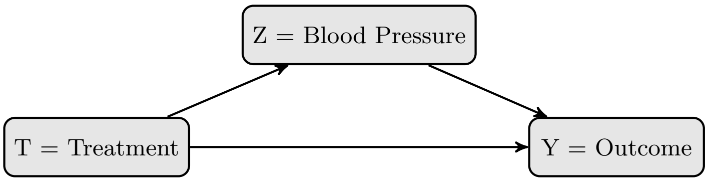


Figure 2.9: Another causal network that can exhibit Simpson's paradox. In this case, "the solution" is not to stratify on Z.

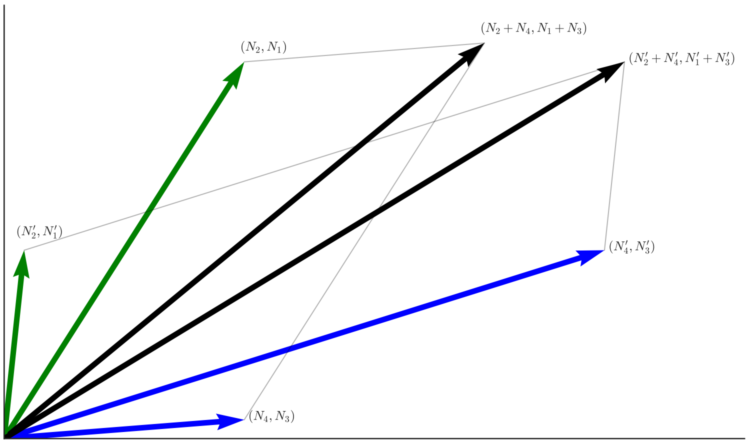


Figure 2.10: Simpson's reversal visualised. The ratios involving N'_i are steeper than those involving N_i for both the blue and green vectors. However, when we sum them, the ratio is steeper for the un-primed variables.

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.

$$\exists \left\{N_1, ...N_4, N'_1...N'_4\right\} \in \mathbb{N} : \frac{N_1}{N_2} < \frac{N'_1}{N'_2}, \quad \frac{N_3}{N_4} < \frac{N'_3}{N'_4} \text{ and } \frac{N_1 + N_3}{N_2 + N_4} > \frac{N'_1 + N'_3}{N'_2 + N'_4} \tag{2.16}$$

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

$\mathrm{P}_{train}\left\{\boldsymbol{v}\right\} = \mathrm{P}\left(\boldsymbol{z}\right)\mathrm{P}\left(\boldsymbol{x}|\boldsymbol{z}\right)$, 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.

