

---

# Causal Bandits: Learning Good Interventions via Causal Inference

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 We study the problem of using causal models to improve the rate at which good  
2 interventions can be learned online in a stochastic environment. Our formalism  
3 combines multi-arm bandits and causal inference to model a novel type of bandit  
4 feedback that is not exploited by existing approaches. We propose a new algorithm  
5 that exploits the causal feedback and prove a bound on its simple regret that is  
6 strictly better (in all quantities) than algorithms that do not use the additional causal  
7 information.

## 8 1 Introduction

9 Medical drug testing, policy setting, and other scientific processes are commonly framed and analysed  
10 in the language of sequential experimental design and, in special cases, as bandit problems (Robbins,  
11 1952; Chernoff, 1959). In this framework, single actions (also referred to as interventions) from a  
12 pre-determined set are repeatedly performed in order to evaluate their effectiveness via feedback from  
13 a single, real-valued reward signal. We propose a generalisation of the standard model by assuming  
14 that, in addition to the reward signal, the learner observes the values of a number of covariates drawn  
15 from a probabilistic causal model (Pearl, 2000). Causal models are commonly used in disciplines  
16 where explicit experimentation may be difficult such as social science, demography and economics.  
17 For example, when predicting the effect of changes to childcare subsidies on workforce participation,  
18 or school choice on grades. Results from causal inference relate observational distributions to  
19 interventional ones, allowing the outcome of an intervention to be predicted without explicitly  
20 performing it. By exploiting the causal information we show, theoretically and empirically, how  
21 non-interventional observations can be used to improve the rate at which high-reward actions can be  
22 identified.

23 The type of problem we are concerned with is best illustrated with an example. Consider a farmer  
24 wishing to optimise the yield of her crop. She knows that crop yield is only affected by temperature,  
25 a particular soil nutrient, and moisture level but the precise effect of their combination is unknown.  
26 In each season the farmer has enough time and money to intervene and control at most one of these  
27 variables: deploying shade or heat lamps will set the temperature to be low or high; the nutrient  
28 can be added or removed through a choice of fertilizer; and irrigation or rain-proof covers will keep  
29 the soil wet or dry. When not intervened upon, the temperature, soil, and moisture vary naturally  
30 from season to season due to weather conditions and these are all observed along with the final crop  
31 yield at the end of each season. How might the farmer best experiment to identify the single, highest  
32 yielding intervention in a limited number of seasons?

33 **Contributions** We take the first step towards formalising and solving problems such as the one  
34 above. In §2 we formally introduce *causal bandit problems* in which interventions are treated as  
35 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 causal 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 discusses 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  $\mathcal{X} = \{X_1, \dots, X_N\}$  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  $\text{Pa}_{X_i}$ , 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  $do(\mathbf{X} = \mathbf{x})$ , assigns the values  $\mathbf{x} = \{x_1, \dots, x_n\}$  to the corresponding variables  $\mathbf{X} = \{X_1, \dots, X_n\} \subset \mathcal{X}$  with the empty intervention (where no variable is set) denoted  $do()$ . The intervention also “mutilates” the graph  $\mathcal{G}$  by removing all edges from  $\text{Pa}_{X_i}$  to  $X_i$  for each  $X_i \in \mathbf{X}$ . The resulting graph defines a probability distribution  $P\{\mathbf{X}^c | do(\mathbf{X} = \mathbf{x})\}$  over  $\mathbf{X}^c := \mathcal{X} - \mathbf{X}$ . 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  $a = do(\mathbf{X} = \mathbf{x})$  by  $\mu_a := \mathbb{E}[Y | do(\mathbf{X} = \mathbf{x})]$  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  $a_t = do(\mathbf{X}_t = \mathbf{x}_t) \in \mathcal{A}$  based on previous observations. It then *observes* sampled values for all non-intervened variables  $\mathbf{X}_t^c$  drawn from  $P\{\mathbf{X}_t^c | do(\mathbf{X}_t = \mathbf{x}_t)\}$ , 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  $R_T = \mu^* - \mathbb{E}[\mu_{\hat{a}_T^*}]$ . This is sometimes referred 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  $P\{Y = 1 | X\} = r(X)$  for some arbitrary but unknown, real-valued function  $r$ . The set of allowed actions in this case is  $\mathcal{A} = \{do(X = k) : k \in \{1, \dots, K\}\}$ . 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 the farming example. It also suffices to witness the regret gap between algorithms that make use of causal models and those which do not.

