Casual Bandits

Abstract

T: This can use some work

We study the problem of learning an optimal intervention. Our approach combines the theory of multi-armed bandits and causal inference. The main contribution is the new setting as well as algorithms that minimise both the simple and cumulative regret.

1. Introduction

A classical problem in economics and operations research is how to choose interventions that maximise utility. We study an idealised model of such interventions using a combination of the multi-armed bandit framework (Robbins, 1952) and the theory/language of causal inference (Pearl, 2000).

The idea is best illustrated with an example. Consider a farmer wishing to optimise the yield of her crop. In each season she can invest in a green house (to control the temperature), cover her plants with nets (to prevent loss to birds), use expensive fertilizers, or install a high-tech watering system. We will assume for simplicity that she does not have the resources to make more than one intervention. Of course, the watering system will only be useful if the rainfall is low and the nets will be useful if the bird population is large. These things are random events that change from season to season, which makes the impact of each intervention difficulty to measure. The goal of this paper is to design an optimal strategy for the farmer to learn which intervention will be most profitable.

A widely used framework for sequential decision making is the multi-armed bandit. In the classic multi-armed bandit setting there is a finite set of available actions, each associated with a distribution over rewards which is unknown but stationary. At each timestep the agent selects an action and receives a reward sampled i.i.d from the corresponding reward distribution. The performance of bandit algorithms is described by the regret: the difference in the expected reward obtained by the algorithm and the reward that could be obtained if the optimal action was selected at every timestep.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

An alternate approach to selecting actions is causal inference. Frameworks for causal inference provide a mechanism to specify assumptions that allow observational distributions over variables to be mapped to interventional ones. This allows an agent to predict the outcome of an action based on non-experimental data. This approach is common in social science, demography, and economics where explicit experimentation may be difficult. For example, predicting the effect of changes to childcare subsidies on workforce participation or school choice on student grades.

We take a first step towards unifying these approaches by considering a variant of the stochastic multi-armed bandit problem where we have prior knowledge of the causal structure governing the available actions.

A natural way to connect the causal framework with the bandit setting is to model the problem as a causal directed acyclic graph. Each possible assignment of variables to values is an action (bandit arm). The reward could be a general function of the action selected and the final state of the graph. However for simplicity, we will consider the reward to be the value of a single specified node minus the cost of the selected action. The number of actions grows exponentially with the number of variables in the graph, making it important to use algorithms that take account of the graph structure to reduce the search space.

Problems framed in this way take on characteristics of different bandit settings depending on the assumptions we make about what subset of actions can be taken, what variables are observable and whether they are observed before or after an action is selected. If feedback is received only on the reward node then the do-calculus can be applied to eliminate some actions immediately, before any experiments are performed and then a standard bandit algorithm can be run on the remaining actions.

If we receive feedback on additional nodes the problem can be more interesting. In addition to being able to eliminate some actions prior to sampling any data as in the previous case, taking one action may give us some information on actions that were not selected.

We consider a bandit problem where the actions and reward are represented by a specific causal graph that demonstrates this interesting structure. We develop an algorithm to leverage the information provided by this structure and

demonstrate it substantially outperforms standard bandit algorithms applied to the same problem where the number of actions is large.

There has been substantial recent work into extending bandit algorithms to incorporate additional assumptions and deal with more complex feedback structures. Algorithms with strong guarantees have been developed for linear bandits [], generalized linear bandits, gaussian process bandits [], etc. There is also an active line of research into bandits with feedback defined by a graph. Actions are modelled as nodes in the graph and the agent observes rewards for each action connected to the selected action []. The novelty of our work is that we assume prior knowledge of the causal structure but not the functional form of the relationship between variables.

Partial monitoring is a very general framework for for decoupling the feedback from the action and reward. It can be used to classify problems into one of four categories, trivial with no regret, easy with $R_T = \tilde{\Theta}\left(\sqrt{T}\right)$, hard with $R_T = \Theta\left(T^{2/3}\right)$ and hopeless with $R_T = \Omega\left(T\right)$ (Bartók et al., 2014). Partial monitoring algorithms yield results that are optimal with respect the horizon T but not other parameters, such as K, which is the key focus of incorporating causal structure.

