Causal Bandits: Learning Good Interventions via Causal Inference

Finnian Lattimore
Australian National University and Data61/NICTA
finn.lattimore@gmail.com

Tor Lattimore
University of Alberta
tor.lattimore@gmail.com

Mark D. Reid
Australian National University and Data61/NICTA
mark.reid@anu.edu.au

August 23, 2017

Abstract

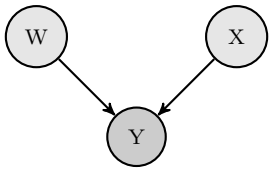
We study the problem of using causal models to improve the rate at which good interventions can be learned online in a stochastic environment. Our formalism combines multi-arm bandits and causal inference to model a novel type of bandit feedback that is not exploited by existing approaches. We propose a new algorithm that exploits the causal feedback and prove a bound on its simple regret that is strictly better (in all quantities) than algorithms that do not use the additional causal information. This is an extended version of our paper published in NIPS 2016, Lattimore et al. [24].

1 Introduction

Medical drug testing, policy setting, and other scientific processes are commonly framed and analysed in the language of sequential experimental design and, in special cases, as bandit problems [29, 12]. In this framework, single actions (also referred to as interventions) from a pre-determined set are repeatedly performed in order to evaluate their effectiveness via feedback from a single, real-valued reward signal. We propose a generalisation of the standard model by assuming that, in addition to the reward signal, the learner observes the values of a number of covariates drawn from a probabilistic causal model [28]. Causal models are commonly used in disciplines where explicit experimentation may be difficult such as social science, demography and economics. For example, when predicting the effect of changes to childcare subsidies on workforce participation, or school choice on grades. Results from causal inference relate observational distributions to interventional ones, allowing the outcome of an intervention to be predicted without explicitly performing it. By exploiting the causal information we show, theoretically and empirically, how non-interventional observations can be used to improve the rate at which high-reward actions can be identified.

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 fertilizer; 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 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?

Contributions We take the first step towards formalising and solving problems such as the one above. In §2 we formally introduce *causal bandit problems* in which interventions are treated as arms in a bandit



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

Figure 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.

problem but their influence on the reward — along with any other observations — is assumed to conform to a known causal graph. We show that our causal bandit framework subsumes the classical bandits (no additional observations) and contextual stochastic bandit problems (observations are revealed before an intervention is chosen) before focusing on the case where, like the above example, observations occur *after* each intervention is made.

Our focus is on the simple regret, which measures the difference between the return of the optimal action and that of the action chosen by the algorithm after T rounds. In §3.1 we analyse a specific family of causal bandit problems that we call *parallel bandit* problems in which N factors affect the reward independently and there are $2N$ possible interventions. We propose a simple causal best arm identification algorithm for this problem and show that up to logarithmic factors it enjoys minimax optimal simple regret guarantees of $\tilde{\Theta}(\sqrt{m/T})$ where m depends on the causal model and may be much smaller than N. In contrast, existing best arm identification algorithms suffer $\Omega(\sqrt{N/T})$ simple regret (Thm. 4 by **?**]). This shows theoretically the value of our framework over the traditional bandit problem. Experiments in §**??** further demonstrate the value of causal models in this framework.

In the general casual bandit problem interventions and observations may have a complex relationship. In §3.2 we propose a new algorithm inspired by importance-sampling that a) enjoys sub-linear regret equivalent to the optimal rate in the parallel bandit setting and b) captures many of the intricacies of sharing information in a causal graph in the general case. As in the parallel bandit case, the regret guarantee scales like $O(\sqrt{m/T})$ where m depends on the underlying causal structure, with smaller values corresponding to structures that are easier to learn. The value of m is always less than the number of interventions N and in the special case of the parallel bandit (where we have lower bounds) the notions are equivalent.

2 Problem Setting

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 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 paper we focus on the case where the causal graph is known. Extending this work to simultaneously learning the casual graph is discussed in §3.4.