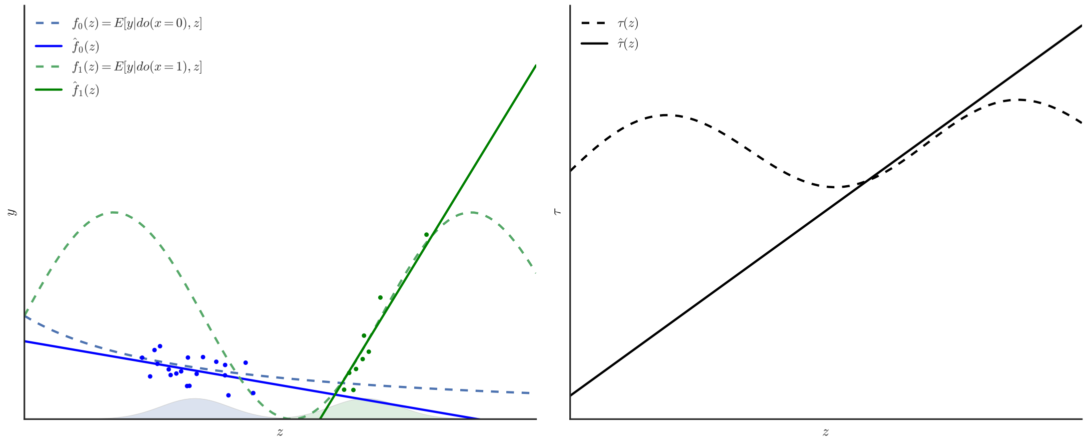


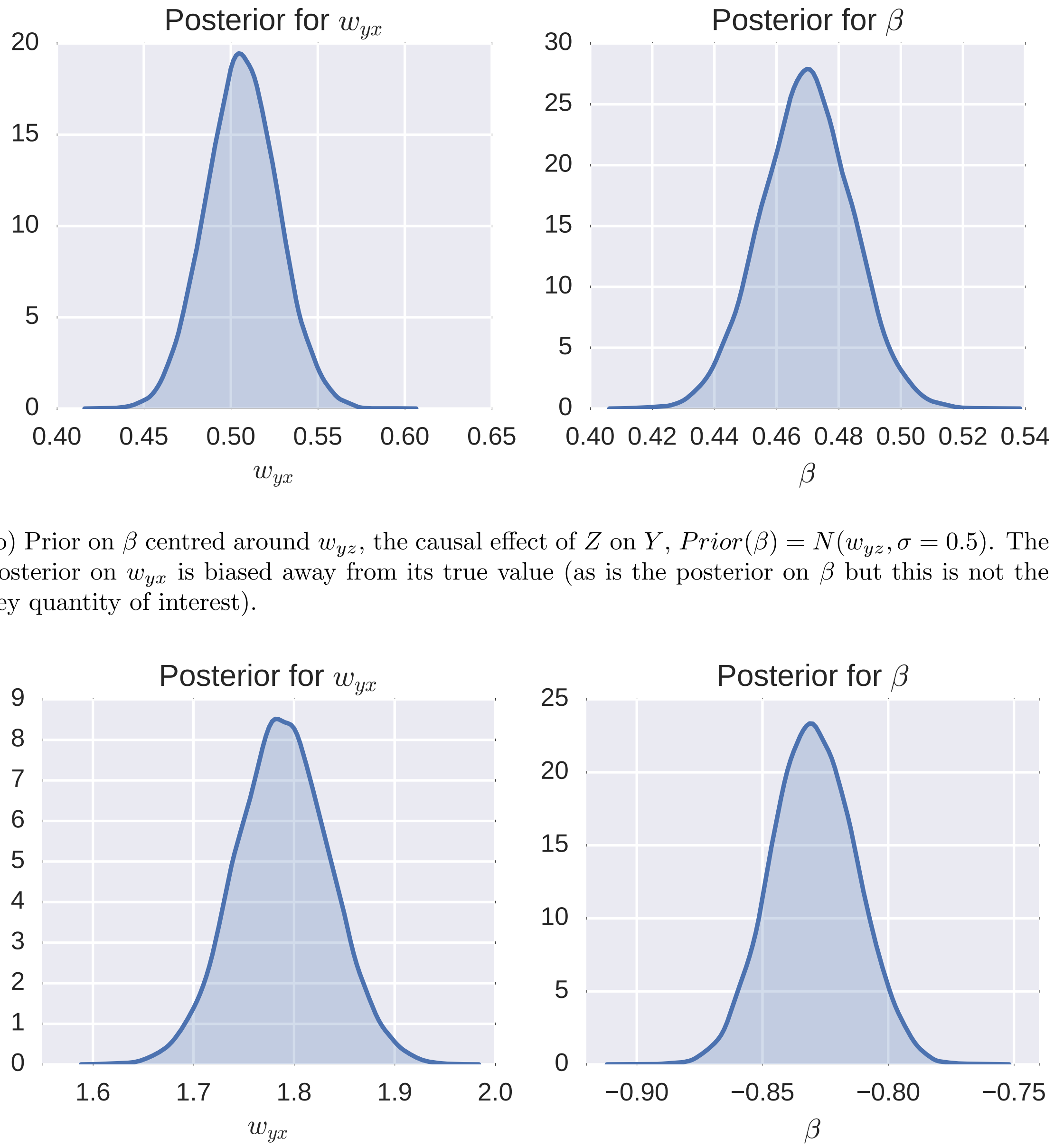
Figure 2.18: Parametric regression may yield poor estimates of causal effects if the model is missspecified, even if the model fits well over the domain of the training data. In this example, $\mathrm{P}(Z|X=0) \sim N(\mu_0, \sigma_0^2)$ and $\mathrm{P}(Z|X=1) \sim N(\mu_1, \sigma_1^2)$ with little overlap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

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

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].