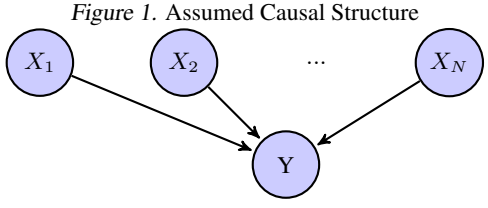
ALSO NEED TO MENTION ANY OTHER COMBINATIONS OF BANDITS+CAUSAL (eg the Elias NIPS paper and Generalized Thompson Sampling paper)

Key to Elias' paper is: observing the action an agent would take if it were allowed to make its natural choice can provide some information about hidden confounders that influence both the reward and the choice of action. Therefore, incorporating an agents natural choice as context may outperform a standard bandit that does not use that context. (Note: even in the presence of hidden confounders, including the agents natural choice as context only may improve the results. It is easy to come up with a counter example in which it does not).

2. Problem Setup

Assume we have a known causal model with binary variables $\boldsymbol{X} = \{X_1, \ldots, X_N\}$ that independently cause a target variable of interest $Y \in \mathbb{R}$ (see Figure 1).

Figure 1. Assumed Causal Structure



The game proceeds over T identical rounds (or time-steps). In each round t the learner can choose either to do nothing or they can choose a variable $I_t \in \{1, \ldots, N\}$ and an intervention $J_t \in \{0, 1\}$. After the learner has chosen an intervention they observe $X_{t,i} \in \{0, 1\}$ for all i where

$$X_{t,i} \sim \begin{cases} \text{Dirac}(J_t) & \text{if } I_t = i \\ \text{Bernoulli}(q_i) & \text{otherwise}\,. \end{cases}$$

where $\boldsymbol{q} \in [0, 1]^N$ is a (possibly unknown) vector of probabilities with $q_i = \mathrm{P}\{X_i = 1\}$. Note that if the learner did not choose an intervention then I_t is undefined and $X_{t,i} \sim \text{Bernoulli}(q_i)$ for all variables $i \in \{1, \ldots, N\}$. Finally the learner observes the reward $Y_t = r(X_t) + \eta_t$ where

$$r : \{0, 1\}^N \to \mathbb{R}$$

is arbitrary (and unknown) and η_t is a 1-subgaussian noise term (with the distribution possibly dependent on X_t).

The expected reward for intervening on variable i by setting it to j is defined by

$$\begin{aligned} \mu_{i,j} &= \mathbb{E}\left[r(X) | do(X_i = j)\right] \\ &= \sum_{\boldsymbol{x} \in \{0,1\}^N : x_i = j} r(\boldsymbol{x}) \prod_{k \neq i} q_k^{x_k} (1 - q_k)^{1 - x_k}\,. \end{aligned}$$

The optimal intervention is $(i^*, j^*) = \arg\max_{i,j} \mu_{i,j}$ and the corresponding optimal reward is $\mu^* = \mu_{i^*,j^*}$. Note that the expected reward of the optimal intervention is at least as large as the expected reward for doing nothing.

It is worth mentioning that the problem may be treated as a multi-armed bandit with $2N$ arms, one corresponding to each intervention. As we shall shortly see, this approach is usually not practical because the resulting algorithms do not exploit the structure in the covariates.

We consider two standard performance measures. The first is the cumulative regret, which measures the difference between the reward expected under the omnipotent strategy that knows the optimal action in advance and the expected reward of the learner.

$$R_T = T\mu^* - \mathbb{E}\left[\sum_{t=1}^{T} Y_t\right]\,. \tag{1}$$

The second performance measure is the simple regret, which measures the performance of the learner on the final round only.

$$r_T = \mu^* - \mathbb{E}\left[\mu_{I_T, J_T}\right]\,. \tag{2}$$

Both versions of the regret are useful in different circumstances. The first is most useful when the learner is truly

online and can change their policy over time. The second is useful when the exploration budget is limited and a fixed policy should eventually be chosen.

T: this remark feels out of place, but where to put it?

