Chapter 4

The interventionist viewpoint

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

4.1 Randomised experiments

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

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

Figure 4.1: causal network for a randomised experiment

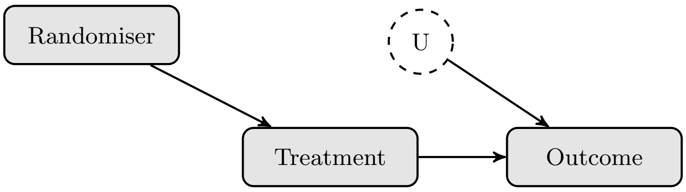
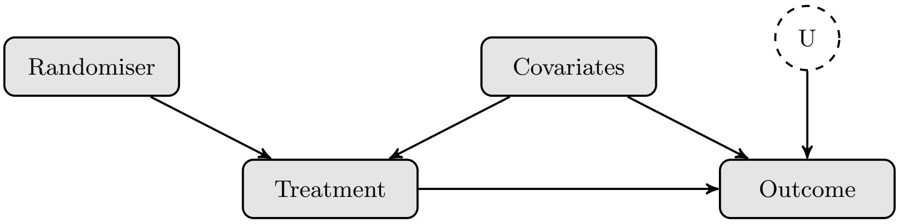


Figure 4.2: causal network for a stratified randomised experiment if the probability an individual is assigned a given treatment depends on some covariates.



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

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

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

4.1.1 Limitations of randomised experiments

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

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

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

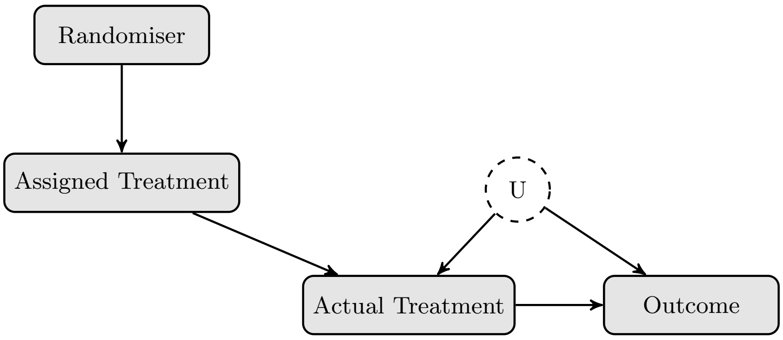
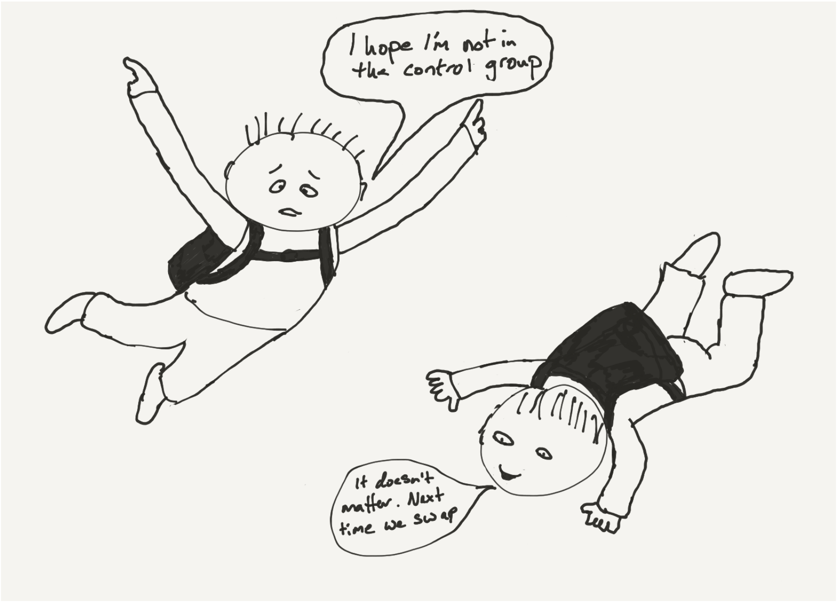


Figure 4.4: Experiments are not always ethical; an illustration of a randomised cross-over trial of parachutes for the prevention of morbidity and mortality associated with falls from large heights.