(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

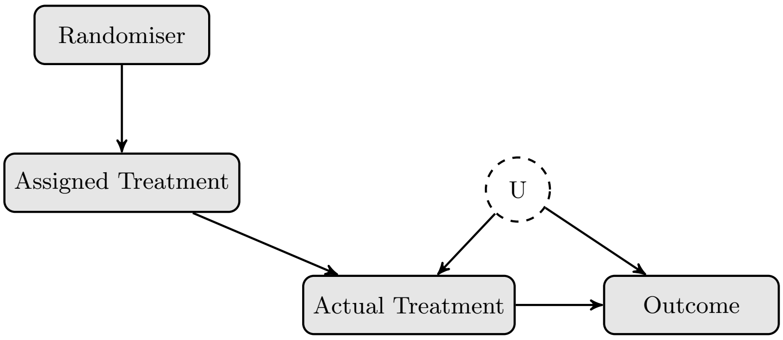


Figure 3.3: causal network for a randomised experiment with imperfect compliance

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)**.**

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

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

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].

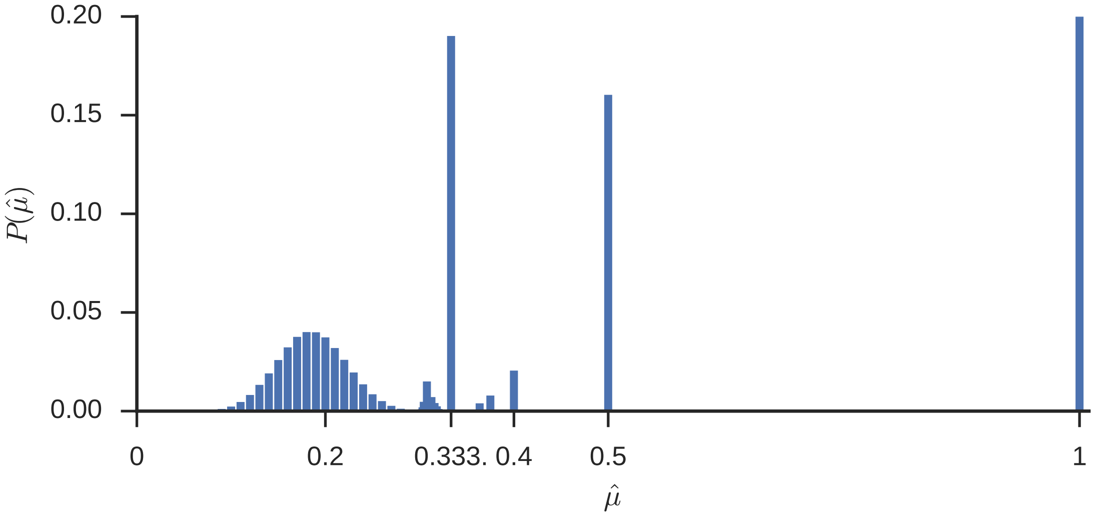


Figure 3.5: Standard empirical estimators can be biased if the number of samples, n, is not fixed in advance, but is a random variable that depends on the values of previous samples. This example plots the distribution (over 10^6 simulations) of $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim Bernoulli(0.2)$. 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** $t \in 1, \ldots, T$ **do**