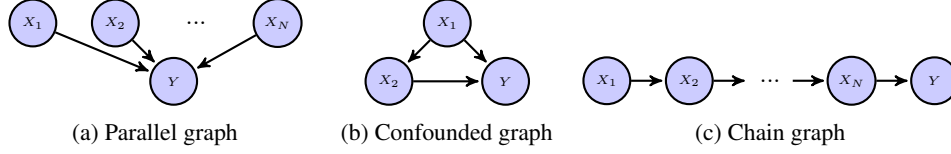


Figure 1: Causal Models

The causal model for this class of problems has  $N$  binary variables  $\{X_1, \dots, X_N\}$  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:  $\mathcal{A} = \{do()\} \cup \{do(X_i = j) : 1 \leq i \leq N \text{ and } j \in \{0, 1\}\}$ . 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  $\mathbf{q} = (q_1, \dots, q_N) \in [0, 1]^N$  so that  $q_i = \mathbb{P}\{X_i = 1\}$ . The value of the reward variable is distributed as  $\mathbb{P}\{Y = 1 | \mathbf{X}\} = r(\mathbf{X})$  where  $r : \{0, 1\}^N \rightarrow [0, 1]$  is an arbitrary, fixed, and unknown function. In the farming example, this choice of  $Y$  models the success or failure of a seasons crop, which depends stochastically on the various environment variables.

**The Parallel Bandit Algorithm** The algorithm operates as follows. For the first  $T/2$  rounds it chooses  $do()$  to collect observational data. As the only link from each  $X_1, \dots, X_N$  to  $Y$  is a direct, causal one,  $\mathbb{P}\{Y | do(X_i = j)\} = \mathbb{P}\{Y | X_i = j\}$ . Thus we can create good estimators for the returns of the actions  $do(X_i = j)$  for which  $\mathbb{P}\{X_i = j\}$  is large. The actions for which  $\mathbb{P}\{X_i = j\}$  is small may not be observed (often) so estimates 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  $\mathbf{q}$  and, in particular, how many of the variables are unbalanced (i.e., small  $q_i$  or  $(1 - q_i)$ ). For  $\tau \in [2 \dots N]$  let  $I_\tau = \{i : \min\{q_i, 1 - q_i\} < \frac{1}{\tau}\}$ . Define

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

$I_\tau$  is the set of variables considered unbalanced and we tune  $\tau$  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  $\mathbf{q} = (\frac{1}{2}, \dots, \frac{1}{2})$  we have  $m(\mathbf{q}) = 2$  and when  $\mathbf{q} = (0, \dots, 0)$  we have  $m(\mathbf{q}) = N$ . We do not assume that  $\mathbf{q}$  is known, thus Algorithm 1 also utilizes the samples captured during the observational phase to estimate  $m(\mathbf{q})$ . Although very simple, the following two theorems show that this algorithm is effectively optimal.

