontdoor Adjustment

The high-level intuition for why we can identify the causal effect of  $T$  on  $Y$  in the graph in Figure 6.1 (even when we can't adjust for the confounder  $W$  because it is unobserved) is as follows: a mediator like  $M$  is very helpful; we can isolate the association that flows through  $M$  by focusing our statistical analysis on  $M$ , and the only association that flows through  $M$  is causal association (association flowing along directed paths from  $T$  to  $Y$ ). We illustrate this intuition in Figure 6.2, where we depict only the *causal* association. In this section, we will focus our analysis on  $M$  using a three step procedure (see Figure 6.3 for our corresponding illustration):

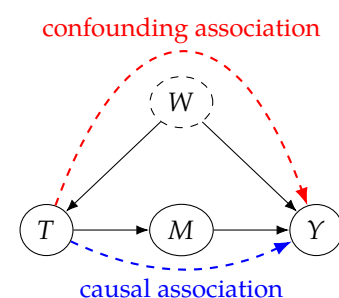
1. Identify the causal effect of  $T$  on  $M$ .
2. Identify the causal effect of  $M$  on  $Y$ .
3. Combine the above steps to identify the causal effect of  $T$  on  $Y$ .

**Step 1** First, we will identify the effect of  $T$  on  $M$ :  $P(m \mid do(t))$ . Because  $Y$  is a collider on the  $T - M$  path through  $W$ , it blocks that backdoor path. So there are no unblocked backdoor paths from  $T$  to  $M$ . This means that the only association that flows from  $T$  to  $M$  is the causal association that flows along the edge connecting them. Therefore, we have the following identification via the backdoor adjustment (Theorem 4.2, using the empty set as the adjustment set):<sup>1</sup>

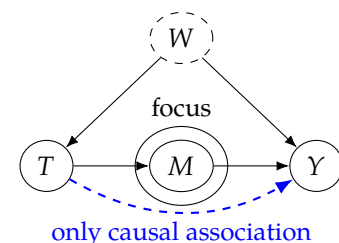
$$P(m \mid do(t)) = P(m \mid t) \quad (6.1)$$

**Step 2** Second, we will identify the effect of  $M$  on  $Y$ :  $P(y \mid do(m))$ . Because  $T$  blocks the backdoor path  $M \leftarrow T \leftarrow W \rightarrow Y$ , we can simply

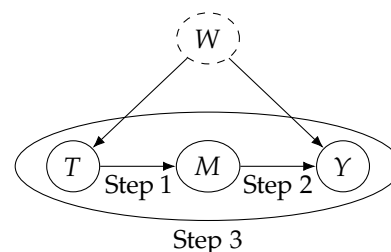
6.1 Frontdoor Adjustment . . .	52
6.2 <i>do</i> -calculus . . . . .	55
Application: Frontdoor Adjustment . . . . .	57
6.3 Determining Identifiability from the Graph . . . . .	58



**Figure 6.1:** Causal graph where  $W$  is unobserved, so we cannot block the backdoor path. We depict the flow of causal association and the flow of confounding association with dashed lines.



**Figure 6.2:** In contrast to Figure 6.1, when we focus our analysis on  $M$ , we are able to isolate only the *causal* association.



**Figure 6.3:** Illustration of steps to get to the frontdoor adjustment.

<sup>1</sup> **Active reading exercise:** Write a proof for Equation 6.1 without using the backdoor adjustment. Instead, start from the truncated factorization (Proposition 4.1) like we did in Section 4.3.1. Hint: The proof can be quite short. We provide a proof in Appendix A.1, in case you get stuck.

adjust for  $T$ . Therefore, using the backdoor adjustment again, we have the following:

$$P(y \mid do(m)) = \sum_t P(y \mid m, t) P(t) \quad (6.2)$$

**Step 3** Now that we know how changing  $T$  changes  $M$  (step 1) and how changing  $M$  changes  $Y$  (step 2), we can combine these two to get how changing  $T$  changes  $Y$  (through  $M$ ):

$$P(y \mid do(t)) = \sum_m P(m \mid do(t)) P(y \mid do(m)) \quad (6.3)$$

The first factor on the right-hand side corresponds to setting  $T$  to  $t$  and observing the resulting value of  $M$ . The second factor corresponds to setting  $M$  to exactly the value  $m$  that resulted from setting  $T$  and then observing what value of  $Y$  results. We must sum over  $m$  because  $P(m \mid do(t))$  is probabilistic, so we must sum over its support. In other words, we must sum over all possible realizations  $m$  of the random variables whose distribution is  $P(M \mid do(t))$ .

If we then plug in Equations 6.1 and 6.2 into Equation 6.3, we get the *frontdoor adjustment* (keep reading to see the definition of the frontdoor criterion):

**Theorem 6.1** (Frontdoor Adjustment) *If  $(T, M, Y)$  satisfy the frontdoor criterion and we have positivity, then*

$$P(y \mid do(t)) = \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t') P(t') \quad (6.4)$$

The causal graph we've been using (Figure 6.4) is an example of a simple graph that satisfies the *frontdoor criterion*. To get the full definition, we must first define *complete/full mediation*: a set of variables  $M$  completely mediates the effect of  $T$  on  $Y$  if all causal (directed) paths from  $T$  to  $Y$  go through  $M$ . We now give the general definition of the frontdoor criterion:

**Definition 6.1** (Frontdoor Criterion) *A set of variables  $M$  satisfies the frontdoor criterion relative to  $T$  and  $Y$  if the following are true:*

1.  $M$  completely mediates the effect of  $T$  on  $Y$  (i.e. all causal paths from  $T$  to  $Y$  go through  $M$ ).
2. There is no unblocked backdoor path from  $T$  to  $M$ .
3. All backdoor paths from  $M$  to  $Y$  are blocked by  $T$ .<sup>2</sup>

Although Equations 6.1 and 6.2 are straightforward applications of the backdoor adjustment, we hand-waved our way to Equation 6.3, which was key to the frontdoor adjustment (Theorem 6.1). We'll now walk through how to get Equation 6.3. **Active reading exercise:** Feel free to stop reading here and do this yourself.

We are about to enter *Equationtown* (Figure 6.5), so if you are satisfied with the intuition we gave for step 3 and prefer to not see a lot of equations, feel free to skip to the end of the proof (denoted by the  $\square$  symbol).

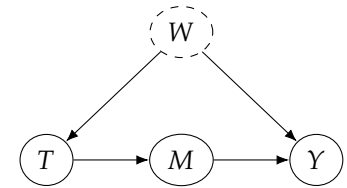


Figure 6.4: Simple causal graph that satisfies the frontdoor criterion

<sup>2</sup> **Active reading exercise:** Think of a graph other than Figure 6.4 that satisfies the frontdoor criterion. Also, for each condition, think of a graph that does *not* satisfy only that condition.

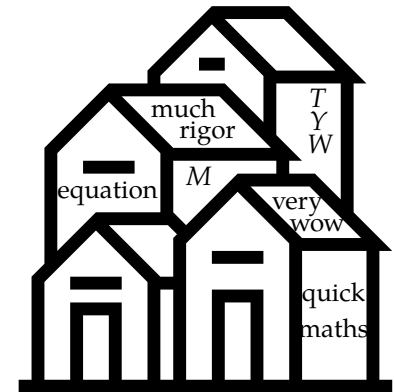


Figure 6.5: Equationtown

*Proof.* As usual, we start with the truncated factorization, using the causal graph in Figure 6.4. From the Bayesian network factorization (Definition 3.1), we have the following:

$$P(w, t, m, y) = P(w) P(t | w) P(m | t) P(y | w, m) \quad (6.5)$$

Then, using the truncated factorization (Proposition 4.1), we remove the factor for  $T$ :

$$P(w, m, y | do(t)) = P(w) P(m | t) P(y | w, m) \quad (6.6)$$

Next, we marginalize out  $w$  and  $m$ :

$$\sum_m \sum_w P(w, m, y | do(t)) = \sum_m \sum_w P(w) P(m | t) P(y | w, m) \quad (6.7)$$

$$P(y | do(t)) = \sum_m P(m | t) \sum_w P(y | w, m) P(w) \quad (6.8)$$

Even though we've removed all the *do* operators, recall that we are not done because  $W$  is unobserved. So we must also remove the  $w$  from the expression. This is where we have to get a bit creative.

We want to be able to combine  $P(y | w, m)$  and  $P(w)$  into a joint factor over both  $y$  and  $w$  so that we can marginalize out  $w$ . To do this, we need to get  $m$  behind the conditioning bar of the  $P(w)$  factor. This would be easy if we could just swap  $P(w)$  out for  $P(w | m)$  in Equation 6.8.<sup>3</sup> The key thing to notice is that we actually can include  $m$  behind the conditioning bar if  $t$  were also there because  $T$  d-separates  $W$  from  $M$  in Figure 6.6. In math, this means that the following equality holds:

$$P(w | t) = P(w | t, m) \quad (6.9)$$

Great, so how do we get  $t$  into this party? The usual trick of conditioning on it and marginalizing it out:

$$P(y | do(t)) = \sum_m P(m | t) \sum_w P(y | w, m) P(w) \quad (6.8 \text{ revisited})$$

$$= \sum_m P(m | t) \sum_w P(y | w, m) \sum_{t'} P(w | t') P(t') \quad (6.10)$$

$$= \sum_m P(m | t) \sum_w P(y | w, m) \sum_{t'} P(w | t', m) P(t') \quad (6.11)$$

$$= \sum_m P(m | t) \sum_{t'} P(t') \sum_w P(y | w, m) P(w | t', m) \quad (6.12)$$

Great, but now we can't combine  $P(y | w, m)$  and  $P(w | t', m)$  because  $P(y | w, m)$  is missing this newly introduced  $t'$  behind its conditioning bar. Luckily, we can fix that<sup>4</sup> and combine the two factors:

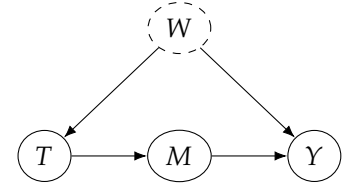
$$= \sum_m P(m | t) \sum_{t'} P(t') \sum_w P(y | w, m) P(w | t', m) \quad (6.13)$$

$$= \sum_m P(m | t) \sum_{t'} P(t') \sum_w P(y | w, t', m) P(w | t', m) \quad (6.14)$$

$$= \sum_m P(m | t) \sum_{t'} P(t') \sum_w P(y, w | t', m) \quad (6.15)$$

$$= \sum_m P(m | t) \sum_{t'} P(t') P(y | t', m) \quad (6.16)$$

<sup>3</sup> **Active reading exercise:** Why would it be easy to marginalize out  $w$  if it were the case that  $P(w) = P(w | m)$ ? And why does this equality not hold?



**Figure 6.6:** Simple causal graph that satisfies the frontdoor criterion

<sup>4</sup> **Active reading exercise:** Why is  $P(y | w, m)$  equal to  $P(y | w, t', m)$ ?