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

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

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

4.2 Multi armed bandits

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

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

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

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

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

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

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

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

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

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

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

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

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

4.2.1 Stochastic bandits: Approaches and results

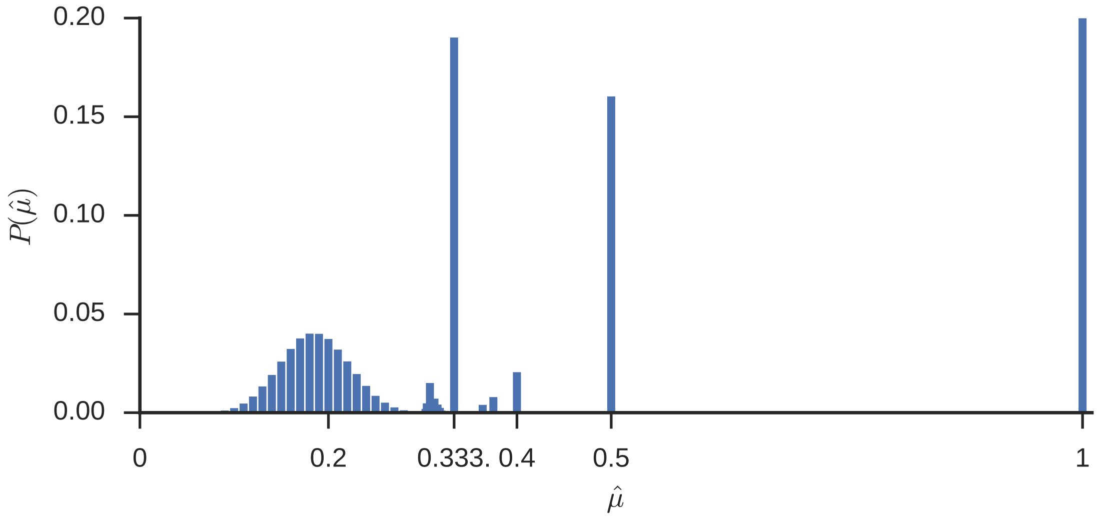
The adaptive nature of multi-armed bandit algorithms complicates the design and analyse of estimators. The action selected by an algorithm at a given timestep can depend on the history of previous actions and rewards. As a result, the probability that each action is selected evolves over time, the actions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

