Budgeted and Non-Budgeted Causal Bandits

Vineet Nair
Technion Israel Institute of Technology
vineet@cs.technion.ac.il

Vishakha Patil[1]
Indian Institute of Science
patilv@iisc.ac.in

Gaurav Sinha[1]
Adobe Research
gasinha@adobe.com

Abstract

Learning good interventions in a causal graph can be modeled as a stochastic multi-armed bandit problem with side-information. First, we study this problem when interventions are more expensive than observations, and a budget is specified. If there are no backdoor paths from the *intervenable* nodes to the reward node, then we propose an algorithm to minimize simple regret that optimally trades-off observations and interventions based on the cost of interventions. We also propose an algorithm that accounts for the cost of interventions, utilizes causal side-information and minimizes the expected cumulative regret without exceeding the budget. Our algorithm performs better than standard algorithms that do not take side-information into account. Finally, we study the problem of learning best interventions without budget constraint in general graphs and give an algorithm that achieves constant expected cumulative regret in terms of the instance parameters when the parent distribution of the reward variable for each intervention is known. Our results are experimentally validated and compared to the best-known bounds in the current literature.

1 Introduction

Causal Bayesian Networks (CBN) (Pearl, 2009) have become the popular choice to model causal relationships in many real-world systems such as online advertising, gene interaction networks, brain functional connectivity, etc. The underlying directed acyclic graph (DAG) of a CBN is called its *causal graph*. The nodes of this graph are labeled by random variables[2] representing the underlying system, and edges between these variables capture direct causal relationships. Once the causal graph is known, any external manipulations on the system that forcibly fixes some target variables can be modeled via an operation called *intervention*. An intervention simulates the effect of the manipulation on other system variables by disconnecting the target variables from their parents[3] and setting them to the desired value.

Two key questions in causal learning are: 1) learning the causal graph itself, and 2) finding the intervention that optimizes some variable of interest (often called *reward* variable) assuming that the causal graph is known. In this work, we focus on the second question by modeling the causal learning problem as an extension of the Stochastic Multi-armed Bandit Problem (MAB) (Robbins, 1952). The MAB problem is a popular model used to capture decision-making in uncertain environments where a decision-maker is faced with k choices (called *arms*), and at each time step, the decision-maker has to choose one out of the k arms (*pull an arm*). The arm that is pulled gives a reward drawn from an underlying distribution that is unknown to the decision-maker beforehand.

We study the MAB problem with dependencies between the arms modeled via a causal graph. This model, called *causal bandits*, was studied in the recent works of Bareinboim et al. (2015); Lattimore et al. (2016); Sen et al. (2017a,b); Lee and Bareinboim (2018); Yabe et al. (2018); Lee and Bareinboim (2019); Lu et al. (2020), where the interventions are modelled as the arms of the bandit and the influence of the arms on the reward is assumed to conform to a known causal graph. In addition to the possible interventions allowed, the set of arms also contains an empty intervention called the observational arm,

[1]These authors have made equal contribution and their names are alphabetically ordered.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

[2]The joint distribution of these random variables factorizes over the graph.

[3]A process known as *causal surgery*.

where the algorithm does not perform any intervention on the causal graph. The goal of a causal bandit algorithm is to learn the intervention that maximizes the reward.

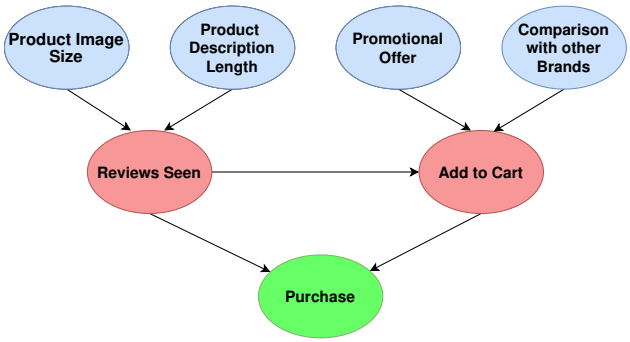


Figure 1: Causal Graph - Product Marketing

We explain the causal bandit problem with a motivating example from the marketing domain, for which a simple causal graph is shown in Figure 1. An e-commerce company sells a product online and makes a profit whenever a customer purchases their product. This corresponds to the green (reward) node labeled *Purchase* in the causal graph. The product web page is rendered using some values of the blue (intervenable) nodes on every new customer visit, chosen from an underlying distribution. For example, a customer might see a large image, short description, no promotions, and comparison with a competing brand. The red nodes capture actions taken by the customer before they make any purchase decision. For example, based on the rendered product page, a customer might want to get more information from the reviews before deciding to add the product to their shopping cart. Note that while the blue nodes are actionable and can be manipulated to increase the chances of purchase, the red nodes are not directly manipulable and can only be passively observed for any given values of the blue nodes. For example, the company might take an action by always offering a promotion, seeing which customers might decide to skip going through the reviews and directly add the product to their shopping cart. The objective here is to learn the intervention (on blue nodes) that maximizes the chances of the product being purchased.

However, in many situations, interventions are costly (Kocaoglu et al., 2017a,b; Lindgren et al., 2018; Addanki et al., 2020). Consider the marketing example above where observational data from this graph can be collected via independent customer visits, whereas to get an interventional sample, one needs to render a specific page configuration that would require additional expenditure. But recent works suggest that, in many scenarios, the effect of interventions can be efficiently estimated using observational samples (Tian and Pearl, 2002; Bhattacharyya et al., 2020; Pearl, 2009). Hence, in the causal bandit framework, for a fixed budget, there is a trade-off between the more economical observational arm and the high-cost interventional arms. This is because the observational arm, though less rewarding, aids in the exploration of the possibly high rewarding interventional arms. This motivates the study of observation/intervention trade-off in the budgeted bandit setting.

1.1 Our Contributions:

We study the problem of finding the best intervention in a causal graph in two settings: with and without budget constraints. Further, we study these problems with two common objectives in the MAB literature: simple regret minimization and cumulative regret minimization.

Budgeted Setting: In Sections 3 and 4, we consider a class of causal graphs that we call *no-backdoor graphs* (see Section 3 for the definition). A special instance of the class of no-backdoor graphs is the parallel graph model defined in Lattimore et al. (2016): $\mathcal{G}$ consists of $M+1$ nodes, $\mathcal{X} = \{Y, X_1, \ldots, X_M\}$, and the only edges in $\mathcal{G}$ are from each X_i to Y. For this, Lattimore et al. (2016) propose an algorithm called the *parallel bandit* algorithm (`PB-ALG` henceforth). We observe that `PB-ALG` works for the more general class of no-backdoor graphs.

