Online Learning for Causal Bandits

Vinayak Sachidananda [1] **joint work w/ Prof. Emma Brunskill** [2]

Abstract

The Causal Multi-Arm Bandit framework (Lattimore & Reid, 2016) allows for modeling sequential decision problems in causal environments. In previous works, online learning in the Causal MAB framework has not been analyzed. We propose an algorithm, Online Causal Thompson Sampling (OC-TS), for online decision making in such environments and perform simulations to understand the performance of OC-TS compared to offline algorithms.

1. Introduction

The Multi-Armed Bandit (MAB) model is a framework for analyzing the exploration/exploitation trade-off inherent in sequential decision problems. The motivating example for studying MABs comes from clinical trials where we have N treatments and would like to learn the efficacy of each in ailing a particular disease. Given that patients arrive sequentially, we would like to use our past experiences of treatments to decide which treatment should be administered. The difficulty of the problem arises in deciding whether to try new or poorly understood treatments (exploring) or sticking with the current best treatment (exploiting). Inherently, we would like to learn about all of the treatments with as little exploration as possible in order to minimize the cumulative regret over our entire sequence of decisions. Today, the MAB problem has numerous applications in Advertising, Healthcare, Education, Finance, and a variety of other domains.

In our problem setting, we are concerned with learning bandit policies in causal environments. Our motivation for learning in causal environments stems from a wide range of decision making settings in Healthcare, Education, and Advertising where actions have inherent causal dependencies. This type of problem is best explained by an example. Consider a Healthcare example where a physician would like to understand how to best influence the health of a patient by minimizing their likelihood of contracting gum disease. The physician can recommend that the patient either floss daily, avoid sugary foods, or use mouthwash. It is reasonable to believe that these actions have causal dependencies of two forms. First, if the physician makes too many suggestions concurrently the patient may be overwhelmed and not follow the suggestions. Secondly, it is possible that if the patient is recommended one of the possible actions, conditioned on the patient's new habits, she may inadvertently follow one of the other recommendations without being prompted by the physician to do so. Despite the ubiquity of this type of decision making setting, there has been limited work on the modeling and analysis of online causal decision systems.

In this type of decision-making environment, there is a clear incentive to learn the causal dependencies between actions and use this information to make high value interventions. The difficulty of this problem is that we now must: (1) learn the structure of our causal graphical model and (2) perform the online estimation of action values that is standard in bandit problems. This interplay is well studied in the algorithm we propose which makes the assumption that these two effects can be independently factorized and learned effectively. For notation, we denote Pa_Y to be the parents of the reward variable Y in our causal graph and performing an intervention a is equivalent to $do(X = a)$. Specifically, our algorithm modifies Thompson Sampling to sample from (1) the conditional posterior of $Pa_Y|do(X = a)$ as well as the conditional posterior of the $Y|Pa_Y$. To the best of our knowledge, this is the first algorithm for online learning in causal bandit environments. Due to the general difficulty of this estimation task, we will primarily analyze 'sparse' causal graphs or causal graphs, with $|V|$ nodes, in which the number of edges $|E|$ is much smaller than $\binom{|V|}{2}$. Furthermore, we will study causal graphical models with Bernoulli distributions to provide faster convergence of our sample means and posterior approximations.

2. Background

We will provide an overview of the Multi-Armed Bandit problem, causal graphical models, and the Thompson Sampling algorithm. This will provide a formal definition of our environment as well as introduce the notion of posterior

[1]Department of Electrical Engineering, Stanford University [2]Department of Computer Science, Stanford University. Correspondence to: Sachidananda, V. <vsachi@stanford.edu>.

sampling based methods for bandit learning.

2.1. Multi-Armed Bandits

Consider the Bernoulli Multi-Armed Bandit problem in which we have N arms and at each time step $t = 1, 2, \ldots, T$ over a finite horizon T, one of the N arms is played. After an arm i is played, a reward $Y_t \in \{0, 1\}$, which is independent of previous plays and i.i.d. from the selected arm's distribution, will be observed. An MAB algorithm must efficiently use observations from the previous $t-1$ plays to estimate reward distributions for each of the N arms and choose which arm to play at time t. One measure of an MAB algorithm's performance is it's ability to maximize expected reward over a horizon T: $\mathbb{E}[\sum_{t=1}^{T} \mu_{A(t)}]$ where $A(t)$ is the arm played at time t and $\mu_{A(t)}$ is the selected arm's expected reward. Another commonly used measure for providing bounds on algorithmic performance is the concept of expected cumulative regret over a finite horizon T: $\mathbb{E}[\mathcal{R}(T)] = \mathbb{E}[\sum_{t=1}^{T}(\mu^* - \mu_{A(t)})] = \sum_i \Delta_i \, \mathbb{E}[k_i(T))]$, where $\mu^* = \max_i \mu_i$, $\Delta_i := \mu^* - \mu_i$ and $k_i(t)$ denotes the number of times arm i has been played up to time t. In deriving regret bounds for our MAB algorithms, we will prove an upper bound on the expected cumulative regret over a finite horizon T.