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

expected reward for the action $a = do(\boldsymbol{X} = \boldsymbol{x})$ by $\mu_a := \mathbb{E}[Y|do(\boldsymbol{X} = \boldsymbol{x})]$ and the optimal expected reward by $\mu^* := \max_{a \in \mathcal{A}} \mu_a$.

Definition 1 (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 $\mathcal{X}$ and incurs a known cost $C(a)$ on the learner. One variable $Y \in \mathcal{X}$ is designated as the *reward variable*.

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

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

We refer to the set of variables that can be observed prior to selecting an action $\boldsymbol{X}^c$ as contextual variables and the set of variables observed after the action is chosen, $\boldsymbol{X}^o$, as post-action feedback variables. Note that $\boldsymbol{X}^c$ and $\boldsymbol{X}^o$ need not be disjoint. A variable may be observed both prior to and after the agent selects an action, and the action may change its value. The notation $\mathrm{P}(\cdot|\boldsymbol{X}_t^c, do(\boldsymbol{X}_t = \boldsymbol{x}_t))$ denotes distributions conditional on having observed $\boldsymbol{X}_t^c$ and *then* intervened to set $\boldsymbol{X}_t = \boldsymbol{x}_t$. The values of variables in $\boldsymbol{X}_t^c$ that are non-decendents of $\boldsymbol{X}_t$ remain unchanged by the intervention. The objective of the learner is to minimise either the simple or cumulative regret.

The causal bandit problem takes on characteristics of different bandit settings depending on the action-space $\mathcal{A}$ and correspondign costs, which variables are observable prior to selecting an action and on which variables we receive post-action feedback. If we can (cheaply) intervene on all of the parents of Y simultaneusly then any context or alternative actions are irrelevent and the problem reduces to a standard multi-armed bandit problem. This is formalised in theorem 2.

Theorem 2. *Let $\mathcal{A}'$ be the set of all possible assignments of values to the parents of Y. If $\mathcal{A}' \subseteq \mathcal{A}$ and $C(a') \leq C(a)\ \forall (a' \in \mathcal{A}', a \in \mathcal{A}/\mathcal{A}')$ then the optimal action $a^* \in \mathcal{A}'$ and the problem reduces to a standard multi-armed bandit (over actions in in $\mathcal{A}'$).*

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

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

□

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 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 1. The causal structure implies:

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

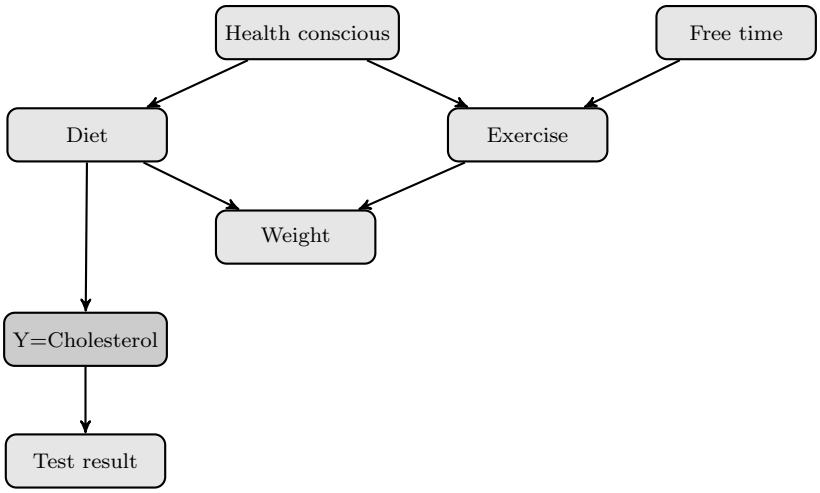


Figure 2: Example causal graph (based on Koller and Friedman [23]) where the outcome of interest (reward) is cholesterol level . The do-calculus can be applied to eliminate some actions immediately without the need to do any experiments. For example, no actions involving 'Test Result' need to be considered and interventions on 'Diet' do not need to be considered in conjunction with any other variables.