    Count the number of times each arm has been selected previously $n_{t,i} = \sum_{s=1}^{t-1} \mathbb{1}\{a_s = i\}$
    Calculate the mean reward for each arm $\hat{\mu}_{t,i} = \frac{1}{n_{t,i}} \sum_{s=1}^{t-1} \mathbb{1}\{a_s = i\} \, Y_t$
    Select arm $a_t \in \arg\max_{i=\{1,\ldots,k\}} \left(\hat{\mu}_{t,i} + \sqrt{\frac{2\log t}{n_{t,i}}}\right)$
```

Let $\Delta_i = \mu_i - \mu^*$ be 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 $\bar{R}_T \in \Omega\left(\sqrt{kT}\right)$

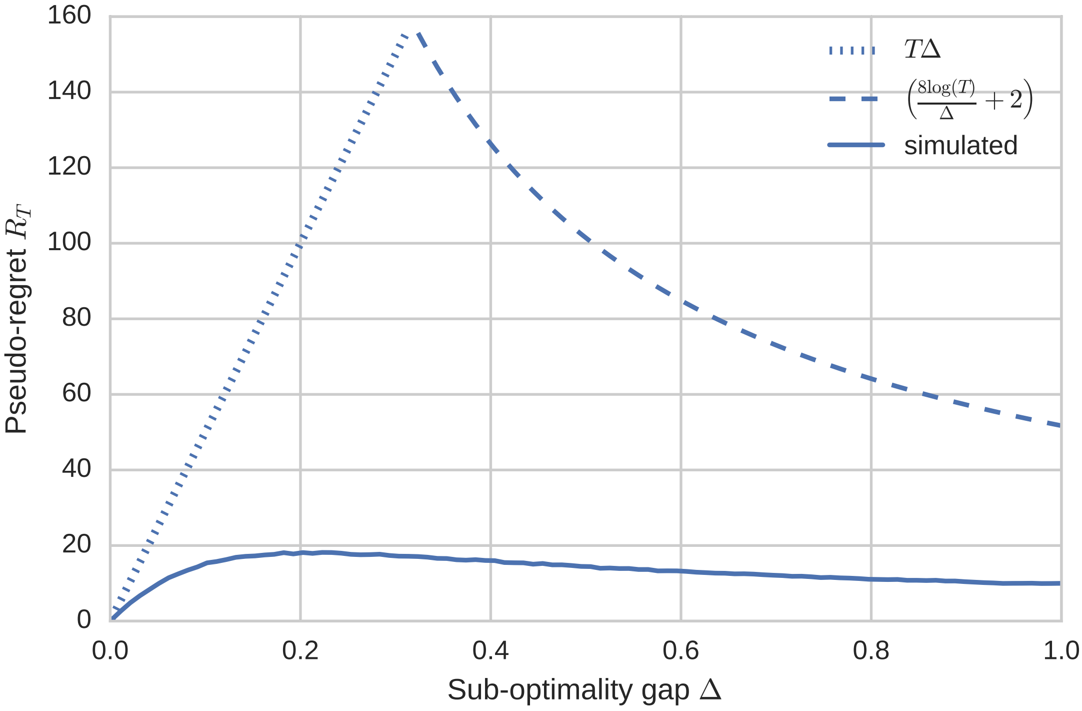


Figure 3.6: The regret bound in equation 3.4 grows as the differences between the expected rewards for each arm shrink. The solid curve shows the mean (cumulative) regret for the UCB-1 algorithm, over a 1000 simulations for a 2-armed, Bernoulli bandit with fixed horizon, $T = 500$, as a function of the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