2.2. Causal Multi-Armed Bandits

In our causal model, following the terminology of (Koller & Friedman, 2009) we have a directed acyclic graph $\mathcal{G}$, a set of variables $\mathcal{X} = \{X_1, X_2, \ldots, X_N\}$, and a joint distribution $\mathbb{P}$ over $\mathcal{X}$ that factorizes over $\mathcal{G}$. Each variable is Bernoulli and the existence of an edge from variable X_i to X_j conditions the probability distribution of X_j i.e. $\mathbb{P}(X_j = x_j | X_i = x_i) \neq \mathbb{P}(X_j = x_j)$. We denote the parents of a variable X_i, or the subset of $\mathcal{X}$ such that there exists an edge from X_j to X_i, to be Pa_{X_i}. An intervention (of size m) is represented as $do(X = x)$ which sets the values of $x = \{x_1, x_2, ..., x_m\}$ to the corresponding variables in $\mathcal{X}$. When an intervention is performed, all edges between X_i and Pa_{X_i} are mutilated and the graph G can be represented by an altered probability distribution $\mathbb{P}\left(X^c | do(X = x)\right)$ where $X^c = \mathcal{X} - X$.

Our agent in the online causal bandit problem is given a set of allowed actions $\mathcal{A}$ and limited knowledge of the graph $\mathcal{G}$. One of the variables, $Y \in \mathcal{X}$, in our graph $\mathcal{G}$ is the reward variable and can be represented by a Bernoulli Random Variable. The expected reward for an action $a \in \mathcal{A}$ can be represented as $\mu_a = \mathbb{E}[Y | do(X = a)]$ and the expected reward for the optimal action $\mu_{a^*} = \max_{a \in \mathcal{A}} \mathbb{E}[Y | do(X = a)]$. After choosing an action a, the agent observes realized values of observable variables in $\mathcal{X}$, X_t, and a reward Y_t. Using these realizations, the agent updates his best estimates of expected rewards for each $a \in \mathcal{A}$ and repeats this procedure for T episodes. We define cumulative regret after T episodes as being $R(T) = \mu_{a^*} T - \sum_{t=1}^{T} \mu_{\hat{a}_t^*}$ where $\hat{a}_t^*$ represents the estimate of the optimal action at time t.

This environment formulation is modeled off of the setting in (Lattimore & Reid, 2016). Distinctively, we focus on an online learning representation and measure performance in cumulative regret rather than simple regret. Note that the problem formulation is different than contextual bandits in that we observe our set of variables X_t only after selecting an action and, therefore, cannot construct a best action through observations of X_t.

2.3. Thompson Sampling

The Thompson Sampling (TS) algorithm initializes with a Bayesian prior of the reward distribution of each arm. After each observation Y_t, TS will update the posterior distribution of the arm that was selected in timestep t. If we initialize our prior to be uniform over the unit interval, our best estimate $\hat{\mu}_i$ of the expected reward for each arm i can be derived simply by taking the point $\theta \in (0, 1)$ that maximizes our posterior distribution (and likelihood function): $\hat{\mu}_i = \text{argmax}_{\theta \in (0,1)} \, P_i(\theta \mid x_1, x_2, \ldots, x_n) \propto f_i(x_1, x_2, \ldots, x_n \mid \theta)$. Furthermore, a plausible expected reward for arm i, θ_i^t, can be sampled from our estimated posterior $P_i^t(\theta \mid x_1, x_2, \ldots, x_n)$.

Generally in the case of Bernoulli bandits, TS will start with a uniform prior over support $[0, 1]$ which can be represented by a $(1, 1)$ beta distribution. Given that we will have to update our posterior distribution at the end of each trial, a beta distribution which has the form $f(x, \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)+\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$ $\alpha, \beta \in \mathbb{Z}_{++}$, is a friendly reward distribution choice due to it's simplistic posterior updates. Concretely, we can represent each arm's posterior distribution on $P_i^t(\theta \mid x_1, x_2, \ldots, x_n)$ as $Beta(S_i(t-1) + 1, F_i(t-1) + 1)$ where $S_i(t-1)$ represents the number of successes from Bernoulli trials for plays from arm i and conversely $F_i(t-1)$ represents the number of failures from Bernoulli trials for plays from arm i. In order to perform an update on the posterior distribution at timestep t for the selected arm i at time t, we simply increment $S_i(t) = S_i(t-1) + \mathbb{1}_{\{Y_t = 1\}}$ and $F_i(t) = F_i(t-1) + \mathbb{1}_{\{Y_t = 0\}}$. Note that a dirichlet distribution allows for extension when there are more than two classes and will be used for modeling causal dependencies in our proposed algorithm.

