Regret Analysis of Bandit Problems with Causal Background Knowledge

Yangyi Lu
Department of Statistics
University of Michigan
yylu@umich.edu

Amirhossein Meisami
Adobe Inc.
meisami@adobe.com

Ambuj Tewari
Department of Statistics
University of Michigan
tewaria@umich.edu

Zhenyu Yan
Adobe Inc.
wyan@adobe.com

Abstract

We study how to learn optimal interventions sequentially given causal information represented as a causal graph along with associated conditional distributions. Causal modeling is useful in real world problems like online advertisement where complex causal mechanisms underlie the relationship between interventions and outcomes. We propose two algorithms, causal upper confidence bound (C-UCB) and causal Thompson Sampling (C-TS), that enjoy improved cumulative regret bounds compared with algorithms that do not use causal information. We thus resolve an open problem posed by Lattimore et al. (2016). Further, we extend C-UCB and C-TS to the linear bandit setting and propose causal linear UCB (CL-UCB) and causal linear TS (CL-TS) algorithms. These algorithms enjoy a cumulative regret bound that only scales with the feature dimension. Our experiments show the benefit of using causal information. For example, we observe that even with a few hundreds of iterations, the regret of causal algorithms is less than that of standard algorithms by a factor of three. We also show that under certain causal structures, our algorithms scale better than the standard bandit algorithms as the number of interventions increases.

1 INTRODUCTION

In a multi-armed bandit (MAB) problem, an agent adaptively learns to pull arms from a finite set of arms based on the past knowledge. At each pull, it observes a single reward corresponding to the arm pulled and its goal is to maximize the cumulative reward received within a time horizon. Bandit models are widely used in various applications, such as education (Williams et al., 2016), clinical trials (Villar et al., 2015; Tewari and Murphy, 2017) and marketing (Burtini et al., 2015; Mersereau et al., 2009).

There are many well-studied stochastic bandit algorithms, such as upper confidence bound (UCB) (Auer et al., 2002) and Thompson Sampling (TS) (Agrawal and Goyal, 2012), that can both achieve a regret bound $\tilde{O}(\sqrt{KT})$[1], where K is the number of arms and T is the time horizon. However, in many real world applications where we search for good interventions, the number of actions (interventions) is extremely large. An intervention here is defined as a forcible change to the value of a set of variables.

As an example of a real world problem with a large space of available interventions, we focus on the email campaign problem. Online advertising companies are constantly looking for an optimal trade-off between exploration and exploitation efforts in order to convert a potential buyer to an actual buyer. In case of email campaigns, the overall target is to maximize the user interaction with the emails that could be defined as opening an email, clicking on a link or eventually buying a product. To achieve these goals, marketers adjust several variables in the process. For instance, they may know that the length of subject, the template, the time of day to send, the product and the type (promotion, online events, etc.) of an email can affect whether a customer who receives the email will click the links inside or not. Every possible assignment of values to these variables can be an intervention leading to an extremely large number of interventions. Therefore, strategic utilization of such interventions is necessary for maximizing the cumulative user conversion throughout the campaign horizon.

A natural approach to deal with a large number of interventions is to exploit relationships between the way dif-

[1] $\tilde{O}$ ignores constant and poly-logarithmic factors.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

ferent interventions affect the outcome. In this paper, we focus on causal relations among interventions. In particular, we use causal graphs (Pearl, 2000) to represent relationships between interacting variables in a complex system. We study the following problem: using previously acquired knowledge about the causal graph structure, how to quickly learn good interventions sequentially (Sen et al., 2017; Hyttinen et al., 2013)? Our goal is to optimize over a given set of interventions in a sequential decision making framework where the dependence among reward distribution of these interventions is captured through a causal structure.

Lattimore et al. (2016) proposed two causal bandit algorithms, but they only provided simple regret guarantees and their bounds scale with the number of interventions in the worst case. Indeed, one of the open problems in their paper is to design algorithms that enjoy a $\tilde{O}(\sqrt{T})$ cumulative regret bound, and utilize the causal structure at the same time. Cumulative regret is appropriate when both exploration and exploitation are needed, while simple regret is useful when it is important to identify a good intervention at the end of a pure exploration phase. In many real world problems, we are not simply looking for the best intervention as quickly as possible without consideration of outcomes obtained during the exploration phase. In email campaign or clinical trials problems, a good policy should lead to high revenue and conversions or good health outcomes cumulatively, which are not what a pure exploration method can achieve. Therefore we focus on cumulative regret in this paper.

1.1 OUR CONTRIBUTIONS

We propose two natural and efficient algorithms, causal UCB (C-UCB) and causal TS (C-TS), by incorporating the available causal knowledge in UCB and TS for multi-armed bandit problems. We use causal knowledge to greatly reduce the amount of exploration needed to achieve low cumulative regret.

Suppose there are N variables that are related to the reward and each of them takes on k distinct values, which means changing the value of any of these variables can affect the reward distribution. Note the number of interventions can be as large as $(k+1)^N$, which means that standard bandit algorithms are only guaranteed to achieve $\tilde{O}(\sqrt{(k+1)^N T})$ regret. Our proposed causal algorithms exploit the causal knowledge to achieve $\tilde{O}(\sqrt{(k+1)^n T})$ regret[2], where n is the number of variables that have *direct* causal effects on the reward. These bounds suggest that causal UCB and TS algorithms are preferable to standard UCB and TS algorithms when $n \ll N$.

We further extend the causal bandit algorithms to linear bandit setting, that leads to our causal linear UCB (CL-UCB) and causal linear TS (CL-TS) algorithms. We show that CL-UCB and CL-TS both achieve $\tilde{O}(d\sqrt{T})$ regret, where d is the dimension of the coefficient vector in the linear reward model.

To complement our upper bounds, we also provide a lower bound for standard UCB algorithm. For some structured bandit instances with $n < N$, we show a lower bound on the cumulative regret of standard UCB which comes arbitrarily close to $\Omega(\sqrt{(k+1)^N T})$, which is much larger than the upper bounds of our proposed algorithms that utilize causal structures. It demonstrates that a standard MAB algorithm is *provably* worse than causal algorithms in certain cases.

Our experiments show the benefit of using causal structure: we observe (see Section 5, Figure 2) that within hundreds of iterations, our causal algorithms are already achieving regret within $1/3$ of the standard algorithms' regret. In addition, we validate numerically that for certain causal graph structure, C-UCB, C-TS, CL-UCB and CL-TS indeed scale better than standard multi-armed bandit algorithms as the size of intervention set grows.

1.2 RELATED WORK

Causal bandit problems can be treated as multi-armed bandit problems by simply ignoring the causal structure information and the extra observations. So existing bandit algorithms such as UCB (Auer et al., 2002) and TS (Agrawal and Goyal, 2012) can be applied. However, causal information should help us learn about an intervention based on the performance of other interventions, which can accelerate the whole learning process.

Combinatorial bandits (Cesa-Bianchi and Lugosi, 2012) also deal with an action set that is exponentially large. For example, the action set is usually a subset of the d-dimensional binary hypercube. In combinatorial bandits, the goal is to feasibly learn in the large action space by assuming certain structure (e.g., linear) in the reward dependence on actions and the availability of an efficient optimization solver over the action set. However, our emphasis is on reducing the statistical complexity by exploiting the given causal structures.

We build on the work of Lattimore et al. (2016). They studied the problem of identifying the best interventions in a stochastic bandit environment with known causal graph and some conditional probabilities of variables in the graph. They proposed two algorithms depending on

[2]Our regret bounds for confidence bound based algorithms will be frequentist while for Thompson sampling they will be Bayesian.

the type of causal graphs: parallel graph/general graph, and proved two simple regret bounds accordingly. Both bounds scale with a measure for causal graph's underlying distribution, which is small if every intervention has similar effect on the reward and can be as large as the number of interventions otherwise. Moreover, their algorithm for general graph contains as many parameters as the number of interventions, which are hard to tune. We focus on the cumulative regret and our algorithms are universal for all directed acyclic causal graphs defined in Section 2 with no tuning parameters other than that of standard MAB algorithms.

Another work (Sen et al., 2017) also considered best intervention identification via importance sampling, and their interventions are soft. Instead of forcing a node to take a specific value, soft intervention only changes the conditional distribution of a node given its parent nodes. However, they also only considered simple regret and their bounds scale with the number of interventions. Sachidananda and Brunskill (2017) studied the most closest setting as our paper. They showed the effectiveness of their causal Thompson Sampling method, but did not provide any regret analysis. Lee and Bareinboim (2018) empirically showed that a brute-force way to apply standard bandit algorithms on all interventions can suffer huge regret. Therefore they proposed a way to carefully choose an intervention subset by observing the causal graph structures. Our lower bound (Theorem 4) provides a theoretical explanation for the phenomenon they observe, namely that brute-force algorithms that try all possible interventions can incur huge regret.

2 PROBLEM SETUP

We follow standard terminology and notation (Koller and Friedman, 2009) to state the causal bandit problem introduced by Lattimore et al. (2016). A directed acyclic graph $\mathcal{G}$ is used to model the causal structure over a set of random variables $\mathcal{X} = \{X_1, \ldots, X_N\}$. Let P denote the joint distribution over $\mathcal{X}$ that factorizes over $\mathcal{G}$. For simplicity, we assume each variable can take on k distinct values, but extending our algorithm to various k values for different variables poses no difficulty. The parents of a variable X_i, denoted by Pa_{X_i}, is the set of all variables X_j such that there is an edge from X_j to X_i in graph $\mathcal{G}$. A size m intervention (action) is denoted by $\text{do}(\mathbf{X} = \mathbf{x})$, which assigns the values $\mathbf{x} = \{x_1, \ldots, x_m\}$ to the corresponding variables $\mathbf{X} = \{X_1, \ldots, X_m\} \subset \mathcal{X}$. An empty intervention is $\text{do}()$. The intervention on $\mathbf{X}$ also removes all edges from Pa_{X_i} to X_i for each $X_i \in \mathbf{X}$. Thus the resulting underlying probability distribution that defines the graph is denoted by $P(\mathbf{X}^c|\text{do}(\mathbf{X} = \mathbf{x}))$ over $\mathbf{X}^c := \mathcal{X} \setminus \mathbf{X}$.

