Causal Bandits: Learning Good Interventions via Causal Inference

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

Tor Lattimore
Indiana University, Bloomington
tor.lattimore@gmail.com

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

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.

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 (Robbins, 1952; Chernoff, 1959). 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 (Pearl, 2000). 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?

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

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 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 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 Audibert and Bubeck (2010)). This shows theoretically the value of our framework over the traditional bandit problem. Experiments in §5 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 §4 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.

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 (Jamieson et al., 2014). However, as we show in §3, ignoring the extra information available in the non-intervened variables yields sub-optimal performance.

A well-studied class of bandit problems with side information are "contextual bandits" Langford and Zhang (2008); Agarwal et al. (2014). 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. (2015), Kocák et al. (2014) have considered very general models related to partial monitoring games (Bartók et al., 2014) where rewards on unplayed actions are revealed according to a feedback graph. As we discuss in §6, 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 (Bubeck et al., 2009). The partial monitoring approach taken by Wu et al. (2015) 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 (2009) utilize 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. (2012) 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. (2015) 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 (2014) present an analysis and extension of Thompson sampling assuming actions are causal interventions. Their focus is on causal induction (*i.e.*, learning an unknown causal model) instead of exploiting a known causal model. Combining their handling of causal induction with our analysis is left as future work.

The truncated importance weighted estimators used in §4 have been studied before in a causal framework by Bottou et al. (2013), 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.

2 Problem Setup

We now introduce a novel class of stochastic sequential decision problems which we call *causal bandit problems*. In these problems, rewards are given for repeated interventions on a fixed causal model Pearl (2000). Following the terminology and notation in Koller and Friedman (2009), a *causal model* is given by a directed acyclic graph $\mathcal{G}$ over a set of random variables $\mathcal{X} = \{X_1, \ldots, X_N\}$ and a joint distribution P over $\mathcal{X}$ that factorises over $\mathcal{G}$. We will assume each variable only takes on a finite number of distinct values. An edge from variable X_i to X_j is interpreted to mean that a change in the value of X_i may directly cause a change to the value of X_j. 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 $do()$. The intervention also "mutilates" the graph $\mathcal{G}$ by removing all edges from $\mathcal{P}\mathrm{a}_i$ to X_i for each $X_i \in \boldsymbol{X}$. The resulting graph defines a probability distribution $\mathrm{P}\left\{\boldsymbol{X}^c | do(\boldsymbol{X} = \boldsymbol{x})\right\}$ over $\boldsymbol{X}^c := \mathcal{X} - \boldsymbol{X}$. Details can be found in Chapter 21 of Koller and Friedman (2009).

A learner for a casual bandit problem is given the casual model's graph $\mathcal{G}$ and a set of *allowed actions* $\mathcal{A}$. One variable $Y \in \mathcal{X}$ is designated as the *reward variable* and takes on values in $\{0, 1\}$. We denote the expected reward for the action $a = do(\boldsymbol{X} = \boldsymbol{x})$ by $\mu_a := \mathbb{E}\left[Y | do(\boldsymbol{X} = \boldsymbol{x})\right]$ and the optimal expected reward by $\mu^* := \max_{a \in \mathcal{A}} \mu_a$. The causal bandit game proceeds over T rounds. In round t, the learner *intervenes* by choosing $a_t = do(\boldsymbol{X}_t = \boldsymbol{x}_t) \in \mathcal{A}$ based on previous observations. It then *observes* sampled values for all non-intervened variables $\boldsymbol{X}_t^c$ drawn from $\mathrm{P}\left\{\boldsymbol{X}_t^c | do(\boldsymbol{X}_t = \boldsymbol{x}_t)\right\}$, including the *reward* $Y_t \in \{0, 1\}$. After T observations the learner outputs an estimate of the optimal action $\hat{a}_T^* \in \mathcal{A}$ based on its prior observations.