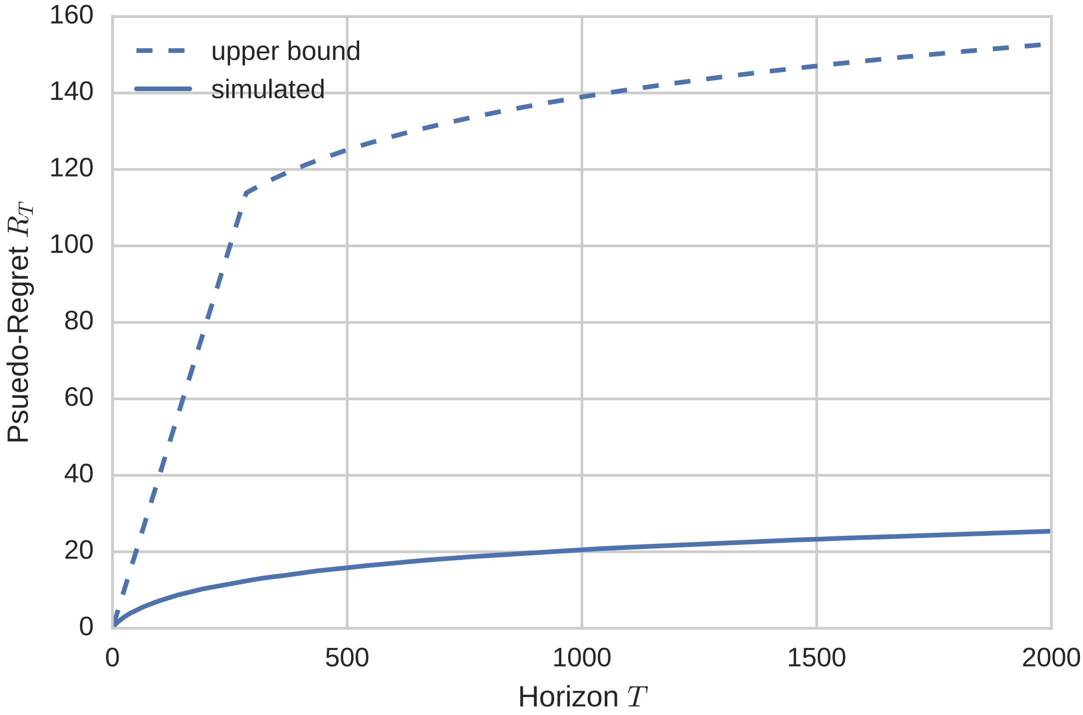


Figure 3.7: The actual performance of the UCB algorithm can be substantially better than suggested by the upper bound, particularly for small T. The solid curve shows the mean expected regret associated with the sequence of arms chosen by UCB-1 with $k = 2$ arms and the rewards sampled from $Bernoulli([.3, .7])$ 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]

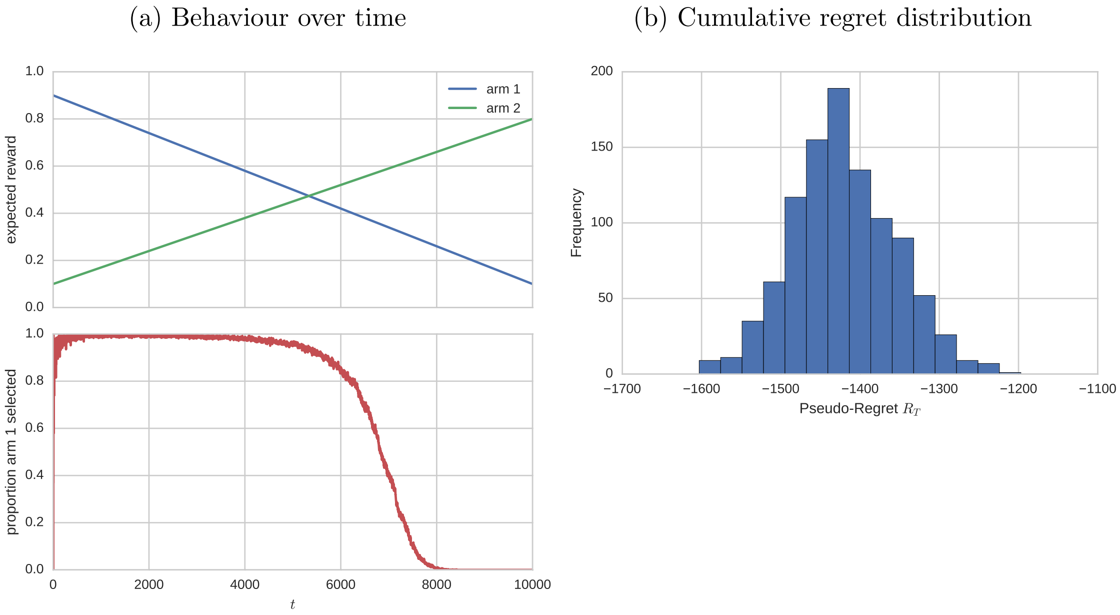
The seminal algorithm for adversarial bandits is Exp-3 [16], which, like UCB, obtains worst case pseudo-regret of $\mathcal{O}\left(\sqrt{TK\log(T)}\right)$ [17]. Optimal algorithms, with $\bar{R}_T = \mathcal{O}\left(\sqrt{TK}\right)$, have also been demonstrated for the oblivious adversarial setting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

