Learning how to act: making good decisions with machine learning

Finnian Lattimore

A thesis submitted for the degree of Doctor of Philosophy
The Australian National University
November 2017

Except where acknowledged in the customary manner, the material presented in this thesis is, to the best of my knowledge, original and has not been submitted in whole or part for a degree in any university.

Finnian Lattimore

Finnian Lattimore
November, 2017

In vain the Sage, with retrospective eye,
Would from th' 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

Acknowledgements

There are many people without whose support this thesis would not have been possible. I would like to thank my supervisors Dr Cheng Soon Ong, Dr Mark Reid and Dr Tiberio Caetano, Prof. Robert Williamson for chairing my panel, my brother Tor for many interesting and insightful discussions on bandit problems, and my parents for helping me proofread and the many days given to care for my lovely daughters. I would also like to thank Victoria, Inger, Natalie and all the people at ANU thesis bootcamp for giving me the techniques I needed to get past the blank page and get words out. Finally I want to thank my wonderful husband for his unwavering support and willingness to take on the household chaos, and my two beautiful daughters Freya and Anouk for their patience while mummy worked on her ”tesis”.

I am also grateful for the financial support provided by an Australian Postgraduate Award and a NICTA/Data61 top up scholarship.

This research is supported by an Australian Research Training Program (RTP) Scholarship.

Abstract

This thesis is about machine learning and statistical approaches to decision making. How can we learn from data to anticipate the consequence of, and optimally select, interventions or actions? Problems such as deciding which medication to prescribe to patients, who should be released on bail, and how much to charge for insurance are ubiquitous, and have far reaching impacts on our lives. There are two fundamental approaches to learning how to act: reinforcement learning, in which an agent directly intervenes in a system and learns from the outcome, and observational causal inference, whereby we seek to infer the outcome of an intervention from observing the system.

The goal of this thesis to connect and unify these key approaches. I introduce causal bandit problems: a synthesis that combines causal graphical models, which were developed for observational causal inference, with multi-armed bandit problems, which are a subset of reinforcement learning problems that are simple enough to admit formal analysis. I show that knowledge of the causal structure allows us to transfer information learned about the outcome of one action to predict the outcome of an alternate action, yielding a novel form of structure between bandit arms that cannot be exploited by existing algorithms. I propose an algorithm for causal bandit problems and prove bounds on the simple regret demonstrating it is close to mini-max optimal and better than algorithms that do not use the additional causal information.

Contents

Chapter 1

Introduction

1.1 Motivation

Many of the most important questions in science, commerce and 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 multiple sclerosis? Would delaying teenage pregnancies improve the outcome for their children.

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 - despite the fact that 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 where 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 [10] 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 machine learning (in particular deep learning), machine learning algorithms are effective only within 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 predictions about 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 in, it is critical to ensure 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?

The notion of causality has been widely debated in science and philosophy [83, 111, 123, 119, 105, 173, 75, 39] but is still viewed as poorly defined. This has led to a reluctance among applied researchers in many fields to make concrete claims about 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 which variables are controlled for (or included in a regression). Avoiding formalising causation, which is the real question of interest, requires the reader to determine via "common sense" the implications of the reported associations.

There are two ways in which an association detected in a data set may be non-causal. The first is that the variables concerned may not be related at all, and the association has arisen by chance in that data sample. Given finite data on enough variables, there is a high probability of finding some that appear correlated even though they are completely unrelated. For example, based on data from the years 1999 to 2009, the age of Miss America is strongly correlated with the number of murders (in the US) by steam, hot vapours and hot objects [168]. We would not expect this relationship to hold in a new sample of data. This form of spurious correlation also has serious repercussions. It lies at the heart of major problems with the process of scientific research: researchers are incentivised to detect effects and thus to explore many possible paths in the process of analysing data and studies that fail to find an effect are less likely to be published. Consequently, the likelihood that reported effects have arisen by chance is underestimated, leading to the conclusion that "most published scientific results are false" [87]. This issue is also highlighted by recent crises in replication [115]. This issue can be ameliorated by obtaining more data and by separating learning models from evaluating their performance, for example by evaluating models on a strict hold-out set or on the extent to which their results can be replicated.

However, a strong association, observed in multiple independent studies may still not be causal. The correlation can arise because both variables are consequences of some other, unmeasured factor. For example, the reading ability of children under twelve is strongly correlated with their height, because older children are taller and can read better. However height is not a cause of reading ability because interventions to increase a child's height, for example by giving them growth hormones, would not be expected to improve their reading. Similarly, extra lessons in reading will not make a child grow taller. This problem is fundamentally different to the issue of spurious correlations arising by chance in finite data sets. Obtaining more (even infinitely many more) samples without directly intervening in the system to manipulate the variables does not allow us to separate causation from correlation.

The key distinction between a real, but non-causal, association and a causal relationship is in what happens if we intervene in the system and change one of the variables. In this thesis, I take an interventionist viewpoint of causality: any model or approach designed to predict the outcome of intervening in a system is causal. This viewpoint captures the types of questions that motivate this thesis. How can we change the way we do things to obtain better outcomes?

Causality is often linked to explanation; understanding how and why things happen. I view explanation in terms of compression and generalisation: the amount of information about the world a model can capture. This creates a hierarchy in the degree to which models are explanatory, rather than a simple binary distinction. A standard predictive model encodes all the information needed to predict some output given inputs provided the system generating the data does not change. A high-level causal model might be able to predict the outcome of a specific intervention holding all else fixed. More detailed causal models could predict the outcome for a wide range of combinations of interventions conditional on a range of contexts. By considering conditional interventions within our definition of causal questions we also capture mediation: the study of pathways through which one variable causes another [166]. Finally, a model that can distil how elements interact into mathematical equations like Newton's laws can be used to predict what will happen in an astounding range of settings, including many never previously observed.[1]

Gelman [67], Gelman and Imbens [68] make a distinction between forward causal inference, the types of "what if" questions I focus on in this thesis, and reverse causal questions, asking why something occurs. The former aims to identify the effect of a known cause. The latter can be viewed as identifying causes of an effect. They regard forward causal inference as well defined within the counterfactual and graphical model frameworks for causal inference, that we describe in section 2.1. However, they state that "A reverse causal question does not in general have a well-defined answer, even in a setting where all possible data are made available." I view this as overly pessimistic, depending on how "all possible data" is defined. The goal of identifying the causes of an effect can be formalised within the graphical causal model framework. Solving this problem is certainly much more challenging than identifying the effect of a specific intervention on a given outcome, since it requires us to test or infer the effect of interventions on many different variables. These practical difficulties may well be overwhelming, particularly in fields such as social science and economics where data sets are often relatively small, systems are complex, the variables are difficult to directly manipulate and even relatively simple

[1] Although Newton's laws are not fully general.

"what if" questions are hard to resolve conclusively. However, this does not mean that the problem of identifying causes of effects is ill-posed in principle. It can be viewed as a form of causal discovery: the attempt to learn the structure of the causal relationships between variables, on which there is a rich literature, see Spirtes and Zhang [153] for a recent review.

There has traditionally been a large gap between researchers in machine learning who focus on prediction, using largely non-interpretable models and researchers in statistics, social science and economics who (at least implicitly) aim to answer causal questions and tend to use highly theory-driven models. However, there is relatively little awareness, particularly within the machine learning and data science communities, of what constitutes a causal problem and the implications of this for the training and evaluation of models. In the next section we emphasise the subtlety that can exist in determining if a problem is causal by examining some typical examples.

1.3 What makes a problem causal?

Machine learning is in the midst of a boom, driven by the availability of large data sets and the computation resources to process them. Machine learning techniques are being applied to a huge range of problems, in both industry and academia. The following examples are intended to capture the breadth of problems that machine learning algorithms are actively being applied to. Which, if any, of these problems require causal inference?

- Speech recognition (for systems like Siri or Google Assistant)
- Image classification
- Forecasting the weather
- Identifying spam emails
- Automated essay marking
- Predicting the risk of death in patients with pneumonia.
- Predicting who will re-offend on release from prison
- Customer churn modelling
- Demand prediction for inventory control
- Playing Go

The question is disingenuous because I have not posed the problems in sufficient detail to determine if causality is an important consideration. In particular, I failed to specify how any model we might build would be used: what actions would be taken in response to its predictions.

Consider speech recognition. You say something, which causes sound waves, which are converted to a digital signal that 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 classification we do not particularly care about building a strong model for exactly how the thing that was photographed translates to an array of pixels, provided we can be fairly confident that the process will not change. If we develop a discriminative model that is highly accurate at classifying cats from dogs, we do not need to understand its internal workings (assuming 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, causality can be ignored. Your decision will not affect if it actually rains. However, longer term climate forecasts might (theoretically) lead to 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 both involve processing text to determine an underlying (complex) attribute such as its topic or quality. In both cases, there is inherent competition between the algorithm and the people generating the text. As a result, decisions made by the algorithm are likely to change the relationship between the features it relies on and the true label. Spammers and students will modify their writing in order to optimise their results. A standard supervised learning approach can only work if the resulting change in the mapping from features to label is sufficiently gradual. There are two key ways of ensuring this. The first is to limit people's ability to observe (and thus react to) decisions made by the algorithm. The second is to use a model in which the features are related to the outcome in such a way that they cannot be manipulated independently.

This example also highlights a connection between causal models and transparency in machine learning. If we are using a non-causal model to make decisions affecting people, there will be a trade-off between the performance and transparency of the model; not because the requirement for transparency restricts us to simple models, but because revealing how the model works allows people to change their behaviour to game it.

What about predicting the risk of death in patients with pneumonia? Suppose the goal is to build a 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 for serious cases, this appears to be straightforward prediction. 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 the model is used to make admission decisions. Caruana et al. [40] found exactly this effect in a real data set. The model learned that people suffering asthma were *less* likely to die from pneumonia. This was because doctors treated such patients very aggressively, thus actually lowering their risk. The issue is not with the model; it performed very well at the task for which it was trained, which is to predict who would be likely to die under the original admission and treatment protocols. However, using it to decide how to *change* these protocols could kill. The actual question of interest in this case is what happens to patients with characteristics X if they are assigned treatment according to decision rule (or policy) $\pi(X)$.

Predicting recidivism among paroled prisoners or customer churn also fit within the class of problems where the goal is to identify a group for which a problem will occur in order to target treatment (additional support and monitoring for people on parole, loyalty rewards to improve customer retention, hospitalisation for the severely ill). Predictive models can be applied to such problems where the most effective treatment is known for a given target group, and where deciding who to treat on the basis of the model predictions will not change the relationship between the features and outcome.

Demand prediction seems like a relatively straightforward prediction problem. Models use features such as location, pricing, marketing, time of year and weather, to forecast the demand for a product. It seems unlikely that using the model to ensure stock is available will itself change demand. However, depending on the way demand is measured, 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 a case with some subtleties. At every turn, the AI agent has a number of actions available. The state of the board 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 reachable board states will lead to a win.[2] Supervised learning can also be applied to learn a policy $\mathrm{P}(a|s)$, the probability of a player selecting action a, given board state s. This allows the agent to estimate the likelihood of winning from a given starting state by simulating many times the remainder of the game, drawing actions from $\mathrm{P}(a|s)$ for both players. Google's Alpha Go, which in May 2017 beat the then strongest human player [113], incorporates a combination of these approaches [147]. The supervised learning was enhanced by having the agent play (variants) of itself many times, so that its estimates of value for each board state and of the likelihood an opponent will play a given move are based on a combination of replicating the way humans play and on the moves that led to a win when playing itself.

The problem of playing go is causal from the interventionalist perspective. The agent wishes to learn the probability of a win given an action they take. However, there are some special characteristics of the go problem that make it amenable to a primarily supervised

[2] This is a challenging pattern recognition problem. There are around 2×10^{170} legal board positions in Go, ([163]), so the algorithm cannot simply memorise the proportion of times each state leads to a win. It must identify higher level features of the board state that are associated with winning.

learning approach. The actions the agent has to explore are the same ones as human players explored to generate the training data, and both have the same objective - to win the game. In addition, the state of the board encapsulates all the information relevant to selecting a move. These factors make it reasonable to conclude that selecting moves with an algorithm will not change the value of a board state or the probability of given move by the opponent by a sufficiently large margin to invalidate the training data.

Having considered these examples we can now identify some general aspects of problems that require causal (as opposed to purely predictive) inference. A predictive model may be sufficient if, given the variable(s) being predicted, it is clear which action is optimal and if selecting actions on the basis of the model does not change the mapping from features to outcomes. The second requirement is particularly difficult to satisfy when an algorithm is making important decisions affecting individual people. Think about problems like credit scoring and parole decisions. There are strong ethical grounds for demanding transparency, but if the goals of society and the individuals are not perfectly aligned and there is any possibility that people can manipulate features independently of the outcome, there will be a conflict between model accuracy and transparency. It is rare to build a model without any intent to make some kind of decision based on its results. Thus, I argue we should assume a causal model is required until we can justify otherwise.

1.4 Observe or intervene: two very different approaches to causal problems

As we have shown, problems involving causality are ubiquitous in many fields. As a result, techniques for addressing them have developed in parallel within many disciplines, including statistics, economics, social science, epidemiology and machine learning. Although the focus and terminology can differ substantially between fields, these techniques all address the underlying goal of estimating the effect of, or optimally selecting, interventions in some system. Methods for learning about interventions can be usefully categorised into two broad approaches, reinforcement learning and observational causal inference.

In reinforcement learning, under which we include traditional randomised experiments, we learn the outcome of actions by taking them. We take the role of an agent, capable of intervening in the system, and aim to develop algorithms that allow the agent to select actions optimally with respect to some goal. A particular strand of research within reinforcement learning are multi-armed bandit problems. They describe settings in which there is a set of available actions, the agent repeatedly decides which to select and then observes the outcome of the chosen action. They capture problems, such as a doctor deciding which treatment to prescribe to a patient or a search engine selecting which advertisement to display to a user, where the agent faces the same set of choices repeatedly and is able to assess the value of the outcome they observe.

The approach of learning the outcome of an action by taking it plays a key role in advancing our knowledge of the world. However, we frequently have access to large bodies of data that have been collected from a system in which we did not have any control over what actions were taken, or perfect knowledge of the basis on which those actions were chosen. Estimating the effect of an action from such observational data sets is the problem addressed by observational causal inference. Observational causal inference can be viewed as a form of transfer learning. The goal is to leverage data obtained from one system, the system we have observed, to estimate key characteristics of another, the system after we select an action via some policy that may differ from the process driving which actions occur in the original system. This is impossible without some assumptions about how the original system and the system after intervention are related to one-another. The key to observational inference is to model how actions change the state of the world in such a way that we can map information collected in one setting to another.

Both multi-armed bandits and observational causal inference can be seen as extensions to the concept of randomised controlled trials. Bandit algorithms deal with the sequential nature of the decision making process, and causal inference with problems where randomisation is not feasible, affordable or ethical. The similarities between the problems addressed by these techniques raise the question of whether there are problems best addressed by a combination of these approaches, and if so, how they can be combined.

1.5 This thesis and its contributions

I view causal problems as one of the greatest current challenges for machine learning. They incorporate a large set of problems of huge practical significance, that require us to go beyond pattern recognition, but are well short of general artificial intelligence. For the major advances in representation and pattern recognition developed within machine learning to be effectively applied in many areas of medicine, economics, social sciences and industry, we need to understand how to leverage our improved approaches to prediction to tackle causal problems.

Contributions The goal of this thesis is to connect and unify the key approaches to solving causal problems from both the observational and interventional viewpoints. My major contribution unifies the causal graphical model approach for inference in observational settings with the sequential experimental approach encapsulated by multi-armed bandits. This synthesis allows us to represent knowledge of how variables are related to one-another in a very natural way and induces an interesting and novel form of structure between the different actions modelled in the bandit problem. I develop a new algorithm that can exploit this structure as a first step towards a unified approach to decision making under uncertainty.

I also make a number of additional connective contributions that are not encompassed by my work on causal bandit problems. I demonstrate the role of a formal causal framework within Bayesian approaches to inference and show how assigning a prior based on human causal intuition without considering the causal structure of the problem can introduce bias. I highlight the connections between approaches to off-policy evaluation in bandit problems, causal effect estimation from observational data, and covariate shift. Finally, I clarify the implicit causal structure underlying various bandit settings and the counterfactual nature of *regret* - the measure by which bandit algorithms are assessed.

Thesis overview This thesis is divided into three key chapters: learning from observational data, learning from interventional data and unifying the approaches. Chapter 2 covers learning to act from observational data, where the goal is to learn the outcome of an external intervention in a system from data obtained by observing it without control over which actions are selected. In §2.1, I describe the key existing frameworks for causal inference from observational data, discuss how they relate to one-another and introduce the notation required to describe causal problems. Section §2.2 describes the key tools these frameworks provide that enable us to answer causal questions, in particular, the do-calculus (§2.2.1.2) and, in sections §2.2.2 and §2.2.3, discusses how we define causal effects, the traditional approaches to estimation and how they relate to covariate shift and off-policy evaluation. Section §2.3 highlights the role graphical causal models can play in Bayesian inference.

Chapter 3 deals with the interventionalist viewpoint, including traditional randomised experiments (§3.1) and multi-armed bandit problems. In §3.2, I describe the key problems and results within the bandit literature, including stochastic bandits (§3.2.1), pure exploration problems (§3.2.2), adversarial bandits (§3.2.3) and contextual bandits (§3.2.4). I clarify the causal structure of (stochastic) contextual bandit problems in §3.2.4.1. In §3.2.5, I review the literature on off-policy evaluation for bandit problems and show how it is a somewhat special case of causal effect estimation from observational data. Finally, in §3.3 I discuss the counterfactual nature of bandit regret.

In chapter 4, I introduce causal bandit problems that unify causal graphical models and multi-armed bandit problems. Bandit arms are related to interventions in a causal graphical model in a very natural way: each multi-armed bandit arm (or action) corresponds to a particular assignment of values to variables within the causal graphical model. I show how the causal bandit framework can be used to describe a number of existing problems that lie in the intersection between the observational and interventional approaches to causality and demonstrate when causal bandit problems reduce to different existing bandit settings depending what information is observable and when.

In §4.2, I focus on causal bandit problems for which the values of variables in the causal graph are observed after an action is selected. I demonstrate that this leads to a novel form of structure between the bandit arms that cannot be exploited by existing bandit algorithms. In §4.2.1, I describe and develop an algorithm for a special case of the causal bandit problem, which I refer to as the *parallel bandit problem.* I demonstrate via upper and lower bounds on the regret that the algorithm is close to optimal for these problems and that the introduction of the causal structure leads to substantially faster learning. In

§4.2.2, I develop and prove regret bounds for an algorithm that can be applied to arbitrary causal graphs, albeit with stronger assumptions on what must be known a priori, and introduce a measure that captures the underlying difficulty of causal bandit problems, which depends on the causal graph and can be viewed as an "effective number of arms". I also show how this measure can be used to quantify the value of optimised interventional data over purely observational data. Section 4.2.3 provides experiments demonstrating the performance of both the parallel and general causal bandit algorithms on causal bandit problems. Finally in §4.2.4, I discuss extensions and potential future work.

Chapter 2

Learning from observational data

The goal of causal inference is to learn the effect of taking an action. We can do this directly via experimental approaches, however any given agent only has a limited capacity to manipulate the world. We are generating and storing data on almost every aspect of our lives at an unprecedented rate. As we incorporate sensors and robotics into our cities, homes, cars, everyday products and even our bodies, the breadth and scale of this data will only increase. However, only a tiny fraction of this data will be generated in a controlled way with the specific goal of answering a single question. An agent that can only learn from data when it had explicit control (or perfect knowledge of) the process by which that data was generated will be severely limited. This makes it critical that we develop effective methods that enable us to predict the outcome of an intervention in some system by observing, rather than acting on it. This is the problem of observational causal inference. The key feature that distinguishes observational from interventional data is that the learning agent does not control the action about which they are trying to learn.