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

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

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

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



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

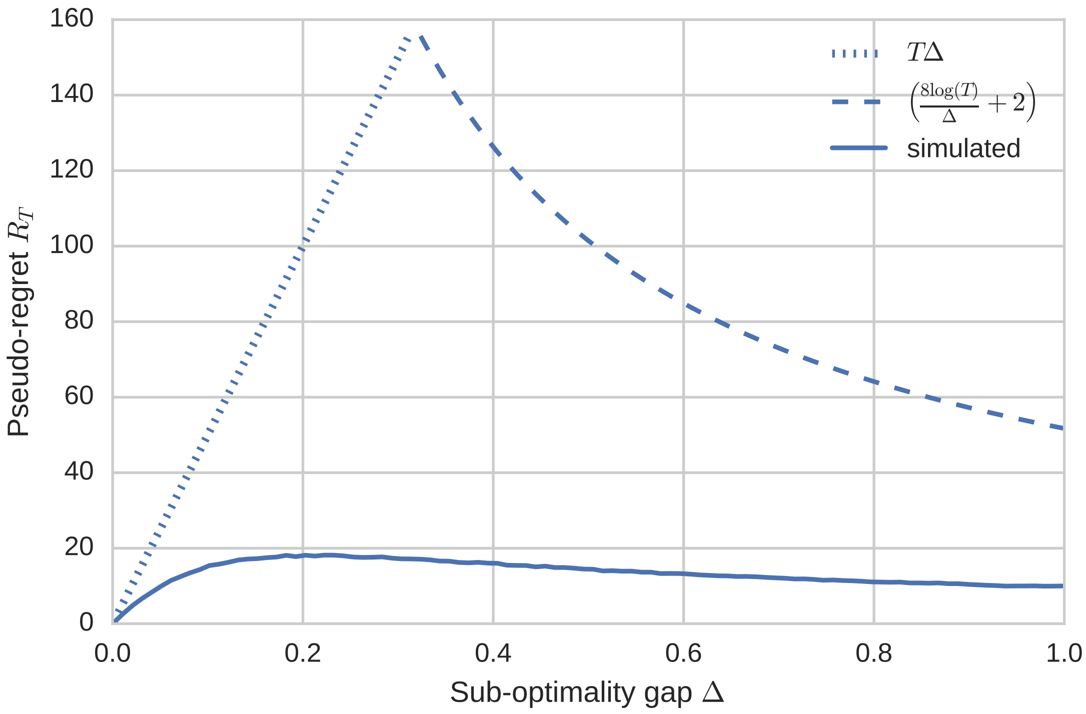
Algorithm 1 UCB-1

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

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

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

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

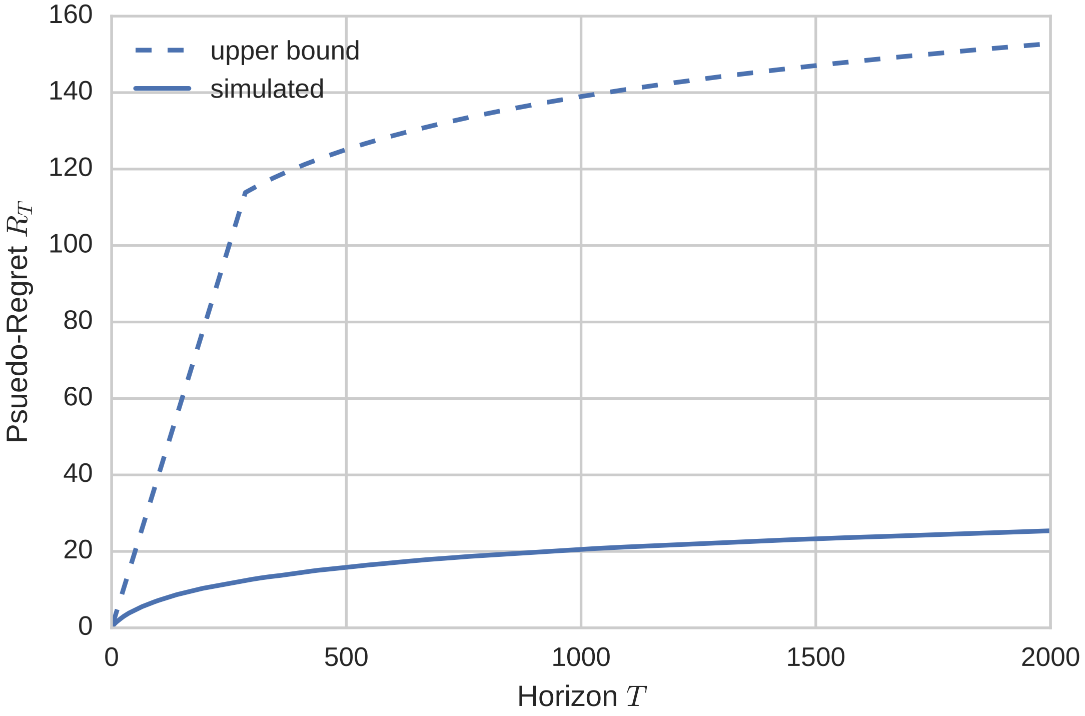


Somewhat unintuitively, the regret increases as the value of the arms gets closer together. This is because it becomes harder for the algorithm to identify the optimal arm. As the differences $\Delta_i \rightarrow 0$, the regret bound in 4.4 blows up, however the regret itself does not - since although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