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 $\mu_1 > \mu_2$.

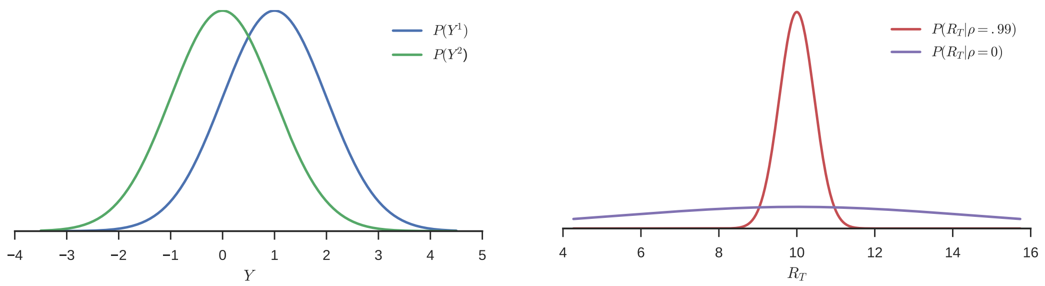
$$P(Y_t^1, Y_t^2) \sim N(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}) \tag{3.14}$$

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 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

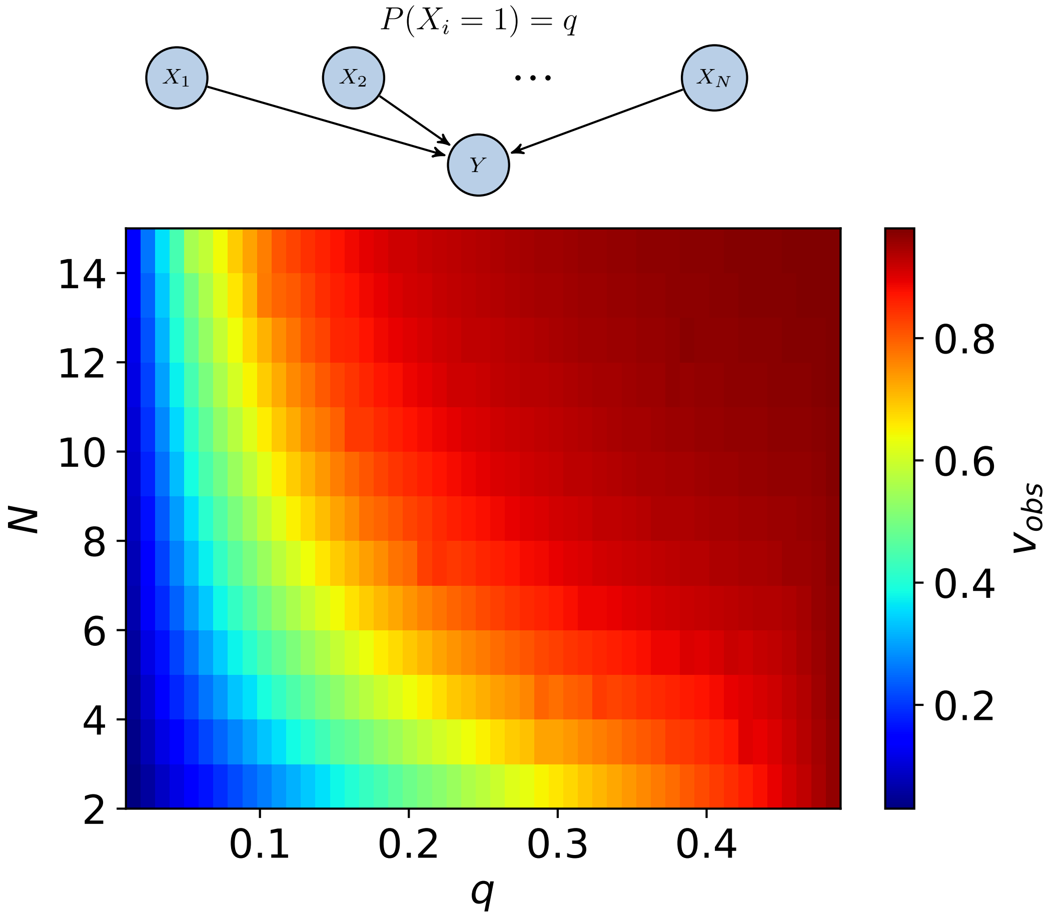


Figure 4.5: An example of quantifying the value of purely observational data. The heat map shows the value of v_{obs} for an instance of the parallel bandit problem where all the variables have equal probability, q, of taking the value 1, as a function of q and the number of variables N. If the variables 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.

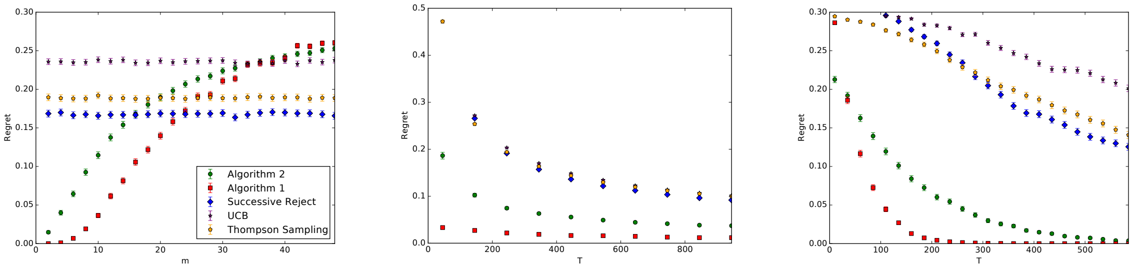