3. Related Work

We will review two general frameworks for bandit learning in causal environments. First, we consider scenarios with

unobserved confounders in observational distributions and examine algorithms for learning optimal policies during experimentation (Bareinboim & Pearl, 2015). Also, we will review experiments aimed at learning the effect of interventions in causal graphs and, by exploiting the learned causal structure, making interventions that maximize reward variables (Lattimore & Reid, 2016).

3.1. Bandits with Unobserved Confounders

In (Bareinboim & Pearl, 2015), it is demonstrated that in the presence of confounding variables, it is useful to understand an agent's natural action if no intervention had taken place. To demonstrate this effect, in the presented experiments we are able to recover sufficient information to make an optimal decision only by looking at the agent's natural decision.

The proposed algorithm, Causal Thompson Sampling, aims to maximize $\text{argmax}_{a\in\mathcal{A}}\,\mathbb{E}[Y_{X=a} = 1|X = x]$, where x corresponds to the action the agent would have taken without intervention. Intuitively, this translates to choosing the action that maximizes the expected reward conditioned on the agent's intuited action. In contrast, Thompson Sampling without causal conditioning aims to maximize $\text{argmax}_{a\in\mathcal{A}}\,\mathbb{E}[Y|do(X = a)]$, which translates to choosing the action that maximizes the expected reward unconditioned on context or natural predilection. Operationally, the Causal Thompson Sampling Algorithm uses the observational distribution P_{obs} to seed intuitive actions, i.e. $\mathbb{E}[Y_{X=a}|X = x]\ \forall a = x$, and explores non-intuitive decisions during experimentation. Since this algorithm makes use of contextual variables for learning causal dependencies, it is unable to learn causal dependencies that cannot be explained by observed variables. As a result, the algorithm is not suited for our general causal bandit setting.

3.2. Learning Good Interventions via Causal Inference

In (Lattimore & Reid, 2016), it is demonstrated that by exploiting causal structure we can achieve regret bounds for bandit problems that are $\widetilde{\Theta}\left(\sqrt{\frac{m}{T}}\right)$ where m refers to the number of unbalanced variables and can be significantly less than $|\mathcal{A}|$. Something to emphasize, however, is that simple regret is optimized in the experimentation and, therefore, there is no cost to exploration until the very last timestep.

Two types of causal bandits are analyzed in this paper: (1) parallel bandits and (2) bandits with general causal graphs. For parallel bandits, we do not have any dependencies between variables $X_t = \{X_1, X_2, \ldots, X_N\}$ and can learn the effect of variables on a reward variable Y_t by observing X_t and Y_t for $\frac{T}{2}$ timesteps. In the second $\frac{T}{2}$ timesteps we can sample unbalanced variables or variables which have shown little variation in the values they take in the previous observations.

In the case of general causal graphs, we assume the distribution over the parents of our reward variable Y conditioned on taking action $a \in \mathcal{A}$, $\mathbb{P}\{Pa_Y|a\}$, is known. Using this assumption, we can then construct a mixture distribution over all interventions, $Q = \sum_{a\in\mathcal{A}} \eta_a\, \mathbb{P}\{Pa_Y|a\}$ where η is a measure defined over all interventions $a \in \mathcal{A}$ (specified by the experimenter). We sample T actions from η and use a truncated importance sampling estimator $R_a(X) = \frac{\mathbb{P}\{Pa_Y|a\}}{Q\{Pa_Y(X)\}}$ to estimate the returns μ_a for all $a \in \mathcal{A}$. Note that $Pa_Y(X)$ is the realized set of variables in X that are parents of Y and $R_a(X)$ is truncated to a fixed B_a, to provide concentration guarantees. We can estimate an action's expected rewards $\mu_a = \frac{1}{T}\sum_{t=1}^{T} Y_t R_a(X_t)\mathbb{1}_{\{R_a(X_t)\le B_a\}}$.