This matches the result stated in Theorem 6.1, so we've completed the derivation of the frontdoor adjustment without using the backdoor adjustment. However, we still need to show that Equation 6.3 is correct to justify step 3. To do that, all that's left is to recognize that these parts match Equations 6.1 and 6.2 and plug those in:

$$= \sum_m P(m \mid do(t)) P(y \mid do(m)) \quad (6.17)$$

$$P(m \mid do(t)) = P(m \mid t) \quad (6.1)$$

$$P(y \mid do(m)) = \sum_t P(y \mid m, t) P(t) \quad (6.2)$$

□

And we're done! We just needed to be a bit clever with our uses of d-separation and marginalization. Part of why we went through that proof is because we will prove the frontdoor adjustment using *do*-calculus in Section 6.2. This way you can easily compare a proof using the truncated factorization to a proof using *do*-calculus to prove the same result.

## 6.2 *do*-calculus

As we saw in the last section, it turns out that satisfying the backdoor criterion (Definition 4.1) isn't necessary to identify causal effects. For example, if the frontdoor criterion (Definition 6.1) is satisfied, that also gives us identifiability. This leads to the following questions: can we identify causal estimands when the associated causal graph satisfies neither the backdoor criterion nor the frontdoor criterion? If so, how? Pearl's *do*-calculus [24] gives us the answer to these questions.

[24]: Pearl (1995), 'Causal diagrams for empirical research'

As we will see, the *do*-calculus gives us tools to identify causal effects using the causal assumptions encoded in the causal graph. It will allow us to identify any causal estimand that is identifiable. More concretely, consider an arbitrary causal estimand  $P(Y \mid do(T = t), X = x)$ , where  $Y$  is an arbitrary set of outcome variables,  $T$  is an arbitrary set of treatment variables, and  $X$  is an arbitrary (potentially empty) set of covariates that we want to choose how specific the causal effect we're looking at is. Note that this means we can use *do*-calculus to identify causal effects where there are multiple treatments and/or multiple outcomes.

In order to present the rules of *do*-calculus, we must define a bit of notation for augmented versions of the causal graph  $G$ . Let  $G_{\overline{X}}$  denote the graph that we get if we take  $G$  and remove all of the *incoming* edges to nodes in the set  $X$ ; recall from Section 4.2 that this is known as the *manipulated graph*. Let  $G_{\underline{X}}$  denote the graph that we get if we take  $G$  and remove all of the *outgoing* edges from nodes in the set  $X$ . The mnemonic meaning to help you remember this is to think of parents as drawn above their children in the graph, so the bar above  $X$  is cutting its incoming edges and the bar below  $X$  is cutting its outgoing edges. Combining these two, we'll use  $G_{\overline{X}\underline{Z}}$  to denote the graph with the incoming edges to  $X$  and the outgoing edges from  $Z$  removed. And recall from Section 3.7 that we use  $\perp\!\!\!\perp_G$  to denote d-separation in  $G$ . We're now ready; *do*-calculus consists of just three rules:

**Theorem 6.2** (Rules of *do*-calculus) *Given a causal graph  $G$ , an associated*



distribution  $P$ , and disjoint sets of variables  $Y$ ,  $T$ ,  $Z$ , and  $W$ , the following rules hold.

Rule 1:

$$P(y \mid do(t), z, w) = P(y \mid do(t), w) \quad \text{if } Y \perp_{G_{\overline{T}}} Z \mid T, W \quad (6.18)$$

Rule 2:

$$P(y \mid do(t), do(z), w) = P(y \mid do(t), z, w) \quad \text{if } Y \perp_{G_{\overline{T}, \underline{Z}}} Z \mid T, W \quad (6.19)$$

Rule 3:

$$P(y \mid do(t), do(z), w) = P(y \mid do(t), w) \quad \text{if } Y \perp_{G_{\overline{T}, \overline{Z(W)}}} Z \mid T, W \quad (6.20)$$

where  $Z(W)$  denotes the set of nodes of  $Z$  that aren't ancestors of any node of  $W$  in  $G_{\overline{T}}$ .

Now, rather than recreate the proofs for these rules from Pearl [24], we'll give intuition for each of them in terms of concepts we've already seen in this book.

[24]: Pearl (1995), 'Causal diagrams for empirical research'

**Rule 1 Intuition** If we take Rule 1 and simply remove the intervention  $do(t)$ , we get the following ([Active reading exercise](#): what familiar concept is this?):

$$P(y \mid z, w) = P(y \mid w) \quad \text{if } Y \perp_G Z \mid W \quad (6.21)$$

This is just what d-separation gives us under the Markov assumption; recall from Theorem 3.1 that d-separation in the graph implies conditional independence in  $P$ . This means that Rule 1 is simply a generalization of Theorem 3.1 to interventional distributions.

**Rule 2 Intuition** Just as with Rule 1, we'll remove the intervention  $do(t)$  from Rule 2 and see what this reminds us of ([Active reading exercise](#): what concept does this remind you of?):

$$P(y \mid do(z), w) = P(y \mid z, w) \quad \text{if } Y \perp_{G_{\underline{Z}}} Z \mid W \quad (6.22)$$

This is exactly what we do when we justify the backdoor adjustment (Theorem 4.2) using the backdoor criterion (Definition 4.1). As we saw at the ends of Section 3.8 and Section 4.4. Association *is* causation if the outcome  $Y$  and the treatment  $Z$  are d-separated by some set of variables that are conditioned on  $W$ . So rule 2 is a generalization of the backdoor adjustment to interventional distributions.

**Rule 3 Intuition** This is the trickiest rule to understand. Just as with the other two rules, we'll first remove the intervention  $do(t)$  to make thinking about this simpler:

$$P(y \mid do(z), w) = P(y \mid w) \quad \text{if } Y \perp_{G_{\overline{Z(W)}}} Z \mid W \quad (6.23)$$

To get the equality in this equation, it must be the case that removing the intervention  $do(z)$  (which is like taking the manipulated graph and reintroducing the edges going into  $Z$ ) introduces no new association that can affect  $Y$ . Because  $do(z)$  removes the incoming edges to  $Z$  to give us  $G_{\overline{Z}}$ , the main association that we need to worry about is association flowing from  $Z$  to  $Y$  in  $G_{\overline{Z}}$  (causal association). Therefore, you might

expect that the condition that gives us the equality in Equation 6.23 is  $Y \perp\!\!\!\perp_{G_Z} Z \mid W$ . However, we have to refine this a bit to prevent inducing association by conditioning on the descendants of colliders (recall from Section 3.6). Namely,  $Z$  could contain colliders in  $G$ , and  $W$  could contain descendants of these colliders. Therefore, to not induce new association through colliders in  $Z$  when we reintroduce the incoming edges to  $Z$  to get  $G$ , we must limit the set of manipulated nodes to those that are not ancestors of nodes in the conditioning set  $W$ :  $Z(W)$ .

**Completeness of *do*-calculus** Maybe there could exist causal estimands that are identifiable but that can't be identified using only the rules of *do*-calculus in Theorem 6.2. Fortunately, Shpitser and Pearl [25] and Huang and Valtorta [26] independently proved that this is not the case. They proved that *do*-calculus is complete, which means that these three rules are sufficient to identify all identifiable causal estimands. Because these proofs are constructive, they also admit algorithms that identify any causal estimand in polynomial time.

**Nonparametric Identification** Note that all of this is about *nonparametric* identification; in other words, *do*-calculus tells us if we can identify a given causal estimand using only the causal assumptions encoded in the causal graph. If we introduce more assumptions about the distribution (e.g. linearity), we can identify more causal estimands. That would be known as *parametric* identification. We don't discuss parametric identification in this chapter, though we will in later chapters.

[25]: Shpitser and Pearl (2006), 'Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models'

[26]: Huang and Valtorta (2006), 'Pearl's Calculus of Intervention is Complete'

## 6.2.1 Application: Frontdoor Adjustment

Recall the simple graph we used that satisfies the frontdoor criterion (Figure 6.7), and recall the frontdoor adjustment:

$$P(y \mid do(t)) = \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t') P(t') \quad (6.4 \text{ revisited})$$

At the end of Section 6.1, we saw a proof for the frontdoor adjustment using just the truncated factorization. To get an idea for how *do*-calculus works and the intuition we use in proofs that use it, we'll now do the frontdoor adjustment proof using the rules of *do*-calculus.

*Proof.* Our goal is to identify  $P(y \mid do(t))$ . Because we have the intuition we described in Section 6.1 that the full mediator  $M$  will help us out, the first thing we'll do is introduce  $M$  into the equation via the marginalization trick:

$$P(y \mid do(t)) = \sum_m P(y \mid do(t), m) P(m \mid do(t)) \quad (6.24)$$

Because the backdoor path from  $T$  to  $M$  in Figure 6.7 is blocked by the collider  $Y$ , all of the association that flows from  $T$  to  $M$  is causal, so we can apply Rule 2 to get the following:

$$= \sum_m P(y \mid do(t), m) P(m \mid t) \quad (6.25)$$

Now, because  $M$  is a full mediator of the causal effect of  $T$  on  $Y$ , we should be able to replace  $P(y \mid do(t), m)$  with  $P(y \mid do(m))$ , but this will

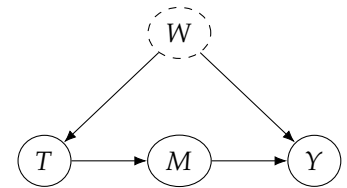


Figure 6.7: Simple causal graph that satisfies the frontdoor criterion

take two steps of *do*-calculus. To remove  $do(t)$ , we'll need to use Rule 3, which requires that  $T$  have no causal effect on  $Y$  in the relevant graph. We can get to a graph like that by removing the edge from  $T$  to  $M$  (Figure 6.9); in *do*-calculus, we do this by using Rule 2 (in the opposite direction as before) to  $do(m)$ . We can do this because the existing  $do(t)$  makes it so there are no backdoor paths from  $M$  to  $Y$  in  $G_{\overline{T}}$  (Figure 6.8).

$$= \sum_m P(y \mid do(t), do(m)) P(m \mid t) \quad (6.26)$$

Now, as we planned, we can remove the  $do(t)$  using Rule 3. We can use Rule 3 here because there is no causation flowing from  $T$  to  $Y$  in  $G_{\overline{M}}$  (Figure 6.9).

$$= \sum_m P(y \mid do(m)) P(m \mid t) \quad (6.27)$$

All that's left is to remove this last *do*-operator. As we discussed in Section 6.1,  $T$  blocks the only backdoor path from  $M$  to  $Y$  in the graph (Figure 6.10). This means, that if we can condition on  $T$ , we can get rid of this last *do*-operator. As usual, we do that by conditioning on and marginalizing out  $T$ . Rearranging a bit and using  $t'$  for the marginalization since  $t$  is already present:

$$= \sum_m P(m \mid t) \sum_{t'} P(y \mid do(m), t') P(t' \mid do(m)) \quad (6.28)$$

Now, we can simply apply Rule 2, since  $T$  blocks the backdoor path from  $M$  to  $Y$ :

$$= \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t') P(t' \mid do(m)) \quad (6.29)$$

And finally, we can apply Rule 3 to remove the last  $do(m)$  because there is no causal effect of  $M$  on  $T$  (i.e. there is no directed path from  $M$  to  $T$  in the graph in (Figure 6.10).

$$= \sum_m P(m \mid t) \sum_{t'} P(y \mid m, t') P(t') \quad (6.30)$$

□

That concludes our proof of the frontdoor adjustment using *do*-calculus. It follows a different path than the proof we gave at the end of Section 6.1, where we used the truncated factorization, but both proofs rely heavily on intuition we get from looking at the graph.