2.1 Causal models

Observational causal inference aims to infer the outcome of an intervention in some system from data obtained by observing (but not intervening on) it. As previously mentioned, this is a form of transfer learning; we need to infer properties of the system post-intervention from observations of the system pre-intervention. Mapping properties from one system to another requires some assumptions about how these two systems are related, or in other words, a way of describing actions and how we anticipate a system will respond to them. Three key approaches have emerged: counterfactuals, structural equation models and causal Bayesian networks.

Counterfactuals [137] were developed from the starting point of generalising from randomised trials to less controlled settings. They describe causal effects in terms of differences between counterfactual variables, what would happen if we took one action versus what would happen if we took another. Counterfactual assertions can be expressed very naturally in human languages and are prevalent in everyday conversations; “if I had worked harder I would have got better grades” and “she would have been much sicker if she hadn’t taken antibiotics”. Structural equation models have been developed and applied primarily within economics and related disciplines. They can be seen as an attempt to capture key aspects of the people’s behaviour with mathematics. Questions around designing policies or interventions play a central role in economics. Thus they have transformed simultaneous equations into a powerful framework and associated set of methods for estimating causal effects. The is also a rich strand of work on using the assumptions that can be encoded in structural equation models, also known as functional causal models to discover the structure and direction of causal relationships - see for example [112, 124]. Causal Bayesian networks [119] are a more recent development and arise from the addition of a fundamental assumption about the meaning of a link to Bayesian networks. They inherit and leverage the way Bayesian networks encode conditional independencies between variables to localise the impact of an intervention in a system in a way that allows formalisation of the conditions under which causal effects can be inferred from observational data.

An understanding of causal Bayesian networks and their properties (in particular the do calculus, see section 2.2.1.2) is sufficient to appreciate my main technical contributions in chapter 4, as well as the importance of formal causal reasoning in Bayesian inference that I highlight in section 2.3. However, the literature on causal inference techniques remains split between the different frameworks. Much of the recent work on estimating causal effects within machine learning, as well as widely used methodologies such as propensity scoring, are described using the counterfactual framework. Methods developed within economics, in particular instrumental variable based approaches, or those requiring parametric or functional assumptions, are often based around structural equation models. This makes it worthwhile for researchers interested in causality to develop an understanding of all these viewpoints.

In the next sections, we describe causal Bayesian networks, counterfactuals and structural equation models: the problems they allow us to solve, the assumptions they rely on and how they differ. By describing all three frameworks, how they relate to one-another, and when they can be viewed as equivalent, we will make it easier for researchers familiar with one framework to understand the others and to transfer ideas and techniques between them. However, sections 2.1.3 (Structural Equation models), and 2.1.4 (Unifying the models) are not crucial to understanding my technical contributions and may be safely skipped. In order to demonstrate the notation and formalisms each framework provides, we will use them to describe the following simple examples.

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 four year olds. We have a large cross-sectional data set on a group of twenty year olds that records if they attended preschool, if they graduated high school and their parents socio-economic status (SES). We will let $X \in \{0, 1\}$ indicate if an individual attended preschool, $Y \in \{0, 1\}$ indicate if they graduated high school and $Z \in \{0, 1\}$ represent if they are from a low or high SES background respectively.[1]

2.1.1 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.

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

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.

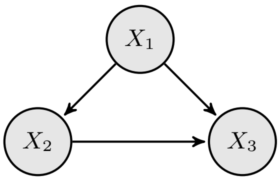


Figure 2.1: A general Bayesian network for the joint distribution over three variables. This network does not encode any conditional independencies between its variables and can thus represent any distribution over three variables.

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

[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 [170].

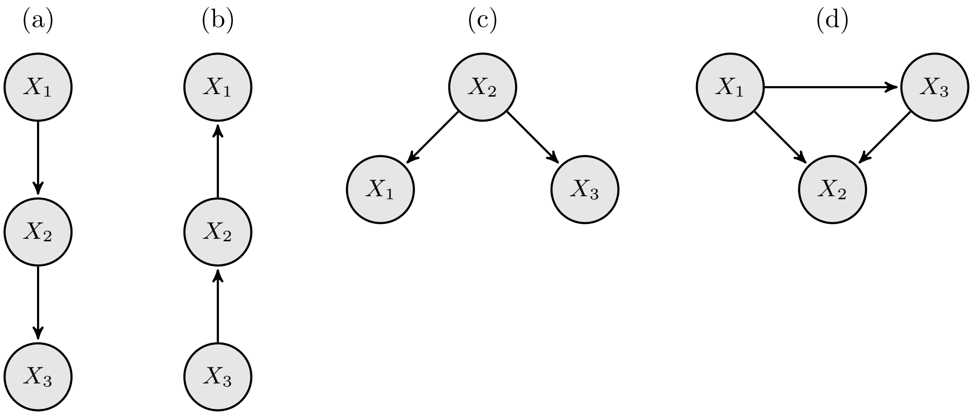


Figure 2.2: Some valid Bayesian networks for a distribution P over (X_1, X_2, X_3) in which X_3 is conditionally independent of X_1 given X_2, denoted $X_3 \perp\!\!\!\perp X_1 | X_2$. Graphs (a), (b) and (c) are all a *perfect map* for P as the graphical structure implies exactly the same set of independencies exhibited by the distribution. Graph (d), like figure 2.1 does not imply any conditional independencies, and is thus a valid (but not very useful) Bayesian network representation for any distribution over three variables.

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 $X_i \rightarrow X_j$, by definition, implies X_i causes X_j. This means an intervention to change the value of X_i can be expected to affect X_j, but interventions on X_j will not affect X_i. 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 [119].

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 to represent a probability distribution/mass 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 $\mathrm{P}\left(Y = 1 | do(X = x)\right)$ as a function of x,
  - the probability mass function for a discrete Y : $\mathrm{P}\left(Y | do(X = x)\right)$,
  - 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 [119])**.** *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. Let $\mathcal{P}\mathrm{a}_{X_i}$ denote the parents of the variable X_i in G.*

$$\mathrm{P}\left(X_1...X_N|do(Z = z)\right) = \mathbb{1}\{Z = z\} \prod_{X_i \notin Z} \mathrm{P}\left(X_i|\, \mathcal{P}\mathrm{a}_{X_i}\right) \tag{2.3}$$

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 truncated 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 $\mathrm{P}\left(Y|do(X = 1)\right)$, relative to if no one was treated $\mathrm{P}\left(Y|do(X = 0)\right)$? The treatment X is a potential cause of Y, along with other unobserved variables, such as the age, gender and the disease subtype of the patient. Since X is assigned via deliberate randomisation, it cannot be affected by any latent variables. The causal Bayesian network for this scenario is shown in figure 2.3. This network represents the (causal) factorisation $\mathrm{P}\left(X, Y\right) = \mathrm{P}\left(X\right)\mathrm{P}\left(Y|X\right)$, so from equation (2.3), $\mathrm{P}\left(Y|do(X)\right) = \mathrm{P}\left(Y|X\right)$. In this example, the interventional distribution is equivalent to the observational one.

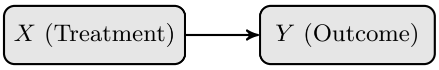


Figure 2.3: Causal Bayesian network for example 1

In example 2 we are interested in $\mathrm{P}\left(Y|do(X = 1)\right)$, the expected high-school graduation rate if we introduce universal preschool. We could compare it to outlawing preschool $\mathrm{P}\left(Y|do(X = 0)\right)$ or the current status quo $\mathrm{P}\left(Y\right)$. 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 affects 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 with high socio-economic status are more

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

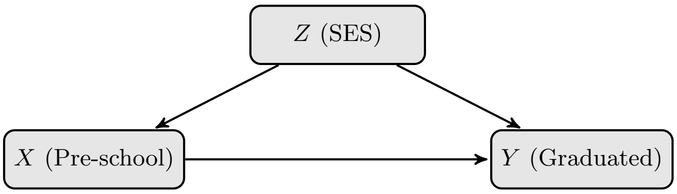


Figure 2.4: Causal Bayesian network for example 2

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 [119]. We discuss them further in section 2.2.1.2.

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 seven 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 thirty parameters.[3]

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. The global Markov property and d-separation also hold in causal networks. D-separation, which characterises which conditional independencies must hold in any distribution that can be represented by a given Bayesian network G, is key to many important results and algorithms for causal inference. We include a brief review of D-separation in section 2.2.1.1.

[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.

2.1.1.1 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 [131, 39]. 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 (figure 2.5). 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.

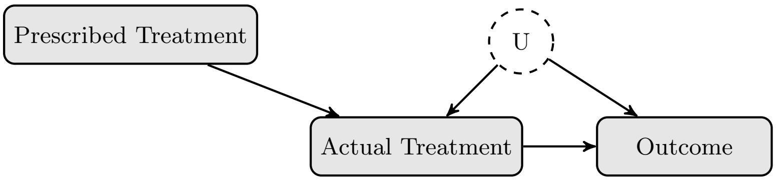


Figure 2.5: Randomised treatment with imperfect compliance

Another key issue with causal Bayesian networks is that they cannot handle cyclic dependencies 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 periods 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, there might then not be sufficient data on each variable for inference. Such problems have mostly been studied within the dynamical 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 [126] compare the equilibrium approach to reasoning about cyclic problems with structural equation models, which we discuss in section 2.1.3 and that can be seen as Bayesian causal networks with additional functional assumptions.

2.1.2 Counterfactuals

The Neyman-Rubin model [137, 138, 135, 139, 140] 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 [171]. Causal effects are differences in counterfactual variables: what the difference is 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 if everyone received the placebo is $\mathrm{P}\left(Y^1\right) - \mathrm{P}\left(Y^0\right)$. We can estimate (from an experimental or observational study):

- $\mathrm{P}\left(Y = 1|X = 1\right)$, the probability that those who took the treatment will recover
- $\mathrm{P}\left(Y = 1|X = 0\right)$, 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 same as the outcome *had* they not taken the treatment. Equivalently stated:

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

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 [135]. 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 from wealthy families could be more likely to attend preschool but also more likely to do better in school regardless, i.e $X \not\!\perp\!\!\!\perp Y^0$ and $X \not\!\perp\!\!\!\perp Y^1$. 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 [135, 119])**.** *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 preschool is independent of the likelihood of graduating high-school had a person attended, then the average rate of high-school graduation given a universal preschool 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) [138]. 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 preschool. 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 objections to counterfactuals arising from the way they describe alternate universes that were never realised. In particular, statements involving joint distributions over counterfactual variables may not be able to be validated empirically Dawid [46]. 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).

2.1.3 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 [175, 71]. 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 [119].

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 2.6 illustrates the SEM for example 1.

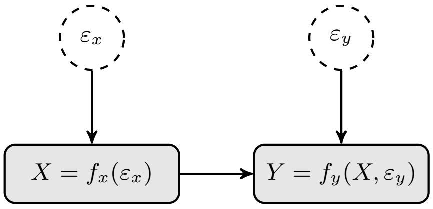


Figure 2.6: SEM for example 1

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 2.6 is identical to that of 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]

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

$$\mathrm{P}\left(Y=y|do(X=1)\right)=\sum_{\varepsilon_y}\mathrm{P}\left(\varepsilon_y\right)\mathbb{1}\{f_y(1,\varepsilon_y)=y\} \tag{2.7}$$

The observational distribution of Y given X is,

$$\mathrm{P}\left(Y=y|X=1\right)=\sum_{\varepsilon_x}\sum_{\varepsilon_y}\mathrm{P}\left(\varepsilon_x|X=1\right)\mathrm{P}\left(\varepsilon_y|\varepsilon_x\right)\mathbb{1}\{f_y(1,\varepsilon_y)=y\} \tag{2.8}$$

$$=\sum_{\varepsilon_y}\mathrm{P}\left(\varepsilon_y\right)\mathbb{1}\{f_y(1,\varepsilon_y)=y\}\text{ , as }\varepsilon_x\perp\!\!\!\perp\varepsilon_y \tag{2.9}$$

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 2.7. Intervening to send all children to preschool replaces the equation $X=f_x(Z,\varepsilon_x)$ with $X=1$, leaving all the other functions and distributions in the model unchanged.

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

$$=\sum_{z}\mathrm{P}\left(z\right)\underbrace{\sum_{\varepsilon_y}\mathrm{P}\left(\varepsilon_y\right)\mathbb{1}\{f_y(1,z,\varepsilon_y)=y\}}_{\mathrm{P}(Y=y|X=1,Z=z)} \tag{2.11}$$

Equation 2.11 corresponds to equations 2.4 and 2.5. It is not equivalent to the observational distribution given by:

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

Structural equation models are generally applied with strong constraints on the functional form of the relationship between the variables and noise, which 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 *causal* assumptions about the relationships between variables. When visualised as a network, the absence of an arrow between two variables encodes the assumption

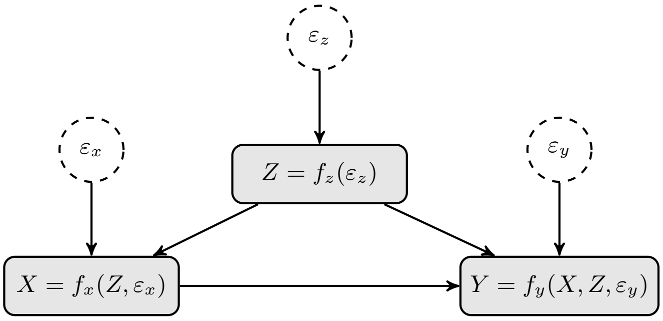


Figure 2.7: SEM for example 2

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 [74, 119].

Granger causality A discussion of approaches to (observational) causal inference would not be complete without a mention of Granger causality, [70]. The fundamental idea behind Granger causality is to leverage the assumption that the future does not cause the past to test the existence and direction of a causal link between two time series. The basic approach is to test, for a pair of time series variables X and Y, if $Y_t \perp\!\!\!\perp (X_1, ..., X_{t-1})|(Y_1, ..., Y_{t-1})$ - that is if the history of X helps to predict Y given the history of Y. The original formulation considered only pairs of variables and linear causal relationships but recent work has generalised the key idea to multiple variables and non-linear relationships. Unlike the previous models we have discussed, Granger causality does not provide us with a means to specify our assumptions about the causal structure between variables. Rather it aims to infer the causal structure of a structural equation model from observational data - subject to some assumptions. I would categorise Granger causality as method for *causal discovery* in time series data, see the discussion of causal discovery versus causal effect estimation in section 2.2.

2.1.4 Comparing and unifying the models

Remarkably for models developed relatively independently in fields with very different approaches and problems, causal Bayesian networks, counterfactuals and structural equation models can be nicely unified for interventional queries (those that can be expressed with the do-notation) [119]. These queries, and the assumptions required to answer them, can be mapped between the frameworks in a straightforward way, allowing techniques developed within one framework to be immediately applied within another. 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:

$$\begin{aligned} \mathrm{P}\left(Y^0\right) &\equiv P(Y|do(X=0)) \\ \mathrm{P}\left(Y^1\right) &\equiv P(Y|do(X=1)) \end{aligned} \tag{2.13}$$

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 made the slightly weaker assumption:

$$X \perp\!\!\!\perp Y^0 \text{ and } X \perp\!\!\!\perp Y^1 \tag{2.14}$$

It is possible to relax the independence of errors assumption for SEMs to correspond exactly with the form of equation (2.14) without losing any of the power provided by d-separation and graphical identification rules [130]. 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 [135] 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 [122, 84, 167] and in legal decisions, particularly on issues such as discrimination [119]. The graphical approach to representing causal knowledge can be extended to cover these types of questions via Single World Intervention Graphs [130], which explicitly represent counterfactual variables in the graph.

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 [119, 152] focuses on asymptotic, non-parametric estimation and rigorous theoretical foundations. The Neyman-Rubin framework builds on the understanding of randomised experiments and generalises to quasi-experimental and observational settings, with a particular focus on non-random assignment to treatment. Treatment variables are typically discrete (often binary). 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 [135]. 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 [76]. The central approach to estimation is regression - which naturally handles continuous variables while discrete variables are typically encoded as indicator variables. In addition, the parametric structural equation models favoured in economics can be extended to analyse cyclic (otherwise referred to as non-recursive) models. However, these differences are not fundamental to the frameworks. Functional assumptions can be specified on the conditional distributions of (causal) Bayesian networks, counterfactuals can readily represent continuous treatments (eg Y^x), and structural equation models can represent complex non-linear relationships between both continuous and discrete variables.

2.1.5 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 [178], demonstrated in real data by Cohen and Nagel [45], and popularised by Simpson [148]. The following example is adapted from Pearl [119]. Suppose a doctor has two treatments, A and B, which she offers to patients to prevent heart disease. She keeps track of the 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 is concerned with $\mathrm{P}\left(Y|do(T)\right)$, not the standard conditional distributions $\mathrm{P}\left(Y|T\right)$. Unfortunately, the data in tables 2.1 and 2.2 is insufficient to enable estimation of the interventional distribution $\mathrm{P}\left(Y|do(T)\right)$ 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 2.8.

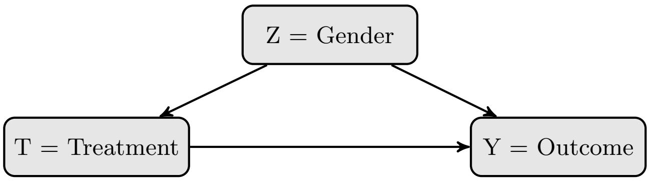


Figure 2.8: An example of causal network that can give rise to Simpson's Paradox. In this case, we should select treatment on the basis of the gender-specific results.

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 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 the 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.

$$\mathrm{P}\left(Y|do(T)\right) = \mathrm{P}\left(Y|T, female\right)\mathrm{P}\left(female\right) + \mathrm{P}\left(Y|T, male\right)\mathrm{P}\left(male\right) \tag{2.15}$$

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 2.9 shows a plausible causal graph for this setting. It differs from the graph in figure 2.8 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.

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

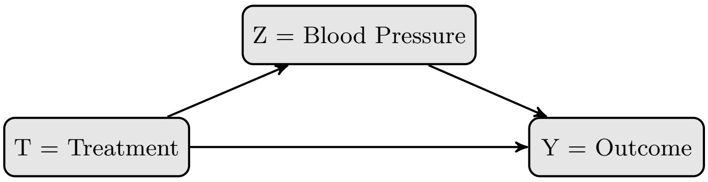


Figure 2.9: Another causal network that can exhibit Simpson's paradox. In this case, "the solution" is not to stratify on Z.

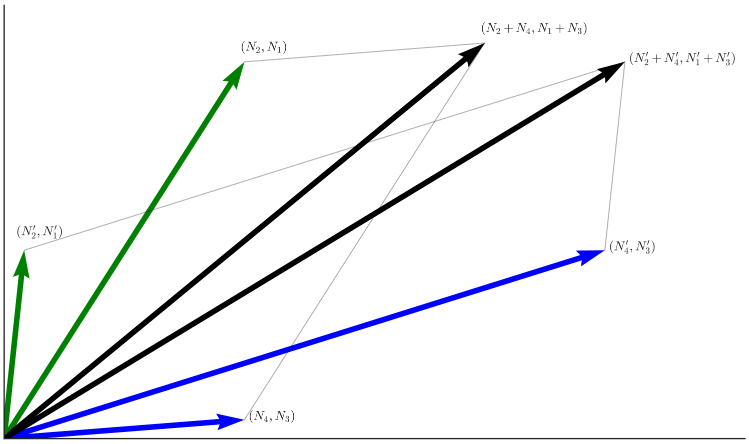


Figure 2.10: Simpson's reversal visualised. The ratios involving N_i' are steeper than those involving N_i for both the blue and green vectors. However, when we sum them, the ratio is steeper for the un-primed variables.

statistical viewpoint, there is no paradox. The reversal just stems from the mathematical property of ratios expressed in equation 2.16 and represented graphically in figure 2.10. 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.

$$\exists \left\{N_1, ...N_4, N_1'...N_4'\right\} \in \mathbb{N} : \frac{N_1}{N_2} < \frac{N_1'}{N_2'}, \ \frac{N_3}{N_4} < \frac{N_3'}{N_4'} \text{ and } \frac{N_1+N_3}{N_2+N_4} > \frac{N_1'+N_3'}{N_2'+N_4'} \tag{2.16}$$

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 as we discuss in the next section.