Remark 1. In order to be consistent with the literature on causal inference we use the notation $do()$ to denote the action of doing nothing and $do(X_{t,I_t} = J_t)$ the action of intervening on the I_tth variable and setting it to equal J_t. In the bandit community it is implicit that algorithms selecting actions are intervening in the system. So it is sufficient to index actions according to the variable and value. However, in causal graphs, it is essential to differentiate observing (or conditioning) on a variable taking a certain value, from intervening to set that variable. Although in the specific causal graph we consider, observation and intervention are the same, we deliberately introduce the do-notation (Pearl, 2000) that makes this distinction clear so as to help provide a bridge between the bandit and causal inference communities.

3. Upper Bounds on the Simple Regret

In this section we develop and analyse an algorithm for minimising the simple regret in the causal bandit problem.

From a causal perspective the key insight is that the assumed causal structure implies that the law of Y_t given we intervene to set $X_{t,i} = j$ is the same as the law of Y_t conditioned on $X_{t,i} = j$. Formally,

$$\mathrm{P}\left\{Y_t \leq y | do(X_{t,i} = j)\right\} = \mathrm{P}\left\{Y_t \leq y | do(), X_{t,i} = j\right\} . \tag{3}$$

This follows from application of the do-calculus (Pearl, 2000) to our specific causal graph. It is not the case in general. For example if there was a variable X' that caused both $X_{t,i}$ and Y (or $X_{t,i}$ and any other variable $X_{t,l}$), that would introduce a backdoor path from $X_{t,i} \to Y_t$ and we would have to condition on X' to derive the interventional distribution of Y_t from the observational one.

T: Or just give a reference or hide in the appendix depending on space in the end

Probably should define causal model and back-door rule to properly show this.

We can also learn about the reward for intervening on one variable from rounds in which we actually set a different variable.

$$P(Y_t | do(X_{t,i} = j)) = \sum_{j'} P(Y_t | do(X_{t,l} = j'), X_{t,i} = j) P(X_{t,l} = j') \tag{4}$$

The main idea of the algorithm is that by making no intervention (doing nothing) it is possible to estimate simultaneously the returns of all interventions. Unfortunately interventions that occur with low probability ($\mathrm{P}\left\{X_{t,i} = j\right\}$ is small) suffer from a high approximation error and should be explored separately by actually making the intervention. The algorithm works by optimally balancing the collection of observational data (when no action is taken) and making interventions to learn about low probability events. The problem is more challenging because $\boldsymbol{q}$ is unknown and must also be estimated.

T: make this nicer

Algorithm 1 Causal Best Arm Identification

Input: T, N
for $t \in 1, \ldots, (T-1)/2$ **do**
  Choose the action $do()$ and observe $\boldsymbol{X}_t$ and r_t
end for
Estimate μ using observation data:

$$(\forall i, j) \qquad \hat{\mu}_{i,j} = \frac{\sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = j\} r_t}{\sum_{t=1}^{T/2} \mathbb{1}\{X_{t,i} = j\}} .$$

Compute $\hat{q}_i = \frac{2}{T} \sum_{t=1}^{T/2} X_{t,i}$
Compute $\hat{s}_i = \min\left\{\hat{q}_i, 1 - \hat{q}_i\right\}$
Compute $\hat{s'} = sorted(\hat{s}) : \hat{s'}_1 \leq \hat{s'}_2 \leq \ldots \leq \hat{s'}_N$
Compute $\hat{m} = \min\left\{1 \leq i \leq N : \hat{s'}_{i+1} \geq \frac{1}{i}\right\}$
$i'(i) =$ the index of $\hat{s}_i$ in $\hat{s'}$
Compute A as the subset of infrequently observed arms $\left\{(i,j) : i'(i) \leq \hat{m}, j = \mathbb{1}\left\{\hat{q}_i \leq \frac{1}{2}\right\}\right\}$ with $|A| = \hat{m}$
for $(i,j) \in A$ **do**
  for $t \in 1, \ldots, \frac{T-1)}{2\hat{m}}$ **do**
    Choose action $do(X_{t,i} = j)$ and observe r_t
  end for
  Recompute $\hat{\mu}_{i,j} = \frac{2\hat{m}}{T} \sum_{t=1}^{T/2\hat{m}} r_t(X_{t,i} = j)$
end for
Estimated optimal action is $\hat{i}^*, \hat{j}^* = \arg\max_{i,j} \hat{\mu}_{i,j}$
Choose action $do(X_{t,\hat{i}^*} = \hat{j}^*)$