### 6.3 Determining Identifiability from the Graph

It's nice to know that we can identify any causal estimand that is possible to identify using *do*-calculus, but this isn't as satisfying as knowing whether a causal estimand is identifiable by simply looking at the causal graph. For example, the backdoor criterion (Definition 4.1) and the frontdoor criterion (Definition 6.1) gave us simple ways to know for sure that a causal estimand is identifiable. However, there are plenty of

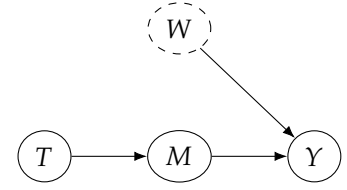


Figure 6.8:  $G_{\overline{T}}$

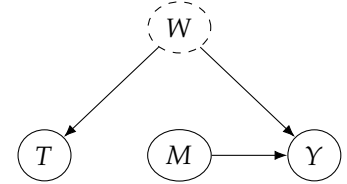


Figure 6.9:  $G_{\overline{M}}$

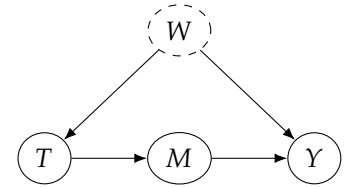


Figure 6.10:  $G$

**Active reading exercise:** Assuming the backdoor criterion, prove the backdoor adjustment using the rules of *do*-calculus.

causal estimands that are identifiable, even though the corresponding causal graphs don't satisfy the backdoor or frontdoor criterion. More general graphical criteria exist that will tell us that these estimands are identifiable. We will discuss these more general graphical criteria for identifiability in this section.

**Single Variable Intervention** When we care about causal effects of an intervention on a single variable, Tian and Pearl [27] provide a relatively simple graphical criterion that is sufficient for identifiability: the *unconfounded children criterion*.

[27]: Tian and Pearl (2002), 'A General Identification Condition for Causal Effects'

**Definition 6.2** (Unconfounded Children Criterion) *This criterion is satisfied if it is possible to block all backdoor paths from the treatment variable  $T$  to all of its children with a single conditioning set.*

This criterion generalizes the backdoor criterion (Definition 4.1) and the frontdoor criterion (Definition 6.1). Like them, it is a *sufficient* condition for identifiability:

**Theorem 6.3** (Unconfounded Children Identifiability) *Let  $Y$  be the set of outcome variables and  $T$  be a single variable. If the unconfounded children criterion and positivity are satisfied, then  $P(Y = y \mid \text{do}(T = t))$  is identifiable [27].*

The intuition for unconfounded children criterion implies identifiability is similar to the intuition for the frontdoor criterion; if we can isolate all of the causal association flowing out of treatment, we have identifiability. To see this intuition, first, consider that all of the causal association from  $T$  must flow through its children. We can isolate this causal association if there is no confounding between  $T$  and any of its children.<sup>5</sup> This isolation of all of the causal association is what gives us identifiability of the causal effect of  $T$  on any other node in the graph. This intuition might lead you to suspect that this criterion is *necessary* in the very specific case where the outcome set  $Y$  is all of the other variables in the graph other than  $T$ ; it turns out that this is true [27]. But this condition is *not necessary* if  $Y$  is a smaller set than that.

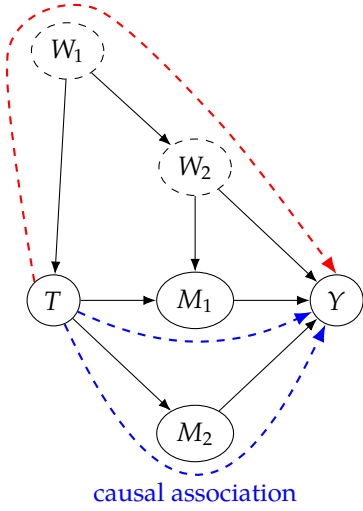
<sup>5</sup> This is analogous to what we saw with the frontdoor criterion in Section 6.1, where we could isolate the causal association flowing through the full mediator  $M$  if the  $T - M$  relationship is unconfounded (no unblocked backdoor paths).

To give you a more visual grasp of the intuition for why the unconfounded children criterion is sufficient for identification, we give an example graph in Figure 6.11. In Figure 6.11a, we visualize the flow of confounding association and causal association that flows in this graph. Then, we depict the isolation of the causal association in that graph in Figure 6.11b.

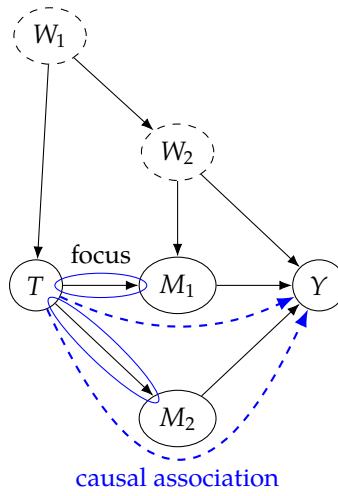
**Necessary Condition** Although the unconfounded children criterion is not necessary for identifiability, there is a necessary condition that is fairly close: For each child  $M$  of  $T$  that is an ancestor of  $Y$ , it is possible to block all backdoor paths from  $T$  to  $M$  [18, p. 92]. The intuition for this is that because the causal association that flows from  $T$  to  $Y$  must go through children of  $T$  that are ancestors of  $Y$ , to be able to isolate this causal association, the effect of  $T$  on these mediating children must be unconfounded. Unfortunately, this condition is *not sufficient*. In sum, the unconfounded children criterion is sufficient but not necessary and this seemingly very similar condition is necessary but not sufficient. It feels like we're close to a necessary *and* sufficient condition for identifiability,

[18]: Pearl (2009), *Causality*

non-causal association



(a) Visualization of the flow of confounding association and causal association.



(b) Visualization of the isolation of the causal association flowing from  $T$  to its children, allowing the unconfounded children criterion to imply identifiability.

**Figure 6.11:** Example graph that satisfies the unconfounded children criterion.

right? Unfortunately, we won't get there in this book, but we'll point you to the relevant articles.

**Necessary and Sufficient Conditions for Multiple Variable Interventions** Shpitser and Pearl [25] provide a necessary and sufficient criterion for identifiability of  $P(Y = y \mid do(T = t))$  when  $Y$  and  $T$  are arbitrary sets of variables: *the hedge criterion*. However, this is outside the scope of this book, as it requires more complex objects such as hedges, C-trees, and other leafy objects. Moving further along, Shpitser and Pearl [28] provide a necessary and sufficient criterion for the most general type of causal estimand: *conditional causal effects*, which take the form  $P(Y = y \mid do(T = t), X = x)$ , where  $Y$ ,  $T$ , and  $X$  are all arbitrary sets of variables.

[25]: Shpitser and Pearl (2006), 'Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models'

[28]: Shpitser and Pearl (2006), 'Identification of Conditional Interventional Distributions'

In the previous chapter, we covered identification. Once we identify some causal estimand by reducing it to a statistical estimand, we still have more work to do. We need to get a corresponding estimate. In this chapter, we'll cover a variety of estimators that we can use to do this. This isn't meant to be anywhere near exhaustive as there are many different estimators of causal effects, but it is meant to give you a solid introduction to them.

All of the estimators that we include full sections on are model-assisted estimators (recall from Section 2.4). And they all work with arbitrary statistical models such as the ones you might get from scikit-learn [29].

## 7.1 Preliminaries

Recall from Chapter 2 that we denote the individual treatment effect (ITE) with  $\tau_i$  and average treatment effect (ATE) with  $\tau$ :

$$\tau_i \triangleq Y_i(1) - Y_i(0) \quad (7.1)$$

$$\tau \triangleq \mathbb{E}[Y_i(1) - Y_i(0)] \quad (7.2)$$

ITEs are the most specific kind of causal effects, but they are hard to estimate without strong assumptions (on top of those discussed in Chapters 2 and 4). However, we often want to estimate causal effects that are a bit more individualized than the ATE. For example, say we've observed an individual's covariates  $x$ ; we might like to use those to estimate a more specific effect for that individual (and anyone else with covariates  $x$ ). This brings us to the *conditional average treatment effect* (CATE)  $\tau(x)$ :

$$\tau(x) \triangleq \mathbb{E}[Y_i(1) - Y_i(0) \mid X = x] \quad (7.3)$$

The  $X$  that is conditioned on does not need to consist of all of the observed covariates, but this is often the case when people refer to CATEs. We call that *individualized average treatment effects* (IATEs).

ITEs and "CATEs" (what we call IATEs) are sometimes conflated, but they are not the same. For example, two individuals could have the same covariates, but their potential outcomes could be different because of other unobserved differences between these individuals. If we encompass everything about an individual that is relevant to their potential outcomes in the vector  $I$ , then ITEs and "CATEs" are the same if  $X = I$ . In a causal graph,  $I$  corresponds to all of the exogenous variables in the magnified graph that have causal association flowing to  $Y$ .<sup>1</sup>

**Unconfoundedness** Throughout this chapter, whenever we are estimating an ATE, we will assume that  $W$  is a sufficient adjustment set, and whenever we are estimating a CATE, we will assume that  $W \cup X$  is a sufficient adjustment set. In other words, for ATE estimation, we assume that  $W$  satisfies the backdoor criterion (Definition 4.1); equivalently for

7.1 Preliminaries . . . . .	61
7.2 Conditional Outcome Modeling (COM) . . . . .	62
7.3 Grouped Conditional Outcome Modeling (GCOM) . . . . .	63
7.4 Increasing Data Efficiency . . . . .	64
TARNet . . . . .	64
X-Learner . . . . .	65
7.5 Propensity Scores . . . . .	66
7.6 Inverse Probability Weighting (IPW) . . . . .	67
7.7 Doubly Robust Methods . . . . .	69
7.8 Other Methods . . . . .	69
7.9 Concluding Remarks . . . . .	70
Confidence Intervals . . . . .	70
Comparison to Randomized Experiments . . . . .	70

<sup>1</sup> This paragraph contains a lot of information. [Active reading exercise:](#)

1) Convince yourself that ITEs and "CATEs" (what we call IATEs) are the same if  $X = I$ .

2) Convince yourself that  $I$  corresponds to the exogenous variables in the magnified graph that have causal association flowing to  $Y$ .

ATE estimation, we assume that we have conditional exchangeability given  $W$  (Assumption 2.2). And similarly for CATE estimation, assuming  $W \cup X$  is a sufficient adjustment set means that we are assuming that  $W \cup X$  satisfies the backdoor criterion / gives us unconfoundedness. This unconfoundedness assumption gives us parametric identification<sup>2</sup> and allows us to focus on estimation in this chapter.

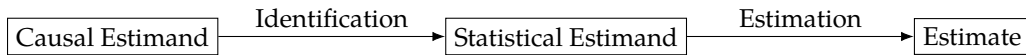
<sup>2</sup> By “parametric identification,” we mean identification under the parametric assumptions of our statistical models. For example, these assumptions are for extrapolation if we don’t have positivity.