**Theorem 1.** Algorithm 1 satisfies

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

**Theorem 2.** For all  $T$ ,  $\mathbf{q}$  and all strategies, there exists a reward function such that

$$R_T \in \Omega\left(\sqrt{\frac{m(\mathbf{q})}{T}}\right).$$

---

#### Algorithm 1 Parallel Bandit Algorithm

---

- 1: **Input:** Total rounds  $T$  and  $N$ .
  - 2: **for**  $t \in 1, \dots, T/2$  **do**
  - 3:    Perform empty intervention  $do()$
  - 4:    Observe  $\mathbf{X}_t$  and  $Y_t$
  - 5: **for**  $a = do(X_i = x) \in \mathcal{A}$  **do**
  - 6:    Count times  $X_i = x$  seen:  $T_a = \sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = x\}$
  - 7:    Estimate reward:  $\hat{\mu}_a = \frac{1}{T_a} \sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = x\} Y_t$
  - 8:    Estimate probabilities:  $\hat{p}_a = \frac{2T_a}{T}$ ,  $\hat{q}_i = \hat{p}_{do(X_i=1)}$
  - 9:    Compute  $\hat{m} = m(\hat{\mathbf{q}})$  and  $A = \{a \in \mathcal{A} : \hat{p}_a \leq \frac{1}{\hat{m}}\}$ .
  - 10: Let  $T_A := \frac{T}{2|A|}$  be times to sample each  $a \in A$ .
  - 11: **for**  $a = do(X_i = x) \in A$  **do**
  - 12:    **for**  $t \in 1, \dots, T_A$  **do**
  - 13:     Intervene with  $a$  and observe  $Y_t$
  - 14:    Re-estimate  $\hat{\mu}_a = \frac{1}{T_A} \sum_{t=1}^{T_A} Y_t$
  - 15: **return** estimated optimal  $\hat{a}_T^* \in \arg \max_{a \in \mathcal{A}} \hat{\mu}_a$
-

The proofs of Theorems 1 and 2 may be found in Sections 7 and 8 respectively. By utilizing knowledge of the causal structure, Algorithm 1 effectively only has to explore the  $m(\mathbf{q})$  '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,  $P\{Y|X_i = j\} \neq P\{Y|do(X_i = j)\}$  (correlation is not causation). However, if all the variables are observable, any causal distribution  $P\{X_1 \dots X_N | do(X_i = j)\}$  can be expressed in terms of observational distributions via the truncated factorization formula Pearl (2000).

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

where  $\mathcal{P}_{a_{X_k}}$  denotes the parents of  $X_k$  and  $\delta$  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  $P\{Y|a\}$  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  $k \in 2, \dots, N$  and  $P\{X_1\} = 0$ . We can learn the reward for all the interventions  $do(X_i = 1)$  simultaneously by selecting  $do(X_1 = 1)$ , but not from  $do(\cdot)$ . In addition, variance of the observational estimator for  $a = do(X_i = j)$  can be high even if  $P\{X_i = j\}$  is large. Given the causal graph in Figure 1b,  $P\{Y|do(X_2 = j)\} = \sum_{X_1} P\{X_1\} P\{Y|X_1, X_2 = j\}$ . Suppose  $X_2 = X_1$  deterministically, no matter how large  $P\{X_2 = 1\}$  is we will never observe  $(X_2 = 1, X_1 = 0)$  and so cannot get a good estimate for  $P\{Y|do(X_2 = 1)\}$ .

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  $P\{\mathcal{P}_{a_Y} | a\}$  (but not  $P\{Y|a\}$ ) 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  $\eta$  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  $Q = \sum_{a \in \mathcal{A}} \eta_a P\{\mathcal{P}_{a_Y} | a\}$  to be the mixture distribution over the interventions with respect to  $\eta$ .

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

---

### Algorithm 2 General Algorithm

---

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

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

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

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

---

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.

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

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

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

230 **Theorem 3.** *If Algorithm 2 is run with  $B \in \mathbb{R}^{\mathcal{A}}$  given by  $B_a = \sqrt{\frac{m(\eta)T}{\log(2T|\mathcal{A}|)}}$ .*

$$R_T \in \mathcal{O} \left( \sqrt{\frac{m(\eta)}{T} \log(2T|\mathcal{A}|)} \right).$$

231 The proof is in Section 9. Note the regret has the same form as that obtained for Algorithm 1, with  
 232  $m(\eta)$  replacing  $m(q)$ . Algorithm 1 assumes only the graph structure and not knowledge of the  
 233 conditional distributions on  $X$ . Thus it has broader applicability to the parallel graph than the generic  
 234 algorithm given here. We believe that Algorithm 2 with the optimal choice of  $\eta$  is close to minimax  
 235 optimal, but leave lower bounds for future work.

