Introduction

Problems requiring choosing an action under uncertainty are rife in all areas of human endeavour. For many problems, actions may be chosen sequentially, allowing the agent to learn from the outcome of early choices to improve later ones. 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 and independent of the reward distribution of other actions. At each timestep the agent selects an action and receives a reward sampled iid from the corresponding reward distribution.

An 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.

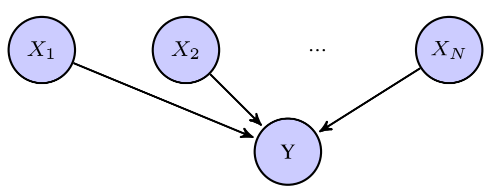
We take a first step towards unifying these approaches by considering a variant of the stochastic multi-armed bandit problem where we have prior knowledge of the causal structure governing the available actions. This structure creates dependencies between the rewards of different arms such that selecting one action can provide information on the reward for other actions.

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.

Problem Formulation

Assume we have a known causal model with binary variables $\boldsymbol{X} = \{X_1..X_N\}$ that independently cause a target variable of interest Y, figure 1. We can run sequential experiments on the system, where at each timestep t we can select a variable on which to intervene and subsequently observe the complete result, $(\boldsymbol{X}_t, Y_t)$. As an example, consider a farmer wishing to optimize the yield of her crop. She can invest in a green house to control temperature, a watering system to control soil moisture, fertilizers to set soil nutrients, etc. We assume only a single intervention is feasible due to cost and that each of these variables are independent of one-another (this may not always be the case - temperature could be related to rainfall for example). After having selected which variable to control, she plants her crops and observes the values of the remaining input variables and the yield. This repeats across many growing seasons, and the goal is to maximize the total cumulative yield.

Figure 1: Assumed Causal Structure



Let $q \in [0,1]^N$ be a fixed vector where $q_i = P(X_i = 1)$. In each time-step t upto a known end point T:

1. The learner chooses an $I_t \in \{1, \ldots, N\}$ and $J_t \in \{0, 1\}$.
2. Then $X_t \in \{0,1\}^N$ is sampled from a product of Bernoulli distributions, $X_{t,i} \sim \mathrm{Bernoulli}(q_i)$

3. The learner observes $\tilde{X}_t \in \{0,1\}^K$, which is defined by

$$\tilde{X}_{t,i} = \begin{cases} X_{t,i} & \text{if } i \neq I_t \\ J_t & \text{otherwise} \,. \end{cases}$$

4. The learner receives reward $Y_t \sim \text{Bernoulli}(r(\tilde{X}))$ where $r : \{0,1\}^K \to [0,1]$ is unknown and arbitrary. THIS LOOKS WRONG SHOULD BE N not K.

The expected reward of taking action i, j is $\mu_{i,j} = \mathbb{E}[r(X)|do(X_i = j)]$. The optimal reward and action are denoted μ^* and (i^*, j^*) respectively, where $(i^*, j^*) = \arg\max_{i,j} \mu_{i,j}$ and $\mu^* = \mu(i^*, j^*)$. The n-step cumulative expected regret is

$$R_n = \mathbb{E} \sum_{t=1}^{n} \left(\mu^* - \mu_{I_t, J_t}\right) .$$

The problem can be treated as a classical multi-armed bandit with $K = 2N$ arms. However, this does not utilize the information provided by the causal assumption.

Now need to expand upon how the assumption gives us extra information.

The difficulties faced due to bias.

To deal with this issue we use a simple explore-exploit algorithm. Our algorithm will explore for h time-steps, sampling actions in a way that depends on our prior knowledge of q but is independent of the observed rewards. We then select the arm with the highest expected reward for the remaining $T - h$ time steps.

Results

Summarize results here and note differences to classic bandit results

Known and Balanced q

We begin with the simplest case were we assume that $q_i = \frac{1}{2} \forall i$. During the exploration phase we sample actions uniformly at random. In this case, this is equivalent to purely observing, that is taking no action and allowing all input variables to take their value randomly as $X_{t,i} \sim \text{Bernoulli}(\frac{1}{2})$.

We will have $n_i \sim Binomial(h, \frac{1}{2})$ observations for each arm i at the end of the exploration stage. Note that this is independent of the number of arms K.

Assume we have K bernoulli arms with means ordered from highest to lowest $\mu_1 ... \mu_K$. Let $\Delta = [\Delta_1 ... \Delta_K]$ be the differences from the optimal reward μ_1.