We study the causal bandit problem for no-backdoor graphs in the budgeted bandit setting (Tran-Thanh et al., 2012), where a budget B is specified, and the ratio of the cost of the intervention to the cost of observation is $\gamma \geq 1$. The goal of an algorithm is to find the best intervention such that the total cost of arm pulls does not exceed B. In Section 3, we first study this problem with the goal of minimizing simple regret. Note that `PB-ALG` does not take into account the cost of interventions and is only optimal in the non-budgeted setting. We show that when γ is higher than a threshold (unknown to the algorithm), the simple algorithm `OBS-ALG` that plays the observational arm every time achieves better simple regret in terms of B than `PB-ALG`. Next, we propose γ`-NB-ALG` (Algorithm 1), which determines this unknown threshold online and successfully manages to trade-off interventions with observations for a specified budget.

In Section 4, we study the cumulative regret minimization (CRM) problem in the above setting and propose the `CRM-NB-ALG` algorithm (Algorithm 2). `CRM-NB-ALG` is based on the `Fractional-KUBE` algorithm (`F-KUBE` henceforth) given in Tran-Thanh et al. (2012) for budgeted bandits with no side-information. `CRM-NB-ALG`

achieves constant regret if the observational arm is the optimal arm and otherwise achieves logarithmic regret, which is better than that of F-KUBE in terms of instance-specific constants.

Non-Budgeted Setting: In Section 5, we study the problem of minimizing the cumulative regret for general causal graphs in the non-budgeted setting. We assume that the distribution of parents of the reward variable for each intervention is known to the algorithm. This assumption, though limiting in the practical setting, is also made in the recent work of Lu et al. (2020) (which studies the same problem) and Lattimore et al. (2016). Lu et al. (2020) proposed an algorithm called C-UCB, which has a worst-case regret guarantee of $O(\sqrt{k^n T})$ where k is the number of distinct values that each of the n parents of the reward variable can take. For the same problem, we propose C-UCB-2 (Algorithm 3) and show it has constant expected cumulative regret in terms of instance parameters, which is a significant improvement over previously known results.

2 Model and Notation

A CBN is a directed acyclic graph $\mathcal{G}$ whose nodes are labelled by random variables $\mathcal{X} = \{X_1, \ldots, X_n\}$, and a joint probability distribution $\mathbb{P}$ over $\mathcal{X}$ that factorizes over $\mathcal{G}$. For each $i \in [n]$, the range of X_i is a finite subset of $\mathbb{R}$. A node X_j is called a parent of node X_i if there is an edge from X_j to X_i in $\mathcal{G}$, i.e., changes in X_j directly affect X_i. The set of parents of X_i is denoted as $Pa(X_i)$. An intervention of size m corresponds to $\mathbf{X} \subset \mathcal{X}$ such that $|\mathbf{X}| = m$, where the variables in $\mathbf{X}$ are set to $\mathbf{x} = (x_1, \ldots, x_m)$, and this intervention is denoted as $do(\mathbf{X} = \mathbf{x})$. For each $X_i \in \mathbf{X}$, the intervention also removes all the edges from $Pa(X_i)$ to X_i, and the resulting graph defines a probability distribution $\mathbb{P}(\mathbf{X}^c | do(\mathbf{X} = \mathbf{x}))$ over $\mathbf{X}^c = \mathcal{X} \setminus \mathbf{X}$. The empty intervention, also called *observation*, corresponding to $\mathbf{X} = \phi$ is denoted as $do()$. A causal bandit algorithm is given as input a causal graph $\mathcal{G}$, the set of allowed interventions $\mathcal{A}$ (which corresponds to the set of arms), and a designated reward variable $Y \in \mathcal{X}$ where $Y \in \{0, 1\}$. The distribution $\mathbb{P}$ is unknown to the algorithm.

An algorithm for this problem is a sequential decision-making process that at each time t performs an intervention $a_t \in \mathcal{A}$ and observes reward $Y_t \in \{0, 1\}$. For each intervention $a \in \mathcal{A}$, where $a = do(\mathbf{X} = \mathbf{x})$, the expected reward of a is denoted $\mu_a = E[Y \mid do(\mathbf{X} = \mathbf{x})]$. We study a budgeted as well as a non-budgeted variant of this problem.

2.1 Budgeted Causal Bandits

In the budgeted variant of our problem, we associate a cost $\gamma > 1$ with each arm pull of a non-empty intervention $a \in \mathcal{A} \setminus \{do()\}$, whereas the cost of pulling the observation arm $do()$ is one. Hence, γ is the ratio of the cost of a non-empty intervention to the cost of the observational arm. The algorithm, in addition to $\mathcal{G}$ and $\mathcal{A}$, is specified a budget $B \in \mathbb{R}_+$. This model is similar to the budget-limited bandit model considered in Tran-Thanh et al. (2010); Tran-Thanh et al. (2012) where each arm has an associated cost per pull. Different cost models such as the linear cost model (Lindgren et al., 2018) and identity cost model (Addanki et al., 2020) have been studied in the causal discovery literature. The identity cost model is equivalent to considering a uniform cost γ across all interventions. In this work, we are interested in the trade-off between observations and interventions, which is more perceptible in the identity cost model. This is equivalent to considering a uniform cost γ across all interventions. Moreover, the algorithm presented for the budgeted setting in this paper can be extended with a bit of effort to different cost models. We study this problem from the perspective of two objectives: minimization of simple regret and cumulative regret, both being well-studied in the bandit community.

Simple regret: Let ALG be an algorithm for the above problem that outputs arm a_B when the budget given is B. Then the simple regret of ALG with budget B, denoted $r_{\texttt{ALG}}(B)$, is

$$r_{\texttt{ALG}}(B) = \max_{a \in \mathcal{A}} \mu_a - \mu_{a_B} \tag{1}$$

An algorithm whose objective is to minimize the simple regret is a pure-exploration algorithm. Its goal is to identify the best arm without restricting the number of times a sub-optimal arm may be played using the budget B. In many applications, we may require that a sub-optimal arm should not be pulled too many times right from the start. This motivates the definition of cumulative regret.

Cumulative regret: Let $G_{\texttt{ALG}}(B)$ be the expected reward accumulated by algorithm ALG with budget B, and let $G_B = \max_{\texttt{ALG}} G_{\texttt{ALG}}(B)$. Then, the cumulative regret of an algorithm ALG with budget B, denoted $R_{\texttt{ALG}}(B)$, is

$$R_{\texttt{ALG}}(B) = G_B - G_{\texttt{ALG}}(B) \tag{2}$$

An algorithm for cumulative regret minimization has to carefully trade-off between exploration vs. exploitation. Hence, an algorithm with good simple regret guarantees may not have good cumulative regret guarantees and vice-versa.

