figure out theoremstle is undefined

Learning how to act: making good decisions with machine learning

Finnian Lattimore

July 9, 2017

In vain the grave, with retrospective Eye,
Would from the apparent what conclude
the why, Infer the Motive from the Deed,
and show That what we chanced, was
what we meant, to do.

Alexander Pope

Contents

1 Introduction **5**
- 1.1 Motivation . . . 5
- 1.2 What is causality and why do we care? . . . 6
  - 1.2.1 Defining causality . . . 7
- 1.3 Identifying when we have a causal problem . . . 7
- 1.4 Approaches to causality . . . 10

2 Causal models **12**
- 2.1 Causal Bayesian networks . . . 12
- 2.2 Counterfactuals . . . 17
- 2.3 Structural Equation models . . . 20
- 2.4 Comparing and unifying the models . . . 22
- 2.5 What does a causal model give us? Resolving Simpson's paradox . . . 23

3 Two key questions **26**
- 3.1 Causal effect estimation . . . 26
  - 3.1.1 Independence in Bayesian networks: D-separation . . . 27
- 3.2 The Do Calculus . . . 29
  - 3.2.1 Identifiability . . . 31
- 3.3 Estimation . . . 32
  - 3.3.1 Defining causal effects . . . 32
  - 3.3.2 Estimating causal effects by adjusting for confounding variables . . . 33
- 3.4 Causal Discovery . . . 39
  - 3.4.1 Conditional independence based methods . . . 40
  - 3.4.2 Discovery with functional models . . . 44

4 The interventionist viewpoint **46**
- 4.1 Randomised experiments . . . 46
  - 4.1.1 Limitations of randomised experiments . . . 47
- 4.2 Multi armed bandits . . . 49
  - 4.2.1 Stochastic bandits: Approaches and results . . . 50
  - 4.2.2 Pure-exploration problems . . . 53
  - 4.2.3 Adversarial Bandits . . . 54
  - 4.2.4 Contextual bandits . . . 55
  - 4.2.5 Learning from logged bandit data . . . 59
  - 4.2.6 Adding structure to actions . . . 59

5 Causal Bandits: Unifying the approaches **61**
- 5.1 The framework . . . 61
- 5.2 Causal bandits with post action feedback . . . 63
  - 5.2.1 The parallel bandit problem . . . 64

5.2.2 General graphs 66
5.2.3 Experiments 68
5.2.4 Discussion & Future work 71
5.2.5 Proofs 72

Todo

1. Add something about Granger causality - does this belong in chapter 2 or chapter 3?
2. Sell my PhD as two key contribution. 1) A technical one, the other educational and unifying (across viewpoints)
3. Add a chapter in 'PhD prose' making some statement about the intersection of prediction and causality. When is a problem causal, what bounds does predictive accuracy give on causal models? To what degree is predictive accuracy a metric on which causal models can be judged?
4. Benchmark estimation algorithms
5. clarify causal interpretation of Dirichlet models
6. Empirical evaluation of bandit algorithms (updated to include latest suggestions). Maybe a proper software bench marking solution?
7. If comparing bandit performance, it would be interesting to include some that violate the stochastic bandit assumption. For example, if the rewards change slowly over time or jump suddenly a limited number of times, or with some small probability.

Chapter 1

Introduction

My thesis in a sentence: Unifying causal inference with multi-armed bandits.

This thesis contributes to knowledge by: Introducing a framework connecting causal graphical models with multi-armed bandits as a first step towards a unified approach to estimating the effect of interventions.

My key research questions are:

- To understand and the difference between prediction and causal inference in machine learning and clarify which problems require causal approaches.
- To summarise the key strands of causal inference research from economics and social sciences for the machine learning community
- To make connections between learning to act from observational versus experimental data. In particular, between causal graphical models and multi-armed bandits.

1.1 Motivation

Many of the most important questions in science and in our personal lives are about the outcomes of doing something. Will asking people to pay upfront at the doctors reduce long term health expenditure? If we developed a drug to suppress particular genes, could we cure MS and would delaying teen-aged pregnancies improve the outcome for their kids.

These are hard questions because they require more than identifying a pattern in data. Correlation is not causation. Causal inference has proven so difficult that there is barely any consensus on even enduring questions like the returns to education or the long-term consequences of early life events – like teenage pregnancy, where the variables involved are susceptible to human intuition and understanding.

We now live in a world of data. Hours of our lives are spent online, where every click can be recorded, tiny computers and sensors are cheap enough to incorporate into everything and the US Institute of Health is considering if all infants should be genetically sequenced at birth. Such data gives us a window into many aspects of our lives at an unprecedented scale and detail but it is messy, complicated and often generated as a by product of some other purpose. It does not come from the controlled world of a randomised experiment.

The rise of big data sets and powerful computers has seen an explosion in the application of machine learning. From health care, to entertainment and self driving cars; machine learning algorithms will transform many industries. It has been suggested that the impressive ability of statistical machine learning to detect complex patterns in huge data sets heralds the end of theory (Reference) and that we may be only a short step from the Singularity, where artificial intelligence exceeds our own and then grows exponentially.

However, despite the huge advances in specific areas of machine learning (in particular deep learning), machine learning algorithms are effective only within very narrow problem settings. Getting them to generalise to even slightly different problems or data sets remains very challenging. Deciding how we should act or what policies we should implement requires us to be able to predict how a system will behave if we change it. The correlations detected by standard machine learning algorithms do not enable us to do this, no matter how many petabytes of data they are based on. As machine learning algorithms are incorporated into more and more of the decision making processes that shape the world we live it, it is critical to ensure that we understand the distinction between causality and prediction and that we develop techniques for learning how to act that are as effective as those we have for pattern recognition.

1.2 What is causality and why do we care?

The notion of causality has been widely debated in science and philosophy [] but is still widely viewed as poorly defined []. This has led to a reluctance among applied researchers in many fields to refer to causality in their work, leading them instead to report that variables are *related*, *correlated* or *associated*. However, the magnitude, direction and even existence of an association depends on what other variables we adjust for (or include in a regression). Avoiding formalising causation, the real question of interest, leaves it up to the reader to determine via common sense if the association reported is the *right one*.

We discuss more detailed definitions of causality in section c

The what-if type questions from the why. [66] []. Why do whites to better than blacks in school (in America). Suggests that reverse causal inference questions are more interesting and motivate most of social science.

I do not find this distinction useful. We can only change the future - history is useful only as far as it tells us something about the future. Reverse causal questions can be reposed as forward ones, when making the translation a reverse causal query will be effectively asking about many possible interventions (rather than just one). Problems highlighted as intractable in the reverse causal sense are also intractable in the forward inference form, typically because concern situations for which we do not have a sufficient number of similar instances to allow statistical reasoning. For example, the war question posed in Gelman.

A distinction between forward causal inference, what happens if we do X and reverse casual inference

To explain or to predict [141] The two cultures [29]

There are two reasons why correlation is not causation . The first is related to variance and over-fitting. Observations are noisy. With a finite data set with enough variables we will be able to find some that are completely unrelated but correlate purely by chance. **??**. The second arises from bias, typically introduced by an un-observed confounding variable. In this case, variables are correlated not by chance. We would expect the relationship to hold if we

sampled more data. However, they are not causally related in that intervening to set one would not likely effect the other. EXAMPLE WITH FIGURE.

1.2.1 Defining causality

- widely debated in science and philosophy (REFERENCES)
- what is explanation?
- any model that aims to predict the outcome of an action or intervention in a system
- I do not see the distinction between explanation and (causal) prediction. Explanation is all about the ability to compress and to generalise. The more a model can do this, the more we view as providing an understanding of the why.
- mediation?

1.3 Identifying when we have a causal problem

Examples of typical machine learning problems. Are they causal?

Consider the following problems, which span a wide range of the types of questions machine learning is currently being applied to. Which of them require casual inference? How can we identify characteristics of a problem that make causal modelling important?

- Speech recognition (for systems like Siri or Google)
- Machine translation
- Image classification
- Forecasting the weather
- Playing Go
- Identifying spam emails
- Automated essay marking
- Predicting the risk of death in patients with pneumonia.
- Predicting who will re-offend on release from prison
- Predicting which customers will cease to be your customers
- Demand prediction for inventory control
- Predicting who will click on an ad
- Financial trading
- Recommending movies
- Online search
- Self driving cars
- Pricing insurance

The above problems are not posed with enough detail to know if causality is an important consideration. In particular, I failed to specify what actions the might be taken in response to model.

Consider speech recognition. You say something, which causes to sound waves, which are converted to a digital signal which Siri maps to words. Whatever action Siri takes is unlikely to change the distribution of words you use, and even less likely to change the function that maps sound waves to text (unless she sends you a DVD on elocution). A similar argument could be made for many applications of machine translation and image classification.

In image recognition, we do not particularly care about building a strong model for exactly how the thing that was photographed translates to the image we see. We can be fairly confident that the process will not change. If we develop a discriminating model that is highly accurate at classifying cats from dogs, we do not care a lot about its internal workings (provided we have strong grounds to believe that the situations in which we will be using our model will match those under which it was trained).

What about forecasting the weather? If you are using a short term forecast to decide whether to pack an umbrella it's clear causality can be ignored - your decision will not effect if it actually rains. However, longer term climate forecasts might (theoretically) lead us to take action on emissions which would then change the weather system. For this we need a (causal) model that allows us to predict the outcome under various different interventions.

Identifying spam and automated essay marking systems are similar. The decision made by the algorithm is likely to change the relationship between the features used by the algorithm and the true label. Spammers and students will modify their writing in order to optimise their results. The standard machine learning approach can only work if the resulting change is sufficiently gradual and fresh ground truth (probably human labelled) training data is provided. (What would the nature of the features have to be such that change did not occur? - they would have to be causes of the label).

What about predicting the risk of death in patients with pneumonia? Suppose we wish to use the model to decide who should be treated in hospital and who can be sent home with antibiotics. If we assume that in hospital treatment is more effective this seems like a straightforward prediction problem. It is not. Depending on how the decision to admit was previously made and what features are included (or omitted) in the model, the relationship between those features and the outcome may change if we start using the model to decide whom to admit. (xxx et al) found exactly this effect. Their model learnt that (among other things) people suffering asthma were *less* likely to die from pneumonia. They realised this was because doctors were treating such patients very aggressively, thus actually lowering their risk. There is no problem with this model if you want to predict who would be likely to die whilst maintaining the original addition and treatment protocols. However, using it to decide on what basis to admit people could kill. The key is understanding exactly what question you are asking. In this case we are care about what happens to patients with characteristics X if we treat them according to decision rule Z.

Predicting which customers will leave or who will re-offend if granted parole also fit within the category of problems where you wish to identify a group for which a problem will occur and target some treatment to them (loyalty reward, deny parole or more support whilst on parole, etc). For all these problems the assumptions required to treat them as pure prediction problems are;

1. The treatment is assumed to be effective (at least better than nothing)
2. Deciding who to treat on the based of the model predictions won't change the relationship between features and outcome

Demand prediction seems relatively straightforward. These models use features such as location, pricing, marketing, time of year, weather, etc to forecast the demand for a product. It seems unlikely that using the model to ensure stock is available will itself change demand. However there is a potential data censoring issue. If demand is modelled by the number of sales, then if a product is out of stock demand will appear to be zero. Changing availability does then change demand.

Playing Go (and other games) is another case with some subtleties. At every turn, the AI agent has a number of actions available. The board state following each action is deterministic and given by the rules of the game. The agent can apply supervised machine learning based on millions of previous games to estimate the probability that each of these board states will lead to a win. ... this is interesting maybe come back to it ... an alternate approach would be to try to forecast the probability of a win given each action given the current board state as context ... One approach to causal inference is indeed to learn about actions from taking actions (or observing the actions that other have taken). When can we learn from the actions others have taken? When there is no confounding. And does this hold with Go? Probably because the board state encapsulates everything that should determine what move is played. Learning directly from actions and trying to generalise (can in some instances reduce the problem to standard ML)

Having considered these examples we can now identify some general aspects of the problem that determine whether or not we require a causal model.

- Does acting on the predictions of the model change the mapping from features to target? (at least if the decision process is open to scrutiny). In general, if we believe that humans are generally trying to optimise to various goals of their own then for any system interacting with them the answer to this will be yes.
- Covariate shift clearly comes in here. Because there are areas where mechanisms are understood it is relatively easy to argue that covariate shift is not occurring and that results will be transferable. The mechanism is known but the function may be complex. Can we write down something that causal models are invariant to in terms of shift that is not the case for non-causal models? Yes, if the way in which features get their values changes, then causal models will be invariant to that in a way that non-causal ones are not.
- To decide between actions we only need to rank them (not estimate their actual effect).
- The predicted outcome in the absence of an intervention provides a single point. We can use this to find which problems are most serious if left alone - and prioritise those for modelling changes.
- Any decision we take does not significantly impact the system from which the data was drawn to make it (for repeat decision making)
- Does acting on the result of the prediction change the predictive distribution p(y—x)? IE change people's behaviour.
- Ethics - ... People's viewpoint on if its OK...

I hope these examples gave you a feel for the richness and subtitles of causal inference. We will return to some of them in more detail once we have established some more concrete language and tools to approach them with.

1.4 Approaches to causality

There are two broad approaches to deciding how to act. Reinforcement learning and causal inference. In reinforcement learning we estimate the effect of actions by taking them. We assume there is an agent capable of intervening in the system and try to find policies for the agent to follow in selecting actions so as to maximise some kind of reward. This is a very powerful and general framework (REFERENCES TO GENERAL AI). However, POINT OUT SOME OF THE DIFFICULTIES WITH REINFORCEMENT LEARNING. We are frequently presented with large bodies of data that have been collected from a system in which we did not have any control over what actions were taken.

We will call data sets where we do not have control over the decision making process that generated the data observational. Versus experimental data sets, where we do have control (experimental data sets do not always have to be randomised -although that is a powerful approach to ensure we have control. There is a space in the middle where we have partially controlled the process by which agent select actions. Randomised data with imperfect compliance would be an example.

An agent (capable of intervening in the system) chooses an action from those available The agent making the decision included in the model. This is

Reinforcement learning addresses the problem of learning from explicitly taking actions. There is typically some state or environment. An agent chooses an action from those available in the current state. The state then evolves stochastically as a function of the selected action and the agent receives some feedback or reward that is a function of the new state. This setting differs from the standard classification problem in that the agent must learn from feedback on the selected action, rather than being presented with the correct action for a given state. A common modelling assumption is that the state evolves only as a function of the previous state and the action chosen, and given these, is independent of the previous history of states and actions. This is known as a Markov decision process or MDP. A particularly well studied and understood model is the singe state MDP. In this case, there are a set of actions, each associated with a fixed but unknown reward distribution and at each time step our agent selects an action and receives corresponding feedback. This is known as the multi-armed bandit problem.

Causal inference makes use of assumptions to allow the outcome of actions to be predicted from observational data. The key to causal inference is a framework that can model how actions change the state of the world. This framework then allows us to map information collected in one setting to another.

Both approaches can be seen as extensions to the concept of randomised controlled trials. Bandit algorithms deal with the sequential nature of the decision making process, causal inference with the problem that full randomisation is not always feasible, affordable or ethical. The similarities between the problems that these techniques have been developed to address raises the question of if there are problems best addressed by a combination of these approaches and how they can be combined. The goal of my thesis is to explore these questions. In the next sections I review the key literature in causal inference and bandits. I then present a general approach to how causal models might be incorporated into bandit settings and conclude by demonstrating an algorithm that leverages causal assumptions to improve performance in a specific bandit setting.

There are two key approaches to causal problems. The first is to learn the outcome of actions by directly intervening in the system and seeing what happens. We then get feedback on how good those actions were. This is the approach taken in reinforcement learning. ADVANTAGES AND DISADVANTAGES OF THIS APPROACH. The second broad approach is causal inference.

Here

??

insert figure showing data generating process and observed data defining ML

Two broad approaches

- Build a model to map the natural behaviour of the system to what will happen for some action
- Take the action and see what happens

The first is causal inference

The second is reinforcement learning

Both generalise from randomised experiment Reinforcement learning to sequential decisions, causal inference to non-experimental conditions

Both these fields relate to the problem of making optimal decisions and both can be seen as generalising randomised controlled experiments. Causal inference is the study of how to estimate the effect of an action in the absence of randomisation. Reinforcement learning studies how we can do better if the decisions are to be made sequentially.

Both approaches involve assumptions the latter that we can group context and actions.

Limitations of causal inference

Limitations of experiments What are the issues with standard randomised experiments?

Chapter 2

Causal models

Causal inference aims to infer the outcome of an intervention in some system from data obtained by observing (but not intervening in) it. To do this we need terminology to describe actions and how we anticipate the system will respond to them. Three key approaches have emerged: counterfactuals, structural equation models and causal Bayesian networks. In this chapter we will examine the problems these approaches allow us to solve, the assumptions they rely on and how they differ. We will also use them to describe the following simplified examples. The aim is to demonstrate the notations and formalisms needed to tackle more interesting problems later on.

Example 1. Suppose a pharmaceutical company wants to assess the effectiveness of a new drug on recovery from a given illness. This is typically tested by taking a large group of representative patients and randomly assigning half of them to a treatment group (who receive the drug) and the other half to a control group (who receive a placebo). The goal is to determine the clinical impacts of the drug by comparing the differences between the outcomes for the two groups (in this case, simplified to only two outcomes - recovery or non-recovery). We will use the variable X (1 = drug, 0 = placebo) to represent the treatment each person receives and Y (1 = recover, 0 = not recover) to describe the outcome.

Example 2. Suppose we want to estimate the impact on high school graduation rates of compulsory preschool for all 4 year olds. We have a large cross-sectional data set on a group of 20 year olds that records if they attended pre-school, if they graduated high school and their parents socio-economic status (SES). We will let $X \in \{0, 1\}$ indicate if an individual attended pre-school, $Y \in \{0, 1\}$ indicate if they graduated high school and $Z \in \{0, 1\}$ represent if they are from a low or high SES background respectively.[1]

2.1 Causal Bayesian networks

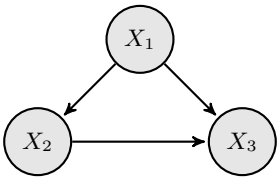
Causal Bayesian networks are an extension of Bayesian networks. A Bayesian network is a graphical way of representing how a distribution factorises. Any joint probability distribution can be factorised into a product of conditional probabilities. There are multiple valid factorisations, corresponding to permutations of variable ordering.

$$P(X_1, X_2, X_3, ...) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)... \tag{2.1}$$

[1]There has been substantial empirical work on the effectiveness of early childhood education including a landmark randomised trial, the Perry Preschool project, which ran from 1962-1967 [166].

We can represent this graphically by drawing a network with a node for each variable and adding links from the variables on the right hand side to the variable on the left for each conditional probability distribution, see figure 2.1. If the factorisation simplifies due to conditional independencies between variables, this is reflected by missing edges in the corresponding network. There are multiple valid Bayesian network representations for any probability distribution over more than one variable, see figure 2.2 for an example.

Figure 2.1: A general Bayesian network for the joint distribution over three variables. This network does not encode any conditional independencies between its variables and can thus represent any distribution over three variables.



The statement that a given graph G is a Bayesian network for a distribution P tells us that the distribution can be factorised over the nodes and edges in the graph. There can be no missing edges in G that do not correspond to conditional independencies in P, (the converse is not true G can have extra edges). If we let $parents_{X_i}$ represent the set of variables that are parents of the variable X_i in G then we can write the joint distribution as;

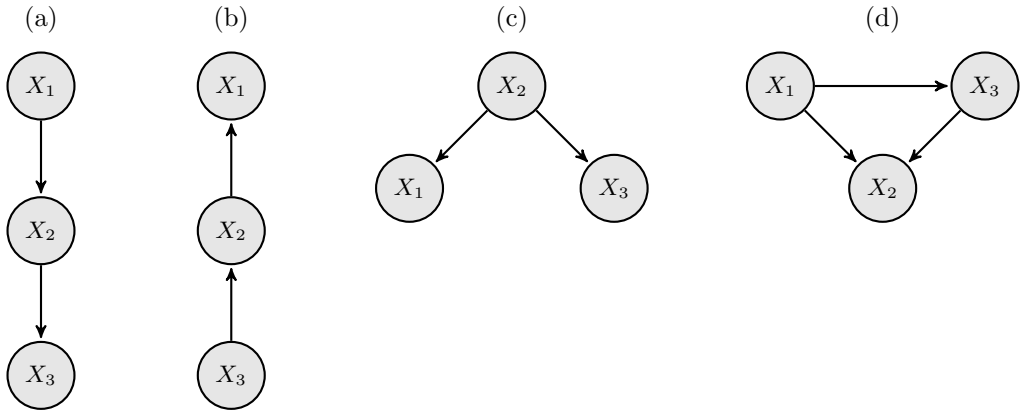
$$\mathrm{P}\left(X_1, ..., X_N\right) = \prod_{i=1...N} \mathrm{P}\left(X_i | parents_{X_i}\right) \tag{2.2}$$

A causal Bayesian network is a Bayesian network in which a link $X_i \to X_j$, by definition, implies X_i causes X_j. This means that if we intervene and change the value of X_i, we expect X_j to change, but if we intervene to change X_j, X_i will not change. We need some notation to describe interventions and represent distributions over variables in the network after an intervention. In this thesis I use the do operator introduced by Pearl [116].

Definition 3. The do-notation

- $do(X = x)$ denotes an intervention that sets the random variable(s) X to x.
- $\mathrm{P}\left(Y | do(X)\right)$ is the distribution of Y conditional on an *intervention* that sets X. This notation is somewhat overloaded. It may be used represent a probability distribution/mass

Figure 2.2: Some valid Bayesian networks for a distribution that an be factorised as $\mathrm{P}\left(X_1, X_2, X_3\right) = \mathrm{P}\left(X_1\right)\mathrm{P}\left(X_2 | X_1\right)\mathrm{P}\left(X_3 | X_2\right)$ (which implies $X_3 \perp\!\!\!\perp X_1 | X_2$)



function or a family of distribution functions depending on whether the variables are discrete or continuous and whether or not we are treating them as fixed. For example it could represent

- the probability $\mathrm{P}(Y = 1|do(X = x))$ as a function of x,
- the probability mass function for a discrete Y : $\mathrm{P}(Y|do(X = x))$,
- the probability density function for a continuous Y : $f_Y(y|do(X = x))$,
- a family of density/mass function for Y parameterised by x.

Where the distinction is important and not clear from context we will use one of the more specific forms above.

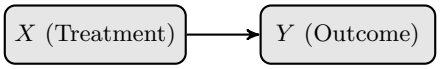
Theorem 4 (Truncated product formula [116])**.** *If G is a causal network for a distribution P defined over variables $X_1...X_N$, then we can calculate the distribution after an intervention where we set $Z \subset X$ to z, denoted $do(Z = z)$ by dropping the terms for each of the variables in Z from the factorisation given by the network.*

$$\mathrm{P}(X_1...X_N|do(Z = z)) = \begin{cases} \prod_{i \notin Z} \mathrm{P}(X_i|parents_{X_i}) & \text{if } (X_1...X_N) \text{ consistent with } Z = z \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

Theorem 4 does not hold for standard Bayesian networks because there are multiple valid networks for the same distribution. The truncated product formula will give different results depending on the selected network. The result is possible with causal Bayesian networks because it follows directly from the assumption that the direction of the link indicates causality. In fact, from the interventionist viewpoint of causality, the truncation product formula defines what it means for a link to be causal.

Returning to example 1, and phrasing our query in terms of interventions; what would the distribution of outcomes look like if everyone was treated $\mathrm{P}(Y|do(X = 1))$, relative to if no one was treated $\mathrm{P}(Y|do(X = 0))$? The treatment X is a potential cause of Y, along with other unobserved variables, such as the age, gender and the disease sub-type of the patient. Since X is assigned via deliberate randomisation we know that it is not affected by any latent variables. The causal Bayesian network for this scenario is shown in figure 2.3. This network represents the (causal) factorisation $\mathrm{P}(X, Y) = \mathrm{P}(X)\,\mathrm{P}(Y|X)$, so from equation (2.3), $\mathrm{P}(Y|do(X)) = \mathrm{P}(Y|X)$. In this example, the interventional distribution is equivalent to the observational one.

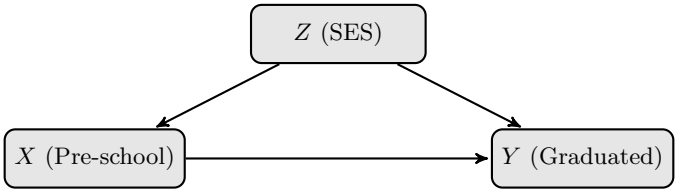
Figure 2.3: Causal Bayesian network for example 1



In example 2 we are interested in $\mathrm{P}(Y|do(X = 1))$, the expected high-school graduation rate if we introduce universal preschool. We could compare it to outlawing preschool $\mathrm{P}(Y|do(X = 0))$ or the current status quo $\mathrm{P}(Y)$. It seems reasonable to assume that preschool attendance affects the likelihood of high school graduation [2] and that parental socio-economic status would affect *both* the likelihood of preschool attendance and high school graduation. If we assume that socio-economic status is the only such variable (nothing else effects both attendance *and* graduation), we can represent this problem with the causal Bayesian network in figure 2.4. In this case, the interventional distribution is not equivalent to the observational one. If parents

[2] The effect does not have to be homogeneous, it may depend non-linearly on characteristics of the child, family and school.

Figure 2.4: Causal Bayesian network for example 2



with high socio-economic status are more likely to send their children to preschool and these children are more likely to graduate high school regardless, comparing the graduation rates of those who attended preschool with those who did not will overstate the benefit of preschool. To obtain the interventional distribution we have to estimate the impact of preschool on high school graduation for each socio-economic level separately and then weight the results by the proportion of the population in that group,

$$\mathrm{P}\left(Y|do(X=1)\right) = \sum_{z \in Z} \mathrm{P}\left(Y|X=1,Z\right)\mathrm{P}\left(Z\right) \tag{2.4}$$

We have seen from these two examples that the expression to estimate the causal effect of an intervention depends on the structure of the causal graph. There is a very powerful and general set of rules that specifies how we can transform observational distributions into interventional ones for a given graph structure. These rules are referred to as the Do-calculus [116]. We discuss them further in section 3.2.

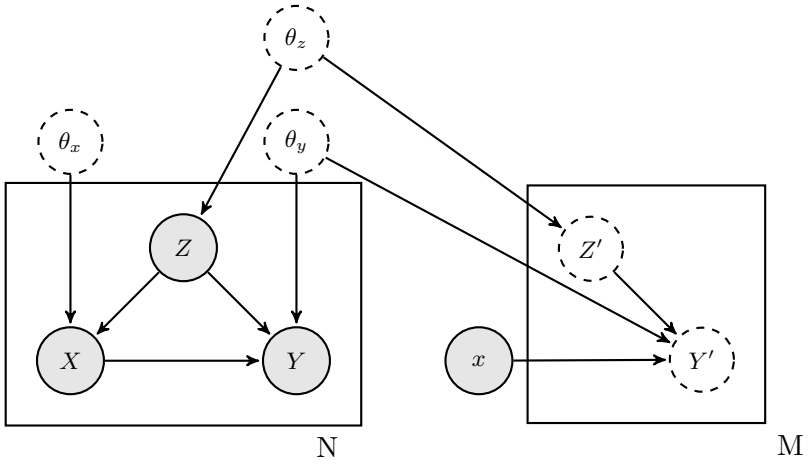
Formalising the definition of an intervention within the framework of causal graphical models provides us with an explicit mechanism to map information from one data generating process, the system pre-intervention, to another, the system post-intervention. The power of defining an intervention in this way stems from the number of things that are invariant between the two processes. All the (conditional) distributions for variables in the graph that were not directly set by the intervention are assumed not be changed by it.

We could represent problems of the type where we try to infer properties of the post-interventional system based on data generated by the pre-interventional distribution by explicitly representing both systems and what they have in common, see figure 2.5. This does not require any special framework or notation. The graphs in figure 2.5 are ordinary Bayesian networks. However, without a causal framework, we have to make assumptions about what will be invariant to the intervention specifically for each such problem we encounter. For complex problems, it is very difficult to conceptualise the assumptions we expect to hold without the benefit of a causal framework.

A causal Bayesian network represents much more information than a Bayesian network with identical structure. A causal network encodes all possible interventions that could be specified with the do-notation. For example, if the network in figure 2.4 were an ordinary Bayesian network and all the variables were binary, the associated distribution could be described by 7 parameters. The equivalent causal Bayesian network additionally represents the post-interventional distributions for six possible single variable interventions and twelve possible two variable interventions. Encoding all this information without the assumptions implicit in the causal Bayesian network would require an additional 30 parameters [3].

[3] After each single variable intervention we have a distribution over two variables, which can be represented by three parameters. After each two variable intervention, we have a distribution over one variables which requires one parameter. This takes us to a total of $6 \times 3 + 12 \times 1 = 30$ additional parameters.

Figure 2.5: Causal inference with ordinary Bayesian networks. The plate on the left represents the observed data generated prior to the intervention and the plate on the right the data we anticipate obtaining after an intervention that sets the pre-interventional variable X to x. The assumptions characterised by this plate model correspond to those implied by the causal Bayesian network in figure 2.4 for the intervention $do(X = x)$. As the networks in this figure are ordinary Bayesian networks, we could have represented the same information with a different ordering of the links within each plate. However, we would then have a complex transformation relating the parameters between the two plates rather than a simple invariance.

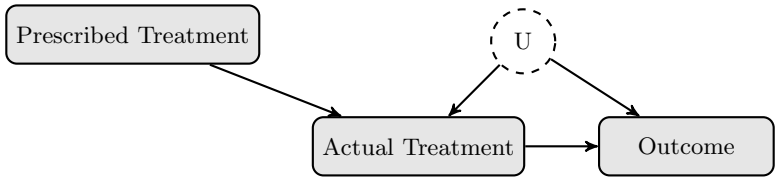


Causal Bayesian networks are Bayesian networks, so results that apply to Bayesian networks carry directly across; the local Markov property states that variables are independent of their non-effects given their direct causes. Similarly the global Markov property and d-separation also hold in causal networks.

Limitations of causal Bayesian networks

A number of criticisms have been levelled at this approach to modelling causality. One is that the definition of an intervention only in terms of setting the value of one or more variables is too precise and that any real world intervention will affect many variables in complex and non-deterministic ways [128, 38]. However, by augmenting the causal graph with additional variables that model how interventions may take effect, the deterministic do operator can model more complex interventions. For example, in the drug treatment case, we assumed that all subjects complied, taking the treatment or placebo as assigned by the experimenter. But what if some people failed to take the prescribed treatment. We can model this within the framework of deterministic interventions by adding a node representing what they were prescribed (the intervention) which probabilistically influences the treatment they actually receive, see figure 2.6. Note that the fact that we no longer directly assign the treatment opens the possibility that an unobserved latent variable could affect both the actual treatment taken and the outcome.