Regret during explore phase

Since the probability we play each arm is constant and uniform during the exploration phase, the expected regret is simply proportional to the average sub-optimality Δ.

$$R_1 = h \sum_i P(i) \Delta_i = \frac{h}{K} \sum_i \Delta_i = hE[\Delta] < h \tag{1}$$

Regret during exploit phase

The regret during this phase is proportional to the expected sub-optimality of the arm with the highest empirical mean at the end of the explore phase.

$$\hat{i^*} = argmax_i[\hat{\mu}_i] \tag{2}$$

$$R_2 = (T-h)E[\Delta_{\hat{i^*}}] = (T-h)\sum_i P(\hat{\mu}_i \geq \hat{\mu}_j \forall j)\Delta_i \tag{3}$$

The difficulty with this approach is that it is hard to get bounds that are tight for all Δ. Instead, we will bound the probability that we select an arm with a sub-optimality gap greater than some D.

$$R_2 \leq (T-h)\left(P(\Delta_{\hat{i^*}} \leq D)D + P(\Delta_{\hat{i^*}} > D)\Delta_{max}\right) \tag{4}$$

The goal now is to get a bound for $P(\Delta_{\hat{i^*}} > D)$ in terms of Hoeffdings type bounds for each arm.

Suppose $i = \hat{i^*} \implies \hat{\mu}_i > \hat{\mu}_1$. If we haven't over-estimated μ_i too much, $\hat{\mu}_i - \mu_i < \frac{D}{2}$, and haven't under-estimated μ_1 too much, $\mu_1 - \hat{\mu_1} < \frac{D}{2}$, then $\Delta_{\hat{i^*}} = \mu_1 - \mu_i < D$

$$P(\Delta_{\hat{i^*}} > D) \leq P(\mu_1 - \hat{\mu_1} > \frac{D}{2}) + \sum_{i=2}^{K} P(\hat{\mu}_i - \mu_i > \frac{D}{2}) \tag{5}$$

If we used the empirical mean as an estimator for μ_i, the bound will depend on the number of times we actually observed each arm, which will be a random variable drawn from a multinomial distribution. Instead we will use an importance weighted estimator.

$$\hat{\mu}_i = \frac{1}{h}\sum_{t=1}^{h} \frac{Y_t \mathbb{1}\{\text{arm } i \text{ active}\}}{q_i} \tag{6}$$

where $q_i = P(\text{arm } i \text{ active})$

Hoeffdings gives $P(\hat{\mu}_i - \mu_i > \epsilon) \leq e^{-2h\epsilon^2 q_i^2}$. In this case we have assumed $q_i = \frac{1}{2}\forall i$. Putting this into equation 5:

$$P(\Delta_{\hat{i^*}} > D) \leq Ke^{-hD^2/8} \tag{7}$$

$$R_2 \leq (T-h)[(1-Ke^{-hD^2/8})D + Ke^{-hD^2/8}] < (T-h)[D + Ke^{-hD^2/8}] \tag{8}$$

Let $D = \sqrt{\frac{8}{h}\log(hK)}$

$$R_2 \leq (T-h)\left(\sqrt{\frac{8}{h}\log(hK)} + \frac{1}{h}\right) \tag{9}$$

Total Regret

Putting together the regret from the exploration and exploitation phases,

$$R_T \leq \frac{h}{K}\sum_i \Delta_i + (T-h)\left(\sqrt{\frac{8}{h}\log(hK)} + \frac{1}{h}\right) \tag{10}$$

$$\leq h + T\left(\sqrt{\frac{8}{h}\log(TK)} + \frac{1}{h}\right) \tag{11}$$

Now if we let $h = T^{2/3}(\log(KT))^{1/3}$,

$$R_T \leq 4T^{\frac{2}{3}}(log(KT))^{\frac{1}{3}} + T^{\frac{1}{3}}(log(KT))^{-\frac{1}{3}} \tag{12}$$

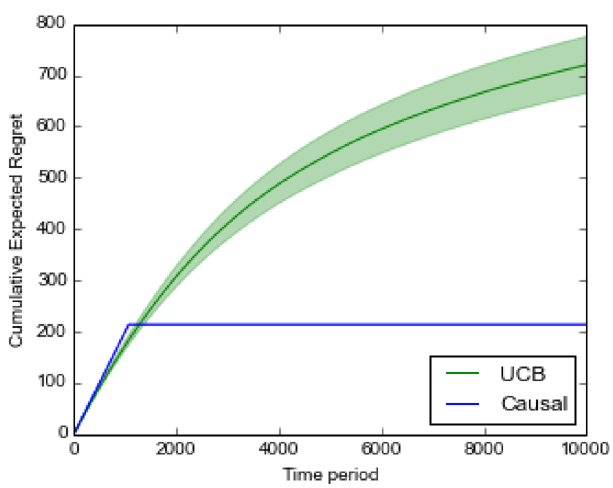
If $T \geq 2$ and $K \geq 2$, the first term dominates and,

$$R_T \leq 5T^{\frac{2}{3}}(log(KT))^{\frac{1}{3}} \tag{13}$$

The distribution independent lower bound for optimised UCB is $O(\sqrt{TK})$ (see Bubeck sect 2.4.3) so we would expect our algorithm to do better if $K >> T^{\frac{1}{3}}$

Empirical results

Figure 2: Comparison of the UCB and causal-explore-exploit for K=20 and T=10000. Note, $K \sim T^{1/3}$ Plot shows average and standard deviation over 10000 trials.