Theorem 2. *Define* $m = \min\left\{1 \leq i \leq N : q_{i+1} \geq \frac{1}{i}\right\}$. *Then Algorithm 1 satisfies*

$$r_T \in \mathcal{O}\left(\sqrt{\frac{m}{T} \log\left(\frac{NT}{m}\right)}\right) .$$

Note that algorithms designed for finite-armed bandits would explore interventions more uniformly and achieve a regret of $\Omega\left(\sqrt{N/T}\right)$. This is a significant improvement, since $m \leq N$ is usually substantially smaller than N.

T: Do we have a proof for this in another doc?

Proof. □

4. Upper Bound on the Cumulative Regret

We now tackle the case where the objective is to minimise the cumulative regret. This is significantly more challenging since the algorithm can no longer explore naively for $O(T)$ rounds without suffering linear regret.

We propose a simple explore-exploit based algorithm that leverages (3). Without loss of generality, we assume $q_i \in [0, \frac{1}{2}]$ and $q_1 \leq q_2 ... \leq q_N$.

Algorithm 2 Causal Explore-Exploit

Input: $T, \boldsymbol{q}$
Let $m = min\left\{1 \leq i \leq N : q_{i+1} \geq \frac{1}{i}\right\}$
Let $h = T^{2/3} m^{1/3} \log(TK)^{1/3}$
Let $A = \{(i,j) : i \leq m, j = 1\}$ be the set of infrequently observed arms
for $t = 1$ **to** $h/2$ **do**
  Choose the action $do()$ and observe $\boldsymbol{X}_t$ and r_t
end for
Compute for all arms $(i,j) \notin A$:

$$\hat{\mu}_{i,j} = \frac{2}{h} \frac{\sum_{t=1}^{h/2} \mathbb{1}\{X_{i,t} = j\} r_t}{q_i^j (1-q_i)^{1-j}}$$

for $(i,j) \in A$ **do**
  for $t' = 1$ **to** $h/2m$ **do**
    Choose the action $do(X_{i,t'} = j)$ and observe r_t
  end for
  Compute $\hat{\mu}_{i,j} = \frac{2m}{h} \sum_{t'=1}^{h/2m} \mathbb{1}\{X_{i,t'} = j\} r_{t'}$
end for
Compute $(\hat{i}^*, \hat{j}^*) = \arg\max_{(i,j)} \hat{\mu}_{i,j}$
for $t = h$ **to** T **do**
  Choose the action $do(X_{\hat{i}^*,t} = \hat{j}^*)$
end for

Theorem 3. *Define* $m = min\left\{1 \leq i \leq N : q_{i+1} \geq \frac{1}{i}\right\}$ *Then Algorithm 2 satisfies*

$$R(T) \in \mathcal{O}\left(T^{2/3} m^{1/3} \log(KT)^{1/3}\right).$$

The lower bound for the standard bandit problem is $R_t \in \Omega\left(\sqrt{TK}\right)$ (Auer et al., 1995). Comparing these results shows exploiting the extra information provided by the causal structure should outperform standard bandit algorithms when the number of arms is large, $K >> m^{2/3} T^{1/3}$. The parameter m summarizes the vector $\boldsymbol{q}$, and represents the number of actions that occur rarely naturally and thus must be explicitly explored. If $q_1, ..., q_N = 0$, the problem is completely unbalanced and $m = N$. If $q_1, ..., q_N = \frac{1}{2}$, the problem is completely balanced and $m = 1$.

Algorithm 2 tries to learn the rewards for all the arms during an exploration phase h and then picks the arm with the highest empirical mean for all remaining timesteps. During its exploration phase, it learns all the frequently occurring actions by observation and the remaining, infrequently occurring actions, by explicitly playing them. This leads to Chernoff type high probability bounds on the difference between the empirical and true rewards for all arms of the form $P(\hat{\mu}_{i,j} - \mu_{i,j} > D) \leq e^{-hD^2/m}$. By choosing optimal values for D and h we obtain the regret bound given in **??**. A full proof is given in the supplementary materials.