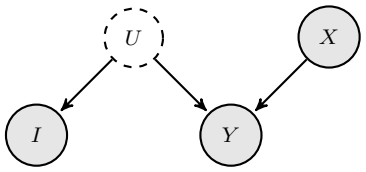
Thus we gain information about the reward for the action $do(W = 0)$ from selecting the action $do()$ or $do(X = x)$ and then observing $W = 0$. We only get this form of side information for actions that don't specify the value of every variable, for example those in the bottom half of the table in figure 1. If additional variables are only observed before an intervention is selected the causal bandit problem can be treated as a stochastic contextual bandit problem, which are already reasonably well understood [1].

Two other problems that sit in the space between causal inference and bandit problems, bandits with unobserved confounders [7] and compliance aware bandits [13], can also be viewed as specific causal bandit problems. In the work on bandits with unobserved confounders it is assumed that the reward given each action may depend on some latent variable U which we cannot observe directly. However, prior to selecting an action, we can observe I, the action that would have been selected under an alternate (stationary) policy, which may depend on U, see figure 3a. In this case the set of contextual variables $\boldsymbol{X}^c = I$, the set of post-action feedback variables $\boldsymbol{X}^o = \{Y\}$ and the action space consists of all possible assigments of values to a single node X, $\mathcal{A} = do(X = x)$. This setting reduces to a contextual bandit problem in our causal bandit framework. However in thier work on bandits with unobserved confounders Forney and Bareinboim [17] also leverage the fact that I represents the action that would have been selected under an alternate policy to fuse data collected under the previous (observational) policy with data collected under the new policy. This information is not encoded in the causal bandit graph in figure 3a, as I could be any variable that is influenced by the unobserved context U, and thus cannot be exploited by a standard contextual bandit algorithm.

Compliance aware bandits describe situations in which the action recommended by the bandit algorithm is not always followed. For example a patient may refuse to take a treatment or an advertiser may have complex rules about how many ads a given customer can recieve which prevends some of the suggestions from the ad recomendation engine from being followed. After an action is selected the algorithm can observe the action that was actually taken in addition to the reward. Della Penna et al. [13] analyse this setting with binary treatments both with and without the presense of a latent confounding variable U, see figure 3b. In this case, there are no contextual variables and the action space is again the set of assignments to a single variable but there is post-action feedback, which reveals the value of the action that was actually taken.[1]

[1]There are some interesting variants of the compliance aware bandit setting that, to my knowledge, have not been analysed. The first is if the confounding variable U is observable, either as context or post-action feedback. The second is if we extend the allowable action set to include acting directly on X, albeit at a higher cost than acting on A. It is also worth noting the connection between this setting and instrumental variables [6]. By making some functional assumptions about the relationships between the variables, we can use A as an instrumental variable to bound or estimate the (casual) effect of X on Y. The

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



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

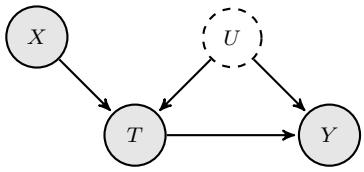


Figure 3

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

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 [21]. However, as we show in §3.1, ignoring the extra information available in the non-intervened variables yields sub-optimal performance.

Our framework bears a superficial similarity to contextual bandit problems, 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. [2], Kocák et al. [22] have considered very general models related to partial monitoring games [8] where rewards on un-played actions are revealed according to a feedback graph. As we discuss in §3.4, 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 [10]. The partial monitoring approach taken by Wu et al. [32] 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 [33] 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. [5] 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. [7] 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 [27] 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 exploit-

estimation will be somewhat complicated in the bandit setting because the action chosen at each timestep is dependent on the previous sequence actions and rewards.