4. Online Causal Thompson Sampling

We propose an algorithm for online learning in the Causal Bandit setting. Online Causal Thompson Sampling (OCTS), uses the observation that $\mu_a = \mathbb{E}[Y|do(X = a)]$ can be decomposed by partitioning over our set of variables Pa_Y. Since each element of $Pa_Y = \{X_1, X_2, ..., X_N\}$ is a Bernoulli Random Variable, $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_{2^N}\}$ partitions our sample space into 2^N disjoint events. As a result, we have that $\mu_a = \mathbb{E}[Y|do(X = a)] = \sum_{k=1}^{2^N} \mathbb{P}\left(Y = 1|Pa_Y = Z_k\right)\mathbb{P}\left(Pa_Y = Z_k|do(X = a)\right)$.

Algorithm 1 Online Causal Thompson Sampling

Input: $\text{Beta}_{\mu_{Z_k}} = (1,1), \text{Dirichlet}_{\rho_a} = \mathbf{1}, S^0_{Z_k}, F^0_{Z_k} = 0$
for timestep $t = 1, 2, \ldots, T$ **do**
  for action $a = 1, 2, \ldots, 2N$ **do**
    $\mathbb{P}\left(Pa_Y = Z_k|do(X = a)\right) \sim dirichlet_{\rho_a}[k]$
    $\mathbb{P}\left(Y = 1|Pa_Y = Z_k\right) \sim beta_{\mu_{Z_k}}[0]$
    $\mu_{\hat{a}_t} = \sum_{k=1}^{2^N} P\left(Y = 1|Pa_Y = Z_k\right) P\left(Pa_Y = Z_k|do(X = a)\right)$
  end for
  $\hat{a}^*_t = \text{argmax}_{\hat{a}_t}\ \mu_{\hat{a}_t}$
  $\{X^c\} \sim P\left(X^c|do(X = \hat{a}^*_t)\right)$
  $Z_k = \{X = \hat{a}^*_t\} \cup \{X^c\}$
  $Y_t \sim Bern(P\left(Y = 1|Pa_Y = Z_k\right))$
  $\text{Dirichlet}_{\rho_{\hat{a}^*_t}}[k]$ += 1
  if $Y_t = 1$ **then**
    $S^t_{Z_k} = S^{t-1}_{Z_k} + 1$
  else
    $F^t_{Z_k} = F^{t-1}_{Z_k} + 1$
  end if
  $\text{Beta}^t_{\mu_{Z_k}} = (S^t_{Z_k} + 1, F^t_{Z_k} + 1)$
end for

Using this observation, we can modify Thompson Sampling to concurrently learn the reward distribution for each com-

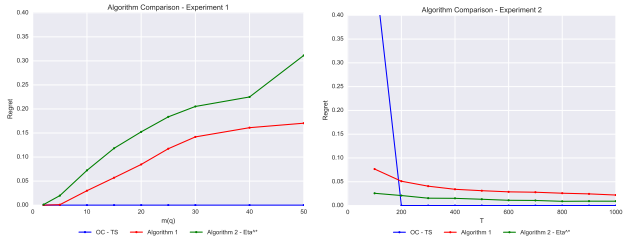
bination of variables, $\mathbb{P}(Y = 1|Pa_Y = Z_k)$, and the dependencies in our causal graph, $\mathbb{P}(Pa_Y = Z_k|do(X = a))$. We consider two variants of the algorithm with varying degrees of knowledge of the causal graph $\mathcal{G}$. In the first setting, we know only $\{Pa_Y\}$, the subset of $\mathcal{X}$ such that there exists an edge from each element of $\{Pa_Y\}$ to Y. In the second setting, we assume knowledge of $\mathbb{P}(Pa_Y = Z_k|do(X = a))$, the probability distributions of $\{Pa_Y\}$ conditioned on performing the intervention $do(X = a)$.

For the setting where $\mathbb{P}(Pa_Y = Z_k|do(X = a))$ is known we can use the true conditional distributions for $\mathbb{P}(Pa_Y = Z_k|do(X = a))$ when calculating $\mu_{\hat{a}_t}$ rather than sampling from our Dirichlet distributions as shown above. This modification greatly reduces the parameter set the algorithm is learning and empirically results in a significant decrease in cumulative regret. However, in many applications knowing $\mathbb{P}(Pa_Y = Z_k|do(X = a))$ may be unrealistic.

5. Experiments (Replication from (Lattimore & Reid, 2016))

Included are some experiments comparing both the simple and cumulative regret of our online causal bandits algorithm and the two offline algorithms proposed in (Lattimore & Reid, 2016) (denoted as Algorithm 1 and Algorithm 2 as in the original paper). Note that for the experiments in this section, we use a modified version of OC-TS that for the first $|\mathcal{A}|$ timesteps plays each action exactly once. As a result, in the experiments where the time horizon $T < |\mathcal{A}|$ we see that OC-TS has a high simple regret at time T as it is still collecting information about the environment. After the first $|\mathcal{A}|$ plays, OC-TS plays actions in an online manner and performs optimally for the conducted experiments.