Figure 2.6: Randomised treatment with imperfect compliance



Another key issue with causal Bayesian networks is that they cannot handle cyclic dependen-

cies between variables. Such feedback loops are common in real-life systems, for example the relationship between supply and demand in economics or predator and prey in ecology. We might regard the underlying causal mechanisms in these examples to be acyclic; the number of predators at one time influences the number of prey in the next period and so on. However, if our measurements of these variables must be aggregated over time frames that are longer than the scale at which these interactions occur the result is a cyclical dependency. Even were we able to measure on shorter timescales, we might then not have sufficient data on each variable for inference. Such problems have mostly been studied within the dynamic systems literature, typically focusing on understanding the stationary or equilibrium state of the system and making very specific assumptions about functional form in order to make problems tractable. Poole and Crowley [122] compare the equilibrium approach to reasoning about cyclic problems with structural equation models, which we discuss in section 2.3 and which can be seen as Bayesian causal networks with additional functional assumptions.

2.2 Counterfactuals

The Neyman-Rubin model [133, 134, 131, 135, 136] defines causality in terms of potential outcomes, or counterfactuals. Counterfactuals are statements about imagined or alternate realities, are prevalent in everyday language and may play a role in the development of causal reasoning in humans [167]. Causal effects are differences in counterfactual variables: what is the difference between what would have happened if we did one thing versus what would have happened if we did something else.

In example 1, the causal effect of the drug relative to placebo for person i is the difference between what would have happened if they were given the drug, denoted ${y_i}^1$ versus what would have happened if they got the placebo, ${y_i}^0$. The fundamental problem of causal inference is that we can only observe one of these two outcomes, since a given person can only be treated or not treated. The problem can be resolved if, instead of people, there are units that can be assumed to be identical or that will revert exactly to their initial state some time after treatment. This type of assumption often holds to a good approximation in the natural sciences and explains why researchers in these fields are less concerned with causal theory.

Putting aside any estimates of individual causal effects, it is possible to learn something about the distributions under treatment or placebo. Let Y^1 be a random variable representing the potential outcome if treated. The distribution of Y^1 is the distribution of Y if everyone was treated. Similarly Y^0 represents the potential outcome for the placebo. The difference between the probability of recovery, across the population, if everyone was treated and the probability of recovery given placebo is $\mathrm{P}\left(Y^1\right) - \mathrm{P}\left(Y^0\right)$. We can estimate (from an experimental or observational study):

- $\mathrm{P}\left(Y|X=1\right)$, the probability that those who took the treatment will recover
- $\mathrm{P}\left(Y|X=0\right)$, the probability that those who were not treated will recover

Now, for those who took the treatment, the outcome *had* they taken the treatment Y^1 is the same as the observed outcome. For those who did not take the treatment, the observed outcome is the sames as the outcome *had* they not taken the treatment. Equivalently stated:

$$\mathrm{P}\left(Y^0|X=0\right) = \mathrm{P}\left(Y|X=0\right)$$
$$\mathrm{P}\left(Y^1|X=1\right) = \mathrm{P}\left(Y|X=1\right)$$

If we assume $X \perp\!\!\!\perp Y^0$ and $X \perp\!\!\!\perp Y^1$:

$$\begin{aligned} \mathrm{P}\left(Y^1\right) &= \mathrm{P}\left(Y^1|X=1\right) = \mathrm{P}\left(Y|X=1\right) \\ \mathrm{P}\left(Y^0\right) &= \mathrm{P}\left(Y^0|X=0\right) = \mathrm{P}\left(Y|X=0\right) \end{aligned}$$

This implies the counterfactual distributions are equivalent to the corresponding conditional distributions and, for a binary outcome Y, the causal effect is,

$$\mathrm{P}\left(Y^1\right) - \mathrm{P}\left(Y^0\right) = \mathrm{P}\left(Y|X=1\right) - \mathrm{P}\left(Y|X=0\right)$$

The assumptions $X \perp\!\!\!\perp Y^1$ and $X \perp\!\!\!\perp Y^0$ are referred to as ignorability assumptions [131]. They state that the treatment each person receives is independent of whether they would recover if treated and if they would recover if not treated. This is justified in example 1 due to the randomisation of treatment assignment. In general the treatment assignment will not be independent of the potential outcomes. In example 2, the children who attended preschool may be more likely to have graduated high school had they in fact not attended than the children who actually did not attend, $X \not\perp\!\!\!\perp Y^0$. Similarly, had the poorer children who did not attend pre-school attended they might not have done as well as the children who did in fact attend, $X \not\perp\!\!\!\perp Y^1$. A more general form of the ignorability assumption is to identify a set of variables Z such that $X \perp\!\!\!\perp Y^1|Z$ and $X \perp\!\!\!\perp Y^0|Z$.

Theorem 5 (Ignorability)**.** *If $X \perp\!\!\!\perp Y^1|Z$ and $X \perp\!\!\!\perp Y^0|Z$,*

$$\mathrm{P}\left(Y^1\right) = \sum_{z \in Z} \mathrm{P}\left(Y|X=1,Z\right)\mathrm{P}\left(Z\right) \tag{2.5}$$

$$\mathrm{P}\left(Y^0\right) = \sum_{z \in Z} \mathrm{P}\left(Y|X=0,Z\right)\mathrm{P}\left(Z\right) \tag{2.6}$$

Assuming that within each socio-economic status level, attendance at pre-school is independent of the likelihood of graduating high-school had a person attended, then the average rate of high-school graduation given a universal pre-school program can be computed from equation 2.5. Note, that this agrees with the weighted adjustment formula in equation 2.4.

Another assumption introduced within the Neyman-Rubin causal framework is the Stable Unit Treatment Value Assumption (SUTVA) [134]. This is the assumption that the potential outcome for one individual (or unit) does not depend on the treatment assigned to another individual. As an example of a SUTVA violation, suppose disadvantaged four year olds were randomly assigned to attend pre-school. The subsequent school results of children in the control group, who did not attend, could be boosted by the improved behaviour of those who did and who now share the classroom with them. SUTVA violations would manifest as a form of model misspecification in causal Bayesian networks.

There are complex philosophical objections to counterfactuals arising from the way they describe alternate universes that were never realised. This makes it quite easy to (accidentally) make statements about counterfactuals that cannot be tested with empirical data. Consider the following example based on Dawid [47]. Again we have a drug where the outcome for an individual if treated is represented by the counterfactual variable Y^1 and the outcome if not

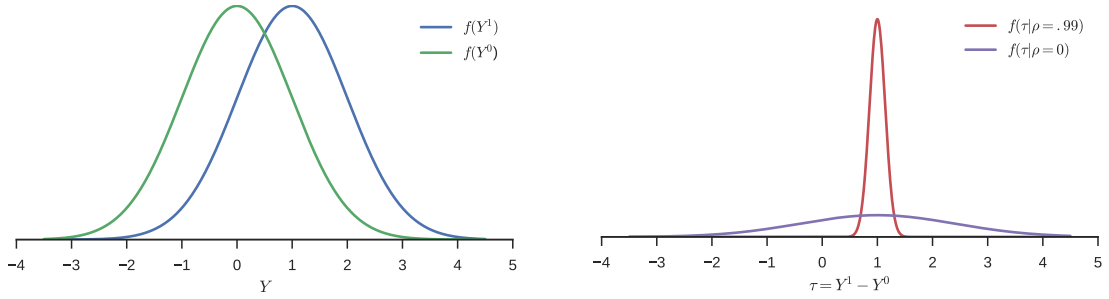
treated is Y^0. Suppose these counterfactual variables Y^1 and Y^0 are jointly normal with equal variance (for simplicity).

$$P(Y^1, Y^0) \sim N(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2\end{bmatrix}) \tag{2.7}$$

Their difference is also normal. Let $\tau = Y^1 - Y^0$,

$$P(\tau) = N(\mu_1 - \mu_0, 2\sigma^2(1-\rho)) \tag{2.8}$$

Figure 2.7: The distribution of individual treatment effects is not identifiable, even from a randomised controlled trial.



(a) Marginal distributions over the potential outcomes Y^1 and Y^0 for $\mu_1 = 1$, $\mu_0 = 0$ and $\sigma = 1$. The blue curve shows the distribution of Y if everyone were to be treated and the blue curve the distribution if no-one was treated.

(b) Two very different distributions of individual causal effects consistent with the potential outcome distributions.

From a (large) randomised controlled trial we can estimate the marginal distributions over the counterfactual variables, see figure 2.7a. These represent the distributions we would expect over the outcome Y if everyone were treated or not treated respectively. However, the distribution over the individual causal effects depends on ρ, see figure 2.7b. The key problem is that we can never observe the joint distribution over Y^1 and Y^0. As a result, ρ and thus the variance of τ is not identifiable, even from experimental data. Dawid [47] argues that we should avoid using counterfactuals as they are defined in terms of (metaphysical) individual causal effects. He further points out that the interventional distributions in figure 2.7a, along with a loss function, contain all the information required to decide how to treat a new patient.

This result is unintuitive. It seems on the face of it that the distribution of individual causal effects is relevant to our decision making. If $\rho = 1$ then almost everyone benefits slightly from the treatment whilst if $\rho = 0$, there is a wide range, with some people benefiting a lot and others suffering significant harm. This confusion can be resolved by thinking about personalised rather than individual causal effects. It is entirely possible that potentially observable characteristics (such as gender, age, genetics, etc) affect how people will respond to the treatment. We can partition people into sub-populations on the basis of these characteristics and measure different *personalised* causal effects for each group. The variance of the potential outcome distributions $f(Y^1)$ and $f(Y^0)$ provides bounds on how much can be gained from further personalisation. The metaphysical nature of individual causal effects only arises when we are at the point where the only remaining variation is due to inherent randomness (or variables that we could not even in principle measure).

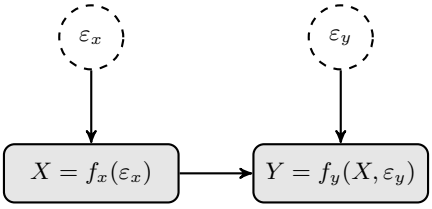
One way of looking at counterfactuals is as a natural language short hand for describing highly specific interventions like those denoted by the do-notation. Rather than talking about the distribution of Y given we intervene to set $X = x$ and hold everything else about the system constant we just say what would the distribution of Y be had X been x. This is certainly convenient, if rather imprecise. However, the ease with which we can make statements with counterfactuals that cannot be tested with empirical data warrants careful attention. It is important to be clear what assumptions are being made and whether or not they could be validated (at least in theory).

2.3 Structural Equation models

Structural equation models (SEMs) describe a deterministic world, where some underlying mechanism or function determines the output of any process for a given input. The mechanism (but not the output) is assumed to be independent of what is fed into it. Uncertainties are not inherent but arise from unmeasured variables. Linear structural equation models have a long history for causal estimation [170, 68]. More recently, they have been formalised, generalised to the non-linear setting and connected to developments in graphical models to provide a powerful causal framework [116].

Mathematically, each variable is a deterministic function of its direct causes and a noise term that captures unmeasured variables. The noise terms are required to be mutually independent. If there is the possibility that an unmeasured variable influences more than one variable of interest in a study, it must be modelled explicitly as a latent variable. Structural equation models can be represented visually as a network. Each variable is a node and arrows are drawn from causes to their effects. Figure 2.8 illustrates the SEM for example 1.

Figure 2.8: SEM for example 1



This model encodes the assumption that the outcome y_i for an individual i is caused solely by the treatment x_i they receive and other factors ε_{y_i} that are independent of X. This is justifiable on the grounds that X is random. The outcome of a coin flip for each patient should not be related to any of their characteristics (hidden or otherwise). Note that the causal graph in figure 2.8 is identical to that the Bayesian network for the same problem, figure 2.3. The latent variables ε_x and ε_y are not explicitly drawn in figure 2.3 as they are captured by the probabilistic nature of the nodes in a Bayesian network.

Taking the *action* $X = 1$ corresponds to replacing the equation $X = f_x(\varepsilon_x)$ with $X = 1$. The function f_y and distribution over ε_y does not change. This results in the interventional distribution [4],

$$\mathrm{P}\left(Y = y | do(X = 1)\right) = \sum_{\varepsilon_y} \mathrm{P}\left(\varepsilon_y\right) \mathbb{1}\{f_y(1, \varepsilon_y) = y\} \tag{2.9}$$

[4] We have assumed the variables are discrete only for notational convenience

The observational distribution of Y given X is,

$$\mathrm{P}\left(Y = y|X = 1\right) = \sum_{\varepsilon_x}\sum_{\varepsilon_y} \mathrm{P}\left(\varepsilon_x|X = 1\right)\mathrm{P}\left(\varepsilon_y|\varepsilon_x\right)\mathbb{1}\{f_y(1, \varepsilon_y) = y\} \tag{2.10}$$

$$= \sum_{\varepsilon_y} \mathrm{P}\left(\varepsilon_y\right)\mathbb{1}\{f_y(1, \varepsilon_y) = y\}\,, \text{ as } \varepsilon_x \perp\!\!\!\perp \varepsilon_y \tag{2.11}$$

The interventional distribution is the same as the observational one. The same argument applies to the intervention $do(X = 0)$ and so the causal effect is simply the difference in observed outcomes as found via the causal Bayesian network and counterfactual approaches.

The SEM for example 2 is shown in figure 2.9. Intervening to send all children to pre-school replaces the equation $X = f_x(Z, \varepsilon_x)$ with $X = 1$, leaving all the other functions and distributions in the model unchanged.

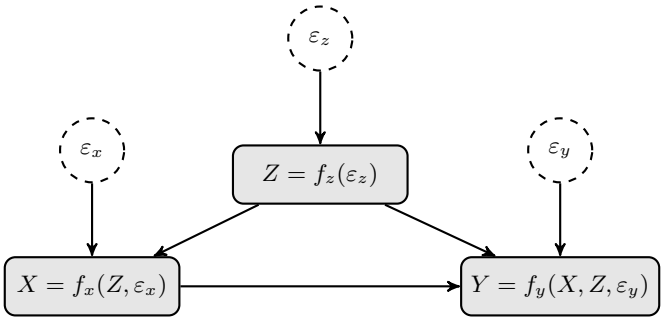
$$\mathrm{P}\left(Y = y|do(X = 1)\right) = \sum_{z}\sum_{\varepsilon_y} \mathrm{P}\left(z\right)\mathrm{P}\left(\varepsilon_y\right)\mathbb{1}\{f_y(1, z, \varepsilon_y) = y\} \tag{2.12}$$

$$= \sum_{z} \mathrm{P}\left(z\right) \underbrace{\sum_{\varepsilon_y} \mathrm{P}\left(\varepsilon_y\right)\mathbb{1}\{f_y(1, z, \varepsilon_y) = y\}}_{\mathrm{P}(Y=y|X=1,Z=z)} \tag{2.13}$$

Equation 2.13 corresponds to equations 2.4 and 2.5. It is not equivalent to the observational distribution given by:

$$\mathrm{P}\left(Y = y|X = 1\right) = \sum_{z}\sum_{\varepsilon_y} \mathrm{P}\left(z|X = 1\right)\mathrm{P}\left(\varepsilon_y\right)\mathbb{1}\{f_y(1, z, \varepsilon_y) = y\} \tag{2.14}$$

Figure 2.9: SEM for example 2



Structural equation models are generally applied with strong constraints on the functional form of the relationship between the variables and noise is typically assumed to be additive, $X_i = f_i(\cdot) + \varepsilon_i$. A structural equation model with N variables resembles a set of N simultaneous equations, with each variable playing the role of the dependent (left hand side) variable in one equation. However a SEM is, by definition, more than a set of simultaneous equations. By declaring it to be structural we are saying that it represents assumptions about the relationships between variables. When we visualise the model as a network the absence of an arrow between two variables encodes the assumption that one does not cause the other. The similarity between the notation used to describe and analyse structural equation models and simultaneous equations, combined with a reluctance to make explicit statements about causality, has led to some confusion in the interpretation of SEMs [71, 116].

2.4 Comparing and unifying the models

Remarkably for models developed relatively independently in fields with very different approaches and problems, the models we have discussed can be nicely unified for interventional queries (those that can be expressed with the do-notation). If the network for a structural equation model is acyclic, that is if starting from any node and following edges in the direction of the arrows you cannot return to the starting point, then it implies a recursive factorisation of the joint distribution over its variables. In other words, the network is a causal Bayesian network. All of the results that apply to causal Bayesian networks also apply to acyclic structural equation models. Taking an action that sets a variable to a specific value equates to replacing the equation for that variable with a constant. This corresponds to dropping a term in the factorisation and the truncated product formula (equation 2.3). Thus, the interventional query $P(Y|do(X))$ is identical in these two frameworks. We can also connect this to counterfactuals via:

$$\begin{aligned} Y^0 &\equiv P(Y|do(X=0)) \\ Y^1 &\equiv P(Y|do(X=1)) \end{aligned} \tag{2.15}$$

The assumption $\varepsilon_X \perp\!\!\!\perp \varepsilon_Y$, stated for our structural equation model, translates to $X \perp\!\!\!\perp (Y^0, Y^1)$ in the language of counterfactuals. When discussing the counterfactual model, we actually made the slightly weaker assumption:

$$X \perp\!\!\!\perp Y^0 \text{ and } X \perp\!\!\!\perp Y^1 \tag{2.16}$$

It is possible to relax the independence of errors assumption for SEMs to correspond exactly with the form of equation (2.16) without losing any of the power provided by d-separation and graphical identification rules [127]. The correspondence between the models for interventional queries (those that can be phrased using the do-notation) makes it straightforward to combine key results and algorithms developed within any of these frameworks. For example, you can draw a causal graphical network to determine if a problem is identifiable and which variables should be adjusted for to obtain an unbiased causal estimate. Then use propensity scores [131] to estimate the effect. If non-parametric assumptions are insufficient for identification or lead to overly large uncertainties, you can specify additional assumptions by phrasing your model in terms of structural equations. The frameworks do differ when it comes to causal queries that involve joint or nested counterfactuals and cannot be expressed with the do-notation. These types of queries arise in the study of mediation [119, 81, 163] and in legal decisions, particularly on issues such as discrimination [116].

In practice, differences in focus and approach between the fields in which each model dominates eclipse the actual differences in the frameworks. The work on causal graphical models [116, 151] focuses on asymptotic, non-parametric estimation and rigorous theoretical foundations. The Neyman-Rubin framework builds on the understanding of randomised experiment and generalises to quasi-experimental and observational settings, with a particular focus on non-random assignment to treatment. This research emphasises estimation of average causal effects and provides practical methods for estimation, in particular, propensity scores; a method to control for multiple variables in high dimensional settings with finite data [131]. In economics, inferring causal effects from non-experimental data to support policy decisions is central to the field. Economists are often interested in more informative measures of the distribution of causal effects than the mean and make extensive use of structural equation models, generally with strong parametric assumptions [72]. In addition, the parametric structural equation models

favoured in economics can be extended to analyse cyclic (otherwise referred to as non-recursive) models.

2.5 What does a causal model give us? Resolving Simpson's paradox

We will now demonstrate our new notation and frameworks for causal inference to resolve a fascinating paradox, noted by Yule [173], demonstrated in real data by Cohen and Nagel [44] and popularised by Simpson [146]. The following example is adapted from Pearl [116]. Suppose a doctor has two treatments, A and B, which she offers to patients to prevent heart disease. She keeps track of which medication her patients choose and whether or not the treatment is successful. She obtains the results in table 2.1.

Table 2.1: Treatment results

Treatment	Success	Fail	Total	Success Rate
A	87	13	100	87%
B	75	25	100	75%

Drug A appears to perform better. However, having read the latest literature on how medications affect men and women differently, she decides to break down her results by gender to see how well the drugs perform for each group and obtains the data in table 2.2.

Table 2.2: Treatment results by gender

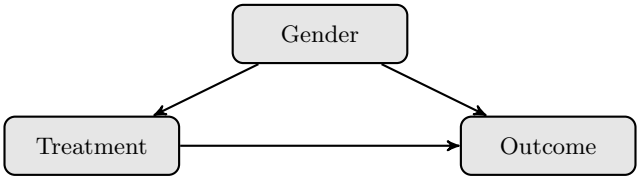
Gender	Treatment	Success	Fail	Total	Success Rate
M	A	12	8	20	60%
M	B	56	24	80	70%
F	A	75	5	80	94%
F	B	19	1	20	95%

Once the data is broken down by gender, drug B looks better for both men *and* women. Suppose the doctor must choose only one drug to prescribe to all her patients in future (perhaps she must recommend which to subsidise under a national health scheme). Should she choose A or B? The ambiguity in this question lies at the heart of Simpson's paradox. How does causal modelling resolve the paradox? The key is that the doctor is trying to choose between *interventions*. She wants to know what the success rate will be if she changes her practice to give all the patients one drug, rather than allowing them to choose as currently occurs.

Let's represent the treatment by the variable T, the gender of the patient by Z and whether or not the treatment was successful by Y. The doctor cares about $\mathrm{P}(Y|do(T))$, not the standard conditional distributions $\mathrm{P}(Y|T)$. Unfortunately, the data in tables 2.1 and 2.2 is insufficient to enable estimation of the interventional distribution $\mathrm{P}(Y|do(T))$ or determine if $do(T = A)$ is better or worse than $do(T = B)$. Some assumptions about the causal relationships between the variables are required. In this example, it seems reasonable to conclude that gender may affect the treatment chosen and the outcome. Assuming there are no other such confounding variables (for example income) then we obtain the causal network in figure 2.10.

With this model, women are more likely to choose drug A and are also more likely to recover than men regardless of the treatment they receive. Knowing a patient took drug A indicates they are more likely to be female. When we compare the group of people who took A against those who took B, the effect of the higher proportion of females in the first group conceals the

Figure 2.10: An example of Simpson's Paradox

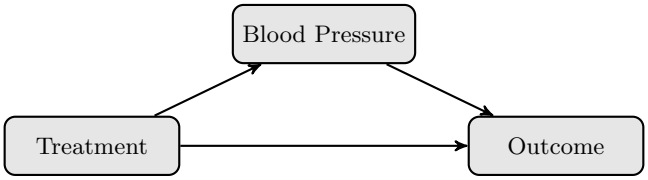


greater benefit of drug B leading to an apparent reversal in effectiveness. However, when the doctor intervenes to set the treatment each person receives there will no longer be a link from gender to treatment. So in this case she should choose which drug to prescribe from the gender specific table (and weight by the proportion of the population that belongs to each gender). Drug B is the better choice.

$$\mathrm{P}\left(Y|do(T)\right) = \mathrm{P}\left(Y|T, female\right)\mathrm{P}\left(female\right) + \mathrm{P}\left(Y|T, male\right)\mathrm{P}\left(male\right) \tag{2.17}$$

Is the solution to Simpson's paradox to always to break down the data by as many variables as possible? No. Suppose we have the identical data as in 2.1 and 2.2 but replace the column name 'gender' with 'blood pressure', 'M' with 'high' and 'F' with 'normal'. This is a drug designed to prevent heart disease. One pathway to doing so might well be to lower blood pressure. Figure 2.11 shows a plausible causal graph for this setting. It differs from the graph in figure 2.10 only in the direction of a single link. Now, however, table 2.2 tells us that people who took treatment A had better blood pressure control and better overall outcomes. In this setting $\mathrm{P}\left(Y|do(T)\right) = \mathrm{P}\left(Y|T\right)$ and drug A is the better choice.

Figure 2.11: An example of Simpson's Paradox

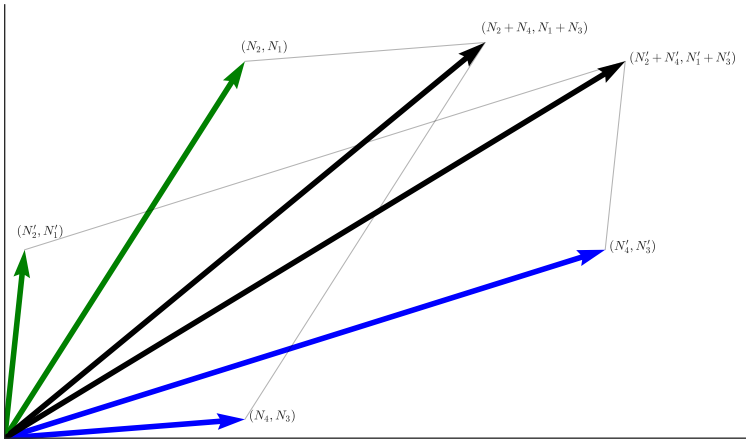


Note that we have not changed the data itself, only the description of the variables that it is associated with. This illustrates that the resolution to Simpson's paradox lies fundamentally not in the data, but in the assumptions we are willing to make. From a purely statistical viewpoint there is no paradox. The reversal just stems from the mathematical property of ratios expressed in equation 2.18 and represented graphically in figure 2.12. The paradox only arises when we attempt to use the data to select an intervention and is resolved when we apply a causal approach to do so.

$$\exists\left\{N_1, ...N_4, N_1'...N_4'\right\} \in \mathbb{N} : \frac{N_1}{N_2} < \frac{N_1'}{N_2'}, \quad \frac{N_3}{N_4} < \frac{N_3'}{N_4'} \text{ and } \frac{N_1 + N_3}{N_2 + N_4} > \frac{N_1' + N_3'}{N_2' + N_4'} \tag{2.18}$$

There are many other plausible causal graphs for both scenarios above. Perhaps income affects drug choice as well as gender, or gender might affect treatment choice and blood pressure control given treatment, etc. Causal modelling provides a powerful tool to specify such assumptions and to determine how to estimate causal effects for a given model see section 3.1.

Figure 2.12: Simpson's reversal visualised. The ratios involving N_i' are steeper than than those involving N_i for both the blue and green vectors. However, when we sum them, the ratio is steeper for the un-primed variables.



Chapter 3

Two key questions

We can roughly categorise the problems studied within causal inference into two groups, causal effect estimation and causal discovery. In causal effect estimation we assume (at least implicitly) that key aspects of the causal graph are known. The goal is then to estimate the effect of an intervention or range of interventions in the system. Causal effect estimation is implicit in countless studies in economics, social science and epidemiology of everything from the effect of education on earnings [37], diet on cancer [27] and breastfeeding on intelligence [84] to the effect of pet ownership on survival after a heart attack [58]. Almost every time someone runs a regression model the key quantity of interest is a causal effect. Given how it underlies so much of our scientific progress, there is a enormous potential in properly understanding when we can draw causal conclusions, exactly what assumptions are required to do so and how we can best leverage those assumptions to infer as much information as we can from our data.

Causal discovery aims to leverage much broader assumptions to learn the structure of causal graphs from data. This is critical in fields where we are generating a lot of data but have limited theoretical knowledge from which to draw on to determine how variables are related to one another. Causal discovery algorithms are being applied in bioinformatics [24, 137, 124, 7, 153, 61, 149, 158], medical imaging [125] and climate science [162]. An effective and generalisable approach for causal discovery would be a major step towards the automation of the scientific endeavour.

3.1 Causal effect estimation

Estimating causal effects from observational data comes down to determining if and how we can write expressions for the interventional distributions of interest in terms of observational quantities, which can be measured. We did this on an ad-hoc basis to resolve the examples discussed in chapter 2. In this chapter we describe a principled approach to mapping observational quantities to interventional ones and discuss some of the key issues involved in estimating such expressions from finite sample data. We assume the basic structure of the graph is known. That is, we assume that we can draw a network containing (at a minimum):

- the target/outcome variable we care about,
- the focus/treatment variables on which we are considering interventions,
- any variables which act to confound two or more of the other variables we have included, and
- any links between variables we have included.

Some of these variables may be latent in that the available data does not record their value, however their position in the network is assumed to be known. For example, consider estimating the impact of schooling on wages. Some measure of inherent ability could influence both the number of years of schooling people choose to pursue and the wages they receive. Even if we have no data to directly assess people's inherent ability we must include it in the graph because it influences two of the variables we are modelling.

How can the structure of the causal graph be leveraged to compute interventional distributions from observational ones? Given the graph corresponding to the observational distribution, the graph after any intervention can be obtained by removing any links into variables directly set by the intervention. The joint interventional distribution is the product of the factors associated with the interventional graph, as given by the truncated product formula 2.3. If there are no latent variables the interventional distribution of interest can be obtained by marginalising over the joint (interventional) distribution. However, if there are latent variables the joint interventional distribution will contain terms that cannot be estimated from the observed data.

The key to estimating causal effects in the presence of latent variables lies in combining the assumption of how an intervention changes the graph, encoded by the truncated product formula, with information the graph structure provides about conditional independencies between variables. By leveraging conditional independencies we can effectively localise the effect of an intervention to a specific part of a larger graph. This gives rise to the do-calculus [116]. The do calculus consists of three rules. They are derived from the causal information encoded in a causal network and the properties of d-separation and do not require any additional assumptions other than that of specifying the causal network.

3.1.1 Independence in Bayesian networks: D-separation

Many causal algorithms are based on leveraging the independence properties encoded in Bayesian networks. Therefor, in this section, we briefly review the key properties of Bayesian networks. A more thorough introduction (including proofs) can be found in [95]. Recall that a Bayesian network is a way of representing the joint distribution over its variables in a way which highlights conditional independencies between them.

Theorem 6. ***(Local Markov condition)*** *Given a Bayesian network G with nodes $X_1...X_N$, each variable X_i is independent of its non-decedents given its parents in G for all distributions $P(X_1...X_N)$ that are compatible with G.*

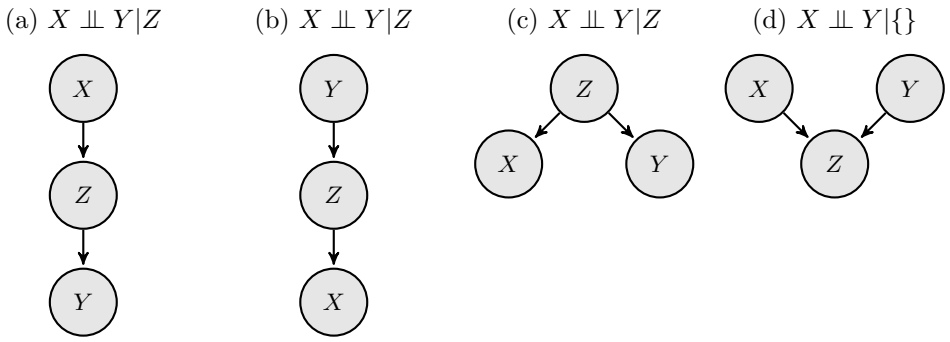
The set of conditional independence relations given by the local Markov condition can enforce additional independencies that also hold in all distributions that are compatible with G. D-separation is an algorithm that extends the local Markov property to find these additional independencies. It provides us with a simple way of reading from a network if a given conditional independence statement is true in all distributions compatible with that network.

The statement that X is conditionally independent of Y given Z implies that if we know Z, learning the value of Y gives us no additional information about X. From a graphical perspective you can think of this as Z blocks the flow of information from X to Y in the network. Figure 3.1 shows all possible network paths from a variable X to Y via Z. In figures (a) to (c) the path is blocked if we condition on Z and unblocked otherwise. In figure (d) the path is unblocked if we condition on Z and blocked otherwise.

The structure in figure 3.1d is referred to as a collider or v-structure. The somewhat counter-intuitive result that conditioning on Z introduces dependence between X and Y is called the *explaining away phenomena*. As an example, consider a scholarship available to female or disad-

vantaged students. Let X be gender, Y be family background and Z receipt of the scholarship. There are roughly equal numbers of boys and girls in both poor and wealthy families so X and Y are independent. However, if we know a student is receiving a scholarship then learning that they are male increases the probability that they are disadvantaged.