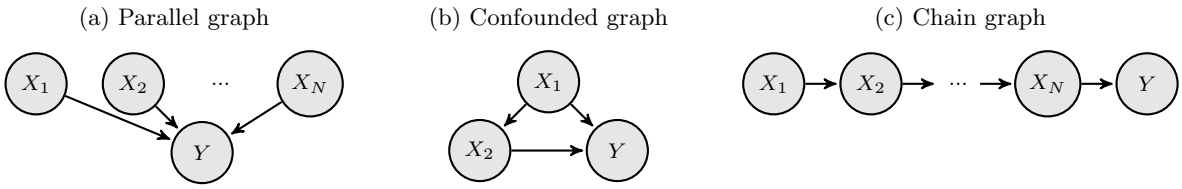


Figure 4: Causal Models

ing 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 §3.2 have been studied before in a causal framework by Bottou et al. [9], 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.

3 Causal bandits with post action feedback

We now focus on causal bandit problems with post-action feedback, in which the value of all the variables are observed after an intervention is selected, the cost of all allowable actions is equal and the goal of the learner is to minimise the simple regret.

3.1 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 §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 4a. All variables are observable and the set of allowable actions are all size 0 and size 1 interventions: $\mathcal{A} = \{do()\} \cup \{do(X_i = j) \colon 1 \leq i \leq N \text{ and } j \in \{0, 1\}\}$

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

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

The Parallel Bandit Algorithm The algorithm operates as follows. For the first $T/2$ rounds it chooses $do()$ to collect observational data. As the only link from each $X_1, \ldots, X_N$ to Y is a direct, causal one, $\mathrm{P}(Y | do(X_i = j)) = \mathrm{P}(Y | X_i = j)$. Thus we can create good estimators for the returns of the actions $do(X_i = j)$ for which $\mathrm{P}(X_i = j)$ is large. The actions for which $\mathrm{P}(X_i = j)$ is small may not be observed (often) so estimates of their returns could be poor. To address this, the remaining $T/2$ rounds are evenly split to

estimate the rewards for these infrequently observed actions. The difficulty of the problem depends on $\boldsymbol{q}$ and, in particular, how many of the variables are unbalanced (*i.e.*, small q_i or $(1-q_i)$). For $\tau \in [2...N]$ let $I_\tau = \left\{i : \min\{q_i, 1-q_i\} < \frac{1}{\tau}\right\}$. Define

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

Algorithm 1 Parallel Bandit Algorithm

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

I_τ is the set of variables considered unbalanced and we tune τ to trade off identifying the low probability actions against not having too many of them, so as to minimise the worst-case simple regret. When $\boldsymbol{q} = (\frac{1}{2}, \ldots, \frac{1}{2})$ we have $m(\boldsymbol{q}) = 2$ and when $\boldsymbol{q} = (0, \ldots, 0)$ we have $m(\boldsymbol{q}) = N$. We do not assume that $\boldsymbol{q}$ is known, thus Algorithm 1 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 3. *Algorithm 1 satisfies*

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

Theorem 4. *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 3 and 4 follow by carefully analysing the concentration of $\hat{p}_a$ and $\hat{m}$ about their true values and may be found in Sections 3.5.1 and 3.5.2 respectively.

By utilising knowledge of the causal structure, Algorithm 1 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.

3.2 General graphs

We now consider the more general problem where the graph structure is known, but arbitrary. For general graphs, $\mathrm{P}(Y|X_i = j) \neq \mathrm{P}(Y|do(X_i = j))$ (correlation is not causation). However, if all the variables

are observable, any causal distribution $\mathrm{P}\left(X_1 ... X_N | do(X_i = j)\right)$ can be expressed in terms of observational distributions via the truncated factorisation formula [28].

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

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 4c and suppose each variable deterministically takes the value of its parent, $X_k = X_{k-1}$ for $k \in 2, \ldots, N$ and $\mathrm{P}\left(X_1\right) = 0$. We can learn the reward for all the interventions $do(X_i = 1)$ simultaneously by selecting $do(X_1 = 1)$, but not from $do()$. In addition, variance of the observational estimator for $a = do(X_i = j)$ can be high even if $\mathrm{P}\left(X_i = j\right)$ is large. Given the causal graph in Figure 4b, $\mathrm{P}\left(Y|do(X_2 = j)\right) = \sum_{X_1} \mathrm{P}\left(X_1\right) \mathrm{P}\left(Y|X_1, X_2 = j\right)$. Suppose $X_2 = X_1$ deterministically, no matter how large $\mathrm{P}\left(X_2 = 1\right)$ is we will never observe $(X_2 = 1, X_1 = 0)$ and so cannot get a good estimate for $\mathrm{P}\left(Y|do(X_2 = 1)\right)$.