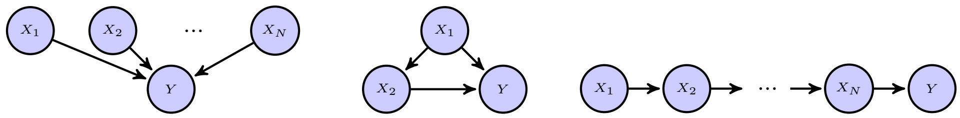
The objective of the learner is to minimise the simple regret $R_T = \mu^* - \mathbb{E}\left[\mu_{\hat{a}_T^*}\right]$. This is sometimes refered to as a "pure exploration" (Bubeck et al., 2009) or "best-arm identification" problem (Gabillon et al., 2012) and is most appropriate when, as in drug and policy testing, the learner has a fixed experimental budget after which its policy will be fixed indefinitely.

Although we will focus on the intervene-then-observe ordering of events within each round, other scenarios are possible. If the non-intervened variables are observed before an intervention is selected our framework reduces to stochastic contextual bandits, which are already reasonably well understood (Agarwal et al., 2014). Even if no observations are made during the rounds, the causal model may still allow offline pruning of the set of allowable interventions thereby reducing the complexity.

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. Intuitively though, one would expect to perform better by making use of the extra structure and observations.

3 Regret Bounds for Parallel Bandit

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



(a) Parallel graph (b) Confounded graph (c) Chain graph

Figure 1: Causal Models

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

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

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 minimize 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 utilizes 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.

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 $A = \left\{a \in \mathcal{A} \colon \hat{p}_a \le \frac{1}{\hat{m}}\right\}$.
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 $\hat{a}_T^* \in \arg\max_{a \in \mathcal{A}} \hat{\mu}_a$

Theorem 1. *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 2. *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 1 and 2 follow by carefully analysing the concentration of $\hat{p}_a$ and $\hat{m}$ about their true values and may be found in the supplementary material. By utilizing 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.

4 Regret Bounds for General Graphs

We now consider the more general problem where the graph structure is known, but arbitrary. For general graphs, $\mathrm{P}\{Y|X_i = j\} \neq \mathrm{P}\{Y|do(X_i = j)\}$ (correlation is not causation). However, if all the variables are observable, any causal distribution $\mathrm{P}\{X_1...X_N|do(X_i = j)\}$ can be expressed in terms of observational distributions via the truncated factorization formula (Pearl, 2000).

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

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

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

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

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 realization 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.

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

The distribution over actions, η plays the role of allocating samples to actions and is optimized to minimize 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 3. *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 the supplementary materials. 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 minimax 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 3 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 4. $m(\eta^*) \leq |\mathcal{A}|$. *Proof.* By definition, $m(\eta^*) \leq m(\eta)$ for all η. Let $\eta_a = 1/|\mathcal{A}| \, \forall a$.

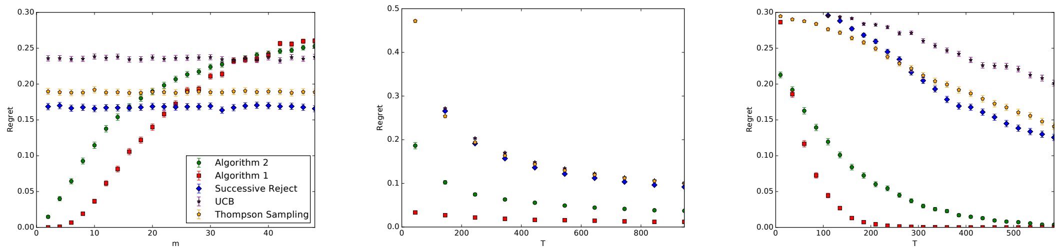
$$m(\eta) = \max_a \mathbb{E}_a \left[\frac{\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}{\mathrm{Q}\left\{\mathcal{P}\mathrm{a}_Y(X)\right\}} \right] \leq \max_a \mathbb{E}_a \left[\frac{\mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}{\eta_a \, \mathrm{P}\left\{\mathcal{P}\mathrm{a}_Y(X)|a\right\}}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

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 4. The proof is in the supplementary materials.

Remark 5. The choice of B_a given in Theorem 3 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. (2013) who also use importance weighted estimators to learn from observational data.