2.2 Answering Causal Questions

We can roughly categorise the problems studied within causal inference from observational data 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 [29] and breastfeeding on intelligence [88] to the effect of pet ownership on survival after a heart attack [59]. 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, the assumptions required to do so, and how to best leverage those assumptions to infer as much information as possible from the available data.

Causal discovery aims to leverage much broader assumptions to learn the structure of causal graphs from data. This is critical in fields where there is abundant data, but limited theoretical knowledge on how variables are related to one another. Causal discovery algorithms are being applied in bioinformatics [26, 141, 128, 7, 155, 62, 151, 160], medical imaging [129] and climate science [165]. An effective and generalisable approach for causal discovery would be a major step towards the automation of the scientific endeavour. In this thesis, I have focused on problems where the structure of the causal graph is known. Extending my work to problems where the causal structure is unknown, leveraging the work on causal discovery, is a rich and fascinating line of potential future work, which I discuss briefly in section 4.2.4.

2.2.1 Mapping from observational to interventional distributions

A central component of estimating causal effects from observational data is 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 section 2.1. In this section we describe a principled approach to mapping observational quantities to interventional ones and then, in section 2.2.3, discuss 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 ability 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 [119]. 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.

2.2.1.1 Independence in Bayesian networks: D-separation

Many causal algorithms are based on leveraging the independence properties encoded in Bayesian networks. Therefore, in this section, we briefly review the key properties of Bayesian networks. A more thorough introduction (including proofs) can be found in [97]. Recall that a Bayesian network is a way of representing the joint distribution over its variables in a way that 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-descendents 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 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 provides 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 2.11 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 2.11d is referred to as a collider or v-structure. The 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 disadvantaged 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.

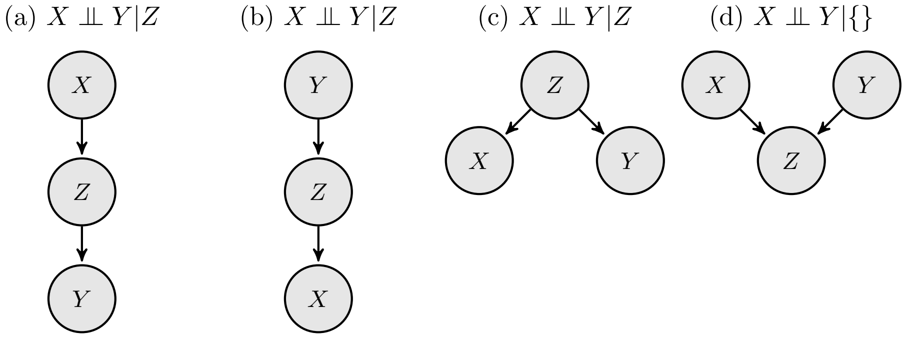

(a) $X \perp\!\!\!\perp Y|Z$
X
Z
Y
(b) $X \perp\!\!\!\perp Y|Z$
Y
Z
X
(c) $X \perp\!\!\!\perp Y|Z$
Z
X
Y
(d) $X \perp\!\!\!\perp Y|\{\}$
X
Y
Z

Figure 2.11: 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 2.11)

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. The presence of deterministic relationships between variables is one instance where this occurs. If there is a Bayesian network G with some deterministic nodes, then 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 the independencies that hold between counterfactuals via twin networks [119, 130] and demonstrates the caution required in using d-connectedness 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 [66].

2.2.1.2 The Do Calculus

The do-calculus is a set of three rules [118] 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 $G_{\overline{\boldsymbol{X}}}$:

$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}), \boldsymbol{Z}=\boldsymbol{z}, \boldsymbol{W}=\boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}), \boldsymbol{Z}=\boldsymbol{z}) \tag{2.17}$$

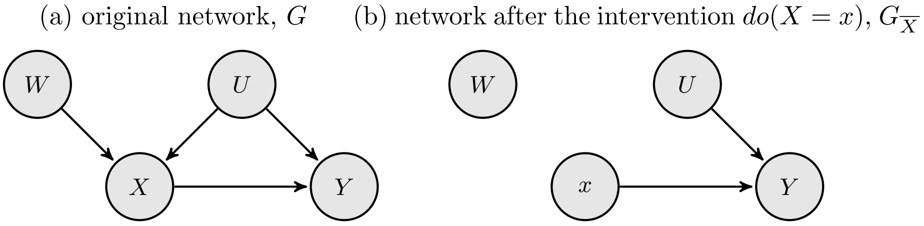

(a) original network, G (b) network after the intervention $do(X=x)$, $G_{\overline{X}}$

Figure 2.12: Rule 1 example. $(Y \perp\!\!\!\perp W|X)$ in $G_{\overline{\boldsymbol{X}}} \implies \mathrm{P}\left(Y|do(X), W\right) = \mathrm{P}\left(Y|do(X)\right)$

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 $G_{\underline{\boldsymbol{X}}}$:

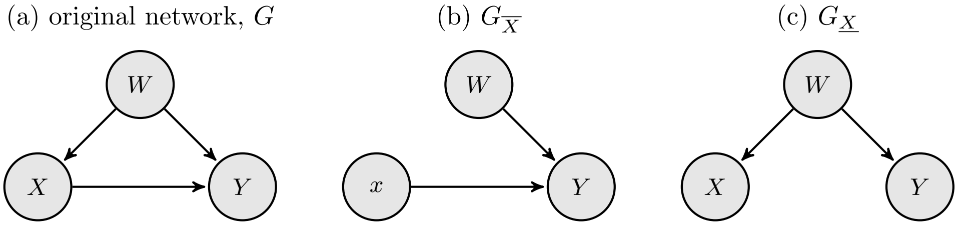
$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),\boldsymbol{W}) = P(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},\boldsymbol{W}) \tag{2.18}$$

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 therefore 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 2.18 does not cover cases where acting on one set of variables allows us to replace acting on another set with conditioning (see figure 2.14). The general form of rule 2 is given in equation 2.19.

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

$$P(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x}),do(\boldsymbol{Z}=\boldsymbol{z}),\boldsymbol{W}) = P(\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x},do(\boldsymbol{Z}=\boldsymbol{z}),\boldsymbol{W}) \tag{2.19}$$


(a) original network, G

(b) $G_{\overline{X}}$

(c) $G_{\underline{X}}$

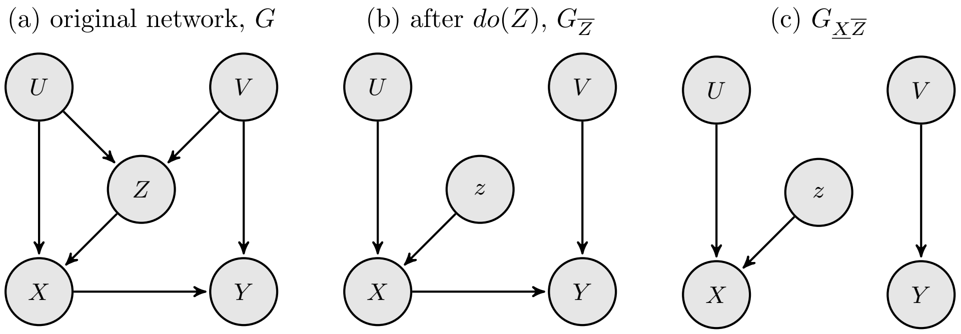
Figure 2.13: An example of rule 2 with a single intervention $(Y \perp\!\!\!\perp X|W)$ in $G_{\underline{\boldsymbol{X}}} \implies \mathrm{P}\left(Y|do(X),W\right) = \mathrm{P}\left(Y|X,W\right)$. In this example, observing X provides information about Y both directly and indirectly, because knowing X tells us something about W, which also influences Y. If we condition on W, we block this indirect path.

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 2.20. 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 $G_{\overline{\boldsymbol{X}}}$:

$$\mathrm{P}\left(\boldsymbol{Y}|do(\boldsymbol{X}=\boldsymbol{x})\right) = \mathrm{P}\left(\boldsymbol{Y}\right) \tag{2.20}$$

The general case of rule 3 is easier to state by explicitly representing the intervention in



(a) original network, G
(b) after $do(Z)$, $G_{\overline{Z}}$
(c) $G_{\underline{X}\overline{Z}}$
U
V
Z
X
Y
U
V
z
X
Y
U
V
z
X
Y

Figure 2.14: An example of applying equation 2.19. In this case $(Y \perp\!\!\!\perp X|Z)$ in $G_{\underline{X}\overline{Z}} \implies \mathrm{P}\left(Y|do(X=x), do(Z=z)\right) = \mathrm{P}\left(Y|X=x, do(Z=z)\right)$. 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 $X - U - Z - V - Y$.

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 2.15b). 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 [97].

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

$$P(\boldsymbol{Y}|do(\boldsymbol{Z}=\boldsymbol{z}), do(\boldsymbol{X}=\boldsymbol{x}), \boldsymbol{W}=\boldsymbol{w}) = P(\boldsymbol{Y}|do(\boldsymbol{Z}=\boldsymbol{z}), \boldsymbol{W}=\boldsymbol{w}) \tag{2.21}$$

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 affect $\boldsymbol{Y}$.

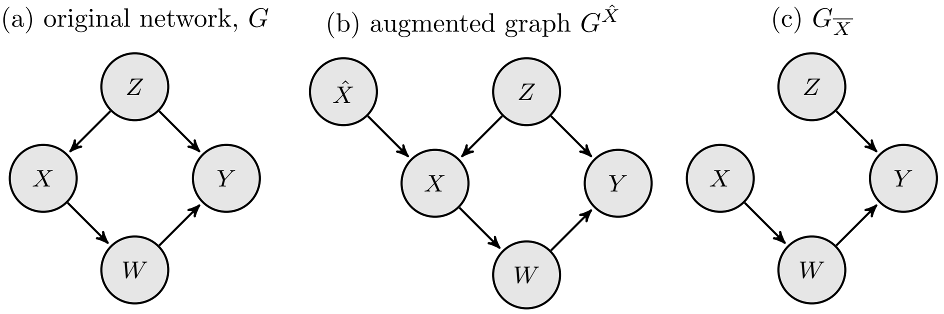


Figure 2.15: Example application of equation 2.21. $(Y \perp\!\!\!\perp \hat{X}|W, Z) \implies \mathrm{P}\left(Y|do(X), W, Z\right) = \mathrm{P}\left(Y|W, Z\right)$. We have to condition on Z because conditioning on W blocks the path $\hat{X} - X - W - Y$ but opens $\hat{X} - X - Z - Y$.

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

The question of non-parametric identifiability is solved. The do calculus is complete [146, 82]. A problem is identifiable if and only if the interventional distribution of interest can be transformed into terms containing only observational quantities via repeated application of the do-calculus. There is a polynomial time algorithm [145] 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, it can determine if the query is identifiable given the assumptions encoded by the network, and
2. if it is identifiable, return the required expression.

Figure 2.16 shows some examples of identifiable and non-identifiable queries. I have created a javascript implementation of the identifiability algorithm [145] on which you can test your own queries http://finnhacks42.github.io.

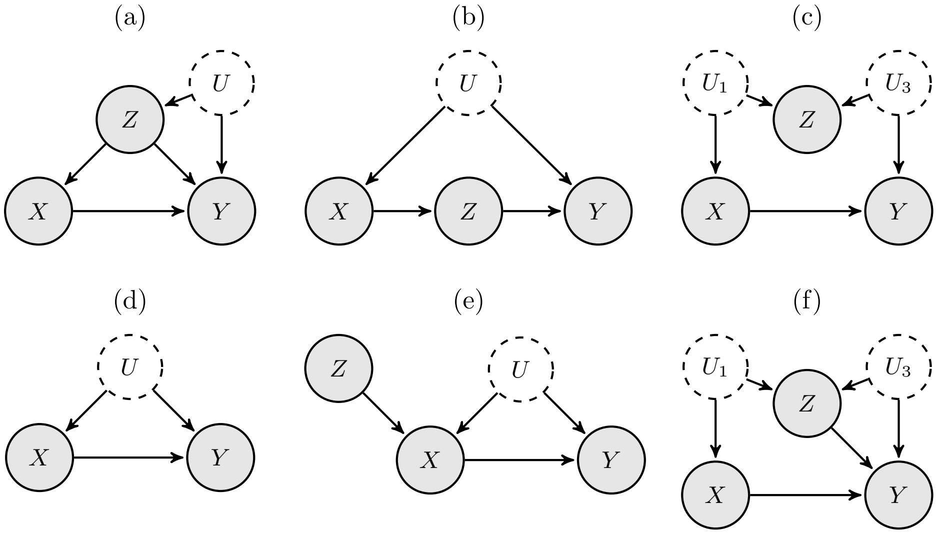


Figure 2.16: Examples of identifiable and non-identifiable queries. In sub figures (a), (b) and (c) the causal query $\mathrm{P}(Y|do(X))$ is identifiable. In sub figures (d), (e) and (f) it is not.

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 that 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 [43, 162, 50].

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.

2.2.2 Defining causal effects

So far we have described causal effect estimation in term of identifying the interventional distribution $\mathrm{P}\left(Y|do(X)\right)$ 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.

$$x* = \arg\max_{x} \mathbb{E}_{y \sim \mathrm{P}(Y|do(X=x))}\left[\mathcal{U}(y)\right] \tag{2.22}$$

Frequently however, studies wish to define and estimate a causal effect without reference to a specific utility function. There are several 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 [5] is defined as:

$$ACE = \mathbb{E}\left[Y|do(X=1)\right] - \mathbb{E}\left[Y|do(X=0)\right] \tag{2.23}$$

Assuming the expectations in equation 2.23 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.

$$ACE(x) = \frac{d}{dx}\mathbb{E}\left[Y|do(X=x)\right] \tag{2.24}$$

The average causal effect is often introduced as the average over individual causal effects as discussed in section 2.1.2. Individual causal effects are deterministic and cannot be expressed as properties of the interventional distribution. However, we can personalise

[5] Also referred to as the average treatment effect (ATE)

the average causal effect by stating it with respect to some observed context. I will refer to this as the personalised causal effect (PCE).[6]

$$PCE(z) = \mathbb{E}\left[Y|do(X=1), z\right] - \mathbb{E}\left[Y|do(X=0), z\right] \tag{2.25}$$

In some cases, the average causal effect for some subgroup of the population is of prime interest. 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.

$$ATT = \mathbb{E}_{z\sim \mathrm{P}(Z|x=1)}\left[Y|do(X=1)\right] - \mathbb{E}_{z\sim \mathrm{P}(Z|x=1)}\left[Y|do(X=0)\right] \tag{2.26}$$

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.

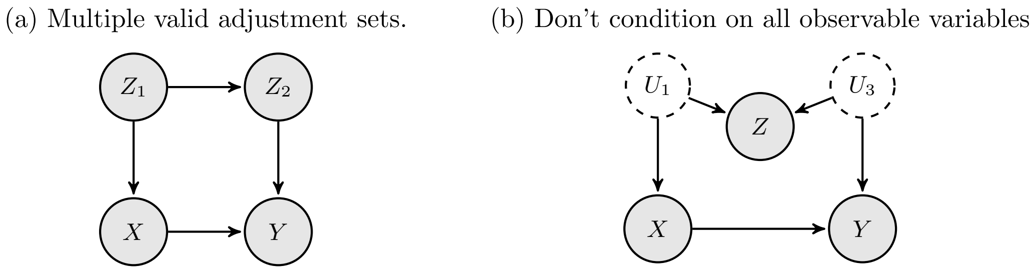
Another way of conceptualising causal effects is as a property indicating the strength of the causal link between two variables. This notion is complex to formalise when the relationship between variables is non-linear. Suppose $Y = X \oplus Z$ with $P(Z=1) = \frac{1}{2}$, the interventional distributions over X are identical after marginalising out Z. Janzing et al. [90] 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.

2.2.3 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 2.16e, which is not identifiable without parametric assumptions, however they can provide tight bounds. Adjusting for confounding equates to identifying a set of variables $\boldsymbol{Z}$ such that the ignorability assumption discussed in section 2.1.2 holds. This corresponds to a simple graphical test known as the backdoor criterion [119]. The setting is also referred to as unconfounded.

Theorem 11 (The backdoor criterion)**.** *[119] 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*

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


(a) Multiple valid adjustment sets.

(b) Don't condition on all observable variables

Figure 2.17: Identifying an optimal adjustment set is not always intuitively obvious. There may be multiple valid adjustment sets. In sub-figure (a) Z_1 or Z_2 or $\{Z_1, Z_2\}$ all block the backdoor path from X to Y. Adding additional variables (even pre-treatment variables) to a valid adjustment set can result in an invalid set as in sub-figure (b). In this case the empty set is a valid adjustment set for the causal effect of X on Y. However adding Z would open a backdoor path $X - U_1 - Z - U_2 - Y$

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;

$$\mathrm{P}\left(\boldsymbol{y}|do(\boldsymbol{x})\right) = \sum_{\boldsymbol{z}} \mathrm{P}\left(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{z}\right) \mathrm{P}\left(\boldsymbol{z}\right) \tag{2.27}$$

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 that satisfies the backdoor criterion. There may be more than one valid adjustment set, (figure 2.17a). 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 [120]. Adjusting for the wrong variables (even pre-treatment variables) can introduce or magnify bias, see figure 2.17b. 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. [142] on whether estimates of the causal effect of smoking on neonatal mortality should adjust for birth weight.

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 2.27. The expected value of Y after the intervention $do(X = x)$ is given by equation 2.28 and the average causal effect for a binary intervention $x \in \{0, 1\}$ is given by equation 2.29.

$$\mathbb{E}\left[Y|do(X = x)\right] = \mathbb{E}_{z\sim\mathrm{P}(\boldsymbol{Z})}\left[\mathbb{E}\left[Y|x, \boldsymbol{z}\right]\right] \tag{2.28}$$

$$ACE = \mathbb{E}_{z\sim\mathrm{P}(\boldsymbol{Z})}\left[\mathbb{E}\left[Y|1, \boldsymbol{z}\right] - \mathbb{E}\left[Y|0, \boldsymbol{z}\right]\right] \tag{2.29}$$

Assuming x and $\boldsymbol{z}$ are discrete, equation 2.28, 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 [85] 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 perspective.

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 $E_{out}(h)$,

$$E_{out}(h) = \mathbb{E}_{(\boldsymbol{x},y)\sim \mathrm{P}(\boldsymbol{x})\,\mathrm{P}(y|\boldsymbol{x})}\left[L(h(\boldsymbol{x}), y)\right] \tag{2.30}$$

We cannot directly compute the generalisation error as $\mathrm{P}(\boldsymbol{x}, y)$ is unknown, as 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.

$$E_{in}(h) = \frac{1}{n}\sum_{i=1}^{n} L(h(\boldsymbol{x}_i), y_i) \tag{2.31}$$

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;

$$\begin{aligned} E_{out} &= \mathbb{E}_{(\boldsymbol{x},\boldsymbol{z},y)\sim\delta(\boldsymbol{x}-\boldsymbol{x'})\,\mathrm{P}(\boldsymbol{z})\,\mathrm{P}(y|\boldsymbol{x},\boldsymbol{z})}\left[L_2(h(\boldsymbol{x}, \boldsymbol{z}), y)\right], &(2.32)\\ &= \mathbb{E}_{(\boldsymbol{z},y)\sim \mathrm{P}(\boldsymbol{z})\,\mathrm{P}(y|\boldsymbol{x},\boldsymbol{z})}\left[L_2(h(\boldsymbol{x}, \boldsymbol{z}), y)\right], &(2.33)\end{aligned}$$

Johansson et al. [91] 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

$\mathrm{P}_{train}\left\{\boldsymbol{v}\right\} = \mathrm{P}\left(\boldsymbol{z}\right)\mathrm{P}\left(\boldsymbol{x}|\boldsymbol{z}\right)$, but at test time the data will be sampled from $\mathrm{P}_{test}\left\{\boldsymbol{v}\right\}\mathrm{P}\left(y|\boldsymbol{v}\right)$, where $\mathrm{P}_{test}\left\{\boldsymbol{v}\right\} = \delta(\boldsymbol{x} - \boldsymbol{x'})\,\mathrm{P}\left(\boldsymbol{z}\right)$.[7] With this connection to covariate shift in mind, let us return to regression, matching and propensity scores.