In this causal bandit problem, the reward variable Y is real-valued. A learner is given the causal model's graph $\mathcal{G}$[3], a set of interventions (actions) $\mathcal{A}$ and conditional distributions of parent variables of Y given an intervention $a \in \mathcal{A}$: $P(\text{Pa}_Y|a)$. We denote the expected reward for action $a = \text{do}(\mathbf{X} = \mathbf{x})$ and the optimal action a^* by:

$$\mu_a := \mathbb{E}\left[Y|\text{do}(\mathbf{X} = \mathbf{x})\right]$$
$$a^* := \text{argmax}_{a \in \mathcal{A}}\, \mu_a.$$

We assume $\mu_a \in [0, 1]$ for every $a \in \mathcal{A}$. In round t, the learner pulls $a_t = \text{do}(\mathbf{X}_t = \mathbf{x}_t)$ based on previous round knowledge and causal information, then observes the reward Y_t and the values of Pa_Y, denoted by $\mathbf{Z}_{(t)} = \{z_1(t), \ldots, z_n(t)\}$, where n is the number of reward's parent variables. However, in the work of Lattimore et al. (2016), they need to observe the values of all variables after taking an action. Thus, comparing to them, the problem we face is more challenging. We know there are k^n different value assignments on Pa_Y, for convenience, we denote them by $\mathbf{Z}_1, \ldots, \mathbf{Z}_{k^n}$, where each $\mathbf{Z}_i$ is a vector of length n.

The objective of the learner is to minimize the expected cumulative regret $\mathbb{E}\left[R_T\right] = T\mu^* - \sum_{t=1}^{T} \mathbb{E}\left[\mu_{a_t}\right]$ using causal knowledge.

Bayesian Regret: Let $\omega \in \Omega$ denote the entire parameters of the distribution of $Y|_{\text{Pa}_Y = \mathbf{Z}}$. Reward can be expressed by $Y = \mathbb{E}\left[Y|_{\text{Pa}_Y = \mathbf{Z}}\right] + \epsilon$, where ϵ is a 1-subgaussian error variable. Thus, the cumulative regret R_T for a given ω can be formally written as $R_T(\omega)$. We particularly focus on the case where ω is random with distribution Q and bound the following Bayesian regret: $BR_T = \mathbb{E}_{\omega \sim Q} \mathbb{E}_\epsilon R_T(\omega)$.

Worst Case Regret: Using same notations as above, the worst case (frequentist) regret is defined as: $\max_{\omega \in \Omega} \mathbb{E}_\epsilon R_T(\omega)$. $\mathbb{E}R_T$ is used to represent the worst case regret from now for short.

3 CAUSAL BANDIT ALGORITHMS

In this section we propose and analyze algorithms for achieving minimal regret when causal information is known. We generalize standard UCB and standard TS algorithms to their causal counterparts in a natural way. We show how the regret bounds of the causal versions scale with a factor that can be much smaller than what would be the case for the standard algorithms. We also extend linear bandit algorithms to their causal version and demonstrate how it further helps us reduce the cumulative regret.

[3]Even though our algorithms take $\mathcal{G}$ as input, the only information used is the identity of Pa_Y variables.

Algorithm 1 C-UCB

Input: Horizon T, action set $\mathcal{A}$, δ, causal graph $\mathcal{G}$, number of parent variables n, number of values each parent variable can take on: k.
Initialization: Values assignment to parent variables: $\mathbf{Z}_j$, $\hat{\mu}_{\mathbf{Z}_j}(0) = 0$, $T_{\mathbf{Z}_j}(0) = 0$, for $j = 1, \ldots, k^n$.
for $t = 1, \ldots, T$ **do**
  for $j = 1, \ldots, k^n$ **do**
    $$\text{UCB}_{\mathbf{Z}_j}(t-1) = \hat{\mu}_{\mathbf{Z}_j}(t-1) + \sqrt{\frac{2\log(1/\delta)}{1 \vee T_{\mathbf{Z}_j}(t-1)}}.$$
  end for
  $a_t = \text{argmax}_{a \in \mathcal{A}} \sum_{j=1}^{k^n} \text{UCB}_{\mathbf{Z}_j}(t-1) P(\text{Pa}_Y = \mathbf{Z}_j | a)$
  Pull arm a_t and observe reward Y_t and its parent nodes' values $\mathbf{Z}_{(t)}$.
  Update $T_{\mathbf{Z}_j}(t) = \sum_{s=1}^{t} \mathbb{1}_{\{\mathbf{Z}_{(s)} = \mathbf{Z}_j\}}$ and $\hat{\mu}_{\mathbf{Z}_j}(t) = \frac{1}{T_{\mathbf{Z}_j}(t)} \sum_{s=1}^{t} Y_s \mathbb{1}_{\{\mathbf{Z}_{(s)} = \mathbf{Z}_j\}}$, for $j = 1, \ldots, k^n$.
end for

3.1 CAUSAL MAB ALGORITHMS

In the first part of this section we consider causal MAB problem and present causal upper confidence bound algorithm (C-UCB) and causal Thompson Sampling algorithm (C-TS).

3.1.1 Causal UCB (C-UCB)

Without causal knowledge, UCB algorithm updates the confidence interval of the reward mean for each arm. At every round, the learner chooses the arm with the highest upper confidence bound value. However, thanks to causal graph structures, we are able to make use of the expectation decomposition formula

$$\mu_a = \sum_{j=1}^{k^n} \mathbb{E}\left[Y | \text{Pa}_Y = \mathbf{Z}_j\right] P(\text{Pa}_Y = \mathbf{Z}_j | a).$$

At every round t, Algorithm 1 only updates the reward mean and upper confidence bound for every possible value assignment on reward's parent variables denoted by $\text{UCB}_{\mathbf{Z}_j}(t-1)$ as $P(\text{Pa}_Y = \mathbf{Z}_j | a)$ terms are known. It provides the upper confidence bound for each arm by:

$$\mathbf{UCB}_a(t-1) = \sum_{j=1}^{k^n} \text{UCB}_{\mathbf{Z}_j}(t-1) P(\text{Pa}_Y = \mathbf{Z}_j | a).$$

We pull a_t that can maximize $\mathbf{UCB}_a(t-1)$ over all $a \in \mathcal{A}$. There remain fewer upper confidence bounds to construct since usually $k^n < (k+1)^N$, so it is reasonable to expect that the cumulative regret can be reduced.

Theorem 1 (Regret Bound for C-UCB)**.** *Let $Y|_{Pa_Y = \mathbf{Z}_j} = \mathbb{E}\left[Y | Pa_Y = \mathbf{Z}_j\right] + \epsilon$, for $j = 1, \ldots, k^n$, where ϵ is a mean zero, 1-subgaussian distributed random error. If $\delta = 1/T^2$, the regret of policy defined in Algorithm 1 is bounded by*

$$\mathbb{E}\left[R_T\right] = \tilde{O}\left(\sqrt{k^n T}\right).$$

3.1.2 Causal TS (C-TS)

Thompson Sampling (TS) algorithm needs to update the posterior distributions for all arms. In this problem, there are $(k+1)^N$ distributions to update, which leads to huge regret when N is large. Similar to UCB algorithms, causal information can greatly help TS improve the performance when $k^n < (k+1)^N$. Again, by using the expectation decomposition formula $\mu_a = \sum_{j=1}^{k^n} \mathbb{E}\left[Y | \text{Pa}_Y = \mathbf{Z}_j\right] P(\text{Pa}_Y = \mathbf{Z}_j | a)$, our C-TS algorithm only updates the posterior distributions for $Y|_{\text{Pa}_Y = \mathbf{Z}_j}, j = 1, \ldots, k^n$ as the $P(\text{Pa}_Y = \mathbf{Z}_j | a)$ terms are known.

We provide two C-TS algorithms where Algorithm 2 uses Beta distribution as its prior and Algorithm 3 uses Gaussian distribution as its prior. At every round t, both C-TS algorithms sample from the posterior distributions for $Y|_{\text{Pa}_Y = \mathbf{Z}_j}, j = 1, \ldots, k^n$, then construct the estimated reward mean denoted by $\hat{\mu}_a$ for $\forall a \in \mathcal{A}$ using causal information. The intervention arm with the highest estimated reward will be pulled, reward Y_t and parent node values $\mathbf{Z}_{(t)}$ will be revealed accordingly. Parameters for Beta or Gaussian distribution are updated according to Beta-Bernoulli and Gaussian-Gaussian prior-posterior updating formulas.

Theorem 2 (Bayesian Regret Bound for C-TS)**.** *Let $Y|_{Pa_Y = \mathbf{Z}_j} = \mathbb{E}\left[Y | Pa_Y = \mathbf{Z}_j\right] + \epsilon$, for $j = 1, \ldots, k^n$, where ϵ is a mean zero, 1-subgaussian distributed random error. Then the Bayesian regret of policies in Algorithm 2 and Algorithm 3 are both be bounded by:*

$$BR_T = \tilde{O}\left(\sqrt{k^n T}\right).$$

3.2 CAUSAL LINEAR BANDIT ALGORITHMS

Previous section demonstrates how we use causal knowledge to improve the multi-armed bandit algorithms. In our setting, the reward Y directly depends on its n parent nodes, then a natural extension is to consider the linear modeling case: $Y|_{\text{Pa}_Y = \mathbf{Z}} = f(\mathbf{Z})^T \theta + \epsilon$, where f denotes the feature function applied on the parent nodes of Y, θ denotes the linear coefficient and ϵ is a zero mean, 1-subgaussian distributed random error.