Figure 3.1: All possible two edge paths from X to Y via Z



Definition 7 (unblocked path)**.** A path from X to Y is a sequence of edges linking adjacent nodes starting at X and finishing at Y, $(X, V_1, V_2 ... V_k, Y)$. It is unblocked if every triple, $X - V_1 - V_2, V_1 - V_2 - V3, ..., V_{k-1} - V_k - Y$ in the path is unblocked (each triple will belong to one of the cases in figure 3.1)

Definition 8 (d-separation)**.** The variables $\boldsymbol{X}$ are d-separated from $\boldsymbol{Y}$ given $\boldsymbol{Z}$ in the network G if, there are no unblocked paths from any $X \in \boldsymbol{X}$ to any $Y \in \boldsymbol{Y}$ after conditioning on $\boldsymbol{Z}$.

Theorem 9 (d-separation and conditional independence)**.** *If a set of variables $\boldsymbol{Z}$ d-separates $\boldsymbol{X}$ and $\boldsymbol{Y}$ in a Bayesian network G then $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})$ in all distributions P compatible with G. Conversely, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-connected (not d-separated) given $\boldsymbol{Z}$ then it is possible to construct a distribution P' that factorises over G in which they are dependent.*

Theorem 9 says that independencies implied by d-separation on a graph hold in every distribution that can be factored over that graph and that if $(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})$ in *all* distributions that can be factored over G then they are d-separated in G. If we denote the independencies implied by d-separation in a graph by $\mathcal{I}(G)$ and the set of independencies in a distribution by $\mathcal{I}(P)$ then $\mathcal{I}(G) \subseteq \mathcal{I}(P)$.

If $\mathcal{I}(G) = \mathcal{I}(P)$ then G is called a perfect map for P. However, it is possible to construct distributions that do not have a perfect map, that is they contain conditional independencies that cannot be represented by d-separation. A particular case in which this occurs is when there are deterministic relationships between variables. If we have a Bayesian network G in which we specify that some nodes are deterministic we cannot conclude that if X and Y are d-connected then there exists a distribution P' *consistent* with G in which they are dependent. This does not conflict with theorem 9 as *consistent* in this setting requires that P' both factorises over G and satisfies the specified the deterministic relations between variables. This subtlety led to confusion in assessing what independencies hold between counterfactuals via twin networks [116, 127] and demonstrates the caution required in using d-connecteness to assert lack of independence. D-separation can be extended to compute the additional independencies implied by a graph in which certain nodes are known to be deterministic [65].

3.2 The Do Calculus

The do-calculus is a set of three rules [115] that can be applied to simplify the expression for an interventional distribution. If by repeated application of the do-calculus, along with standard probability transformations, we can obtain an expression containing only observational quantities then we can use it to estimate the interventional distribution from observational data. Let $\boldsymbol{X}$,$\boldsymbol{Y}$,$\boldsymbol{Z}$ and $\boldsymbol{W}$ be disjoint sets of variables in a causal graph G. We denote the graph G after the an intervention $do(\boldsymbol{X})$, which has the effect of removing all edges into variables in $\boldsymbol{X}$, as $G_{\overline{\boldsymbol{X}}}$

Rule 1: (adding or removing evidence)

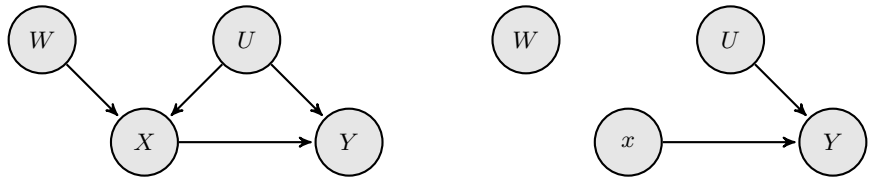
Rule 1 allows us to remove (or insert) observational evidence from the right hand side of a conditional interventional distribution. It follows directly from the fact that the relationship between d-separation in a network and independence in the corresponding probability distribution still applies after an intervention.

If $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}|\boldsymbol{Z}, \boldsymbol{X})$ in $G_{\overline{\boldsymbol{X}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{Z} = \boldsymbol{z}, \boldsymbol{W} = \boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{Z} = \boldsymbol{z}) \tag{3.1}$$

Figure 3.2: Rule 1 example. $(Y \perp\!\!\!\perp W|X)$ in $G_{\overline{\boldsymbol{X}}} \implies \mathrm{P}\,(Y|do(X), W) = \mathrm{P}\,(Y|do(X))$

(a) original network, G (b) network after the intervention $do(X = x)$, $G_{\overline{X}}$



Rule 2: (exchanging actions with observations)

Rule 2 describes when conditioning on $\boldsymbol{X} = \boldsymbol{x}$ and intervening, $do(\boldsymbol{X} = \boldsymbol{x})$, have the same effect on the distribution over $\boldsymbol{Y}$. Let $G_{\underline{\boldsymbol{X}}}$ denote the causal graph G with all edges *leaving* X removed.

If $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{W})$ in $G_{\underline{\boldsymbol{X}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{W}) = P(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{W}) \tag{3.2}$$

The intuition behind this is that interventional distributions differ from observational ones due to the presence of indirect paths between X and Y. Observing a variable X provides information about Y both directly and indirectly, by changing our belief about the distribution of the parents of X. However setting X tells us nothing about its parents and therefor affects Y only via direct paths out of X. Removing edges *leaving* X removes all the direct paths out of X. If X is then independent of Y (conditional on W), that indicates there are no indirect paths. This implies conditioning on X is equivalent to setting X (given W).

Equation 3.2 does not cover cases where acting on one set of variables allows us to replace acting on another set with conditioning (see figure 3.4). The general form of rule 2 is given in equation 3.3.

If $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}|\boldsymbol{W}, \boldsymbol{Z})$ in $G_{\underline{\boldsymbol{X}}\overline{\boldsymbol{Z}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x}), do(\boldsymbol{Z} = \boldsymbol{z}), \boldsymbol{W}) = P(\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}, do(\boldsymbol{Z} = \boldsymbol{z}), \boldsymbol{W}) \tag{3.3}$$

Figure 3.3: An example of rule 2 with a single intervention $(Y \perp\!\!\!\perp X|W)$ in $G_{\underline{\boldsymbol{X}}} \implies \mathrm{P}(Y|do(X), W) = \mathrm{P}(Y|X, W)$. In this example, observing X provides information about Y both directly and indirectly, because knowing X tells us something about W which also influences Y. If we condition on W, we block this indirect path.

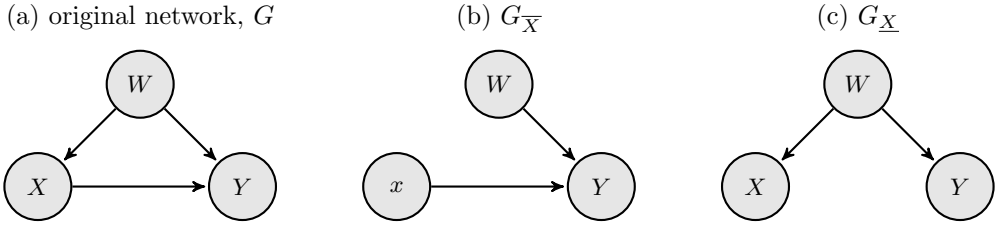
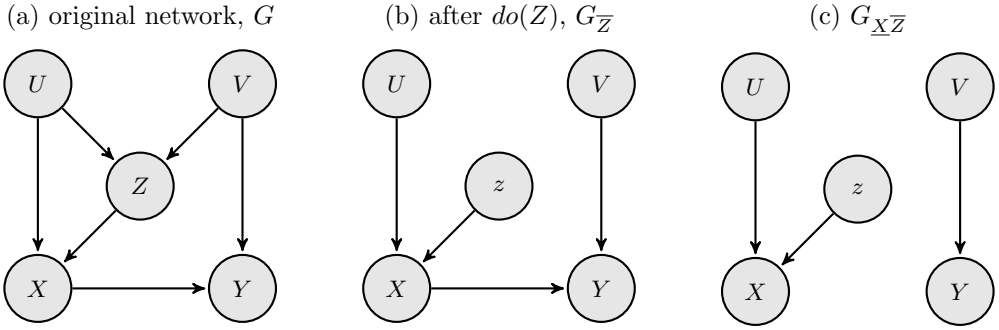


Figure 3.4: An example of applying equation 3.3. In this case $(Y \perp\!\!\!\perp X|Z)$ in $G_{\underline{X}\overline{Z}} \implies \mathrm{P}(Y|do(X = x), do(Z = z)) = \mathrm{P}(Y|X = x, do(Z = z))$. Observing, rather than intervening, on Z would not have allowed us to exchange $do(X = x)$ for $X = x$. Conditioning on Z does block the indirect path $X - Z - V - Y$ but opens $X - U - Z - V - Y$.



Rule 3: (adding or removing actions)

This rule describes cases where the intervention $do(\boldsymbol{X} = \boldsymbol{x})$ has no effect on the distribution of the outcome $\boldsymbol{Y}$. A simple case of rule 3 is given in equation 3.4. If $\boldsymbol{Y}$ is independent of $\boldsymbol{X}$ in G after removing links entering $\boldsymbol{X}$ then can be no direct path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ and any intervention on $\boldsymbol{X}$ will not affect $\boldsymbol{Y}$.

if $(\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X})$ in $G_{\overline{\boldsymbol{X}}}$:

$$\mathrm{P}(\boldsymbol{Y}|do(\boldsymbol{X} = \boldsymbol{x})) = \mathrm{P}(\boldsymbol{Y}) \tag{3.4}$$

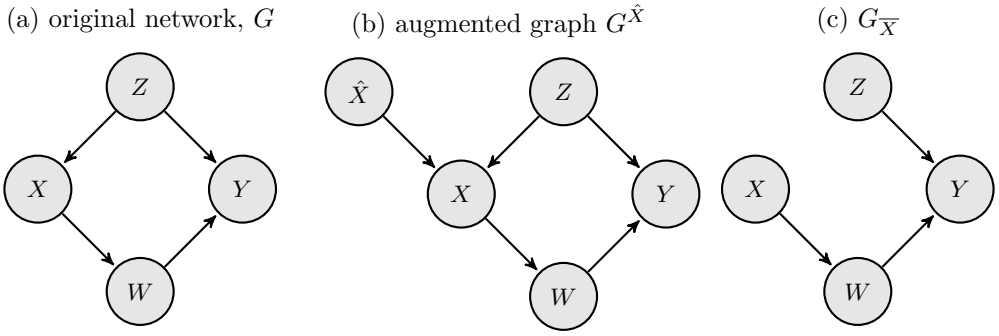
The general case of rule 3 is easier to state by explicitly representing the intervention in the graphical model. Let $G^{\hat{\boldsymbol{X}}}$ denote the graph G after adding a variable $\hat{X}_i$ as a parent of each variable $X_i \in \boldsymbol{X}$ (see figure 3.5b). The variable $\hat{X}_i$ can be thought of as representing the mechanism by which X_i takes its value, either by being set via intervention or as a stochastic function of its other parents [95].

if $(\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}}|\boldsymbol{Z}, \boldsymbol{W})$ in $G^{\hat{\boldsymbol{X}}}_{\overline{\boldsymbol{Z}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{Z} = \boldsymbol{z}), do(\boldsymbol{X} = \boldsymbol{x}), \boldsymbol{W} = \boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{Z} = \boldsymbol{z}), \boldsymbol{W} = \boldsymbol{w}) \tag{3.5}$$

The statement that $\boldsymbol{Y} \perp\!\!\!\perp \hat{\boldsymbol{X}}$ (without conditioning on $\boldsymbol{X}$) implies that there is no unblocked path from $\boldsymbol{X}$ to $\boldsymbol{Y}$ in G which *includes* an arrow leaving $\boldsymbol{X}$. These are the only paths by which intervening in $\boldsymbol{X}$ can effect $\boldsymbol{Y}$.

Figure 3.5: Example application of equation 3.5. $(Y \perp\!\!\!\perp \hat{X}|W,Z) \implies \mathrm{P}\left(Y|do(X),W,Z\right) = \mathrm{P}\left(Y|W,Z\right)$. We have to condition on Z because conditioning on W blocks the path $\hat{X}-X-W-Y$ but opens $\hat{X}-X-Z-Y$.



3.2.1 Identifiability

A natural question to ask is, given a set of assumptions about the causal graph, is it possible to estimate a given interventional distribution from observational data? This is the identifiablity problem. It asks if we can obtain an unbiased point estimate for the causal query of interest in the infinite data limit. A query is non-parametrically identifiable if it is identifiable without assumptions about the functional form of the dependencies between variables in the graph.

Definition 10 (Non-parametric identifiablity)**.** Let G be a causal graph containing observed variables $\boldsymbol{V}$ and latent variables $\boldsymbol{U}$ and let $\mathrm{P}\left(\cdot\right)$ be any positive distribution over $\boldsymbol{V}$. A causal query of the form $\mathrm{P}\left(\boldsymbol{Y}|do(\boldsymbol{X}),\boldsymbol{W}\right)$, where $\boldsymbol{Y}$,$\boldsymbol{X}$ and $\boldsymbol{W}$ are disjoint subsets of $\boldsymbol{V}$, is non-parametrically identifiable if it is uniquely determined by $\mathrm{P}\left(\cdot\right)$ and G.

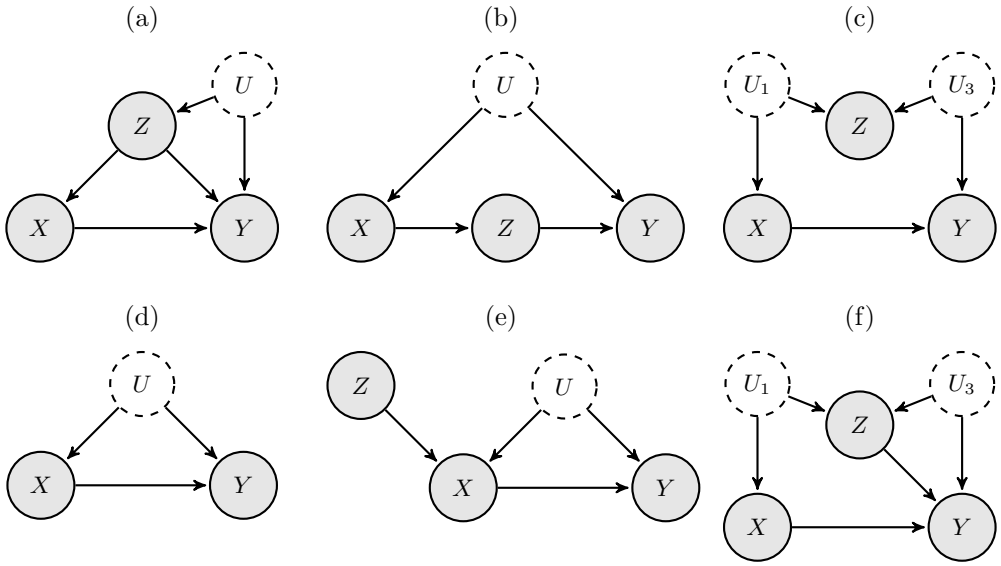
The question of non-parametric identifiability is solved. The do calculus is complete [144, 80]. A problem is identifiable if and only if the interventional distribution of interest can be transformed into term containing only observational quantities via repeated application of the do-calculus. There is a polynomial time algorithm [143] based on these properties that, for a given network and interventional (do-type) query, can:

1. determine if the query can be translated into an expression involving only distributions over observed variables. In other words, determine if the query is identifiable given the assumptions encoded by the network, and
2. if it is identifiable, return the required expression.

Figure 3.6 shows some examples of identifiable and non identifiable queries. I have created a javascript implementation of the identifiability algorithm [143] on which you can test your own queries http://finnhacks42.github.io.

Many interesting questions relating to identifiability remain open. What is the minimal (by some metric) additional information that would be required to make a non-identifiable query identifiable? What if we assume various restrictions on the functional form of the relationships between the variables? Some queries which are not identifiable non-parametrically can be identified by additional assumptions such as linearity. A complete algorithm for the problem of linear identifiability is yet to be found, despite a rich body of work [41, 160, 50].

Figure 3.6: Examples of identifiable and non-identifiable queries. In subfigures (a), (b) and (c) the causal query $\mathrm{P}(Y|do(X))$ is identifiable. In subfigures (d), (e) and (f) it is not.



Although identifiability is a natural and important question to ask, it does not partition causal questions into solvable and unsolvable. Estimators for identifiable queries can be slow to converge and we may be able to obtain useful bounds on causal effects in cases where point estimates are not identified.

3.3 Estimation

3.3.1 Defining causal effects

So far we have described causal effect estimation in term of identifying the interventional distribution $\mathrm{P}(Y|do(X))$ from observational data. This interventional distribution is in fact a family of distributions parameterised by the value, x, to which the treatment variable X is set. From a decision theoretic viewpoint, we can select an optimal action x by specifying a utility function $\mathcal{U} : y \in \mathcal{Y} \rightarrow \mathbb{R}$ that assigns a value to each outcome y and then selecting the action that maximises the expected utility.

$$x* = \arg\max_x \mathbb{E}_{y \sim \mathrm{P}(Y|do(X=x))} [\mathcal{U}(y)] \tag{3.6}$$

Frequently however, studies wish to define and estimate a causal effect without reference to a specific utility function. There are a variety of ways of defining causal effects that can be viewed as different ways of summarising the family of interventional distributions. For a binary treatment variable X, the average causal effect, ACE [1] is defined as:

$$ACE = \mathbb{E}[Y|do(X=1)] - \mathbb{E}[Y|do(X=0)] \tag{3.7}$$

[1] also referred to as the average treatment effect (ATE)

Assuming the expectations in equation 3.7 are well defined, the ACE captures the shift in the mean outcome that arises from varying X. It does not capture changes in variance or higher moments of the distribution. The ACE can be generalised to non-discrete interventions by considering the effect on the expectation of Y of an infinitesimal change in x. If X is linearly related to Y then the ACE is constant and equivalent to the corresponding coefficient in the linear structural equation model.

$$ACE(x) = \frac{d}{dx}\mathbb{E}\left[Y|do(X = x)\right] \tag{3.8}$$

The average causal effect is often introduced as the average over individual causal effects as discussed in section 2.2. Individual causal effects are deterministic and cannot be expressed as properties of the interventional distribution. However we can personalise the average causal effect by stating it with respect to some observed context. I will refer to this as the personalised causal effect (PCE).[2]

$$PCE(z) = \mathbb{E}\left[Y|do(X = 1), z\right] - \mathbb{E}\left[Y|do(X = 0), z\right] \tag{3.9}$$

In some cases we may be interested in the average causal effect for some sub-group of the population. A particularly common example of this is the average treatment effect of the treatment of the treated (ATT). This would be the key quantity of interest if we had to decide whether or not to continue providing a program or treatment for which we could not control the treatment assignment process.

$$ATT = \mathbb{E}_{z\sim \mathrm{P}(Z|x=1)}\left[Y|do(X = 1)\right] - \mathbb{E}_{z\sim \mathrm{P}(Z|x=1)}\left[Y|do(X = 0)\right] \tag{3.10}$$

Causal effects can also be written in terms of counterfactuals. The ACE is $\mathbb{E}\left[Y^1 - Y^0\right]$. We could estimate the ratio of expectations $\frac{\mathbb{E}[Y^1]}{\mathbb{E}[Y^0]}$ instead of the difference. However, the quantity $\mathbb{E}\left[\frac{Y^1}{Y^0}\right]$ depends on the joint distribution over the counterfactual variables (Y^1, Y^0) and thus cannot be computed from the interventional distribution.

Another way of conceptualising causal effects is as a property indicating the strength of the causal link between two variables. This notion is complex to formalise when the relationship between variables is non-linear. Suppose $Y = X \oplus Z$ with $P(Z = 1) = \frac{1}{2}$, the interventional distributions over X are identical after marginalising out Z. Janzing et al. [86] propose a number of postulates that a notion of causal strength could satisfy, demonstrate why previous measures fail these postulates and propose an alternative based on information flow.

3.3.2 Estimating causal effects by adjusting for confounding variables

Probably the two most frequently applied approaches to estimating causal effects from observational data are instrumental variables and adjusting for confounding factors. Instrumental variables correspond to the graph in figure 3.6e, which is not identifiable without parametric assumptions, however they can provide tight bounds. Adjusting for confounding equates to

[2] This quantity is sometimes called the conditional average treatment effect (CATE), however that term is also used for the sample rather than population effect.

Figure 3.7

(a) There can be multiple valid adjustment sets. (b) Conditioning on Z opens the backdoor path from X to Y. Z_1 or Z_2 or $\{Z_1, Z_2\}$ all block the backdoor path $X - U_1 - Z - U_2 - Y$

identifying a set of variables $\boldsymbol{Z}$ such that the ignorability assumption discussed in section 2.2 holds. This corresponds to a simple graphical test known as the backdoor criterion [116]. The setting is also referred to as unconfounded.

Theorem 11 (The backdoor criterion)**.** *[116] Let $\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{Y}$ be disjoint sets of vertices in a causal graph G. If $\boldsymbol{Z}$ blocks (see Definition 7) for every path from X_i to Y_j that contains a link into X_i, for every pair $(X_i \in \boldsymbol{X}, Y_j \in \boldsymbol{Y})$, and no node in $\boldsymbol{Z}$ is a decedent of a node in $\boldsymbol{X}$ then the backdoor criterion is satisfied and;*

$$\mathrm{P}\left(\boldsymbol{y}|do(\boldsymbol{x})\right) = \sum_{\boldsymbol{z}} \mathrm{P}\left(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}\right) \mathrm{P}\left(\boldsymbol{z}\right) \tag{3.11}$$

The backdoor criterion derives from rule 2 of the do-calculus. Selecting which covariates should be adjusted for to estimate a causal effect reduces to identifying a set which satisfies the backdoor criterion. There may be more than one valid adjustment set, see figure 3.7a. The seemingly simple problem of determining if a variable should be adjusted for when estimating causal effects has been the subject of substantial debate and controversy [117]. Adjusting for the wrong variables (even pre-treatment variables) can introduce or magnify bias, see figure 3.7b. Causal graphs and the back door criterion provide a clear mechanism for deciding which variables should be adjusted for. For a practical example, see the discussion in Schisterman et al. [139] on whether birth weight should be adjusted for to estimate the causal effect of smoking on neonatal mortality.

Given that a set of variables $\boldsymbol{Z}$ satisfies the backdoor criterion (or equivalently the conditional ignorability assumption), the interventional distribution is asymptotically identifiable and can be estimated from equation 3.11. The expected value of Y after the intervention $do(X = x)$ is given by equation 3.12 and the average causal effect for a binary intervention $x \in \{0, 1\}$ is given by equation 3.13.

$$\mathbb{E}\left[Y|do(X = x)\right] = \mathbb{E}_{z \sim \mathrm{P}(\boldsymbol{Z})}\left[\mathbb{E}\left[Y|x, \boldsymbol{z}\right]\right] \tag{3.12}$$

$$ACE = \mathbb{E}_{z \sim \mathrm{P}(\boldsymbol{Z})}\left[\mathbb{E}\left[Y|1, \boldsymbol{z}\right] - \mathbb{E}\left[Y|0, \boldsymbol{z}\right]\right] \tag{3.13}$$

Assuming x and $\boldsymbol{z}$ are discrete, equation 3.12, and thus the ACE, can be estimated by selecting the data for which $X = x$, stratifying by $\boldsymbol{Z}$, then computing the mean outcome within each

stratum and finally weighting the results by the number of samples in each strata. However this approach is not workable for most real world problems with finite samples as the number of strata grows exponentially with the dimension of $\boldsymbol{Z}$ and it cannot handle continuous covariates. There is a substantial body of work within in the statistics and econometrics literature on estimating average causal effects assuming conditional ignorability, see Imbens [82] for a comprehensive review. The key approaches are based on matching on covariates, propensity score methods and regression. We now examine these approaches from a machine learning perceptive.

In standard supervised learning, we have a training set $(\boldsymbol{x_1}, y_1), ..., (\boldsymbol{x_n}, y_n)$ assumed to be sampled i.i.d from an unknown distribution $\mathrm{P}(\boldsymbol{x}, y) = \mathrm{P}(\boldsymbol{x})\,\mathrm{P}(y|\boldsymbol{x})$. The goal is to select a hypothesis $h \in \mathcal{H} : \mathcal{X} \to \mathcal{Y}$ such that, on unseen data $\sim \mathrm{P}(\boldsymbol{x}, y)$, $h(\boldsymbol{x})$ is close (by some metric) in expectation to y. In other words we wish to minimise the generalisation error $E_{out}(h)$,

$$E_{out}(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim \mathrm{P}(\boldsymbol{x})\,\mathrm{P}(y|\boldsymbol{x})}\left[L(h(\boldsymbol{x}), y)\right] \tag{3.14}$$

We cannot directly compute the generalisation error as $\mathrm{P}(\boldsymbol{x}, y)$ is unknown, we only have access to a sample. We could search over $\mathcal{H}$ and select a hypothesis $h^*(\boldsymbol{x})$ that minimises some loss function on the sample data.

$$E_{in}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(\boldsymbol{x}_i), y_i) \tag{3.15}$$

The VC-dimension of the hypothesis space provides (typically loose) bounds on the probability that $E_{out} >> E_{in}$. However, in practice, the generalisation error is usually estimated empirically from a hold-out set of the sample that was not used to train the model, or via cross-validation.

In the causal effect estimation under ignorability, we have training data $(\boldsymbol{x}_1, \boldsymbol{z}_1, y_1), ..., (\boldsymbol{x}_n, \boldsymbol{z}_n, y_n)$ sampled i.i.d from $\mathrm{P}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{z})\,\mathrm{P}(y|\boldsymbol{x}, \boldsymbol{z})$. Estimating $\mathbb{E}\left[Y|do(\boldsymbol{X} = \boldsymbol{x})\right]$ corresponds to selecting a hypothesis $h \in \mathcal{H} : \mathcal{X} \times \mathcal{Z} \to \mathcal{Y}$ that minimises;

$$E_{out} = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{z},y)\sim\delta(\boldsymbol{x}-\boldsymbol{x'})\,\mathrm{P}(\boldsymbol{z})\,\mathrm{P}(y|\boldsymbol{x},\boldsymbol{z})}\left[L_2(h(\boldsymbol{x}, \boldsymbol{z}), y)\right], \tag{3.16}$$

$$= \mathbb{E}_{(\boldsymbol{z},y)\sim\mathrm{P}(\boldsymbol{z})\,\mathrm{P}(y|\boldsymbol{x},\boldsymbol{z})}\left[L_2(h(\boldsymbol{x}, \boldsymbol{z}), y)\right], \tag{3.17}$$

Johansson et al. [89] identified that this is equivalent to the covariate shift problem. If we let $\boldsymbol{v} = (\boldsymbol{x}, \boldsymbol{z})$ then we have training data sampled from $\mathrm{P}_{train}\{\boldsymbol{v}\}\,\mathrm{P}(y|\boldsymbol{v})$ where $\mathrm{P}_{train}\{\boldsymbol{v}\} = \mathrm{P}(\boldsymbol{z})\,\mathrm{P}(\boldsymbol{x}|\boldsymbol{z})$ but at test time the data will be sampled from $\mathrm{P}_{test}\{\boldsymbol{v}\}\,\mathrm{P}(y|\boldsymbol{v})$, where $\mathrm{P}_{test}\{\boldsymbol{v}\} = \delta(\boldsymbol{x} - \boldsymbol{x'})\,\mathrm{P}(\boldsymbol{z})$. [3] With this connection to covariate shift in mind, let us return to regression, matching and propensity scores.

[3] It is not obvious that the question of estimating causal effects under ignorability entirely reduces to covariate shift. Take the case where we have a binary intervention $x \in \{0, 1\}$. Suppose we learn $h(1, \boldsymbol{z}) = \mathbb{E}\left[Y|x = 1, \boldsymbol{z}\right] + g(\boldsymbol{z})$ and $h(0, \boldsymbol{z}) = \mathbb{E}\left[Y|x = 0, \boldsymbol{z}\right] + g(\boldsymbol{z})$, then the estimated average causal effect equals the true average causal effect for any function g, $\mathbb{E}\left[h(1, \boldsymbol{z}) - h(0, \boldsymbol{z})\right] = \mathbb{E}\left[Y|x = 1, \boldsymbol{z}\right] - \mathbb{E}\left[Y|x = 0, \boldsymbol{z}\right]$. More generally, if the goal is to select an optimal action x^* from a continuous space of possible interventions we need algorithms capable of leveraging any structure in the relationship between x and y as well as a means of focusing the loss on regions of the sample likely to affect x^*.

Regression

The regression approach is to learn a function that is a good approximation to the output surface $\mathbb{E}[Y|X,Z]$. Let $f_1(z) = \mathbb{E}[Y|X=1,Z=z]$. The expectation of Y after the intervention $X=1$ is then obtained by taking the expectation with respect to Z, $\mathbb{E}[Y|do(X=1)] = \mathbb{E}_{z\sim \mathrm{P}(Z)}[\mathbb{E}[Y|X=1,z]]$. We can learn a parametric regression model $\hat{f}_1(z)$ via empirical risk minimisation.

$$\hat{f}_1(z) = h_1(z;\hat{\theta}_{obs}), \text{ where } \hat{\theta}_{obs} = \arg\min_{\theta\in\Theta}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i=1\}\,L\left(h_1(z_i;\theta),y_i\right)\right] \tag{3.18}$$

This estimator is consistent with respect to the observational distribution. As the sample size tends to infinity, $\hat{\theta}_{obs}$ approaches the parameter within the hypothesis space that minimises the expected loss given data sampled from the observational distribution.

$$\lim_{n\to\infty}\hat{\theta}_{obs} = \arg\min_{\theta\in\Theta}\mathbb{E}_{(z,y)\sim \mathrm{P}(z|x=1)\,\mathrm{P}(y|x=1,z)}\left[L\left(h_1(z;\theta),y\right)\right] \tag{3.19}$$

If the model is correctly specified such that $f_1(z) = h_1(z;\theta^*)$ for some $\theta^*\in\Theta$ then the empirical risk minimisation estimate is consistent with respect to the loss over any distribution of Z [155], including the interventional one.

$$\lim_{n\to\infty}\hat{\theta}_{obs} = \theta^* = \arg\min_{\theta\in\Theta}\mathbb{E}_{(z,y)\sim \mathrm{P}(z)\,\mathrm{P}(y|x=1,z)}\left[L\left(h_1(z;\theta),y\right)\right] \tag{3.20}$$

The average causal effect can then be estimated by,

$$\hat{\tau}_{reg} = \sum_{i=1}^{n}\left(\hat{f}_1(z_i) - \hat{f}_0(z_i)\right) \tag{3.21}$$

Regression thus has a direct causal interpretation if the parametric model is correctly specified and the covariates included form a valid backdoor adjustment set for the treatment variable of interest in the corresponding structural equation model.