2.2 Non-budgeted Causal Bandits

In the non-budgeted variant of the problem, the cost associated with every intervention is the same, i.e., we can assume $\gamma = 1$ for all $a \in \mathcal{A}$. In Section 5, we study this problem with the objective of minimizing the expected cumulative regret when the time horizon T is unknown but finite. The regret notion is defined as in Equation 2 with $B = T$. Observe that $G_T = T \max_{a \in \mathcal{A}} \mu_a$. The cumulative regret of an algorithm ALG after T rounds, denoted $R_{\texttt{ALG}}(T)$, is then defined as

$$R_{\texttt{ALG}}(T) = T \cdot \max_{a \in \mathcal{A}} \mu_a - \sum_{t \in [T]} \mu_{a_t} \ . \tag{3}$$

The goal of any algorithm for such a setting is to minimize the expected cumulative regret $E[R_{\texttt{ALG}}(T)]$ where the expectation is taken over the randomness in the rewards as well as in the algorithm.

3 Simple Regret with Budget: No-backdoor Graphs

In this section and Section 4, we assume that the interventions are of size 1. Formally, let $\{X_1, \ldots, X_M\} \subseteq \mathcal{X}$ be the set of intervenable nodes such that $X_i \in \{0, 1\}$ for all $i \in [M]$. An intervention in this setting is defined as explicitly setting the value of a single node X_i as either 0 or 1. When not intervened upon, $X_i \sim \text{Bernoulli}(p_i)$. Hence, we have $2M + 1$ interventions in total: $2M$ interventions correspond to setting each of the M variables X_i to either 0 or 1, denoted $do(X_i = 0)$ and $do(X_i = 1)$ respectively, and the last intervention corresponds to the empty intervention, $do()$. Moreover, we assume that there are no backdoor paths from X_i to the reward variable. This implies $E[Y \mid do(X_i = x)] = E[Y \mid X_i = x] = \mu_{i,x}$ (see Section 3.3.1 Pearl (2000)). We call a causal graph $\mathcal{G}$ satisfying this property as a *no-backdoor graph* (NB-graph).

For ease of notation, we denote the intervention $do(X_i = x)$ by $a_{i,x}$ where $i \in [M]$ and $x \in \{0, 1\}$, and the empty intervention as a_0. The set of interventions is then $\mathcal{A} = \{a_{i,x} \mid i \in [M], x \in \{0, 1\}\} \uplus \{a_0\}$. The expected reward for the intervention $a_{i,x}$ and a_0 are $\mu_{i,x} = E[Y \mid X_i = x]$ and $\mu_0 = E[Y]$ respectively. Throughout Sections 3 and 4, $i \in [M]$ and $x \in \{0, 1\}$. Also, we use a to denote an intervention in $\mathcal{A}$ when we do not differentiate between $a_{i,x}$ and a_0. We study the budgeted causal bandit problem for no-backdoor graphs. As stated in Section 2, an algorithm for this problem is given as input the graph $\mathcal{G}$, the set of intervenable nodes $\{X_1, \ldots, X_M\}$, a budget B, and γ which is the cost for pulling an arm $a_{i,x}$. The algorithm does not know p_i for any i. Note that if $\gamma \geq B$ then trivially the algorithm can only make observations.

As stated above, $X_i \sim \text{Bernoulli}(p_i)$. Let $p_{i,1} = \mathbb{P}(X_i = 1) = p_i$ and $p_{i,0} = 1 - p_{i,1}$. We assume $p = \min_{(i,x)}\{p_{i,x}\} > 0$, which is reasonable in situations where the best arm is observable, and if the best arm is not observable then observational samples are not useful. Also, let $\mathbf{p} = (p_1 \ldots p_M)$. PB-ALG, the algorithm by Lattimore et al. (2016) minimizes the expected simple regret for the parallel graph model in the non-budgeted setting. We observe that PB-ALG works for any no-backdoor graph model. In particular, PB-ALG when applied to the budgeted setting plays the observational arm a_0 for the first $\frac{B}{(1+\gamma)}$ rounds and in the remaining $\frac{B}{(1+\gamma)}$ rounds plays the interventional arms that were observed only a few times during the arm pulls of a_0 (because the probability of observing them when a_0 is pulled is low). The simple regret guarantee of PB-ALG depends on the quantity $m(\mathbf{p})$ which captures the number of X_i's such that $\min(p_i, 1 - p_i) \ll 1/2$, (i.e. the number of arms that would be observed fewer number of times when the arm a_0 is pulled in the initial rounds). Formally, for $\tau \in [2, M]$, let $I_\tau = \left\{i \mid \min_x\{p_{i,x}\} < \frac{1}{\tau}\right\}$. Then, $m(\mathbf{p}) = \min\{\tau \mid |I_\tau| \leq \tau\}$ and the simple regret of the PB-ALG algorithm is $O\left(\sqrt{\frac{\gamma m(\mathbf{p})}{B} \log(\frac{MB}{\gamma m})}\right)$.

In Section 3.1, we show that for $\gamma = \Omega(\frac{1}{p \cdot m(\mathbf{p})})$ the simple algorithm that plays the observation arm for B rounds achieves better expected simple regret than PB-ALG. Since p_i for all i is unknown, the threshold $\frac{1}{p \cdot m(\mathbf{p})}$ is a priori unknown to an algorithm. Hence, in Section 3.2, we propose an algorithm that estimates this threshold online and trades-off between interventions and observations dependent on γ and the threshold to minimize the expected simple regret.

3.1 Observational Algorithm

Here, we analyze the simple-regret of the observational algorithm (OBS-ALG) which plays the arm a_0 for all the rounds, and at the end of B rounds outputs the arm $a \in \mathcal{A}$ with the highest empirical mean estimate. The empirical estimate of $\mu_{i,x}$ is computed as the average of the rewards accrued in those rounds where X_i was sampled as x. Theorem 1 shows the dependence of the expected simple regret of OBS-ALG on p.

Theorem 1. *The expected simple regret of OBS-ALG with budget B is $O\left(\sqrt{\frac{1}{pB} \log(pMB)}\right)$.*

The proof of Theorem 1 is in Appendix B.1 in the extended version of the paper[4]. Theorem 1 is proved by crucially leveraging the fact that the arm a_0 aides

[4] https://arxiv.org/pdf/2012.07058.pdf

in the exploration of all the other $2M$ arms, which is the side-information available in NB-graphs. Observe that the guarantee of OBS-ALG is better than that of PB-ALG in Lattimore et al. (2016) if $\gamma = \Omega(\frac{1}{p \cdot m(\mathbf{p})})$. Intuitively, the quantity $\frac{1}{p \cdot m(\mathbf{p})}$ captures the tipping point where intervening on the least observable arm is more costly than observing it.