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

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

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



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

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

4.2.2 Pure-exploration problems

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

Definition 15 (Simple Regret)**.**

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

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

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

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

4.2.3 Adversarial Bandits

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

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

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

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

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

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

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

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

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

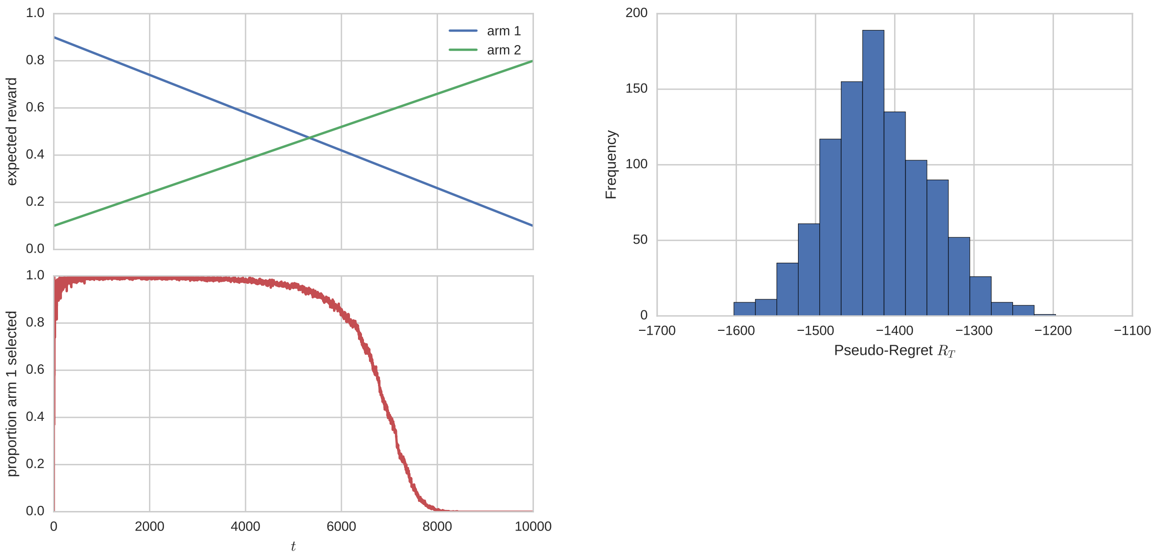
Adversarial bandits appear to be more applicable to real world problems because they do not assume that the rewards associated with each arm are constant over time or independent of the previous actions of the learner. However, pseudo-regret, as defined in equation 4.8, is not a good measure of an algorithms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

4.2.4 Contextual bandits

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

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

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



(a) Behaviour over time

(b) Cumulative regret distribution

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

1. the world stochastically generates the vector of rewards for each action and the context, $(\boldsymbol{X}_t, [Y_t^1, ..., Y_t^k]) \sim \mathrm{P}(\boldsymbol{x}, \boldsymbol{y})$ and reveals $\boldsymbol{X}_t$ to the learner

2. the learner selects an action $A_t \in \{1, ..., k\}$, based on the context as well as actions and rewards from previous time-steps,

3. the learner observes and receives (only) the reward for the selected action $Y_t = Y_t^{A_t}$

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

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

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

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

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

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

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

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

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

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

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

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

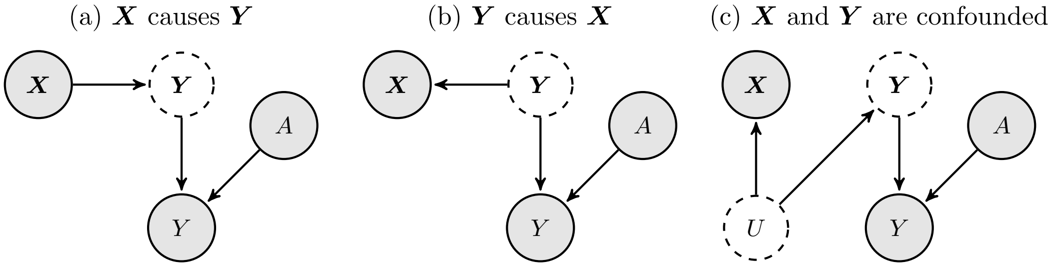
$$\arg\max_{h\in\mathcal{H}} \sum_{t=1}^{\tau} Y_t \mathbb{1}\{h(\boldsymbol{X}_t) = A_t\} \tag{4.10}$$