5 Experiments

We compare Algorithms 1 and 2 with the Successive Reject algorithm of Audibert and Bubeck (2010), Thompson Sampling and UCB under a variety of conditions. Thomson sampling and UCB are optimized to minimize cumulative regret. We apply them in the fixed horizon, best arm identification setting by running them upto 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 5.



(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 2: Experimental results

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 2a 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.

Figure 2b demonstrates the performance of the algorithms in the worst case environment for standard bandits, where the gap between the optimal and sub-optimal arms, $\varepsilon = \sqrt{N/(8T)}$, is just too small to be learned. This gap is learnable by the causal algorithms, for which the worst case ε depends on $m \ll N$. In Figure 2c 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 minimax optimal asymptotically (Li et al., 2014).

6 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 3 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 $\mathrm{P}\{\mathcal{P}\mathrm{a}_Y | a\}$ 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 2 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 (Even-Dar et al., 2002). 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 sucessive rejects algorithm for large m, as seen in Figure 2a. 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 minimax regret, which at least for the parallel bandit is already optimal. For this reason we prefer to emphasize the main point that causal structure 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 (Wu et al., 2015) 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 (Alon et al., 2015; Kocák et al., 2014), where at each timestep the feedback graph G_t is selected stochastically, dependent on $\boldsymbol{q}$, and revealed after an action has been chosen. The feedback graph is distinct from the causal graph. A link $A \to B$ in G_t indicates that selecting the action A reveals the reward for action B. For this parallel bandit problem, G_t will always be a star graph with the action $do()$ connected to half the remaining actions. However, Alon et al. (2015); Kocák et al. (2014) give adversarial algorithms, which when applied to the parallel bandit problem obtain the standard bandit regret. A malicious adversary can select the same graph each time, such that the rewards for half the arms are never revealed by the informative action. This is equivalent to a nominally stochastic selection of feedback graph where $\boldsymbol{q} = \boldsymbol{0}$.

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 dataset, 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 Eberhardt et al. (2005); Eberhardt (2010); Hauser and Bühlmann (2014); Hu et al. (2014) could potentially be incorporated into a causal bandit algorithm. In particular, Hu et al. (2014) 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 dataset would be much more challenging.

References

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 *ICML*, pages 1638–1646.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits. In *COLT*, pages 23–35.

Audibert, J.-Y. and Bubeck, S. (2010). Best arm identification in multi-armed bandits. In *COLT*, pages 13–p.

Avner, O., Mannor, S., and Shamir, O. (2012). Decoupling exploration and exploitation in multi-armed bandits. In *ICML*, pages 409–416.

Bareinboim, E., Forney, A., and Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. In *NIPS*, pages 1342–1350.

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.

Bottou, L., Peters, J., Quinonero-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. *JMLR*, 14(1):3207–3260.

Bubeck, S., Munos, R., and Stoltz, G. (2009). Pure exploration in multi-armed bandits problems. In *ALT*, pages 23–37.

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

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

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

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.

Gabillon, V., Ghavamzadeh, M., and Lazaric, A. (2012). Best arm identification: A unified approach to fixed budget and fixed confidence. In *NIPS*, pages 3212–3220.

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.

Hu, H., Li, Z., and Vetta, A. R. (2014). Randomized experimental design for causal graph discovery. In *NIPS*, pages 2339–2347.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *COLT*, pages 423–439.

Kocák, T., Neu, G., Valko, M., and Munos, R. (2014). Efficient learning by implicit exploration in bandit problems with side observations. In *NIPS*, pages 613–621.

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

Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *NIPS*, pages 817–824.

Li, L., Munos, R., and Szepesvari, C. (2014). On minimax optimal offline policy evaluation. *arXiv preprint arXiv:1409.3653*.

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.

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

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

Wu, Y., György, A., and Szepesvári, C. (2015). Online Learning with Gaussian Payoffs and Side Observations. In *NIPS*, pages 1360–1368.

Yu, J. Y. and Mannor, S. (2009). Piecewise-stationary bandit problems with side observations. In *ICML*, pages 1177–1184.