236 **Choosing the Sampling Distribution** Algorithm 2 depends on a choice of sampling distribution  
 237  $\mathbb{Q}$  that is determined by  $\eta$ . In light of Theorem 3 a natural choice of  $\eta$  is the minimiser of  $m(\eta)$ .

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

238 Since the mixture of convex functions is convex and the maximum of a set of convex functions is  
 239 convex, we see that  $m(\eta)$  is convex (in  $\eta$ ). Therefore the minimisation problem may be tackled  
 240 using standard techniques from convex optimisation. An interpretation of  $m(\eta^*)$  is the minimum  
 241 achievable worst-case variance of the importance weighted estimator. In the experimental section we  
 242 present some special cases, but for now we give two simple results. The first shows that  $|\mathcal{A}|$  serves as  
 243 an upper bound on  $m(\eta^*)$ .

244 **Proposition 4.**  $m(\eta^*) \leq |\mathcal{A}|$ . *Proof.* By definition,  $m(\eta^*) \leq m(\eta)$  for all  $\eta$ . Let  $\eta_a = 1/|\mathcal{A}| \forall a$ .

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

245 The second observation is that, in the parallel bandit setting,  $m(\eta^*) \leq 2m(q)$ . This is easy to see  
 246 by letting  $\eta_a = 1/2$  for  $a = do()$  and  $\eta_a = \mathbb{1}\{\mathbb{P}\{X_i = j\} \leq 1/m(q)\}/2m(q)$  for the actions  
 247 corresponding to  $do(X_i = j)$ , and applying an argument like that for Proposition 4. The proof is in  
 248 Section 9.1.

249 **Remark 5.** The choice of  $B_a$  given in Theorem 3 is not the only possibility. As we shall see in the  
 250 experiments, it is often possible to choose  $B_a$  significantly larger when there is no heavy tail and  
 251 this can drastically improve performance by eliminating the bias. This is especially true when the  
 252 ratio  $R_a$  is never too large and Bernstein's inequality could be used directly without the truncation.  
 253 For another discussion see the article by Bottou et al. (2013) who also use importance weighted  
 254 estimators to learn from observational data.

## 255 5 Experiments

256 We compare Algorithms 1 and 2 with Successive Elimination on the parallel bandit problem under a  
 257 variety of conditions, including where the importance weighted estimator used by Algorithm 2 is not  
 258 truncated, which is justified in this setting by Remark 5. Throughout we use a model in which  $Y$   
 259 depends only on a single variable  $X_1$  (this is unknown to the algorithms).  $Y_t \sim \text{Bernoulli}(\frac{1}{2} + \varepsilon)$  if  
 260  $X_1 = 1$  and  $Y_t \sim \text{Bernoulli}(\frac{1}{2} - \varepsilon')$  otherwise, where  $\varepsilon' = q_1\varepsilon/(1 - q_1)$ . This leads to an expected

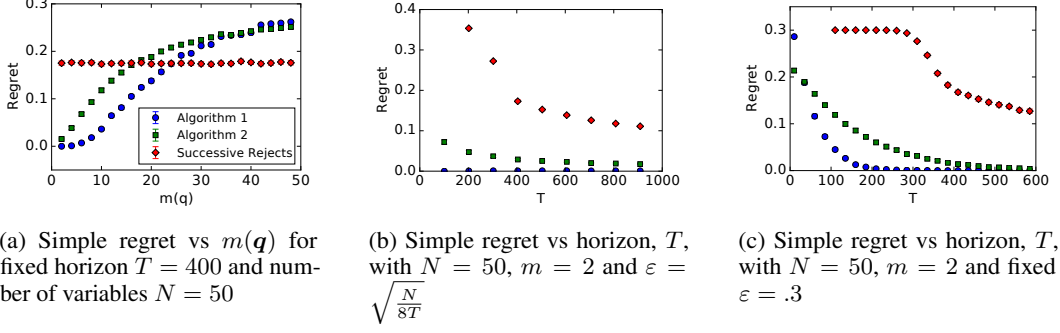


Figure 2: Experimental results

261 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  $q_i = 0$   
 262 for  $i \leq m$  and  $\frac{1}{2}$  otherwise. Note that changing  $m$  and thus  $\mathbf{q}$  has no effect on the reward distribution.