We can write the expected reward mean for $\forall a \in \mathcal{A}$ as:

$$\mu_a = \langle \sum_{j=1}^{k^n} f(\mathbf{Z}_j) P(\text{Pa}_Y = \mathbf{Z}_j | a), \theta \rangle.$$

Algorithm 2 C-TS with Beta Prior (If $Y \in [0,1]$)

Input: Horizon T, action set $\mathcal{A}$, causal graph $\mathcal{G}$, all $P(Pa_Y|a)$, number of parent variables n, number of values each parent variable can take on: k.
Initialization: Value assignments to parent variables: $\mathbf{Z}_j$, $S^0_{\mathbf{Z}_j} = F^0_{\mathbf{Z}_j} = 1$, for $j = 1, \ldots, k^n$.
for $t \in \{1, \ldots, T\}$ **do**
  Sample $\hat{\theta}_j(t)$ from beta distn with parameters $(S^{t-1}_{\mathbf{Z}_j}, F^{t-1}_{\mathbf{Z}_j})$, for $j = 1, \ldots, k^n$.
  for action $a \in \mathcal{A}$ **do**
    $\hat{\mu}_a = \sum_{j=1}^{k^n} \hat{\theta}_j(t) P(\text{Pa}_Y = \mathbf{Z}_j|a)$
  end for
  $a_t = \text{argmax}_a \, \hat{\mu}_a$
  Pull arm a_t and observe reward $\tilde{Y}_t$ and its parent nodes values of $\mathbf{Z}_{(t)}$. Perform a Bernoulli trial with success probability Y_t and observe the output Y_t.
  if $Y_t = 1$ **then**
    $S^t_{\mathbf{Z}_{(t)}} = S^{t-1}_{\mathbf{Z}_{(t)}} + 1$
  else
    $F^t_{\mathbf{Z}_{(t)}} = F^{t-1}_{\mathbf{Z}_{(t)}} + 1$
  end if
end for

To this point, we demonstrate that linearly modeling the reward's parent nodes is just a special case of standard linear bandit problem, where the feature vector for $a \in \mathcal{A}$ is $m_a := \sum_{j=1}^{k^n} f(\mathbf{Z}_j) P(\text{Pa}_Y = \mathbf{Z}_j|a)$. Thus, we easily extend C-UCB and C-TS to this particular linear bandit setting.

Causal linear UCB (CL-UCB) algorithm (Algorithm 4) and causal linear TS (CL-TS) algorithm (Algorithm 5) are straightforward linear UCB and linear TS algorithms. It is helpful in the sense that the regret dependence on $\sqrt{k^n}$ can be further reduced to the dimension of linear coefficient θ denoted by d while linear reward over parent variables holds.

Theorem 3 (Regret Bound for CL-UCB & CL-TS adapted from Chapter 19 in Lattimore and Szepesvári (2020))**.** *Assume that $\|\theta\|_2 \leq 1$ and $\|f(\mathbf{Z})\|_2 \leq 1$, the dimension of θ and $f(\mathbf{Z})$ are both d, then run CL-UCB with $\beta = 1 + \sqrt{2\log(T) + d\log\left(1 + \frac{T}{d}\right)}$ and CL-TS, the regret of CL-UCB and Bayesian regret of CL-TS can both be bounded by*

$$\mathbb{E}\left[R_{T_{CL-UCB}}\right], BR_{T_{CL-TS}} = \tilde{O}\left(d\sqrt{T}\right).$$

Remark: Our algorithms are easily adapted to a more general setup, e.g. there exist a set of observable variables $\mathbf{W}$ that d-separates the manipulable variables and the reward variable and $P(\mathbf{W}|a)$ are known for all realizations $\mathbf{W}, a$. In this senario, one can replace Pa_Y and $P(\text{Pa}_Y|a)$ in above algorithms with $\mathbf{W}$ and $P(\mathbf{W}|a)$ and achieve $\tilde{O}\left(\sqrt{|\mathbf{W}|T}\right)$ regret, where $|\mathbf{W}|$ refers to the number of realizations of $\mathbf{W}$. This is beneficial when $|\mathbf{W}| \ll \mathcal{A}$ or the reward's direct parents are not known nor observable, but the variables $\mathbf{W}$ are.

Algorithm 3 C-TS with Gaussian Prior

Input: Horizon T, action set $\mathcal{A}$, causal graph $\mathcal{G}$, all $P(Pa_Y|a)$, number of parent variables n, number of values each parent variable can take on: k.
Initialization: Value assignments to parent variables: $\mathbf{Z}_j$, $k_{\mathbf{Z}_j} = 0$, $\hat{\mu}_{\mathbf{Z}_j} = 0$, for $j = 1, \ldots, k^n$.
for $t \in \{1, \ldots, T\}$ **do**
  Sample $\hat{\theta}_j(t) \sim N(\hat{\mu}_{\mathbf{Z}_j}, \frac{1}{k_{\mathbf{Z}_j}+1})$,for $j = 1, \ldots, k^n$.
  for action $a \in \mathcal{A}$ **do**
    $\hat{\mu}_a = \sum_{j=1}^{k^n} \hat{\theta}_j(t) P(\text{Pa}_Y = \mathbf{Z}_j|a)$
  end for
  $a_t = \text{argmax}_a \, \hat{\mu}_a$
  Pull arm a_t and observe the parent nodes values of Y denoted by $\mathbf{Z}_{(t)}$ and reward Y_t.
  Update $k_{\mathbf{Z}_t} := k_{\mathbf{Z}_t} + 1$
  Update $\hat{\mu}_{\mathbf{Z}_t} := \frac{\hat{\mu}_{\mathbf{Z}_t} k_{\mathbf{Z}_t} + Y_t}{k_{\mathbf{Z}_t}+1}$
end for

4 LOWER BOUND FOR NON-CAUSAL METHODS

In this section, we show that it is necessary to use an algorithm that utilizes the causal structure. We prove that there exists a simple bandit environment with causal information, for which the regret of the standard UCB algorithm scales *at least* exponentially with the size N of *all variables*. For the same environment, our regret upper bounds of C-UCB and C-TS scale *at most* exponentially with the size n of *parents*. Since it is possible to have $N \gg n$, this demonstrates the necessity of using causal bandits algorithms.

We now describe the environment. The bandit environment ν has N variables $X_1, \ldots, X_N$, each can take a value from $\{1, 2\}$. The marginal distribution for X_i is $P(X_i = 1) = p_i$, for $i = 1, \ldots, N$. The reward node Y is generated by $Y = \Delta X_1 + \epsilon$, where Δ is a positive coefficient to be determined and $\epsilon \sim \mathcal{N}(0, 1)$. Actions are denoted by $do(X_1 = i_1, \ldots, X_N = i_N)$, where $i_1, \ldots, i_N \in \{0, 1, 2\}$, and 0 is an additional dimension for the case that we do not set any value for a variable.

In this example, there are three types of actions:

- Type 0: Actions with $i_1 = 0$.

Algorithm 4 Causal Linear UCB (CL-UCB)

Input: horizon T, action set $\mathcal{A}$, all $P(Pa_Y|a)$.
Initialization: $V_0 = I_d$, $\hat{\theta}_0 = 0_d$, $g = 0_d$, $\beta = 1 + \sqrt{2\log(T) + d\log\left(1 + \frac{T}{d}\right)}$.
for $t = 1, \ldots, T$ **do**
  for $a \in \mathcal{A}$ **do**
    $\text{UCB}_a(t) = \max_{\theta \in \mathcal{C}_t} \langle \theta, m_a \rangle = \langle \hat{\theta}_{t-1}, m_a \rangle + \beta \|m_a\|_{V_{t-1}^{-1}}$, where $\mathcal{C}_t = \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}_{t-1} \right\|_{V_{t-1}} \leq \beta \right\}$
  end for
  $a_t = \text{argmax}_{a \in \mathcal{A}} \text{UCB}_a(t)$
  Pull arm a_t and observe parent node $Z_{(t)}$ and reward Y_t.
  Update $V_t = V_{t-1} + m_{a_t} m_{a_t}^T$, $g = g + m_{a_t} Y_t$, $\hat{\theta}_t = V_t^{-1} g$
end for

Algorithm 5 Causal Linear TS (CL-TS)

Input: Horizon T, action set $\mathcal{A}$, all $P(Pa_Y|a)$, standard deviation parameter v.
Initialization: $V_0 = I_d$, $\hat{\theta} = 0_d$, $g = 0_d$.
for $t \in \{1, \ldots, T\}$ **do**
  Sample $\tilde{\theta}_t \sim N(\hat{\theta}, v^2 V_t^{-1})$
  for action $a \in \mathcal{A}$ **do**
    $\hat{\mu}_a(t) = \langle m_a, \tilde{\theta}_t \rangle$
  end for
  $a_t = \text{argmax}_a \hat{\mu}_a(t)$
  Pull arm a_t and observe the parent nodes values of Y denoted by $Z_{(t)}$ and reward Y_t.
  Update $V_t = V_{t-1} + m_{a_t} m_{a_t}^T$, $g = g + m_{a_t} Y_t$ and $\hat{\theta} = V_t^{-1} g$
end for

- Type 1: Actions with $i_1 = 1$.
- Type 2: Actions with $i_1 = 2$.

The expected reward for three types actions are $2\Delta - p_1\Delta$, Δ and 2Δ respectively. Type 2 actions are optimal arms, while the gaps for type 0 and type 1 are $p_1\Delta$ and Δ respectively.

Now we present the lower bound of standard UCB for this environment.

Theorem 4 (Lower Bound for Standard UCB)**.** *For any $\epsilon > 0$, there exists a constant $C_\epsilon > 0$ such that the following holds. In the bandit environment ν described above, running standard UCB algorithm for T steps will incur regret at least $C_\epsilon \sqrt{3^N} T^{1/2-\epsilon}$.*

This theorem can be generalized to provide lower bounds for a broad class of MAB algorithms (p-order policies, see appendix), including standard TS. We give a proof outline of this theorem. The main idea is to apply Theorem 5 in Lattimore and Szepesvári (2020). This is an algorithm-dependent lower bound that shows if an algorithm has a uniform regret upper bound for all instances in the unstructured bandit environment class (defined below), then it must have a particular instance-dependent regret lower bound. We first show ν belongs to the unstructured bandit environment class. Next, since the standard UCB has a uniform regret upper bound for this class, we can apply Theorem 5 in Lattimore and Szepesvári (2020) to obtain a lower bound of standard UCB for ν.