2.2.3.1 Regression

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

$$\hat{f}_1(z) = h_1(z;\hat{\theta}_{obs}),\text{ where } \hat{\theta}_{obs} = \underset{\theta\in\Theta}{\arg\min}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i=1\}\,L\left(h_1(z_i;\theta),y_i\right)\right] \tag{2.34}$$

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.

$$\lim_{n\to\infty}\hat{\theta}_{obs} = \underset{\theta\in\Theta}{\arg\min}\,\mathbb{E}_{(z,y)\sim\mathrm{P}(z|x=1)\,\mathrm{P}(y|x=1,z)}\left[L\left(h_1(z;\theta),y\right)\right] \tag{2.35}$$

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 [157], including the interventional one.

$$\lim_{n\to\infty}\hat{\theta}_{obs} = \theta^* = \underset{\theta\in\Theta}{\arg\min}\,\mathbb{E}_{(z,y)\sim\mathrm{P}(z)\,\mathrm{P}(y|x=1,z)}\left[L\left(h_1(z;\theta),y\right)\right] \tag{2.36}$$

The average causal effect can then be estimated by:

$$\hat{\tau}_{reg} = \sum_{i=1}^{n}\left(\hat{f}_1(z_i) - \hat{f}_0(z_i)\right) \tag{2.37}$$

[7] 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 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.

2.2.3.2 Propensity scores

If the parametric model is missspecified 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 2.18.

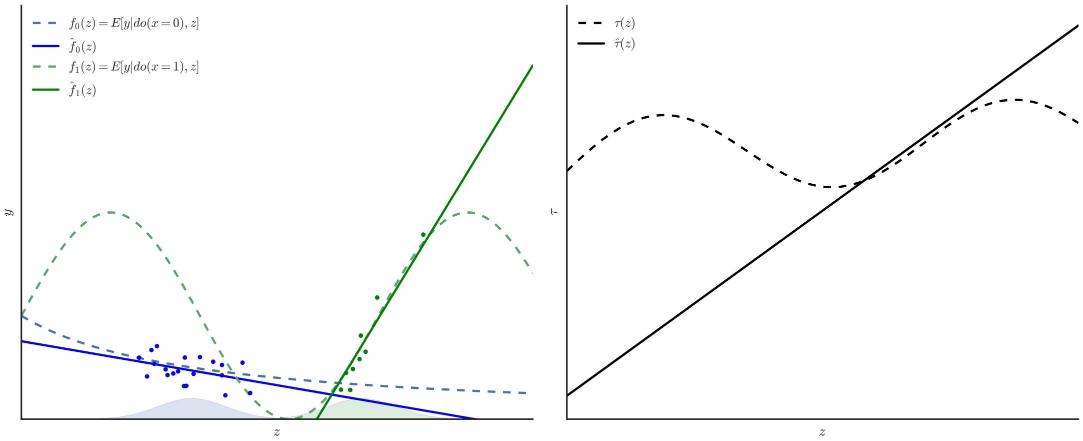


Figure 2.18: Parametric regression may yield poor estimates of causal effects if the model is missspecified, even if the model fits well over the domain of the training data. In this example, $\mathrm{P}(Z|X=0) \sim N(\mu_0, \sigma_0^2)$ and $\mathrm{P}(Z|X=1) \sim N(\mu_1, \sigma_1^2)$ with little overlap 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.

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 [80, 97].

$$\mathbb{E}_{\boldsymbol{v} \sim \mathrm{P}(\boldsymbol{v})}[f(v)] = \mathbb{E}_{\boldsymbol{v} \sim \mathrm{Q}(\boldsymbol{v})}\left[f(v)\frac{\mathrm{P}(\boldsymbol{v})}{\mathrm{Q}(\boldsymbol{v})}\right] \tag{2.38}$$

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 [157].

$$\hat{\theta}_{iw} = \underset{\theta \in \Theta}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = 1\} \, L\left(h_1(z_i;\theta), y_i\right) \frac{\mathrm{P}\left(z_i\right) \delta(x_i - 1)}{\mathrm{P}\left(z_i\right) \mathrm{P}\left(x_i = 1 | z_i\right)} \right] \tag{2.39}$$

$$= \underset{\theta \in \Theta}{\arg\min} \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x_i = 1\} \, L\left(h_1(z_i;\theta), y_i\right) \frac{1}{e(z_i)} \right], \tag{2.40}$$

where $e(\boldsymbol{z})$ is the propensity score, defined by [135];

$$e(\boldsymbol{z}) \equiv \mathrm{P}\left(x = 1 | \boldsymbol{z}\right) \tag{2.41}$$

The estimator in equation 2.39 is an example of a doubly robust estimator [134, 93]. Doubly robust methods are asymptotically unbiased so long as either the regression model h or propensity score e are correctly specified [133].

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 [135] 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 2.43 [85]. In some settings, such as stratified randomised trials [86] or learning from logged bandit feedback [30], the propensity score may be known. However, in general, it must be estimated from data. Frequently this is undertaken 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 [20]. Lunceford et al. [109] review the theoretical properties of key propensity score based estimators, including stratification and inverse propensity weighting.

$$\mathbb{E}\left[Y|do(X = x)\right] = \mathbb{E}_{z \sim \mathrm{P}(\boldsymbol{Z})}\left[\mathbb{E}\left[Y|x, \boldsymbol{z}\right]\right] = \mathbb{E}_{z \sim \mathrm{P}(\boldsymbol{Z}|\boldsymbol{x})}\left[\mathbb{E}\left[Y|x, \boldsymbol{z}\right] \frac{1}{e(\boldsymbol{z})}\right] \tag{2.42}$$

$$\hat{\tau}_{ip} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\mathbb{1}\{x_i = 1\} \, y_i}{e(\boldsymbol{z}_i)} - \frac{\mathbb{1}\{x_i = 0\} \, y_i}{1 - e(\boldsymbol{z}_i)} \right) \tag{2.43}$$

2.2.3.3 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 $\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i=1\}\,\hat{f}_1(z_i)=\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{x_i=1\}\,y_i$ [8], and equation 2.37 can be re-written as:

$$\hat{\tau}_{reg}=\sum_{i=1}^{n}\Big[\mathbb{1}\{x_i=1\}\left(y_i-\hat{f}_0(z_i)\right)+\mathbb{1}\{x_i=0\}\left(\hat{f}_1(z_i)-y_i\right)\Big] \tag{2.44}$$

This formulation of the regression estimator highlights the missing data aspect of causal 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 Imbiens [2] 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 $x_i \neq x_j$.

$$\hat{\tau}_{match}=\sum_{i=1}^{n}\left[\mathbb{1}\{x_i=1\}\left(y_i-\frac{1}{k}\sum_{j\in J_k(i)}y_j\right)+\mathbb{1}\{x_i=0\}\left(\frac{1}{k}\sum_{j\in J_k(i)}y_j-y_i\right)\right] \tag{2.45}$$

This estimator is equivalent to equation 2.44 with k nearest neighbour regression. There are many variants of matching estimators using 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 [44, 136]. Although intuitive, matching estimators in general have poor large sample properties [1]. 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) [85].

The practical performance of the estimation approaches discussed in this section 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 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 a theoretical 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. [157] propose inverse propensity weighted cross validation for the covariate shift problem. There is relatively little theory

[8] [85] state this holds for most implementations

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 [174, 28, 121, 114]. 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.

The performance of methods for causal effect estimation can be tested on simulated data [60, 179, 79, 49] or by comparing estimates from observational studies with the results from corresponding experiments [99, 58, 78, 77, 47, 150, 11]. 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.

2.3 Bayesian reasoning and causality

The Bayesian approach to modelling encapsulates a very general approach to combining prior knowledge and assumptions with data to draw inferences. You can specify a complex model that captures your assumptions about the underlying mechanisms of your system, describe uncertainty about the parameters (and priors over their values), and then update your beliefs about the values of these parameters based on the observed data. If you believe the model to be a mechanistic representation of the way the world works then it can be used to infer the consequences of interventions.

Why then do we need a special framework for causal inference? Why not encode all the assumptions required to infer the outcome of an intervention in a given system in a Bayesian model? The short answer is we could. However, causal graphical models represent a useful level of abstraction for many problems. 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. We do not need to specify the functional relationships between the variables or priors over their distributions. The power of defining an intervention in this way stems from the number of things that are assumed 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. The causal framework also provides, via the do-calculus, a rigorous method to draw a boundary around which variables we must measure and model to identify causal effects.

We can represent problems where the goal is to infer properties of the post-interventional system based on data generated by the pre-interventional distribution by explicitly representing both systems and their common features (figure 2.19). This does not require any special framework or notation. The graphs in figure 2.19 are ordinary Bayesian networks. However, without a causal framework, we have to make assumptions about what will be invariant to the intervention for each such problem encountered. For complex problems, it is very difficult to conceptualise the assumptions we expect to hold without the benefit of a causal framework.

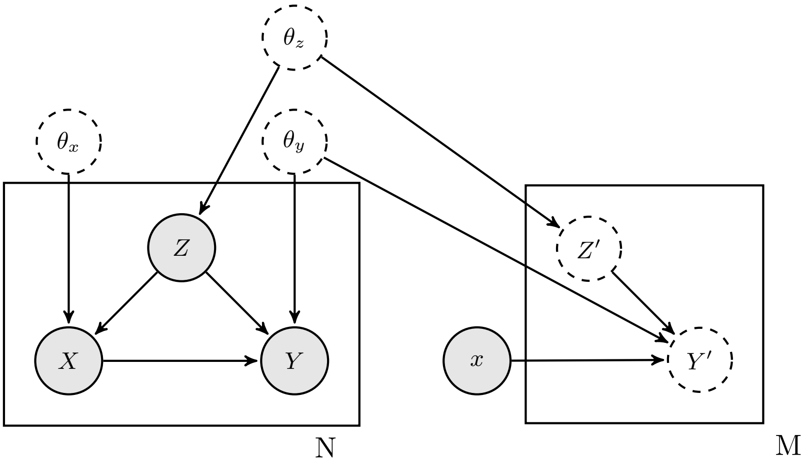


Figure 2.19: Causal inference with ordinary Bayesian networks. The plate on the left represents the observed data generated prior to the intervention and the plate on the right the data we anticipate obtaining after an intervention that sets the pre-interventional variable X to x. The assumptions characterised 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, this would then entail a complex transformation relating the parameters between the two plates, rather than a simple invariance.

Causal graphical models can also be combined with Bayesian approaches to estimation. For example, we can use a causal graphical model to identify a set of variables that form a valid adjustment set and then use Bayesian regression to estimate posterior distribution over a given causal effect. However, as I highlight in the next section, the way priors are specified is critical.

2.3.1 Be careful with that prior

Suppose data is generated by the linear Gaussian Bayesian network below. The goal is to estimate the causal effect of X on Y, that is identify the value of the coefficient w_{yx}. The variable U is latent.

$$\mathrm{P}(U) = N(0, v_u) \tag{2.46}$$

$$\mathrm{P}(Z|U) = N(w_{zu}U, v_z) \tag{2.47}$$

$$\mathrm{P}(X|Z) = N(w_{xz}Z, v_x) \tag{2.48}$$

$$\mathrm{P}(Y|U, Z, X) = N(w_{yu}U + w_{yz}Z + w_{yx}X, v_y) \tag{2.49}$$

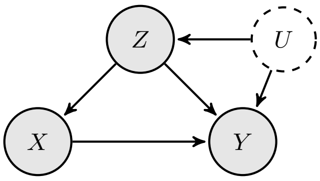


Figure 2.20

As each variable is a linear function of its parents, with Gaussian noise, resulting in a joint distribution $\mathrm{P}(U, Z, X, Y)$ that is multivariate normal. Marginalising out U and conditioning on X and Z yields,

$$Y \sim N(w_{yx}X + \beta Z, \varepsilon) \tag{2.50}$$

where,

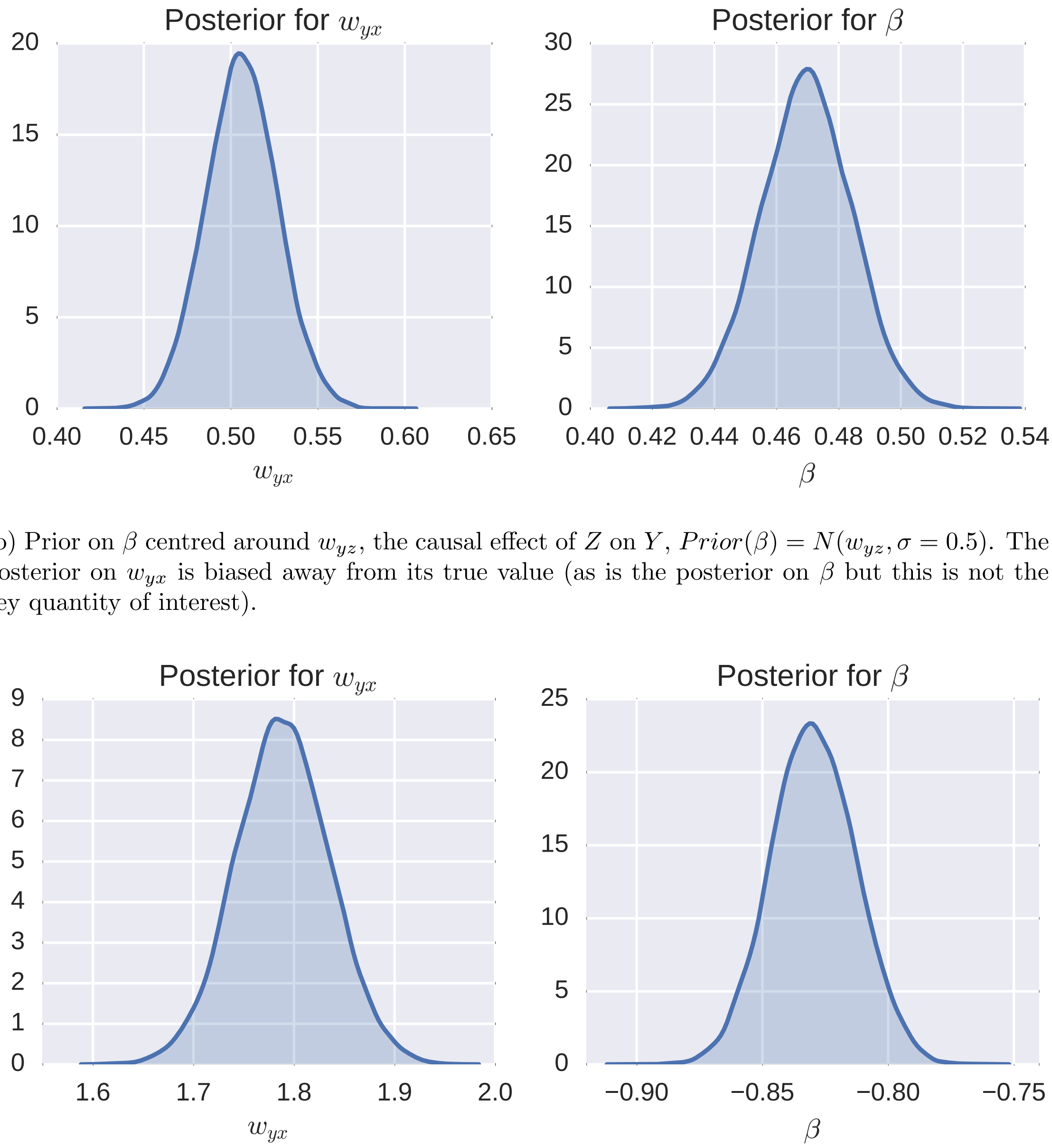
$$\beta = w_{yz} + \frac{w_{yu}w_{zu}}{w_{zu}^2 + \frac{v_z}{v_u}}, \text{ and,} \tag{2.51}$$

$$\varepsilon = v_y + w_{yu}^2 v_u - \frac{v_u^2 w_{zu}^2 w_{yu}^2}{v_z^2 + v_u^2 w_{zu}^2} \tag{2.52}$$

The causal effect of X on Y is identifiable (even without the linear Gaussian assumptions) as Z satisfies the backdoor criterion and the expectation of the coefficient for X in a standard OLS regression of Y against X and Z is w_{yx}. However, the causal effect of Z on Y is not identifiable due the presence of the unobserved confounder U. The coefficient β captures both the causal relationship between Z and Y, w_{yz} and the indirect relationship through U, which we cannot separate without observing U.

Undertaking a Bayesian regression of X and Z against Y and using intuition based on the causal relationship between Z and Y to select a prior for β will lead to selection of something centred around w_{yz}. However, this is incorrect and can introduce bias into our estimate of the coefficient for w_{yx}, as demonstrated in figure 2.21. The prior for β should be based on the belief about how Z is expected to be *associated* with Y, after marginalising out any confounding variables.

This result is counter intuitive. When fitting a model to estimate the effect of marketing on sales, it is natural to put a prior on the coefficient for price that is strongly negative because we believe increasing price decreases sales. However, our prior should be on the expected association between price and sales, holding fixed all the other variables in the model, not the causal relationship between price and sales. An alternative is to explicitly include U as a latent variable in the model. However, this comes at a computational cost with little benefit for estimation, unless we have good knowledge of the relationships between U, Z and Y. More fundamentally, at whatever level we decide to stop adding latent variables and specify a prior, we can introduce this form of bias if priors are selected naively on the basis of causal intuition. The issue is not limited to linear-Gaussian models. This is an important insight for the many applications of Bayesian modelling that make use of human-elicited priors, since people tend to think causally. The solutions, either explicitly modelling latent variables to higher levels or allowing for them in setting priors, have the effect of broadening the priors over nuisance parameters in the model.



(a) No prior on β, the posterior on w_{yx} is centred around its true value

(b) Prior on β centred around w_{yz}, the causal effect of Z on Y, $Prior(\beta) = N(w_{yz}, \sigma = 0.5)$. The posterior on w_{yx} is biased away from its true value (as is the posterior on β but this is not the key quantity of interest).

Figure 2.21: An example demonstrating how selecting a prior for a nuisance parameter centred around the causal effect of that parameter on the outcome, rather than its association with the outcome, can bias estimates for actual parameter of interest. We sample $n = 1000$ data points from the joint distribution defined by $P(U)P(Z|U)P(X|Z)P(Y|U,Z,X) \sim N(0,1)N(2U,0.09)N(0.5Z,1)N(3U - Z + 0.5X, 0.25)$ and fit the model $Y \sim N(w_yxX + \beta Z, \varepsilon)$ in Stan and plot the posterior distributions over w_yx and β. Figure (a) shows the results with no prior (equivalently an improper prior) for both w_yx and β. Figure (b) shows the results when we place a Gaussian prior centred around the w_yz on the distribution for β. All models were fit using Stan [38]

Chapter 3

Learning from interventions

The previous sections focused on aspects of the problem of estimating the likely effect of an intervention from data gathered prior to making the intervention. 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. First, we determine the nature of the intervention and thereby control the data points used to estimate causal effects. Selecting data points optimally for learning is the focus of the optimal experimental design literature within statistics [127] and the active learning literature in machine learning [144]. 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 [158].