5.1. Simple Regret Experiments



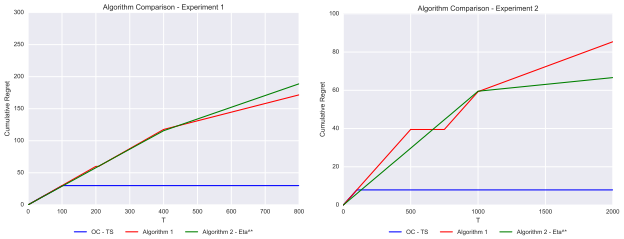
(a) Simple regret vs m(q) for fixed horizon $T = 400$, number of variables $N = 50$, and fixed $\epsilon = 0.3$

(b) Simple regret vs horizon, T, with $N = 50$, $m = 2$, and $\epsilon = \sqrt{\frac{N}{8T}}$

Figure 1. Simple Regret Comparison

These experiment settings are the same as those in (Lattimore & Reid, 2016). In all of the experiments, Y depends only on a single variable X_1 (unknown to the algorithms). $Y_t \sim Bernoulli(\frac{1}{2} + \epsilon)$ if $X_1 = 1$ and $Y_t \sim Bern(\frac{1}{2} + \epsilon')$ if $X_1 = 0$ where $\epsilon' = q_1\epsilon/(1 - q_1)$. We have an expected reward of $\frac{1}{2} + \epsilon$ for $do(X_1 = 1)$, $\frac{1}{2} - \epsilon'$ for $do(X_1 = 0)$ and $\frac{1}{2}$ for all other actions. We set $q_i = 0$ for $i \leq m$ and $\frac{1}{2}$ otherwise. We notice that our online algorithm outperforms both of the offline variants whenever $T > |\mathcal{A}|$.

5.2. Cumulative Regret Experiments



(a) Cumulative regret for fixed horizon $T = 800$, number of variables $N = 50$, $m(q) = 25$, and fixed $\epsilon = 0.3$

(b) Cumulative regret for fixed horizon $T = 2000$, number of variables $N = 50$, $m(q) = 2$, and fixed $\epsilon = \sqrt{\frac{N}{8T}}$

Figure 2. Cumulative Regret Comparison

Now, we measure the cumulative regret for same experiments with a fixed horizon T for each scenario (large ϵ and small ϵ). For each algorithm, we follow standard behavior for the first $\frac{T}{2}$ timesteps and fix the best estimated action for the second $\frac{T}{2}$ timesteps. We notice that our online algorithm's regret converges for both experiments unlike the offline variants.

6. Experiments: Chain and Confounded Causal Graphs

Included are some experiments for non-parallel causal graphs. Note that for the experiments in this section, we use a modified version of OC-TS that for the first $|\mathcal{A}|$ timesteps plays each action exactly once. As a result, in the experiments where the time horizon $T < |\mathcal{A}|$ we see that OC-TS has a high simple regret at time T as it is still collecting information about the environment. After the first $|\mathcal{A}|$ plays, OC-TS plays actions in an online manner and performs well for the conducted experiments.

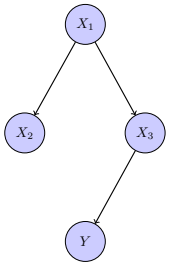
6.1. Chain Causal Graph

In the chain causal graph setting, Y depends only on a single variable X_3. $Y_t \sim Bernoulli(\frac{1}{2} + \epsilon)$ if $X_3 = 1$ and $Y_t \sim Bern(\frac{1}{2})$ if $X_1 = 0$. We have an expected reward of $\frac{1}{2} + \epsilon$

X_1	$P(X_2=1\|X_1)$	$P(X_3=1\|X_1)$
0	P_u	P_u
1	1-P_u	1-P_u

$P(Y=1\|X_3)$	
$X_3=0$	0.5
$X_3=1$	0.8

(a) Graph Conditional Probabilities



(b) Graph Depiction

Figure 3. Chain Causal Graph Problem Formulation

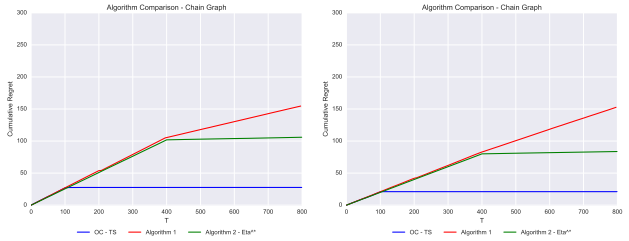
for $do(X_3 = 1)$, $(1-P_u)(\frac{1}{2}+\epsilon)+P_u(\frac{1}{2})$ for $do(X_1 = 1)$, and $P_u(\frac{1}{2}+\epsilon)+(1-P_u)(\frac{1}{2})$ for all other actions. We set $q_i = 0$ for $i \leq m, i \notin \{2,3\}$; $q_i = P_u$ for $i \in \{2,3\}$; and $\frac{1}{2}$ otherwise.