observational data even when $v_{obs} = 1$. However, the gap between the problem dependent and worst case regret is not known in advance - as the reward distributions are unknown - so the additional benefit of adaptive control cannot be computed ahead of time.

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].

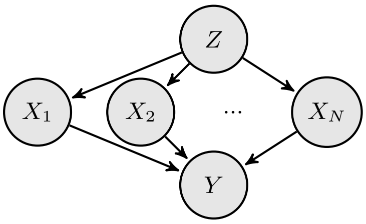


Figure 4.7: Confounded graph

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

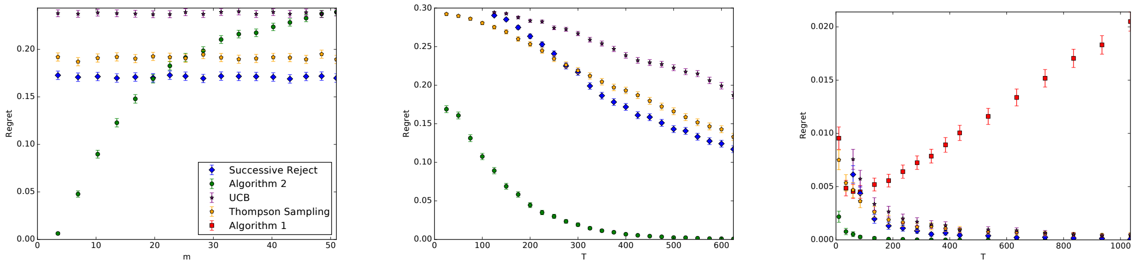
m and ranges between 2 and N. The values for the CPD's have been chosen such that the reward distribution is independent of m and so that we can analytically calculate η^*. This allows us to just show the dependence on m, removing the noise associated with different models selecting values for η^* with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