## 7.2 Conditional Outcome Modeling (COM)

We are interested in estimating the ATE  $\tau$ . We’ll start with recalling the adjustment formula (Theorem 2.1), which can be derived as a corollary of the backdoor adjustment (Theorem 4.2), as we saw in Section 4.4.1:

$$\tau \triangleq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_W [\mathbb{E}[Y | T = 1, W] - \mathbb{E}[Y | T = 0, W]] \quad (7.4)$$

On the left-hand side of Equation 7.4, we have a causal estimand, and on the right-hand side, we have a statistical estimand (i.e. we have identified this causal quantity). Then, the next step in the Identification-Estimation Flowchart (see Figure 7.1 reproduced from Section 2.4) is to get an estimate of this (statistical) estimand.



**Figure 7.1:** The Identification-Estimation Flowchart – a flowchart that illustrates the process of moving from a target causal estimand to a corresponding estimate, through identification and estimation.

The most straightforward thing to do is to just fit a statistical model (machine learning model) to the conditional expectation  $\mathbb{E}[Y | T, W]$  and then approximate  $\mathbb{E}_W$  with an empirical mean over the  $n$  data points ( $\frac{1}{n} \sum_i$ ). And this is exactly what we did in the simple examples of estimation in Sections 2.5 and 4.6.2. To make this more clear, we introduce  $\mu$  in place of this conditional expectation:

$$\mu(1, w) - \mu(0, w) \triangleq \mathbb{E}[Y | T = 1, W = w] - \mathbb{E}[Y | T = 0, W = w] \quad (7.5)$$

Then, we can fit a statistical model to  $\mu$ . We will denote that these fitted models are approximations of  $\mu$  with a hat:  $\hat{\mu}$ . We will refer to a model  $\hat{\mu}$  as a *conditional outcome model*. Now, we can cleanly write the model-assisted estimator (for the ATE) that we’ve described:

$$\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i)) \quad (7.6)$$

We will refer to estimators that take this form as *conditional outcome model (COM) estimators*. Because minimizing the mean-squared error (MSE) of predicting  $Y$  from  $(T, X)$  pairs is equivalent to modeling this conditional expectation [see, e.g., 10, Section 2.4], there are many different models we can use for  $\hat{\mu}$  in Equation 7.6 to get a COM estimator (see, e.g., scikit-learn [29]).

For CATE estimation, because we assumed that  $W \cup X$  is a sufficient adjustment set, rather than just  $W$ ,<sup>3</sup> we must additionally add  $X$  as

**Active reading exercise:** What are the two different approximations we make in this estimator and what parts of the statistical estimand in Equation 7.4 do each of them replace?

[10]: Hastie et al. (2001), *The Elements of Statistical Learning*

[29]: Pedregosa et al. (2011), ‘Scikit-learn: Machine Learning in Python’

<sup>3</sup> **Active reading exercise:** Why do we additionally add  $X$  to the adjustment set when we are interested in CATEs?



an input to our conditional outcome model. More precisely, for CATE estimation, we define  $\mu$  as follows:

$$\mu(t, w, x) \triangleq \mathbb{E}[Y \mid T = t, W = w, X = x] \quad (7.7)$$

Then, we train a statistical model  $\hat{\mu}$  to predict  $Y$  from  $(T, W, X)$ . And this gives us the following COM estimator for the CATE  $\tau(x)$ :

$$\hat{\tau}(x) = \frac{1}{n_x} \sum_{i: x_i = x} (\hat{\mu}(1, w_i, x) - \hat{\mu}(0, w_i, x)) \quad (7.8)$$

where  $n_x$  is the number of data points that have  $x_i = x$ . When we are interested in the IATE (CATE where  $X$  is all of the observed covariates),  $n_x$  is often 1, which simplifies our estimator to a simple difference between predictions:

$$\hat{\tau}(x_i) = \hat{\mu}(1, w_i, x_i) - \hat{\mu}(0, w_i, x_i) \quad (7.9)$$

Even, though IATEs are different from ITEs ( $\tau(x_i) \neq \tau_i$ ), if we really want to give estimates for ITEs, it is relatively common to take this estimator as our estimator of the ITE  $\tau_i$  as well:

$$\hat{\tau}_i = \hat{\tau}(x_i) = \hat{\mu}(1, w_i, x_i) - \hat{\mu}(0, w_i, x_i) \quad (7.10)$$

Though, this will likely be unreliable due to severe positivity violation.<sup>4</sup>

**The Many-Faced Estimator** COM estimators have many different names in the literature. For example, they are often called G-computation estimators, parametric G-formula, or standardization in epidemiology and biostatistics. Because we are fitting a single statistical model for  $\mu$  here, “COM estimator” is sometimes referred to as an “S-learner,” where the “S” stands for “single.”

**Active reading exercise:** Write down the causal estimand and statistical estimand that lead us to the estimator in Equation 7.8, and proof that they’re equal under unconfoundedness and positivity. In other words, identify the CATE.

<sup>4</sup> **Active reading exercise:** Why is there a severe positivity violation here? Does this only apply in Equation 7.10 or also in Equation 7.9? What if there were multiple units with  $x_i = x$ ?

## 7.3 Grouped Conditional Outcome Modeling (GCOM)

In order to get the estimate in Equation 7.6, we must train a model that predicts  $Y$  from  $(T, W)$ . However,  $T$  is often one-dimensional, whereas  $W$  can be high-dimensional. But the input to  $\hat{\mu}$  for  $t$  is the only thing that changes between the two terms inside the sum  $\hat{\mu}(1, w_i) - \hat{\mu}(0, w_i)$ . Imagine concatenating  $T$  to a 100-dimensional vector  $W$  and then feeding that through a neural network that we’re using for  $\hat{\mu}$ . It seems reasonable that the network could ignore  $T$  while focusing on the other 100 dimensions of its input. This would result in an ATE estimate of zero. And, indeed, there is some evidence of COM estimators being biased toward zero [30].

So how can we ensure that the model  $\hat{\mu}$  doesn’t ignore  $T$ ? Well, we can just train two different models  $\hat{\mu}_1(w)$  and  $\hat{\mu}_0(w)$  that model  $\mu_1(w)$  and  $\mu_0(w)$ , respectively, where

$$\mu_1(w) \triangleq \mathbb{E}[Y \mid T = 1, W = w] \quad \text{and} \quad \mu_0(w) \triangleq \mathbb{E}[Y \mid T = 0, W = w]. \quad (7.11)$$

Using two separate models for the values of treatment ensures that  $T$  cannot be ignored. To train these statistical models, we first group the

[30]: Künzel et al. (2019), ‘Metalearners for estimating heterogeneous treatment effects using machine learning’

data into a group where  $T = 1$  and a group where  $T = 0$ . Then, we train  $\hat{\mu}_1(w)$  to predict  $Y$  from  $W$  in the group where  $T = 1$ . And, similarly, we train  $\hat{\mu}_0(w)$  to predict  $Y$  from  $W$  in the group where  $T = 0$ . This gives us a natural derivative of COM estimators (Equation 7.6), *grouped conditional outcome model (GCOM) estimators*:<sup>5</sup>

$$\hat{\tau} = \frac{1}{n} \sum_i (\hat{\mu}_1(w_i) - \hat{\mu}_0(w_i)) \quad (7.12)$$

And just as we saw, in Equation 7.8, we can add  $X$  as an input to  $\hat{\mu}_1$  and  $\hat{\mu}_0$  to get a GCOM estimator for the CATE  $\tau(x)$ :

$$\hat{\tau}(x) = \frac{1}{n_x} \sum_{i: x_i=x} (\hat{\mu}_1(w_i, x) - \hat{\mu}_0(w_i, x)) \quad (7.13)$$

While GCOM estimation seems to fix the problem that COM estimation can have regarding bias toward zero treatment effect, it does have an important downside. In COM estimation, we were able to make use of all the data when we estimate the single model  $\hat{\mu}$ . However, in *grouped conditional outcome model* estimation, we only use the  $T = 1$  group to estimate  $\hat{\mu}_1$ , and we only use the  $T = 0$  group to estimate  $\hat{\mu}_0$ . Importantly, we are missing out on making the most of our data by not using all of the data to estimate  $\hat{\mu}_1$  and all of the data to estimate  $\hat{\mu}_0$ .

<sup>5</sup> Künzel et al. [30] call a GCOM estimator a “T-learner” where the “T” is for “two” because it requires fitting two different models:  $\hat{\mu}_1$  and  $\hat{\mu}_0$ .

## 7.4 Increasing Data Efficiency

In this section, we’ll cover two ways to address the problem of data efficiency that we mentioned is present in GCOM estimation at the end of the last section: TARNet (Section 7.4.1) and X-Learner (Section 7.4.2).

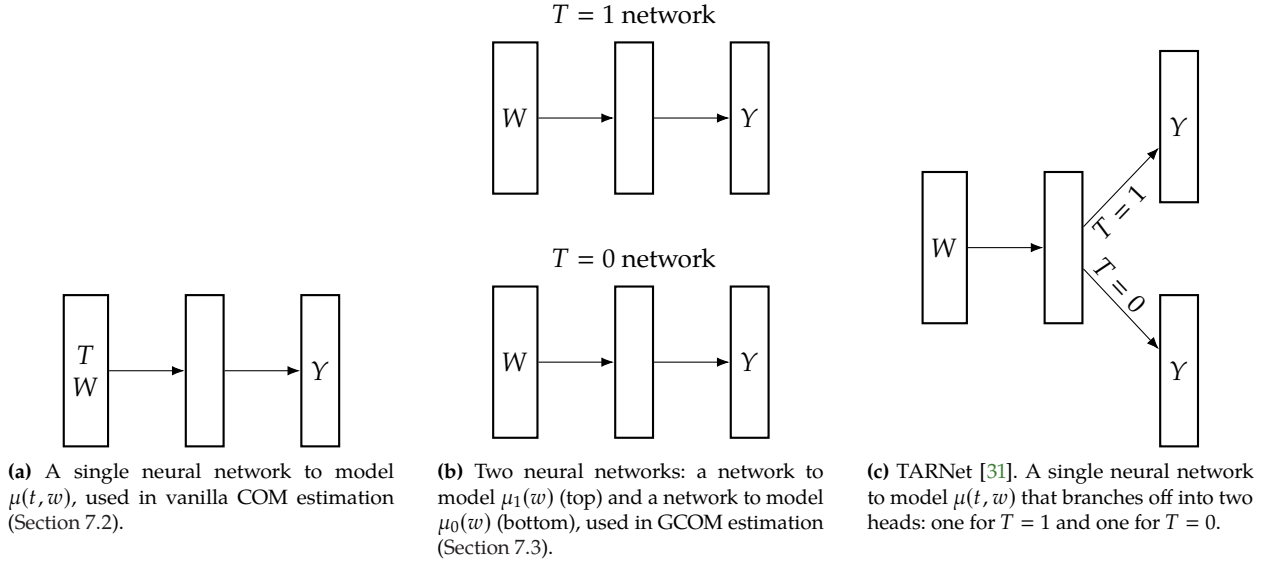
### 7.4.1 TARNet

Consider that we’re using neural networks for our statistical models; starting with that, we’ll contrast, vanilla COM estimation, GCOM estimation, and TARNet. In vanilla COM estimation, the neural network is used to predict  $Y$  from  $(T, W)$  (see Figure 7.2a). This has the problem of potentially yielding ATE estimates that are biased toward zero, as the network might ignore the scalar  $T$ , especially when  $W$  is high-dimensional. To ensure that  $T$  can’t be ignored in GCOM estimation by using two separate neural networks for the two treatment groups (Figure 7.2b). However, this is inefficient as we only use the treatment group data for training one network and the control group data for training the other network.