Now we give more details. The unstructured Gaussian bandit environment class is defined as follows.

Definition 1 (Unstructured Gaussian bandit environment class)**.** *A Gaussian bandit environment class $\mathcal{E}$ is unstructured if $\mathcal{A}$ is finite and there exists set of Gaussian distributions $\mathcal{M}_a := \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 \leq 1\}$ for each $a \in \mathcal{A}$ such that*

$$\mathcal{E} = \{\nu = (P_a : a \in \mathcal{A}) : P_a \in \mathcal{M}_a, \forall a \in \mathcal{A}\}.$$

Note this environment class is the Cartesian product over all distributions in $\mathcal{M}_a$ for each arm. This is a large class, and in particular it contains the environment ν, which we formalize in the claim below.

Claim 1. *Denote a unstructured K-arm Gaussian bandit environment class by $\mathcal{E}_K(\mathcal{N})$. Given any causal graph $\mathcal{G}$ and conditional probabilities $P(Pa_Y|a), \forall a \in \mathcal{A}$ where Y is the reward variable and Pa_Y are its parents, for any bandit instance ν' that satisfies:*

- *arms are K interventions over a set of variables that are consistent with $\mathcal{G}$ and the corresponding conditional probabilities, and*
- *the conditional reward given parent values are independent Gaussian distributions:*

$$Y|_{Pa_Y = \mathbf{Z}} = \mathbb{E}[Y|Pa_Y = \mathbf{Z}] + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, 1)$,

we have that $\nu' \in \mathcal{E}_K(\mathcal{N})$.

The proof of this claim is given in the appendix. We can finish the proof of Theorem 4 by applying Theorem 16.4 in Lattimore and Szepesvári (2020) (presented in Theorem 5 in the appendix) on the environment ν.

Note that Theorem 5 in Lattimore and Szepesvári (2020) cannot be applied to causal algorithms. Causal algorithms proposed in this paper can only perform well

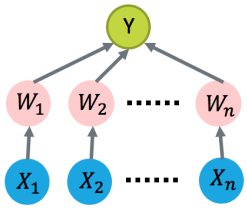

Figure 1: Causal Graph for Pure Simulation: only blue variables can be intervened.

on environments equipped with the fixed input causal graph $\mathcal{G}$ and the corresponding conditional probabilities, and thus causal algorithms cannot provide uniform (sublinear) regret upper bounds for all environments in the unstructured bandit environment class.

5 EXPERIMENTS

We compare the performance of standard bandit with causal bandit algorithms to validate that causal information plays an important role in bandit algorithms. We also show that when the reward is truly generated by a linear combination of the reward's parent node, CL-TS and CL-UCB can further achieve smaller regrets comparing with C-TS and C-UCB that only use causal structures but not the linear property.

5.1 PURE SIMULATION

We set up a pure simulation environment that will allow us to run scaling experiments in order to qualitatively test the scaling predictions of our theory. Throughout our pure simulations, we use a model in which there is a reward variable Y, reward's parent variables $W_1, \ldots, W_n$, taking values from $\{1, 2\}$, and non-parent variables $X_1, \ldots, X_n$, taking values from $\{1, \ldots, m\}$. Reward Y directly depends on its parent variables $W_1, \ldots, W_n$, while each parent variable W_i directly depends on the corresponding non-parent variable X_i $(i = 1, \ldots, n)$. The causal graph is displayed in Figure 1.

Intervention set: Denote an intervention by

$$a = \text{do}(X_1 = i_1, \ldots, X_n = i_n),$$

where $i_1, \ldots, i_n \in \{0, 1, \ldots, m\}$, 0 is an additional dimension for the case that we do not set any value for a variable. That means only non-parent variables can be intervened, the parent variables of the reward are not under control.

Reward Y is generated by: $Y = \langle f(W_1, \ldots, W_n), \theta \rangle + \epsilon$, where f is a function applied on parent variables, θ is a n-dimensional vector, ϵ is a sub-gaussian random error.

5.1.1 A Gentle Start: $m = 3, n = 4$

We begin with a simple case where $m = 3, n = 4$. The marginal distributions for X_1, X_2, X_3, X_4 and conditional probabilities for $W_i = 1|X_i, i = 1, \ldots, 4$ are displayed in Table 1 (Section B).

For simplicity, we set $f(W_1, W_2, W_3, W_4) := (W_1, W_2, W_3, W_4)$, and the error is a Gaussian variable $\epsilon \sim N(0, 0.1^2)$.

UCB algorithms: The true linear coefficient θ is $(0.25, 0.25, -0.25, -0.25)$. To approximate the expected regret, for each UCB algorithm we plot the average regret over 20 simulations.

TS algorithms: We plot both of the regret under $\theta = (0.25, 0.25, -0.25, -0.25)$ and the Bayesian regret. For the frequentist one, the procedure is same as UCB algorithms described above. For the Bayesian one, the "true" parameter θ is generated from its prior distribution $N(0, 0.1^2 I_4)$ for 20 times as Monte Carlo simulation. Then we plot the averaged regret over these 20 simulations to approximate the Bayesian regret.

Regret comparison plots are displayed in Figure 2.

5.1.2 Scaling with Non-Parent Variables' Range: m

In this section, we fix $n = 4$ while changing the domain range of non-parent variables m from 2 to 6 and see how it affects the performance of all six algorithms.

In each simulation, the marginal probabilities for each non-parent variable X_i: $\{P(X_i = j)\}_{j=1}^m$ are generated from independent Dirichlet distributions with parameter $\alpha = \mathbb{1}_m$ and the conditional probabilities $P(W_i = 1|X_i = j), i = 1, \ldots, n; j = 1, \ldots, m$ are generated randomly from $[0, 1]$. Throughout we fix the $\theta = (0.25, 0.25, -0.25, -0.25)$. For each algorithm, the final regret is averaged over 20 simulations. Regret comparison plot is displayed in Figure 3.

5.1.3 Scaling with Size of Parent Variables: n

In this section we fix $m = 3$ while changing the number of parent/non-parent variables n from 2 to 6. Since X_i takes value from $\{1, 2, 3\}$ and W_i takes value from $\{1, 2\}$, by adding additional pair $W_i \sim X_i$, the intervention size increases much faster than the number of value assignments on parent variables. We compare the performance of six algorithms.

In each simulation, the marginal probabilities for each non-parent variable $\{P(X_i = j)\}_{j=1}^m$ and conditional probabilities for each parent variable $P(W_i = 1|X_i = j), j = 1, \ldots, m$ are sampled in the same way as Section 5.1.2. To keep the reward at the same scale as m

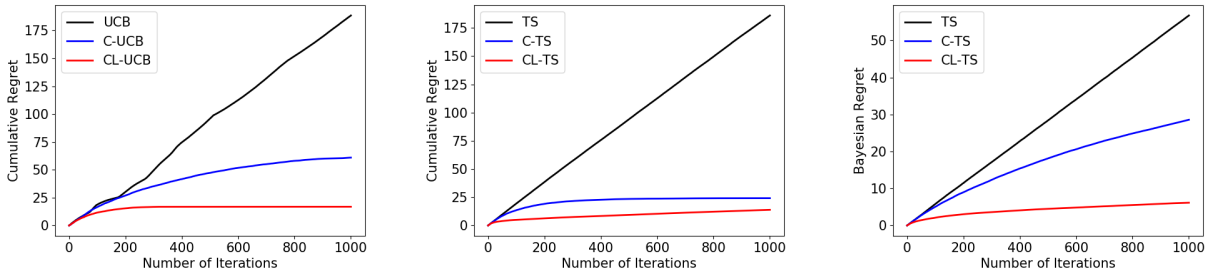

Figure 2: Regret comparison for $m = 3, n = 4$. Left: UCB regrets. Middle: TS regrets. Right: TS Bayesian regrets.

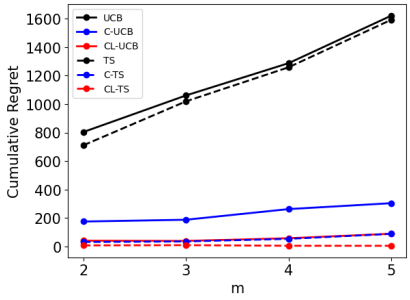

Figure 3: Cumulative regret v.s. m, fix $n = 4$, time horizon $T = 5000$.

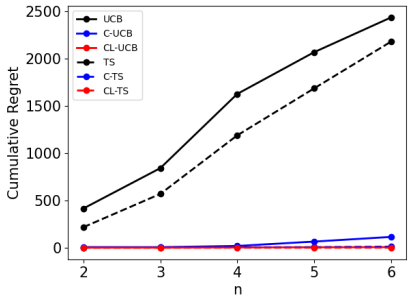

Figure 4: Cumulative regret v.s. n, fix $m = 3$, time horizon $T = 10000$.

varies, we use $\theta = (1, 0, \ldots, 0)$, where only the first element of linear coefficient is 1 and other elements are all zeros. For each algorithm, the final regret is averaged over 20 simulations. Regret comparison plot is displayed in Figure 4.

5.1.4 Conclusion of Pure Simulation

In Figure 2, the left and middle plots demonstrate the performance of algorithms for a fixed causal bandit environment. We observe that for UCB and TS, causal linear algorithms outperform the "non-linear" causal algorithms moderately and all causal algorithms outperform the standard bandit algorithms significantly. In the third plot, we demonstrate the performance in terms of Bayesian regret for three TS algorithms, and their performance order matches with the first two plots.

In Figure 3, we fix n and the time horizon T and compare the performance of the algorithm as m increases. The regret of C-UCB, C-TS, CL-UCB and CL-TS do not vary as m increases as their regret only depends on the size of parent variable value assignments. However, the regret of UCB and TS keeps increasing as m grows. Thus, we validate that the performance of our causal algorithms are not affected by the number of interventions on non-parent variables.