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

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

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

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


(a) $\boldsymbol{X}$ causes $\boldsymbol{Y}$

(b) $\boldsymbol{Y}$ causes $\boldsymbol{X}$

(c) $\boldsymbol{X}$ and $\boldsymbol{Y}$ are confounded

4.2.5 Learning from logged bandit data

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

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

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

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

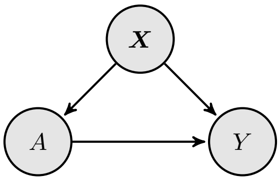
As in estimating average causal effects under ignorability, we have a covariate shift problem, with training data sampled from $\mathrm{P}\left(\boldsymbol{x}\right)\mathrm{P}_{\pi}\left\{a|\boldsymbol{x}\right\}\mathrm{P}\left(y|\boldsymbol{x}, a\right)$ but generalisation error measured with respect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

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

4.2.6 Adding structure to actions

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

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



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

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

Chapter 5

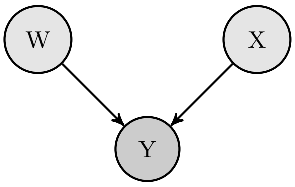
Causal Bandits: Unifying the approaches

5.1 The framework

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

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

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



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

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

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

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

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

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

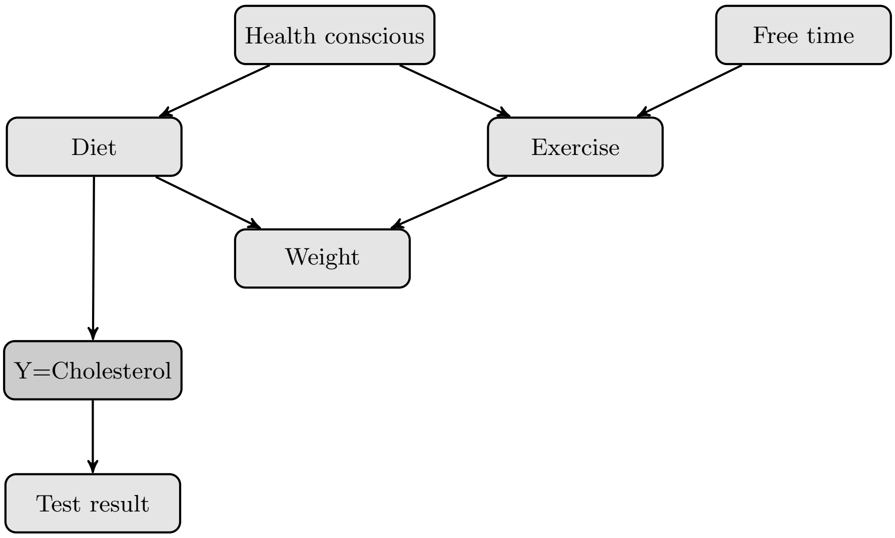
We refer to the set of variables that can be observed prior to selecting an action $\boldsymbol{X}^c$ as contextual variables and the set of variables observed after the action is chosen, $\boldsymbol{X}^o$, as post-action feedback variables. Note that $\boldsymbol{X}^c$ and $\boldsymbol{X}^o$ need not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

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

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

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

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



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

5.2 Causal bandits with post action feedback

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

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

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

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

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

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

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

5.2.1 The parallel bandit problem

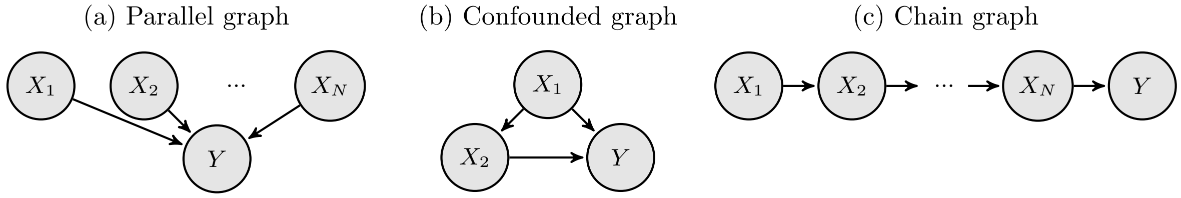
In this section we propose and analyse an algorithm for achieving the optimal regret in a natural special case of the causal bandit problem which we call the *parallel bandit*. It is simple enough to admit a thorough analysis but rich enough to model the type of problem discussed in §5.1, including the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