To solve the general problem we need an estimator for each action that incorporates information obtained from every other action and a way to optimally allocate samples to actions. To address this difficult problem, we assume the conditional interventional distributions $\mathrm{P}\left(\mathcal{P}\mathrm{a}_Y \, | a\right)$ (but not $\mathrm{P}\left(Y|a\right)$) are known. These could be estimated from experimental data on the same covariates but where the outcome of interest differed, such that Y was not included, or similarly from observational data subject to identifiability constraints. Of course this is a somewhat limiting assumption, but seems like a natural place to start. The challenge of estimating the conditional distributions for all variables in an optimal way is left as an interesting future direction. Let η be a distribution on available interventions $a \in \mathcal{A}$ so $\eta_a \geq 0$ and $\sum_{a \in \mathcal{A}} \eta_a = 1$. Define $Q = \sum_{a \in \mathcal{A}} \eta_a \, \mathrm{P}\left(\mathcal{P}\mathrm{a}_Y \, | a\right)$ to be the mixture distribution over the interventions with respect to η.

Algorithm 2 General Algorithm

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

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

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

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

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

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

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

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

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

Theorem 5. *If Algorithm 2 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 3.5.3.

Note the regret has the same form as that obtained for Algorithm 1, with $m(\eta)$ replacing $m(q)$. Algorithm 1 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 2 with the optimal choice of η is close to mini-max optimal, but leave lower bounds for future work.

Choosing the Sampling Distribution Algorithm 2 depends on a choice of sampling distribution Q that is determined by η. In light of Theorem 5 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 6. $m(\eta^*) \leq |\mathcal{A}|$. *Proof.* By definition, $m(\eta^*) \leq m(\eta)$ for all η. Let $\eta_a = 1/|\mathcal{A}| \, \forall a$.

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

The second observation is that, in the parallel bandit setting, $m(\eta^*) \leq 2m(\boldsymbol{q})$. This is easy to see by letting $\eta_a = 1/2$ for $a = do()$ and $\eta_a = \mathbb{1}\{\mathrm{P}\left(X_i = j\right) \leq 1/m(\boldsymbol{q})\}/2m(\boldsymbol{q})$ for the actions corresponding to $do(X_i = j)$, and applying an argument like that for Proposition 6. The proof is in section 3.5.4.

Remark 7. The choice of B_a given in Theorem 5 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. [9] who also use importance weighted estimators to learn from observational data.

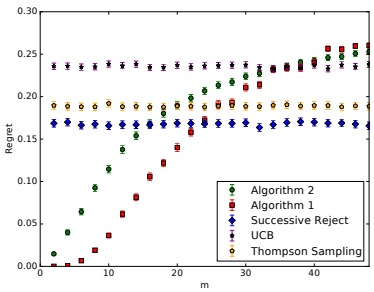
3.3 Experiments

We compare Algorithms 1 and 2 with the Successive Reject algorithm of Audibert and Bubeck [3], Thompson Sampling and UCB under a variety of conditions. Thomson sampling and UCB are optimised to minimise

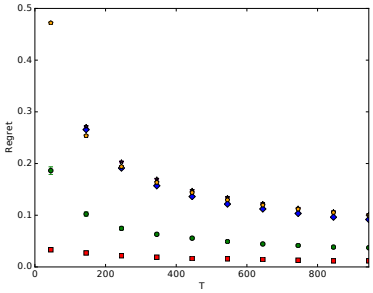
cumulative regret. We apply them in the fixed horizon, best arm identification setting by running them up to horizon T and then selecting the arm with the highest empirical mean. The importance weighted estimator used by Algorithm 2 is not truncated, which is justified in this setting by Remark 7.