(a) Simple regret vs horizon T; $N = 50, m(q) = 30, P_u = 0.1$, and fixed $\epsilon = 0.3$

(b) Simple regret vs horizon T; $N = 50, m(q) = 30, P_u = 0.3$, and fixed $\epsilon = 0.3$

Figure 4. Simple Regret Comparison

We measure the simple regret for the chain causal graph experiments as we vary the horizon T for each $P_u \in \{0.1, 0.3\}$. We notice that our online algorithm performs optimally for all horizons $T > |\mathcal{A}|$.



(a) Cumulative regret for fixed horizon $T = 800$, $N = 50$, $m(q) = 30, P_u = 0.1$, and fixed $\epsilon = 0.3$

(b) Cumulative regret for fixed horizon $T = 800$, $N = 50$, $m(q) = 30, P_u = 0.3$, and fixed $\epsilon = 0.3$

Figure 5. Cumulative Regret Comparison

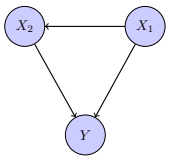
Now, we measure the cumulative regret for the same experiments with a fixed horizon T for each $P_u \in \{0.1, 0.3\}$. For each algorithm, we follow standard behavior for the first $\frac{T}{2}$ timesteps and fix the best estimated action for the second $\frac{T}{2}$ timesteps. We notice that our online algorithm's cumulative regret converges much faster than the two offline algorithms.

6.2. Confounded Causal Graph

X_1	$P(X_2=1\|X_1)$
0	P_u
1	1-P_u

$P(Y=1\|X_1,X_2)$	$X_1=0$	$X_1=1$
$X_2=0$	0.3	0.5
$X_2=1$	0.5	0.7

(a) Graph Conditional Probabilities

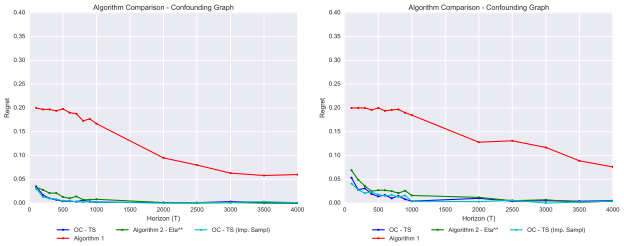


(b) Graph Depiction

Figure 6. Confounded Causal Graph Problem Formulation

In the confounded causal graph setting, Y depends only on two variables X_1 and X_2. Conditional reward probabilities, $P(Y = 1|X_1, X_2)$ are given in Figure 6(a) along with $P(X_2 = 1|X_1)$.

Our best possible action for both scenarios, $P_u = 0.1$ and $P_u = 0.3$, is the intervention $do(X_1 = 1)$. When $P_u = 0.3$, the difference in conditional rewards for the actions $do(X_1 = 1)$ and $do(X_2 = 1)$ is very small. We set $q_i = 0$ for $i \leq m, i \notin \{2,3\}$; $q_i = P_u$ for $i \in \{2,3\}$; and $\frac{1}{2}$ otherwise. Note that we compare two versions of OC-TS: (1) where each action is played once for the first $|\mathcal{A}|$ timesteps and (2) where is each action is drawn from η^*, the same sampling distribution as Algorithm 2 from (Lattimore & Reid, 2016): $\eta^* = \text{argmin}_\eta\, m(\eta) = \text{argmin}_{a\in\mathcal{A}} \max_{a\in\mathcal{A}} \mathbb{E}[\frac{\mathbb{P}(Pa_Y(X)|a)}{\sum_{b\in\mathcal{A}} \eta_b\, \mathbb{P}(Pa_Y(X)|b)}]$.



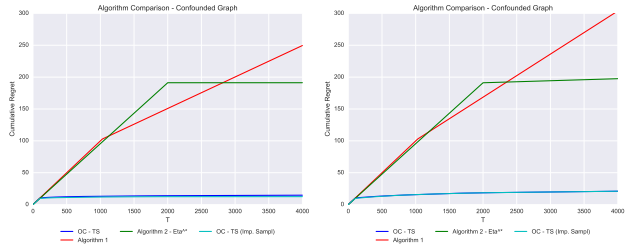
(a) Simple regret vs horizon T; $N = 50, m(q) = 30, P_u = 0.1$, and fixed $\epsilon = 0.3$

(b) Simple regret vs horizon T; number of variables $N = 50$, $m(q) = 30, P_u = 0.3$, and fixed $\epsilon = 0.3$

Figure 7. Simple Regret Comparison

We measure the simple regret for the confounded causal

graph as we vary the horizon T for each $P_u \in \{0.1, 0.3\}$. We notice that our online algorithm performs well for all horizons $T > |\mathcal{A}|$.