263 We compare the performance of the Algorithm 1, which is specific to the parallel problem, but does  
 264 not require knowledge of  $\mathbf{q}$ , with that of Algorithm 2 and the Successive Reject algorithm of Audibert  
 265 and Bubeck (2010). For each experiment, we show the average regret over 10,000 simulations with  
 266 error bars displaying three standard errors.

267 In figure 2a we fix the number of variables  $N$  and the horizon  $T$  and compare the performance  
 268 of the algorithms as  $m$  increases. The regret for the Successive Reject algorithm is constant as it  
 269 depends only on the reward distribution and has no knowledge of the causal structure. For the causal  
 270 algorithms it increases approximately with  $\sqrt{m}$ . As  $m$  approaches  $N$ , the gain the causal algorithms  
 271 obtain from knowledge of the structure is outweighed by fact they do not leverage the observed  
 272 rewards to focus sampling effort on actions with high pay-offs.

273 Figure 2b demonstrates the performance of the algorithms in the worst case environment for standard  
 274 bandits, where the gap between the optimal and sub-optimal arms,  $\varepsilon = \sqrt{N/(8T)}$ , is just too small  
 275 to be learned. This gap is learn-able by the causal algorithms, for which the worst case  $\varepsilon$  depends  
 276 on  $m \ll N$ . In figure 2c we fix  $N$  and  $\varepsilon$  and observe that, for sufficiently large  $T$ , the regret decays  
 277 exponentially. The decay constant is larger for the causal algorithms as they have observed a greater  
 278 effective number of samples for a given  $T$ .

279 For the parallel bandit problem, the regression estimator used in the specific algorithm outperforms the  
 280 truncated importance weighted estimator in the more general algorithm, despite the fact the specific  
 281 algorithm must estimate  $\mathbf{q}$  from the data. This is an interesting phenomenon that has been noted  
 282 before in off-policy evaluation where the regression (and not the importance weighted) estimator is  
 283 known to be minimax optimal asymptotically (Li et al., 2014).

## 284 6 Discussion & Future Work

285 Algorithm 2 for general causal bandit problems estimates the reward for all allowable interventions  
 286  $a \in \mathcal{A}$  over  $T$  rounds by sampling and applying interventions from a distribution  $\eta$ . Theorem 3 shows  
 287 that this algorithm has (up to log factors) simple regret that is  $\mathcal{O}(\sqrt{m(\eta)/T})$  where the parameter  
 288  $m(\eta)$  measures the difficulty of learning the causal model and is always less than  $N$ . The value of  
 289  $m(\eta)$  is a uniform bound on the variance of the reward estimators  $\hat{\mu}_a$  and, intuitively, problems where  
 290 all variables' values in the causal model "occur naturally" when interventions are sampled from  $\eta$   
 291 will have low values of  $m(\eta)$ .

292 The main practical drawback of Algorithm 2 is that both the estimator  $\hat{\mu}_a$  and the optimal sampling  
 293 distribution  $\eta^*$  (i.e., the one that minimises  $m(\eta)$ ) require knowledge of the conditional distributions  
 294  $P\{\mathcal{P}_{\mathcal{A}_Y} | a\}$  for all  $a \in \mathcal{A}$ . In contrast, in the special case of parallel bandits, Algorithm 1 uses  
 295 the  $do()$  action to effectively estimate  $m(\eta)$  and the rewards then re-samples the interventions with  
 296 variances that are not bound by  $\hat{m}(\eta)$ . Despite these extra estimates, Theorem 2 shows that this  
 297 approach is optimal (up to log factors). Finding an algorithm that only requires the causal graph and  
 298 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 successive 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  $\eta$ , 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  $\mathbf{q}$  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 each round the learner chooses an action and receives a Gaussian reward signal for all actions, but with a known variance that depends on the chosen action. In this way the learner can gain information about actions it does not take with varying levels of accuracy. The reduction follows by substituting the importance weighted estimators in place of the Gaussian reward. In the case that  $\mathbf{q}$  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  $\mathbf{q}$ , and revealed after an action has been chosen. The feedback graph is distinct from the causal graph. A link  $A \rightarrow B$  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  $\mathbf{q} = \mathbf{0}$ .