We can achieve a middle ground between vanilla COM estimation and GCOM estimation using Shalit et al. [31]’s TARNet. With TARNet, we use a single network that takes only  $W$  as input but then branches off into two separate heads (sub-networks) for each treatment group. We then use this model for  $\mu(t, w)$  to get a COM estimator. This has the advantage of learning a treatment-agnostic representation (TAR) of  $W$  using *all* of the data while still forcing the model to not ignore  $T$  by branching into two heads for the different values of  $T$ . In other words, TARNet uses the knowledge we have about  $T$  (as a uniquely important variable) in

[31]: Shalit et al. (2017), ‘Estimating individual treatment effect: generalization bounds and algorithms’

its architecture. Still, the sub-networks for each of these heads are only trained with the data for the corresponding treatment group, rather than all of the data.<sup>6</sup>



**Figure 7.2:** Coarse neural networks architectures for vanilla COM estimation (left), GCOM estimation (middle), and TARNet (right). In this figure, we use each arrow to denote a sub-network that has an arbitrary number of layers.

## 7.4.2 X-Learner

We just saw that one way to increase data efficiency relative to GCOM estimation is to use TARNet, a COM estimator that shares some qualities with GCOM estimators. However, TARNet still doesn't use all of the data for the full model (neural network). In this section, we will start with GCOM estimation and build on it to create a class of estimators that use all of the data for both models that are part of the estimators. An estimator in this class is known as an X-learner [30]. Unlike TARNet, X-learners are neither COM estimators nor GCOM estimators.

There are three steps to X-learning, and the **first step** is the exact same as what's used in GCOM estimation: estimate  $\hat{\mu}_1(x)$  using the treatment group data and estimate  $\hat{\mu}_0(x)$  using the control group data.<sup>7</sup> As before, this can be done with any models that minimize MSE. For simplicity, in this section, we'll be considering IATEs ( $X$  is all of the observed variables) where  $X$  satisfies the backdoor criterion ( $X$  contains  $W$  and no descendants of  $T$ ).

The **second step** is the most important part as it is both where we end up using all of the data for both models and where the "X" comes from. We specify  $\hat{\tau}_{1,i}$  for the treatment group ITE estimates and  $\hat{\tau}_{0,i}$  for the control group ITE estimates:

$$\hat{\tau}_{1,i} = Y_i(1) - \hat{\mu}_0(x_i) \quad (7.14)$$

$$\hat{\tau}_{0,i} = \hat{\mu}_1(x_i) - Y_i(0) \quad (7.15)$$

Here,  $\hat{\tau}_{1,i}$  is estimated using the treatment group outcomes and the imputed counterfactual that we get from  $\hat{\mu}_0$ . Similarly,  $\hat{\tau}_{0,i}$  is estimated

<sup>6</sup> **Active reading exercise:** Which parts of TARNet are like Figure 7.2a and which parts are like Figure 7.2b? What advantages/disadvantages do Figures 7.2a to 7.2c have relative to each other?

[30]: Künzel et al. (2019), 'Metalearners for estimating heterogeneous treatment effects using machine learning'

<sup>7</sup> Recall that  $\hat{\mu}_1(w)$  and  $\hat{\mu}_0(w)$  are approximations of  $\mathbb{E}[Y \mid T = 1, W = w]$  and  $\mathbb{E}[Y \mid T = 0, W = w]$ , respectively.

using the control group outcomes and the imputed counterfactual that we get from  $\hat{\mu}_1$ . If you draw a line between the observed potential outcomes and a line between the imputed potential outcomes, you can see the “X” shape. Importantly, this “X” tells us that each treatment group ITE estimate  $\hat{\tau}_{1,i}$  uses both treatment group data (its observed potential outcome under treatment), *and* control group data (in  $\hat{\mu}_0$ ). Similarly,  $\hat{\tau}_{0,i}$  is estimated with data from both treatment groups.

However, each ITE estimate only uses a single data point from its corresponding treatment group. We can fix this by fitting a model  $\hat{\tau}_1(x)$  to predict  $\hat{\tau}_{1,i}$  from the corresponding treatment group  $x_i$ ’s. Finally, we have a model  $\hat{\tau}_1(x)$  that was fit using all of the data (treatment group data just now and control group data when  $\mu_0$  was fit in step 1). Similarly, we can fit a model  $\hat{\tau}_0(x)$  to predict  $\hat{\tau}_{0,i}$  from the corresponding *control* group  $x_i$ ’s. The **output of step 2** is two different estimators for the IATE:  $\hat{\tau}_1(x)$  and  $\hat{\tau}_0(x)$ .

Finally, in **step 3**, we combine  $\hat{\tau}_1(x)$  and  $\hat{\tau}_0(x)$  together to get our IATE estimator:

$$\hat{\tau}(x) = g(x) \hat{\tau}_0(x) + (1 - g(x)) \hat{\tau}_1(x) \quad (7.16)$$

where  $g(x)$  is some weighting function. Künzel et al. [30] report that an estimate of the propensity score (introduced in next section) works well, but that choosing the constant function 0 or 1 also makes sense if the treatment groups are very different sizes. Or that choosing  $g(x)$  to minimize the variance of  $\hat{\tau}(x)$  could also be attractive.

## 7.5 Propensity Scores

Given that the vector of variables  $W$  satisfies the backdoor criterion (or, equivalently, that  $(Y(1), Y(0)) \perp\!\!\!\perp T \mid W$ ), we might wonder if it is really necessary to condition on that whole vector to isolate causal association, especially when  $W$  is high-dimensional. It turns out that it isn’t. If  $W$  satisfies unconfoundedness and positivity, then we can actually get away with only conditioning on the *scalar*  $P(T = 1 \mid W)$ . We’ll let  $e(w)$  denote  $P(T = 1 \mid W = w)$ , as we’ll refer to  $e(w)$  as the *propensity score* since it is the propensity for (probability of) receiving treatment given that  $W$  is  $w$ . The magic of being able to condition on the scalar  $e(W)$  in the place of the vector  $W$  is due to Rosenbaum and Rubin [32]’s propensity score theorem:

**Theorem 7.1** (Propensity Score Theorem) *Given positivity, unconfoundedness given  $W$  implies unconfoundedness given the propensity score  $e(W)$ . Equivalently,*

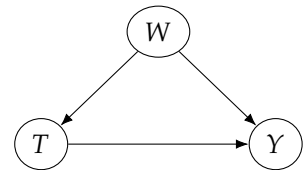
$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W). \quad (7.17)$$

We provide a more traditional mathematical proof in Appendix A.2 and give a graphical proof here. Consider the graph in Figure 7.3. Because the edge from  $W$  to  $T$  is a symbol for the mechanism  $P(T \mid W)$  and because the propensity score completely describes that distribution ( $P(T = 1 \mid W) = e(W)$ ), we can think of the propensity score as a full mediator of the effect of  $W$  on  $T$ . This means that we can redraw this

[30]: Künzel et al. (2019), ‘Metalearners for estimating heterogeneous treatment effects using machine learning’

**Active reading exercise:** In this section, we covered the X-learner for IATE estimation. What would an X-learner for more general CATE estimation ( $X$  is arbitrary and doesn’t necessarily contain all confounders  $W$ ) look like?

[32]: Rosenbaum and Rubin (1983), ‘The central role of the propensity score in observational studies for causal effects’



**Figure 7.3:** Simple graph where  $W$  satisfies the backdoor criterion

graph with  $e(W)$  situated between  $W$  and  $T$ . And in this redrawn graph in Figure 7.4, we can see that  $e(W)$  blocks all backdoor paths that  $W$  blocks, so  $e(W)$  must be a sufficient adjustment set if  $W$  is. Therefore, we have a graphical proof of the propensity score theorem using the backdoor adjustment (Theorem 4.2).

Importantly, this theorem means that we can swap in  $e(W)$  in place of  $W$  wherever we are adjusting for  $W$  in a given estimator in this chapter. For example, this seems very useful when  $W$  is high-dimensional.

Recall The Unconfoundedness-Positivity Tradeoff from Section 2.3.4. As we condition on more non-collider-bias-inducing variables, we decrease confounding. However, this comes at the cost of decreasing overlap because the  $W$  in  $P(T = 1 | W)$  becomes higher and higher dimensional. The propensity score seems to allow us to magically fix that issue since the  $e(W)$  remains a scalar, even as  $W$  grows in dimension. Fantastic, right?

Well, unfortunately, we usually don't have access to  $e(W)$ . Rather, the best we can do is model it. We do this by training a model to predict  $T$  from  $W$ . For example, logistic regression (logit model) is very commonly used to do this. And because this model is fit to the high-dimensional  $W$ , in some sense, we have just shifted the positivity problem to our model for  $e(W)$ .

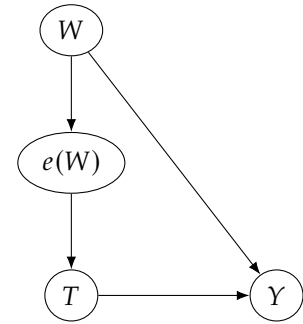


Figure 7.4: Graph illustrating that  $e(W)$  blocks the backdoor path(s) that  $W$  blocks.

## 7.6 Inverse Probability Weighting (IPW)

What if we could resample the data in a way to make it so that association is causation? This is the motivation behind creating “pseudo-populations” that are made up of reweighted versions of the observed population. To get to this, let's recall why association is not causation in general.

Association is not causation in the graph in Figure 7.5 because  $W$  is a common cause of  $T$  and  $Y$ . In other words, the mechanism that generates  $T$  depends on  $W$ , and the mechanism that generates  $Y$  depends on  $W$ . Focusing on the mechanism that generates  $T$ , we can write this mathematically as  $P(T | W) \neq P(T)$ . It turns out that we can reweight the data to get a *pseudo-population* where  $P(T | W) = P(T)$  or  $P(T | W)$  equals some constant; the important part is that we make  $T$  independent of  $W$ . The corresponding graph for such a pseudo-population has no edge from  $W$  to  $T$  because  $T$  does not depend on  $W$ ; we depict this in Figure 7.6.

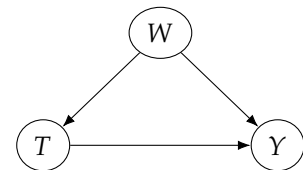


Figure 7.5: Simple graph where  $W$  confounds the effect of  $T$  on  $Y$

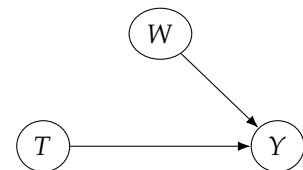


Figure 7.6: Effective graph for pseudo-population that we get by reweighting the data generated according to the graph in Figure 7.5 using inverse probability weighting.