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

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

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

Figure 5.3: Causal Models


(a) Parallel graph

(b) Confounded graph

(c) Chain graph

round. Formally, when not intervened upon we assume that each $X_i \sim \text{Bernoulli}(q_i)$ where $\boldsymbol{q} = (q_1, \dots, q_N) \in [0,1]^N$ 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 $X_1, \dots, X_N$ 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** $t \in 1, \dots, T/2$ **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** $t \in 1, \dots, T_A$ **do**
        Intervene with $a$ and observe $Y_t$
    Re-estimate $\hat{\mu}_a = \frac{1}{T_A} \sum_{t=1}^{T_A} Y_t$
**return** estimated optimal $\hat{a}^*_T \in \arg\max_{a \in \mathcal{A}} \hat{\mu}_a$
```

I_τ is the set of variables considered unbalanced and we tune τ to trade off identifying the low probability actions against not having too many of them, so as to minimise the worst-case simple regret. When $\boldsymbol{q} = (\frac{1}{2}, \ldots, \frac{1}{2})$ we have $m(\boldsymbol{q}) = 2$ and when $\boldsymbol{q} = (0, \ldots, 0)$ we have $m(\boldsymbol{q}) = N$. We do not assume that $\boldsymbol{q}$ is known, thus Algorithm 2 also utilises the samples captured during the observational phase to estimate $m(\boldsymbol{q})$. Although very simple, the following two theorems show that this algorithm is effectively optimal.

Theorem 19. *Algorithm 2 satisfies*

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

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

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

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

5.2.2 General graphs

We now consider the more general problem where the graph structure is known, but arbitrary. For general graphs, $\mathrm{P}\left(Y|X_i = j\right) \neq \mathrm{P}\left(Y|do(X_i = j)\right)$ (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 factorisation formula [116].

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

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 factorisation 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 5.3c and suppose each variable deterministically takes the value of its parent, $X_k = X_{k-1}$ for $k \in 2, \ldots, N$ and $\mathrm{P}\left(X_1\right) = 0$. 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 5.3b, $\mathrm{P}\left(Y|do(X_2 = j)\right) = \sum_{X_1} \mathrm{P}\left(X_1\right)\mathrm{P}\left(Y|X_1, X_2 = j\right)$. 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 $Q = \sum_{a \in \mathcal{A}} \eta_a \operatorname{P}(\mathcal{P}\mathrm{a}_Y \,|a)$ 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 $t \in \{1,\dots,T\}$ 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 $\mathcal{P}\mathrm{a}_Y(X)$ denote the realisation of the variables in X that are parents of Y and define $R_a(X) = \frac{\operatorname{P}\{\mathcal{P}\mathrm{a}_Y(X)|a\}}{\operatorname{Q}(\mathcal{P}\mathrm{a}_Y(X))}$

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

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

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

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

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 5.2.5.

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

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

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

Since the mixture of convex functions is convex and the maximum of a set of convex functions is convex, we see that $m(\eta)$ is convex (in η). Therefore the minimisation problem may be tackled using standard techniques from convex optimisation. The quantity $m(\eta^*)$ may be interpreted as the minimum achievable worst-case variance of the importance weighted estimator. In the experimental section we present some special cases, but for now we give two simple results. The first shows that $|\mathcal{A}|$ serves as an upper bound on $m(\eta^*)$.
Proposition 22. $m(\eta^*) \le |\mathcal{A}|$. *Proof.* By definition, $m(\eta^*) \le m(\eta)$ for all η. Let $\eta_a = 1/|\mathcal{A}|\,\forall a$.

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

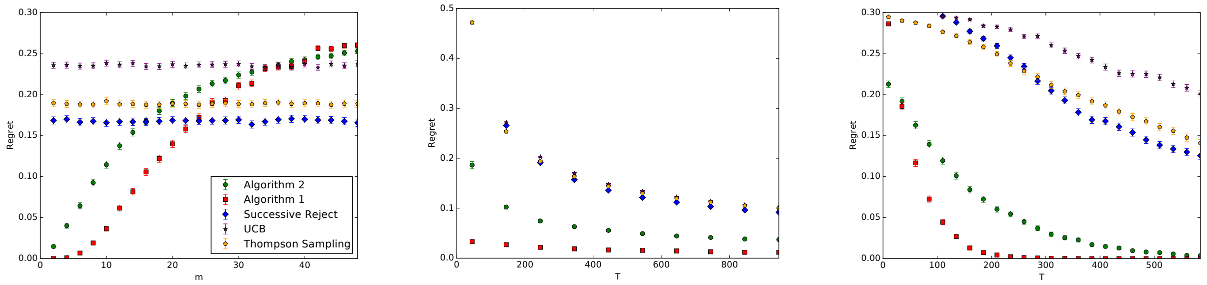
The second observation is that, in the parallel bandit setting, $m(\eta^*) \le 2m(\boldsymbol{q})$. This is easy to see by letting $\eta_a = 1/2$ for $a = do()$ and $\eta_a = \mathbb{1}\{\mathrm{P}\left(X_i = j\right) \le 1/m(\boldsymbol{q})\}/2m(\boldsymbol{q})$ for the actions corresponding to $do(X_i = j)$, and applying an argument like that for Proposition 22. The proof is in section 5.2.5.
Remark 23. The choice of B_a given in Theorem 21 is not the only possibility. As we shall see in the experiments, it is often possible to choose B_a significantly larger when there is no heavy tail and this can drastically improve performance by eliminating the bias. This is especially true when the ratio R_a is never too large and Bernstein's inequality could be used directly without the truncation. For another discussion see the article by Bottou et al. [28] who also use importance weighted estimators to learn from observational data.