**Causal Models with Non-Observable Variables** If we assume knowledge of the conditional *interventional* distributions  $P\{\mathcal{P}_{\mathcal{A}\mathcal{Y}}|a\}$  our analysis applies unchanged to the case of causal models with non-observable variables. Some of the interventional distributions may be non-identifiable meaning we can not obtain prior estimates for  $P\{\mathcal{P}_{\mathcal{A}\mathcal{Y}}|a\}$  from even an infinite amount of observational data. Even if all variables are observable and the graph is known, if the conditional distributions are unknown, then Algorithm 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  $\mathcal{O}(\log \log n)$  multi-variable interventions are required on average to recover a causal graph over  $n$  variables once purely observational data is used to recover the “essential graph”. Simultaneously learning a completely unknown causal model while estimating the rewards of interventions without a large observational 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.
- 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.
- 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., Quinero-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.
- Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. OUP Oxford.
- 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 Szepesvári, 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.

- 401 Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–536.
- 403 Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.
- 404 Wu, Y., Györfy, A., and Szepesvári, C. (2015). Online Learning with Gaussian Payoffs and Side Observations. In *NIPS*, pages 1360–1368.
- 406 Yu, J. Y. and Mannor, S. (2009). Piecewise-stationary bandit problems with side observations. In *ICML*, pages 1177–1184.

## 408 7 Proof of Theorem 1

409 Assume without loss of generality that  $q_1 \leq q_2 \leq \dots \leq q_N \leq 1/2$ . The assumption is non-  
 410 restrictive since all variables are independent and permutations of the variables can be pushed to the  
 411 reward function. The proof of Theorem 1 requires some lemmas, the first of which is an immediate  
 412 consequence of the Chernoff bound (see Exercise 2.1 in Boucheron et al. (2013) for this form of the  
 413 Chernoff bound).

414 **Lemma 6.** Let  $i \in \{1, \dots, N\}$  and  $\delta > 0$ . Then

$$\mathbb{P} \left\{ |\hat{q}_i - q_i| \geq \sqrt{\frac{6q_i}{T} \log \frac{2}{\delta}} \right\} \leq \delta.$$

415 **Lemma 7.** Let  $X_1, X_2, \dots$ , be a sequence of random variables with  $X_i \in [0, 1]$  and  $\mathbb{E}X_i = p$  and  
 416  $\delta \in [0, 1]$ . Then

$$\mathbb{P} \left\{ \exists t \geq n_0 : \left| \frac{1}{t} \sum_{s=1}^t X_s - p \right| \geq \sqrt{\frac{2}{n_0} \log \frac{2}{\delta}} \right\} \leq 4\delta.$$

417 *Proof.* For  $\delta \geq 1/4$  the result is trivial. Otherwise by Hoeffding's bound and the union bound:

$$\begin{aligned} \mathbb{P} \left\{ \exists t \geq n_0 : \left| \frac{1}{t} \sum_{s=1}^t X_s - p \right| \geq \sqrt{\frac{2}{n_0} \log \frac{2}{\delta}} \right\} &\leq \sum_{t=n_0}^{\infty} \mathbb{P} \left\{ \left| \frac{1}{t} \sum_{s=1}^t X_s - p \right| \geq \sqrt{\frac{2}{n_0} \log \frac{2}{\delta}} \right\} \\ &\leq 2 \sum_{t=n_0}^{\infty} \exp \left( -\frac{t}{n_0} \log \frac{2}{\delta} \right) \leq 4\delta. \quad \square \end{aligned}$$

418 **Lemma 8.** Let  $\delta \in (0, 1)$  and assume  $T \geq 48m \log \frac{2N}{\delta}$ . Then

$$\mathbb{P} \{ 2m(\mathbf{q})/3 \leq m(\hat{\mathbf{q}}) \leq 2m(\mathbf{q}) \} \geq 1 - \delta.$$