Reinforcement learning describes the problem of an agent interacting with an environment, learning by observing the outcome of its actions, with the goal of maximising some reward. These problems, which also incorporate planning, are extremely difficult because the value of an action is generally not immediately clear. The state of the environment may evolve over time and according to previously selected actions. Actions available at future time steps can depend on those taken in the past, and rewards may be obtained only after a long sequence of actions. This makes it extremely difficult for the agent to accurately attribute value to each chosen action along the path to obtaining a given reward.

We focus on a simpler class of problems within reinforcement learning known as multi-armed bandit problems. Multi-armed bandit problems also describe an agent aiming to maximise some reward by interacting with an environment. However, the reward is observed immediately after the action is selected and the environment is stateless. Future rewards generated by the environment in response to the agent's actions do not depend

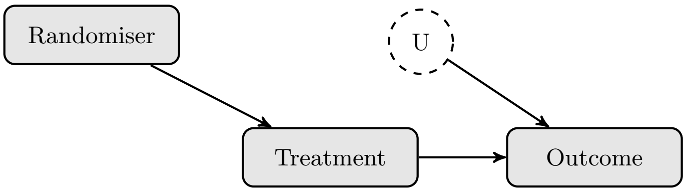


Figure 3.1: causal network for a randomised experiment

on the previous actions of the agent. These assumptions are a reasonable approximation in settings where we face the same decision repeatedly with respect to many independent units or individuals and we have a good proxy for the outcome. In this chapter we describe the interventionalist approaches to learning to act for these types of problems, beginning with classical randomised experiments and then the, more adaptive, multi-armed bandit algorithms. We show how both observational causal inference and bandit algorithms can be viewed as generalisations of randomised experiments and highlight connections between the observational and interventional approaches to causal inference.

3.1 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 3.1. 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 affect both treatment and outcome.

Randomisation does not ensure target and control group are exactly alike. The more other 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 3.1 still holds. We can then 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, (figure 3.2). This necessitates that one condition on these covariates in computing the average causal effect in the same

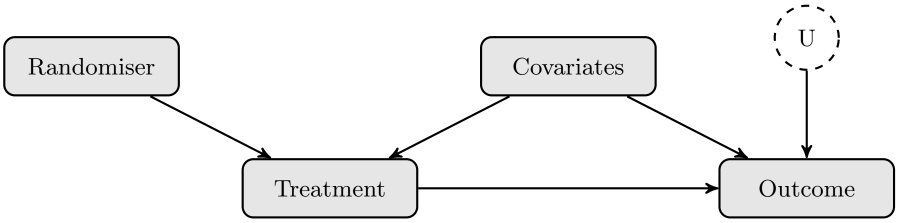


Figure 3.2: causal network for a stratified randomised experiment if the probability an individual is assigned a given treatment depends on some covariates.

way as for estimating causal effects under ignorability (§2.2.3). 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 [86] 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 [119] (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.

3.1.1 Limitations of randomised experiments

The idealised notion of an experiment represented by figure 3.1 does not capture the complexities of randomised experiments in practice. There may be imperfect compliance so that the treatment selected by the randomiser is not always followed, or output censoring in which 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 affect the outcome, then the problem of confounding bias returns.[1] See figure 3.3 for a graphical model of a randomised experiment with imperfect compliance.

It is not always possible or ethical to conduct a randomised controlled trial, as is beautifully demonstrated by the paper of Smith and Pell [149] on randomised cross-over trials of parachute use for the reduction of the mortality and morbidity associated with falls from large heights (figure 3.4). 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 that can be explored. In addition, they are often conducted on a convenient, but unrepresentative, sample of the broader population of interest (for example first year university students). This can result in estimates with

[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 to take the treatment they have been prescribed is likely to be the same if the treatment were made available more generally beyond the experimental trial.

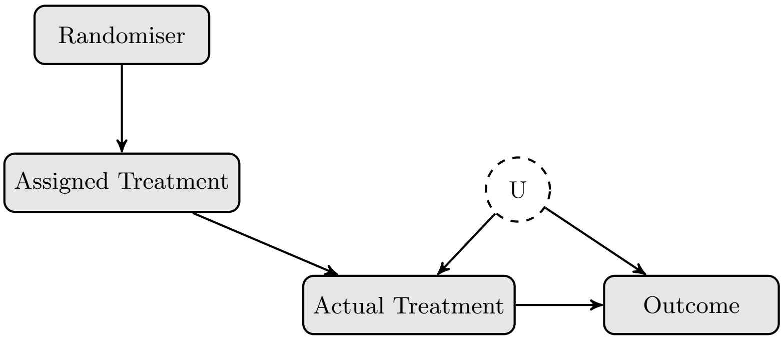


Figure 3.3: causal network for a randomised experiment with imperfect compliance

I hope I'm not in the control group
It doesn't matter. Next time we swap

Figure 3.4: 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.

high *internal validity* [36], in that they should replicate well in a similar population, but low *external validity* in that 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 [24] 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.

3.2 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 [161]. In its classic formulation [132, 98] 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 timestep 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 12 (Stochastic k-armed bandit problem)**.** Let $\mathcal{A} = \{1, ..., k\}$ be the set of available actions (or bandit arms) and $\mathrm{P}(\boldsymbol{y}) = \mathrm{P}\left(Y^1, ..., Y^k\right)$ 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 timesteps 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^*. Note we have used counterfactual notation (see section 2.1.2) to denote the rewards for each action, Y_t^i is the reward the algorithm would have received had it selected action i at timestep t. I discuss the (potentially) counterfactual nature of regret further in section 3.3.

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 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 timestep. This difference is known as the (cumulative) regret.[3]

$$R_T = \sum_{t=1}^{T} Y_t^{i^*} - \sum_{t=1}^{T} Y_t^{a_t} \tag{3.1}$$

[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, researchers often refer to the expected regret as "the regret".

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 [17, 15, 14]. The expectation of the regret, as defined by equation 3.1, is referred to as the pseudo-regret [31] and is given by equation 3.2. A stochastic bandit algorithm is learning if it obtains pseudo-regret that is sub-linear in T.

Definition 13 (Pseudo-Regret)**.**

$$\bar{R}_T(\pi) = \max_{i \in \{1,...,k\}} \mathbb{E}\left[\sum_{t=1}^{T} Y_t^i\right] - \mathbb{E}\left[\sum_{t=1}^{T} Y_t^{a_t}\right] \tag{3.2}$$

$$= T\mu_{i^*} - \mathbb{E}\left[\sum_{t=1}^{T} Y_t^{a_t}\right] \tag{3.3}$$

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).

3.2.1 Stochastic bandits: Approaches and results

The adaptive nature of multi-armed bandit algorithms complicates the design and analysis 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 3.5 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 [63].

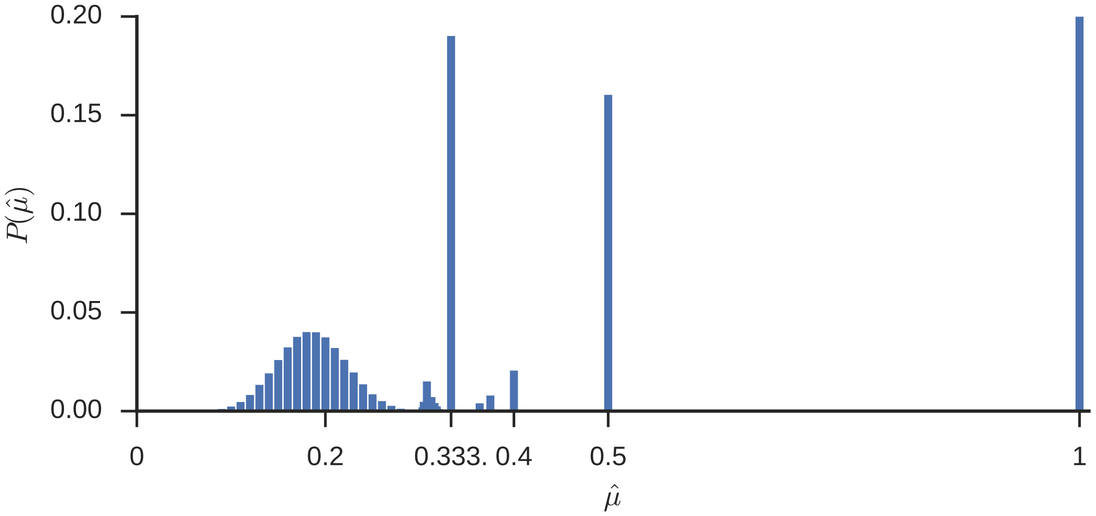


Figure 3.5: Standard empirical estimators can be biased if the number of samples, n, is not fixed in advance, but is a random variable that depends on the values of previous samples. This example plots the distribution (over 10^6 simulations) of $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim Bernoulli(0.2)$. 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. $\mathbb{E}\left[\hat{\mu}\right] = 0.439$. The estimator is substantially biased above $\mathbb{E}\left[X_i\right] = 0.2$ by the early stopping. Note that excluding experiments that were stopped early creates a bias in the opposite direction, $\mathbb{E}\left[\hat{\mu}|n = 100\right] = 0.185$, as trials that obtained positive results early are excluded. This has some interesting 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.

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 [98] 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. [16] developed the UCB-1 algorithm, see algorithm 1, which requires 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 sub-gaussian reward distributions, see Bubeck et al. [31].

Algorithm 1 UCB-1

```
**Input:** horizon $T$.
Play each arm once.
**for** $t \in 1, \dots, T$ **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 $a_t \in \arg\max_{i=\{1,\dots,k\}} \left(\hat{\mu}_{t,i} + \sqrt{\frac{2 \log t}{n_{t,i}}}\right)$
```

Let $\Delta_i = \mu_i - \mu^*$ be the degree to which each arm is sub-optimal. The problem-dependent pseudo-regret for UCB-1 is bounded by equation 3.4 [31],

$$\bar{R}_T \leq \sum_{i:\Delta_i > 0} \left(\frac{8 \log(T)}{\Delta_i} + 2 \right) \tag{3.4}$$

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 equation 3.4 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 3.4 and $T\Delta$ which is maximised when they are equal, see figure 3.6. By solving this equality for Δ one can show the worst case regret is bounded by equation 3.5, see Bubeck et al. [31].

$$\bar{R}_T \in \mathcal{O}\left(\sqrt{kT \log(T)}\right) \tag{3.5}$$

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. [17] show that the worst case regret for the k-armed bandit problem is lower bounded by $\bar{R}_T \in \Omega\left(\sqrt{kT}\right)$

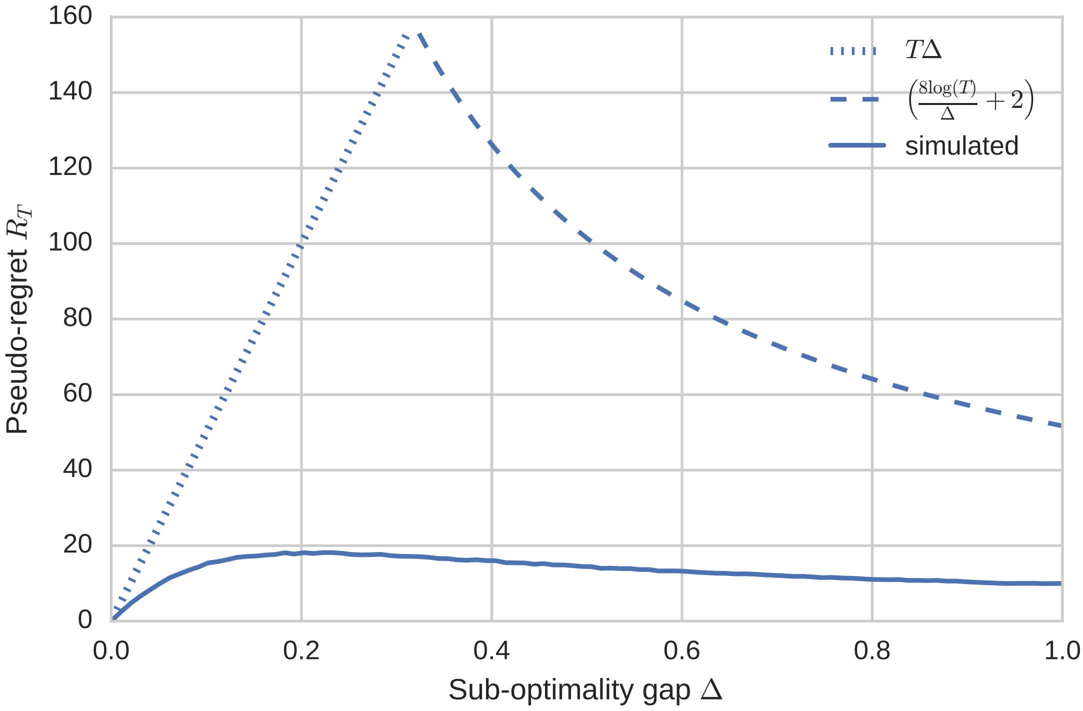


Figure 3.6: The regret bound in equation 3.4 grows as the differences between the expected rewards for each arm shrink. The solid curve shows the mean (cumulative) regret for the UCB-1 algorithm, over a 1000 simulations for a 2-armed, Bernoulli bandit with fixed horizon, $T = 500$, as a function of the difference in the expected reward for the arms Δ. The dashed curves show the corresponding upper bounds; $T\Delta$ and equation 3.4

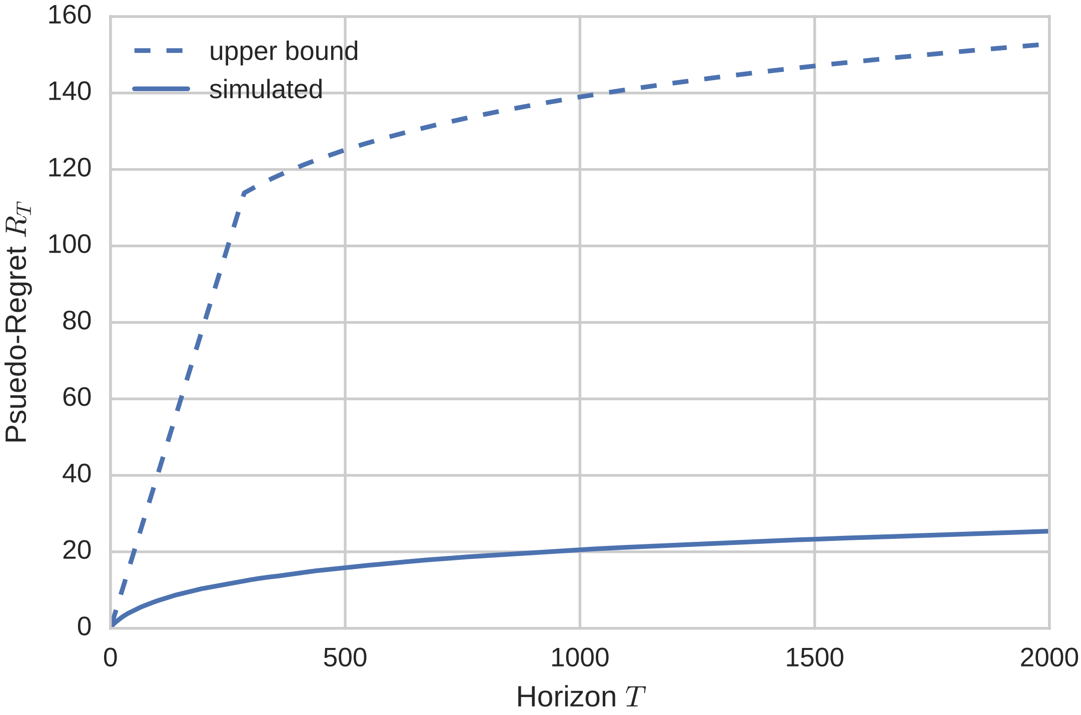


Figure 3.7: The actual performance of the UCB algorithm can be substantially better than suggested by the upper bound, particularly for small T. The solid curve shows the mean expected regret associated with the sequence of arms chosen by UCB-1 with $k = 2$ arms and the rewards sampled from $Bernoulli([.3, .7])$ over 1000 simulations. The dashed curve shows the corresponding upper bound given by the minimum of $T\Delta_{max}$ and equation3.4.

Subtle modifications to the UCB algorithm can eliminate the logarithmic term in equation 3.5. This yields regret $\mathcal{O}\left(\sqrt{TK}\right)$ and closes the gap with the worst case lower bound [12, 103], whilst retaining a good problem-dependent bound of the form achieved by UCB [103].

Finally, there is the heuristic principle of playing each arm with probability proportional to the likelihood that it is optimal. This approach is generally called Thompson sampling as it was the method proposed in the original bandit paper by Thompson [161]. Thompson sampling has strong empirical performance, [42]. However, it is complex to analyse. Kaufmann et al. [95] 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.

3.2.2 Pure-exploration problems

Another problem that has attracted recent attention [32, 13, 61, 94] 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 14 (Simple Regret)**.**

$$R_T = \mu_{i^*} - \mathbb{E}\left[\mu_{\hat{i}^*}\right]. \tag{3.6}$$

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 3.7 ([32]).

$$R_T \in \mathcal{O}\left(\sqrt{K/T}\right) \tag{3.7}$$

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. This means we should focus exploration on arms which are plausibly optimal, creating a form of exploration-exploitation trade-off, albeit subtlety different to that for the cumulative regret.

3.2.3 Adversarial Bandits

Adversarial bandits, described by Auer et al. [17], 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 15 (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 timesteps 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 timesteps 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 3.8 because the sequence of rewards is fixed. However the reward obtained by the algorithm is still a random variable as we are considering randomised algorithms.

$$\bar{R}_T(\pi) = \max_{i \in \{1,...,k\}} \sum_{t=1}^{T} Y_t^i - \mathbb{E}\left[\sum_{t=1}^{T} Y_t^{a_t}\right] \tag{3.8}$$

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]

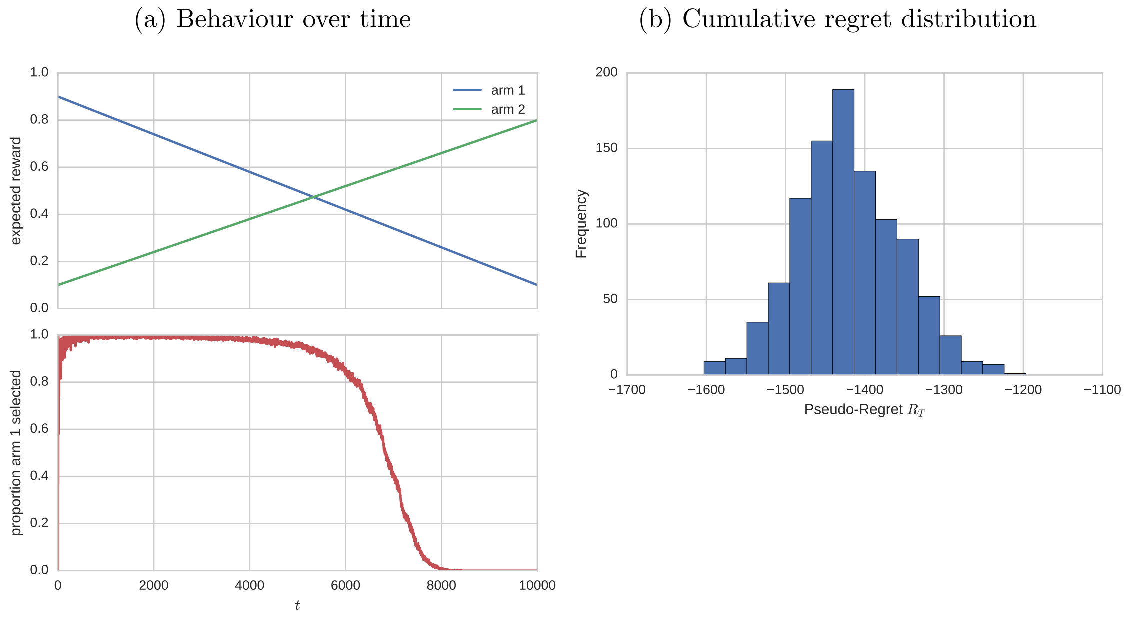
The seminal algorithm for adversarial bandits is Exp-3 [16], which, like UCB, obtains worst case pseudo-regret of $\mathcal{O}\left(\sqrt{TK\log(T)}\right)$ [17]. Optimal algorithms, with $\bar{R}_T = \mathcal{O}\left(\sqrt{TK}\right)$, have also been demonstrated for the oblivious adversarial setting [12]. 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, 19].