Algorithm 2 relies on $\boldsymbol{q}$ and thus m being known. We now consider the case where $\boldsymbol{q}$ is unknown. We begin by considering the simple regret, defined as the expected difference between the mean payoff of the optimal action and that of the action estimated to be optimal in Tth round.

Algorithm 3 Bandit Regret Algorithm

1: **Input:** T, N
2: $\delta = \frac{1}{T^{1/3}}$
3: $T_1 = 48 N log\left(4N/\delta\right)$
4: Run Algorithm 1 to line 11 with input T_1, N.
5: **if** $\hat{m} > \frac{N^{3/2}}{\sqrt{T}}$ **then**
6: Switch to the standard UCB algorithm.
7: **else**
8: $h = T^{2/3} \hat{m}^{1/3} log(TK)^{1/3}$
9: Run Algorithm 1 with input h, N.
10: Compute $(\hat{i}^*, \hat{j}^*) = \arg\max_{(i,j)} \hat{\mu}_{i,j}$
11: **for** $t = h$ **to** T **do**
12: Choose the action $do(X_{\hat{i}^*,t} = \hat{j}^*)$
13: **end for**
14: **end if**

Theorem 4. *Define* $m = min\left\{1 \leq i \leq N : q_{i+1} \geq \frac{1}{i}\right\}$ *Then Algorithm 3 satisfies*

$$R(T) \in \mathcal{O}\left(T^{2/3} m^{1/3} log(KT)^{1/3}\right).$$

5. Lower Bounds

6. Experiments

7. Discussion

In our algorithm, we have only used the side information provided by the $do()$ action about other actions. Since the $do()$ action fully reveals the value of alternate actions we could have incorporated this information via the graph feedback model (Mannor & Shamir, 2011), where at each timestep the feedback graph G_t is selected stochastically, dependent on $\boldsymbol{q}$, and revealed after an action has been chosen. The feedback graph is distinct from the causal graph. A link $A \to B$ in G_t indicates that selecting the action A reveals the reward for action B. For this specific problem, G_t will always be a star graph with the action $do()$ connected to half the remaining actions. The Exp3-IX algo-

Figure 2. Final regret versus number of variables N for UCB with $\alpha = 2$, Causal-Explore-Exploit with $m = 2$ and with $m = N$ and horizon $T = 10,000$. Error bars show standard deviation over 100 simulations. The regret for UCB grows linearly with the number of variables, whist for Causal-Explore-Exploit with fixed m, the growth is sub-logarthmic.

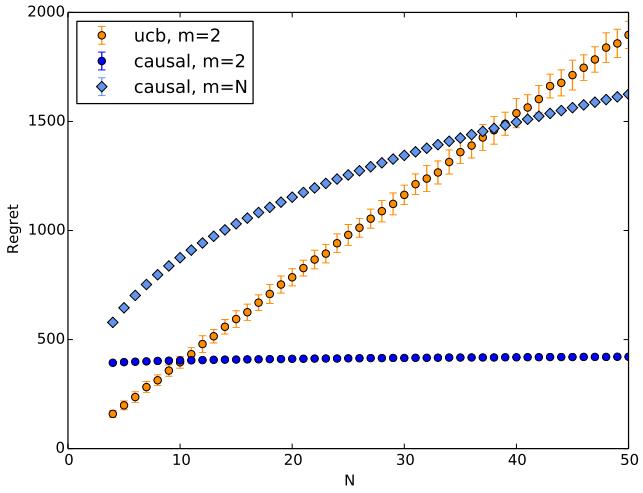
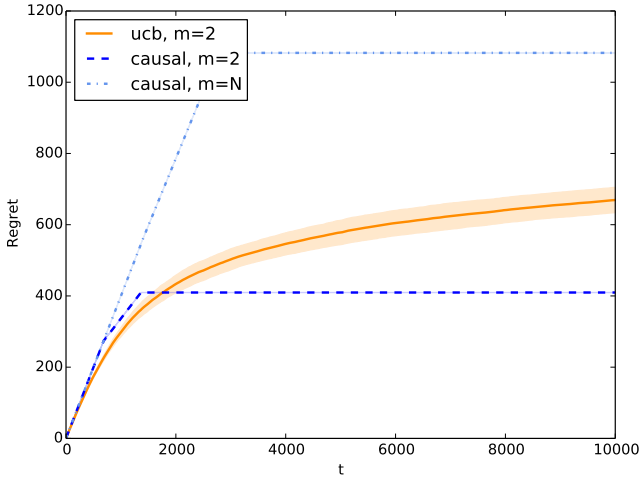