3.2 Observation-Intervention Trade-off

We next propose γ-NB-ALG (Algorithm 1), which trades-off between observations and interventions depending on the value of γ to minimize the expected simple regret. The idea behind γ-NB-ALG is that, if γ is larger than the threshold $\frac{1}{p \cdot m(\mathbf{p})}$, then performing only observations gives a better regret (as stated at the end of Section 3.1), whereas if γ is less than this threshold, then the algorithm follows the strategy of PB-ALG.

γ-NB-ALG pulls arm a_0 for the first $\frac{B}{2}$ rounds (see Step 1). We note here that the regret of the algorithm (in Theorem 2) remains the same in order, if instead of $\frac{B}{2}$, for any constant c, $\frac{B}{c}$ arm pulls of a_0 are made initially. Since $\mathbf{p}$ and p are not known a priori, the algorithm has to estimate the threshold online as done in Step 6 of γ-NB-ALG. Note that at Step 6, $\widehat{\mathbf{p}} = (\widehat{p}_1 \dots \widehat{p}_M)$, where $\widehat{p}_i = \widehat{p}_{i,1}$, and $m(\widehat{\mathbf{p}})$ is defined similar to $m(\mathbf{p})$. If γ is less than the estimated threshold, then γ-NB-ALG chooses to play the interventions that were observed less frequently in the elapsed rounds. At Step 11, γ-NB-ALG determines the set A of such interventions, and at Step 12, plays them for an equal number of times in the remaining rounds. At Steps 13-14, the empirical estimates of only the arms in A are updated.

In Theorem 2, we bound the expected simple regret of γ-NB-ALG, which depends on γ and the value of the threshold.

Theorem 2. *If $\gamma \geq \frac{1}{p \cdot m(\mathbf{p})}$ then the expected simple regret of γ-NB-ALG is $O\left(\sqrt{\frac{1}{pB}\log(pMB)}\right)$, and if $\gamma \leq \frac{1}{p \cdot m(\mathbf{p})}$ then it is $O\left(\sqrt{\frac{\gamma \cdot m(\mathbf{p})}{B}\log\frac{MB}{\gamma \cdot m(\mathbf{p})}}\right)$.*

The proof of Theorem 2 is in Appendix B.2. Observe that the expected simple regret of γ-NB-ALG is equal to that of PB-ALG if $\gamma \leq \frac{1}{p \cdot m(\mathbf{p})}$, and is equal to the that of OBS-ALG if $\gamma > \frac{1}{p \cdot m(\mathbf{p})}$. For $\gamma = O(\frac{1}{p \cdot m(\mathbf{p})})$ the optimality of the regret up to log factors follows from Theorem 2 in Lattimore et al. (2016) where they show a $\Omega\left(\sqrt{\frac{\gamma m}{B}}\right)$ lower bound on the expected simple regret.[5] Experiment 2 in Section 6 shows that the performance of γ-NB-ALG matches or is better than the performance of PB-ALG for all values of γ, which validates our theoretical claim.

[5]The lower bound is shown in non-budgeted setting, which translates to $\Omega\left(\sqrt{\frac{\gamma m}{B}}\right)$ lower bound in our setting if $\gamma = O(\frac{1}{p \cdot m(\mathbf{p})})$.

Algorithm 1 γ-NB-ALG

INPUT: $\mathcal{G}$, B, and γ.

1: Play arm a_0 for the first $B/2$ rounds.
2: For each $a \in \mathcal{A}$, compute
3: $\quad \widehat{\mu}_{i,x} = \frac{\sum_{t=1}^{B/2} Y_t \cdot \mathbb{1}\{X_i = x\}}{\sum_{t=1}^{B/2} \mathbb{1}\{X_i = x\}}$, $\widehat{\mu}_0 = \frac{2\sum_{t=1}^{B/2} Y_t}{B}$
4: For each (i, x), compute
5: $\quad \widehat{p}_{i,x} = \frac{2}{B}\sum_{t=1}^{B/2} \mathbb{1}\{X_i = x\}$, and $\widehat{p} = \min_{i,x} \widehat{p}_{i,x}$
6: **if** $\widehat{p} \cdot m(\widehat{\mathbf{p}}) \geq \frac{1}{\gamma}$ **then**
7: $\quad$ Play arm a_0 for the remaining $B/2$ rounds.
8: $\quad$ For each $a \in \mathcal{A}$, compute
9: $\quad \widehat{\mu}_{i,x} = \frac{\sum_{t=1}^{B} Y_t \cdot \mathbb{1}\{X_i = x\}}{\sum_{t=1}^{B} \mathbb{1}\{X_i = x\}}$, $\widehat{\mu}_0 = \frac{\sum_{t=1}^{T} Y_t}{B}$
10: **else**
11: $\quad$ Compute $A = \{a_{i,x} \mid \widehat{p}_{i,x} < \frac{1}{m(\widehat{\mathbf{p}})}\}$.
12: $\quad$ Play each arm $a_{i,x} \in A$, for $\frac{B}{2\gamma|A|}$ rounds.
13: $\quad$ For each $a_{i,x} \in A$, set
14: $\quad\quad \widehat{\mu}_{i,x} = \frac{2\gamma|A|}{B}\sum_{t=\frac{B}{2}+1}^{\frac{B}{2}+\frac{B}{2\gamma}} Y_t \cdot \mathbb{1}\{a_t = a_{i,x}\}$.
15: **end if**
16: Output $\arg\max_{a \in \mathcal{A}} \widehat{\mu}_a$.

4 Cumulative Regret with Budget: No-backdoor Graphs

In this section, we propose CRM-NB-ALG (Algorithm 2) that minimizes the cumulative regret for the model in Section 3. CRM-NB-ALG is based on Fractional-KUBE (F-KUBE, Tran-Thanh et al. (2012)), which is a budget-limited version of Upper Confidence Bound (UCB, Auer et al. (2002)) but without side-information. In our setting, since arm a_0 aids in the exploration of all other $2M$ arms and has a unit cost, CRM-NB-ALG unlike F-KUBE ensures that arm a_0 is pulled sufficiently many times. Importantly, in CRM-NB-ALG the estimate for an arm $a_{i,x}$ is made using the effective number of pulls of the arm $a_{i,x}$, which is equal to the number of pulls of the arm $a_{i,x}$ plus the number of pulls of arm a_0 where X_i was sampled as x. Formally, let $N_t^{i,x}$ and N_t^0 denote the number of pulls of arm $a_{i,x}$ and a_0 respectively after t rounds, and let a_t denote the arm pulled at round t. The effective number of arm pulls of $a_{i,x}$ after t rounds is equal to $E_t^{i,x} = N_t^{i,x} + \sum_{s=1}^{t} \mathbb{1}\{a_s = a_0 \text{ and } X_i = x\}$.