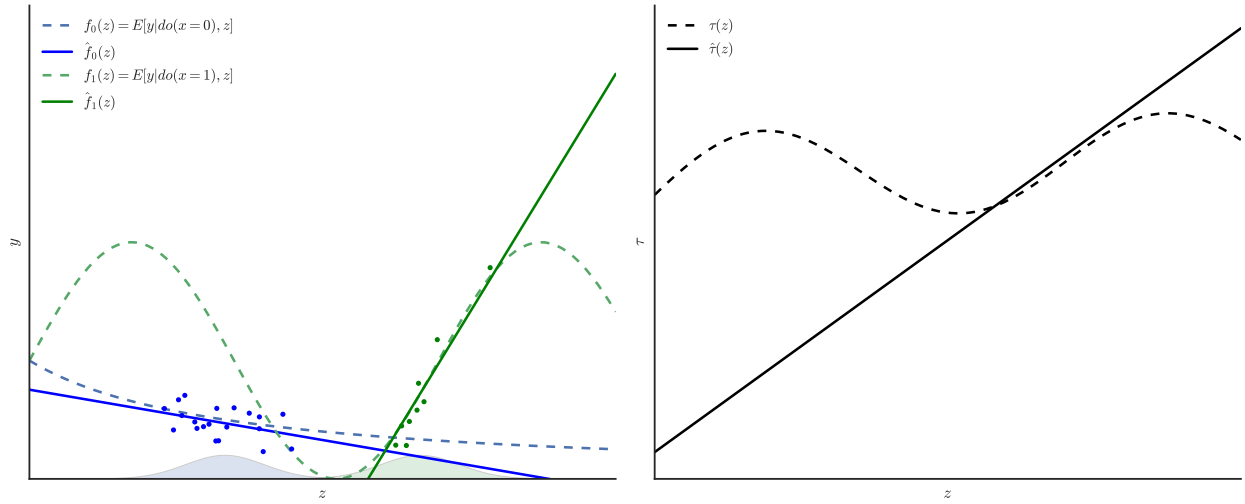
Propensity scores

If the parametric model is miss-specified then the parameter that minimises the loss depends on the distribution from which the covariates z are sampled. The model learned by ERM could perform very well in a validation set (which estimates the generalisation error over the observational distribution of (x,z)) but yield very poor estimates of the causal effect, see figure 3.8.

A general approach to estimating the expectation of some function $f(\cdot)$ with respect to data from some distribution $\mathrm{P}(\cdot)$, when we have data sampled from a different distribution $\mathrm{Q}(\cdot)$ is importance sampling [76, 95].

$$\mathbb{E}_{\boldsymbol{v}\sim \mathrm{P}(\boldsymbol{v})}[f(v)] = \mathbb{E}_{\boldsymbol{v}\sim \mathrm{Q}(\boldsymbol{v})}\left[f(v)\frac{\mathrm{P}(\boldsymbol{v})}{\mathrm{Q}(\boldsymbol{v})}\right] \tag{3.22}$$

Figure 3.8: Parametric regression may yield poor estimates of causal effects if the model is miss-specified, even if the model fits well over the domain of the training data. In this example, $\mathrm{P}(Z|X=0) \sim N(\mu_0, \sigma_0^2)$ and $\mathrm{P}(Z|X=1) \sim N(\mu_1, \sigma_1^2)$ with little overlap in the densities. If $X=0$ then $Y \sim N(f_1(x) = sin(x), \sigma_y^2)$ and if $X=1$ then $Y \sim N(f_0(x) = \frac{1}{x+1}, \sigma_y^2)$. We estimate $f_1(z)$ from the sample in which $X=1$ (green points) and $f_0(z)$ from the sample for which $X=0$ (blue points). In both cases the linear model is a good fit to the data. However, the resulting estimate of the causal effect is very poor for the lower values of z.



This importance weighting approach can be applied to the covariate shift/average causal effect problem by weighting the terms in the empirical risk minimisation estimator [155].

$$\hat{\theta}_{iw} = \arg\min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = 1\} L\left(h_1(z_i; \theta), y_i\right) \frac{\mathrm{P}\left(z_i\right) \delta(x_i - 1)}{\mathrm{P}\left(z_i\right) \mathrm{P}\left(x_i = 1 | z_i\right)} \right] \tag{3.23}$$

$$= \arg\min_{\theta \in \Theta} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = 1\} L\left(h_1(z_i; \theta), y_i\right) \frac{1}{e(z_i)} \right], \tag{3.24}$$

where $e(\boldsymbol{z})$ is the propensity score, defined by [131];

$$e(\boldsymbol{z}) \equiv \mathrm{P}\left(x = 1 | \boldsymbol{z}\right) \tag{3.25}$$

The estimator in equation 3.23 is an example of a doubly robust estimator [138**?**]. Doubly robust methods are asymptotically unbiased as long as either the regression model h or propensity score e are correctly specified [130].

The propensity score can be used to estimate the average causal effect without specifying a regression model for $\mathbb{E}\left[Y|X, \boldsymbol{Z}\right]$. Rosenbaum and Rubin [131] demonstrated that if the ignorability assumption is satisfied by conditioning on $\boldsymbol{Z}$, then it is also satisfied by conditioning on $e(\boldsymbol{z})$. This allows for estimators based on stratifying, matching or regression on the propensity score rather than the covariates $\boldsymbol{Z}$. Inverse propensity weighting can also be combined with empirical estimation of $\mathbb{E}\left[Y|X, Z\right]$ yielding the simple, albeit inefficient, estimator in equation 3.27 [82]. In some settings, such as stratified randomised trials [83] or learning from logged bandit feedback [28] the propensity score may be known. However in general, it must be estimated from data. Frequently this is done with a simple parametric model such as logistic regression, but a wide range of standard machine learning algorithms including bagging and boosting, random

forests and neural networks can also be applied [19]. Lunceford et al. [107] review the theoretical properties of key propensity score based estimators, including stratification and inverse propensity weighting.

$$\mathbb{E}\left[Y|do(X=x)\right] = \mathbb{E}_{z\sim \mathrm{P}(\boldsymbol{Z})}\left[\mathbb{E}\left[Y|x,\boldsymbol{z}\right]\right] = \mathbb{E}_{z\sim \mathrm{P}(\boldsymbol{Z}|\boldsymbol{x})}\left[\mathbb{E}\left[Y|x,\boldsymbol{z}\right]\frac{1}{e(\boldsymbol{z})}\right] \tag{3.26}$$

$$\hat{\tau}_{ip} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\mathbb{1}\{x_i=1\}\,y_i}{e(\boldsymbol{z}_i)} - \frac{\mathbb{1}\{x_i=0\}\,y_i}{1-e(\boldsymbol{z}_i)}\right) \tag{3.27}$$

Matching

There is a straightforward connection between matching and regression for causal effect estimation. If $h \in \mathcal{H} \implies h+a \in \mathcal{H}$ for any constant a and $\hat{f}$ is selected by minimising empirical risk with an L_2 loss then $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i=1\}\,\hat{f}_1(z_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i=1\}\,y_i$ [4], and equation 3.21 can be re-written as:

$$\hat{\tau}_{reg} = \sum_{i=1}^{n}\left[\mathbb{1}\{x_i=1\}\left(y_i-\hat{f}_0(z_i)\right) + \mathbb{1}\{x_i=0\}\left(\hat{f}_1(z_i)-y_i\right)\right] \tag{3.28}$$

This formulation of the regression estimator highlights the missing data aspect of casual effect estimation. For each instance, the regression models are used to estimate the counterfactual outcome had the instance received the alternate treatment. Matching estimates the counterfactual outcome for an instance from the outcome of *similar* instances that received a different treatment. Abadie and Imbens [1] analyse an estimator where both target and control instances are matched and the matching is done with replacement, let $j \in J_k(i)$ be the indices of the k instances closest to i by some metric $d(z_i, z_j)$ such that $x_i \neq x_j$.

$$\hat{\tau}_{match} = \sum_{i=1}^{n}\left[\mathbb{1}\{x_i=1\}\left(y_i - \frac{1}{k}\sum_{j\in J_k(i)} y_j\right) + \mathbb{1}\{x_i=0\}\left(\frac{1}{k}\sum_{j\in J_k(i)} y_j - y_i\right)\right] \tag{3.29}$$

This estimator is equivalent to equation 3.28 with k nearest neighbour regression. There are many variants of matching estimators utilising different distance metrics, matching with or without replacement (and in the latter case, greedy or optimal matching) and with or without discarding matches beyond some threshold [43, 132]. Although intuitive, matching estimators in general have poor large sample properties [2]. An exception is where the goal is to estimate the average treatment effect of treatment on the treated in settings where there is a large set of control instances (compared to treatment instances) [82].

The practical performance of the estimation approaches discussed in this section will depend on the sample size, dimensionality of the covariates, the complexity of the treatment assignment mechanism and output function, and the degree of prior knowledge available about these

[4] [82] state this holds for most implementations

functions. A key difference between standard machine learning problems and causal effect estimation is that when estimating causal effects we cannot directly apply cross-validation or a hold-out set for model selection because we lack samples from the counterfactual.

The significance of this should not be underestimated. Cross-validation has allowed applied machine learning to succeed with a very atheoretical approach on the basis that we can identify when a model is successful. With causal effect estimation there is no guarantee that a model that performs well at prediction (even out of sample) will accurately estimate the outcome of an intervention. Sugiyama et al. [155] propose inverse propensity weighted cross validation for the covariate shift problem. There is relatively little theory on model selection for estimating the propensity score. To achieve asymptotically unbiased estimates, the covariates should satisfy the backdoor criterion. It is also known that conditioning on instrumental variables, which directly influence X but not Y, increases variance without any reduction in bias and can increase bias if there are unmeasured confounding variables [169, 26, 118, 111]. With doubly robust estimators, one could apply an iterative approach, fitting a propensity score model, using the results for inverse propensity weighted cross-validation of the regression model and then selecting covariates for the propensity model on the assumption the estimated regression function was correct. I have found no examples of such an approach.

The performance of methods for causal effect estimation can be tested on simulated data [59, 177, 75, 49] or by comparing estimates from observational studies with the results from corresponding experiments [97, 57, 74, 73, 48, 148, 10]. Unfortunately there are a relatively small number of examples where comparable observational and experimental data are available. The results are mixed with later studies finding generally better alignment of results but it is hard to ascertain if this is due to improved methodological approaches or over-fitting to the available data-sets.

3.4 Causal Discovery

We now move to the much more general problem of learning a causal graph from observational data. In this setting we make much broader assumptions about the structure of the graph. For example, that it is acyclic or that we have no unmeasured confounding variables. We do not assume the existence or directions of any links between the variables. Amazingly, it is possible to infer some aspects of causal structure with such general assumptions. The set of conditional independence in a non-experimental data set indicates some causal structures are more likely than others. In addition, there can be subtle asymmetries in the relationship between the joint distribution of cause and effect and the distributions of cause given effect and effect given cause. These clues are the key to causal discovery algorithms.

Causal discovery is a much grander goal than causal effect estimation given a known causal network. Arguably, if achieved, it would equate to the automation of scientific discovery. We need simply supply our algorithm with a vast collection of variables (regardless of their relevance to the problem) and it would learn the causal structure and from that allow us to estimate the effects of any intervention we cared to make. Unfortunately, causal discovery is very hard. Even with the assumption that the causal graph is acyclic and there are no latent variables, the number of possible graphs grows exponentially with the square of the number of variables .

In the next sections we briefly survey the key approaches to causal discovery. We roughly divide the methods into those based on those that exploit the connection between the conditional independencies in a joint distribution and the structure of a causal model and those that leverage assumptions about the functional form of the relationships between cause and effect.

3.4.1 Conditional independence based methods

One general approach is to look for clues about the structure of the network in the conditional independence relations in the distribution. For any Bayesian network, G, (causal or not) we can read of conditional independencies in the joint distribution from the structure of the network. If a set of variables Z d-separates X and Y in G then $(X \perp\!\!\!\perp Y|Z)$ in the distribution P. However, we want to work in the other direction, from conditional independence in the distribution to the structure of the network. This requires that we assume the reverse condition: $(X \perp\!\!\!\perp Y|Z)$ in P must imply Z d-separates X and Y in G. This assumption, commonly referred to as ***faithfulness*** **??**, says there are no additional independence relations that are satisfied in P but not in all distributions $\boldsymbol{P'}$ that are compatible with G. Stating that P is faithful to G is equivalent to G is a ***perfect map*** [114] for P.

Faithfulness is an assumption. It does not always hold and we cannot verify it from the observational data we wish to use for causal inference. However, most distributions generated by a causal Bayesian network will be faithful to that network. For faithfulness to be violated, different causal effects must exactly balance one-another out. For example, consider a simple binary variable model of chocolate consumption, income and obesity (figure). If the coefficients in the conditional probability tables are just right then the direct effect of chocolate on obesity will exactly balance the indirect effect through income and obesity will appear independent of chocolate consumption. However, this independence is not stable. It would disappear under a small perturbation to any of the parameters. In discrete systems, violations correspond to the solutions to polynomial equations over values in the CPD tables and thus are a space of measure zero with respect to all possible distributions associated with the graph [95].

Given the faithfulness assumption, our causal discovery problem reduces to finding the set of Bayesian networks that have exactly the dependency structure as we observe in P. This set can also be referred to as the Markov equivalence class compatible with P.

Without hidden common causes

The strong assumption that there are no hidden variables that cause two or more variables in $\boldsymbol{V}$ significantly reduces the 'search space' of Bayesian networks we must consider. This assumption is referred to as causal sufficiency [151].

We will begin with a brute force algorithm (described as the SGS algorithm in Spirtes et al. [151] and IC algorithm in Pearl [116]). While it is impractical for all but the smallest of networks, it demonstrates key concepts that also underlie the more useful and complex algorithms we will discuss later.

The SGS (or IC) Algorithm

Input: A distribution P, over variables $\boldsymbol{V}$, that was generated by and is faithful to an (unknown) Bayesian network G

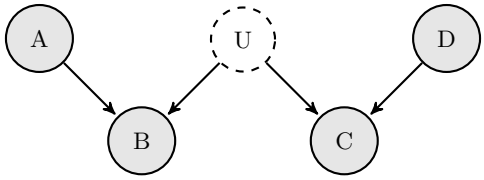
Output: A partially directed network that represents the Markov equivalence class of G

1. Join all pairs of vertices $(a, b) \in \boldsymbol{V}$ with an un-directed link to form a complete graph.
2. For each link $a - b$ search for a set $\boldsymbol{S}_{a,b} \subseteq V \setminus \{a, b\}$ that renders a and b conditionally independent. If such a set (including the empty set) exists then a and b cannot be directly connected in G so delete the link.
3. For all pairs of non-linked variables (α, β) with a common neighbour, c, if $c \notin \boldsymbol{S}_{\alpha,\beta}$, then c must be a collider in the path α, c, β so add arrows to direct the links $\alpha - c$ and $\beta - c$ towards c.
4. Recursively try to orient any edges that remain un-directed to avoid creating cycles (because they are not there by assumption) and additional colliders (because any colliders were found in step 3).

The SGS algorithm utilises the fact that a collider structure (figure 3.1d) induces a distinct conditional independence relation. Assuming you have a consistent conditional independence test, it converges to return a partially directed network that represents the Markov equivalence class for the generating causal model. Unfortunately the number of conditional independence tests required for step 2 grows exponentially (in the worst case) with the number of variables. Not only that, but for each edge that is in the true network, the algorithm will always tests all other possible subsets of variables. If the assumption that there are no hidden common causes or that the distribution is faithful are violated, step 3 of the SGS algorithm can produce double headed arrows.

The PC algorithm Spirtes et al. [151] modifies step 2 of the SGS algorithm to utilise the fact that if two variables (a, b) are conditionally independent given some set, they will also be conditionally independent given a set that contains only variables adjacent to a or b. It also checks for low order conditional independence relations before higher order ones. This allows it to exploit any sparsity in the true network, leading to much better average case performance [151] (although the worst case, where the true network is complete, is still exponential). With finite data, the order in which the links are considered can change the output (unlike for SGS). The effect of wrongly removing a link early on flows through to later conditional independence tests by changing which nodes are considered adjacent.

Figure 3.9: A distribution faithful to this DAG is not faithful to any DAG over the variables $\{A, B, C, D\}$ after marginalising over U.



The PC Algorithm

Input: A distribution P, over variables $\boldsymbol{V} = \{V_1...V_k\}$, that was generated by and is faithful to an (unknown) Bayesian network G

Output: A partially directed network that represents the Markov equivalence class of G

1. As for SGS
2. **for** each link $a - b$:
   $n = 0$
   $\boldsymbol{A}_{a,b} = \{A_1...A_j\}$ be the set of nodes adjacent to a and/or b
   while a and b are connected and $n < j$:
   if any subset of size n of $\boldsymbol{A}$ makes a and b conditionally independent:
   delete the link
   $n = n + 1$
3. as for SGS
4. as for SGS

The PC algorithm also returns a set of Markov equivalent networks consistent with the distribution. Since we have assumed there are no hidden variables, for any single graph in this set we can calculate causal effects from the truncated product formula 2.3. We can then bound the true causal effect by combining the results for the all the networks. This procedure is the IDA algorithm [109] and has been found to outperform standard regularisation techniques at finding causal effects in a high-dimensional yeast gene expression data set [108]. An implementation is available in the R package [91]

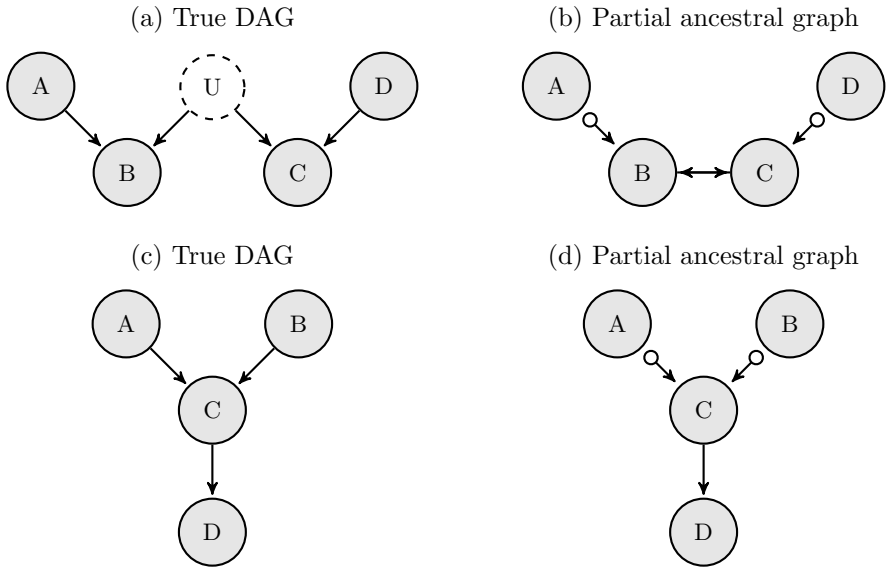
With hidden variables

There are an number of difficulties in extending the approach of the last section to deal with the case where there are latent variables. With an unknown number of hidden variables there are infinity many possible structures to search over. In addition, the space of causal networks is not closed under marginalisation. If we have a distribution that $P'(\boldsymbol{O}, \boldsymbol{U})$ generated by and is faithful to a network G the distribution $P(\boldsymbol{O})$, that results from marginalising over $\boldsymbol{U}$, may not be faithful to any Bayesian network (see figure 3.9). The key to constraining the space of possible models is that many latent structures are equivalent (under transforms of the hidden variables).

Theorem 12. *Verma [164] For every latent structure there is a dependency equivalent structure such that every latent (unobserved) variable is a root node with exactly two children .*

Since we only care about the causal relationships between observed variables, it is sufficient to search over networks where any hidden variables have no parents and directly cause two of

Figure 3.10: FCI examples: true graph and FCI output



the observed variables. Instead of representing hidden variables explicitly we can capture the necessary independence relations with a more general graphical model that supports bi-directed edges that play the role of a hidden confounding variable. These models, referred to as maximal ancestral graphs (MAGs) are closed under marginalisation and conditioning.

For any DAG with latent (and selection) variables there is a unique MAG [126]. This makes it possible to extend the PC algorithm to latent structures, resulting in the FCI algorithm [151]. The logic behind the algorithm is very similar. Certain structures are ruled out as a consequence of being inconstant with the observed conditional independence relations. The output is an equivalence class of MAGs, which can be represented graphically as a partial ancestral graph PAG [150]. Assuming there are no selection variables, the PAG can contain four types of link:

1. $X \to Y$, meaning X causes Y
2. $X \leftrightarrow Y$, meaning there is a latent variable that causes X and Y.
3. $X \circ\!\!\to Y$, either X causes Y or a latent variable causes both.
4. $X \circ\!\!-\!\!\circ Y$, either X causes Y or Y causes X or a latent variable causes both.

The circles indicate where it is ambiguous if there should be an arrowhead (IE where there is one in some MAGs and not in others in the equivalence class). Counter-intuitively it is sometimes possible to rule out or confirm the existence of a confounding variable and fully determine the causal type of a link (see examples in figure 3.10).

The FCI algorithm can be made complete such that it discovers all aspects of the true causal structure that are identifiable from the conditional independence relations of a distribution over observed variables and the faithfulness assumption [174]. More recently Colombo et al. [45] have proposed the RFCI algorithm, which in some cases returns more ambiguous links than FCI but is substantially faster. Claassen et al. [42] point out that the problem of learning sparse causal networks from data is not NP-hard and propose the FCI+ algorithm, that requires $O(N^{2(k+2)})$ conditional independence tests, where k is the maximum node degree over the observed variables.

Latent variables can create constraints on the marginal distribution over the observed variables that cannot be expressed in term of conditional independencies. These generalised constraints

can be expressed and leveraged within nested Markov models [145, 142]

All the algorithms discussed in this section rely on being able to perform conditional independence tests. This is non-trivial with high dimensional data. If the functional relationship between the variables is linear with Gaussian noise then the network represents a multivariate normal distribution and a pair of variables are conditionally independent if and only if the corresponding entry in the inverse correlation matrix is non-zero [95]. Where the functions are non-linear one can apply kernelised independence tests [67, 176]

3.4.2 Discovery with functional models

The algorithms we have considered so far return a Markov equivalence class. They cannot distinguish between two models that result in the same set of conditional independence relations. Consider the very simple case where there are only two variables and the possible causal structures are $X \to Y$ or $Y \to X$. These models have the same dependency structure but in one case $P(Y|do(X)) = P(Y|X)$ and in the other $P(Y|do(X)) = P(Y)$. No algorithm relying purely on conditional independence relations can separate these two cases.

Let us focus only on the two variable case $X \to Y$ or $Y \to X$. What possible clues could there be in the distribution $P(X, Y)$ that could indicate which causal model it was generated from? Recall the functional definition of causality (section 2.3). There are a number of assumptions about the form of the functions that can allow us to identify the causal direction: non-invertable functions, additive noise [78], post-non-linear additive noise [175] or linear models with non-Gaussian noise [77].

The causal direction can also be identified via a connection between casual discovery and semi-supervised learning [88]. Suppose we are trying to learn $\mathrm{P}(Y|X)$. The goal of semi-supervised learning is to improve our estimate of $\mathrm{P}(Y|X)$ by leveraging additional data sampled from $\mathrm{P}(X)$. However, if the true causal model is $X \to Y$ then their is some function mapping values of X to Y which should be invariant to any changes in the input distribution $\mathrm{P}(X)$. Therefor $\mathrm{P}(X)$ should be independent of $\mathrm{P}(Y|X)$ and semi-supervised learning should not perform any better than standard supervised learning. However if the true causal model is $Y \to X$ then variations in the $\mathrm{P}(X)$ can result from both the input distribution over Y and the mapping from X to Y and semi-supervised learning could help. This assumption of independence of mechanism and input can also allow the identification of the causal direction in SEMs even where the functions are deterministic and invertable [46]. Janzing et al. [87] leverage an information geometric viewpoint of the independence of mechanism and input to infer the causal direction between a pair of associated variables.

Instead of positing a functional restriction on the relationship between variables and then developing theory to exploit that assumption, [106] propose learning what the causal relationship looks like from data. They assume there will be a difference between the relationship of $\mathrm{P}(X)$, $\mathrm{P}(Y)$ and $\mathrm{P}(X|Y)$ between $X \to Y$ versus $Y \to X$. Their algorithm requires a data set in which each row is itself a data set consisting of pairs of variables (x_i, y_i) with a label indicating the direction of causality between X and Y. They use a kernel mean embedding to represent the distributions $\mathrm{P}(X)$, $\mathrm{P}(Y)$ and $\mathrm{P}(X|Y)$ as features for each individual sub-data set and train an algorithm to learn the direction of causality. Unfortunately we do not have a large collection of data sets where the causal direction is known to train such a model. Lopez-Paz et al. [106] instead use a simulated data set so their model will necessarily be based on the assumptions they make when generating the data. Nonetheless this approach makes it possible to rapidly construct a model from a wide range of possible assumptions, without doing a lot of theory to design a specific algorithm optimised to that setting.

[120] have extended results from the bi-variate case to the multivariate setting. They show that if we can come up with a condition that guarantees identifiability for the bi-variate case, we can extend that result to get the conditions under which the multivariate case is identifiable. They build on this to develop an algorithm that allows the construction of causal graphs based on the additive noise assumption.

Chapter 4

The interventionist viewpoint

The previous sections all focus on aspects of the question; how can we estimate the effect of an intervention in a system from data collected prior to taking it. There is an obvious alternative. Instead of trying to infer the outcome of an intervention from passive observations one can intervene and see what happens. There are three key differences between observing a system and explicitly intervening in it. Firstly, when we intervene we choose which actions to take and thus have control over which data points we obtain to learn from. Selecting data points optimally for learning is the focus of the optimal experimental design literature within statistics [123] and the active learning literature in machine learning [140]. Secondly, explicitly choosing interventions yields a perfect model of the probability with which each action is selected, given any context, allowing control over confounding bias. Finally, when we are intervening in a system we typically care about the impact of our actions on the system in addition to optimising learning. For example, in a drug trial , assigning people a sub-optimal treatment has real world costs. This leads to a trade-off between exploiting the best known action so far and exploring alternative actions about which we are less certain. This exploration-exploitation trade-off lies at the heart of the field of reinforcement learning [156].

4.1 Randomised experiments

Randomised controlled trials are often presented as the gold standard for determining causal effects. What is it about randomisation that makes it so important when it comes to causality? The graphical model for a randomised controlled experiment is shown in figure 4.1. If we assume perfect compliance (everyone takes the treatment that we select for them) then we have a perfect model for the treatment assignment process. Since treatment is assigned randomly, there can be no other variables that influence it and thus no confounding variables that effect both treatment and outcome.

Randomisation does not ensure target and control group are exactly alike. The more other

Figure 4.1: causal network for a randomised experiment

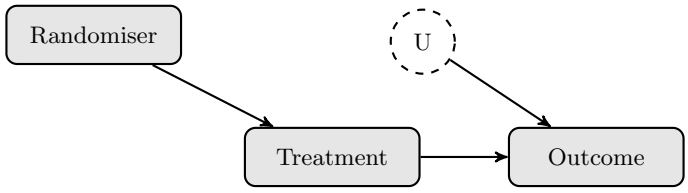
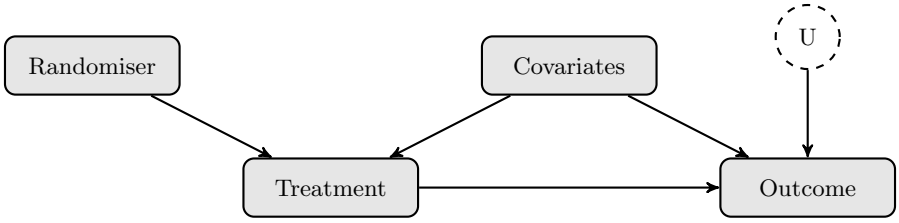


Figure 4.2: causal network for a stratified randomised experiment if the probability an individual is assigned a given treatment depends on some covariates.



features (observed or latent) influence the outcome, the more likely it is that there will be a significant difference in the joint distribution of these variables between the target and control groups in a finite data sample. However, the variance in the outcome, within both the target and control groups, also increases. The net result is increased variance (but not bias) in the estimate of causal effects.

Stratified randomised experiments address the issue of variance due to covariate imbalance by randomly allocating treatment conditional on covariates believed to influence the outcome of interest. If we stratify in such a way that the probability an instance receives a given treatment is independent of its covariates, for example, by grouping instances by each assignment to the covariates and then assigning treatment randomly with fixed probabilities, the causal graphical model in figure 4.1 still holds and we can estimate the average causal effects directly from the differences in outcome across treatments. More complex stratification strategies can introduce a backdoor path from treatment to outcome via the covariates on which treatment is stratified, see figure 4.2, necessitating that one condition on these covariates in computing the average casual effect in the same way as for estimating causal effects under under ignorability §3.3.2. The key difference is that the propensity score is known, as it is designed by the experimenter, and there are guaranteed (rather than assumed) to be no latent confounding variables (that influence both treatment and outcome). See Imbens and Rubin [83] for a discussion of the trade-offs between stratified versus completely random experiments.

The benefit provided by randomisation in breaking the link between the treatment variable and any latent confounders should not be understated. The possibility of unobserved confounders cannot be empirically ruled out from observational data [116] (there is no test for confounding). This means causal estimates from non-experimental data are always subject to the criticism that an important confounder may have been overlooked or not properly adjusted for. However, randomised experiments do have some limitations.

4.1.1 Limitations of randomised experiments

The idealised notion of an experiment represented by figure 4.1 does not capture the complexities of randomised experiments in practice. There may be imperfect compliance, the treatment selected by the randomiser is not always followed, or output censoring, the experimenter is not able to observe the outcome for all units (for example if people drop out). If compliance or attrition is not random but associated with (potentially latent) variables that also effect the outcome then the problem of confounding bias returns.[1] See figure 4.3 for a the graphical model of a randomised experiment with imperfect compliance.

[1]Non-compliance is a problem if the goal is to estimate the causal effect of the treatment on the outcome but not if the goal is to estimate the causal effect of prescribing the treatment. The latter makes sense in a context where the process by which people decide whether or not to take the treatment they have been prescribed is likely to be the same if the treatment were made available more generally beyond experimental trial.

Figure 4.3: causal network for a randomised experiment with imperfect compliance

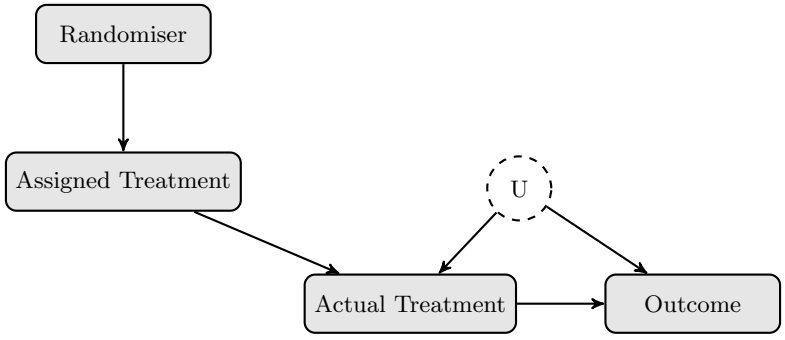


Figure 4.4: Experiments are not always ethical; an illustration of a randomised cross-over trial of parachutes for the prevention of morbidity and mortality associated with falls from large heights.