In Figure 4, we fix m and time horizon T and compare the performance of all algorithms as n grows. The regret of four causal algorithms does not vary a lot as n increases. We show in our theorem that in worst case, the regret of C-TS and C-UCB grow with $\sqrt{k^n}$ and the regret of CL-TS and CL-UCB grow with d for fixed time horizon. And we also observe in this simulation that for certain coefficient such as $\theta = (1, 0, \ldots, 0)$, the growth is even slower. Clearly the regret of standard UCB and TS algorithms keeps increasing as n grows.

5.2 EMAIL CAMPAIGN DATA

The experimental set up in this section is inspired by the email campaign data from Adobe. The reward variable is binary: whether the commercial links inside the email are clicked or not by the recipient. Features under control are "product", such as Photoshop, Acrobat XI Pro, Adobe Stock, etc., "purpose", such as awareness, promotion, operation, nurture, etc., "send out time" that includes morning, afternoon and evening. Even though these features are highly correlated with the reward variable, but they are not the direct causes. The variables that are actually causing the email links clicking are: the subject length, two different email templates, send out time, so we set these variables as the reward's parents. The three features in blue that can be intervened are further connected with reward's parent variables as in Figure 5.

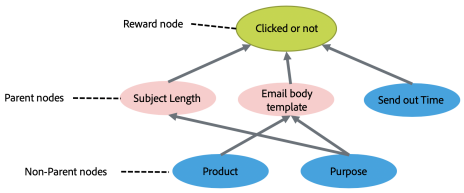

Figure 5: Causal Graph for Email Campaign: only blue nodes are under control.

Each combination of product and purpose has an email pool, once they are fixed, the company picks out emails from the pool. Thus, subject length and email body template cannot be intervened, they depend on the emails picking out from the pool, which is a random process.

From historical knowledge, email with "subject length" fewer than 7 words are more likely to be opened, so we denote "subject length" by Z_1, taking values from $\{1, 2\}$, representing "less than 7 words" or not. "Template" is denoted by Z_2, taking values from $\{1, 2\}$, representing template indices "1" or "2". "Send out time" is denoted by Z_3, taking values from $\{1, 2, 3\}$, representing "morning", "afternoon" and "evening". We consider "Photoshop" (1), "Acrobat XI Pro"(2), "Adobe Stock" (3) for the "product" variable, denoted by X_1; "Operational" (1), "Promo" (2), "Nurture" (3) and "Awareness" (4) for purpose variable, denoted by X_2.

The marginal probabilities for X_1 and X_2 and Z_3, conditional distributions for Z_1, Z_2 are displayed in Table 2 (Section B). The reward follows a Bernoulli distribution, with parameter $1 - (Z_1 + Z_2 + Z_3)/9$.

Interventions are denoted by $\text{do}(X_1 = i_1, X_2 = i_2, X_3 = i_3)$, where $i_1, i_3 \in \{0, 1, 2, 3\}$, $i_2 \in \{0, 1, 2, 3, 4\}$, 0 means no intervention on a variable.

In Figure 6, we compare the performance of UCB, C-UCB, TS (beta prior) and C-TS (beta prior). We plot the average regret over 20 simulations to approximate the expected cumulative regret for each method. Clearly both of C-UCB and C-TS outperforms UCB and TS significantly. Besides, we observe that TS algorithms generally perform better than the UCB algorithms. This phenomenon is also consistent with previous empirical discoveries (Chapelle and Li, 2011).

6 DISCUSSION & FUTURE WORK

We proposed C-UCB and C-TS algorithms and showed that their regret can be bounded by $\tilde{O}\left(\sqrt{k^n T}\right)$. We further extended linear bandit algorithms to their causal versions and showed the regret bound of CL-UCB and CL-TS can be reduced to $\tilde{O}\left(d\sqrt{T}\right)$. There are several interesting directions for future work.

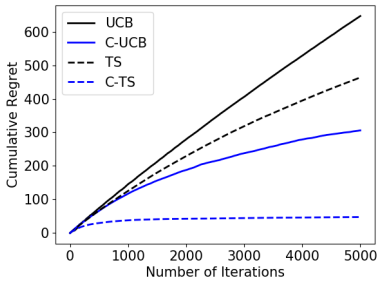

Figure 6: Regret Comparison of UCB, TS, C-UCB and C-TS for email campaign problem.

Extension to MDPs: We plan to extend our causal bandit framework to the MDP (Markov decision process) setting. The key feature of causal MDP is that there is an additional dimension: *state*, which can be affected by the previous intervention and the reward behaves differently under different status. This phenomenon is typical in many practical settings, including mobile health, online advertising and online education.

Learning causal structure: In many cases the causal structure is not known beforehand or only partially understood. Therefore it is desirable to develop methods that can recover the underlying causal structure and minimize the cumulative regret at the same time. An ideal algorithm that can efficiently learn the causal structure and the bandit together should achieve lower regret than normal bandit algorithms when the time horizon T is large. Ortega and Braun (2014) empirically shows that TS can recover causal structures in some cases. Combining causal learning algorithm with those that minimize cumulative regret is an interesting direction to investigate.

ACKNOWLEDGEMENT

Part of this work was done while Yangyi Lu was visiting Adobe. Ambuj Tewari would like to acknowledge the support of an Adobe Data Science Research Award, a Sloan Research Fellowship, and NSF grant CAREER IIS-1452099.

References

Agrawal, S. and Goyal, N. (2012). Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1.

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

Burtini, G., Loeppky, J. L., and Lawrence, R. (2015). Improving online marketing experiments with drifting multi-armed bandits. In *17th International Conference on Enterprise Information Systems (ICEIS (1))*, pages 630–636.

Cesa-Bianchi, N. and Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422.

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

Hyttinen, A., Eberhardt, F., and Hoyer, P. O. (2013). Experiment selection for causal discovery. *The Journal of Machine Learning Research*, 14(1):3041–3071.

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

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

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lee, S. and Bareinboim, E. (2018). Structural causal bandits: where to intervene? In *Advances in Neural Information Processing Systems*, pages 2568–2578.

Mersereau, A. J., Rusmevichientong, P., and Tsitsiklis, J. N. (2009). A structured multiarmed bandit problem and the greedy policy. *IEEE Transactions on Automatic Control*, 54(12):2787–2802.

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*. Cambridge University Press.

Sachidananda, V. and Brunskill, E. (2017). Online learning for causal bandits. `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. (2017). Identifying best interventions through online importance sampling. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3057–3066. JMLR. org.

Tewari, A. and Murphy, S. A. (2017). From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer.

Villar, S. S., Bowden, J., and Wason, J. (2015). Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 30(2):199.

Williams, J. J., Kim, J., Rafferty, A., Maldonado, S., Gajos, K. Z., Lasecki, W. S., and Heffernan, N. (2016). Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 379–388. ACM.

A Proof for Theorems

We prove Theorem 2 before Theorem 1, since the former one includes more technical steps and main parts of the two proofs are similar.

A.1 Proof of Theorem 2 (C-TS)

Proof. By definition, $\mu_a := E[Y|a] = \sum_{i=1}^{k^n} E[Y|Pa_Y = Z_i] P(Pa_Y = Z_i|a)$, $a^* = \text{argmax}_a \mu_a$.

Define:

$$T_Z(t) := \sum_{s=1}^{t} \mathbb{1}_{\{Z_{(s)}=Z\}},$$
$$\hat{\mu}_Z(t) := \frac{1}{T_Z(t)} \sum_{s=1}^{t} Y_s \mathbb{1}_{\{Z_{(s)}=Z\}},$$
$$\mu_Z := E[Y|Pa_Y = Z],$$

where $Z_{(s)}$ denotes the observed values of parent nodes for Y, in round s. Note that $\hat{\mu}_Z(t) = 0$ when $T_Z(t) = 0$.

Let E be the event that for all $t \in [T]$, $i \in [k^n]$ such that $\max_{a \in \mathcal{A}} P(Pa_Y = Z_i|a) > 0$, we have

$$|\hat{\mu}_{Z_i}(t-1) - \mu_{Z_i}| \leq \sqrt{\frac{2log(1/\delta)}{1 \vee T_{Z_i}(t-1)}}.$$

For fixed t and i, by Sub-Gaussian property, we can show

$$P\left(|\hat{\mu}_{Z_i}(t) - \mu_{Z_i}| \geq \sqrt{\frac{2\log(1/\delta)}{1 \vee T_{Z_i}(t)}}\right) = \mathbb{E}\left[P\left(|\hat{\mu}_{Z_i}(t) - \mu_{Z_i}| \geq \sqrt{\frac{2\log(1/\delta)}{1 \vee T_{Z_i}(t)}} \middle| Z_{(1)}, \ldots, Z_{(t)}\right)\right]$$
$$\leq \mathbb{E}[2\delta] = 2\delta.$$

By union bound, we have $P(E^c) \leq 2\delta T k^n$.

The Bayesian regret can be written as

$$BR_T = \mathbb{E}\left[\sum_{t=1}^{T} (\mu_{a^*} - \mu_{a_t})\right] = \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[\mu_{a^*} - \mu_{a_t}|\mathcal{F}_{t-1}]\right],$$

where $\mathcal{F}_{t-1} = \sigma(a_1, Z_1, Y_1, \ldots, a_{t-1}, Z_{t-1}, Y_{t-1})$.