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 3.8, does not fully capture an algorithm's performance in such cases because it

[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.

[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).

is defined with respect to playing the single arm with the best average return over the game. In settings where the rewards change over time, the pseudo-regret can be negative (see figure 3.8) so upper bounds on the pseudo-regret do not fully reflect how sub-optimal the algorithm may be. An example of a setting that lies between stochastic and adversarial bandit problems is the non-stationary setting, in which the rewards are generated stochastically from a distribution that varies over time. Adversarial bandit algorithms may perform better in such settings than standard stochastic policies to the extent that they explore more (to avoid the adversary simulating their behaviour) and thus adapt quicker to changes in the reward distribution. Adversarial algorithms also have stronger worst case regret guarantees, since even the weak regret for stochastic bandits is not guaranteed to be sub-linear in such settings. However, if there are constraints on how rapidly or frequently the reward distributions can change over time, it is better to use algorithms specifically developed to exploit such information and compare them against a stronger notion of regret (see for example [65, 64, 27]).



(a) Behaviour over time

(b) Cumulative regret distribution

Figure 3.8: 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).

3.2.4 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, the optimal action depends on some context. The best treatment to offer an individual patient could depend on their age, gender, disease subtype 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 [101]. However, close variants of the underlying problem have also been posed under the names; "associative reinforcement learning" [92], "bandits with concomitant variables"[172] and "bandit problems with side information" [169].

Definition 16 (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 timesteps,
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.

$$\bar{R}_T = \max_{h \in \mathcal{H}} \mathbb{E}\left[\sum_{t=1}^{T} Y_t^{h(\boldsymbol{X}_t)}\right] - \mathbb{E}\left[\sum_{t=1}^{T} Y_t^{A_t}\right] \tag{3.9}$$

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 $\mathcal{O}\left(\sqrt{NkT}\right)$,

[6] Contextual bandits can also be defined in the adversarial setting analogously to definition 15

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 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; that is, 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 [101]. At each timestep, 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 the 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 [41]. 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 the component of the reward that 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, [101], 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 timesteps. The epsilon-greedy algorithm obtains worst

[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, $N = 2^n$

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 $\mathcal{O}\left(\sqrt{kTlog(|\mathcal{H}|)}\right)$ for both stochastic and adversarial contextual bandit problems [18]. However, it involves maintaining a list of weights for each hypothesis h, resulting in time and memory requirements 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,

$$\arg\max_{h\in\mathcal{H}} \sum_{t=1}^{\tau} Y_t \mathbb{1}\{h(\boldsymbol{X}_t) = A_t\} \tag{3.10}$$

This expression equates to identifying the empirically best 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 [51].

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* [106] and *Generalised Thompson Sampling* [6].

3.2.4.1 The causal structure of contextual bandits

The definition of contextual bandits (see definition 16) does not make any assumptions about the *causal* relationship between the context $\boldsymbol{X}$ and the reward $\boldsymbol{Y}$ (see figure 3.9). However, the context should be relevant, such that $\mathrm{P}(\boldsymbol{y}|\boldsymbol{x}) \neq \mathrm{P}(\boldsymbol{y})$, otherwise including it is equivalent to adding irrelevant features to a supervised learning problem. Bareinboim et al. [23] 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 3.9b. It is even possible to have (useful) context $\boldsymbol{X}$ that is a consequence of $\boldsymbol{Y}$ as in figure 3.9c and example 17.

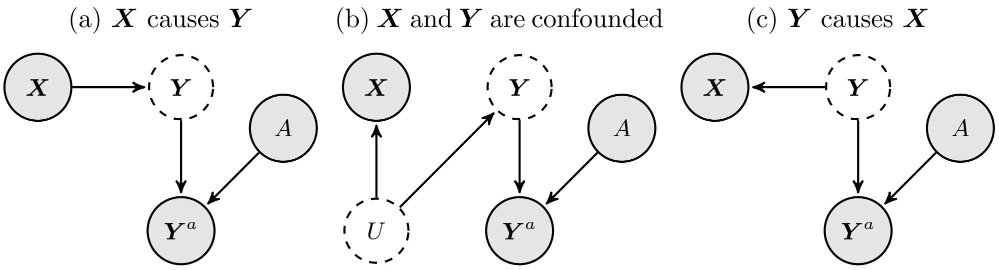

(a) $\boldsymbol{X}$ causes $\boldsymbol{Y}$
(b) $\boldsymbol{X}$ and $\boldsymbol{Y}$ are confounded
(c) $\boldsymbol{Y}$ causes $\boldsymbol{X}$
$\boldsymbol{X}$
$\boldsymbol{Y}$
A
$\boldsymbol{Y}^a$
U

Figure 3.9: 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 $\boldsymbol{Y}$. The causal structures in figures 3.9a and 3.9b also make sense if the elements of $\boldsymbol{Y}$ are genuine counterfactuals, that is only the reward of the selected action is ever realised. The context $\boldsymbol{X}$ can only be useful in causal graphs of the form in figure 3.9c if the components of $\boldsymbol{Y}$ are realised prior to the agent selecting an action, i.e when the components of $\boldsymbol{Y}$ are not truly counterfactual, it just that the agent is limited to only observing the outcome of the action they choose. 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 timestep t and allow the action A_t to depend on the context $\boldsymbol{X}_t$ and the previous observations $(\boldsymbol{X}, A, Y)_1^{t-1}$.

Example 17. Imagine a modified game of roulette. As in standard roulette, it consists of a spinning wheel with a ball that can land in one of 37 numbered pockets. However, unlike the standard game where punters place bets before the wheel is spun, in this variant the wheel is spun first. Five seconds before it comes to rest the face of the wheel darkens to hide the position of the ball from the players. They then place bets after it has stopped (and are each allowed only to check if the position they have bet on was correct). At the point in time the players are selecting their actions, the rewards for each possible choice are fully determined (all zero except for the pocket containing the ball). Suppose a canny gambler realises the ball makes a tiny sound, from rocking back and forth, if it is close to the left or right side of the wheel such that the pocket holding it is close to horizontal. The existence (or not) of this sound is a consequence of the position of the ball and thus the reward state and the gambler can use it as context to improve the likelihood of obtaining a reward.

It is difficult to come up with realistic examples in which the full sequence of rewards is generated, but only the one associated with the selected action is observable to the learner. In the typically cited applications of contextual bandit algorithms, such as selecting how to treat patients or which ad to use, only the reward of the selected action is ever realised, and the alternate rewards are counterfactual variables.

3.2.5 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 [100, 156, 107, 52, 30, 159]. 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 alternative 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 of the bandit nature of the feedback; only the reward of the selected action is observed. In addition, if π is allowed to depend on 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. [100] 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 policy 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, as we discussed in section 2.2.3. The causal structure can be represented in figure 3.10 There is an (implicit) assumption that all variables that affect 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 section 4.2. of Pearl [119]. 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,

$$\mathbb{E}\left[y|\pi'\right] = \mathbb{E}\left[y|do(a \sim \pi'(\boldsymbol{x}))\right] = \mathbb{E}_{(\boldsymbol{x},a)\sim \mathrm{P}(\boldsymbol{x})\,\mathrm{P}_{\pi'}\{a|\boldsymbol{x}\}}\left[y|\boldsymbol{x}, a\right] \tag{3.11}$$

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 Dudik et al. [52], more attractive.

Swaminathan and Joachims [159] point out that the problem of identifying the optimal policy (subject to some risk minimisation goal) is not as simple as estimating the expected

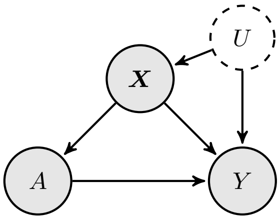


Figure 3.10: Causal graphical model for learning from logged feedback data under the assumption the original policy π for selecting actions was stationary and dependent only on some observed context $\boldsymbol{X}$, $a \sim \pi(\boldsymbol{x})$. The outcome Y may depend on $\boldsymbol{X}$ directly, indirectly through a latent variable U or both.

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.

3.2.6 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 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 [117], X-armed bandits [33] and Gaussian process bandits [154], or that consider more complex feedback, for example the recent work on graph feedback [110, 104, 9, 35, 96, 8] and partial monitoring [125, 25].

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 these previous approaches.

3.3 The counterfactual nature of regret

We conclude this chapter on bandits with a note on the counterfactual nature of regret, which I have not seen discussed with respect to bandit regret. However, the issues raised are very closely related to the work by Dawid [46] on problems associated with the use of counterfactual variables for (observational) causal inference.

It is possible, as we demonstrated in example 17, to construct bandit problems in which the full vector of rewards $[Y_t^1, ..., Y_t^k]$ is actually generated at each timestep (and, at least in principle, is observable to someone although not the agent). However, in most real applications for bandit problems, only the outcome associated with the selected action is ever realised and so the remaining components of the reward vector are counterfactual, justifying the use of the counterfactual notation to denote them. There are complex philosophical objections to counterfactuals arising from the way they describe alternate universes that were never realised. This makes it easy to make statements using counterfactuals that cannot be confirmed empirically (even with infinite experimental data). We now show with a simple example that the standard definition of regret is a fundamentally counterfactual quantity. Recall that the cumulative regret is defined by,

$$R_T = \sum_{t=1}^{T} Y_t^{i^*} - \sum_{t=1}^{T} Y_t^{a_t} \tag{3.12}$$

$$= \sum_{t=1}^{T} \left(Y_t^{i^*} - Y_t^{a_t} \right) \tag{3.13}$$

Take a stochastic, two-armed, Gaussian bandit with the joint distribution of the (counterfactual) rewards $\mathrm{P}\left(Y_t^1, Y_t^2\right)$ given by equation 3.14. Suppose without loss of generality that arm 1 is the optimal arm, such that $\mu_1 > \mu_2$.

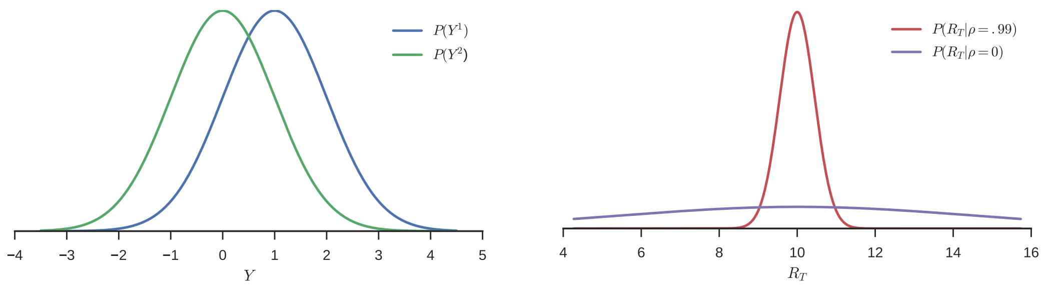
$$P(Y_t^1, Y_t^2) \sim N(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}) \tag{3.14}$$

Consider an algorithm that always selects arm 2. The distribution over the difference of jointly normal random variables is also normal, letting, $\tau_t = Y_t^1 - Y_t^2$ yields $\mathrm{P}\left(\tau_t\right) = N(\mu_1 - \mu_2, 2\sigma^2(1-\rho))$. Thus the distribution over the regret for this algorithm is given by,

$$R_T \sim N\left(T(\mu_1 - \mu_2), 2\sigma^2 T(1-\rho)\right) \tag{3.15}$$

The parameters of the marginal distributions, $\mathrm{P}\left(Y_t^1\right)$ and $\mathrm{P}\left(Y_t^2\right)$, can be estimated directly by simply sampling from each arm. However, the covariance ρ cannot, because we never simultaneously observe the rewards for both arms. As a result, even with full knowledge of the reward distributions for each arm, the distribution over regret cannot, in general, be computed without untestable assumptions about the covariance between counterfactuals (see figure 3.11).

This result is somewhat perturbing. The stochastic bandit problem can be defined without recourse to counterfactual variables by having the world stochastically generate only a reward for the selected action at each timestep, and the behaviour of standard bandit algorithms depends only on the marginal reward distributions for each arm. The expectation of the regret as defined by equation 3.2 also remains unchanged as both its definition and the learner's actions depend only on the marginal distributions. It seems therefore unfortunate that we should have to assume, for example, that the rewards for alternate actions are independent of one-another to be able to analyse the variance of the regret.



(a) Marginal distributions over the rewards for each action for $\mu_1 = 1$, $\mu_2 = 0$ and $\sigma = 1$.

(b) The distribution over R_T for $T = 10$ for two different values of ρ

Figure 3.11: The distribution over the regret as defined by equation 3.12 depends on unobservable properties of the joint distribution over counterfactual rewards. The same (marginal) distributions over the rewards can correspond to quite different regret distributions.

This is particularly so, as this assumption is likely to be violated in many realistic bandit problems. For example, a given user may be more (or less) likely to buy something no matter which advertisement they are served, leading to a positive correlation between counterfactual rewards. Equally, an illness might have two (unobservable) subtypes, with each medication (bandit arm) effective only against one, resulting in a negative correlation between the counterfactual rewards.

We could focus only on analysing the expected regret, since this depends only on the marginal distributions. However, there are many real problems for which we do care how tightly concentrated the regret is around its expectation. For example, if we are risk averse, we may prefer an algorithm with slightly higher expected regret but a lower probability of suffering extremely large regret. This raises the question, is it possible to construct an alternate definition of regret that can capture how consistently bandit algorithms behave, but that does not depend on any properties of the joint distribution over counterfactual rewards? A natural candidate would be:

$$R_T = \sum_{i=1}^{k} N_i(T)\Delta_i, \tag{3.16}$$

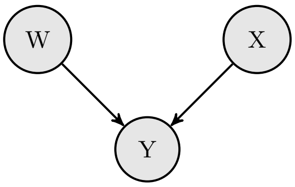
where $N_i(T)$ is the number of times arm i was played up to timestep T and Δ_i is the degree to which arm i is sub-optimal, $\Delta_i = \mu^* - \mu_i$. The expectation of this variant of regret is the same as for the version defined in equation 3.12. It depends on the randomness of the reward distribution only indirectly through the number of times each action is selected, which in turn depends only on the marginal distributions. Furthermore, this quantity has already been analysed in existing work on the concentration of bandit regret, [15, 14] as a more tractable proxy to the standard regret. In conclusion, when selecting measures of bandit performance, it is worth noting whether they rely on counterfactual assumptions and considering if these assumptions are justifiable or needed for the specific problem of interest.

Chapter 4

Causal Bandits: Unifying the approaches

As we have seen in the previous two chapters, critical insights and methods for learning how to act have emerged from both the observational and experimental approaches to causality. From observational inference, we have the do-calculus, which allows us to map information between observational and interventional distributions and we have a range of estimators designed for these types of mapping problems including propensity scores and doubly robust methods. The bandit community has developed carefully tuned algorithms to balance exploration and exploitation, estimation approaches for identifying and selecting the optimal action (rather than estimating the rewards for all actions) and techniques to bound the bias of estimators due to the adaptive nature of decision making processes.

However, despite the commonality of the underlying problem, there has been relatively little work on the intersection between observational causal inference and adaptive decision making processes. Although some ideas, such as doubly robust estimators and inverse propensity score weighting have transferred across (from causal inference to off-policy evaluation), there are many more opportunities for fruitful work in this space. In most key real world decision making processes, from algorithms recommending drug treatments to robots deciding how to open doors, we have access to both potentially huge observational data sets, as well as the ability to intervene in the system at some level. It is critical that we develop methods that allow us to incorporate the causal knowledge we have about the world, along with data sets collected under a range of different conditions, into our decision making processes. In this chapter, I introduce a very general framework that connects causal graphical models with multi-armed bandit problems and demonstrate how it can be leveraged to make better decisions.



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

Figure 4.1: 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.

4.1 A synthesis that connects bandit problems and observational causal inference

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 4.1 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 causal graph is discussed in §4.2.4.

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. Where there is no intervention, the temperature, soil, and moisture vary naturally from season to season due to weather conditions. 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 now formalise the definition of causal bandit problems. 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).

Definition 18 (Causal bandit problem)**.** A learner for a causal bandit problem is given the causal model's graph G over variables $\mathcal{X}$ and a set of allowed actions $\mathcal{A}$. Each action

$a \in \mathcal{A}$, denoted $do(\boldsymbol{X} = \boldsymbol{x})$, assigns values $\boldsymbol{x}$ to a corresponding set of variables $\boldsymbol{X} \subset \mathcal{X}$ and incurs a known cost $C(a)$ on the learner. One variable $Y \in \mathcal{X}$ is designated as the *reward variable.*

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

1. *observes* the value of a subset of the variables $\boldsymbol{X}^c_t$,
2. *intervenes* by choosing $a_t = do(\boldsymbol{X}_t = \boldsymbol{x}_t) \in \mathcal{A}$ based on previous observations and rewards,
3. *observes* values for another subset of variables $\boldsymbol{X}^o_t$ drawn from $\mathrm{P}\left(\boldsymbol{X}^o_t | \boldsymbol{X}^c_t, do(\boldsymbol{X}_t = \boldsymbol{x}_t)\right)$,
4. *obtains reward* $r_t = Y_t - C(a_t)$, where Y_t is sampled from $\mathrm{P}\left(Y_t | \boldsymbol{X}^c_t, do(\boldsymbol{X}_t = \boldsymbol{x}_t)\right)$

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 notation $\mathrm{P}\left(\cdot | \boldsymbol{X}^c_t, do(\boldsymbol{X}_t = \boldsymbol{x}_t)\right)$ denotes distributions conditional on having observed $\boldsymbol{X}^c_t$ and *then* intervened to set $\boldsymbol{X}_t = \boldsymbol{x}_t$. The values of variables in $\boldsymbol{X}^c_t$ that are non-descendents of $\boldsymbol{X}_t$ remain unchanged by the intervention. The objective of the learner is to minimise either the simple (equation 3.6) or cumulative regret (equation 3.2).

The empty intervention (where no variable is set) is denoted $do()$. 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 $\mathcal{G}$. We denote the expected reward for the action $a = do(\boldsymbol{X} = \boldsymbol{x})$ by $\mu_a := \mathbb{E}\left[Y | do(\boldsymbol{X} = \boldsymbol{x})\right]$ and the optimal expected reward by $\mu^* := \max_{a \in \mathcal{A}} \mu_a$.