It is not always possible or ethical to conduct a randomised controlled trial as is beautifully demonstrated by the paper of Smith and Pell [147] on randomised cross-over trials of parachute use for the reduction of the mortality and morbidity associated with falls from large heights 4.4. When experimentation is possible, it is frequently difficult or expensive. This means experimental data sets are often much smaller than observational ones, limiting the complexity of models we can explore. In addition, they are often conducted on a convenient but unrepresentative sample of the broader population of interest (for example first year students at research universities). This can result in estimates with high *internal validity* [36] in that they should replicate well in a similar population, but very low *external validity*; the results may not carry over to the general population of interest. The question of whether an experiment conducted on one population can be mapped to another is referred to as the transportability problem [22] and relies on very similar assumptions and arguments to causal inference and the do-calculus.

Finally, non-adaptive randomised experiments are not optimal from either an active or reinforcement learning perspective. As an experiment proceeds information is obtained about the expectation and variance of each intervention (or treatment). Fixed experimental designs cannot make use of this information to select which intervention to try next. This results in both poorer estimates for a fixed number of experimental samples and more sub-optimal actions

during the course of the experiment.

4.2 Multi armed bandits

Multi-armed bandits address the problem of designing experiments that can adapt as samples are observed. Their introduction is generally attributed to Thompson [159]. In its classic formulation [129, 96] the (stochastic) k-armed bandit describes a sequential decision making problem, with k possible actions or arms. Each arm i is associated with a fixed but unknown reward distribution. [2] For each time step up to some horizon T the learner selects an action and receives a reward, sampled i.i.d from the marginal distribution corresponding to that action. The goal of the learner is to maximise the total reward they receive. This problem captures the exploration-exploitation trade-off, the learner must balance playing arms that have yielded good results previously with exploring arms about which they are uncertain.

Definition 13 (Stochasitic k-armed bandit problem)**.** Let $\mathcal{A} = \{1, ..., k\}$ be the set of available actions (or bandit arms) and $\mathrm{P}(\boldsymbol{y}) = \mathrm{P}(y_1, ..., y_k)$ be a joint distribution over the rewards for each action. The multi-armed bandit problem proceeds over T rounds. In each round t,

1. the learner selects an action $a_t \in \{1, ..., k\}$, based on the actions and rewards from previous time-steps and a (potentially stochastic) *policy* π
2. the world stochastically generates the rewards for each action, $[Y_{t,1}, ..., Y_{t,k}] \sim \mathrm{P}(\boldsymbol{y})$
3. the learner observes and receives (only) the reward for the selected action Y_{t,a_t}

At the end of the game the total reward obtained by the learner is $\sum_{t=1}^{T} Y_{t,a_t}$. We denote the expected reward for the action i by μ_i and the action with the highest expected reward by i^*.

The total reward a bandit algorithm/policy can expect to achieve depends on the distributions from which the rewards for each action are sampled. To account for this, the performance of bandit algorithms is quantified by the the difference between the reward obtained by the algorithm and the reward that would have been obtained by an oracle that selects the arm with the highest expected reward at every time step. This difference is known as the (cumulative) regret [3].

$$R_T = \sum_{t=1}^{T} Y_{t,i^*} - \sum_{t=1}^{T} Y_{t,a_t} \tag{4.1}$$

Both the rewards and the actions selected by the algorithm are random variables. The majority of work in the bandit literature focuses on analysing and optimising some form of the expected regret, however there has been some work that also considers the concentration of the regret [16, 14, 13]. The expectation of the regret, as defined by equation 4.1, is referred to as the pseudo-regret [30] and is given by equation 4.2. A stochastic bandit algorithm is learning if it obtains pseudo-regret that is sub-linear in T.

[2] In order to quantify the performance of bandit algorithms, some assumptions are required on the distributions from which the rewards are generated. It sufficient (but not necessary) to assume they are sub-Gaussian.

[3] The term regret is somewhat overloaded in the reinforcement learning literature. There are alternative definitions that arise in the related problems of adversarial bandits and learning from expert advice. In addition, researches often refer to the expected regret as "the regret".

Definition 14 (Pseudo-Regret)**.**

$$\bar{R}_T(\pi) = \max_{i \in \{1,\dots,k\}} \mathbb{E}\left[\sum_{t=1}^{T} Y_{t,i}\right] - \mathbb{E}\left[\sum_{t=1}^{T} Y_{t,a_t}\right] \tag{4.2}$$

$$= n\mu_{i^*} - \mathbb{E}\left[\sum_{t=1}^{T} Y_{t,a_t}\right] \tag{4.3}$$

The regret is invariant to adding a constant to the expected rewards for all actions. However, it still depends on key characteristics of the reward distributions for each action. Bandit algorithms are designed given assumptions about the form of the distributions, such as that they come from a given family (i.e Bernoulli bandits, Gaussian bandits) or that the rewards are bounded in some range. Given these assumptions, the performance of the algorithm is characterised in two ways; by the *problem dependent regret*, which typically depends on how far each arm is from optimal and by the *worst case regret*, which is the maximum regret over all possible configurations of the reward distributions (for a given horizon T and number of arms k).

4.2.1 Stochastic bandits: Approaches and results

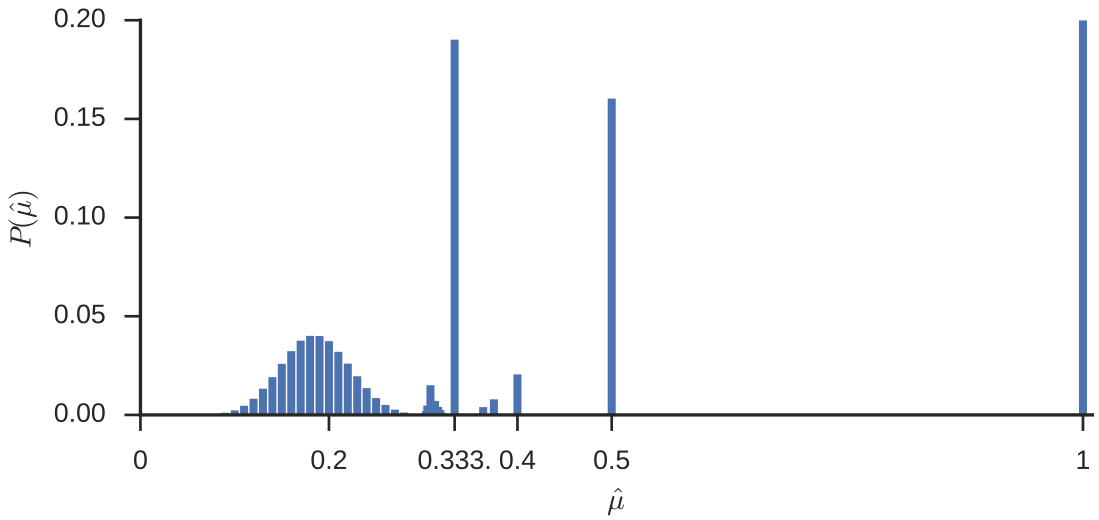
The adaptive nature of multi-armed bandit algorithms complicates the design and analyse of estimators. The action selected by an algorithm at a given timestep can depend on the history of previous actions and rewards. As a result, the probability that each action is selected evolves over time, the actions are not sampled i.i.d from a fixed distribution and the number of times each action is selected is a random variable. The expectation and variance guarantees of standard estimators do not hold in this setting, see figure 4.5 for a concrete example. This makes it very difficult to obtain an analytical expression for the expected regret for a given algorithm and problem. Instead, the focus is on computing bounds on the expected regret.

There are a few key principles that are used to guide the development of bandit algorithms. The simplest is to explicitly separate exploration from exploitation and base estimation of the expected rewards of each arm only on the data generated during exploration steps. A common example in practice is uniform exploration (or A/B testing) for some fixed period followed by selecting the action found to be best during the exploration phase. This results in simpler analysis, particularly if the number of exploration steps is fixed in advance, however it is sub-optimal, even if the exploration period is adaptive [62].

Another key approach is *optimism in the face of uncertainty*. Applied to stochastic bandits, the optimism in the face of uncertainty principle suggests computing a plausible upper-bound for the expected reward of each arm, and selecting the arm with the highest upper bound. The optimism principle encourages exploitation and exploration because a high upper bound on the expected reward for an action implies either the expected reward or the uncertainty about the reward for that action is high. Thus selecting it yields either a good reward or useful information.

Lai and Robbins [96] leveraged the optimism in the face of uncertainty principle to develop an algorithm for specific families of reward distributions, including the exponential family. They showed that, for a given bandit problem, the pseudo-regret increased with $\mathcal{O}\left(log(T)\right)$ asymptotically and proved this is asymptotically efficient. However, their algorithm is complex and memory intensive to compute as, at each timestep, it relies on the entire sequence of rewards for each arm. Agrawal [4] developed a simpler algorithm that computed upper bounds based only on the mean of previous samples for each arm, whist retaining the logarithmic dependence on T. Finally, Auer et al. [15] developed the UCB-1 algorithm, see algorithm 1, which requires

Figure 4.5: Standard empirical estimators can be biased if the number of samples, n, is not fixed in advance, but is a random variable that depends on the values of previous samples. This example plots the distribution (over 10^6 simulations) of $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim Bernoulli(0.2)$. In each simulation, we stop taking samples if the average value of X_i up to that point exceeds a threshold of 0.3 or n reaches 100. $\mathbb{E}[\hat{\mu}] = 0.439$. The estimator is substantially biased above $\mathbb{E}[X_i] = 0.2$ by the early stopping. Note that excluding experiments that were stopped early creates a bias in the opposite direction, $\mathbb{E}[\hat{\mu}|n = 100] = 0.185$, as trials that obtained positive results early are excluded. This has some interesting potential real world implications. Early stopping of clinical trials is controversial. A researcher conducting a meta-analysis who wished to avoid (rather than bound) bias due to early stopping would have to exclude not only those trials which were stopped early but those which *could* have been stopped early.



only that the reward distributions are bounded, and proved finite-time regret bounds. We now assume the rewards are bounded in $[0, 1]$. The algorithm and regret bounds can be generalised to submission reward distributions, see Bubeck and Cesa-Bianchi [30].

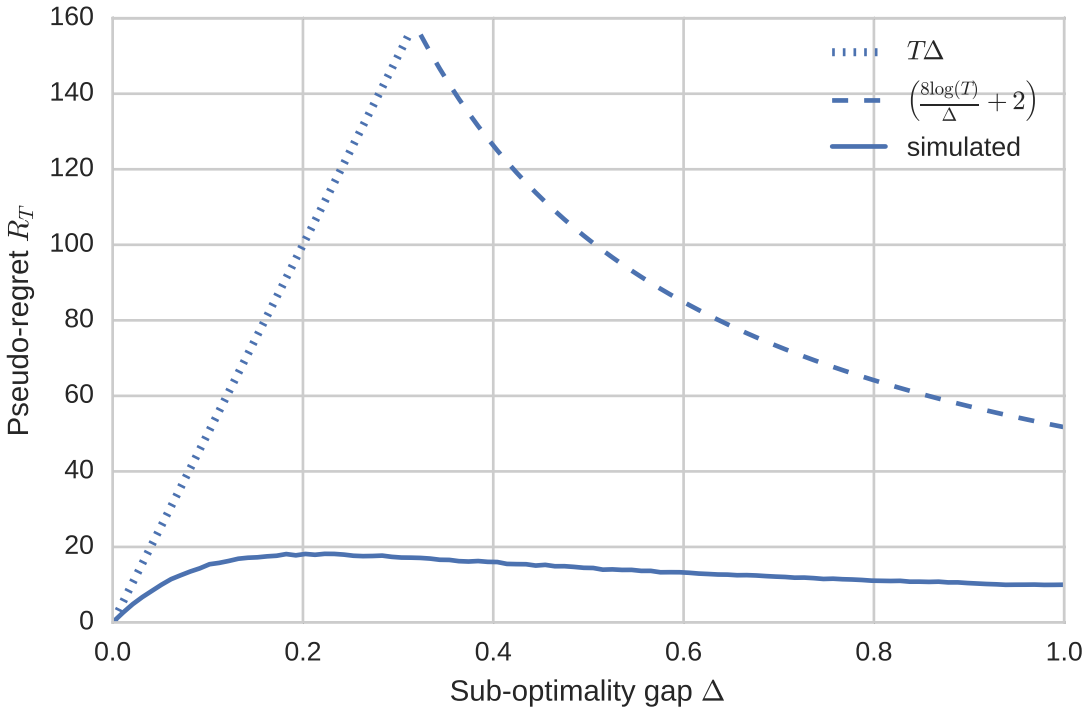
Algorithm 1 UCB-1

Input: horizon T.
Play each arm once.
for $t \in 1, \ldots, T$ **do**
  Count the number of times each arm has been selected previously $n_{t,i} = \sum_{s=1}^{t-1} \mathbb{1}\{a_s = i\}$
  Calculate the mean reward for each arm $\hat{\mu}_{t,i} = \frac{1}{n_{t,i}} \sum_{s=1}^{t-1} \mathbb{1}\{a_s = i\} Y_t$
  Select arm $a_t \in \arg\max_{i=\{1,\ldots,k\}} \left(\hat{\mu}_{t,i} + \sqrt{\frac{2 \log t}{n_{t,i}}}\right)$

Let $\Delta_i = \mu_i - \mu^*$ be degree to which each arm is sub-optimal. The problem dependent pseudo-regret for UCB-1 is bounded by equation 4.4 [30],

$$\bar{R}_T \leq \sum_{i:\Delta_i > 0} \left(\frac{8 \log(T)}{\Delta_i} + 2\right) \tag{4.4}$$

Figure 4.6: The regret bound in equation 4.4 grows as the differences between the expected rewards for each arm shrink. The solid curve shows the mean (cumulative) regret, over a 1000 simulations for a 2-armed, Bernoulli bandit with fixed horizon, $T = 500$, as a function of the difference in the expected reward for the arms Δ. The dashed curves show the corresponding upper bounds; $T\Delta$ and equation 4.4

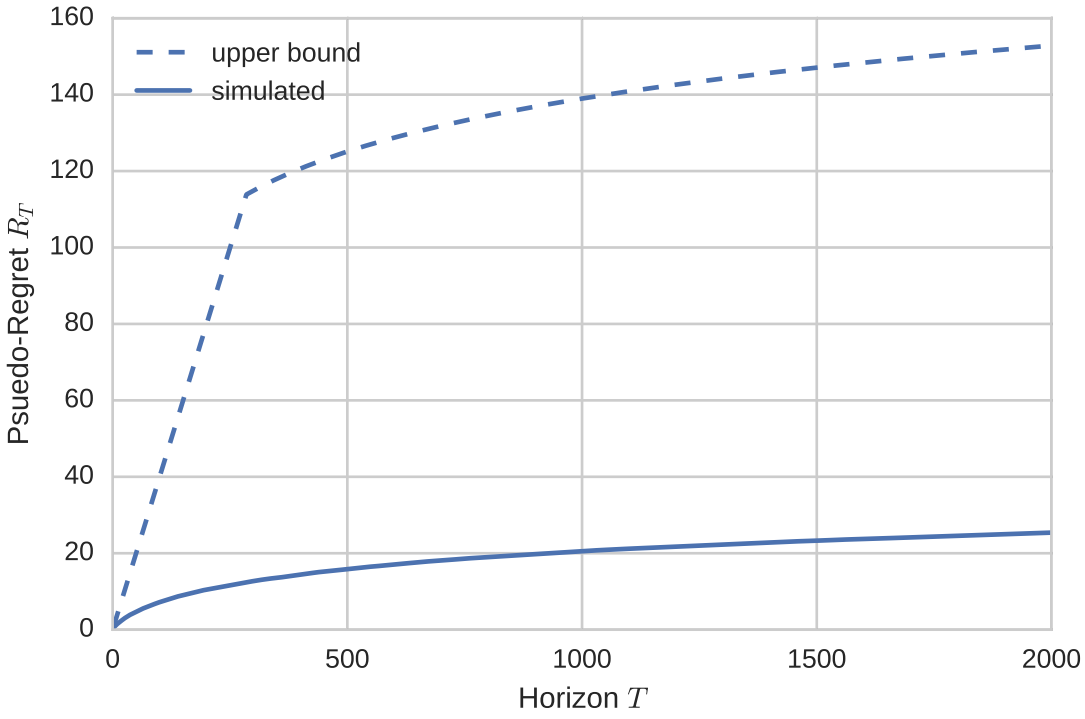


Somewhat unintuitively, the regret increases as the value of the arms gets closer together. This is because it becomes harder for the algorithm to identify the optimal arm. As the differences $\Delta_i \to 0$, the regret bound in 4.4 blows up, however the regret itself does not - since although we may not be able to distinguish arms with very small Δ_i from the optimal arm, we also do not lose much by selecting them. The worst case occurs if all arms have the same expected reward μ except for the optimal arm which has reward $\mu^* = \mu + \Delta$, where Δ is just too small for the algorithm to learn to identify which arm is optimal given the horizon T. The regret cannot exceed what would be obtained by selecting the a sub-optimal arm in every timestep, $T\Delta$, so the worst case regret is bounded by the minimum of equation 4.4 and $T\Delta$ which is maximised when they are equal, see figure 4.6. By solving this equality for Δ one can show the worst case regret is bounded by equation 4.5, see Bubeck and Cesa-Bianchi [30].

$$\bar{R}_T \in \mathcal{O}\left(\sqrt{kT\log(T)}\right) \tag{4.5}$$

The form of the dependence on the number of arms k and horizon T differs between the problem dependent and worst case regret. The problem dependent regret grows linearly with the number of arms, k, and logarithmically with T. The difference stems from the fact the problem dependent regret defines how the regret grows for a given set of reward distributions as T increases, whereas in the worst case regret, the gap between expected rewards is varied as a function of T. Auer et al. [16] show that the worst case regret for the k-armed bandit problem is lower bounded by $\bar{R}_T \in \Omega\left(\sqrt{kT}\right)$

Figure 4.7: The actual performance of the UCB algorithm can be substantially better than suggested by the upper bound, particularly for small T. The solid curve shows the mean expected regret associated with the sequence of arms chosen by UCB-1 with $k = 2$ arms and the rewards sampled from $bernoulli([.3, .7])$ over 1000 simulations. The dashed curve shows the corresponding upper bound given by the minimum of $T\Delta_{max}$ and equation4.4.



Subtle modifications to the UCB algorithm can eliminate the logarithmic term equation 4.5. This yields regret $\mathcal{O}\left(\sqrt{TK}\right)$ and closes the gap with the worst case lower bound [11, 101], whilst retaining a good problem dependent bound of the form achieved by UCB [101].

Finally, there is the heuristic principle of playing each arm with probability proportional to the likelyhood that it is optimal. This approach is generally called Thompson sampling as it was the method proposed in the original bandit paper by Thompson [159]. Thompson sampling has strong empirical performance, [40]. However, it is complex to analyse, Kaufmann et al. [93] demonstrate that it obtains optimal problem dependent bounds, Agrawal and Goyal [5] show that it obtains worst case regret of $\mathcal{O}\left(\sqrt{kT\log(T)}\right)$, equivalent to UCB.

4.2.2 Pure-exploration problems

Another problem that has attracted a lot of recent attention [31, 12, 60, 92] within the stochastic multi-armed bandit framework is *pure exploration* or *best arm identification*. In this setting, the horizon T represents a fixed budget for exploration after which the algorithm outputs a single best arm i. The performance of the algorithm is measured by the simple regret; the expected difference between the mean reward of the (truly) optimal arm and the mean reward of the arm selected by the algorithm.

Definition 15 (Simple Regret)**.**

$$R_T = \mu_{i^*} - \mathbb{E}\left[\mu_{\hat{i}^*}\right]. \tag{4.6}$$

The best arm identification problem arises naturally in applications where there is a testing or evaluation phase, during which regret is not incurred, followed by a commercialisation or exploitation phase. For example, many strategies might be assessed via simulation prior to one being selected and deployed. The worst case simple regret for a k-armed bandit is lower bounded by equation 4.7 ([31]).

$$R_T \in \mathcal{O}\left(\sqrt{K/T}\right) \tag{4.7}$$

Pure-exploration does not mean simply playing the arm with the widest uncertainty bounds. The goal is to be sure the arm we believe is optimal is in fact optimal at the end of the exploration period.

4.2.3 Adversarial Bandits

Adversarial bandits, described by Auer et al. [16], are an alternate, widely studied, setting that relaxes the assumption that rewards are generated stochastically. Instead, simultaneously with the learner selecting an action a_t, a potentially malicious adversary selects the reward vector $\boldsymbol{Y}_t$. As in the stochastic setting, the learner then receives reward only for the selected action.

Definition 16 (Adversarial k-armed bandit problem)**.** Let $\mathcal{A} = \{1, ...k\}$ be the set of available actions. In each round $t \in 1, ..., T$,

1. the world (or adversary) generates, but does not reveal, a vector or rewards $\boldsymbol{Y_t} = [Y_{t,1}, ..., Y_{t,k}]$.
2. the learner selects an action $a_t \in \{1, ..., k\}$, based on the actions and rewards from previous time-steps and a (potentially stochastic) *policy* π
3. the learner observes and receives (only) the reward for the selected action Y_{t,a_t}

Adversaries that generate rewards independently of the sequence of actions selected by the learner in previous time steps are referred to as *oblivious*, as opposed to *non-oblivious* adversaries, which can generate rewards as a function of the history of the game. In the case of oblivious adversaries, we can also define the adversarial bandit problem by assuming the adversary generates the entire sequence of reward vectors before the game commences.

For oblivious adversarial bandits we can define regret analogously to stochastic bandits as the difference between the reward obtained by playing the single arm with the highest reward in every round and the expected reward obtained by the algorithm [4]. We do not have to take the expectation over the first term of equation 4.8 because sequence of rewards is fixed, however the reward obtained by the algorithm is still a random variable as we are considering randomised algorithms.

$$\bar{R}_T(\pi) = \max_{i \in \{1,...,k\}} \sum_{t=1}^{T} Y_{t,i} - \mathbb{E}\left[\sum_{t=1}^{T} Y_{t,a_t}\right] \tag{4.8}$$

[4] This is also referred to as the weak regret, since in the adversarial case, it can make more sense to compare against the best sequence of arms rather than the best single arm.

The policy (or algorithm) used by the learner is available to the adversary before the game begins, and there are no limitations placed on the amount of computation the adversary can perform in selecting the reward sequences. This implies the adversary can ensure that any learner with a deterministic policy suffers regret $\mathcal{O}(T)$ by forecasting their entire sequence of actions. For example, if the learner will play $a_1 = 1$ in the first round, then the adversary sets the reward $\boldsymbol{Y_1} = [0, 1, 1, ...1]$, forecasts what action the learner will play in round 2, given they received a reward of 0 in round 1, and again generates the reward vector such that the action the learner will select obtains no reward, and all other actions obtain the maximum reward. This implies adversarial bandit policies must be sufficiently random to avoid such exploitation [5]

The seminal algorithm for adversarial bandits is Exp-3 [15], which, like UCB, obtains worst case pseudo-regret of $\mathcal{O}\left(\sqrt{TK\log(T)}\right)$ [16]. Optimal algorithms, with $\bar{R}_T = \mathcal{O}\left(\sqrt{TK}\right)$, have also been demonstrated for the oblivious adversarial setting [11]. The focus, for adversarial bandits, is on analysing the worst case regret because the problem dependent regret is not well defined without additional assumptions. However there has been recent work on developing algorithms that are optimised for both the adversarial and stochastic settings, in that they are sufficiently cautious to avoid linear regret in the adversarial setting but can nonetheless obtain good problem dependent regret in more favourable environments [34, 18].

Adversarial bandits appear to be more applicable to real world problems because they do not assume that the rewards associated with each arm are constant over time or independent of the previous actions of the learner. However, pseudo-regret, as defined in equation 4.8, is not a good measure of an algorithms performance in such cases because it is defined with respect to playing the single arm with the best average return over the game. In settings were the rewards change over time, the pseudo-regret can be negative, see figure 4.8, so upper bounds on the pseudo-regret do not reflect how sub-optimal the algorithm may be. Adversarial bandit algorithms may perform better in non-stationary settings than standard stochastic policies to the extend that they explore more (to avoid the adversary simulating their behaviour), however it is preferable to develop algorithms specifically for non-stationary settings (subject to assumptions about how rapidly or frequently rewards can change), see for example [63, 64, 25].

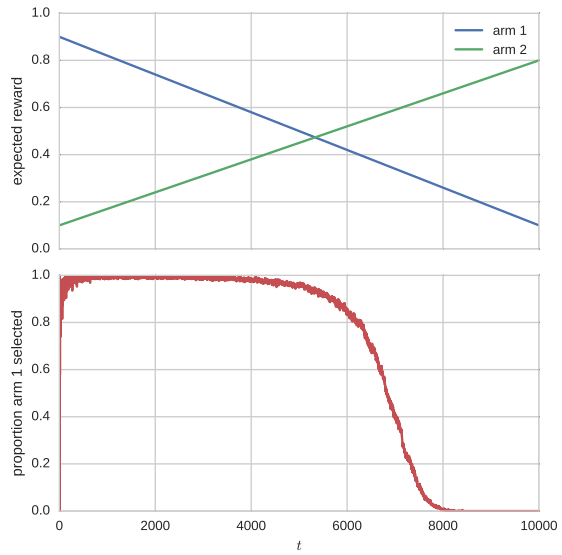
4.2.4 Contextual bandits

In the standard multi-armed bandit setting, each decision is identical and the goal is to learn a single best action. However, in most real life (sequential) decision making processes, which action is optimal depends on some context. The best treatment to offer an individual patient could depend on their age, gender, disease sub-type or genetics and will not always align with the treatment that is best on average (or for the majority of people). Similarly, decisions on which ad or content to display on a webpage or which product to recommend can be *personalised* based on the previous behaviour of the user. A movie recommender system that learned a single "best" movie for everyone would not be very useful. Contextual bandits are a generalisation of multi-armed bandits that make use of this additional contextual information. The term contextual bandit was coined by Langford and Zhang [99], however close variants of the underlying problem have also been posed under the names; "associative reinforcement learning" [90], "bandits with concomitant variables"[168] and "bandit problems with side information" [165].

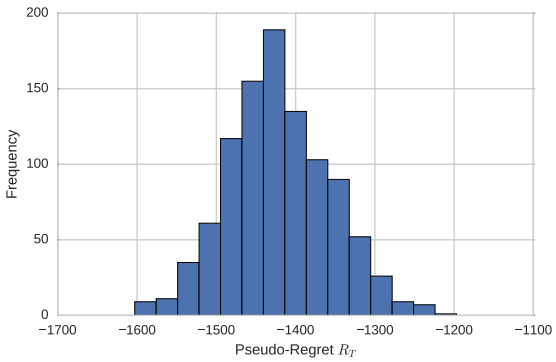
[5] The UCB algorithm, defined by algorithm 1, is deterministic if the order in which arms are played during the first k rounds is fixed and the method for selecting which arm to play when multiple-arms have the same upper-confidence bound is not-random (for example, select the arm one with the lowest index i).

Figure 4.8: The pseudo-regret can be negative if rewards are non-stationary. This example shows the results of 1000 simulations of running the UCB-1 algorithm on a 2-armed Bernoulli bandit problem where the expected rewards change linearly over time, up to a horizon $T = 10,000$. Figure (a) shows the expected rewards of each arm, and the proportion of time that arm-1 is played, as a function of time. The single best-arm is arm-1 as it has the highest expected reward (averaged over t). An oracle that selects arm-1 in every round obtains an expected reward of $5,000$. However, despite not being designed to do so, the UCB-algorithm can adapt to the changing reward distribution to obtain consistently higher rewards. The distribution of regret over the 1000 simulations is shown in figure (b).

(a) Behaviour over time



(b) Cumulative regret distribution



Definition 17 (Stochastic Contextual Bandit [6])**.** Let $\mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$ be the joint distribution over the rewards for each action and some context $\boldsymbol{X} \in \mathcal{X}$. In each round $t \in \{1, ...T\}$,

1. the world stochastically generates the vector of rewards for each action and the context, $(\boldsymbol{X}_t, [Y_t^1, ..., Y_t^k]) \sim \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$ and reveals $\boldsymbol{X}_t$ to the learner
2. the learner selects an action $A_t \in \{1, ..., k\}$, based on the context as well as actions and rewards from previous time-steps,
3. the learner observes and receives (only) the reward for the selected action $Y_t = Y_t^{A_t}$

Standard multi-armed bandits learn to select the action a that, with high probability, maximises $\mathbb{E}[Y|a]$. Contextual bandits learn to select actions that maximise $\mathbb{E}[Y|\boldsymbol{x}, a]$. The reward for contextual bandits should be compared to an oracle that acts optimally based on the context. To achieve this, even when the context is continuous, the regret is defined with respect to a class of hypothesis that map from context to action, $h \in \mathcal{H} : \mathcal{X} \to \{1, ..., k\}$. The pseudo-regret is the difference between the expected regret obtained by an oracle that selects actions based on the single best hypothesis or policy h at each timestep, and the expected reward obtained by the algorithm.

$$\bar{R}_T = \max_{h \in \mathcal{H}} \mathbb{E}\left[\sum_{t=1}^{T} Y_t^{h(\boldsymbol{X}_t)}\right] - \mathbb{E}\left[\sum_{t=1}^{T} Y_t^{A_t}\right] \tag{4.9}$$

[6] Contextual bandits can also be defined in the adversarial setting analogously to definition 16

If the context is discrete, $\mathcal{X} = \{1, ..., N\}$, the contextual bandit problem can be reduced to the standard multi-armed bandit problem by creating a separate standard bandit instance for each value of the context. This approach results in a worst case regret of $\mathcal{O}\left(\sqrt{NkT}\right)$, with respect to the hypothesis class $\mathcal{H} = \mathcal{X} \times \mathcal{A}$, consisting of all possible mappings from context to action[7]. This is optimal with respect to this class of hypothesis. However, as this reduction treats the problem of learning the correct action for each context completely independently, it cannot leverage any structure in the in the relationships between different contexts and actions. As in the supervised learning setting, the existence of some form of low-dimensional structure is key to learning in realistic problems, where the context is continuous or high-dimensional [8]. We expect some form of smoothness; values of context that are similar should lead to comparable rewards for a given action. We need algorithms that can leverage such assumptions.

An alternate reduction to the standard bandit problem, which allows us to constrain the hypothesis space to explore, is to treat each hypothesis h as a bandit arm [99]. At each time-step, we select $h \in \mathcal{H}$ based on the rewards previously observed for each hypothesis, take action $h(\boldsymbol{x})$ and observe the associated reward. Although this approach removes the explicit dependence on the size of the context, the regret grows linearly with the size of the hypothesis class considered, limiting our ability to learn any complex mappings from context to actions. The key problem with this approach is each sample is used to update our knowledge about only one hypothesis, as opposed to the supervised learning setting, where each data point is (implicitly) used to compute the loss for every hypothesis simultaneously.