At the end of t rounds, CRM-NB-ALG computes $\widehat{\mu}_{i,x}(t)$ and $\widehat{\mu}_0(t)$, which are the empirical estimates of $\mu_{i,x}$

and μ_0 respectively, as follows: $\widehat{\mu}_{i,x}(t)$ is equal to

$$\frac{\sum_{s=1}^{t} Y_s \cdot \mathbb{1}\{(a_s = a_{i,x}) \text{ or } (a_s = a_0 \text{ and } X_i = x)\}}{E_t^{i,x}}$$

$$\widehat{\mu}_0(t) = \frac{1}{N_t^0} \sum_{s=1}^{t} Y_s \cdot \mathbb{1}\{a_s = a_0\})$$

Based on these estimates, `CRM-NB-ALG` computes the weighted UCB estimates $\overline{\mu}_{i,x}(t)$ and $\overline{\mu}_0(t)$ for the arms $a_{i,x}$ and a_0 respectively, as follows:[6]

$$\overline{\mu}_{i,x}(t) = \frac{1}{\gamma}\left(\widehat{\mu}_{i,x}(t) + \sqrt{\frac{8\log t}{E_t^{i,x}}}\right)$$

$$\overline{\mu}_0(t) = \widehat{\mu}_0(t) + \sqrt{\frac{8\log t}{N_t^0}}$$

In each round, `CRM-NB-ALG` first ensures that arm a_0 is pulled at least $\beta^2 \log T$ times (steps 4-5), where β is set as in steps 11-14, and otherwise pulls the arm with the highest weighted UCB estimate (steps 6-8). Ensuring that arm a_0 is pulled at least $\beta^2 \log T$ times at the end of T rounds delicately balances the explore-exploit trade-off between the causal side-information obtained by pulling arm a_0, which ensures free exploration of the other $2M$ interventions, and the loss experienced in pulling the arm a_0 (if a_0 is the sub-optimal arm). The reason for setting β as in steps 11-14 is explained after Theorem 3, which bounds the expected cumulative regret of `CRM-NB-ALG`. Observe that `CRM-NB-ALG` halts once it has exhausted its entire budget B. Crucially though, the decisions of `CRM-NB-ALG` do not depend on the budget, i.e., it is budget-oblivious. But note that the decisions of the algorithm do take into account the cost of an intervention, i.e., the algorithm is not cost-oblivious.

Before stating Theorem 3, we introduce a few more notations which are used in the theorem. Let $v_{i,x} = \frac{\mu_{i,x}}{\gamma}$, and $v_0 = \mu_0$, and $a^* = \arg\max_{a \in \mathcal{A}}\{v_a\}$. Further, let $\Delta_a = \mu_{a^*} - \mu_a$ and $d_a = v_{a^*} - v_a$ for each $a \in \mathcal{A}$. Note that there could be $a \in \mathcal{A}$ such that $\Delta_a < 0$. In Theorem 3 we analyze the instance-dependent regret guarantee of `CRM-NB-ALG` , that is, the regret expression shows how the regret grows logarithmically with B given that the other instance-dependent parameters remain fixed.

Theorem 3. *If $a^* = a_0$ then the expected cumulative regret of the algorithm is $O(1)$ and otherwise the expected cumulative regret of the algorithm is of order* $\sum_{\Delta_{i,x}>0} \Delta_{i,x}\left(max\left(0, 1 + 8\ln B\left(\frac{1}{d_{i,x}^2} - \frac{p_{i,x}}{3d_0^2}\right)\right) + \frac{\pi^2}{3}\right) + \Delta_0\left(\frac{50\ln B}{d_0^2} + 1 + \frac{\pi^2}{3}\right)$.

[6] If we have different costs for each intervention, say $\gamma_{i,x}$ corresponding to arm $a_{i,x}$, then the estimate $\overline{\mu}_{i,x}(t)$ can be calculated by replacing γ with $\gamma_{i,x}$.

Algorithm 2 `CRM-NB-ALG`

INPUT: $\mathcal{G}$, Set of nodes $\{X_1, \ldots, X_M\}$, B, γ

1: Pull each arm once and set $t = 2M + 2$
2: Update $B_{t-1} = B - 2\gamma M - 1$ and let $\beta = 1$.
3: **while** $B_t \geq 1$ **do**
4: **if** $N_{t-1}^0 < \beta^2 \log t$ or $B_t < \gamma$ **then**
5: Pull $a_t = a_0$
6: **else**
7: Pull $a_t = \arg\max_{a \in \mathcal{A}} \overline{\mu}_a(t-1)$
8: **end if**
9: Update $N_t^a = N_{t-1}^a + \mathbb{1}\{a_t = a\}$
10: Update E_t^a, $\widehat{\mu}_a(t)$ and $\overline{\mu}_a(t)$ for all $a \in \mathcal{A}$.
11: Let $\widehat{\mu}^* = \max_{i,x} \widehat{\mu}_{i,x}(t)$.
12: **if** $\widehat{\mu}_0(t) < \frac{\widehat{\mu}^*}{\gamma}$ **then**
13: Update $\beta = \min(\frac{2\sqrt{2}}{\widehat{\mu}^*/\gamma - \widehat{\mu}_0(t)}, \sqrt{\log t})$
14: **end if**
15: Update $B_{t+1} = \begin{cases} B_t - 1 & \text{if } a_t = a_0 \\ B_t - \gamma & \text{if } a_t \neq a_0 \end{cases}$
16: Set $t = t + 1$
17: **end while**

The optimal arm a^* is equal to a_0 if the ratio of the expected reward of any intervention to expected reward of a_0 is at most γ. In particular, if $\frac{\max_{i,x}\mu_{i,x}}{\mu_0} \leq \gamma$, then $a^* = a_0$, and in this case the expected cumulative regret of `CRM-NB-ALG` is bounded by a *constant*. The proof of Theorem 3 is given in Appendix B.3. For the value of β set as in steps 11-14, we show that if $a^* \neq a_0$ then $\frac{8}{9d_0^2} \leq E[\beta^2] \leq \frac{50}{d_0^2}$ (see Lemma B.5 in Appendix B.3). This, in particular, ensures that if $a^* \neq a_0$ then the expected number of pulls of a sub-optimal arm $a_{i,x}$ is at most $\max\left(0, 1 + 8\ln B\left(\frac{1}{d_{i,x}^2} - \frac{p_{i,x}}{3d_0^2}\right)\right) + \frac{\pi^2}{3}$. Hence, note that if $\frac{1}{d_{i,x}^2} \geq \frac{p_{i,x}}{3d_0^2}$, then this sub-optimal arm is pulled at most a constant number of times. In Section 6, we show via simulation that `CRM-NB-ALG` performs much better than `F-KUBE` even for small values of γ. Note that `F-KUBE` does not take side information into account.