The key insight is to notice that by definition of Thompson Sampling,

$$P(a^* = \cdot|\mathcal{F}_{t-1}) = P(a_t = \cdot|\mathcal{F}_{t-1}). \tag{1}$$

Further, define $\text{UCB}_a(t) := \sum_{j=1}^{k^n} \text{UCB}_{Z_j}(t) P(Pa_Y = Z_j|a)$, we can bound the conditional expected difference between optimal arm and the arm played at round t using equation 1 by

$$\begin{aligned}
&\mathbb{E}[\mu_{a^*} - \mu_{a_t}|\mathcal{F}_{t-1}] \\
&= \mathbb{E}[\mu_{a^*} - \text{UCB}_{a_t}(t-1) + \text{UCB}_{a_t}(t-1) - \mu_{a_t}|\mathcal{F}_{t-1}] \\
&= \mathbb{E}[\mu_{a^*} - \text{UCB}_{a^*}(t-1) + \text{UCB}_{a_t}(t-1) - \mu_{a_t}|\mathcal{F}_{t-1}].
\end{aligned}$$

Next by tower rule, we have

$$BR_T = \mathbb{E}\left[\sum_{t=1}^{T} (\mu_{a^*} - \text{UCB}_{a^*}(t-1) + \text{UCB}_{a_t}(t-1) - \mu_{a_t})\right].$$

On event E^c, by the original definition of BR_T we have $BR_T \leq 2T$. On event E, the first term is negative showing by the definition of $\text{UCB}_{\mathbf{Z}_j}, j = 1, \ldots, k^n$ and

$$\mu_{a^*} - \text{UCB}_{a^*}(t-1) = \sum_{j=1}^{k^n} \left(\mathbb{E}\left[Y|Pa_Y = Z_j\right] - \text{UCB}_{Z_j}(t-1)\right) P(Pa_Y = Z_j|a^*) \leq 0,$$

because $\mathbb{E}\left[Y|Pa_Y = Z_j\right] - \text{UCB}_{Z_j}(t-1) \leq 0$ on event E. Also on event E, the second term can be bounded by

$$\begin{aligned}
\mathbb{1}_E \sum_{t=1}^{T} \left(\text{UCB}_{a_t}(t-1) - \mu_{a_t}\right) &= \mathbb{1}_E \sum_{t=1}^{T} \sum_{j=1}^{k^n} \left(\text{UCB}_{Z_j}(t-1) - \mathbb{E}\left[Y|Pa_Y = Z_j\right]\right) P(Pa_Y = Z_j|a_t) \\
&\leq \mathbb{1}_E \sum_{t=1}^{T} \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(t-1)}} P(Pa_Y = Z_j|a_t) \\
&\leq \mathbb{1}_E \sum_{t=1}^{T} \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(t-1)}} \left(P(Pa_Y = Z_j|a_t) - \mathbb{1}_{\{Z_{(t)} = Z_j\}} + \mathbb{1}_{\{Z_{(t)} = Z_j\}}\right).
\end{aligned} \tag{2}$$

The second part of equation 2 can be bounded by

$$\begin{aligned}
\mathbb{1}_E \sum_{t=1}^{T} \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(t-1)}} \mathbb{1}_{\{Z_{(t)} = Z_j\}} &\leq \mathbb{1}_E \sum_{j=1}^{k^n} \int_0^{T_{Z_j}(T)} \sqrt{\frac{8\log(1/\delta)}{s}} ds \\
&\leq \sum_{j=1}^{k^n} \sqrt{32 T_{Z_j}(T) \log(1/\delta)} \\
&\leq \sqrt{32 k^n T \log(1/\delta)}.
\end{aligned}$$

For the first part of equation 2, we define $X_t := \sum_{s=1}^{t} \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(s-1)}} \left(P(Pa_Y = Z_j|a_s) - \mathbb{1}_{\{Z_{(s)} = Z_j\}}\right)$, $X_0 := 0$. Note that $\{X_t\}_{t=0}^{T}$ is a martingale sequence and we have

$$\begin{aligned}
|X_t - X_{t-1}|^2 &= \left| \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(t-1)}} \left(P(Pa_Y = Z_j|a_t) - \mathbb{1}_{\{Z_{(t)} = Z_j\}}\right) \right|^2 \\
&\leq 32 \log(1/\delta).
\end{aligned}$$

By applying Azuma's inequality we have

$$P(|X_T| > \sqrt{k^n T \log(T)} \log(T)) \leq \exp\left(-\frac{k^n \log^3(T)}{32 \log(1/\delta)}\right).$$

We take $\delta = 1/T^2$, combine the first and second part of equation 2, we show that with probability $1 - P(E^c) - \exp\left(-\frac{k^n \log^2(T)}{64}\right) = 1 - 2k^n/T - \exp\left(-\frac{k^n \log^2(T)}{64}\right)$,

$$R_T \leq 16\sqrt{k^n T \log(T)} \log(T).$$

Thus the Bayesian regret can be bounded by:

$$\begin{aligned}
\mathbb{E}\left[R_T\right] &\leq P(E^c) \times 2T + \exp\left(-\frac{k^n \log^2(T)}{64}\right) \times 2T + \sqrt{64 k^n T \log(T)} \log(T) \\
&\leq C\sqrt{k^n T \log(T)} \log(T).
\end{aligned}$$

where C is a constant and the above inequality holds for large T. Thus we have proved that $\mathbb{E}\left[R_T\right] = \tilde{O}\left(\sqrt{k^n T}\right)$. $\square$

A.2 Proof of Theorem 1 (C-UCB)

Proof. Let E be the event that for all $t \in [T]$, $j \in [k^n]$, we have

$$\left|\hat{\mu}_{Z_j}(t-1) - \mathbb{E}\left[Y|Pa_Y = Z_j\right]\right| \leq \sqrt{\frac{2\log(1/\delta)}{1 \vee T_{Z_j}(t-1)}}.$$

Use same proof idea in Theorem 2, we have $P(E^c) \leq 2\delta T k^n$. Define $\text{UCB}_a(t) := \sum_{j=1}^{k^n} \text{UCB}_{Z_j}(t) P(Pa_Y = Z_j|a)$, the regret can be rewritten as

$$\begin{aligned} R_T &= \sum_{t=1}^{T} (\mu_{a^*} - \mu_{a_t}) \\ &= \sum_{t=1}^{T} \left(\mu_{a^*} - \text{UCB}_{a_t}(t-1) + \text{UCB}_{a_t}(t-1) - \mu_{a_t}\right). \end{aligned}$$

On event E^c, $R_T \leq 2T$. On event E we can show

$$\begin{aligned} \mu_{a^*} - \text{UCB}_{a_t}(t-1) &= \sum_{j=1}^{k^n} \mathbb{E}\left[Y|Pa_Y = Z_j\right] P(Pa_Y = Z_j|a^*) - \sum_{j=1}^{k^n} \text{UCB}_{Z_j}(t-1) P(Pa_Y = Z_j|a_t) \\ &\leq \sum_{j=1}^{k^n} \text{UCB}_{Z_j}(t-1) P(Pa_Y = Z_j|a^*) - \sum_{j=1}^{k^n} \text{UCB}_{Z_j}(t-1) P(Pa_Y = Z_j|a_t) \leq 0, \end{aligned}$$

where the last inequality follows by the way to choose a_t in Algorithm 1, the second last inequality follows by the definition of event E. Thus on event E we have

$$\begin{aligned} R_T &\leq \sum_{t=1}^{T} \left(\text{UCB}_{a_t}(t-1) - \mu_{a_t}\right) \\ &= \sum_{t=1}^{T} \sum_{j=1}^{k^n} \left(\text{UCB}_{Z_j}(t-1) - \mathbb{E}\left[Y|Pa_Y = Z_j\right]\right) P(Pa_Y = Z_j|a_t) \\ &\leq \sum_{t=1}^{T} \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(t-1)}} P(Pa_Y = Z_j|a_t) \\ &\leq \sum_{t=1}^{T} \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(t-1)}} \left(P(Pa_Y = Z_j|a_t) - \mathbb{1}_{\{Z_{(t)} = Z_j\}} + \mathbb{1}_{\{Z_{(t)} = Z_j\}}\right). \end{aligned} \tag{3}$$

The second part of Equation 3 can be bounded by

$$\begin{aligned} \sum_{t=1}^{T} \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(t-1)}} \mathbb{1}_{\{Z_{(t)} = Z_j\}} &\leq \sum_{j=1}^{k^n} \int_0^{T_{Z_j}(T)} \sqrt{\frac{8\log(1/\delta)}{s}} ds \\ &\leq \sum_{j=1}^{k^n} \sqrt{32 T_{Z_j}(T) \log(1/\delta)} \\ &\leq \sqrt{32 k^n T \log(1/\delta)}. \end{aligned}$$

For the first part of equation 3, we define $X_t := \sum_{s=1}^{t} \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(s-1)}} \left(P(Pa_Y = Z_j|a_s) - \mathbb{1}_{\{Z_{(s)} = Z_j\}}\right)$, $X_0 := 0$. Note that $\{X_t\}_{t=0}^{T}$ is a martingale sequence.

$$\begin{aligned} |X_t - X_{t-1}|^2 &= \left| \sum_{j=1}^{k^n} \sqrt{\frac{8\log(1/\delta)}{1 \vee T_{Z_j}(t-1)}} \left(P(Pa_Y = Z_j|a_t) - \mathbb{1}_{\{Z_{(t)} = Z_j\}}\right) \right|^2 \\ &\leq 32\log(1/\delta). \end{aligned}$$

By applying Azuma's inequality we have

$$P(|X_T| > \sqrt{k^n T \log(T)} \log(T)) \leq \exp\left(-\frac{k^n \log^3(T)}{32 \log(1/\delta)}\right).$$

We take $\delta = 1/T^2$, combine the first and second part of equation 3, with probability $1 - P(E^c) - \exp\left(-\frac{k^n \log^2(T)}{64}\right) = 1 - 2k^n/T - \exp\left(-\frac{k^n \log^2(T)}{64}\right)$, the regret can be bounded by

$$R_T \leq 16\sqrt{k^n T \log(T)} \log(T).$$

Thus the expected regret can be bounded by:

$$\begin{aligned}\mathbb{E}[R_T] &\leq P(E^c) \times 2T + \exp\left(-\frac{k^n \log^2(T)}{64}\right) \times 2T + \sqrt{64 k^n T \log(T)} \log(T) \\ &\leq C\sqrt{k^n T \log(T)} \log(T)\end{aligned}$$

where C is a constant, above inequality holds for large T. Thus we prove $\mathbb{E}[R_T] = \tilde{O}\left(\sqrt{k^n T}\right)$ □