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

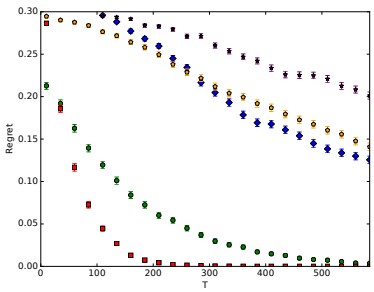
In Figure 5a we fix the number of variables N and the horizon T and compare the performance of the algorithms as m increases. The regret for the Successive Reject algorithm is constant as it depends only on the reward distribution and has no knowledge of the causal structure. For the causal algorithms 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: Experimental results

Figure 5b 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 5c 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 [26].

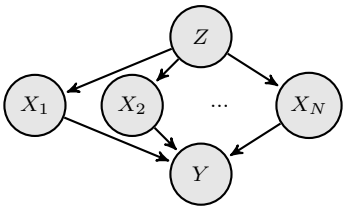


Figure 6: Confounded graph

We now compare the general algorithm with a range of standard bandit algorithms on the confounded graph in Figure 6. 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 7a we fix N and T and $P(Z = 1) = .4$. For some $2 \leq N_1 \leq N$ we define

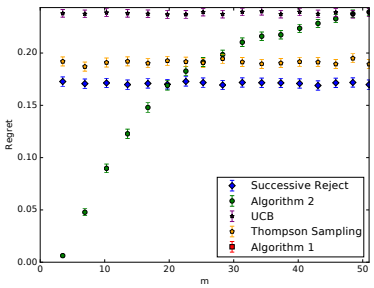
$$P(X_i = 1|Z = 0) = \begin{cases} 0 & \text{if } i \in \{1, ...N_1\} \\ .4 & \text{otherwise} \end{cases}$$
$$P(X_i = 1|Z = 1) = \begin{cases} 0 & \text{if } i \in \{1, ...N_1\} \\ .65 & \text{otherwise} \end{cases}$$

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.

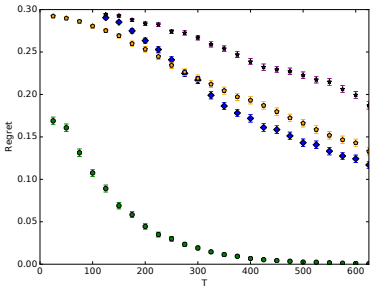
In Figure 7b we fix the model and number of variables, N, and vary the horizon T. $P(Z)$ and $P(X|Z)$ are the same as for the previous experiment. In Figure 7c we additionally show the performance of Algorithm 1, but exclude actions on Z from the set of allowable actions to demonstrate that Algorithm 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:

$$P(X_i = 1|Z = 0) = \begin{cases} .166 & \text{if } i \in \{1, ..., 6\} \\ .2 & \text{if } i = 7 \\ .7 & \text{otherwise} \end{cases}$$
$$P(X_i = 1|Z = 1) = \begin{cases} .166 & \text{if } i \in \{1, ..., 6\} \\ .8 & \text{if } i = 7 \\ .3 & \text{otherwise} \end{cases}$$

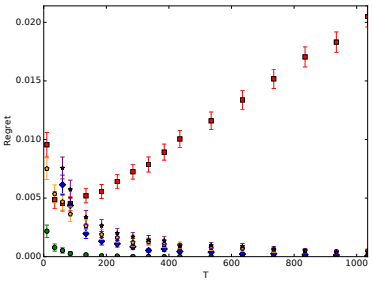
In this setting X_7 tends to agree with Z and X_N tends to disagree. It is sub-optimal to act on either X_7 or X_N, while all other actions are optimal. The first group of X variables with $i \leq 6$ will be identified by the parallel bandit as the most unbalanced ones and played explicitly. All remaining 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 7: Experimental results on the confounded graph

3.4 Discussion & Future work

Algorithm 2 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 5 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 2 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 1 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 4 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 [16]. 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 5a.