5 Cumulative Regret without Budget: General Graphs

In this section, we study the non-budgeted version of the causal bandit problem for general graphs (see Section 2.2) with the goal of minimizing the expected cumulative regret. This problem was studied in the recent work of Lu et al. (2020) who gave a `UCB` based algorithm, called `C-UCB`, which has a worst-case regret bound of $\sqrt{k^n T}$ when each of the n parent nodes of Y (the reward variable) in the graph can take one of k values. For the same problem, we propose an algorithm called `C-UCB-2`, which has constant regret in

terms of instance-parameters. Additionally, C-UCB in Lu et al. (2020) takes the time horizon T as input, but our algorithm C-UCB-2 works for any *unknown* (but finite) time horizon.

Let $Y_1, \ldots, Y_n$ be the parents of Y, hence we have $|Pa(Y)| = n$. Further, let $S \subset \mathbb{R}$ and $|S| = k$ be the set of values that a parents node of Y can take. We denote the realization of $Y_i = y_i$, where $y_i \in S$ for each $i \in [n]$, as $Pa(Y) = \mathbf{y}$ where $\mathbf{y} = (y_1 \ldots y_n) \in S^n$. We assume the following: (a) the algorithm receives as input $\mathbb{P}(Pa(Y) = \mathbf{y}|do(a))$ for all a, and (b) the distributions $\mathbb{P}(Pa(Y) = \mathbf{y}|do(a))$ for all a have the same non-zero support. Assumption (a) is also made in Lu et al. (2020); Lattimore et al. (2016) whereas Assumption (b) is made in other existing literature on causal bandits (see Sen et al. (2017a)). Let $c_{\mathbf{y}} = \min_a(\mathbb{P}(Pa(Y) = \mathbf{y} \mid do(a))$. Observe that the expected reward μ_a of any intervention $a \in \mathcal{A}$ satisfies

$$\mu_a = \sum_{\mathbf{y} \in S^n} E[Y \mid Pa(Y) = \mathbf{y}] \cdot \mathbb{P}(Pa(Y) = \mathbf{y} \mid do(a)) \ .$$

We denote $E[Y \mid Pa(Y) = \mathbf{y}]$ as $\mu_{\mathbf{y}}$. In C-UCB-2 (Algorithm 3), $\zeta_a = \sum_{c_{\mathbf{y}} > 0} \frac{\mathbb{P}(Pa(Y)=\mathbf{y}|do(a))}{c_{\mathbf{y}}}$ for each a, and $N_{\mathbf{y},t}$ denotes the number of times $Pa(Y)$ have been sampled as $\mathbf{y}$ in t rounds. Let $Pa_t(Y)$ denote the realization of $Pa(Y)$ at time t. Then, $N_{\mathbf{y},t} = \sum_{s=1}^{t} \mathbb{1}\{Pa_s(Y) = \mathbf{y}\}$ and $\widehat{\mu}_{\mathbf{y}}(t)$ is the empirical estimate of $\mu_{\mathbf{y}}$ at the end of t rounds defined as,

$$\widehat{\mu}_{\mathbf{y}}(t) = \frac{1}{N_{\mathbf{y},t}} \sum_{s=1}^{t} Y_s \cdot \mathbb{1}\{Pa_s(Y) = \mathbf{y}\}.$$

if $N_{\mathbf{y},t} \geq 1$ and otherwise $\widehat{\mu}_{\mathbf{y}}(t) = 0$. The algorithm also computes the empirical estimate $\widehat{\mu}_a(t)$ and the UCB estimate $\overline{\mu}_a(t)$ for all a using $\widehat{\mu}_{\mathbf{y}}(t)$ at the end of every round as follows:

$$\widehat{\mu}_a(t) = \sum_{\mathbf{y} \in S^n} \widehat{\mu}_{\mathbf{y}}(t) \cdot \mathbb{P}(Pa(Y) = \mathbf{y}|do(a)), \quad \text{and}$$

$$\overline{\mu}_a(t) = \widehat{\mu}_a(t) + \sqrt{\frac{\log(k^n t^2/2)}{t}} \zeta_a \ .$$

The quantity $\sqrt{\frac{\log(k^n t^2/2)}{t}} \zeta_a$ is called the confidence radius around the empirical estimate $\widehat{\mu}_a(t)$ at the end of t rounds. We remark here that the difference between our algorithm C-UCB-2 and C-UCB by Lu et al. (2020) is that C-UCB maintains a UCB estimate for each parent value tuple $\mathbf{y}$, whereas C-UCB-2 maintains a UCB estimate for each intervention $a \in \mathcal{A}$. We note that ζ_a sums only over $c_{\mathbf{y}} > 0$, and since the parent distributions have the same non-zero support, ζ_a is summed over same $\mathbf{y}$ for all $a \in \mathcal{A}$. This is used in proof of Theorem 4 (see Lemma B.8), which bounds the expected cumulative regret of C-UCB-2 . In Theorem 4, $\Delta_a = \max_{b \in \mathcal{A}} \mu_b - \mu_a$.

Algorithm 3 C-UCB-2

INPUT: $\mathcal{G}$, $\mathbb{P}(Pa(Y) = \mathbf{y}|do(a))$ for all $a \in \mathcal{A}$.

1: Play each intervention in round robin and for $t = |A|$ update $N_{\mathbf{y},t}, \widehat{\mu}_{\mathbf{y}}(t), \widehat{\mu}_a(t), \overline{\mu}_a(t)$.
2: Set $t = |A| + 1$
3: **loop**
4: Play $a_t = \arg\max_{a \in A} \overline{\mu}_a$.
5: Update $N_{\mathbf{y},t}, \widehat{\mu}_{\mathbf{y}}(t), \widehat{\mu}_a(t), \overline{\mu}_a(t)$.
6: $t = t + 1$
7: **end loop**

Theorem 4. *Let* $\delta = \min_{\mathbf{y} \in S^n}\{c_{\mathbf{y}} > 0\}$, $L_1 = \min\left\{t \in \mathbb{N} \mid t \geq \frac{2\log(k^n t^2)}{\delta^2}\right\}$, $L_{2,a} = \min\left\{t \in \mathbb{N} \mid t \geq \frac{4\log(k^n t^2/2)}{\Delta_a^2}\zeta_a^2\right\}$ *for all* $a \in \mathcal{A}$, *and* $L_a = max\{L_1, L_{2,a}\}$. *Then the expected cumulative regret of C-UCB-2 after* T *rounds is at most* $\sum_{a \in \mathcal{A}} \Delta_a \left(L_a + \frac{2\pi^2}{3}\right)$.

Observe that in Theorem 4, L_a is a constant based on problem instance parameters, and hence Theorem 4 proves that C-UCB-2 achieves instance dependent constant regret. Theorem 4 is proved by showing that the expected number of pulls of a sub-optimal arm $a \in \mathcal{A}$ after time L_a is at most $\frac{2\pi^2}{3}$ (proof in Appendix B.4). In Section 6, we show via simulations that the expected cumulative regret of C-UCB-2 is better than that of C-UCB, and the experiment also validates that the regret of C-UCB-2 is a constant.