It turns out that the propensity score is key to this reweighting. All we have to do is reweight each data point with treatment  $T$  and confounders  $W$  by its inverse probability of receiving its value of treatment given that it has its value of  $W$ . This is why this technique is called *inverse probability weighting (IPW)*. For individuals that received treatment 1, this weight is  $\frac{1}{e(W)}$ , and for individuals that received treatment 0, this weight is  $\frac{1}{1-e(W)}$ .<sup>8</sup> If the treatment were continuous, the weight would be  $\frac{1}{P(T|W)}$ , which happens to also be the reciprocal of the generalization of the propensity score to continuous treatment.

Why does what we described in the above paragraph work? Well, recall that our goal is to undo confounding by “removing” the edge that goes

<sup>8</sup> **Active reading exercise:** Why is the denominator  $1 - e(W)$  when  $T = 0$ . Hint: recall the precise definition of  $e(W)$ .

from  $W$  to  $T$  (i.e. move from Figure 7.5 to Figure 7.6). And the mechanism that edge describes is  $P(T | W)$ . By weighting the data points by  $\frac{1}{P(T|W)}$ , we are effectively canceling it out. That's the intuition. Formally, we have the following identification equation:

$$\mathbb{E}[Y(t)] = \mathbb{E} \left[ \frac{\mathbb{1}(T = t)Y}{P(t | W)} \right] \quad (7.18)$$

where  $\mathbb{1}(T = t)$  is an indicator random variable that takes on the value 1 if  $T = t$  and 0 otherwise. We provide a proof of Equation 7.18 using the familiar adjustment formula  $\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y | t, W]]$  (Theorem 2.1) in Appendix A.3.

Assuming binary treatment, the following identification equation for the ATE follows from Equation 7.18:

$$\tau \triangleq \mathbb{E}[Y(1) - Y(0)] = \mathbb{E} \left[ \frac{\mathbb{1}(T = 1)Y}{e(W)} \right] - \mathbb{E} \left[ \frac{\mathbb{1}(T = 0)Y}{1 - e(W)} \right] \quad (7.19)$$

Now that we have a statistical estimand in the form of IPW, we can get an IPW estimator. Replacing expectations by empirical means and  $e(W)$  by a propensity score model  $\hat{e}(W)$ , we get the following equivalent formulations of the basic IPW estimator<sup>9</sup> for the ATE:

$$\hat{\tau} = \frac{1}{n} \sum_i \left( \frac{\mathbb{1}(t_i = 1)y_i}{\hat{e}(w_i)} - \frac{\mathbb{1}(t_i = 0)y_i}{1 - \hat{e}(w_i)} \right) \quad (7.20)$$

$$= \frac{1}{n_1} \sum_{i:t_i=1} \frac{y_i}{\hat{e}(w_i)} - \frac{1}{n_0} \sum_{i:t_i=0} \frac{y_i}{1 - \hat{e}(w_i)} \quad (7.21)$$

where  $n_1$  and  $n_0$  are the number of treatment group units and control group units, respectively.

**Weight Trimming** As you can see in Equations 7.20 and 7.21, if the propensity scores are very close to 0 or 1, the estimates will blow up. In order to prevent this, it is not uncommon to trim the propensity scores that are less than  $\epsilon$  to  $\epsilon$  and those that are greater than  $1 - \epsilon$  to  $1 - \epsilon$  (effectively trimming the weights to be no larger than  $\frac{1}{\epsilon}$ ), though this introduces its own problems such as bias.

**CATE Estimation** We can extend the ATE estimator in Equation 7.20 to get an IPW estimator for the CATE  $\tau(x)$  by just restricting to the data points where  $x_i = x$ :

$$\hat{\tau}(x) = \frac{1}{n_x} \sum_{i:x_i=x} \left( \frac{\mathbb{1}(t_i = 1)y_i}{\hat{e}(w_i)} - \frac{\mathbb{1}(t_i = 0)y_i}{1 - \hat{e}(w_i)} \right) \quad (7.22)$$

where  $n_x$  is the number of data points with  $x_i = x$ . However, the estimator in Equation 7.22 may quickly run into the problem of using very small amounts of data, leading to high variance. More general CATE estimation with IPW estimators is more complex and outside the scope of this book. See, for example, Abrevaya et al. [34] and references therein.

<sup>9</sup> This estimator is originally from Horvitz and Thompson [33].

[33]: Horvitz and Thompson (1952), 'A Generalization of Sampling Without Replacement from a Finite Universe'

**Active reading exercise:** What would be the corresponding formulations of the basic IPW estimator for  $\mathbb{E}[Y(t)]$ ?

[34]: Abrevaya et al. (2015), 'Estimating Conditional Average Treatment Effects'



## 7.7 Doubly Robust Methods

We've seen that we can estimate causal effects by modeling  $\mu(t, w) \triangleq \mathbb{E}[Y \mid t, w]$  (Sections 7.2 to 7.4) or by modeling  $e(w) \triangleq P(T = 1 \mid w)$  (Section 7.6). What if we modeled both  $\mu(t, w)$  and  $e(w)$ ? Well, we can and estimators that do this are sometimes *doubly robust*. A doubly robust estimator has the property that it is a consistent<sup>10</sup> estimator of  $\tau$  if either  $\hat{\mu}$  is a consistent estimator of  $\mu$  or  $\hat{e}$  is a consistent estimate of  $e$ . In other words, only one of  $\hat{\mu}$  and  $\hat{e}$  needs to be well-specified. Additionally, the rate at which a doubly robust estimator converges to  $\tau$  is the *product* of the rate at which  $\hat{\mu}$  converges to  $\mu$  and the rate at which  $\hat{e}$  converges to  $e$ . This makes double robustness very useful when we are using flexible machine learning models in high-dimensions because, in this setting, each of our individual models ( $\hat{\mu}$  and  $\hat{e}$ ) converge more slowly than the ideal rate of  $n^{-1/2}$ .

However, there is some controversy over how well doubly robust methods work in practice if not at least one of  $\hat{\mu}$  or  $\hat{e}$  is well-specified [35]. Though, this might be contested as we get better at using doubly robust estimators with flexible machine learning models (see, e.g., [36]). Meanwhile, the estimators that currently seem to do the best all flexibly model  $\mu$  (unlike pure IPW estimators) [37]. This is why we began this chapter with estimators that model  $\mu$  and dedicated several sections to such estimators.

Doubly robust methods are largely outside the scope of this book, so we refer the reader to an introduction by Seaman and Vansteelandt [38], along with other seminal works on the topic: [39–41]. Additionally, there is a large body of doubly robust work on methods that have performed reasonably well in competitions [37]; this category is known as *targeted maximum likelihood estimation (TMLE)*. [42–44].

## 7.8 Other Methods

As this chapter is only an introduction to estimation in causal inference, there are some methods that we've entirely left out. We'll briefly describe some of the most popular ones in this section.

**Matching** In matching methods, we try to match units in the treatment group with units in the control group and throw away the non-matches to create comparable groups. We can match in raw covariate space, coarsened covariate space, or propensity score space. There are different distance functions for deciding how close two units are. Furthermore, there are different criteria for deciding whether a given distance is close enough to count as a match (one criterion requires an exact match), how many matches each treatment group unit can have, how many matches each control group unit can have, etc. See, for example, Stuart [45] for a review.

**Double Machine Learning** In double machine learning, we fit three models in two stages: two in the first stage and a final model in the second stage. First stage:

1. Fit a model to predict  $Y$  from  $W$  to get the predicted  $\hat{Y}$ .<sup>11</sup>

<sup>10</sup> An estimator is consistent if it converges in probability to its estimand as the number of samples  $n$  grows.

[35]: Kang and Schafer (2007), 'Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data'

[36]: Zivich and Breskin (2020), *Machine learning for causal inference: on the use of cross-fit estimators*

[37]: Dorie et al. (2019), 'Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition'

[38]: Seaman and Vansteelandt (2018), 'Introduction to Double Robust Methods for Incomplete Data'

[39]: Tsiatis (2007), *Semiparametric theory and missing data*

[40]: Robins et al. (1994), 'Estimation of Regression Coefficients When Some Regressors are not Always Observed'

[41]: Bang and Robins (2005), 'Doubly Robust Estimation in Missing Data and Causal Inference Models'

[42]: Van Der Laan and Rubin (2006), 'Targeted maximum likelihood learning'

[43]: Schuler and Rose (2017), 'Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies'

[44]: Van der Laan and Rose (2011), *Targeted learning: causal inference for observational and experimental data*

[45]: Stuart (2010), 'Matching Methods for Causal Inference: A Review and a Look Forward'

<sup>11</sup> **Active reading exercise:** How is this model different from  $\hat{\mu}$ ?



2. Fit a model to predict  $T$  from  $W$  to get the predicted  $\hat{T}$ .

Then, in the second stage, we “partial out”  $W$  by looking at  $Y - \hat{Y}$  and  $T - \hat{T}$ . In a sense, we have deconfounded the effect of treatment on the outcome with this partialling out. Then, we fit a model to predict  $Y - \hat{Y}$  from  $T - \hat{T}$ . This gives us our causal effect estimates. For more on this topic, see, for example [46–49].

**Causal Trees and Forests** Another popular estimation method is to recursively partition the data into subsets that have the same treatment effects [50]. This forms a *causal tree* where the leaves are subsets of the population with similar causal effects. Since random forests generally perform better than decision trees, it would be great if this kind of strategy can be extended to random forests. And it can. This extension is known as *causal forests* [51], which are part of more general class known as *generalized random forests* [52]. Importantly, these methods were developed with the goal in mind of yielding valid confidence intervals for the estimates.

## 7.9 Concluding Remarks

### 7.9.1 Confidence Intervals

So far, in this chapter, we have only discussed point estimates for causal effects. We haven’t discussed how we can gauge our uncertainty due to data sampling. We haven’t discussed how to calculate confidence intervals on these estimates. This is a machine learning perspective, after all; who cares about confidence intervals... Jokes aside, because we are allowing for arbitrary machine learning models in all of the estimators we discuss, it is actually quite difficult to get valid confidence intervals.

**Bootstrapping** One way to get confidence intervals is to use bootstrapping. With bootstrapping, we repeat the causal effect estimation process many times, each time with a different sample (with replacement) from our data. This allows us to build an empirical distribution for the estimate. We can then compute whatever confidence interval we like from that empirical distribution. Unfortunately, bootstrapped confidence intervals are not always valid. For example, if we take a bootstrapped 95% confidence interval, it might not contain the true value (estimand) 95% of the time.

**Specialized Models** Another way to get confidence intervals is to analyze very specific models, rather than allowing for arbitrary models. Linear models are the simplest example of this; it is easy to get confidence intervals in linear models. Similarly, if we use a linear model as the second stage model in double machine learning, we can get confidence intervals. Noticeably, causal trees and causal forests were developed with the goal in mind of getting confidence intervals.