5.2.3 Experiments

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

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

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



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

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

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

Figure 5.4: Experimental results

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

For the parallel bandit problem, the regression estimator used in the specific algorithm outperforms the truncated importance weighted estimator in the more general algorithm, despite the fact the specific algorithm must estimate $\boldsymbol{q}$ from the data. This is an interesting phenomenon that has been noted before in off-policy evaluation where the regression (and not the importance weighted) estimator is known to be mini-max optimal asymptotically [105].

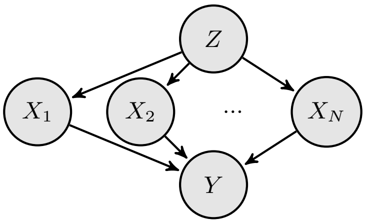


Figure 5.5: Confounded graph

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

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

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

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

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

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

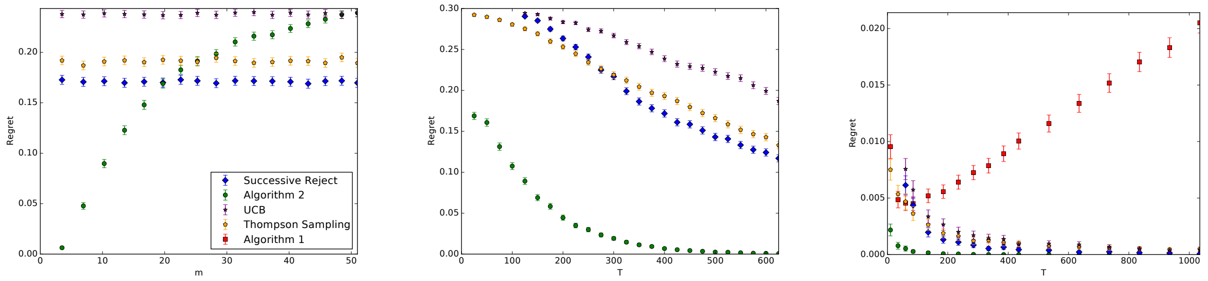
As in the parallel bandit case, we let Y depend only on X_1, $P(Y|do(X_1=1)) = \frac{1}{2}+\varepsilon$ and $P(Y|do(X_1=0)) = \frac{1}{2}-\varepsilon'$, where $\varepsilon' = \varepsilon P(X_1=1)/P(X_1=0)$. The value of N_1 determines m and ranges between 2 and N. The values for the CPD's have been chosen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

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

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

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

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



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

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

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

Figure 5.6: Experimental results on the confounded graph

5.2.4 Discussion & Future work

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

The main practical drawback of Algorithm 3 is that both the estimator $\hat{\mu}_a$ and the optimal sampling distribution η^* (*i.e.*, the one that minimises $m(\eta)$) require knowledge of the conditional distributions $\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y | a\right\}$ 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 5.4a.

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

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

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

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

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

Partially or Completely Unknown Causal Graph A much more difficult generalisation would be to consider causal bandit problems where the causal graph is completely unknown or known to be a member of class of models. The latter case arises naturally if we assume free access to a large observational data set, from which the Markov equivalence class can be found via causal discovery techniques. Work on the problem of selecting experiments to discover the correct causal graph from within a Markov equivalence class [54, 53, 70, 79] could potentially be incorporated into a causal bandit algorithm. In particular, Hu and Vetta [79] show that only $\mathcal{O}(\log\log n)$ multi-variable interventions are required on average to recover a causal graph over n variables once purely observational data is used to recover the "essential graph". Simultaneously learning a completely unknown causal model while estimating the rewards of interventions without a large observational data set would be much more challenging.

5.2.5 Proofs

Proof of Theorem 19

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

The proof of Theorem 19 requires some lemmas.

Lemma 24. *Let $i \in \{1, \ldots, N\}$ 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 [69]),

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

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

□

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

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

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

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

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

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

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

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

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

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

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

Via Hoeffding's inequality and the union bound,

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

For arms not in A,

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

Again applying Hoeffding's and the union bound

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