6 Simulations

In this section, we experimentally validate our theoretical results. The rationale behind the choice of parameters and the causal graphs used in the experiments are explained in Appendix C.

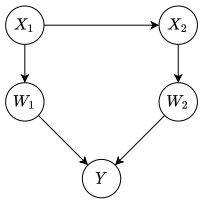
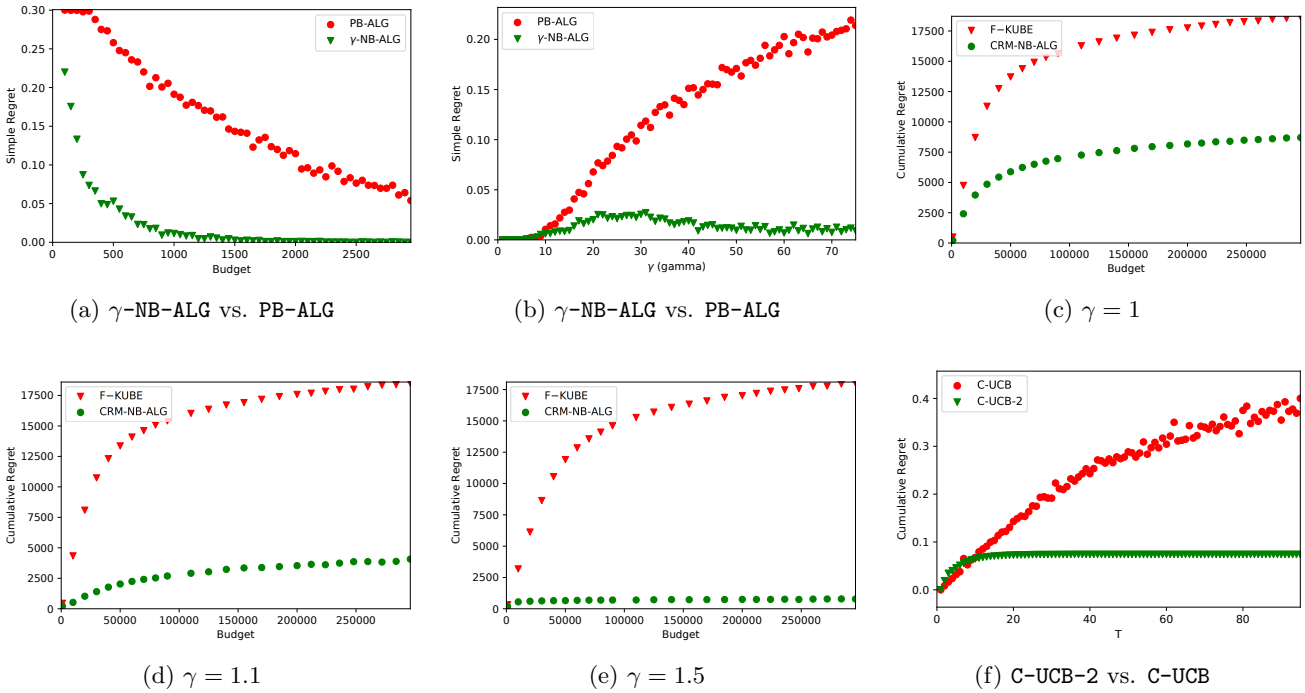


Figure 2: General Graphs

Experiment 1 (γ-NB-ALG **vs.** PB-ALG): This experiment compares the performance of γ-NB-ALG and PB-ALG on a parallel graph with $M = 50$, i.e the reward variable Y has 50 parents $X_1, \ldots, X_{50}$: $X_i \sim Bernoulli(p_i)$ for $i \in [50]$, $p_1 = p_2 = 0.02$, and



(a) γ-NB-ALG vs. PB-ALG (b) γ-NB-ALG vs. PB-ALG (c) $\gamma = 1$

(d) $\gamma = 1.1$ (e) $\gamma = 1.5$ (f) C-UCB-2 vs. C-UCB

Figure 3

$p_i = 0.5$ for $i \in [3, 50]$. For this choice of p_i's the PB-ALG algorithm asymptotically achieves the best regret. The rewards variable Y depends on X_i's as follows (unknown to both the algorithms): if $X_1 = 1$ then $Y \sim Bernoulli(0.5 + \epsilon)$ and otherwise $Y \sim Bernoulli(0.5 - \epsilon')$, where $\epsilon = 0.3$, and $\epsilon' = \frac{p_1 \epsilon}{1-p_1} \sim 0.006$.

The experiment in Figure 3a compares the simple regret of the two algorithms when $\gamma = 60$ and the budget is increased to 3000. The regret is computed by averaging it over 1000 independent runs. The experiment in Figure 3b illustrates the effect on the simple regret of the algorithms as γ increases from 1 to 75. Observe that till a threshold value of γ both algorithms have very close simple regret and post the threshold, γ-NB-ALG trades off between observations and interventions to yield a much better simple regret.

Experiment 2 (F-KUBE vs. CRM-NB-ALG): This experiment compares the performance of F-KUBE and CRM-NB-ALG . The model is as in Experiment 1, except $\epsilon = 0.5$. This was done to ensure better visualization. Figures 3c, 3d, and 3e illustrate the cumulative regrets of both the algorithms for γ equal to $1, 1.1$ and 1.5 respectively as the budget is increased. The regret is computed by averaging over 50 independent runs. Notice that CRM-NB-ALG yields a much better regret in all three cases and its regret is constant for $\gamma = 1.5$.

Experiment 3 (C-UCB vs. C-UCB-2): This experiment compares the performance of C-UCB and C-UCB-2. The underlying causal graph of our model is shown in Figure 2. Since node X_2 has a backdoor path, this model is not a no-backdoor graph. The conditional probability distribution for each node (given its parents) is given in Table 1, Appendix C. We model the conditional distribution of reward variable Y as: $Y|w_1, w_2 = \theta_1 w_1 + \theta_2 w_2 + \epsilon$, where ϵ is a Gaussian random variable $\mathcal{N}(0, 0.01)$. The parameters θ_1 and θ_2 are fixed to 0.25. Figure 3f shows a comparison of cumulative regret incurred by both algorithms for values of T in the range $[5, 100]$. The regret is computed by averaging over 500 independent runs. Notice that the regret of C-UCB is much higher than C-UCB-2 and also grows with time. Moreover, the regret of C-UCB-2 grows a little initially and then becomes constant as proved in Theorem 4.