A.3 Proof of Theorem 3 (CL-TS)

Lemma 1. *(Lattimore and Szepesvári, 2020) Notations same as algorithm 4 and algorithm 5. Let $\delta \in (0, 1)$. Then with probability at least $1 - \delta$ it holds that for all $t \in \mathbb{N}$,*

$$\left\|\hat{\theta}_t - \theta\right\|_{V_t(\lambda)} \leq \sqrt{\lambda}\|\theta\|_2 + \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det V_t(\lambda)}{\lambda^d}\right)}.$$

Furthermore, if $\|\theta^*\| \leq m_2$, then $P(\exists t \in \mathbb{N}^+ : \theta^* \notin \mathcal{C}_t) \leq \delta$ with

$$\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \left\|\hat{\theta}_{t-1} - \theta\right\|_{V_{t-1}(\lambda)} \leq m_2\sqrt{\lambda} + \sqrt{2\log\left(\frac{1}{\delta}\right) + \log\left(\frac{\det V_{t-1}(\lambda)}{\lambda^d}\right)}\right\}.$$

Lemma 2. *(Lattimore and Szepesvári, 2020) Let $x_1, \ldots, x_T \in \mathbb{R}^d$ be a sequence of vectors with $\|x_t\|_2 \leq L < \infty$ for all $t \in [T]$, then*

$$\sum_{t=1}^{T}\left(1 \wedge \|x_t\|^2_{V_{t-1}^{-1}}\right) \leq 2\log(\det V_T) \leq 2d\log\left(1 + \frac{TL^2}{d}\right),$$

where $V_t = I_d + \sum_{s=1}^{t} x_s x_s^T$.

Proof. W define $\beta = 1 + \sqrt{2\log(T) + d\log\left(1 + \frac{T}{d}\right)}$ and $V_t = I_d + \sum_{s=1}^{t} m_{a_s} m_{a_s}^T$ same as Algorithm 5, where $m_a := \sum_{i=1}^{k^n} f(Z_i)P(Pa_Y = Z_i|a)$. Define upper confidence bound $\text{UCB}_t : \mathcal{A} \to \mathbb{R}$ by

$$\text{UCB}_t(a) = \max_{\theta \in \mathcal{C}_t}\langle\theta, m_a\rangle = < \hat{\theta}_{t-1}, m_a > + \beta\|m_a\|_{V_{t-1}^{-1}},$$

where $\mathcal{C}_t = \left\{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_{t-1}\right\|_{V_{t-1}} \leq \beta\right\}$. By Lemma 1, we have

$$P\left(\exists t \leq T : \left\|\hat{\theta}_{t-1} - \theta\right\|_{V_{t-1}} \geq 1 + \sqrt{2\log(T) + \log(\det V_t)}\right) \leq \frac{1}{T}.$$

And note $\|m_a\|_2 \leq 1$, thus by geometric means inequality we have

$$\det V_t \leq \left(trace(\frac{V_t}{d})\right)^d \leq \left(1 + \frac{T}{d}\right)^d.$$

Thus, by $\|\theta\|_2 \leq 1$,

$$P\left(\exists t \leq T : \left\|\hat{\theta}_{t-1} - \theta\right\|_{V_{t-1}} \geq 1 + \sqrt{2\log(T) + d\log\left(1 + \frac{T}{d}\right)}\right) \leq \frac{1}{T}.$$

Let E_t be the event that $\left\|\hat{\theta}_{t-1} - \theta\right\|_{V_{t-1}} \leq \beta$, $E := \cap_{t=1}^{T} E_t$, $a^* := \text{argmax}_a \sum_{i=1}^{k^n} \langle f(Z_i), \theta\rangle P(Pa_Y = Z_i|a)$, which is a random variable in this setting because θ is random. Then

$$\begin{aligned}
BR_T &= \mathbb{E}\left[\sum_{t=1}^{T}\left\langle\sum_{i=1}^{k^n} f(Z_i)\left(P\left(Pa_Y = Z_i|a^*\right) - P\left(Pa_Y = Z_i|a_t\right)\right), \theta\right\rangle\right] \\
&= \mathbb{E}\left[\mathbb{1}_{E^c}\sum_{t=1}^{T}\left\langle\sum_{i=1}^{k^n} f(Z_i)\left(P(Pa_Y = Z_i|a^*) - P(Pa_Y = Z_i|a_t)\right), \theta\right\rangle\right] \\
&+ \mathbb{E}\left[\mathbb{1}_{E}\sum_{t=1}^{T}\left\langle\sum_{i=1}^{k^n} f(Z_i)\left(P(Pa_Y = Z_i|a^*) - P(Pa_Y = Z_i|a_t)\right), \theta\right\rangle\right] \\
&\leq 2TP(E^c) + \mathbb{E}\left[\mathbb{1}_{E}\sum_{t=1}^{T}\left\langle\sum_{i=1}^{k^n} f(Z_i)\left(P(Pa_Y = Z_i|a^*) - P(Pa_Y = Z_i|a_t)\right), \theta\right\rangle\right] \\
&\leq 2 + \mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{E_t}\left\langle\sum_{i=1}^{k^n} f(Z_i)\left(P(Pa_Y = Z_i|a^*) - P(Pa_Y = Z_i|a_t)\right), \theta\right\rangle\right]. \qquad (4)
\end{aligned}$$

Again, we know from equation 1 such that $P(a^* = \cdot|\mathcal{F}_{t-1}) = P(a_t = \cdot|\mathcal{F}_{t-1})$, where $\mathcal{F}_{t-1} = \sigma(Z_1, a_1, Y_1, \ldots, Z_{t-1}, a_{t-1}, Y_{t-1})$. Thus we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}_{E_t}\left\langle\sum_{i=1}^{k^n} f(Z_i)\left(P\left(Pa_Y = Z_i|a^*\right) - P\left(Pa_Y = Z_i|a_t\right)\right), \theta\right\rangle\middle|\mathcal{F}_{t-1}\right] \\
=&\mathbb{1}_{E_t}\mathbb{E}\left[\left\langle\sum_{i=1}^{k^n} f(Z_i)\left(P\left(Pa_Y = Z_i|a^*\right) - P\left(Pa_Y = Z_i|a_t\right)\right), \theta\right\rangle\middle|\mathcal{F}_{t-1}\right] \\
=&\mathbb{1}_{E_t}\mathbb{E}\left[\left\langle\sum_{i=1}^{k^n} f(Z_i)P\left(Pa_Y = Z_i|a^*\right), \theta\right\rangle - UCB_t(a^*) + UCB_t(a_t) - \left\langle\sum_{i=1}^{k^n} f(Z_i)P(Pa_Y = Z_i|a_t), \theta\right\rangle\middle|\mathcal{F}_{t-1}\right] \\
\leq&\mathbb{1}_{E_t}\mathbb{E}\left[UCB_t(a_t) - \left\langle\sum_{i=1}^{k^n} f(Z_i)P(Pa_Y = Z_i|a_t), \theta\right\rangle\middle|\mathcal{F}_{t-1}\right] \\
\leq&\mathbb{1}_{E_t}\mathbb{E}\left[\left\langle\sum_{i=1}^{k^n} f(Z_i)P(Pa_Y = Z_i|a_t), \hat{\theta}_{t-1} - \theta\right\rangle\middle|\mathcal{F}_{t-1}\right] + \beta\left\|\sum_{i=1}^{k^n} f(Z_i)P(Pa_Y = Z_i|a)\right\|_{V_{t-1}^{-1}} \\
\leq&2\beta\left\|\sum_{i=1}^{k^n} f(Z_i)P(Pa_Y = Z_i|a)\right\|_{V_{t-1}^{-1}}.
\end{aligned}$$

Substituting into the second term of equation 4,

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{1}_{E_t}\left\langle\sum_{i=1}^{k^n}f(Z_i)\left(P(Pa_Y=Z_i|a^*)-P(Pa_Y=Z_i|a_t)\right),\theta\right\rangle\right]\\
\leq&2\mathbb{E}\left[\beta\sum_{t=1}^{T}\left(1\wedge\left\|\sum_{i=1}^{k^n}f(Z_i)P(Pa_Y=Z_i|a)\right\|_{V_{t-1}^{-1}}\right)\right]\\
\leq&2\sqrt{T\mathbb{E}\left[\beta^2\sum_{t=1}^{T}\left(1\wedge\left\|\sum_{i=1}^{k^n}f(Z_i)P(Pa_Y=Z_i|a)\right\|^2_{V_{t-1}^{-1}}\right)\right]}\text{ (By Cauchy-Schwartz)}\\
\leq&2\sqrt{2dT\beta^2\log\left(1+\frac{T}{d}\right)}\text{ (By Lemma 2).}
\end{aligned}$$

Putting together we prove

$$BR_T\leq 2+2\sqrt{2dT\beta^2\log\left(1+\frac{T}{d}\right)}=\tilde{O}\left(d\sqrt{T}\right). \tag{5}$$

□

A.4 Proof of Theorem 3 (CL-UCB)

Proof. Define $\beta=1+\sqrt{2\log(T)+d\log\left(1+\frac{T}{d}\right)}$, by Lemma 1 and above proof for CL-TS we have

$$\begin{aligned}
P(\exists t\leq T:\left\|\hat{\theta}_{t-1}-\theta^*\right\|_{V_{t-1}}\geq\beta)\leq\frac{1}{T},\\
P(\exists t\in\mathbb{N}^+:\theta^*\notin\mathcal{C}_t)\leq\frac{1}{T},
\end{aligned}$$

where $\mathcal{C}_t=\left\{\theta\in\mathbb{R}^d:\left\|\theta-\hat{\theta}_{t-1}\right\|_{V_{t-1}}\leq\beta\right\}$.

