Exploring Online Learning Algorithms for Causal Bandits

Yash Shah 160050002

Gaurav Didwania 160050020

Rupesh 160050042

Kumar Saurav 160050057

Problem statement

A causal bandit instance can be described as follows. We are given a set of random variables $\mathcal{X} = \{X_1,\ X_2\ \ldots\ X_N\}$, a reward variable Y and a set of allowed *actions* $\mathcal{A}$. The dependencies between $\mathcal{X}$ and Y are represented using a causal graph $\mathcal{G}$, and the amount of information we have beforehand on $\mathcal{G}$ can be varied. Each action $a_t \in \mathcal{A}$ is an *intervention* of the form $do(\mathbf{X}_t = \mathbf{x}_t)$ (where $\mathbf{X}_t = \{X_{j_1},\ X_{j_2}\ \ldots X_{j_m}\},\ X_{j_i} \in \mathcal{X}$ $\forall i \in [m]$) that involves assigning X_{j_i} the corresponding value in $\mathbf{x}_t$ and removing all incoming edges for X_{j_i} in $\mathcal{G}$ to obtain a *mutilated* graph $\mathcal{G}_{a_t}$ (the empty intervention, $do()$, is also permitted). *After* the intervention has been performed, values of the non-intervened variables $\mathcal{X}_c = \mathcal{X} \setminus \mathbf{X}_t$ and the reward y_t are sampled from the distribution of $\mathcal{G}_{a_t}$. This process is repeated upto a fixed horizon T, and the objective is to choose $\{a_t\}_{t=1}^{T}$ such that the cumulative (or simple) regret is minimised.

For the purpose of this project, we restrict ourselves to those causal bandit instances where the *structure* of $\mathcal{G}$ is known beforehand but the joint probability distribution (i.e. $\mathbb{P}(\mathcal{X} \cup \{Y\})$ is unknown. Moreover, each $X \in \mathcal{X}$ and Y is a $\{0,1\}$-valued random variable, and the set of allowed actions $\mathcal{A}$ is the set of all possible size-1 interventions (i.e. $|\mathbf{X}_t| = |\mathbf{x}_t| = 1$, and hence $|\mathcal{A}| = 2N$). We shall minimise cumulative regret in our algorithms, which is defined as $R_T = \mu^* T - \mathbb{E}\left[\sum_{t=1}^{T} y_t\right]$.

Prior work and limitations

The causal bandit problem was first formulated in Lattimore et al. [2016], where they assumed the causal graph $\mathcal{G}$ to be completely known beforehand. They suggested algorithms to minimise simple regret, which is defined as $R = \mu^* - \max_{a \in \mathcal{A}} \mu_a$ where $\mu_a = \mathbb{E}[Y|a]$, in two different scenarios — (i) when $\mathcal{G}$ was parallel i.e. all non-reward variables were independent of each other, and (ii) when $\mathcal{G}$ was arbitrary. Lattimore et al. [2016] used the graph structure by looking at parents of Y and optimising a distribution over $\mathcal{A}$, with the distribution itself changing to minimise the time taken to reach the best action.

Sachidananda and Brunskill [2017] extend Thompson Sampling to the causal bandit setting of Lattimore et al. [2016] by breaking the procedure into two parts — given an intervention $a = do(\mathbf{X} = \mathbf{x})$, they first try to estimate a distribution over the possible assignments of parent variables of Y given a, and then they estimate the reward distribution given a particular parent configuration; both distributions are updated after each trial as new samples are observed.

For many real-life causal bandit instances, it makes sense to minimise cumulative regret for a fixed horizon which is not explored in detail in these papers. A comprehensive analysis and comparison of the proposed algorithms with classical multi-armed bandit algorithms is also lacking. Moreover, in most practical scenarios, the *structure* of the causal graph is known beforehand — we know how variables affect each other as well as the reward — it is the exact *probability distribution* which is unknown. We believe that these algorithms can be modified to exploit this additional information, which is one of the few things we plan to explore in this project.

Objectives

We plan to perform an in-depth analysis of the problem and the solutions proposed in Lattimore et al. [2016] and Sachidananda and Brunskill [2017] by starting with the algorithms discussed in class, extending them to the causal bandit setting, and comparing how these fare against those mentioned in the aforementioned papers. In all the algorithms which we plan to explore or devise, we will be aiming to minimise *cumulative regret* unlike in Sachidananda and Brunskill [2017] where *simple regret* was minimised.

- In general, a conventional multi-armed bandit instance is different from a causal bandit one in the sense that at each timestep t in the latter, along with the reward y_t we observe values of the set of non-intervened variables $\mathcal{X}_c$. Since each $X \in \mathcal{X}_c$ affects Y directly or indirectly, ignoring this additional information might lead to suboptimal performance. Nevertheless, we shall begin by treating each intervention of a causal bandit as an action of a classical multi-armed bandit and analyse how the algorithms discussed in class (UCB, KL-UCB, Thompson Sampling) perform.
- We then plan to propose a simple but novel algorithm (along the lines of ϵ_t-greedy) that uses the sampled values of non-intervened variables $\mathcal{X}_c$ to estimate the graph's joint probability distribution and exploits it to pick an optimal intervention for each timestep t.
- Later, we wish to compare how each of the algorithms mentioned above fares against the OC-TS algorithm proposed in Sachidananda and Brunskill [2017], by first reproducing their results and then adopting it for minimising cumulative regret.
- We then aim to extend the OC-TS algorithm for our problem setting (known graph structure, unknown distribution) by estimating the joint distribution using sampled values of $\mathcal{X}_c$ and then using the non-dirichlet form of OC-TS as mentioned in Sachidananda and Brunskill [2017].
- Finally, if time permits, we will try to come up with an entirely new (and possibly better) algorithm.

Implementation details

We will mainly use Python for implementation, with the code repository[1] of Lattimore et al. [2016] as a starting point. We plan to test the performance of the proposed algorithms on randomly created causal graphs of varying structure and distribution.

References

Finnian Lattimore, Tor Lattimore, and Mark D. Reid. Causal bandits: Learning good interventions via causal inference. In *NIPS*, 2016. URL `https://papers.nips.cc/paper/6195-causal-bandits-learning-good-interventions-via-causal-inference.pdf`.

Vin Sachidananda and Emma Brunskill. Online learning for causal bandits. 2017. URL `https://web.stanford.edu/class/cs234/past_projects/2017/2017_Sachidananda_Brunskill_Causal_Bandits_Paper.pdf`.

[1] `https://github.com/finnhacks42/causal_bandits`