### 7.9.2 Comparison to Randomized Experiments

You might read somewhere that some of these adjustment techniques ensure that we’ve addressed confounding and isolated a causal effect. Of course, this is not true when there is unobserved confounding. These

[46]: Chernozhukov et al. (2018), ‘Double/debiased machine learning for treatment and structural parameters’

[47]: Felton (2018), Chernozhukov et al. on *Double / Debiased Machine Learning*

[48]: Syrgkanis (2019), *Orthogonal/Double Machine Learning*

[49]: Foster and Syrgkanis (2019), *Orthogonal Statistical Learning*

[50]: Athey and Imbens (2016), ‘Recursive partitioning for heterogeneous causal effects’

[51]: Wager and Athey (2018), ‘Estimation and Inference of Heterogeneous Treatment Effects using Random Forests’

[52]: Athey et al. (2019), ‘Generalized random forests’

methods only address *observed* confounding. If there are any unobserved confounders, these methods don't fix that like randomization does (Chapter 5). These adjustment methods aren't magic. And it's hard to know when it is reasonable to assume we've observed all confounders. That's why it is important to run a sensitivity analysis where we gauge how robust our causal effect estimates are to unobserved confounding. This is the topic of the next chapter.

**Active reading exercise:** What kind of estimator did we use back in the estimation examples in Sections 2.5 and 4.6.2?

# Counterfactuals

# 8

## 8.1 Coming Soon

8.1 Coming Soon . . . . . 72

**More Chapters Coming**

**9**

# APPENDIX

## A.1 Proof of Equation 6.1 from Section 6.1

**Claim** Given the causal graph is Figure A.1,  $P(m \mid do(t)) = P(m \mid t)$ .

*Proof.* We first apply the Bayesian network factorization (Definition 3.1):

$$P(w, t, m, y) = P(w) P(t \mid w) P(m \mid t) P(y \mid w, m) \quad (\text{A.1})$$

Next, we apply the truncated factorization (Proposition 4.1):

$$P(w, m, y \mid do(t)) = P(w) P(m \mid t) P(y \mid w, m) \quad (\text{A.2})$$

Finally, we marginalize out  $w$  and  $y$ :

$$\sum_w \sum_y P(w, m, y \mid do(t)) = \sum_w \sum_y P(w) P(m \mid t) P(y \mid w, m) \quad (\text{A.3})$$

$$P(m \mid do(t)) = \left( \sum_w P(w) \right) P(m \mid t) \left( \sum_y P(y \mid w, m) \right) \quad (\text{A.4})$$

$$= P(m \mid t) \quad (\text{A.5})$$

□

## A.2 Proof of Propensity Score Theorem (7.1)

**Claim**  $(Y(1), Y(0)) \perp\!\!\!\perp T \mid W \implies (Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$ .

*Proof.* Assuming  $(Y(1), Y(0)) \perp\!\!\!\perp T \mid W$ , we will prove  $(Y(1), Y(0)) \perp\!\!\!\perp T \mid e(W)$  by showing that  $P(T = 1, \mid Y(t), e(W))$  does not depend on  $Y(t)$ , where  $Y(t)$  is either potential outcome.

First, because  $T$  is binary, can turn this probability into an expectation:

$$P(T = 1, \mid Y(t), e(W)) = \mathbb{E}[T \mid Y(t), e(W)] \quad (\text{A.6})$$

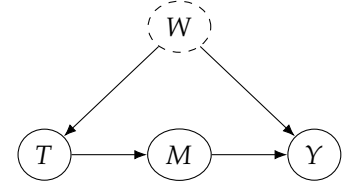
Then, using the law of iterated expectations, we can introduce  $W$ :

$$= \mathbb{E}[\mathbb{E}[T \mid Y(t), e(W), W] \mid Y(t), e(W)] \quad (\text{A.7})$$

A.1 Proof of Equation 6.1 from Section 6.1 . . . . . 75

A.2 Proof of Propensity Score Theorem (7.1) . . . . . 75

A.3 Proof of IPW Estimand (7.18) 76



**Figure A.1:** Causal graph where  $W$  is unobserved, so we cannot block the backdoor path  $T \leftarrow W \rightarrow Y$ .

Because we have now conditioned on all of  $W$  and  $e(W)$  is a function of  $W$ , it is redundant, so we can remove  $e(W)$  from the inner expectation:

$$= \mathbb{E} [\mathbb{E}[T \mid Y(t), W] \mid Y(t), e(W)] \quad (\text{A.8})$$

Now, we apply the unconfoundedness assumption we started with to remove  $Y(t)$  from the inner expectation:

$$= \mathbb{E} [\mathbb{E}[T \mid W] \mid Y(t), e(W)] \quad (\text{A.9})$$

Again, using the fact that  $T$  is binary, we can reduce the inner expectation to  $P(T = 1 \mid W) \triangleq e(W)$ , something that is already conditioned on:

$$= \mathbb{E} [P(T = 1 \mid W) \mid Y(t), e(W)] \quad (\text{A.10})$$

$$= \mathbb{E} [e(W) \mid Y(t), e(W)] \quad (\text{A.11})$$

$$= e(W) \quad (\text{A.12})$$

Because this does not depend on  $Y(t)$ , we've proven that  $T$  is independent of  $Y(t)$  given  $e(W)$ .  $\square$

### A.3 Proof of IPW Estimand (7.18)

**Claim** Under unconfoundedness and positivity,  $\mathbb{E}[Y(t)] = \mathbb{E} \left[ \frac{\mathbb{1}(T=t)Y}{P(t|W)} \right]$ .

*Proof.* We will start with the statistical estimand that we get from the adjustment formula (Theorem 2.1). Given unconfoundedness and positivity, the adjustment formula tells us

$$\mathbb{E}[Y(t)] = \mathbb{E}[\mathbb{E}[Y \mid t, W]] \quad (\text{A.13})$$

We'll assume the variable are discrete to break these expectations into sums (replace with integrals if continuous):

$$= \sum_w \left( \sum_y y P(y \mid t, w) \right) P(w) \quad (\text{A.14})$$

To get  $P(t \mid w)$  in there, we multiply by  $\frac{P(t|w)}{P(t|w)}$ :

$$= \sum_w \sum_y y P(y \mid t, w) P(w) \frac{P(t \mid w)}{P(t \mid w)} \quad (\text{A.15})$$

Then, noticing that  $P(y \mid t, w) P(t \mid w) P(w)$  is the joint distribution:

$$= \sum_w \sum_y y P(y, t, w) \frac{1}{P(t \mid w)} \quad (\text{A.16})$$



$\sum_y y P(y, t, w)$  is nearly  $\sum_y y P(y) = \mathbb{E}[Y]$ , but because of  $T = t$  and  $W = w$  are in the probability, the terms of this sum are only non-zero if  $T = t$  and  $W = w$ . Therefore, we get the indicator random variable for this event in the expectation that is over all three random variables ( $T$ ,  $W$ , and  $Y$ ):

$$= \sum_w \mathbb{E} [\mathbb{1}(T = t, W = w) Y] \frac{1}{P(t | w)} \quad (\text{A.17})$$

Now, the  $\sum_w \frac{1}{P(t | w)}$  that remains is a weighted expectation over  $W$ . Integrating this means that because we are now marginalizing over  $W$ ,  $w$  becomes a random variable ( $W$ ) and the  $W = w$  inside the indicator becomes redundant. This gives us the following:

$$= \mathbb{E} \left[ \frac{\mathbb{1}(T = t) Y}{P(t | W)} \right] \quad (\text{A.18})$$

□

Note: For some people, it might be more natural to skip straight from Equation A.16 to Equation A.18.

# Bibliography

Here are the references in citation order.

- [1] Tyler Vigen. *Spurious correlations*. <https://www.tylervigen.com/spurious-correlations>. 2015 (cited on page 3).
- [2] Jerzy Splawa-Neyman. 'On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.' Trans. by D. M. Dabrowska and T. P. Speed. In: *Statistical Science* 5.4 (1923 [1990]), pp. 465–472 (cited on page 6).
- [3] Donald B. Rubin. 'Estimating causal effects of treatments in randomized and nonrandomized studies.' In: *Journal of educational Psychology* 66.5 (1974), p. 688 (cited on pages 6, 7).
- [4] Jasjeet S. Sekhon. 'The Neyman-Rubin Model of Causal Inference and Estimation via Matching Methods'. In: *Oxford handbook of political methodology* (2008), pp. 271– (cited on page 6).
- [5] Paul W. Holland. 'Statistics and Causal Inference'. In: *Journal of the American Statistical Association* 81.396 (1986), pp. 945–960. doi: [10.1080/01621459.1986.10478354](https://doi.org/10.1080/01621459.1986.10478354) (cited on pages 8, 42).
- [6] Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. *Overlap in Observational Studies with High-Dimensional Covariates*. 2017 (cited on page 13).
- [7] Miguel A Hernán and James M Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020 (cited on pages 14, 27).
- [8] Miguel Angel Luque-Fernandez, Michael Schomaker, Daniel Redondo-Sanchez, Maria Jose Sanchez Perez, Anand Vaidya, and Mireille E Schnitzer. 'Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application'. In: *International Journal of Epidemiology* 48.2 (Dec. 2018), pp. 640–653. doi: [10.1093/ije/dyy275](https://doi.org/10.1093/ije/dyy275) (cited on pages 16, 45).
- [9] Salim S. Virani et al. 'Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association'. In: *Circulation* (Mar. 2020), pp. 640–653. doi: [10.1161/cir.0000000000000757](https://doi.org/10.1161/cir.0000000000000757) (cited on page 16).
- [10] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001 (cited on pages 17, 62).
- [11] Dominik Janzing, David Balduzzi, Moritz Grosse-Wentrup, and Bernhard Schölkopf. 'Quantifying causal influences'. In: *Ann. Statist.* 41.5 (Oct. 2013), pp. 2324–2358. doi: [10.1214/13-AOS1145](https://doi.org/10.1214/13-AOS1145) (cited on page 18).
- [12] Stephen L. Morgan and Christopher Winship. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. 2nd ed. Analytical Methods for Social Research. Cambridge University Press, 2014 (cited on page 18).
- [13] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive Computation and Machine Learning. The MIT Press, 2009 (cited on pages 21, 29, 47).
- [14] J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017 (cited on page 21).
- [15] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016 (cited on page 25).
- [16] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988 (cited on page 29).
- [17] Judea Pearl. 'Causal inference in statistics: An overview'. In: *Statist. Surv.* 3 (2009), pp. 96–146. doi: [10.1214/09-SS057](https://doi.org/10.1214/09-SS057) (cited on page 42).