(a) Cumulative regret for fixed horizon $T = 4000$, $N = 50$, $m(q) = 30$, $P_u = 0.1$, and fixed $\epsilon = 0.3$

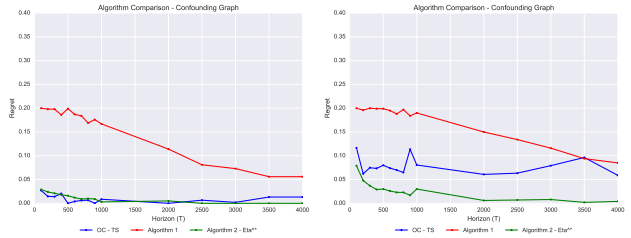
(b) Cumulative regret for fixed horizon $T = 4000$, $N = 50$, $m(q) = 30$, $P_u = 0.3$, and fixed $\epsilon = 0.3$

Figure 8. Cumulative Regret Comparison

Now, we measure the cumulative regret for the same experiments with a fixed horizon T for each scenario (large ϵ and small ϵ). For each algorithm, we follow standard behavior for the first $\frac{T}{2}$ timesteps and fix the best estimated action for the second $\frac{T}{2}$ timesteps.

7. Limitations of OC-TS

One limitation of our proposed algorithm is the need to seed initial exploration either through exhaustive search or minimizing the maximal importance sampling ratio as seen in Section 6. In particular, for the confounded Causal Graph as seen in Section 6.2 we have provided empirical results below where we do not 'warm-start' our algorithm. We observe that for our experimental parameters, $P_u = 0.3$ results in a gap $\mu_a^* - \max_{a \in \mathcal{A}, a \neq a^*} \mu_a = 0.04$ compared to $\mu_a^* - \max_{a \in \mathcal{A}, a \neq a^*} \mu_a = 0.08$ for $P_u = 0.1$. For $P_u = 0.1$ but not $P_u = 0.3$, we are able to play the optimal action with high probability without seeding our initial exploration.

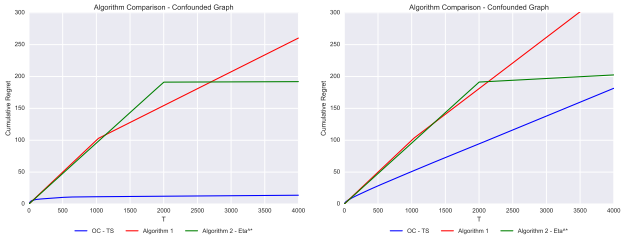


(a) Simple regret vs horizon T; $N = 50, m(q) = 30, P_u = 0.1$, and fixed $\epsilon = 0.3$

(b) Simple regret vs horizon T; $N = 50, m(q) = 30, P_u = 0.3$, and fixed $\epsilon = 0.3$

Figure 9. Simple Regret (No Exhaustive Initial Exploration)

In the figures above, we can see that without performing an exhaustive initial exploration, the performance of our algorithm is severely detrimented. This is still an area in which we are currently researching and is an area of future work. In particular, we are interested in understanding how important initial exploration is as the actions space and number of causal dependencies scale.



(a) Cumulative regret for fixed horizon $T = 800$, $N = 50$, $m(q) = 30$, $P_u = 0.1$, and fixed $\epsilon = 0.3$

(b) Cumulative regret for fixed horizon $T = 800$, $N = 50$, $m(q) = 30$, $P_u = 0.3$, and fixed $\epsilon = 0.3$

Figure 10. Cumulative Regret (No Exhaustive Initial Exploration)

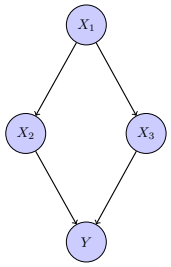
8. Experiments: Sparse Causal Graphs

Included are some experiments comparing online bandits algorithms in the setting of sparse causal graphs. For dense causal graphs we contrast the performance of of Online Causal Thompson Sampling with knowledge of $\{Pa_Y\}$ and Online Thompson Sampling without causal signaling.