The causal bandit problem takes on characteristics of different bandit settings depending on the action-space $\mathcal{A}$ and corresponding costs, which variables are observable prior to selecting an action, and on which variables we receive post-action feedback. If we can (at no cost) intervene on all of the parents of Y simultaneously then any context or alternative actions are irrelevant, and the problem can be treated as a standard multi-armed bandit problem over the set of actions that fully specify the values of the parents of Y. 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 4.2 as an example. If we receive post-action feedback on additional variables 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 4.1. The causal structure implies:

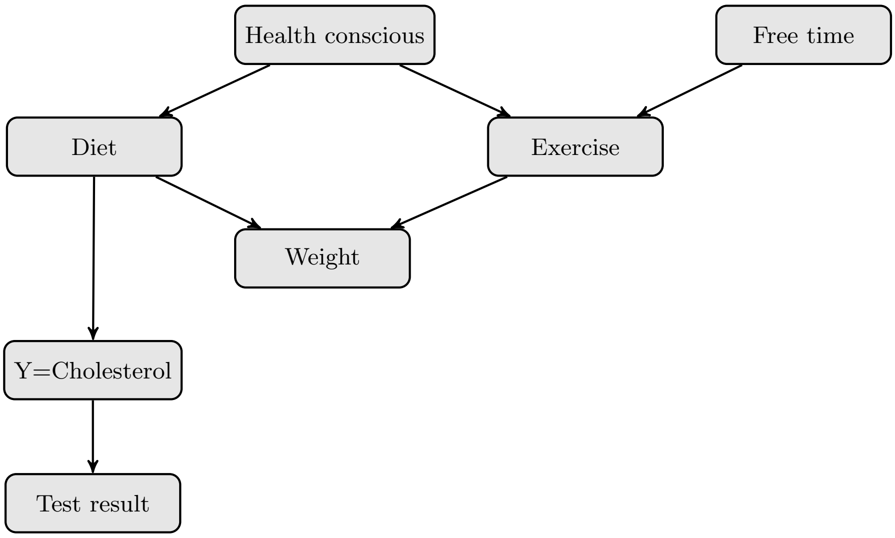


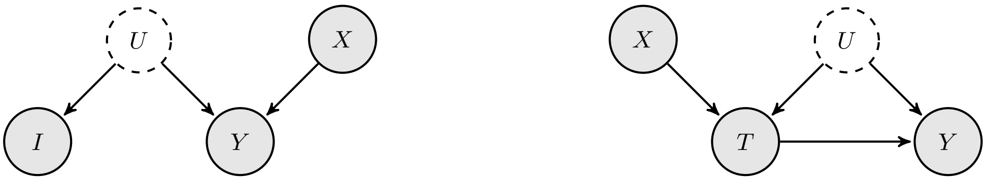
Figure 4.2: Example causal graph (based on Koller and Friedman [97]) where the outcome of interest (reward) is cholesterol level. The do-calculus can be applied to eliminate some actions immediately without the need to do any experiments. For example, no actions involving 'Test Result' need to be considered and interventions on 'Diet' do not need to be considered in conjunction with any other variables.

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

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 do not specify the value of every variable, for example those in the bottom half of the table in figure 4.1.

Two other problems that sit in the space between causal inference and bandit problems - bandits with unobserved confounders [23] and compliance aware bandits [48] - can also be viewed as specific causal bandit problems. In the work on bandits with unobserved confounders, it is assumed that the reward given each action may depend on some latent variable U, which we cannot observe directly. However, prior to selecting an action, we can observe I, the action that would have been selected under an alternate (stationary) policy, which may depend on U, see figure 4.3a. In this case, the set of contextual variables $\boldsymbol{X}^c = I$, the set of post-action feedback variables $\boldsymbol{X}^o = \{Y\}$ and the action space consists of all possible assignments of values to a single node X, $\mathcal{A} = do(X=x)$. This setting reduces to a contextual bandit problem in our causal bandit framework. However in their work on bandits with unobserved confounders, Forney and Bareinboim [57] also leverage the fact that I represents the action that would have been selected under an alternate policy to fuse data collected under the previous (observational) policy with data collected under the new policy. This information is not encoded in the causal bandit graph in figure 4.3a, as I could be any variable that is influenced by the unobserved context U, and thus cannot be exploited by a standard contextual bandit algorithm.

Compliance-aware bandits describe situations in which the action recommended by the bandit algorithm is not always followed. For example, a patient may refuse to take a treatment or an advertiser may have complex rules about how many advertisements a given customer can receive, which prevents some of the suggestions from the ad recommendation engine from being followed. After an action is selected, the algorithm can observe the action that was actually taken, in addition to the reward. Della Penna et al. [48] analyse this setting with binary treatments both with and without the presence of a latent confounding variable U, see figure 4.3b. In this case, there are no contextual variables and the action space is again the set of assignments to a single variable but there is post-action feedback, which reveals the value of the action that was actually taken.[1]


(a) Bandits with unobserved confounders: $\boldsymbol{X}^c = \{I\}$, $\boldsymbol{X}^o = \{Y\}$, $\mathcal{A} = do(X = x)$

(b) Compliance aware bandits: $\boldsymbol{X}^c = \{\}$, $\boldsymbol{X}^o = \{T, Y\}$, $\mathcal{A} = do(X = x)$

Figure 4.3: Causal bandit representation of bandits with unobserved confounders and compliance aware bandits.

The classical K-armed stochastic bandit problem can be recovered in the causal bandit setting 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.

4.2 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, the cost of all allowable

[1] There are some interesting variants of the compliance aware bandit setting that, to my knowledge, have not been analysed. The first is if the confounding variable U is observable, either as context or post-action feedback. The second is if we extend the allowable action set to include acting directly on X, albeit at a higher cost than acting on A. It is also worth noting the connection between this setting and instrumental variables [22]. By making some functional assumptions about the relationships between the variables, we can use A as an instrumental variable to bound or estimate the (causal) effect of X on Y. The estimation will be complicated in the bandit setting because the action chosen at each timestep is dependent on the previous sequence of actions and rewards.

actions is equal and the goal of the learner is to minimise the simple regret. I presented this work at NIPS 2016 [102].

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 [89]. However, as we show in §4.2.1, ignoring the extra information available in the non-intervened variables yields sub-optimal performance.

Causal bandit problems have a superficial similarity to contextual bandit problems, §3.2.4, since the extra observations on non-intervened variables might be viewed as context for selecting an intervention. However, a crucial difference is that we have focused on analysing settings where we receive post-action feedback, in which 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. [8], Kocák et al. [96] have considered very general models related to partial monitoring games [25] where rewards on un-played actions are revealed according to a feedback graph. As we discuss in §4.2.4, 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. The partial monitoring approach taken by Wu et al. [176] could be applied (up to modifications for the simple regret) to the parallel bandit, but the resulting strategy require knowledge about the likelihood of each factor in advance, while our strategy learns this online. Yu and Mannor [177] 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. [21] 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. Partial monitoring is a very general framework for decoupling the feedback from the action and reward. It can be used to classify problems into one of four categories, trivial with no regret, easy with $R_T = \tilde{\Theta}\left(\sqrt{T}\right)$, hard with $R_T = \Theta\left(T^{2/3}\right)$ and hopeless with $R_T = \Omega\left(T\right)$ [25]. Partial monitoring algorithms yield results that are optimal with respect the horizon T but not other parameters, such as the number of actions, which is the key focus of incorporating causal structure.

Two pieces of recent work also consider applying ideas from causal inference to bandit problems. Bareinboim et al. [23] 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 from this thesis 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 [116]

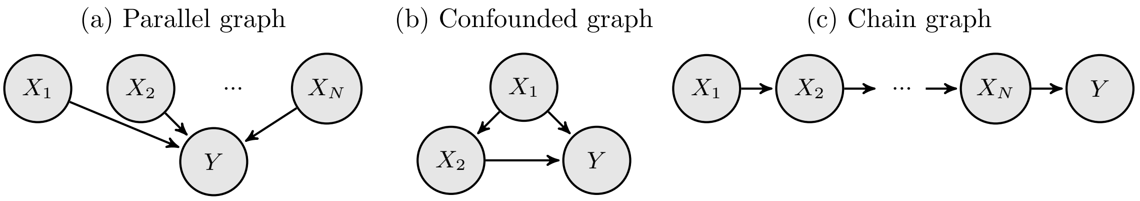


Figure 4.4: Causal Models

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 §4.2.2 have been studied before in a causal setting by Bottou et al. [30], 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.

4.2.1 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 §4.1, including the farming example. It also suffices to witness the regret gap between algorithms that make use of causal models and those that 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 4.4a. 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 above, 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 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 4.2.5.1 and 4.2.5.2 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.

4.2.2 General graphs

We now consider the more general problem where the graph structure is known, but arbitrary. For general graphs, $\mathrm{P}\left(Y|X_i = j\right) \neq \mathrm{P}\left(Y|do(X_i = j)\right)$ (correlation is not causation). However, if all the variables are observable, any causal distribution $\mathrm{P}\left(X_1...X_N|do(X_i = j)\right)$ can be expressed in terms of observational distributions via the truncated product formula [119].

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

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 formula (or the do-calculus if there are latent variables) to write an expression for each $\mathrm{P}\left(Y|a\right)$ 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 4.4c and suppose each variable deterministically takes the value of its parent, $X_k = X_{k-1}$ for $k \in 2, \ldots, N$ and $\mathrm{P}\left(X_1\right) = 0$. 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 $\mathrm{P}\left(X_i = j\right)$ is large. Given the causal graph in Figure 4.4b, $\mathrm{P}\left(Y|do(X_2 = j)\right) = \sum_{X_1} \mathrm{P}\left(X_1\right) \mathrm{P}\left(Y|X_1, X_2 = j\right)$. Suppose $X_2 = X_1$ deterministically, no matter how large $\mathrm{P}\left(X_2 = 1\right)$ is we will never observe $(X_2 = 1, X_1 = 0)$ and so cannot get a good estimate for $\mathrm{P}\left(Y|do(X_2 = 1)\right)$.

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 $\mathrm{P}\left(\mathcal{P}\mathrm{a}_Y\,|a\right)$ (but not $\mathrm{P}\left(Y|a\right)$) 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 $Q = \sum_{a \in \mathcal{A}} \eta_a \operatorname{P}(\mathcal{P}a_Y \,|a)$ 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 $\eta$
    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 $\mathcal{P}a_Y(X)$ denote the realisation of the variables in X that are parents of Y and define $R_a(X) = \frac{\operatorname{P}\{\mathcal{P}a_Y(X)|a\}}{\operatorname{Q}(\mathcal{P}a_Y(X))}$

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

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

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

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 4.2.5.3.

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 $m(\eta)$.

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

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 $\eta_a = 1/|\mathcal{A}| \, \forall a$.

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

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 4.2.5.4.

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. [30] who also use importance weighted estimators to learn from observational data.

Remark 24. If the action space $\mathcal{A}$ is unconstrained, that is consists of all possible assignments of values to variables, (and all actions have equal cost) the optimal action will set the value of all the parents of Y and Algorithm 3 cannot do better than uniform exploration over these arms. In this case, after we use the causal structure to eliminate irrelevant actions prior to taking any samples, the problem of selecting within the remaining actions can be treated as a standard multi-armed bandit problem as we state more formally below.

Theorem 25. *Let $\mathcal{A}'$ be the set of all possible assignments of values to the parents of Y. If $\mathcal{A}' \subseteq \mathcal{A}$ and $C(a') \leq C(a)\ \forall (a' \in \mathcal{A}', a \in \mathcal{A}/\mathcal{A}')$ then the optimal action $a^* \in \mathcal{A}'$ and Algorithm 3 cannot do better than uniform exploration over $\mathcal{A}'$.*

Proof. for any action $a \in \mathcal{A}$,

$$\begin{aligned}
\mathbb{E}\left[Y|\boldsymbol{X}_t^c, a\right] &= \mathbb{E}_{\mathcal{P}\mathrm{a}_Y \sim \mathrm{P}(\mathcal{P}\mathrm{a}_Y|\boldsymbol{X}_t^c, a)}\left[\mathbb{E}\left[Y|\boldsymbol{X}_t^c, a, \mathcal{P}\mathrm{a}_Y\right]\right] \\
&= \mathbb{E}_{\mathcal{P}\mathrm{a}_Y \sim \mathrm{P}(\mathcal{P}\mathrm{a}_Y|\boldsymbol{X}_t^c, a)}\left[\mathbb{E}\left[Y|\,\mathcal{P}\mathrm{a}_Y\right]\right] \\
&= \mathbb{E}_{\mathcal{P}\mathrm{a}_Y \sim \mathrm{P}(\mathcal{P}\mathrm{a}_Y|\boldsymbol{X}_t^c, a)}\left[\mathbb{E}\left[Y|do(\mathcal{P}\mathrm{a}_Y)\right]\right] \\
&\leq \max_{\mathcal{P}\mathrm{a}_Y} \mathbb{E}\left[Y|do(\mathcal{P}\mathrm{a}_Y)\right] = \mathbb{E}\left[Y|a'\right] \text{ for some } a' \in \mathcal{A}'
\end{aligned}$$

This proves the optimal action $a^* \in \mathcal{A}'$

We now consider using importance weighted estimators from Algorithm 3 to estimate the rewards for all actions in $\mathcal{A}'$. The optimal sampling weights η are given by,

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

Note that we now only have to obtain estimates for actions $a \in \mathcal{A}'$, since we know the others to be sub-optimal, so the *max* is only over these actions. However b still sums over all possible actions in the denominator of the importance sampling estimator, to allow for the possibility that playing sub-optimal actions allows more efficient estimation of the optimal actions. We now prove that this is not the case and that, in this specific setting, Algorithm 3 cannot do better than uniform sampling over the actions $a \in \mathcal{A}'$. Each action $a \in \mathcal{A}'$ consists of a given assignment $\boldsymbol{x}_a$ to $\mathcal{P}\mathrm{a}_Y$.

$$\begin{aligned}
a =& do(\mathcal{P}\mathrm{a}_Y = \boldsymbol{x}_a) \implies \mathbb{E}_a \left[\frac{\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}{\sum_{b \in \mathcal{A}} \eta_b \mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|b\right\}}\right] = \frac{1}{\sum_{b \in \mathcal{A}} \eta_b \mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y = \boldsymbol{x}_a|b\right\}} \\
&\implies \eta^* = \arg\max_{\eta} \left[\min_{a \in \mathcal{A}'} \sum_{b \in \mathcal{A}} \eta_b \mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y = \boldsymbol{x}_a|b\right\}\right]
\end{aligned}$$

Let N_a denote $\sum_{b \in \mathcal{A}} \eta_b \mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y = \boldsymbol{x}_a|b\right\}$, which can be viewed as an effective number of samples for action a. Choosing $\eta_b = \mathbb{1}\{b \in \mathcal{A}'\}\frac{1}{|\mathcal{A}'|}$, corresponding to uniform exploration over the optimal arms only, yields $N_a = \frac{1}{|\mathcal{A}'|}$ for all a. To do better, we would need to find weights η such that $N_a > \frac{1}{|\mathcal{A}'|}$ for all a. However,

$$N_a > \frac{1}{|\mathcal{A}'|}\forall a \implies \sum_{a\in\mathcal{A}'} N_a > 1$$

$$\implies \sum_{b\in\mathcal{A}} \eta_b \sum_{a\in\mathcal{A}'} \mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y = \boldsymbol{x}_a | b\right\} > 1 \implies \sum_{b\in\mathcal{A}} \eta_b > 1$$

This violates the fact that η is a distribution over the actions, and thus must have weights that sum to 1, thus completing the proof. □

4.2.2.1 Combining interventional and observational data

In the previous sections we have demonstrated that knowledge of the causal structure between variables allows the rewards of multiple actions to be estimated simultaneously in an online, interventional setting, leading to more rapid identification of the optimal arm. In many cases we may also have access to observational data; in other words data sampled from $\mathrm{P}\left(\boldsymbol{X}, Y | do()\right)$. Under the assumptions we made for algorithm 3, we can incorporate such data in a straightforward way. Let N be the number of observational data points previously collected and T again be the number of rounds in which we can intervene. Combining these two periods, the values for the parents of Y are sampled from;

$$Q'(\mathcal{P}\mathrm{a}_Y(X)) = \frac{N}{N+T}\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|do()\right\} + \frac{T}{N+T}\sum_{b\in\mathcal{A}}\eta_b \mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|b\right\} \tag{4.1}$$

We can then let $R_a(X) = \frac{\mathrm{P}\{\mathcal{P}\mathrm{a}_Y(X)|a\}}{Q'(\mathcal{P}\mathrm{a}_Y(X))}$ and select weights to minimise the worst case variance of the importance weighted estimator with respect to Q'.

$$\eta^* = \arg\min_{\eta} \underbrace{\max_{a\in\mathcal{A}} \mathbb{E}_a\left[\frac{\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}{Q'(\mathcal{P}\mathrm{a}_Y(X))}\right]}_{m'(\eta)} . \tag{4.2}$$

Algorithm 3 is modified to sum over both N and T [2] and obtains regret bounds;

$$R_T \in \mathcal{O}\left(\sqrt{\frac{m'(\eta)}{N+T}\log\left(2(N+T)|\mathcal{A}|\right)}\right) . \tag{4.3}$$

In the worse case, the observational distribution over the parents of Y has no overlap with the distributions for any of the actions in $\mathcal{A}$. This results in $m'(\eta^*)$ that is a factor

[2] In practise, to minimise the problem dependent regret, we would also incorporate an elimination step based on the observational data prior to selecting interventions. See the discussion on making better use of the reward in section 4.2.4

of $\frac{N+T}{T}$ larger than $m(\eta^*)$ (which assumes all actions were selected from the optimal sampling distribution) and results in regret that decays with $\sqrt{T^{-1}}$ as in the case where no observational data is provided.

4.2.2.2 The relative value of observational versus interventional data

Another interesting question is the relative value of observational data versus interventional data in a given setting. Obtaining interventional data often involves substantial fixed costs in setting up a system to control the allocation of interventions. Ideally, we would be able to estimate of the additional value interventional data would provide prior to setting up such a system. The quantity $m(\eta)$ also provides a means to this goal. The regret bound in theorem 21 holds for any η. Thus we can compare the relative value of purely observational data as opposed to optimally designed interventional data by considering the ratio:

$$v_{obs} \in [0,1] = \frac{m(\eta^*)}{m(\eta_b = \mathbb{1}\{b = do()\})} \tag{4.4}$$

If $v_{obs} = 0$, then there exists an action for which the reward cannot be estimated from observational data. Thus we cannot guarantee that we will identify the optimal action regardless of the quantity of available observational data. This does not imply that observational data will not improve estimation in conjunction with interventional data as we discussed in the previous section - just that observational data *alone* is insufficient for best arm identification. If $v_{obs} = 1$, then the worst case regret from purely observational data matches that for interventional data. A value of $v_{obs} = .5$ would imply we would need twice as many samples to obtain the same regret bound from observational data as compared to interventional data.