7 Discussion and Future Work

The MAB problem can be used to model several real-world scenarios where additional information besides the reward of the pulled arms is available and hence the study of the MAB problem with side-information has been an area of significant interest in the research community. One of the most prominent models with side-information is the contextual MAB problem where the algorithm receives extra information (called *context*) before each arm pull (Lu et al., 2010). A class of bandit problems where the side-information obtained conforms to a feedback graph has also been studied in the

literature (Alon et al., 2015). The special case of parallel causal graphs studied by Lattimore et al. (2016) and in this work is in fact captured by such a model, but as shown by Lattimore et al. (2016) their regret bounds are not optimal in this setting.

In this work, we study the the causal bandit problem for no-backdoor graphs in the budgeted bandit framework. In this setting, observations are cheaper compared to interventions, which is practically well-motivated. In Sections 3 and 4 we provided two algorithms, γ-`NB-ALG` and `CRM-NB-ALG` , that minimized the expected simple regret and expected cumulative regret respectively. Sen et al. (2017a) also study the best intervention identification problem via importance sampling under budget constraint. But in contrast to our work, they consider soft interventions on a single node V, and also assume that the interventional distributions and the marginals of the parent distribution of the node V are known. This is incomparable with hard interventions on no-backdoor graphs, where interventions can be performed on different variables and the parent distributions of the intervened nodes are not known. Also, their setting is parameterized by B' and T, where B' is the upper bound on the average cost of sampling and T is the total number of samples that the algorithm draws. This can be mapped to our model by setting the budget to be $B'T$. In our budgeted setting, T is not given as input to the algorithm, and this is important for the trade-off between observations and interventions.

In the non-budgeted setting, we showed that our algorithm, `C-UCB-2`, has constant expected cumulative regret in terms of instance-parameters. We conjecture that the worst-case regret bound of our algorithm matches that in Lu et al. (2020), and resolving that remains open. The work by Sachidananda and Brunskill (2017) studies a similar problem as that in our work and experimentally show the effectiveness of Thompson Sampling but do not provide any theoretical guarantees.

Finally, many of the works in the literature such as those of Lu et al. (2020) and Lattimore et al. (2016) assume that the parent distribution for each intervention is known to the algorithm. We only make this assumption in Section 5. This assumption is limiting in practice and showing a non-trivial regret guarantee for settings without this assumption remains an important open direction.

Acknowledgements

Vineet Nair is thankful to be supported by the European Union's Horizon 2020 research and innovation program under grant agreement No 682203 -ERC-[Inf-Speed-Tradeoff]. Vishakha Patil is grateful for the support of a Google PhD Fellowship. The authors would like to thank the anonymous reviewers for their helpful comments.

References

Addanki, R., Kasiviswanathan, S. P., McGregor, A., and Musco, C. (2020). Efficient intervention design for causal discovery with latents. In *International Conference on Machine Learning, 2020*.

Alon, N., Cesa-Bianchi, N., Dekel, O., and Koren, T. (2015). Online learning with feedback graphs: Beyond bandits. In *Annual Conference on Learning Theory, 2015*, volume 40.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Bareinboim, E., Forney, A., and Pearl, J. (2015). Bandits with unobserved confounders: A causal approach. In *Annual Conference on Neural Information Processing Systems, 2015*, pages 1342–1350.

Bhattacharyya, A., Gayen, S., Kandasamy, S., Maran, A., and Vinodchandran, N. V. (2020). Efficiently learning and sampling interventional distributions from observations. *CoRR*, abs/2002.04232.

Kocaoglu, M., Dimakis, A., and Vishwanath, S. (2017a). Cost-optimal learning of causal graphs. In *International Conference on Machine Learning, 2017*, pages 1875–1884.

Kocaoglu, M., Shanmugam, K., and Bareinboim, E. (2017b). Experimental design for learning causal graphs with latent variables. In *Annual Conference on Neural Information Processing Systems, 2017*, pages 7018–7028.

Lattimore, F., Lattimore, T., and Reid, M. D. (2016). Causal bandits: Learning good interventions via causal inference. In *Annual Conference on Neural Information Processing Systems, 2016*, pages 1181–1189.

Lee, S. and Bareinboim, E. (2018). Structural causal bandits: Where to intervene? In *Annual Conference on Neural Information Processing Systems, 2018*, pages 2573–2583.

Lee, S. and Bareinboim, E. (2019). Structural causal bandits with non-manipulable variables. In *AAAI Conference on Artificial Intelligence, 2019*, pages 4164–4172. AAAI Press.

Lindgren, E. M., Kocaoglu, M., Dimakis, A. G., and Vishwanath, S. (2018). Experimental design for cost-aware learning of causal graphs. In *Annual Conference on Neural Information Processing Systems, 2018*, pages 5284–5294.

Lu, T., Pál, D., and Pál, M. (2010). Contextual multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics, 2010*, pages 485–492.

Lu, Y., Meisami, A., Tewari, A., and Yan, W. (2020). Regret analysis of bandit problems with causal background knowledge. In *Conference on Uncertainty in Artificial Intelligence, 2020*, pages 141–150. PMLR.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, USA.

Pearl, J. (2009). *Causality*. Cambridge university press.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535.

Sachidananda, V. and Brunskill, E. (2017). Online learning for causal bandits. Available at https://web.stanford.edu/class/cs234/past_projects/2017/2017_Sachidananda_Brunskill_Causal_Bandits_Paper.pdf.

Sen, R., Shanmugam, K., Dimakis, A. G., and Shakkottai, S. (2017a). Identifying best interventions through online importance sampling. In *International Conference on Machine Learning, 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3057–3066. PMLR.

Sen, R., Shanmugam, K., Kocaoglu, M., Dimakis, A. G., and Shakkottai, S. (2017b). Contextual bandits with latent confounders: An NMF approach. In *International Conference on Artificial Intelligence and Statistics, 2017*, volume 54 of *Proceedings of Machine Learning Research*, pages 518–527. PMLR.

Tian, J. and Pearl, J. (2002). A general identification condition for causal effects. In *National Conference on Artificial Intelligence and Conference on Innovative Applications of Artificial Intelligence, 2002*, pages 567–573.

Tran-Thanh, L., Chapman, A., Rogers, A., and Jennings, N. R. (2012). Knapsack based optimal policies for budget-limited multi-armed bandits. In *AAAI Conference on Artificial Intelligence, 2012*, pages 1134–1140.

Tran-Thanh, L., Chapman, A. C., de Cote, E. M., Rogers, A., and Jennings, N. R. (2010). Epsilon-first policies for budget-limited multi-armed bandits. In *AAAI Conference on Artificial Intelligence, 2010*.

Yabe, A., Hatano, D., Sumita, H., Ito, S., Kakimura, N., Fukunaga, T., and Kawarabayashi, K. (2018). Causal bandits with propagating inference. In *International Conference on Machine Learning, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5508–5516. PMLR.