Let $\tilde{\theta}_t$ denote a θ that satisfies $\langle\tilde{\theta}_t,a_t\rangle=UCB_t(a_t)$. Again let E_t be the event that $\left\|\hat{\theta}_{t-1}-\theta^*\right\|_{V_{t-1}}\leq\beta$, let $E=\bigcap E_t$, $a^*=\text{argmax}_a\sum_{j=1}^{k^n}\langle f(Z_j),\theta\rangle P(Pa_Y=Z_j|a)$. Then on event E_t, using the fact that $\theta^*\in\mathcal{C}_t$ we have

$$\langle\theta^*,\sum_{j=1}^{k^n}f(Z_j)P(Pa_Y=Z_j|a^*)\rangle\leq UCB_t(a^*)\leq UCB_t(a_t)=\langle\tilde{\theta}_t,\sum_{j=1}^{k^n}f(Z_j)P(Pa_Y=Z_j|a_t)\rangle$$

Thus we can bound the difference of expected reward between optimal arm and a_t by

$$\begin{aligned}
\mu_{a^*}-\mu_{a_t}&=\langle\theta^*,\sum_{j=1}^{k^n}f(Z_j)P(Pa_Y=Z_j|a^*)\rangle-\langle\theta^*,\sum_{j=1}^{k^n}f(Z_j)P(Pa_Y=Z_j|a_t)\rangle\\
&\leq\langle\tilde{\theta}_t-\theta^*,\sum_{j=1}^{k^n}f(Z_j)P(Pa_Y=Z_j|a_t)\rangle\\
&\leq 2\wedge 2\beta\left\|\sum_{j=1}^{k^n}f(Z_j)P(Pa_Y=Z_j|a_t)\right\|_{V_{t-1}^{-1}}\\
&\leq 2\beta\left(1\wedge\left\|\sum_{j=1}^{k^n}f(Z_j)P(Pa_Y=Z_j|a_t)\right\|_{V_{t-1}^{-1}}\right).
\end{aligned}$$

So the expected regret can be further bounded by:

$$\begin{aligned}
\mathbb{E}[R_T] &= \mathbb{E}\left[\sum_{t=1}^{T}(\mu_{a^*}-\mu_{a_t})\right] = \mathbb{E}\left[\mathbb{1}_E\sum_{t=1}^{T}(\mu_{a^*}-\mu_{a_t})\right] + \mathbb{E}\left[\mathbb{1}_{E^c}\sum_{t=1}^{T}(\mu_{a^*}-\mu_{a_t})\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}(\mu_{a^*}-\mu_{a_t})\mathbb{1}_{E_t}\right] + \mathbb{E}\left[\mathbb{1}_{E^c}\sum_{t=1}^{T}(\mu_{a^*}-\mu_{a_t})\right] \\
&\leq 2\beta\sum_{t=1}^{T}\left(1\wedge\left\|\sum_{j=1}^{k^n}f(Z_j)P(Pa_Y=Z_j|a_t)\right\|_{V_{t-1}^{-1}}\right) + 2TP(E^c) \\
&\leq 2+2\beta\sqrt{T\sum_{t=1}^{T}\left(1\wedge\left\|\sum_{j=1}^{k^n}f(Z_j)P(Pa_Y=Z_j|a_t)\right\|^2_{V_{t-1}^{-1}}\right)} \text{ (By Cauchy-Schwartz)} \\
&\leq 2+2\beta\sqrt{2dT\log\left(1+\frac{T}{d}\right)} \text{ (By Lemma 2)}
\end{aligned}$$

□

A.5 Proof of Claim 1

Proof. Denote the reward variable for action a by $Y|_a$ and denote the reward variable given fixed parent values by $Y|_{\text{Pa}_Y=\mathbf{Z}}$. According to the causal information, $Y|_a$ can be represented as a weighted sum of $Y|_{\text{Pa}_Y=\mathbf{Z}}$:

$$Y|_a = \sum_{\mathbf{Z}} P(\text{Pa}_Y=\mathbf{Z}|a)Y|_{\text{Pa}_Y=\mathbf{Z}}. \tag{6}$$

In the statement of claim 1 we know that $Y|_{\text{Pa}_Y=\mathbf{Z}}$ are independent Gaussian distributions, therefore $Y|_a$, a weighted sum of Gaussian distributions still follows a Gaussian distribution. It remains to show the variance of $Y|_a$ is less than 1.

$$\begin{aligned}
\text{Var}(Y|_a) &= \sum_{\mathbf{Z}} P(\text{Pa}_Y=\mathbf{Z}|a)^2\text{Var}(Y|_{\text{Pa}_Y=\mathbf{Z}}) &(7)\\
&\leq \sum_{\mathbf{Z}} P(\text{Pa}_Y=\mathbf{Z}|a)^2 \leq \sum_{\mathbf{Z}} P(\text{Pa}_Y=\mathbf{Z}|a) = 1, &(8)
\end{aligned}$$

where the first inequality above uses the condition that $\text{Var}(Y|_{\text{Pa}_Y=\mathbf{Z}}) \leq 1$. We show that the reward for every arm $Y|_a$ is Gaussian distributed with variance less than 1, thus the bandit environment ν' described in the claim is an instance in Gaussian bandit environment class. □

A.6 Proof of Theorem 4

We first introduce an important concept.

Definition 2 (p-order Policy)**.** *For K-arm unstructured Gaussian bandit environments $\mathcal{E} := \mathcal{E}_K(\mathcal{N})$ and policy π, whose regret, on any $\nu \in \mathcal{E}$, is bounded by CT^p for some $C > 0$ and $p > 0$. We call this policy class $\Pi(\mathcal{E}, C, T, p)$, the class of p-order policies.*

Note that UCB and TS are in this class with $C = C'_\epsilon\sqrt{K}$ and $p = 1/2 + \epsilon$ with some $C'_\epsilon > 0$ for arbitrary small ϵ.

We use the following result to prove our theorem.

Theorem 5 (Finite-time, instance-dependent regret lower bound for p-order policies, Theorem 16.4 in Lattimore and Szepesvári (2020))**.** *Let $\nu \in \mathcal{E}_K(\mathcal{N})$ be a K-arm Gaussian bandit with mean vector $\mu \in \mathbb{R}^K$ and suboptimality gaps $\Delta \in [0,\infty)^K$. Let*

$$\mathcal{E}(\nu) = \{\nu' \in \mathcal{E}_K(\mathcal{N}) : \mu_i(\nu') \in [\mu_i, \mu_i + 2\Delta_i]\}.$$

Suppose π is a p-order policy such that $\exists C > 0$ and $p \in (0,1)$, $R_T(\pi, \nu') \leq CT^p$ for all T and $\nu' \in \mathcal{E}(\nu)$. Then for any $\epsilon \in (0,1]$,

$$\mathbb{E}R_T(\pi,\nu) \geq \frac{2}{(1+\epsilon)^2} \sum_{i:\Delta_i>0} \left(\frac{(1-p)\log(T) + \log(\frac{\epsilon\Delta_i}{8C})}{\Delta_i} \right)^+,$$

where $(x)^+ = \max(x, 0)$ is the positive part of $x \in \mathbb{R}$.

Proof of Theorem 4. Consider the bandit environment ν described in section 4. By claim 1 we know ν is an instance in unstructured Gaussian bandit environment class, so we can further apply Theorem 5. The size of three types of actions are all $3^N/3$. For Type 1 actions, its gap compared to the optimal actions is Δ, for Type 0 actions, gap is $p_1\Delta$. Plugging into the results of Theorem 5, for every p-order policy over $\mathcal{E}(\nu)$, we have

$$\mathbb{E}R_T(\pi,\nu) \geq \frac{1}{2}\frac{3^N}{3} \left(\frac{(1-p)\log(T) + \log(\frac{\Delta}{8C})}{\Delta} \right)^+ + \frac{1}{2}\frac{3^N}{3} \left(\frac{(1-p)\log(T) + \log(\frac{p_1\Delta}{8C})}{p_1\Delta} \right)^+. \tag{9}$$

In particular, choose $\Delta = 8\rho CT^{p-1}$, we get

$$(1-p)\log(T) + \log(\frac{\Delta}{8C}) = \log(\rho),$$
$$(1-p)\log(T) + \log(\frac{p_1\Delta}{8C}) = \log(p_1\rho).$$

Note that $\sup_{\rho>0} \log(\rho)/\rho = \exp(-1) \approx 0.35$, and we next plug above two equations in Equation 9 to get

$$\mathbb{E}R_T(\pi,\nu) \geq \frac{3^N}{3}\frac{0.35}{8CT^{p-1}}.$$

Now consider π to be UCB, by plugging in $C = C'_\epsilon\sqrt{3^N}$ and $p = 1/2 + \epsilon$ we have

$$\mathbb{E}R_T(UCB,\nu) \geq \frac{0.35}{24C'_\epsilon}\sqrt{3^N}T^{1/2-\epsilon}.$$

□

B Probability Tables Used in Experiments

i	1	2	3
$P(X_1 = i)$	0.3	0.4	0.3
$P(X_2 = i)$	0.3	0.3	0.4
$P(X_3 = i)$	0.5	0.3	0.2
$P(X_4 = i)$	0.25	0.25	0.5
$P(W_1 = 1 \mid X_1 = i)$	0.2	0.5	0.8
$P(W_2 = 1 \mid X_2 = i)$	0.3	0.2	0.8
$P(W_3 = 1 \mid X_3 = i)$	0.4	0.6	0.5
$P(W_4 = 1 \mid X_4 = i)$	0.3	0.5	0.6

Table 1: Marginal and conditional probabilities for pure simulation experiment in section 5.1.1, numbers are randomly selected.

i	1	2	3	4
$P(X_1 = i)$	0.2	0.2	0.6	
$P(X_2 = i)$	0.05	0.6	0.3	0.05
$P(Z_3 = i)$	0.5	0.2	0.3	
$P(Z_1 = 1 \mid X_2 = i)$	0.7	0.7	0.3	0.3
$P(Z_2 = 1 \mid X_1 = 3, X_2 = i)$	0.6	0.7	0.6	0.5
$P(Z_2 = 1 \mid X_1 \neq 3, X_2 = i)$	0.8	0.9	0.5	0.2

Table 2: Marginal and conditional probabilities for email campaign causal graph.