The ratio v_{obs} can be computed prior to collecting any data, observational or interventional, if the distribution over the parents of Y given each action are known - as we assumed for Algorithm 3. The approach is also not limited to the comparison of observational data with optimised interventional data. It can equally be applied to evaluate the potential for improvement on any other distribution over actions, for example we might want to evaluate the benefit of replacing a system that uniformly explores all actions with Algorithm 3. We should note however that theorem 21 bounds the worst case regret, which occurs when the rewards for each action are sufficiently close that we must obtain good estimates for the value of all of the actions. Where the rewards are better separated, the *problem-dependent* regret can be reduced by using an adaptive algorithm that ceases to explore actions that are sub-optimal with high probability, as we discuss in §4.2.4. Although v_{obs} accurately captures the relative value of interventional versus observational data for a fixed design algorithm like Algorithm 3, in general, using interventional data with an adaptive algorithm will lead to lower regret than selecting the best action on the basis of

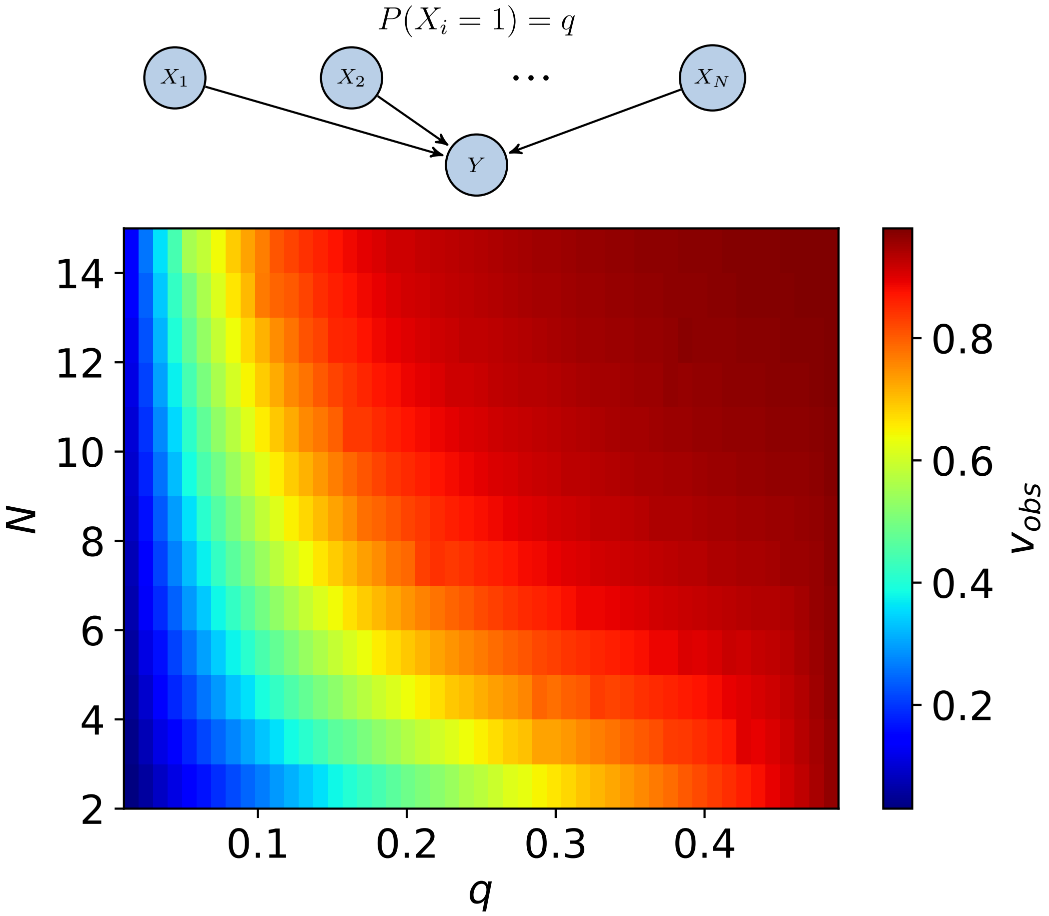


Figure 4.5: An example of quantifying the value of purely observational data. The heat map shows the value of v_{obs} for an instance of the parallel bandit problem where all the variables have equal probability, q, of taking the value 1, as a function of q and the number of variables N. If the variables are perfectly balanced, $q = 0.5$, then purely observing is optimal and $v_{obs} = 1$. If $q = 0$ then we cannot learn the value of actions $do(X_i = 1)$ without intervention, even with infinite observational data, and $v_{obs} = 0$. For intermediate values of q, we see that the improvement in the worst case regret that optimised intervention yields over purely observational data drops as the number of variables, N, increases.

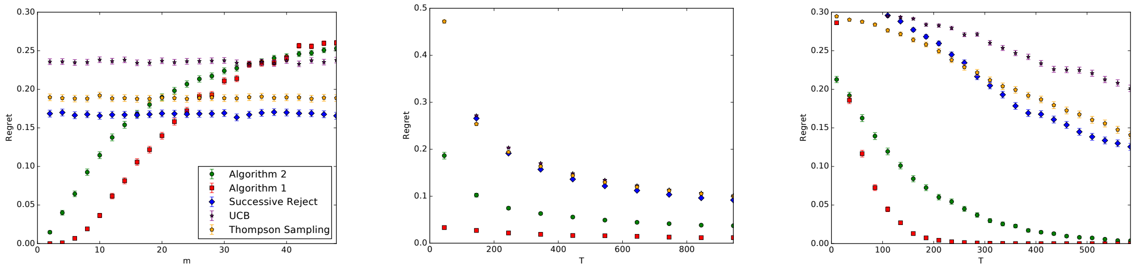
observational data even when $v_{obs} = 1$. However, the gap between the problem dependent and worst case regret is not known in advance - as the reward distributions are unknown - so the additional benefit of adaptive control cannot be computed ahead of time.

4.2.3 Experiments

We compare Algorithms 2 and 3 with the Successive Reject algorithm of Audibert and Bubeck [13], Thompson Sampling and UCB under a variety of conditions. Thompson 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 4.6a 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 4.6: Experimental results

Figure 4.6b 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 learnable by the causal algorithms, for which the worst case ε depends on $m \ll N$. In Figure 4.6c 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 [108].

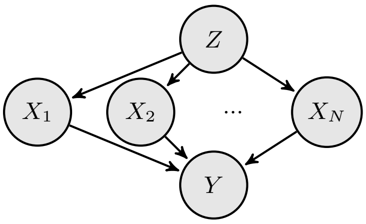


Figure 4.7: Confounded graph

We now compare the general algorithm with a range of standard bandit algorithms on the confounded graph in Figure 4.7. 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 4.8a 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

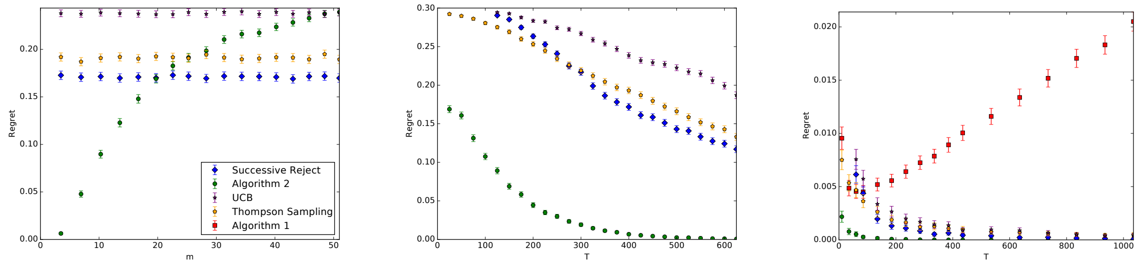
m and ranges between 2 and N. The values for the CPD's have been chosen such that the reward distribution is independent of m and so that we can analytically calculate η^*. This allows us to just show the dependence on m, removing the noise associated with different models selecting values for η^* with the same m (and also worst case performance), but different performance for a given reward distribution.

In Figure 4.8b 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 4.8c 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 4.8: Experimental results on the confounded graph

4.2.4 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 $\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y|a\right\}$ 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 4.6a.

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. This has now been done by Sen et al. [143], leading to problem dependent regret bounds for causal bandit problems. Note that these modifications do not improve the mini-max regret, which at least for the parallel bandit is already optimal. For this reason we focused on emphasising the point that causal structure can and 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 [176] on bandits with side information to get near-optimal cumulative regret guarantees. They consider a finite-armed bandit model with side information where in each 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 [8, 96], where at each timestep 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. [8], Kocák et al. [96] 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 $\boldsymbol{q} = \boldsymbol{0}$.

Lelarge and Ens [104] 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. [35] 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 Unobservable Variables If we assume knowledge of the conditional *interventional* distributions $\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y \,|a\right\}$ 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 $\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y \,|a\right\}$ 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 [54, 53, 73, 81] could potentially be incorporated into a causal bandit algorithm. In particular, Hu et al. [81] 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.

4.2.5 Proofs

4.2.5.1 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 26. *Let $i \in \{1, \ldots, N\}$ 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 $X_{t,i} \sim \mathit{Bernoulli}(q_i)$. Therefore from the Chernoff bound (see equation 6 in Hagerup and Rüb [72]),

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

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

□

Lemma 27. *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 26 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 26 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 27, $T_A \geq T/(4m(\boldsymbol{q}))$ 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,

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

For arms not in A,

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

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

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

which completes the result. □

4.2.5.2 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 [164] we have

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

where $\operatorname{KL}(\mathrm{P}_0, \mathrm{P}_i)$ 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 $i \le m$ we have $q_i \le 1/m$ 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. [17]. This leads to

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

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

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

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

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

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

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

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

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

as required.

4.2.5.3 Proof of Theorem 21

Proof. First note that X_t, Y_t are sampled from Q. We define $Z_a(X_t) = Y_t R_a(X_t) \mathbb{1}\{R_a(X_t) \le B_a\}$ 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 $|Z_{at}| \le B_a$ and

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

Checking the expectation we have

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

where

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

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. □

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

Proposition 28. *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.

□

Chapter 5

Conclusion

The underlying motivation behind much applied statistical and machine learning work is to guide us to make better decisions. In many cases, the actions that we take in response to the model will change the system from which the data was generated. It is critical that we are able to recognise the causal nature of such problems and appropriately model the decision making part of the process. If machine learning is to be as transformative in fields such as economics, medicine and social science as it has been for image recognition, voice processing and machine translation, we must develop methods to estimate the effect of, and optimally select, interventions that are as effective as those we have for pattern recognition. We need to bridge the gap between the theory driven models of economics and science and the black box prediction approach that has been so successful in machine learning. This will involve clarifying what information is (currently) best encoded by theory and what can be successfully inferred from data and developing methods that can incorporate the theory or structure required to allow models to generalise from one setting to another whilst retaining the flexibility to capture complex patterns in empirical data.

A better understanding of causality is also relevant for the discussions around transparency and ethics in machine learning, particularly with respect to the European Union's new General Data Protection Regulation, which requires that automated decision systems that significantly affect individuals provide *"meaningful information about the logic involved."* [69]. The recognition that there is a fundamental trade-off between accuracy and transparency, unless we can build perfect causal models, when the interests of individuals and society diverge has implications for the way we design and regulate systems that have the potential to have major impacts on people's lives. We must develop approaches to ensuring machine learning decisions are reasonable and ethical that allow affected individuals recourse to dispute or improve outcomes but do not undermine the ability of the system to function.

The observational and interventional viewpoints on learning to act contribute complementary components to a general approach. Observational causal inference provides, through the do-calculus, a formal means to map information from observational to interventional settings, as well as between different interventions. Bandit algorithms capture the sequential nature of decision making processes and provide techniques to carefully balance exploration and exploitation.

I have developed an approach that formally connects causal graphical models with bandit problems in a natural way and demonstrated that this conceptualisation encodes some key existing problems in the literature. I showed that knowledge of the causal structure (but without knowledge of the functional relationships) between variables can induce a novel form of structure between alternate actions and that an algorithm that leverages this structure obtains better performance that one that does not. I introduce a metric that captures the difficulty of causal bandit problems and (under strong assumptions) allows the value of optimised interventional data over purely observational data to be computed prior to intervening in the system.

This work represents a first step towards a unified approach to causal inference and optimal decision making. There is much exciting work remaining to be done. Although the causal bandit setting can capture contextual information as well as post-action feedback, I have formally analysed and developed algorithms only for the latter. Additionally, to make the problem tractable, I made the (major) assumption that the interventional distribution over the parents of the outcome was known. This can be relaxed to assuming the interventional distribution over some Markov blanket with respect to the outcome is known. Information can then be shared between actions outside the blanket in the same way as in algorithm 3, whilst actions inside could be learned explicitly. Relaxing this assumption entirely is a much more challenging problem. However, as is demonstrated by the specific example of the parallel causal bandit problem, it is possible to develop algorithms that require only the causal structure of the graph and yield substantially lower regret.

Another interesting line of research is the question of off-policy evaluation for causal-bandit problems. As in the online case, knowledge of the causal structure between variables in the graph provides additional information about the reward for the actions that were not selected at each timestep. The problem differs from typical observational causal inference in several ways: the focus is on identifying an optimal policy, rather than unbiased estimation of all policies; the goal is to explore the value of interventions on a range of different variables, rather than the optimal setting of a single variable; and the data will, in general, be non-stationary in a rather special way due to the adaptive nature of policy that generated it. Existing work on off-policy evaluation focuses (at least implicitly) on estimating causal effects by adjusting for all variables that simultaneously affect both the action selection and outcome. If additional information is available about the causal relationships between variables involved, other approaches to identifying causal effects such as instrumental variables or the “front door method” [119] could be also be applied to off-policy evaluation.

Insights from the bandit literature can also be applied to more classical causal inference problems. In particular, estimators that are geared towards optimal action selection (rather than evaluation of all actions) and approaches to quantify the finite time properties, as opposed to asymptotic efficiency, of estimators, for example [108]. An important line of research, that is relevant to both the observational and interventional approaches to causal inference, is developing methodologies for model evaluation and selection that provide something equivalent to what cross-validation does for supervised learning. My hope is that, in the next years, combining the reinforcement learning approach to decision making with causal graphical models and causal effect estimation techniques developed within statistics, economics and epidemiology will lead to a revolution in our ability to make good data driven decisions.

Bibliography

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

[2] Abadie, A. and Imbiens, G. W. (2002). Simple and Bias-Corrected Matching Estimators for Average Treatment Effects. *NBER Technical Working Paper No. 283*.

[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. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 127—-135.

[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., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online Learning with Feedback Graphs : Beyond Bandits. In *Conference on Learning Theory (COLT)*, pages 23–35.

[9] Alon, N., Cesa-Bianchi, N. N., Gentile, C., and Mansour, Y. (2013). From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1610–1618.

[10] Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired magazine*, 16(7):7–16.

[11] 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.

[12] Audibert, J. and Bubeck, S. (2009). Minimax policies for adversarial and stochastic bandits. In *Conference On Learning Theory (COLT)*, pages 773–818.

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

[14] 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.

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

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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)*, pages 409–416.

[22] Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439):1171–1176.

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

[24] Bareinboim, E. and Lee, S. (2013). Transportability from Multiple Environments with Limited Experiments. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1–9.

[25] 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.

[26] 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.

[27] 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.

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

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

[30] 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.

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

[32] Bubeck, S., Munos, R., and Stoltz, G. (2009). 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. (2011). X-armed bandits. *Journal of Machine Learning Research*, 12:1655–1695.

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

[35] Buccapatnam, S., Eryilmaz, A., and Shroff, N. B. (2014). Stochastic Bandits with Side Observations on Networks. *ACM SIGMETRICS Performance Evaluation Review*, 41(1):289–300.

[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] Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2016). Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37.

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

[40] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015).

Intelligible Models for HealthCare. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pages 1721–1730.

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

[42] Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2249–2257.

[43] Chen, B. (2016). Identification and Overidentification of Linear Structural Equation Models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1579–1587.

[44] 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.

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

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

[47] 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.

[48] Della Penna, N., Reid, M. D., and Balduzzi, D. (2016). Compliance-Aware Bandits. *arXiv preprint arXiv:1602.02852.*

[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.

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

[52] Dudik, M., Langford, J., Li, L., Dudik, M., Langford, J., and Li, L. (2011b). Doubly robust policy evaluation and learning. In *International Conference on Machine Learning (ICML)*, pages 1097–1104.

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

[54] 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 *Uncertainty in Artificial Intelligence (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 (NIPS)*, pages 586–594.

[57] Forney, A. and Bareinboim, E. (2017). Counterfactual Data-Fusion for Online Reinforcement Learners. In *International Conference on Machine Learning (ICML)*.

[58] 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.

[59] 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.

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

[61] 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 (NIPS)*, pages 3212–3220.

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

[63] 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 (NIPS)*, pages 784–792.

[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.

[65] 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.

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

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

[68] Gelman, A. and Imbens, G. (2013). Why ask why? Forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research.

[69] Goodman, B. and Flaxman, S. (2016). EU regulations on algorithmic decision-making and a "right to explanation". In *ICML Workshop on Human Interpretability in Machine Learning*.

[70] Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424.

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

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

[73] 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.

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

[75] Heckman, J. J. (2005). 1. The Scientific Model of Causality. *Sociological methodology*, 35(1):1–97.

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

[77] 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.

[78] 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.

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

[80] 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.

[81] 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 (NIPS)*, pages 2339–2347.

[82] Huang, Y. and Valtorta, M. (2006). Pearl's 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.

[83] Hume, D. (1741). A Treatise of Human Nature.

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

[85] Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(February):4–29.

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

[87] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):0696–0701.

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

[89] 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.

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

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

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

[93] Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):540–543.

[94] 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.

[95] Kaufmann, E., Korda, N., and Munos, R. (2012). Thompson Sampling: An Asymptotically Optimal Finite-Time Analysis. In *Algorithmic Learning Theory (ALT)*, pages 199–213. Springer.

[96] Kocák, T., Neu, G., Valko, M., and Munos, R. (2014). Efficient learning by implicit exploration in bandit problems with side observations. *Advances in Neural Information Processing Systems (NIPS)*, pages 613–621.

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

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

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

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

[101] Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems (NIPS)*, pages 817–824.

[102] Lattimore, F., Lattimore, T., and Reid, M. D. (2016). Causal Bandits: Learning Good Interventions via Causal Inference. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1181–1189.

[103] Lattimore, T. (2015). Optimally Confident UCB : Improved Regret for Finite-Armed Bandits. *arXiv preprint arXiv:1507.07880.*

[104] Lelarge, M. and Ens, I. (2012). Leveraging Side Observations in Stochastic Bandits. *Uncertainty in Artificial Intelligence (UAI).*

[105] Lewis, D. K. (2000). Causation as Influence. *Journal of Philosophy*, 97(April):182–197.

[106] Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). 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.

[107] Li, L., Chu, W., Langford, J., and Wang, X. (2011). Unbiased Offline Evaluation of Contextual-bandit-based News Article Recommendation Algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306.

[108] 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.

[109] 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.

[110] Mannor, S. and Shamir, O. (2011). From Bandits to Experts: On the Value of Side-Observations. *Advances in Neural Information Processing Systems (NIPS)*, pages 684–692.

[111] Mill, J. S. (1893). *A system of logic, ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation.* Harper & brothers.

[112] Mooij, J. M., Peters, J., Janzing, D., and Zscheischler, J. (2016). Distinguishing Cause from Effect Using Observational Data : Methods and Benchmarks. *The Journal of Machine Learning Research*, 17:1–102.

[113] Mozur, P. (2017). Heart 'Bumping,' a Go Master Again Loses to Google.

[114] 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.

[115] Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716–aac4716.

[116] 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.

[117] 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.

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

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

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

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

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

[123] Pearson, K. (1911). The Grammar of Science.

[124] 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.

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

[126] 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.

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

[128] 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.

[129] 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.

[130] 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).

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

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

[133] 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.

[134] Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors Are Not Always Observed. *Journal of the American Statistical Association*, 89(427):846.

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

[136] 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.

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

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

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

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

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

[142] 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.

[143] Sen, R., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017). Identifying Best Interventions through Online Importance Sampling. *International Conference on Machine Learning (ICML)*, pages 1–30.

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

[145] 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.

[146] 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.

[147] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Den, G. V., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Leach, M., Kavukcuoglu, K., Graepel, T., Hassabis, D., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., and Others (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

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

[149] 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.

[150] 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.

[151] 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 <i>DAT1</i> genetic variants and striatal brain activation during reward processing. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 168(6):508–515.

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

[153] Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: concepts and recent methodological advances. *Applied Informatics*, 3(1):3.

[154] 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*.

[155] 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.

[156] 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.

[157] 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.

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

[159] Swaminathan, A. and Joachims, T. (2015). Counterfactual Risk Minimization: Learning from Logged Bandit Feedback. In *International Conference on Machine Learning (ICML)*, pages 814–823.

[160] 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.

[161] 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.

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

[163] Tromp, J. (2016). The Number of Legal Go Positions. In *International Conference on Computers and Games*, pages 183–190. Springer.

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

[165] 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.

[166] VanderWeele, T. (2015). *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press.

[167] 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.

[168] Vigen, T. (2015). *Spurious Correlations*. Hachette Books.

[169] 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.

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

[171] 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.

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

[173] Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford university press.

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

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

[176] 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.

[177] 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.

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

[179] 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.