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

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

420 Then by the union bound and Lemma 6 we have  $\mathbb{P} \{ F \} \leq \delta$ . The result will be completed by showing  
 421 that when  $F$  does not hold we have  $2m(\mathbf{q})/3 \leq m(\hat{\mathbf{q}}) \leq 2m(\mathbf{q})$ . From the definition of  $m(\mathbf{q})$  and  
 422 our assumption on  $\mathbf{q}$  we have for  $i > m$  that  $q_i \geq q_m \geq 1/m$  and so by Lemma 6 we have

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

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

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

425 Therefore  $m(\hat{\mathbf{q}}) \geq 2m(\mathbf{q})/3$  as required.  $\square$

426 *Proof of Theorem 1.* Let  $\delta = m = m(\mathbf{q})/N$ . Then by Lemma 8 we have

$$\mathbb{P} \{2m/3 \leq m(\hat{\mathbf{q}}) \leq 2m\} \geq 1 - \delta.$$

427 Recall that  $A = \{a \in \mathcal{A} : \hat{p}_a \leq 1/m(\hat{\mathbf{q}})\}$ . Then for  $a \in A$  the algorithm estimates  $\mu_a$  from  
 428  $T/(2m(\hat{\mathbf{q}})) \geq T/(4m)$  samples. Therefore by Hoeffding's inequality and the union bound we  
 429 have

$$\mathbb{P} \left\{ \exists a \in A : |\mu_a - \hat{\mu}_a| \geq \sqrt{\frac{8m}{T} \log \frac{2N}{\delta}} \right\} \leq \delta.$$

430 For arms not in  $a$  we have  $\hat{p}_a \geq 1/m(\hat{\mathbf{q}}) \geq 1/(2m)$ . Therefore if  $a = do(X_i = j)$ , then

$$\hat{p}_a = \frac{2}{T} \sum_{t=1}^{T/2} \mathbb{1}\{X_i = j\} \geq \frac{1}{2m}.$$

431 Therefore  $\sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = j\} \geq T/4m$  and by Lemma 7 we have

$$\mathbb{P} \left\{ \sum_{t=1}^{T/2} \mathbb{1}\{X_i = j\} \geq \frac{T}{4m} \text{ and } |\hat{\mu}_a - \mu_a| \geq \sqrt{\frac{8m}{T} \log \frac{2N}{\delta}} \right\} \leq 4\delta/N.$$

432 Therefore with probability at least  $1 - 6\delta$  we have

$$(\forall a \in \mathcal{A}) \quad |\hat{\mu}_a - \mu_a| \leq \sqrt{\frac{8m}{T} \log \frac{N}{\delta}} = \varepsilon.$$

433 If this occurs, then

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

434 Therefore

$$\mu^* - \mathbb{E}[\mu_{\hat{a}_T^*}] \leq 6\delta + \varepsilon \leq \frac{6m}{T} + \sqrt{\frac{32m}{T} \log \frac{NT}{m}},$$

435 which completes the result.  $\square$

## 436 8 Proof of Theorem 2

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

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

440 where  $1/4 \geq \varepsilon > 0$  is some constant to be chosen later. We abbreviate  $R_{T,i}$  to be the expected  
 441 simple regret incurred when interacting with the environment determined by  $\mathbf{q}$  and  $r_i$ . Let  $P_i$  be the  
 442 corresponding measure on all observations over all  $T$  rounds and  $\mathbb{E}_i$  the expectation with respect to  
 443  $P_i$ . By Lemma 2.6 by Tsybakov (2008) we have

$$\mathbb{P}_0 \{\hat{a}_T^* = a^*\} + \mathbb{P}_i \{\hat{a}_T^* \neq a^*\} \geq \exp(-\text{KL}(P_0, P_i)),$$