Suppose that, at each timestep t, after selecting an action, the learner received the reward for chosen action but observed the full vector of rewards $[Y_t^1, ..., Y_t^k]$. This is known as the full information setting. In this case, the learner can simulate running each hypothesis over the history to compute the reward it would have obtained and use the hypothesis with the best empirical reward to select the next action. This is the *follow the leader* algorithm, which obtains optimal regret $\mathcal{O}\left(\sqrt{Tlog(|\mathcal{H}|)}\right)$ for the full-information problem [39]. Unfortunately, in the contextual bandit problem, the (counterfactual) rewards associated with alternate action choices are not observed. As in causal effect estimation, we can view this as a missing data problem. However, the data is missing not at random because which component of the reward is observed depends on the action selected which in turn is a function of the previous history of actions and rewards.

The Epoch-greedy algorithm, [99], addresses these issues by transforming the contextual bandit problem into a data missing at random problem by explicitly separating exploration from exploitation. Epoch-greedy is an explore-exploit algorithm. It selects actions uniformly at random during an exploration phase and leverages this data to estimate the value of each hypothesis, using inverse propensity weighted estimators to "fill in" the missing data. The hypothesis with the highest empirical reward is then used to select actions for the remaining time steps. The epsilon-greedy algorithm obtains worst case regret $\mathcal{O}\left(T^{\frac{2}{3}}(k \log |\mathcal{H}|)^{\frac{1}{3}}\right)$, which has sub-optimal dependence on the horizon T.

The Exp-4 algorithm, developed in the context of learning from expert advice (each $h \in \mathcal{H}$ can be viewed as an expert who recommends which action to take), achieves optimal worst case regret of $\mathcal{O}\left(\sqrt{kTlog(|\mathcal{H}|)}\right)$ in both the stochastic and adversarial settings [17]. However, it involves maintaining a list of weights for each hypothesis h resulting in time and memory requirements

[7] This follows from the fact that we have N standard bandit instances, each suffering regret $\mathcal{O}\left(\sqrt{kT_c}\right)$, where T_c is the number of times context c occurred such that $\sum_{c=1}^{N} T_c = T$. The regret is maximised if $T_c = T/N$ resulting in total regret $\mathcal{O}\left(N\sqrt{kT/N}\right)$.

[8] Even if the context is genuinely discrete, N grows exponentially with the number of variables. For example, with n binary variables, $N = 2^n$

that grow linearly with the size of the hypothesis space and, unlike the epoch-greedy algorithm, it cannot be generalised to infinite dimensional hypothesis spaces in a straightforward way. The ILOVECONBANDITS algorithm combines the best of both worlds to obtain a computationally efficient algorithm with (almost) optimal regret [3]

Both Epoch-greedy and ILOVECONBANDITS involve solving problems of the form,

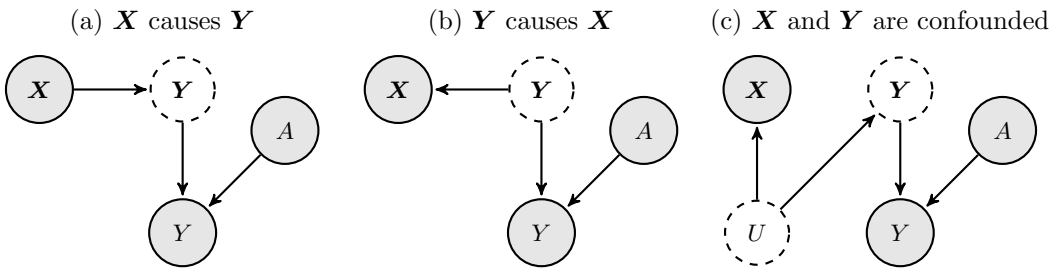
$$\arg\max_{h \in \mathcal{H}} \sum_{t=1}^{\tau} Y_t \mathbb{1}\{h(\boldsymbol{X}_t) = A_t\} \tag{4.10}$$

This expression equates to identifying the best empirical policy based on previous data. The algorithms assume the existence of an oracle that can solve this problem and report complexity in terms of the number of calls required to the oracle. The computational tractability of these algorithms on large (or infinite) hypothesis spaces stems from the fact that this problem (also known as the argmax-oracle), can be reduced to solving a cost sensitive classification problem [52].

Finally, if we have a parametric model for the relationship between context, action and reward that allows (efficient) computation of the posterior or confidence bounds on the reward for each arm given context, we can develop generalised versions of the UCB or Thompson sampling algorithms. For linear pay-off models, both approaches yield algorithms with strong regret guarantees, *Lin-UCB* [103] and *Generalised Thompson Sampling* [6].

It is worth noting that definition 17 does not make any assumptions about the *causal* relationship between the context $\boldsymbol{X}$ and the reward $\boldsymbol{Y}$, see figure 4.9. However, the context should be relevant, such that $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \neq \mathrm{P}(\boldsymbol{y})$, otherwise including it is the equivalent to adding irrelevant features to a supervised learning problem. Bareinboim et al. [21] demonstrate that, in some cases, policies that incorporate observations of the action an agent would have taken were their action not set by the bandit policy can achieve lower regret than those that ignore this information. This is an example of the case represented in figure 4.9c.

Figure 4.9: Several potential causal graphical models for the contextual bandit problem (if actions are selected at random). $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent the context and reward vectors respectively, and $Y = \boldsymbol{Y}^a$ is the reward received by the learner, which is a deterministic function of $\boldsymbol{Y}$ and a. Regardless of the causal structure, observing the context $\boldsymbol{X}$ provides information about the vector of rewards $\boldsymbol{Y}$. To represent a realistic (contextual) bandit problem, where actions are not selected at random, we would need to "unroll" the graphs, such that there was a copy for each time-step t and allow the action A_t to depend on the context $\boldsymbol{X}_t$ and the previous observations $(\boldsymbol{X}, A, Y)_1^{t-1}$.



4.2.5 Learning from logged bandit data

Another topic of interest within the bandit community, which is deeply connected to causal effect estimation from observational data, is learning from logged bandit feedback data or off-policy evaluation[98, 154, 104, 51, 28, 157]. In this setting, the learner has a data set $S = \left\{(\boldsymbol{X}_t, A_t, Y_t^{A_t})\right\}_{t=1}^{T}$, which is assumed to have been generated by a stochastic contextual bandit environment interacting with some unknown, potentially stochastic, policy $\pi(\boldsymbol{x}_t, h_t)$, where h_t is the sequence of observed data up to time t. The goal of the learner is to evaluate the value of an alternate policy, π', for selecting actions, often with the underlying motivation of identifying an optimal policy within some space of policies Π.

This problem differs from the contextual bandit problem in that the learner is not interacting with the environment. As a result, there is no exploration-exploitation trade-off to be made. However, the problem does not reduce to supervised learning, because the label, y is not the desired . In addition, if π is allowed to depend h then the samples are not i.i.d. The majority of the literature considers the case where the original policy π was stationary ($\pi(\boldsymbol{x}_t, h_t) = \pi(\boldsymbol{x})$). Langford et al. [98] do allow the original policy to be adaptive and prove a high probability bound on the accuracy of their estimator for π', albeit with the strong assumption that the original estimator π did not depend on $\boldsymbol{X}$.

If the original policy is assumed to be stationary, the problem of evaluating an alternate policy π' is almost identical to that of causal effect estimation under ignorability, discussed in section 3.3.2. The causal structure can be represented in figure 4.10 There is an (implicit) assumption that all variables that impact the choice of action by π are included in $\boldsymbol{X}$, ensuring that $\boldsymbol{X}$ satisfies the backdoor criterion with respect to identifying the causal effect of $do(A = a)$ on the observed reward Y, for any action $a \in \{1, ..., k\}$. The only difference is that the goal is to evaluate alternate policies π' that may be stochastic and depend on $\boldsymbol{x}$, as opposed to only policies of the form $\pi'(x) = a$, equivalent to $do(A = a)$. However, the identification of such stochastic, conditional policies can be reduced to the identification of $\mathrm{P}\left(y|do(A = a), \boldsymbol{x}\right)$, see Pearl [116], section 4.2. In this case, letting $\mathrm{P}_{\pi'}\left\{a|\boldsymbol{x}\right\}$ denote the distribution over actions under policy π' given context $\boldsymbol{x}$, the expected (per round) reward obtained by π' is given by,

$$\mathbb{E}\left[y|\pi'\right] = \mathbb{E}\left[y|do(a \sim \pi'(\boldsymbol{x}))\right] = \mathbb{E}_{(\boldsymbol{x},a)\sim \mathrm{P}(\boldsymbol{x})\,\mathrm{P}_{\pi'}\{a|\boldsymbol{x}\}}\left[y|\boldsymbol{x}, a\right] \tag{4.11}$$

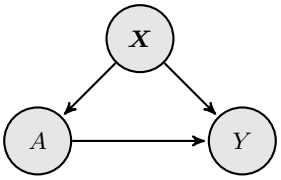
As in estimating average causal effects under ignorability, we have a covariate shift problem, with training data sampled from $\mathrm{P}\left(\boldsymbol{x}\right)\mathrm{P}_{\pi}\left\{a|\boldsymbol{x}\right\}\mathrm{P}\left(y|\boldsymbol{x}, a\right)$ but generalisation error measured with respect to $\mathrm{P}\left(\boldsymbol{x}\right)\mathrm{P}_{\pi'}\left\{a|\boldsymbol{x}\right\}\mathrm{P}\left(y|\boldsymbol{x}, a\right)$. A difference in practice, is that in the applications frequently considered under learning from logged feedback data, such as ad serving or recommender systems, there may be substantial information available about π, in the best case, $\mathrm{P}_{\pi}\left\{a|\boldsymbol{x}\right\}$ is known. This makes estimators utilising inverse propensity weighting, including doubly robust estimators as in Dud\'\ik et al. [51], more attractive.

Swaminathan and Joachims [157] point out that the problem of identifying the optimal policy (subject to some risk minimisation goal) is not as simple as estimating the expected reward associated with each policy in some space and selecting the empirical best because the variance of the estimators for some policies may be much higher than for others.

4.2.6 Adding structure to actions

The classic multi-armed bandit is a powerful tool for sequential decision making. However, the regret grows linearly with the number of (sub-optimal) actions and many real world problems

Figure 4.10: Causal graphical model for learning from logged feedback data under the assumption the original policy π for selecting actions was stationary and dependent only on some observed context $\boldsymbol{X}$, $a \sim \pi(\boldsymbol{x})$



have large or even infinite action spaces. This has led to the development of a wide range of models that assume some structure across the reward distributions for different arms, for example generalised linear bandits [56], dependent bandits [113], X-armed bandits [33] and Gaussian process bandits [152], or that consider more complex feedback, for example the recent work on graph feedback[110, 102, 8, Buccapatnam et al., 94, 9] and partial monitoring [121, 23].

In the next chapter, I propose a very natural connection between causal graphs and bandit problems and show it induces a novel form of additional feedback and structure between arms that cannot be can exploited by any of these previous approaches.

Chapter 5

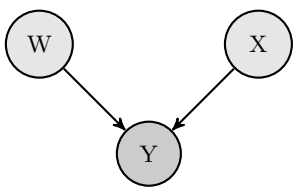
Causal Bandits: Unifying the approaches

5.1 The framework

A natural way to connect the causal framework with the bandit setting is to model the action space as interventions on variables in a causal directed acyclic graph. Each possible assignment of variables to values is a potential action (or bandit arm), see figure 5.1 for a simple example. In some settings, it makes sense to restrict the action space available to the agent to a subset of all the possible actions, for example the set of single variable interventions. The reward could be a general function of the action selected and the final state of the graph. However for simplicity, we will consider the reward to be the value of a single specified node. We refer to these problems as *causal bandit problems*. In this thesis I focus on the case where the causal graph is known. Extending this work to simultaneously learning the casual graph is discussed in §5.2.4.

The type of problem we are concerned with is best illustrated with an example. Consider a farmer wishing to optimise the yield of her crop. She knows that crop yield is only affected by temperature, a particular soil nutrient, and moisture level but the precise effect of their combination is unknown. In each season the farmer has enough time and money to intervene and control at most one of these variables: deploying shade or heat lamps will set the temperature to be low or high; the nutrient can be added or removed through a choice of fertiliser; and irrigation or rain-proof covers will keep the soil wet or dry. When not intervened upon, the temperature, soil, and moisture vary naturally from season to season due to weather conditions

Figure 5.1: A simple causal graphical model and corresponding complete action space. W and Z represent binary variables that can be intervened on and Y represents the reward.



$$\mathcal{A} = \begin{array}{|c|}\hline do(W=0,Z=0) \\ do(W=0,Z=1) \\ do(W=1,Z=0) \\ do(W=1,Z=1) \\ \hline do(W=0) \\ do(W=1) \\ do(Z=0) \\ do(Z=1) \\ do() \\ \hline \end{array}$$

and these are all observed along with the final crop yield at the end of each season. How might the farmer best experiment to identify the single, highest yielding intervention in a limited number of seasons?

We will assume each variable only takes on a finite number of distinct values. (The path to relaxing this assumption would be through levering the work on continuous armed bandits). The *parents* of a variable X_i, denoted $\mathcal{P}\mathrm{a}_{X_i}$, is the set of all variables X_j such that there is an edge from X_j to X_i in $\mathcal{G}$. An *intervention or action (of size n)*, denoted $do(\boldsymbol{X} = \boldsymbol{x})$, assigns the values $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ to the corresponding variables $\boldsymbol{X} = \{X_1, \ldots, X_n\} \subset \mathcal{X}$ with the empty intervention (where no variable is set) denoted $do()$. We denote the expected reward for the action $a = do(\boldsymbol{X} = \boldsymbol{x})$ by $\mu_a := \mathbb{E}[Y|do(\boldsymbol{X} = \boldsymbol{x})]$ and the optimal expected reward by $\mu^* := \max_{a \in \mathcal{A}} \mu_a$.

Definition 18 (Causal bandit problem)**.** A learner for a casual bandit problem is given the casual model's graph G over variables $\mathcal{X}$ and a set of allowed actions $\mathcal{A}$. Each action $a \in \mathcal{A}$ assigns a value to a subset of the variables in $\mathcal{X}$. One variable $Y \in \mathcal{X}$ is designated as the *reward variable* and takes on values in $\{0, 1\}$.

The causal bandit game proceeds over T rounds. In each round t, the learner:

1. *observes* the value of a subset of the variables $\boldsymbol{X}_t^c$,
2. *intervenes* by choosing $a_t = do(\boldsymbol{X}_t = \boldsymbol{x}_t) \in \mathcal{A}$ based on previous observations, and finally
3. *observes* sampled values for another subset of variables $\boldsymbol{X}_t^o$ drawn from $\mathrm{P}\left(\boldsymbol{X}_t^o | do(\boldsymbol{X}_t = \boldsymbol{x}_t)\right)$ including the *reward* $Y_t \in \{0, 1\}$.

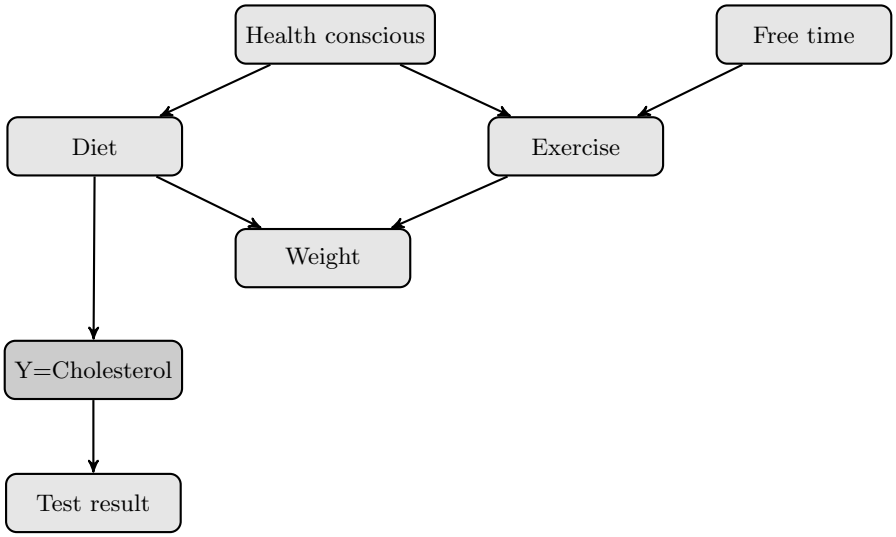
We refer to the set of variables that can be observed prior to selecting an action $\boldsymbol{X}^c$ as contextual variables and the set of variables observed after the action is chosen, $\boldsymbol{X}^o$, as post-action feedback variables. Note that $\boldsymbol{X}^c$ and $\boldsymbol{X}^o$ need not be disjoint. A variable may be observed both prior to and after the agent selects an action, and the action may change its value. The objective of the learner is to minimise either the simple (equation **??**) or cumulative regret (equation **??**).

The causal bandit problem takes on characteristics of different bandit settings depending on the action-space $\mathcal{A}$, which variables are observable prior to selecting an action and on which variables we receive post-action feedback. If feedback is received only on the reward node $\boldsymbol{X}^o = \{Y\}$, as in the standard bandit setting, then the do-calculus can be applied to eliminate some actions immediately, before any experiments are performed and then a standard bandit algorithm can be run on the remaining actions, see figure 5.2 as an example. If we receive post-action feedback on additional nodes the problem can be more interesting. In addition to being able to eliminate some actions prior to sampling any data as in the previous case, taking one action may give us some information on actions that were not selected. Consider again the model in figure 5.1. The causal structure implies:

$$\begin{aligned} P(Y|do(W=0)) &= P(Y|do(), W=0) \\ &= P(Y|do(X=0), W=0)P(X=0) + P(Y|do(X=1), W=0)P(X=1) \end{aligned}$$

Thus we gain information about the reward for the action $do(W = 0)$ from selecting the action $do()$ or $do(X = x)$ and then observing $W = 0$. We only get this form of side information for actions that don't specify the value of every variable, for example those in the bottom half of the table in figure 5.1. If additional variables are only observed before an intervention is selected the causal bandit problem reduces to stochastic contextual bandits, which are already reasonably well understood [3].

Figure 5.2: Example causal graph (based on Koller and Friedman [95]) where the outcome of interest (reward) is cholesterol level . The do-calculus can be applied to eliminate some actions immediately without the need to do any experiments. For example, no actions involving 'Test Result' need to be considered and interventions on 'Diet' do not need to be considered in conjunction with any other variables.



We note that classical K-armed stochastic bandit problem can be recovered in our framework by considering a simple causal model with one edge connecting a single variable X that can take on K values to a reward variable $Y \in \{0, 1\}$ where $\mathrm{P}(Y = 1|X) = r(X)$ for some arbitrary but unknown, real-valued function r. The set of allowed actions in this case is $\mathcal{A} = \{do(X = k) \colon k \in \{1, \ldots, K\}\}$. Conversely, any causal bandit problem can be reduced to a classical stochastic $|\mathcal{A}|$-armed bandit problem by treating each possible intervention as an independent arm and ignoring all sampled values for the observed variables except for the reward. However, the number of actions or arms grows exponentially with the number of variables in the graph making it important to develop algorithms that leverage the graph structure and additional observations.

5.2 Causal bandits with post action feedback

We now focus on causal bandit problems with post-action feedback, in which the value of all the variables are observed after an intervention is selected, and where the goal of the learner is to minimise the simple regret. I presented this work at NIPS 2016 [100].

Related Work As alluded to above, causal bandit problems can be treated as classical multi-armed bandit problems by simply ignoring the causal model and extra observations and applying an existing best-arm identification algorithm with well understood simple regret guarantees [85]. However, as we show in §5.2.1, ignoring the extra information available in the non-intervened variables yields sub-optimal performance.

Our framework bears a superficial similarity to contextual bandit problems, §4.2.4, since the extra observations on non-intervened variables might be viewed as context for selecting an intervention. However, a crucial difference is that in our model the extra observations are only revealed *after* selecting an intervention and hence cannot be used as context.

There have been several proposals for bandit problems where extra feedback is received after an action is taken. Most recently, Alon et al. [9], Kocák et al. [94] have considered very general models related to partial monitoring games [23] where rewards on un-played actions are revealed according to a feedback graph. As we discuss in §5.2.4, the parallel bandit problem can be captured in this framework, however the regret bounds are not optimal in our setting. They also focus on cumulative regret, which cannot be used to guarantee low simple regret [32]. The partial monitoring approach taken by Wu et al. [171] could be applied (up to modifications for the simple regret) to the parallel bandit, but the resulting strategy would need to know the likelihood of each factor in advance, while our strategy learns this online. Yu and Mannor [172] utilise extra observations to detect changes in the reward distribution, whereas we assume fixed reward distributions and use extra observations to improve arm selection. Avner et al. [20] analyse bandit problems where the choice of arm to pull and arm to receive feedback on are decoupled. The main difference from our present work is our focus on simple regret and the more complex information linking rewards for different arms via causal graphs. To the best of our knowledge, our paper is the first to analyse simple regret in bandit problems with extra post-action feedback.

Two pieces of recent work also consider applying ideas from causal inference to bandit problems. Bareinboim et al. [21] demonstrate that in the presence of confounding variables the value that a variable would have taken had it not been intervened on can provide important contextual information. Their work differs in many ways. For example, the focus is on the cumulative regret and the context is observed before the action is taken and cannot be controlled by the learning agent.

Ortega and Braun [112] present an analysis and extension of Thompson sampling assuming actions are causal interventions. Their focus is on causal induction (*i.e.*, learning an unknown causal model) instead of exploiting a known causal model. Combining their handling of causal induction with our analysis is left as future work.

The truncated importance weighted estimators used in §5.2.2 have been studied before in a causal framework by Bottou et al. [28], where the focus is on learning from observational data, but not controlling the sampling process. They also briefly discuss some of the issues encountered in sequential design, but do not give an algorithm or theoretical results for this case.

5.2.1 The parallel bandit problem

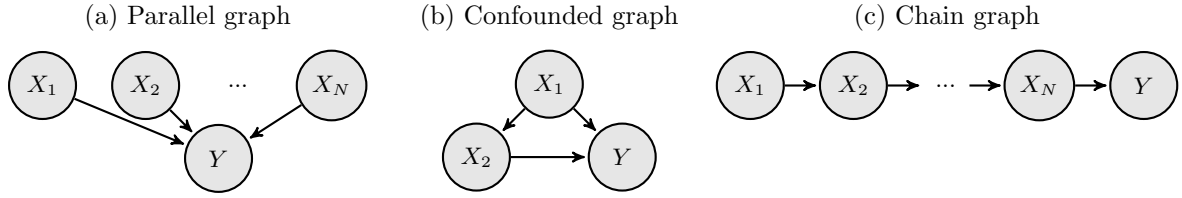
In this section we propose and analyse an algorithm for achieving the optimal regret in a natural special case of the causal bandit problem which we call the *parallel bandit*. It is simple enough to admit a thorough analysis but rich enough to model the type of problem discussed in §5.1, including the farming example. It also suffices to witness the regret gap between algorithms that make use of causal models and those which do not.

The causal model for this class of problems has N binary variables $\{X_1, \ldots, X_N\}$ where each $X_i \in \{0, 1\}$ are independent causes of a reward variable $Y \in \{0, 1\}$, as shown in Figure 5.3a. All variables are observable and the set of allowable actions are all size 0 and size 1 interventions: $\mathcal{A} = \{do()\} \cup \{do(X_i = j) \colon 1 \leq i \leq N \text{ and } j \in \{0, 1\}\}$

In the farming example from the introduction, X_1 might represent temperature (*e.g.*, $X_1 = 0$ for low and $X_1 = 1$ for high). The interventions $do(X_1 = 0)$ and $do(X_1 = 1)$ indicate the use of shades or heat lamps to keep the temperature low or high, respectively.

In each round the learner either purely observes by selecting $do()$ or sets the value of a single variable. The remaining variables are simultaneously set by independently biased coin flips. The value of all variables are then used to determine the distribution of rewards for that

Figure 5.3: Causal Models



round. Formally, when not intervened upon we assume that each $X_i \sim \text{Bernoulli}(q_i)$ where $\boldsymbol{q} = (q_1, \ldots, q_N) \in [0,1]^N$ so that $q_i = \mathrm{P}(X_i = 1)$.

The value of the reward variable is distributed as $\mathrm{P}(Y = 1|\boldsymbol{X}) = r(\boldsymbol{X})$ where $r : \{0,1\}^N \to [0,1]$ is an arbitrary, fixed, and unknown function. In the farming example, this choice of Y models the success or failure of a seasons crop, which depends stochastically on the various environment variables.

The Parallel Bandit Algorithm The algorithm operates as follows. For the first $T/2$ rounds it chooses $do()$ to collect observational data. As the only link from each $X_1, \ldots, X_N$ to Y is a direct, causal one, $\mathrm{P}(Y|do(X_i = j)) = \mathrm{P}(Y|X_i = j)$. Thus we can create good estimators for the returns of the actions $do(X_i = j)$ for which $\mathrm{P}(X_i = j)$ is large. The actions for which $\mathrm{P}(X_i = j)$ is small may not be observed (often) so estimates of their returns could be poor. To address this, the remaining $T/2$ rounds are evenly split to estimate the rewards for these infrequently observed actions. The difficulty of the problem depends on $\boldsymbol{q}$ and, in particular, how many of the variables are unbalanced (*i.e.*, small q_i or $(1 - q_i)$). For $\tau \in [2...N]$ let $I_\tau = \left\{i : \min\{q_i, 1 - q_i\} < \frac{1}{\tau}\right\}$. Define

$$m(\boldsymbol{q}) = \min\{\tau : |I_\tau| \leq \tau\} \ .$$

Algorithm 2 Parallel Bandit Algorithm

Input: Total rounds T and N.
for $t \in 1, \ldots, T/2$ **do**
  Perform empty intervention $do()$
  Observe $\boldsymbol{X}_t$ and Y_t
for $a = do(X_i = x) \in \mathcal{A}$ **do**
  Count times $X_i = x$ seen: $T_a = \sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = x\}$
  Estimate reward: $\hat{\mu}_a = \frac{1}{T_a} \sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = x\} Y_t$
  Estimate probabilities: $\hat{p}_a = \frac{2T_a}{T}$, $\hat{q}_i = \hat{p}_{do(X_i=1)}$
Compute $\hat{m} = m(\hat{\boldsymbol{q}})$ and $A = \left\{a \in \mathcal{A} \colon \hat{p}_a \leq \frac{1}{\hat{m}}\right\}$.
Let $T_A := \frac{T}{2|A|}$ be times to sample each $a \in A$.
for $a = do(X_i = x) \in A$ **do**
  for $t \in 1, \ldots, T_A$ **do**
    Intervene with a and observe Y_t
  Re-estimate $\hat{\mu}_a = \frac{1}{T_A} \sum_{t=1}^{T_A} Y_t$
return estimated optimal $\hat{a}_T^* \in \arg\max_{a \in \mathcal{A}} \hat{\mu}_a$

I_τ is the set of variables considered unbalanced and we tune τ to trade off identifying the low probability actions against not having too many of them, so as to minimise the worst-case simple

regret. When $\boldsymbol{q} = (\frac{1}{2}, \ldots, \frac{1}{2})$ we have $m(\boldsymbol{q}) = 2$ and when $\boldsymbol{q} = (0, \ldots, 0)$ we have $m(\boldsymbol{q}) = N$. We do not assume that $\boldsymbol{q}$ is known, thus Algorithm 2 also utilises the samples captured during the observational phase to estimate $m(\boldsymbol{q})$. Although very simple, the following two theorems show that this algorithm is effectively optimal.

Theorem 19. *Algorithm 2 satisfies*

$$R_T \in \mathcal{O}\left(\sqrt{\frac{m(\boldsymbol{q})}{T}\log\left(\frac{NT}{m(\boldsymbol{q})}\right)}\right).$$

Theorem 20. *For all strategies and T, $\boldsymbol{q}$, there exist rewards such that $R_T \in \Omega\left(\sqrt{\frac{m(\boldsymbol{q})}{T}}\right)$.*

The proofs of Theorems 19 and 20 follow by carefully analysing the concentration of $\hat{p}_a$ and $\hat{m}$ about their true values and may be found in Sections 5.2.5 and 5.2.5 respectively.

By utilising knowledge of the causal structure, Algorithm 2 effectively only has to explore the $m(\boldsymbol{q})$ 'difficult' actions. Standard multi-armed bandit algorithms must explore all $2N$ actions and thus achieve regret $\Omega(\sqrt{N/T})$. Since m is typically much smaller than N, the new algorithm can significantly outperform classical bandit algorithms in this setting. In practice, you would combine the data from both phases to estimate rewards for the low probability actions. We do not do so here as it slightly complicates the proofs and does not improve the worst case regret.

5.2.2 General graphs

We now consider the more general problem where the graph structure is known, but arbitrary. For general graphs, $\mathrm{P}(Y|X_i = j) \neq \mathrm{P}(Y|do(X_i = j))$ (correlation is not causation). However, if all the variables are observable, any causal distribution $\mathrm{P}(X_1...X_N|do(X_i = j))$ can be expressed in terms of observational distributions via the truncated factorisation formula [116].

$$\mathrm{P}(X_1...X_N|do(X_i = j)) = \prod_{k\neq i} \mathrm{P}(X_k|\,\mathcal{P}\mathrm{a}_{X_k})\,\delta(X_i - j)\,,$$

where $\mathcal{P}\mathrm{a}_{X_k}$ denotes the parents of X_k and δ is the Dirac delta function.

We could naively generalise our approach for parallel bandits by observing for $T/2$ rounds, applying the truncated product factorisation to write an expression for each $\mathrm{P}(Y|a)$ in terms of observational quantities and explicitly playing the actions for which the observational estimates were poor. However, it is no longer optimal to ignore the information we can learn about the reward for intervening on one variable from rounds in which we act on a different variable. Consider the graph in Figure 5.3c and suppose each variable deterministically takes the value of its parent, $X_k = X_{k-1}$ for $k \in 2, \ldots, N$ and $\mathrm{P}(X_1) = 0$. We can learn the reward for all the interventions $do(X_i = 1)$ simultaneously by selecting $do(X_1 = 1)$, but not from $do()$. In addition, variance of the observational estimator for $a = do(X_i = j)$ can be high even if $\mathrm{P}(X_i = j)$ is large. Given the causal graph in Figure 5.3b, $\mathrm{P}(Y|do(X_2 = j)) = \sum_{X_1} \mathrm{P}(X_1)\,\mathrm{P}(Y|X_1, X_2 = j)$. Suppose $X_2 = X_1$ deterministically, no matter how large $\mathrm{P}(X_2 = 1)$ is we will never observe $(X_2 = 1, X_1 = 0)$ and so cannot get a good estimate for $\mathrm{P}(Y|do(X_2 = 1))$.

To solve the general problem we need an estimator for each action that incorporates information obtained from every other action and a way to optimally allocate samples to actions. To address this difficult problem, we assume the conditional interventional distributions $\mathrm{P}(\mathcal{P}\mathrm{a}_Y\,|a)$ (but not $\mathrm{P}(Y|a)$) are known. These could be estimated from experimental data on the same covariates but where the outcome of interest differed, such that Y was not included, or similarly from observational data subject to identifiability constraints. Of course this is a somewhat limiting

assumption, but seems like a natural place to start. The challenge of estimating the conditional distributions for all variables in an optimal way is left as an interesting future direction. Let η be a distribution on available interventions $a \in \mathcal{A}$ so $\eta_a \geq 0$ and $\sum_{a\in\mathcal{A}} \eta_a = 1$. Define $Q = \sum_{a\in\mathcal{A}} \eta_a \, \mathrm{P}\left(\mathcal{P}\mathrm{a}_Y \,|a\right)$ to be the mixture distribution over the interventions with respect to η.