X_1	$P(X_2 = 1 \mid X_1)$	$P(X_3 = 1 \mid X_1)$
0	0.1	0.1
1	0.9	0.9

$P(Y = 1 \mid X_2, X_3)$	$X_3 = 0$	$X_3 = 1$
$X_2 = 0$	0.5	0.1
$X_2 = 1$	0.1	0.6

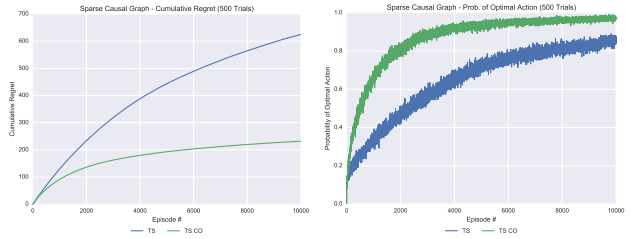
(a) Graph Conditional Probabilities



(b) Graph Depiction

Figure 11. Dense Causal Graph Problem Formulation

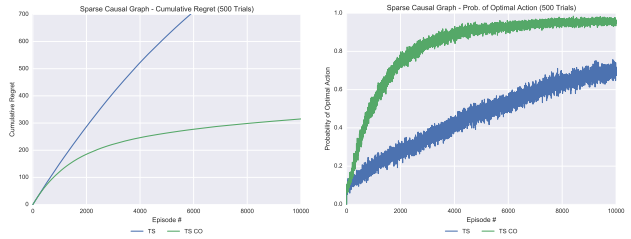
Our causal graph G has variables $\mathcal{X} = \{X_1, X_2, X_3, X_4, X_5\}$ and $|\mathcal{A}| = 10$ since each variable is a Bernoulli Random Variable and we are allowed to make an intervention of size 1. Note that $P(Y = 1 \mid X_2, X_3, X_4, X_5) = P(Y = 1 \mid X_2, X_3)$ and that $P(X_2 = 1 \mid X_1)$, $P(X_2 = 1 \mid X_1)$ are the same as in Figure 3.



(a) Cumulative Regret Comparison (b) Optimal Action Prob. Comparison

Figure 12. Sparse Causal Graph ($|\mathcal{A}| = 10$)

Now we consider a sparser causal graph G which has variables $\mathcal{X} = \{X_1, X_2, \ldots, X_7\}$ and $|\mathcal{A}| = 14$. Note that $P(Y = 1|X_2, X_3, \ldots, X_7) = P(Y = 1|X_2, X_3)$ and that $P(X_2 = 1|X_1)$, $P(X_2 = 1|X_1)$ are the same as in Figure 3.



(a) Cumulative Regret Comparison (b) Optimal Action Prob. Comparison

Figure 13. Sparse Causal Graph ($|\mathcal{A}| = 14$)

9. Conclusion

Through our experiments, we have observed the benefits of online learning in the Causal Bandit setting. By using an algorithm that adapts its arm sampling policy after observing rewards, we can perform directed exploration to evaluate potentially high value interventions and achieve smaller cumulative regret. In comparison to (Lattimore & Reid, 2016) which uses an offline importance sampling based estimator, we have shown empirically that an online model based estimator more efficiently learns the causal environment in a wide range of experiments. Moreover, we have designed an algorithm for concurrently learning the effects of causal dependencies and estimating conditional reward distributions. The drawbacks of our algorithms include sensitivity to initial exploration as well as computational complexity. Our algorithm's runtime is close to an order of magnitude larger for the experiments we conducted. One potential piece of future work would be to evaluate batch methods which can tradeoff between the regret performance of our online method and the computational tractability of offline methods.

10. Future Work

We would like to better understand the fundamental limitations of our algorithm and potential alternatives that may tradeoff performance with robustness and/or computational complexity. One limitation we presented is sensitivity to the initial exploration phase of our algorithm. When the gap between the optimal action and the next best action, $\mu_a^* - \max_{a \in \mathcal{A}, a \neq a^*} \mu_a$, is small we notice that our algorithm's performance is sensitive to initial exploration strategies. We believe this is a byproduct of the cardinality of the action space being large as well as the sparsity of the causal graph and are currently performing research to validate this claim. Another potential piece of future work is understanding the performance of batch methods for causal bandits and whether such methods provide any added robustness in estimation of causal dependencies or conditional reward distributions.

References

Agrawal, S. and Goyal, N. Analysis of thompson sampling for the multi-armed bandit problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, 2012.

Bareinboim, E., Forney A. and Pearl, J. Bandits with unobserved confounders: A causal approach. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS)*, 2015.

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

Lattimore, F., Lattimore T. and Reid, M. Causal bandits: Learning good interventions via causal inference. In *Proceedings of the 30th Conference on Neural Information Processing Systems (NIPS)*, 2016.