Online Learning for Causal Bandits

Vin Sachidananda (joint work w/ Prof. Emma Brunskill)

Introduction

Problem

The Multi-Armed Bandit (MAB) model is used to analyze the exploration/exploitation trade-off inherent in sequential decision problems and reinforcement learning. The motivating example for studying MABs comes from clinical trials where we have N treatments and would like to learn the efficacy of each treatment. Given that patients arrive sequentially, we decide whether to try new or poorly understood treatments (exploring) or stick 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.

Motivation

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. In this type of decision-making environment, there is a clear incentive to learn the causal dependencies between actions and use these learned dependencies to make high value interventions. The difficulty of this problem is that we now must learn both the structure of our causal graphical model as well as 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.

Causal Graphical Model

In our causal model, following the terminology of [Koller, 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_j, or the subset of $\mathcal{X}$ such that there exists an edge from X_i to X_j, to be Pa_{X_j}. 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 on node X_i, all edges between X_i and Pa_{X_i} are mutilated and the graph G can be represented by an altered probability distribution $\mathbb{P}(X^c | do(X = x))$ where $X^c = \mathcal{X} - X$.

Our agent 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 is Bernoulli. 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.

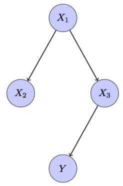
Results

Chain Causal Graph

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

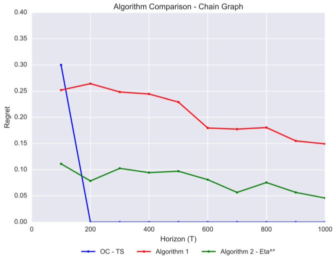
$P(Y = 1 \| X_3)$	
$X_3 = 0$	0.5
$X_3 = 1$	0.8

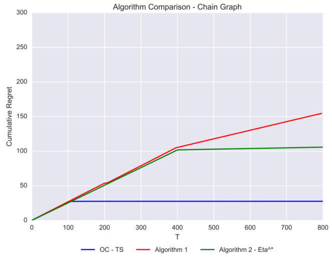
(a) Graph Conditional Probabilities



(b) Graph Depiction

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 $\frac{1}{2} + \epsilon$ 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. For our simulations shown below, $N = 50$, $m(q) = 30$, $P_u = 0.1$ and $\epsilon = 0.3$. Note that the graph on the left depicts simple regret vs. horizon T while the graph on the right depicts cumulative regret vs. horizon T.



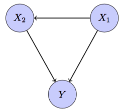


Confounded Causal Graph

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

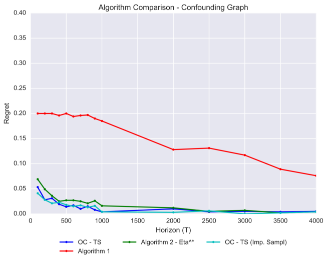
$P(Y = 1 \| X_1, X_2)$	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	0.3	0.5
$X_2 = 1$	0.5	0.7

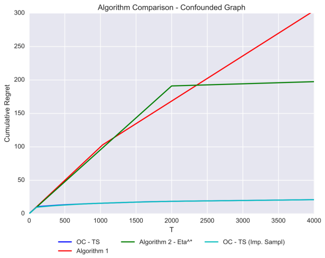
(a) Graph Conditional Probabilities



(b) Graph Depiction

In the confounded causal graph setting, Y depends only on two variables X_1 and X_2. Our best possible action is the intervention $do(X_1 = 1)$. 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. We compare two versions of OC-TS: (1) where each action is played once for first $|\mathcal{A}|$ timesteps and (2) where is each action is drawn from $\eta^* = \text{argmin}_{a \in \mathcal{A}} \max_{a \in \mathcal{A}} \mathbb{E}[\frac{\mathbb{P}(Pa_Y(X)|a)}{\sum_{b \in \mathcal{A}} \eta_b \mathbb{P}(Pa_Y(X)|b)}]$. For our simulations shown below, $N = 50$, $m(q) = 30$, $P_u = 0.3$ and $\epsilon = 0.3$. Note that the graph on the left depicts simple regret vs. horizon T while the graph on the right depicts cumulative regret vs. horizon T.





Algorithm

We propose an online algorithm for the setting of causal bandits with general causal graphs. Online Causal Thompson Sampling (OC-TS), uses the observation that $\mu_a = \mathbb{E}[Y | do(X = a)]$ can be decomposed by partitioning over $Pa_Y \in \mathcal{X}$. For $Pa_Y = \{X_1, X_2, ..., X_N\}$ where each element in the set is a Bernoulli Random Variable, our partition over the sample space is $\mathcal{Z} = \{Z_1, Z_2, \ldots, Z_{2^N}\}$.

$$\mu_a = \mathbb{E}[Y | do(X = a)] = \sum_{k=1}^{2^N} \mathbb{P}(Y = 1 | Pa_Y = Z_k) \, \mathbb{P}(Pa_Y = Z_k | do(X = a))$$

Using this observation, we can use Thompson Sampling to learn $\mathbb{P}(Y = 1 | Pa_Y = Z_k)$ and $\mathbb{P}(Pa_Y = Z_k | do(X = a))$. We provide 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 $do(X = a)$.

Algorithm 1 Online Causal Thompson Sampling - Known $\{Pa_Y\}$ Setting

```
Initialize: Beta$^0_{\mu_{Z_k}} = (1,1)$, Dirichlet$^0_{\rho_a} = 1$, $S^0_{Z_k}, F^0_{Z_k} = 0$
for timestep $t = 1, 2, \ldots, T$ do
    for action $a = 1, 2, \ldots, |\mathcal{A}|$ do
        $\mathbb{P}(Pa_Y = Z_k | do(X = a)) \sim Dirichlet^{t-1}_{\rho_a}[k]$
        $\mathbb{P}(Y = 1 | Pa_Y = Z_k) \sim Beta^{t-1}_{\mu_{Z_k}}[0]$
        $\mu_{\hat{a}_t} = \sum_{k=1}^{2^N} \mathbb{P}(Y = 1 | Pa_Y = Z_k) \mathbb{P}(Pa_Y = Z_k | do(X = a))$
    end for
    $\hat{a}_t^* = \text{argmax}_{\hat{a}_t} \mu_{\hat{a}_t}$
    $\{X^c\} \sim \mathbb{P}(X^c | do(X = \hat{a}_t^*))$
    $Z_k = \{X = \hat{a}_t^*\} \cup \{X^c\}$
    $Y_t \sim Bern(\mathbb{P}(Y = 1 | Pa_Y = Z_k))$
    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
    Beta$^t_{\mu_{Z_k}} = (S^t_{Z_k} + 1, F^t_{Z_k} + 1)$
end for
```

Discussion & Future Work

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, 2016] which uses an offline importance sampling based estimator, we have shown empirically that an online model based estimator more efficiently learns causal reward structures in a wide range of environments.

In terms of future work, we would like to better understand limitations of our algorithm and provide theoretical guarantees on regret. In particular, one limitation we notice is in 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 both the cardinality of the action space being large as well as the sparsity of the causal graph.