Algorithm 3 General Algorithm

Input: T, $\eta \in [0,1]^{\mathcal{A}}$, $B \in [0,\infty)^{\mathcal{A}}$
for $t \in \{1,\ldots,T\}$ **do**
  Sample action a_t from η
  Do action a_t and observe X_t and Y_t
for $a \in \mathcal{A}$ **do**

$$\hat{\mu}_a = \frac{1}{T}\sum_{t=1}^{T} Y_t R_a(X_t) \mathbb{1}\{R_a(X_t) \leq B_a\}$$

return $\hat{a}^*_T = \arg\max_a \hat{\mu}_a$

Our algorithm samples T actions from η and uses them to estimate the returns μ_a for all $a \in \mathcal{A}$ simultaneously via a truncated importance weighted estimator. Let $\mathcal{P}\mathrm{a}_Y(X)$ denote the realisation of the variables in X that are parents of Y and define $R_a(X) = \frac{\mathrm{P}\{\mathcal{P}\mathrm{a}_Y(X)|a\}}{\mathrm{Q}(\mathcal{P}\mathrm{a}_Y(X))}$

$$\hat{\mu}_a = \frac{1}{T}\sum_{t=1}^{T} Y_t R_a(X_t) \mathbb{1}\{R_a(X_t) \leq B_a\} \ ,$$

where $B_a \geq 0$ is a constant that tunes the level of truncation to be chosen subsequently. The truncation introduces a bias in the estimator, but simultaneously chops the potentially heavy tail that is so detrimental to its concentration guarantees.

The distribution over actions, η plays the role of allocating samples to actions and is optimised to minimise the worst-case simple regret. Abusing notation we define $m(\eta)$ by

$$m(\eta) = \max_{a\in\mathcal{A}} \mathbb{E}_a \left[\frac{\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}{\mathrm{Q}\left(\mathcal{P}\mathrm{a}_Y(X)\right)}\right] \ , \text{ where } \mathbb{E}_a \text{ is the expectation with respect to } \mathrm{P}\left\{.|a\right\}$$

We will show shortly that $m(\eta)$ is a measure of the difficulty of the problem that approximately coincides with the version for parallel bandits, justifying the name overloading.

Theorem 21. *If Algorithm 3 is run with* $B \in \mathbb{R}^{\mathcal{A}}$ *given by* $B_a = \sqrt{\frac{m(\eta)T}{\log(2T|\mathcal{A}|)}}$.

$$R_T \in \mathcal{O}\left(\sqrt{\frac{m(\eta)}{T}\log\left(2T|\mathcal{A}|\right)}\right) .$$

The proof is in Section 5.2.5.

Note the regret has the same form as that obtained for Algorithm 2, with $m(\eta)$ replacing $m(q)$. Algorithm 2 assumes only the graph structure and not knowledge of the conditional distributions on X. Thus it has broader applicability to the parallel graph than the generic algorithm given here. We believe that Algorithm 3 with the optimal choice of η is close to mini-max optimal, but leave lower bounds for future work.

Choosing the Sampling Distribution Algorithm 3 depends on a choice of sampling distribution Q that is determined by η. In light of Theorem 21 a natural choice of η is the minimiser of $m(\eta)$.

$$\eta^* = \arg\min_{\eta} m(\eta) = \arg\min_{\eta} \underbrace{\max_{a \in \mathcal{A}} \mathbb{E}_a \left[\frac{\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}{\sum_{b \in \mathcal{A}} \eta_b \mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|b\right\}} \right]}_{m(\eta)}.$$

Since the mixture of convex functions is convex and the maximum of a set of convex functions is convex, we see that $m(\eta)$ is convex (in η). Therefore the minimisation problem may be tackled using standard techniques from convex optimisation. The quantity $m(\eta^*)$ may be interpreted as the minimum achievable worst-case variance of the importance weighted estimator. In the experimental section we present some special cases, but for now we give two simple results. The first shows that $|\mathcal{A}|$ serves as an upper bound on $m(\eta^*)$.

Proposition 22. $m(\eta^*) \leq |\mathcal{A}|$. *Proof.* By definition, $m(\eta^*) \leq m(\eta)$ for all η. Let $\eta_a = 1/|\mathcal{A}| \, \forall a$.

$$m(\eta) = \max_a \mathbb{E}_a \left[\frac{\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}{\mathrm{Q}\left(\mathcal{P}\mathrm{a}_Y(X)\right)} \right] \leq \max_a \mathbb{E}_a \left[\frac{\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}{\eta_a \mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}} \right] = \max_a \mathbb{E}_a \left[\frac{1}{\eta_a} \right] = |\mathcal{A}|$$

The second observation is that, in the parallel bandit setting, $m(\eta^*) \leq 2m(\boldsymbol{q})$. This is easy to see by letting $\eta_a = 1/2$ for $a = do()$ and $\eta_a = \mathbb{1}\{\mathrm{P}\left(X_i = j\right) \leq 1/m(\boldsymbol{q})\}/2m(\boldsymbol{q})$ for the actions corresponding to $do(X_i = j)$, and applying an argument like that for Proposition 22. The proof is in section 5.2.5.

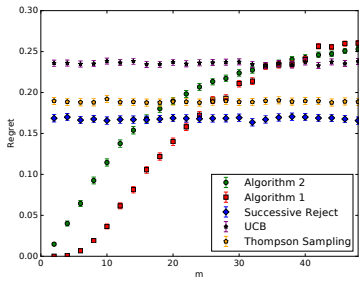
Remark 23. The choice of B_a given in Theorem 21 is not the only possibility. As we shall see in the experiments, it is often possible to choose B_a significantly larger when there is no heavy tail and this can drastically improve performance by eliminating the bias. This is especially true when the ratio R_a is never too large and Bernstein's inequality could be used directly without the truncation. For another discussion see the article by Bottou et al. [28] who also use importance weighted estimators to learn from observational data.

5.2.3 Experiments

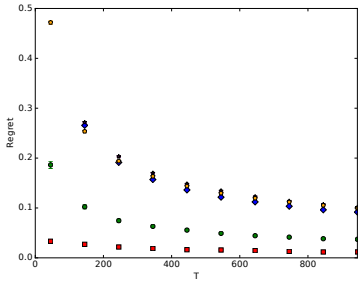
We compare Algorithms 2 and 3 with the Successive Reject algorithm of Audibert and Bubeck [12], Thompson Sampling and UCB under a variety of conditions. Thomson sampling and UCB are optimised to minimise cumulative regret. We apply them in the fixed horizon, best arm identification setting by running them up to horizon T and then selecting the arm with the highest empirical mean. The importance weighted estimator used by Algorithm 3 is not truncated, which is justified in this setting by Remark 23.

Throughout we use a model in which Y depends only on a single variable X_1 (this is unknown to the algorithms). $Y_t \sim \text{Bernoulli}(\frac{1}{2} + \varepsilon)$ if $X_1 = 1$ and $Y_t \sim \text{Bernoulli}(\frac{1}{2} - \varepsilon')$ otherwise, where $\varepsilon' = q_1\varepsilon/(1-q_1)$. This leads to an expected reward of $\frac{1}{2}+\varepsilon$ for $do(X_1 = 1)$, $\frac{1}{2}-\varepsilon'$ for $do(X_1 = 0)$ and $\frac{1}{2}$ for all other actions. We set $q_i = 0$ for $i \leq m$ and $\frac{1}{2}$ otherwise. Note that changing m and thus $\boldsymbol{q}$ has no effect on the reward distribution. For each experiment, we show the average regret over 10,000 simulations with error bars displaying three standard errors. The code is available from `<https://github.com/finnhacks42/causal_bandits>`

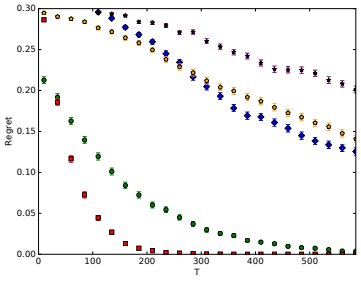
In Figure 5.4a we fix the number of variables N and the horizon T and compare the performance of the algorithms as m increases. The regret for the Successive Reject algorithm is constant as it depends only on the reward distribution and has no knowledge of the causal structure. For the causal algorithms it increases approximately with $\sqrt{m}$. As m approaches N, the gain the causal algorithms obtain from knowledge of the structure is outweighed by fact they do not leverage the observed rewards to focus sampling effort on actions with high pay-offs.



(a) Simple regret vs $m(\boldsymbol{q})$ for fixed horizon $T = 400$ and number of variables $N = 50$



(b) Simple regret vs horizon, T, with $N = 50$, $m = 2$ and $\varepsilon = \sqrt{\frac{N}{8T}}$



(c) Simple regret vs horizon, T, with $N = 50$, $m = 2$ and fixed $\varepsilon = .3$

Figure 5.4: Experimental results

Figure 5.4b demonstrates the performance of the algorithms in the worst case environment for standard bandits, where the gap between the optimal and sub-optimal arms, $\varepsilon = \sqrt{N/(8T)}$, is just too small to be learned. This gap is learn-able by the causal algorithms, for which the worst case ε depends on $m \ll N$. In Figure 5.4c we fix N and ε and observe that, for sufficiently large T, the regret decays exponentially. The decay constant is larger for the causal algorithms as they have observed a greater effective number of samples for a given T.

For the parallel bandit problem, the regression estimator used in the specific algorithm outperforms the truncated importance weighted estimator in the more general algorithm, despite the fact the specific algorithm must estimate $\boldsymbol{q}$ from the data. This is an interesting phenomenon that has been noted before in off-policy evaluation where the regression (and not the importance weighted) estimator is known to be mini-max optimal asymptotically [**?**].

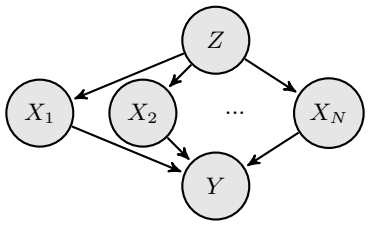


Figure 5.5: Confounded graph

We now compare the general algorithm with a range of standard bandit algorithms on the confounded graph in Figure 5.5. All the variables are binary and the action space consists of the set of single variable interventions plus the do nothing action,

$$\mathcal{A} = \{\{do(X_i = j)\} \cup \{do(Z = j)\} \cup \{do()\} : 1 \leq i \leq N,\ j \in \{0, 1\}\}$$

We choose this setting because it generalises the parallel bandit, while simultaneously being sufficiently simple that we can compute the exact reward and interventional distributions for large N (in general inference in graphical models is exponential in N). As before, we show the average regret over 10,000 simulations with error bars showing three standard errors.

In Figure 5.6a we fix N and T and $P(Z=1)=.4$. For some $2 \leq N_1 \leq N$ we define

$$P(X_i=1|Z=0) = \begin{cases} 0 & \text{if } i \in \{1,...N_1\} \\ .4 & \text{otherwise} \end{cases}$$

$$P(X_i=1|Z=1) = \begin{cases} 0 & \text{if } i \in \{1,...N_1\} \\ .65 & \text{otherwise} \end{cases}$$

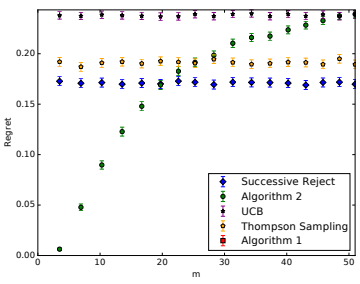
As in the parallel bandit case, we let Y depend only on X_1, $P(Y|do(X_1=1)) = \frac{1}{2}+\varepsilon$ and $P(Y|do(X_1=0)) = \frac{1}{2}-\varepsilon'$, where $\varepsilon' = \varepsilon P(X_1=1)/P(X_1=0)$. The value of N_1 determines m and ranges between 2 and N. The values for the CPD's have been chosen such that the reward distribution is independent of m and so that we can analytically calculate $\eta*$. This allows us to just show the dependence on m, removing the noise associated with different models selecting values for $\eta*$ with the same m (and also worst case performance), but different performance for a given reward distribution.

In Figure 5.6b we fix the model and number of variables, N, and vary the horizon T. $P(Z)$ and $P(X|Z)$ are the same as for the previous experiment. In Figure 5.6c we additionally show the performance of Algorithm 1, but exclude actions on Z from the set of allowable actions to demonstrate that Algorithm 1 can fail in the presence of a confounding variable, which occurs because it incorrectly assumes that $P(Y|do(X)) = P(Y|X)$. We let $P(Z)=.6$, $P(Y|\boldsymbol{X}) = X_7 \oplus X_N$ and $P(X|Z)$ be given by:

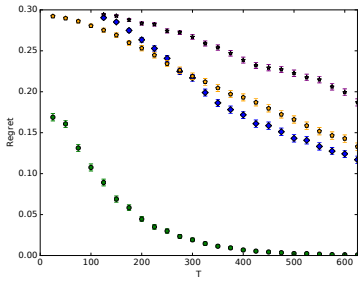
$$P(X_i=1|Z=0) = \begin{cases} .166 & \text{if } i \in \{1,...,6\} \\ .2 & \text{if } i=7 \\ .7 & \text{otherwise} \end{cases}$$

$$P(X_i=1|Z=1) = \begin{cases} .166 & \text{if } i \in \{1,...,6\} \\ .8 & \text{if } i=7 \\ .3 & \text{otherwise} \end{cases}$$

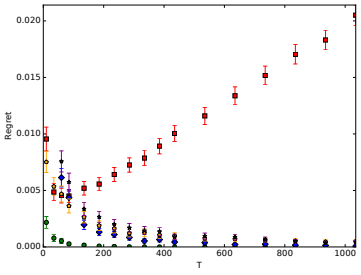
In this setting X_7 tends to agree with Z and X_N tends to disagree. It is sub-optimal to act on either X_7 or X_N, while all other actions are optimal. The first group of X variables with $i \leq 6$ will be identified by the parallel bandit as the most unbalanced ones and played explicitly. All remaining variables are likely to be identified as balanced and estimated from observational estimates. The CPD values have been chosen to demonstrate the worst case outcome, where the bias in the estimates leads Algorithm 1 to asymptotically select a sub-optimal action.



(a) Simple regret vs $m(\eta*)$ for fixed horizon $T = 400$ and number of variables $N = 50$



(b) Simple regret vs horizon, T, with $N = 50$ and $m(\eta*) = 3.1$



(c) Simple regret vs horizon, T, with $N = 21$, $m(\eta*) = 4.3$ with no actions setting Z

Figure 5.6: Experimental results on the confounded graph

5.2.4 Discussion & Future work

Algorithm 3 for general causal bandit problems estimates the reward for all allowable interventions $a \in \mathcal{A}$ over T rounds by sampling and applying interventions from a distribution η. Theorem 21 shows that this algorithm has (up to log factors) simple regret that is $\mathcal{O}(\sqrt{m(\eta)/T})$ where the parameter $m(\eta)$ measures the difficulty of learning the causal model and is always less than N. The value of $m(\eta)$ is a uniform bound on the variance of the reward estimators $\hat{\mu}_a$ and, intuitively, problems where all variables' values in the causal model "occur naturally" when interventions are sampled from η will have low values of $m(\eta)$.

The main practical drawback of Algorithm 3 is that both the estimator $\hat{\mu}_a$ and the optimal sampling distribution η^* (*i.e.*, the one that minimises $m(\eta)$) require knowledge of the conditional distributions $\mathrm{P}\{\mathcal{P}\mathrm{a}_Y | a\}$ for all $a \in \mathcal{A}$. In contrast, in the special case of parallel bandits, Algorithm 2 uses the $do()$ action to effectively estimate $m(\eta)$ and the rewards then re-samples the interventions with variances that are not bound by $\hat{m}(\eta)$. Despite these extra estimates, Theorem 20 shows that this approach is optimal (up to log factors).Finding an algorithm that only requires the causal graph and lower bounds for its simple regret in the general case is left as future work.

Making Better Use of the Reward Signal Existing algorithms for best arm identification are based on "successive rejection" (SR) of arms based on UCB-like bounds on their rewards [55]. In contrast, our algorithms completely ignore the reward signal when developing their arm sampling policies and only use the rewards when estimating $\hat{\mu}_a$. Incorporating the reward signal into our sampling techniques or designing more adaptive reward estimators that focus on high reward interventions is an obvious next step. This would likely improve the poor performance of our causal algorithm relative to the successive rejects algorithm for large m, as seen in Figure 5.4a.

For the parallel bandit the required modifications should be quite straightforward. The idea would be to adapt the algorithm to essentially use successive elimination in the second phase so arms are eliminated as soon as they are provably no longer optimal with high probability. In the general case a similar modification is also possible by dividing the budget T into phases and optimising the sampling distribution η, eliminating arms when their confidence intervals are no longer overlapping. Note that these modifications will not improve the mini-max regret, which at least for the parallel bandit is already optimal. For this reason we prefer to emphasise the main point that causal structure should be exploited when available. Another observation is that Algorithm 3 is actually using a fixed design, which in some cases may be preferred to a sequential design for logistical reasons. This is not possible for Algorithm 2, since the $\boldsymbol{q}$ vector is unknown.

Cumulative Regret Although we have focused on simple regret in our analysis, it would also be natural to consider the cumulative regret. In the case of the parallel bandit problem we can slightly modify the analysis from [171] on bandits with side information to get near-optimal cumulative regret guarantees. They consider a finite-armed bandit model with side information where in reach round the learner chooses an action and receives a Gaussian reward signal for all actions, but with a known variance that depends on the chosen action. In this way the learner can gain information about actions it does not take with varying levels of accuracy. The reduction follows by substituting the importance weighted estimators in place of the Gaussian reward. In the case that $\boldsymbol{q}$ is known this would lead to a known variance and the only (insignificant) difference is the Bernoulli noise model. In the parallel bandit case we believe this would lead to near-optimal cumulative regret, at least asymptotically.

The parallel bandit problem can also be viewed as an instance of a time varying graph feedback problem [9, 94], where at each time step the feedback graph G_t is selected stochastically, dependent on $\boldsymbol{q}$, and revealed after an action has been chosen. The feedback graph is distinct from the causal graph. A link $A \to B$ in G_t indicates that selecting the action A reveals the reward for action B. For this parallel bandit problem, G_t will always be a star graph with the action $do()$ connected to half the remaining actions. However, Alon et al. [9], Kocák et al. [94] give adversarial algorithms, which when applied to the parallel bandit problem obtain the standard bandit regret. A malicious adversary can select the same graph each time, such that the rewards for half the arms are never revealed by the informative action. This is equivalent to a nominally stochastic selection of feedback graph where $\boldsymbol{q} = \mathbf{0}$.

Lelarge and Ens [102] consider a stochastic version of the graph feedback problem, but with a fixed graph available to the algorithm before it must select an action. In addition, their algorithm is not optimal for all graph structures and fails, in particular, to provide improvements for star like graphs as in our case. [Buccapatnam et al.] improve the dependence of the algorithm on the graph structure but still assume the graph is fixed and available to the algorithm before the action is selected.

Causal Models with Non-Observable Variables If we assume knowledge of the conditional *interventional* distributions $\mathrm{P}\{\mathcal{P}\mathrm{a}_Y | a\}$ our analysis applies unchanged to the case of causal models with non-observable variables. Some of the interventional distributions may be non-identifiable meaning we can not obtain prior estimates for $\mathrm{P}\{\mathcal{P}\mathrm{a}_Y | a\}$ from even an infinite amount of observational data. Even if all variables are observable and the graph is known, if the conditional distributions are unknown, then Algorithm 3 cannot be used. Estimating these quantities while simultaneously minimising the simple regret is an interesting and challenging open problem.

Partially or Completely Unknown Causal Graph A much more difficult generalisation would be to consider causal bandit problems where the causal graph is completely unknown or known to be a member of class of models. The latter case arises naturally if we assume free access to a large observational data set, from which the Markov equivalence class can be found via causal discovery techniques. Work on the problem of selecting experiments to discover the correct causal graph from within a Markov equivalence class [**?** 53, 70, 79] could potentially be incorporated into a causal bandit algorithm. In particular, Hu and Vetta [79] show that only $\mathcal{O}(\log \log n)$ multi-variable interventions are required on average to recover a causal graph over n variables once purely observational data is used to recover the "essential graph". Simultaneously learning a completely unknown causal model while estimating the rewards of interventions without a large observational data set would be much more challenging.

5.2.5 Proofs

Proof of Theorem 19

Assume without loss of generality that $q_1 \leq q_2 \leq \ldots \leq q_N \leq 1/2$. The assumption is non-restrictive since all variables are independent and permutations of the variables can be pushed to the reward function.

The proof of Theorem 19 requires some lemmas.

Lemma 24. *Let $i \in \{1, \dots, N\}$ and $\delta > 0$. Then*

$$\mathrm{P}\left(|\hat{q}_i - q_i| \geq \sqrt{\frac{6q_i}{T} \log \frac{2}{\delta}}\right) \leq \delta\,.$$

Proof. By definition, $\hat{q}_i = \frac{2}{T} \sum_{t=1}^{T/2} X_{t,i}$, where $X_{t,i} \sim Bernoulli(q_i)$. Therefore from the Chernoff bound (see equation 6 in Hagerup and Rüb [69]),

$$\mathrm{P}\left(|\hat{q}_i - q_i| \geq \varepsilon\right) \leq 2e^{-\frac{T\varepsilon^2}{6q_i}}$$

Letting $\delta = 2e^{-\frac{T\varepsilon^2}{6q_i}}$ and solving for ε completes the proof.

□

Lemma 25. *Let $\delta \in (0, 1)$ and assume $T \geq 48m \log \frac{2N}{\delta}$. Then*

$$\mathrm{P}\left(2m(\boldsymbol{q})/3 \leq m(\hat{\boldsymbol{q}}) \leq 2m(\boldsymbol{q})\right) \geq 1 - \delta\,.$$

Proof. Let F be the event that there exists and $1 \leq i \leq N$ for which

$$|\hat{q}_i - q_i| \geq \sqrt{\frac{6q_i}{T} \log \frac{2N}{\delta}}\,.$$

Then by the union bound and Lemma 24 we have $\mathrm{P}(F) \leq \delta$. The result will be completed by showing that when F does not hold we have $2m(\boldsymbol{q})/3 \leq m(\hat{\boldsymbol{q}}) \leq 2m(\boldsymbol{q})$. From the definition of $m(\boldsymbol{q})$ and our assumption on $\boldsymbol{q}$ we have for $i > m(\boldsymbol{q})$ that $q_i \geq q_m \geq 1/m(\boldsymbol{q})$ and so by Lemma 24 we have

$$\begin{aligned}\frac{3}{4} &\geq \frac{1}{2} + \sqrt{\frac{3}{T} \log \frac{2N}{\delta}} \geq q_i + \sqrt{\frac{6q_i}{T} \log \frac{2N}{\delta}} \geq \hat{q}_i \\ &\geq q_i - \sqrt{\frac{6q_i}{T} \log \frac{2N}{\delta}} \geq q_i - \sqrt{\frac{q_i}{8m(\boldsymbol{q})}} \geq \frac{1}{2m(\boldsymbol{q})}\,.\end{aligned}$$

Therefore by the pigeonhole principle we have $m(\hat{\boldsymbol{q}}) \leq 2m(\boldsymbol{q})$. For the other direction we proceed in a similar fashion. Since the failure event F does not hold we have for $i \leq m(\boldsymbol{q})$ that

$$\hat{q}_i \leq q_i + \sqrt{\frac{6q_i}{T} \log \frac{2N}{\delta}} \leq \frac{1}{m(\boldsymbol{q})} \left(1 + \sqrt{\frac{1}{8}}\right) \leq \frac{3}{2m(\boldsymbol{q})}\,.$$

Therefore $m(\hat{\boldsymbol{q}}) \geq 2m(\boldsymbol{q})/3$ as required. □

Proof of Theorem 19. Recall that $A = \{a \in \mathcal{A} : \hat{p}_a \leq 1/m(\hat{\boldsymbol{q}})\}$. Then, for $a \in A$, the algorithm estimates μ_a from $T_A \doteq T/(2m(\hat{\boldsymbol{q}}))$ samples. From lemma 25, $T_A \geq T/(4m(\boldsymbol{q}))$ with probability $(1 - \delta)$. Let H be the event $T_A < T/(4m(\boldsymbol{q}))$ and G be the event $\exists a \in A : |\mu_a - \hat{\mu}_a| \geq \sqrt{\frac{2m(\boldsymbol{q})}{T} \log \frac{2N}{\delta}}$

$$\mathrm{P}(G) \leq \mathrm{P}(H) + \mathrm{P}(G|\neg H) \leq \delta + \mathrm{P}(G|\neg H)$$

Via Hoeffding's inequality and the union bound,

$$\mathrm{P}\left(G|\neg H\right) \doteq \mathrm{P}\left(\exists a \in A : |\mu_a - \hat{\mu}_a| \geq \sqrt{\frac{2m(\boldsymbol{q})}{T}\log\frac{2N}{\delta}},\ \text{given } T_A \geq T/(4m(\boldsymbol{q}))\right) \leq \delta$$
$$\implies \mathrm{P}\left(G\right) \doteq \mathrm{P}\left(\exists a \in A : |\mu_a - \hat{\mu}_a| \geq \sqrt{\frac{2m(\boldsymbol{q})}{T}\log\frac{2N}{\delta}}\right) \leq 2\delta\,.$$

For arms not in A,

$$\begin{aligned}\hat{p}_a =& \frac{2}{T}\sum_{t=1}^{T/2}\mathbb{1}\{X_i = j\} \geq 1/m(\hat{\boldsymbol{q}}),\ \text{by definition of not being in } A\\ \geq& \frac{1}{2m(\boldsymbol{q})},\ \ \text{with probability } 1-\delta\\ \implies T_a \doteq& \sum_{t=1}^{T/2}\mathbb{1}\{X_i = j\} \geq \frac{T}{4m(\boldsymbol{q})},\ \ \text{with probability } 1-\delta\end{aligned}$$

Again applying Hoeffding's and the union bound

$$\mathrm{P}\left(\exists a \notin A : |\hat{\mu}_a - \mu_a| \geq \sqrt{\frac{2m(\boldsymbol{q})}{T}\log\frac{2N}{\delta}}\right) \leq 2\delta$$

Therefore, combining this result with the bound for arms $a \in A$, we have with probability at least $1-4\delta$ that,

$$(\forall a \in \mathcal{A}) \qquad |\hat{\mu}_a - \mu_a| \leq \sqrt{\frac{2m(\boldsymbol{q})}{T}\log\frac{2N}{\delta}} \doteq \varepsilon\,.$$

If this occurs, then

$$\mu_{\hat{a}_T^*} \geq \hat{\mu}_{\hat{a}_T^*} - \varepsilon \geq \hat{\mu}_{a^*} - \varepsilon \geq \mu_{a^*} - 2\varepsilon\,.$$

Therefore

$$\begin{aligned}\mu^* - \mathbb{E}[\mu_{\hat{a}_T^*}] &\leq 4\delta + \varepsilon\\ &\leq \frac{8m(\boldsymbol{q})}{T} + \sqrt{\frac{2m(\boldsymbol{q})}{T}\log\frac{NT}{m(\boldsymbol{q})}},\ \ \text{letting } \delta = \frac{2m(\boldsymbol{q})}{T}\\ &\leq \sqrt{\frac{20m(\boldsymbol{q})}{T}\log\frac{NT}{m(\boldsymbol{q})}},\ \ \text{via Jenson's Inequality}\end{aligned}$$

which completes the result. □

Proof of Theorem 20

We follow a relatively standard path by choosing multiple environments that have different optimal arms, but which cannot all be statistically separated in T rounds. Assume without loss of generality that $q_1 \leq q_2 \leq \ldots \leq q_N \leq 1/2$. For each i define reward function r_i by

$$r_0(\boldsymbol{X}) = \frac{1}{2} \qquad\qquad r_i(\boldsymbol{X}) = \begin{cases}\frac{1}{2}+\varepsilon & \text{if } X_i = 1\\ \frac{1}{2} & \text{otherwise}\,,\end{cases}$$

where $1/4 \geq \varepsilon > 0$ is some constant to be chosen later. We abbreviate $R_{T,i}$ to be the expected simple regret incurred when interacting with the environment determined by $\boldsymbol{q}$ and r_i. Let P_i be the corresponding measure on all observations over all T rounds and $\mathbb{E}_i$ the expectation with respect to P_i. By Lemma 2.6 by Tsybakov [161] we have

$$\mathrm{P}_0\left\{\hat{a}_T^* = a^*\right\} + \mathrm{P}_i\left\{\hat{a}_T^* \neq a^*\right\} \geq \exp\left(-\mathrm{KL}(\mathrm{P}_0, \mathrm{P}_i)\right) ,$$

where $\mathrm{KL}(\mathrm{P}_0, \mathrm{P}_i)$ is the KL divergence between measures P_0 and P_i. Let $T_i(T) = \sum_{t=1}^{T} \mathbb{1}\{a_t = do(X_i = 1)\}$ be the total number of times the learner intervenes on variable i by setting it to 1. Then for $i \leq m$ we have $q_i \leq 1/m$ and the KL divergence between P_0 and P_i may be bounded using the telescoping property (chain rule) and by bounding the local KL divergence by the χ-squared distance as by Auer et al. [16]. This leads to

$$\mathrm{KL}(\mathrm{P}_0, \mathrm{P}_i) \leq 6\varepsilon^2 \mathbb{E}_0\left[\sum_{t=1}^{T} \mathbb{1}\{X_{t,i} = 1\}\right] \leq 6\varepsilon^2\left(\mathbb{E}_0 T_i(T) + q_i T\right) \leq 6\varepsilon^2\left(\mathbb{E}_0 T_i(T) + \frac{T}{m}\right) .$$

Define set $A = \{i \leq m : \mathbb{E}_0 T_i(T) \leq 2T/m\}$. Then for $i \in A$ and choosing $\varepsilon = \min\left\{1/4, \sqrt{m/(18T)}\right\}$ we have

$$\mathrm{KL}(\mathrm{P}_0, \mathrm{P}_i) \leq \frac{18T\varepsilon^2}{m} = 1 .$$

Now $\sum_{i=1}^{m} \mathbb{E}_0 T_i(T) \leq T$, which implies that $|A| \geq m/2$. Therefore

$$\sum_{i \in A} \mathrm{P}_i\left\{\hat{a}_T^* \neq a\right\} \geq \sum_{i \in A} \exp\left(-\mathrm{KL}(\mathrm{P}_0, \mathrm{P}_i)\right) - 1 \geq \frac{|A|}{e} - 1 \geq \frac{m}{2e} - 1 .$$

Therefore there exists an $i \in A$ such that $\mathrm{P}_i\left\{\hat{a}_T^* \neq a^*\right\} \geq \frac{\frac{m}{2e} - 1}{m}$. Therefore if $\varepsilon < 1/4$ we have

$$R_{T,i} \geq \frac{1}{2}\mathrm{P}\left\{\hat{a}_T^* \neq a^* | i\right\}\varepsilon \geq \frac{\frac{m}{2e} - 1}{2m}\sqrt{\frac{m}{18T}} .$$

Otherwise $m \geq 18T$ so $\sqrt{m/T} = \Omega(1)$ and

$$R_{T,i} \geq \frac{1}{2}\mathrm{P}\left\{\hat{a}_T^* \neq a^* | i\right\}\varepsilon \geq \frac{1}{4}\frac{\frac{m}{2e} - 1}{2m} \in \Omega(1)$$

as required.