- [18] Judea Pearl. *Causality*. Cambridge University Press, 2009 (cited on pages 44, 47, 48, 59).
- [19] Felix Elwert and Christopher Winship. 'Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable.' In: *Annual review of sociology* 40 (2014), pp. 31–53 (cited on page 44).
- [20] David Galles and Judea Pearl. 'An Axiomatic Characterization of Causal Counterfactuals'. In: *Foundations of Science* 3.1 (1998), pp. 151–182. doi: [10.1023/A:1009602825894](https://doi.org/10.1023/A:1009602825894) (cited on page 47).
- [21] Joseph Y. Halpern. 'Axiomatizing Causal Reasoning'. In: *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*. UAI'98. Madison, Wisconsin: Morgan Kaufmann Publishers Inc., 1998, pp. 202–210 (cited on page 47).
- [22] Elizabeth L. Ogburn and Tyler J. VanderWeele. 'Causal Diagrams for Interference'. In: *Statist. Sci.* 29.4 (Nov. 2014), pp. 559–578. doi: [10.1214/14-STS501](https://doi.org/10.1214/14-STS501) (cited on page 48).
- [23] J. Pearl. 'On the consistency rule in causal inference: axiom, definition, assumption, or theorem?' In: *Epidemiology* 21.6 (Nov. 2010), pp. 872–875 (cited on page 48).
- [24] Judea Pearl. 'Causal diagrams for empirical research'. In: *Biometrika* 82.4 (Dec. 1995), pp. 669–688. doi: [10.1093/biomet/82.4.669](https://doi.org/10.1093/biomet/82.4.669) (cited on pages 55, 56).
- [25] Ilya Shpitser and Judea Pearl. 'Identification of Joint Interventional Distributions in Recursive Semi-Markovian Causal Models'. In: *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2*. AAAI'06. Boston, Massachusetts: AAAI Press, 2006, pp. 1219–1226 (cited on pages 57, 60).
- [26] Yimin Huang and Marco Valtorta. 'Pearl's Calculus of Intervention is Complete'. In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. UAI'06. Cambridge, MA, USA: AUAI Press, 2006, pp. 217–224 (cited on page 57).
- [27] Jin Tian and Judea Pearl. 'A General Identification Condition for Causal Effects'. In: *Eighteenth National Conference on Artificial Intelligence*. Edmonton, Alberta, Canada: American Association for Artificial Intelligence, 2002, pp. 567–573 (cited on page 59).
- [28] Ilya Shpitser and Judea Pearl. 'Identification of Conditional Interventional Distributions'. In: *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. UAI'06. Cambridge, MA, USA: AUAI Press, 2006, pp. 437–444 (cited on page 60).
- [29] F. Pedregosa et al. 'Scikit-learn: Machine Learning in Python'. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cited on pages 61, 62).
- [30] Sören R. Künzle, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 'Metalearners for estimating heterogeneous treatment effects using machine learning'. In: *Proceedings of the National Academy of Sciences* 116.10 (2019), pp. 4156–4165. doi: [10.1073/pnas.1804597116](https://doi.org/10.1073/pnas.1804597116) (cited on pages 63–66).
- [31] Uri Shalit, Fredrik D. Johansson, and David Sontag. 'Estimating individual treatment effect: generalization bounds and algorithms'. In: ed. by Doina Precup and Yee Whye Teh. Vol. 70. *Proceedings of Machine Learning Research*. International Convention Centre, Sydney, Australia: PMLR, June 2017, pp. 3076–3085 (cited on pages 64, 65).
- [32] Paul R. Rosenbaum and Donald B. Rubin. 'The central role of the propensity score in observational studies for causal effects'. In: *Biometrika* 70.1 (Apr. 1983), pp. 41–55. doi: [10.1093/biomet/70.1.41](https://doi.org/10.1093/biomet/70.1.41) (cited on page 66).
- [33] D. G. Horvitz and D. J. Thompson. 'A Generalization of Sampling Without Replacement from a Finite Universe'. In: *Journal of the American Statistical Association* 47.260 (1952), pp. 663–685. doi: [10.1080/01621459.1952.10483446](https://doi.org/10.1080/01621459.1952.10483446) (cited on page 68).
- [34] Jason Abrevaya, Yu-Chin Hsu, and Robert P. Lieli. 'Estimating Conditional Average Treatment Effects'. In: *Journal of Business & Economic Statistics* 33.4 (2015), pp. 485–505. doi: [10.1080/07350015.2014.975555](https://doi.org/10.1080/07350015.2014.975555) (cited on page 68).
- [35] Joseph D. Y. Kang and Joseph L. Schafer. 'Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data'. In: *Statist. Sci.* 22.4 (Nov. 2007), pp. 523–539. doi: [10.1214/07-STS227](https://doi.org/10.1214/07-STS227) (cited on page 69).
- [36] Paul N Zivich and Alexander Breskin. *Machine learning for causal inference: on the use of cross-fit estimators*. 2020 (cited on page 69).

- [37] Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. 'Automated versus Do-It-Yourself Methods for Causal Inference: Lessons Learned from a Data Analysis Competition'. In: *Statist. Sci.* 34.1 (Feb. 2019), pp. 43–68. doi: [10.1214/18-STS667](#) (cited on page 69).
- [38] Shaun R. Seaman and Stijn Vansteelandt. 'Introduction to Double Robust Methods for Incomplete Data'. In: *Statist. Sci.* 33.2 (May 2018), pp. 184–197. doi: [10.1214/18-STS647](#) (cited on page 69).
- [39] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007 (cited on page 69).
- [40] James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. 'Estimation of Regression Coefficients When Some Regressors are not Always Observed'. In: *Journal of the American Statistical Association* 89.427 (1994), pp. 846–866. doi: [10.1080/01621459.1994.10476818](#) (cited on page 69).
- [41] Heejung Bang and James M. Robins. 'Doubly Robust Estimation in Missing Data and Causal Inference Models'. In: *Biometrics* 61.4 (2005), pp. 962–973. doi: [10.1111/j.1541-0420.2005.00377.x](#) (cited on page 69).
- [42] Mark J Van Der Laan and Daniel Rubin. 'Targeted maximum likelihood learning'. In: *The international journal of biostatistics* 2.1 (2006) (cited on page 69).
- [43] Megan S. Schuler and Sherri Rose. 'Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies'. In: *American Journal of Epidemiology* 185.1 (Jan. 2017), pp. 65–73. doi: [10.1093/aje/kww165](#) (cited on page 69).
- [44] Mark J Van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011 (cited on page 69).
- [45] Elizabeth A. Stuart. 'Matching Methods for Causal Inference: A Review and a Look Forward'. In: *Statist. Sci.* 25.1 (Feb. 2010), pp. 1–21. doi: [10.1214/09-STS313](#) (cited on page 69).
- [46] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 'Double/debiased machine learning for treatment and structural parameters'. In: *The Econometrics Journal* 21.1 (2018), pp. C1–C68. doi: [10.1111/ectj.12097](#) (cited on page 70).
- [47] Chris Felton. Chernozhukov et al. on Double / Debiased Machine Learning. [https://scholar.princeton.edu/sites/default/files/bstewart/files/felton.chern\\_.slides.20190318.pdf](https://scholar.princeton.edu/sites/default/files/bstewart/files/felton.chern_.slides.20190318.pdf). 2018 (cited on page 70).
- [48] Vasilis Syrgkanis. *Orthogonal/Double Machine Learning*. <https://econml.azurewebsites.net/spec/estimation/dml.html>. Accessed: 16 September 2020. 2019 (cited on page 70).
- [49] Dylan J. Foster and Vasilis Syrgkanis. *Orthogonal Statistical Learning*. 2019 (cited on page 70).
- [50] Susan Athey and Guido Imbens. 'Recursive partitioning for heterogeneous causal effects'. In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7353–7360. doi: [10.1073/pnas.1510489113](#) (cited on page 70).
- [51] Stefan Wager and Susan Athey. 'Estimation and Inference of Heterogeneous Treatment Effects using Random Forests'. In: *Journal of the American Statistical Association* 113.523 (2018), pp. 1228–1242. doi: [10.1080/01621459.2017.1319839](#) (cited on page 70).
- [52] Susan Athey, Julie Tibshirani, and Stefan Wager. 'Generalized random forests'. In: *Ann. Statist.* 47.2 (Apr. 2019), pp. 1148–1178. doi: [10.1214/18-AOS1709](#) (cited on page 70).

# Alphabetical Index

- adjustment formula, 11
- ancestor, 19
- association, 4
  - causal association, 4, 30
  - confounding association, 4, 30
- associational difference, 8
- average treatment effect (ATE), 8, 61
- backdoor adjustment, 38
- backdoor criterion, 37
- backdoor path, 37
- Bayesian network, 21
  - chain rule, 21
  - factorization, 21
- Berkson's paradox, 27
- blocked path, 25, 26, 28
- causal Bayesian networks, 35
- causal effect
  - average, 8, 61
  - conditional average, 61
  - individual, 7, 61
  - individualized average, 61
  - unit-level, 7
- causal estimand, 15, 33
- causal forests, 70
- causal graph, 22
  - non-strict, 23
  - strict, 22
- causal mechanism, 34, 41
- causal tree, 70
- cause, 22, 41
- child, 19
- collider, 26
- collider bias, 43
- COM estimator, 62
- common cause, 4
- common support, 13
- comparability, 49
- conditional average treatment effect (CATE), 61
- conditional outcome model, 62
- confounder, 4
- consistent, 69
- correlation, 4
- counterfactual, 8
- covariate balance, 49
- curse of dimensionality, 13
- cycle, 19
- d-connected, 29
- d-separated, 29
- d-separation, 29
- data generating process, 27
- descendant, 19
- direct cause, 22, 41
- directed acyclic graph (DAG), 19
- directed graph, 19
- directed path, 19
- do-calculus, 55
- do-operator, 32
- double robustness, 69
- doubly robust, 69
- edge, 19
- endogenous, 41
- Equationtown, 53
- estimand, 15
  - causal, 15, 33
  - statistical, 15, 33
- estimate, 15
- estimation, 15, 16
- estimator, 15
- exchangeability, 9, 50
- exogenous, 41
- extrapolation, 13
- factual, 8
- frontdoor adjustment, 53
- frontdoor criterion, 53
- generalized random forests, 70
- global Markov assumption, 29
- graph, 19
- hedge criterion, 60
- identifiability, 10, 33
- identification, 10, 16, 33
  - nonparametric, 57
  - parametric, 57, 62
- ignorability, 9
- immorality, 20, 26
- individual treatment effect (ITE), 7, 61
- individualized average treatment effect (IATE), 61
- interference, 13
- interventional distribution, 32
- interventional SCM, 42
- inverse probability weighting, 67
- IPW, 67
- local Markov assumption, 20
- lurking variable, 4
- M-bias, 44, 47
- magnification, 43
- magnify, 43
- manipulated graph, 34, 55
- Markov compatibility, 21
- Markovian, 41
- mediator, 43
- minimality, 21
- misspecification, 18
- model-assisted estimation, 16
- model-assisted estimator, 16, 17
- node, 19
- non-Markovian, 41
- nonparametric, 40
- nonparametric identification, 57
- observational data, 33
- observational distribution, 33
- overlap, 13
- parametric identification, 57, 62
- parent, 19
- path, 19
  - blocked, 25, 26
  - blocked , 28

- unblocked, 25, 27
- unblocked , 28
- positivity, 12
- post-intervention, 33
- post-treatment covariates,
  - 44
- potential outcome, 6
- pre-intervention, 33
- propensity score, 66
- pseudo-populations, 67
- randomized control trials
  - (RCTs), 49
- randomized experiments,
  - 49

- semi-Markovian, 41
- Simpson's paradox, 1
- spurious correlation, 3
- statistical estimand, 15, 33
- structural causal model (SCM),
  - 40
- structural equation, 40
- sufficient adjustment set, 38
- SUTVA, 14
- targeted maximum likelihood
  - estimation, 69
- TARNet, 64
- terminology machine gun,
  - 19

- TMLE, 69
- treatment assignment
  - mechanism, 49
- truncated factorization, 35
- unblocked path, 25, 27, 28
- unconfounded children
  - criterion, 59
- unconfoundedness, 11
- undirected graph, 19
- unit-level treatment effect, 7
- vertex, 19
- X-learner, 65