Figure 3. Cumulative regret over time for $N = 17$ for UCB with $\alpha = 2$, Causal-Explore-Exploit with $m = 2$ and Causal-Explore-Exploit with $m = N$. Shaded region shows standard deviation over 100 simulations. The Causal-Explore-Exploit algorithm incurs linear regret during the exploration phase, after which it selects the optimal arm with high probability. For $m = 2$, we have $K \sim m^{2/3}T^{1/3}$ and see that we are in the regime in which Causal-Explore-Exploit outperforms UCB.



rithm (Kocák et al., 2014) was developed for the adversarial version of this problem and has regret $\mathcal{O}\left(\sqrt{\bar{\alpha}T}\right)$, where $\bar{\alpha}$ is the average independence number of G_t. In our case $\bar{\alpha} = \frac{N}{2}$ so we again obtain the regret of the standard bandit algorithm. The issue here is that a malicious adversary can select the same graph each time, such that the rewards for half the arms are never revealed by the informative action. This is equivalent to a, nominally, stochastic selection of feedback graph where $\boldsymbol{q} = \mathbf{0}$

(Lelarge & Ens, 2012) consider a stochastic version of the graph feedback problem, but with a fixed graph available to the algorithm before it must select an action. In addition, their algorithm is not optimal for all graph structures and fails, in particular, to provide improvements for star like graphs as in our case. (Buccapatnam et al.) improve the dependence of the algorithm on the graph structure but still assume the graph is fixed and available to the algorithm before the action is selected.

More generally, assuming causal structure creates more complex types of side information, such as that shown in equation 4. In this case, selecting one action does not fully reveal an alternate action but provides some information towards an estimate. The quality of the estimate notably depends not only on the number of times that action was selected. For example, to get a good estimate for $X_1 = 1$ by intervening on X_2 requires us to sample both $X_2 = 0$ and $X_2 = 1$, in proportions dependent on q_2. This more complex side information does not fit within the graph feedback framework.

8. Future Open Questions

- Known but arbitrary structure
- Learning structure then exploiting

9. Conclusion

References

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R.E. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331, 1995. ISSN 0272-5428. doi: 10.1109/SFCS.1995.492488.

Bartók, Gábor, Foster, Dean P, Pál, Dávid, Rakhlin, Alexander, and Szepesvári, Csaba. Partial monitoring-classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4):967–997, 2014.

Buccapatnam, Swapna, Eryilmaz, Atilla, and Shroff Ness, B. Stochastic Bandits with Side Observations on Net-

works. *ACM SIGMETRICS'14, June 2014, Austin, Texas*. doi: 10.1145/2591971.2591989.

Kocák, Tomáš, Neu, Gergely, Valko, Michal, and Munos, Rémi. Efficient learning by implicit exploration in bandit problems with side observations. *Neural Information Processing Systems*, pp. 1–9, 2014.

Lelarge, Marc and Ens, Inria. Leveraging Side Observations in Stochastic Bandits. *Uai*, 2012.

Mannor, Shie and Shamir, Ohad. From Bandits to Experts: On the Value of Side-Observations. pp. 1–9, 2011. URL http://arxiv.org/abs/1106.2436.

Pearl, Judea. *Causality: models, reasoning and inference*. MIT Press, Cambridge, 2000.

Robbins, Herbert. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–536, 1952. ISSN 0002-9904. doi: 10.1090/S0002-9904-1952-09620-8.