For the parallel bandit the required modifications should be quite straightforward. The idea would be to adapt the algorithm to essentially use successive elimination in the second phase so arms are eliminated as soon as they are provably no longer optimal with high probability. In the general case a similar modification is also possible by dividing the budget T into phases and optimising the sampling distribution η, eliminating arms when their confidence intervals are no longer overlapping. This has now been done by Sen et al. [30], leading to the first problem dependent regret bounds for causal bandit problems. Note that these modifications do not improve the mini-max regret, which at least for the parallel bandit is already optimal. For this reason we focused on emphasising the main point that causal structure can and should be exploited when available. Another observation is that Algorithm 2 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 1, 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 [32] 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 [2, 22], 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. [2], Kocák et al. [22] 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 [25] 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 $\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y \,|a\right\}$ our analysis applies unchanged to the case of causal models with non-observable variables. Some of the interventional distributions may be non-identifiable meaning we can not obtain prior estimates for $\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y \,|a\right\}$ from even an infinite amount of observational data. Even if all variables are observable and the graph is known, if the conditional distributions are unknown, then Algorithm 2 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 [15, 14, 19, 20] could potentially be incorporated into a causal bandit algorithm. In particular, Hu et al. [20] show that only $\mathcal{O}\left(\log\log n\right)$ 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.

3.5 Proofs

3.5.1 Proof of Theorem 3

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 3 requires some lemmas.

Lemma 8. *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 [18]),

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

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

□

Lemma 9. *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 8 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 8 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 3. Recall that $A = \{a \in \boldsymbol{\mathcal{A}} : \hat{p}_a \leq 1/m(\hat{\boldsymbol{q}})\}$. Then, for $a \in A$, the algorithm estimates μ_a from $T_A \doteq T/(2m(\hat{\boldsymbol{q}}))$ samples. From lemma 9, $T_A \geq T/(4m(\boldsymbol{q}))$ with probability $(1 - \delta)$. Let H be the event $T_A < T/(4m(\boldsymbol{q}))$ and G be the event $\exists a \in A : |\mu_a - \hat{\mu}_a| \geq \sqrt{\frac{2m(\boldsymbol{q})}{T} \log \frac{2N}{\delta}}$

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

Via Hoeffding's inequality and the union bound,

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

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

For arms not in A,

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

Again applying Hoeffding's and the union bound

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

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

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

If this occurs, then

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

Therefore

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

which completes the result. □

3.5.2 Proof of Theorem 4

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 [31] we have

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

where $\mathrm{KL}(\mathrm{P}_0, \mathrm{P}_i)$ is the KL divergence between measures P_0 and P_i. Let $T_i(T) = \sum_{t=1}^T \mathbb{1}\{a_t = do(X_i = 1)\}$ be the total number of times the learner intervenes on variable i by setting it to 1. Then for $i \leq m$ we have $q_i \leq 1/m$ and the KL divergence between P_0 and P_i may be bounded using the telescoping property (chain rule) and by bounding the local KL divergence by the χ-squared distance as by Auer et al. [4]. This leads to

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

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

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

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

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

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

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

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

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

as required.

3.5.3 Proof of Theorem 5

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

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

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

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

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

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

Then by the union bound and Bernstein's inequality

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

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

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

By the definition of the truncation we have

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

and

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

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

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

Therefore

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

as required. □

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

Proposition 10. *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.

□

References

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

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

[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*.

[12] Chernoff, H. (1959). Sequential design of experiments. *The Annals of Mathematical Statistics*, pages 755–770.

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

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

[15] 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*.

[16] 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.

[17] Forney, A. and Bareinboim, E. (2017). Counterfactual Data-Fusion for Online Reinforcement Learners. In *ICML*, number June.

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

[19] 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.

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

[21] 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.

[22] 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.

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

[24] 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.

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

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

[27] 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.

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

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

[30] Sen, R., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017). Identifying Best Interventions through Online Importance Sampling. pages 1–30.

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

[32] 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.

[33] 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.