Proof of Theorem 21

Proof. First note that X_t, Y_t are sampled from Q. We define $Z_a(X_t) = Y_t R_a(X_t)\mathbb{1}\{R_a(X_t) \leq B_a\}$ and abbreviate $Z_{at} = Z_a(X_t)$, $R_{at} = R_a(X_t)$ and $\mathrm{P}\{.|a\} = \mathrm{P}_a\{.\}$. By definition we have $|Z_{at}| \leq B_a$ and

$$\mathrm{Var}_Q[Z_{at}] \leq \mathbb{E}_Q[Z_{at}^2] \leq \mathbb{E}_Q[R_{at}^2] = \mathbb{E}_a[R_{at}] = \mathbb{E}_a\left[\frac{\mathrm{P}_a\left\{\mathcal{P}\mathrm{a}_Y(X)\right\}}{\mathrm{Q}\left(\mathcal{P}\mathrm{a}_Y(X)\right)}\right] \leq m(\eta) .$$

Checking the expectation we have

$$\mathbb{E}_Q[Z_{at}] = \mathbb{E}_a\left[Y\mathbb{1}\{R_{at} \leq B_a\}\right] = \mathbb{E}_a Y - \mathbb{E}_a\left[Y\mathbb{1}\{R_{at} > B_a\}\right] = \mu_a - \beta_a ,$$

where

$$0 \leq \beta_a = \mathbb{E}_a[Y\mathbb{1}\{R_{at} > B_a\}] \leq \mathrm{P}_a\left\{R_{at} > B_a\right\}$$

is the negative bias. The bias may be bounded in terms of $m(\eta)$ via an application of Markov's inequality.

$$\beta_a \leq \mathrm{P}_a\left\{R_{at} > B_a\right\} \leq \frac{\mathbb{E}_a[R_{at}]}{B_a} \leq \frac{m(\eta)}{B_a}\,.$$

Let $\varepsilon_a > 0$ be given by

$$\varepsilon_a = \sqrt{\frac{2m(\eta)}{T}\log\left(2T|\mathcal{A}|\right)} + \frac{3B_a}{T}\log\left(2T|\mathcal{A}|\right)\,.$$

Then by the union bound and Bernstein's inequality

$$\mathrm{P}\left(\text{exists } a \in \mathcal{A} : |\hat{\mu}_a - \mathbb{E}_Q[Z_{at}]| \geq \varepsilon_a\right) \leq \sum_{a\in\mathcal{A}} \mathrm{P}\left(|\hat{\mu}_a - \mathbb{E}_Q[Z_{at}]| \geq \varepsilon_a\right) \leq \frac{1}{T}\,.$$

Let $I = \hat{a}_T^*$ be the action selected by the algorithm, $a^* = \arg\max_{a\in\mathcal{A}} \mu_a$ be the true optimal action and recall that $\mathbb{E}_Q[Z_{at}] = \mu_a - \beta_a$. Assuming the above event does not occur we have,

$$\mu_I \geq \hat{\mu}_I - \varepsilon_I \geq \hat{\mu}_{a^*} - \varepsilon_I \geq \mu^* - \varepsilon_{a^*} - \varepsilon_I - \beta_{a^*}\,.$$

By the definition of the truncation we have

$$\varepsilon_a \leq \left(\sqrt{2}+3\right)\sqrt{\frac{m(\eta)}{T}\log\left(2T|\mathcal{A}|\right)}$$

and

$$\beta_a \leq \sqrt{\frac{m(\eta)}{T}\log\left(2T|\mathcal{A}|\right)}\,.$$

Therefore for $C = \sqrt{2}+4$ we have

$$\mathrm{P}\left(\mu_I \geq \mu^* - C\sqrt{\frac{m(\eta)}{T}\log\left(2T|\mathcal{A}|\right)}\right) \leq \frac{1}{T}\,.$$

Therefore

$$\mu^* - \mathbb{E}[\mu_I] \leq C\sqrt{\frac{m(\eta)}{T}\log\left(2T|\mathcal{A}|\right)} + \frac{1}{T}$$

as required. □

Relationship between $m(\eta)$ and $m(\boldsymbol{q})$

Proposition 26. *In the parallel bandit setting, $m(\eta^*) \leq 2m(\boldsymbol{q})$.*

Proof. Recall that in the parallel bandit setting,

$$\mathcal{A} = \{do()\} \cup \{do(X_i = j) \colon 1 \leq i \leq N \text{ and } j \in \{0,1\}\}$$

Let:

$$\eta_a = \mathbb{1}\left\{ \mathrm{P}\left(X_i = j\right) < \frac{1}{m(\boldsymbol{q})} \right\} \frac{1}{2m(\boldsymbol{q})} \text{ for } a \in do(X_i = j)$$

Let $D = \sum_{a \in do(X_i=j)} \eta_a$. From the definition of $m(\boldsymbol{q})$,

$$\sum_{a \in do(X_i=j)} \mathbb{1}\left\{ \mathrm{P}\left(X_i = j\right) < \frac{1}{m(\boldsymbol{q})} \right\} \leq m(\boldsymbol{q}) \implies D \leq \frac{1}{2}$$

Let $\eta_a = \frac{1}{2} + (1 - D)$ for $a = do()$ such that $\sum_{a \in \mathcal{A}} \eta_a = 1$

Recall that,

$$m(\eta) = \max_a \mathbb{E}_a \left[\frac{\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}{\mathrm{Q}\left(\mathcal{P}\mathrm{a}_Y(X)\right)} \right]$$

We now show that our choice of η ensures $\mathbb{E}_a \left[\frac{\mathrm{P}\{\mathcal{P}\mathrm{a}_Y(X)|a\}}{\mathrm{Q}(\mathcal{P}\mathrm{a}_Y(X))} \right] \leq 2m(\boldsymbol{q})$ for all actions a.

For the actions $a : \eta_a > 0$, ie $do()$ and $do(X_i = j) : \mathrm{P}\left(X_i = j\right) < \frac{1}{m(\boldsymbol{q})}$,

$$\mathbb{E}_a \left[\frac{\mathrm{P}\left\{X_1 ... X_N | a\right\}}{\sum_b \eta_b \mathrm{P}\left\{X_1 ... X_N | b\right\}} \right] \leq \mathbb{E}_a \left[\frac{\mathrm{P}\left\{X_1 ... X_N | a\right\}}{\eta_a \mathrm{P}\left\{X_1 ... X_N | a\right\}} \right] = \mathbb{E}_a \left[\frac{1}{\eta_a} \right] \leq 2m(\boldsymbol{q})$$

For the actions $a : \eta_a = 0$, ie $do(X_i = j) : \mathrm{P}\left(X_i = j\right) \geq \frac{1}{m(\boldsymbol{q})}$,

$$\begin{aligned} \mathbb{E}_a \left[\frac{\mathrm{P}\left\{X_1 ... X_N | a\right\}}{\sum_b \eta_b \mathrm{P}\left\{X_1 ... X_N | b\right\}} \right] \leq & \mathbb{E}_a \left[\frac{\mathbb{1}\{X_i = j\} \prod_{k \neq i} \mathrm{P}\left(X_k\right)}{(1/2 + D) \prod_k \mathrm{P}\left(X_k\right)} \right] \\ = & \mathbb{E}_a \left[\frac{\mathbb{1}\{X_i = j\}}{(1/2 + D) \mathrm{P}\left(X_i = j\right)} \right] \leq \mathbb{E}_a \left[\frac{\mathbb{1}\{X_i = j\}}{(1/2)(1/m(\boldsymbol{q}))} \right] \leq 2m(\boldsymbol{q}) \end{aligned}$$

Therefore $m(\eta*) \leq m(\eta) \leq 2m(\boldsymbol{q})$ as required.

□

Bibliography

[1] Abadie, A. and Imbens, G. (2002). Simple and bias-corrected matching estimators for average treatment effects.

[2] Abadie, A. and Imbens, G. W. (2006). Large Sample Properties of Matching Estimators. *Econometrica*, 74(1):235–267.

[3] Agarwal, A., Hsu, D., Kale, S., Langford, J., Li, L., and Schapire, R. E. (2014). Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1638–1646.

[4] Agrawal, R. (1995). Sample Mean Based Index Policies with O (log n) Regret for the Multi-Armed Bandit Problem. *Advances in Applied Probability*, 27(4):1054–1078.

[5] Agrawal, S. and Goyal, N. (2013a). Further optimal regret bounds for thompson sampling. In *Artificial Intelligence and Statistics*, volume 31, pages 99–107.

[6] Agrawal, S. and Goyal, N. (2013b). Thompson Sampling for Contextual Bandits with Linear Payoffs. *ICML*.

[7] Alekseyenko, A. V., Lytkin, N. I., Ai, J., Ding, B., Padyukov, L., Aliferis, C. F., and Statnikov, A. (2011). Causal graph-based analysis of genome-wide association data in rheumatoid arthritis. *Biology Direct*, 6(1):25.

[8] Alon, N. and Cesa-Bianchi, N. (2013). From Bandits to Experts: A tale of Domination and Independence. *arXiv preprint arXiv: \ldots*, pages 1–22.

[9] Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online Learning with Feedback Graphs : Beyond Bandits. *Colt*, pages 1–26.

[10] Anglemyer, A., Horvath, H. T., and Bero, L. (2014). Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials. *The Cochrane database of systematic reviews*, 4(4):MR000034.

[11] Audibert, J. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *Proceedings of the 22nd annual Conference On Learning Theory*, pages 773—-818.

[12] Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13—-p.

[13] Audibert, J. Y. and Munos, R. (2009). Exploration-exploitation trade-off using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902.

[14] Audibert, J.-Y., Munos, R., and Szepesvari, C. (2007). Tuning Bandit Algorithms in Stochastic Environments. *Algorithmic Learning Theory*, pages 150–165.

[15] Auer, P., Cesa-bianchi, N., and Fischer, P. (2002a). Finite time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

[16] Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pages 322–331.

[17] Auer, P., Cesa-bianchi, N., Freund, Y., and Schapire, R. (2002b). The non-stochastic multi-armed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

[18] Auer, P. and Chiang, C.-K. (2016). An algorithm with nearly optimal pseudo-regret for both stochastic and adversarial bandits. In *Conference on Learning Theory (COLT)*, pages 116–120.

[19] Austin, P. C. (2011). An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, 46(3):399–424.

[20] Avner, O., Mannor, S., and Shamir, O. (2012). Decoupling Exploration and Exploitation in Multi-Armed Bandits. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 409–416.

[21] Bareinboim, E., Forney, A., and Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, number November, pages 1342–1350.

[22] Bareinboim, E. and Lee, S. (2013). Transportability from Multiple Environments with Limited Experiments. *Advances in Neural \ldots*, pages 1–9.

[23] Bartók, G., Foster, D. P., Pál, D., Rakhlin, A., and Szepesvári, C. (2014). Partial monitoring-classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997.

[24] Bay, S., Shrager, J., Pohorille, a., and Langley, P. (2002). Revising regulatory networks: from expression data to linear causal models. *Journal of Biomedical Informatics*, 35(5-6):289–297.

[25] Besbes, O., Gur, Y., and Zeevi, A. (2014). Stochastic Multi-Armed-Bandit Problem with Non-stationary Rewards. In *Advances in Neural Information Processing Systems (NIPS)*, pages 199–207.

[26] Bhattacharya, J. and Vogt, W. B. (2012). Do Instrumental Variables Belong in Propensity Scores? *International Journal of Statistics & Economics*, 9(A12):107–127.

[27] Bingham, S. and Riboli, E. (2004). Diet and cancer—the European prospective investigation into cancer and nutrition. *Nature Reviews Cancer*, 4(3):206–215.

[28] Bottou, L., Peters, J., Ch, P., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. (2013). Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising. *Journal of Machine Learning Research*, 14:3207–3260.

[29] Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–231.

[30] Bubeck, S., Cesa-Bianchi, N., and Others (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends\textregistered in Machine Learning*, 5(1):1–122.

[31] Bubeck, S., Munos, R., and Stoltz, G. (2009a). Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pages 23–37. Springer.

[32] Bubeck, S., Munos, R., and Stoltz, G. (2009b). Pure exploration in multi-armed bandits problems. In *Algorithmic Learning Theory*, pages 23–37. Springer.

[33] Bubeck, S., Munos, R., Stoltz, G., and Szepesvári, C. (2010). X-armed bandits. *Multi-Armed Bandits*, pages 1–38.

[34] Bubeck, S. and Slivkins, A. (2012). The best of both worlds : stochastic and adversarial bandits. In *Conference on Learning Theory (COLT)*.

[Buccapatnam et al.] Buccapatnam, S., Eryilmaz, A., and Shroff Ness, B. Stochastic Bandits with Side Observations on Networks. *ACM SIGMETRICS14, June 2014, Austin, Texas*.

[36] Campbell, D., Stanley, J., and Gage, N. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin, Boston.

[37] Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, 3:1801–1863.

[38] Cartwright, N. (2007). *Hunting causes and using them: Approaches in philosophy and economics*. Cambridge University Press.

[39] Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.

[40] Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257.

[41] Chen, B. (2016). Identification and Overidentification of Linear Structural Equation Models. In *NIPS*, number Nips, pages 1579–1587.

[42] Claassen, T., Mooij, J., and Heskes, T. (2013). Learning sparse causal models is not Np-hard. In *Uncertainty in Artificial Intelligence*.

[43] Cochran, W. G. W. and Rubin, D. D. B. (1973). Controlling bias in observational studies: A review. *Sankhyā: The Indian Journal of Statistics, Series A*, 35(4):417–446.

[44] Cohen, M. and Nagel, E. (1934). *An Introduction to Logic and Scientific Method*. Harcourt, Brace and Co., New York.

[45] Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, 40(1):294–321.

[46] Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., and Schölkopf, B. (2010). Inferring deterministic causal relations. In *Uncertainty in Artificial Intelligence*.

[47] Dawid, A. (2000). Causal inference without counterfactuals. *Journal of the American Statistical Association*.

[48] Dehejia, R. H. and Wahba, S. (1999). Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American statistical Association*, 94(448):1053–1062.

[49] Dorie, V., Hill, J., Shalit, U., Cervone, D., and Scott, M. (2016). Is Your SatT where It s At ? A causal Inference Data Analysis Challenge. Atlantic Causal Inference Conference.

[50] Drton, M., Foygel, R., and Sullivant, S. (2011). Global identifiability of linear structural equation models. *The Annals of Statistics*, pages 865–886.

[52] Dudik, M., Hsu, D., Kale, S., Karampatziakis, N., Langford, J., and Reyzin, L. (2011). Efficient Optimal Learning for Contextual Bandits. In *UAI*.

[51] Dud$\backslash$$\ backslash$ik, M., Langford, J., Li, L., Dudik, M., Langford, J., and Li, L. (2011). Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

[53] Eberhardt, F. (2010). Causal Discovery as a Game. In *NIPS Causality: Objectives and Assessment*, pages 87–96.

[] Eberhardt, F., Glymour, C., and Scheines, R. (2005). On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. In *UAI*.

[55] Even-Dar, E., Mannor, S., and Mansour, Y. (2002). PAc bounds for multi-armed bandit and Markov decision processes. In *Computational Learning Theory*, pages 255–270.

[56] Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. (2010). Parametric bandits: The generalized linear case. In *Advances in Neural Information Processing Systems*, pages 586–594.

[57] Fraker, T. and Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources*, 22(2):194–227.

[58] Friedmann, E. and Thomas, S. A. (1995). Pet ownership, social support, and one-year survival after acute myocardial infarction in the Cardiac Arrhythmia Suppression Trial (CasT). *The American journal of cardiology*, 76(17):1213–1217.

[59] Frolich, M. (2001). Nonparametric Covariate Adjustment: Pair-matching versus Local Polynomial Matching.

[60] Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, pages 3212–3220.

[61] Gao, B. and Cui, Y. (2015). Learning directed acyclic graphical structures with genetical genomics data. *Bioinformatics*, (September):btv513.

[62] Garivier, A., Lattimore, T., and Kaufmann, E. (2016). On Explore-Then-Commit strategies. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 784–792. Curran Associates, Inc.

[64] Garivier, A. and Moulines, E. (2011). On Upper-Confidence Bound Policies for Switching Bandit Problems. *International Conference on Algorithmic Learning Theory (ALT)*, pages 174–188.

[63] Garivier, A. A. and Moulines, E. (2008). On Upper-Confidence Bound Policies for Non-Stationary Bandit Problems. *arXiv preprint arXiv:0805.3415*, (22):174–188.

[65] Geiger, D., Verma, T., and Pearl, J. (1990). Identifying independence in bayesian networks.

[66] Gelman, A. (2010). Causality and Statistical Learning.

[67] Gretton, A., Fukumizu, K., Teo, C. H., Song, L., Schölkopf, B., and Smola, A. J. (2007). A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, pages 585–592.

[68] Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica, Journal of the Econometric Society*, 11(1):1–12.

[69] Hagerup, T. and Rüb, C. (1990). A guided tour of chernoff bounds. *Information Processing Letters*, 33(6):305–308.

[70] Hauser, A. and Bühlmann, P. (2014). Two optimal strategies for active learning of causal models from interventional data. *International Journal of Approximate Reasoning*, 55(4):926–939.

[71] Heckman, J., Pinto, R., and Heckman, J. (2015). Causal analysis after Haavelmo. *Econometric Theory*, 31(01):115–151.

[72] Heckman, J. J. (2008). Econometric causality. *International Statistical Review*.

[73] Heckman, J. J., Ichimura, H., Smith, J., and Todd, P. (1998). Characterizing selection bias using experimental data. Technical Report 5, National bureau of economic research.

[74] Heckman, J. J., Ichimura, H., and Todd, P. E. (1997). Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *The review of economic studies*, 64(4):605–654.

[75] Hill, J. L. (2011). Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240.

[76] Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260):663–685.

[77] Hoyer, P., Hyvarinen, A., and Scheines, R. (2012). Causal discovery of linear acyclic models with arbitrary distributions. *arXiv*.

[78] Hoyer, P., Janzing, D., and Mooij, J. (2009). Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*.

[79] Hu, H., Li, Z., and Vetta, A. R. (2014). Randomized Experimental Design for Causal Graph Discovery. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27*, pages 2339–2347. Curran Associates, Inc.

[80] Huang, Y. and Valtorta, M. (2006). Pearls Calculus of Intervention Is Complete. In Richardson, T. S. and Dechter, R., editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

[81] Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, Inference and Sensitivity Analysis for Causal Mediation Effects. *Statistical Science*, 25(1):51–71.

[82] Imbens, G. W. (2004). NOnpAraMetRic EstImaTioN oF aVerAge TreAtmEnt EffEctS uNdeR eXogEneIty : A rEviEw $*$. $Reviewof Economicsandstatistics, 86(February) : 4 - -29$.

[83] Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

[84] Jain, A., Concato, J., and Leventhal, J. M. (2002). How good is the evidence linking breastfeeding and intelligence? *Pediatrics*, 109(6):1044–1053.

[85] Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil\ucb\: An Optimal Exploration Algorithm for Multi-Armed Bandits. In *Proceedings of The 27th Conference on Learning Theory*, pages 423–439.

[86] Janzing, D., Balduzzi, D., Grosse-Wentrup, M., and Scholkopf, B. (2013). Quantifying causal influences. *The Annals of Statistics*, 41(5):2324–2358.

[87] Janzing, D., Mooij, J., Zhang, K., Lemeire, J., Zscheischler, J., Daniusis, P., Steudel, B., and Schölkopf, B. (2012). Information-geometric approach to inferring causal directions. *Artificial Intelligence*, 182-183:1–31.

[88] Janzing, D. and Peters, J. (2012). On causal and anticausal learning. In *International Conference on Machine Learning*.

[89] Johansson, F. D., Shalit, U., and Sontag, D. (2016). Learning Representations for Counterfactual Inference. In *ICML*, volume 48, New York.

[90] Kaelbling, L. P. (1994). Associative reinforcement learning: Functions ink-dnf. *Machine Learning*, 15(3):279–298.

[91] Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, VV(Ii).

[92] Karnin, Z., Koren, T., and Somekh, O. (2013). Almost optimal exploration in multi-armed bandits. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1238–1246.

[93] Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. Number 1, pages 1–16. Springer.

[94] Kocák, T., Neu, G., Valko, M., and Munos, R. (2014). Efficient learning by implicit exploration in bandit problems with side observations. *Neural Information Processing Systems*, pages 1–9.

[95] Koller, D. and Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*. MIT Press.

[96] Lai, T. L. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22.

[97] LaLonde, R. J. (1986). Evaluating the econometric evaluations of training programs with experimental data. *The American economic review*, pages 604–620.

[98] Langford, J., Strehl, A., and Wortman, J. (2008). Exploration scavenging. *Proceedings of the 25th international conference on Machine learning - ICML 08*, pages 528–535.

[99] Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824.

[100] Lattimore, F., Lattimore, T., and Reid, M. D. (2016). Causal Bandits: Learning Good Interventions via Causal Inference. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, number Nips, pages 1181–1189. Curran Associates, Inc.

[101] Lattimore, T. (2015). Optimally Confident Ucb : Improved Regret for Finite-Armed Bandits. (1):1–16.

[102] Lelarge, M. and Ens, I. (2012). Leveraging Side Observations in Stochastic Bandits. *Uai*.

[103] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010a). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, volume 3, pages 661–670. ACM.

[104] Li, L., Chu, W., Langford, J., and Wang, X. (2010b). Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms.

[] Li, L., Munos, R., Szepesvari, C., Szepesvári, C., and Szepesvari, C. (2014). On minimax optimal offline policy evaluation. *arXiv preprint arXiv:1409.3653*, pages 1–15.

[106] Lopez-Paz, D., Muandet, K., and Recht, B. (2014). The Randomized Causation Coefficient. *arXiv preprint arXiv:1409.4366*.

[107] Lunceford, J. K., Lunceford, J. K., Davidian, M., and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment e ects: a comparative study. *Statistics in Medicine*, 2960(19):2937–2960.

[108] Maathuis, M. H., Colombo, D., Kalisch, M., and Bühlmann, P. (2010). Predicting causal effects in large-scale systems from observational data. *Nature Methods*, 7(4):247–248.

[109] Maathuis, M. H., Kalisch, M., and Bühlmann, P. (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164.

[110] Mannor, S. and Shamir, O. (2011). From Bandits to Experts: On the Value of Side-Observations. pages 1–9.

[111] Myers, J. a., Rassen, J. a., Gagne, J. J., Huybrechts, K. F., Schneeweiss, S., Rothman, K. J., Joffe, M. M., and Glynn, R. J. (2011). Effects of adjusting for instrumental variables on bias and precision of effect estimates. *American journal of epidemiology*, 174(11):1213–22.

[112] Ortega, P. A. and Braun, D. A. (2014). Generalized Thompson sampling for sequential decision-making and causal inference. *Complex Adaptive Systems Modeling*, 2(1):2.

[113] Pandey, S., Chakrabarti, D., and Agarwal, D. (2007). Multi-armed bandit problems with dependent arms. *Proceedings of the 24th international conference on Machine learning - ICML 07*, pages 721–728.

[114] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems.

[115] Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, 82(4):669.

[116] Pearl, J. (2000). *Causality: models, reasoning and inference*. MIT Press, Cambridge.

[117] Pearl, J. (2009). Myth, confusion, and science in causal analysis. *Department of Statistics, UCLA*, (January 2000):1–6.

[118] Pearl, J. (2012). On a class of bias-amplifying variables that endanger effect estimates. *arXiv preprint arXiv:1203.3503*, (July):417–424.

[119] Pearl, J. (2014). Interpretation and Identification of Causal Mediation. *Psychological methods*.

[120] Peters, J., Mooij, J., Janzing, D., and Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053.

[121] Piccolboni, A. and Schindelhauer, C. (2001). Discrete prediction games with arbitrary feedback and loss. In *COLT/EuroCOLT*, pages 208–223. Springer.

[122] Poole, D. and Crowley, M. (2013). Cyclic causal models with discrete variables: Markov chain equilibrium semantics and sample ordering. *IJCAI International Joint Conference on Artificial Intelligence*, pages 1060–1068.

[123] Pukelsheim, F. (2006). *Optimal design of experiments*. SIAM.

[124] Ram, R., Chetty, M., and Dix, T. I. (2006). Causal Modeling of Gene Regulatory Network. *2006 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, pages 1–8.

[125] Ramsey, J., Hanson, S., Hanson, C., Halchenko, Y., Poldrack, R., and Glymour, C. (2010). Six problems for causal inference from fMri. *NeuroImage*, 49(2):1545–1558.

[126] Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *Annals of Statistics*, 30(4):962–1030.

[127] Richardson, T. S. and Robins, J. M. (2013). Single World Intervention Graphs (SwiGs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(128).

[128] Rickles, D. (2009). Causality in complex interventions. *Medicine, Health Care and Philosophy*, 12(1):77–90.

[129] Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–536.

[130] Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*, 90(429):122–129.

[131] Rosenbaum, P. and Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

[132] Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

[133] Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*.

[134] Rubin, D. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*.

[135] Rubin, D. (2005). Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331.

[136] Rubin, D. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3):808–840.

[137] Sachs, K. (2005). Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(5721):523–529.

[138] Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120.

[139] Schisterman, E. F., Cole, S. R., and Platt, R. W. (2009). Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology (Cambridge, Mass.)*, 20(4):488.

[140] Settles, B. (2010). Active Learning Literature Survey. Technical Report 1648, University of Wisconsin-Madison.

[141] Shmueli, G. (2010). To Explain or to Predict ? *Statistical science*, 25(3):289–310.

[142] Shpitser, I., J. Evans, R., S. Richardson, T., and M. Robins, J. (2014). Introduction To Nested Markov Models. *Behaviormetrika*, 41(1):3–39.

[143] Shpitser, I. and Pearl, J. (2006a). Identification of conditional interventional distributions. In Richardson, T. S. and Dechter, R., editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*. AUAI Press.

[144] Shpitser, I. and Pearl, J. (2006b). Identification of Joint Interventional Distributions in Recursive semi-Markovian Causal Models. In *Proceedings of the Twenty-First AAAI Conference on Artificial Intelligence*, number July, pages 1219–1226.

[145] Shpitser, I. and Richardson, T. (2012). Parameter and structure learning in nested Markov models. *arXiv preprint arXiv: \ldots*.

[146] Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 238–241.

[147] Smith, G. C. S. and Pell, J. P. (2003). Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. *BMJ: British Medical Journal*, 327(7429):1459.

[148] Smith, J. A. and Todd, P. E. (2001). Reconciling conflicting evidence on the performance of propensity-score matching methods. *American Economic Review*, 91(2):112–118.

[149] Sokolova, E., Hoogman, M., Groot, P., Claassen, T., Vasquez, A. A., Buitelaar, J. K., Franke, B., and Heskes, T. (2015). Causal discovery in an adult AdhD data set suggests indirect link between \textlessi\textgreaterDat1\textless/i\textgreater genetic variants and striatal brain activation during reward processing. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(6):508–515.

[150] Spirtes, P. (1995). Directed cyclic graphical representations of feedback models. *Proceedings of the Eleventh conference on Uncertainty \ldots*.

[151] Spirtes, P., Glymour, C. N., and Scheines, R. (2000). *Causation, prediction, and search*, volume 81. MIT press.

[152] Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009). Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

[153] Statnikov, A., Henaff, M., Lytkin, N. I., and Aliferis, C. F. (2012). New methods for separating causes from effects in genomics data. *BMC genomics*, 13 Suppl 8(Suppl 8):S22.

[154] Strehl, A. L., Langford, J., Li, L., and Kakade, S. M. (2010). Learning from logged implicit exploration data. In *Advances in Neural Information Processing Systems*, pages 2217–2225.

[155] Sugiyama, M., Krauledat, M., and Muller, K.-R. (2007). Covariate Shift Adaptation by Importance Weighted Cross Validation. *Journal of Machine Learning Research*, 8:985–1005.

[156] Sutton, R. S. and Barto, A. G. (1998). *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

[157] Swaminathan, A. and Joachims, T. (2015). Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *ICML*, volume 1, pages 814–823.

[158] Taruttis, F., Spang, R., and Engelmann, J. C. (2015). A statistical approach to virtual cellular experiments: improved causal discovery using accumulation Ida (aIda). *Bioinformatics*, (August):btv461.

[159] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3):285–294.

[160] Tian, J. (2009). Parameter Identification in a Class of Linear Structural Equation Models. In *IJCAI*, pages 1970–1975.

[161] Tsybakov, A. B. (2009). Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats.

[162] Uphoff, E. and Deng, Y. (2013). Causal Discovery in Climate Science Using Graphical Models. In *Third International Workshop on Climate Informatics*, volume 18, pages 2–4.

[163] VanderWeele, T. J. and Hernández-Diaz, S. (2011). Is there a direct effect of pre-eclampsia on cerebral palsy not through preterm birth? *Paediatric and perinatal epidemiology*, 25(2):111–5.

[164] Verma, T. (1993). Graphical aspects of causal models. Technical report.

[165] Wang, C.-c., Member, S., Kulkarni, S. R., and Poor, H. V. (2005). Bandit problems with side observations. *IEEE Transactions on Automatic Control*, 50(3):338–355.

[166] Weikart, D. P. and Others (1970). Longitudinal Results of the Ypsilanti Perry Preschool Project. Final Report. Volume Ii of 2 Volumes.

[167] Weisberg, D. S. and Gopnik, A. (2013). Pretense, counterfactuals, and Bayesian causal models: why what is not real really matters. *Cognitive science*, 37(7):1368–81.

[168] Woodroofe, M. (1979). A one-armed bandit problem with a concomitant variable. *Journal of the American Statistical Association*, 74(368):799–806.

[169] Wooldridge, J. (2009). Should instrumental variables be used as matching variables. Technical Report September 2006, Michigan State University, MI.

[170] Wright, S. (1921). Correlation and causation. *Journal of agricultural research*.

[171] Wu, Y., György, A., and Szepesvári, C. (2015). Online Learning with Gaussian Payoffs and Side Observations. In *Advances in Neural Information Processing Systems*, pages 1360–1368.

[172] Yu, J. Y. and Mannor, S. (2009). Piecewise-stationary bandit problems with side observations. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1177–1184. ACM.

[173] Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, 2(2):121–134.

[174] Zhang, J. (2008). On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias. *Artificial Intelligence*, 172(16-17):1873–1896.

[175] Zhang, K. and Hyvärinen, A. (2008). Distinguishing causes from effects using nonlinear acyclic causal models. *NIPS 2008 Workshop on Causality. URL http://www \ldots*.

[176] Zhang, K., Peters, J., Janzing, D., and Schölkopf, B. (2012). Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv: \ldots*.

[177] Zhao, Z. (2004). Using matching to estimate treatment effects: Data requirements, matching metrics, and Monte Carlo evidence. *review of economics and statistics*, 86(1):91–107.