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

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

If this occurs, then

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

Therefore

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

which completes the result. □

Proof of Theorem 20

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

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

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

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

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 $i \leq m$ we have $q_i \leq 1/m$ and the KL divergence between P_0 and P_i may be bounded using the telescoping property (chain rule) and by bounding the local KL divergence by the χ-squared distance as by Auer et al. [16]. This leads to

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

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

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

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

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

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

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

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

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

as required.

Proof of Theorem 21

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

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

Checking the expectation we have

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

where

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

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

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

Let $\varepsilon_a > 0$ be given by

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

Then by the union bound and Bernstein's inequality

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

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

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

and

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

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

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

Therefore

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

as required. □

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

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

Proof. Recall that in the parallel bandit setting,

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

Let:

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

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

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

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

Recall that,

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

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

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

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

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

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

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

□

Bibliography

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[Buccapatnam et al.] Buccapatnam, S., Eryilmaz, A., and Shroff Ness, B. Stochastic Bandits with Side Observations on Networks. *ACM SIGMETRICS'14, June 2014, Austin, Texas.*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[51] Dud\'\ik, M., Langford, J., Li, L., Dudik, M., Langford, J., and Li, L. (2011). Doubly Robust Policy Evaluation and Learning.

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

[53] Eberhardt, F. (2010). Causal Discovery as a Game. In *NIPS 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 *UAI*.

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

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

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

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

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

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

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

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

[63] Garivier, A. and Moulines, E. (2008). On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv:0805.3415*.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[79] Hu, H. and Vetta, A. (2014). Randomized Experimental Design for Causal Graph Discovery. In *NIPS*, pages 1–9.

[80] 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.

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

[82] Imbens, G. W. (2004). NONPARAMETRIC ESTIMATION OF AVERAGE TREATMENT EFFECTS UNDER EXOGENEITY : A REVIEW *. *Review of Economics and statistics*, 86(February):4–29.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[149] Sokolova, E., Hoogman, M., Groot, P., Claassen, T., Vasquez, A. A., Buitelaar, J. K., Franke, B., and Heskes, T. (2015). Causal discovery in an adult ADHD data set suggests indirect link between <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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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