444 where  $\text{KL}(P_0, P_i)$  is the KL divergence between measures  $P_0$  and  $P_i$ . Let  $T_i(T) =$   
 445  $\sum_{t=1}^T \mathbb{1}\{a_t = do(X_i = 1)\}$  be the total number of times the learner intervenes on variable  $i$  by  
 446 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  
 447 bounded using the telescoping property (chain rule) and by bounding the local KL divergence by the  
 448  $\chi$ -squared distance as by Auer et al. (1995). This leads to

$$\text{KL}(P_0, P_i) \leq 6\varepsilon^2 \mathbb{E}_0 \left[ \sum_{t=1}^T \mathbb{1}\{X_{t,i} = 1\} \right] \leq 6\varepsilon^2 (\mathbb{E}_0 T_i(T) + q_i T) \leq 6\varepsilon^2 \left( \mathbb{E}_0 T_i(T) + \frac{T}{m} \right).$$

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

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

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

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

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

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

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

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

454 as required.

## 455 9 Proof of Theorem 3

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

$$\text{Var}_{\mathbb{Q}}[Z_{at}] \leq \mathbb{E}_{\mathbb{Q}}[Z_{at}^2] \leq \mathbb{E}_a \left[ \frac{\mathbb{P}\{\mathcal{P}_{aY}(X)|a\}}{\mathbb{Q}\{\mathcal{P}_{aY}(X)\}} \right] \leq m(\eta).$$

458 Checking the expectation we have

$$\mathbb{E}_{\mathbb{Q}}[Z_{at}] = \mathbb{E}_a[Y \mathbb{1}\{R_{at} \leq B_a\}] = \mathbb{E}_a Y - \mathbb{E}_a[Y \mathbb{1}\{R_a > B_a\}] = \mu_a - \beta_a,$$

459 where

$$0 \leq \beta_a = \mathbb{E}_a[Y \mathbb{1}\{R_a > B_a\}] \leq \mathbb{P}\{R_a(X) > B_a | a\}$$

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

$$\beta_a \leq \mathbb{P}\{R_a(X) > B_a | a\} \leq \frac{\mathbb{E}_a[R_a(X)]}{B_a} \leq \frac{m(\eta)}{B_a}.$$

462 Let  $\varepsilon_a > 0$  be given by

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

463 Then by the union bound and Bernstein's inequality

$$\mathbb{P}\{\text{exists } a \in \mathcal{A} : |\hat{\mu}_a - \mathbb{E}_{\mathbb{Q}}[\hat{\mu}_a]| \geq \varepsilon_a\} \leq \sum_{a \in \mathcal{A}} \mathbb{P}\{|\hat{\mu}_a - \mathbb{E}_{\mathbb{Q}}[\hat{\mu}_a]| \geq \varepsilon_a\} \leq \frac{1}{T}.$$

464 Assuming this event does not occur and letting  $a^* = \arg \max_{a \in \mathcal{A}} \mu_a$  we have

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

465 By the definition of the truncation we have

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

466 and

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

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

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

468 Therefore

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

469 as required. □

## 470 9.1 Relationship between $m(\eta)$ and $m(\mathbf{q})$

471 **Proposition 9.** *In the parallel bandit setting,  $m(\eta^*) \leq 2m(\mathbf{q})$ .*

472 *Proof.* Recall that in the parallel bandit setting,

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

473 Let:

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

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

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

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

476 Recall that,

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

477 We now show that our choice of  $\eta$  ensures  $\mathbb{E}_a \left[ \frac{\mathbb{P}\{\mathcal{P}_{\text{aY}}(X)|a\}}{\mathbb{Q}\{\mathcal{P}_{\text{aY}}(X)\}} \right] \leq 2m(\mathbf{q})$  for all actions  $a$ .

478 For the actions  $a : \eta_a > 0$ , ie  $do()$  and  $do(X_i = j) : \mathbb{P}\{X_i = j\} < \frac{1}{m(\mathbf{q})}$ ,

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

479 For the actions  $a : \eta_a = 0$ , ie  $do(X_i = j) : \mathbb{P}\{X_i = j\} \geq \frac{1}{m(\mathbf{q})}$ ,

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

480 Therefore  $m(\eta^*) \leq m(\eta) \leq 2m(\mathbf{q})$  as required. □

481

482