1 Generalizing to unbalanced q

When some arms have low natural probability we cannot rely on exploring them adequately by pure observation. We need to explicitly play them during the exploration phase.

We now have an additional trade off to make, which is how much should be observe (learning something about half the arms each timestep) versus playing the low probability arms.

Without loss of generality, we can assume $q_i \in [0, \frac{1}{2}]$ and $q_1 \leq q_2 ... \leq q_N$. Let $m \in [2, N] = \{m : q_m > \frac{1}{m}\}$ Ie if the problem is completely balanced $q_1 ... q_N = \frac{1}{2}$ then $m = 2$. If the problem is completely unbalanced, $q_1 ... q_N = 0$ then $m = N$. Let $q_{ij} = \mathrm{P}\left\{X_i = j\right\}$

Suppose we observe for the first $h/2$ timesteps. This is at worst half the optimal.

We then have estimates

$$\hat{\mu}_{ij} = \frac{\sum_{t=1}^{h/2} \mathbb{1}\{Y_t = 1, X_{t,i} = j\}}{\frac{h}{2} q_{ij}} \tag{14}$$

We take this as our estimate for those arms for which $q_{ij} > \frac{1}{m}$

Let $Z_{t,ij} = \mathbb{1}\{Y_t = 1, X_{t,i} = j\} \sim Bernoulli(q_{ij}\mu_{ij})$,

Chernoff's inequality gives

$$P(\hat{\mu}_i - \mu_i > \frac{D}{2}) \leq e^{-hD^2/24m} \tag{15}$$

We play each of the remaining m arms $\frac{h}{2m}$ times so for them we get

$$P(\hat{\mu}_i - \mu_i > \frac{D}{2}) \leq e^{-hD^2/4m} \tag{16}$$

So for all the arms

$$P(\hat{\mu}_i - \mu_i > \frac{D}{2}) \leq e^{-hD^2/24m} \tag{17}$$

and

$$P(\Delta_{\hat{i^*}} > D) \leq Ke^{-hD^2/24m} \tag{18}$$

If we let $D = \sqrt{\frac{24m\log(hK)}{h}}$

$$\begin{aligned} R_T &\leq h + T\left(\sqrt{\frac{24m\log(hK)}{h}} + \frac{1}{h}\right) && (19) \\ &\leq h + T\left(\sqrt{\frac{24m\log(TK)}{h}} + \frac{1}{h}\right) && (20) \end{aligned}$$

Let $h = T^{2/3}m^{1/3}\log(TK)^{1/3}$

$$R_T \leq 6T^{2/3}m^{1/3}log(KT)^{1/3} + T^{1/3}m^{-1/3}log(KT)^{-1/3} \tag{21}$$
$$\leq 7T^{2/3}m^{1/3}log(KT)^{1/3} \tag{22}$$

Again comparing to the UCB bound of $\sqrt{KT}$, we expect the causal algorithm to do better if $K > m^{2/3}T^{1/3}$, or alternatively if $m \leq \frac{K^{3/2}}{\sqrt{T}}$

To get results that degrade to similar order bounds as UCB when the arms are very unbalanced, I will need to drop the explore/exploit strategy.

1.1 Generalizing to unknown q

We now consider the case where $\boldsymbol{q}$, and thus m is not known in advance. We begin by considering the simple regret

1.1.1 Bounding the Simple Regret

- N is number of variables.
- q_i is probability that $X_{i,t} = 1$ for any t
- $s_i = \min\{q_i, 1 - q_i\}$
- $\mu : \{1, \ldots, N\} \times \{0, 1\} \to [0, 1]$.
- $I_t \in \{1, \ldots, N\} \times \{0, 1\}$ is the intervention

Let $R_s(h)$ be the simple regret after h time steps, which is

$$R_s(h) = \mathbb{E}\left[\mu^* - \mu_{I_h}\right] ,$$

where μ^* is the mean payoff of the optimal intervention and I_h is the intervention chosen in the hth round. We propose a two part algorithm that is provably order optimal when its performance is measured in terms of the simple regret.

Input: h, N

for $t \in 1, \ldots, h/2$ **do**

  Do nothing and observe $X_{i,t}$ for $i \in \{1, \ldots, N\}$ and r_t

end for

Compute for all $i \in \{1, \ldots, N\}$ and $j \in \{0, 1\}$:

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

Compute $\hat{q}_i = \frac{2}{h} \sum_{t=1}^{h/2} X_{i,t}$

Compute $\hat{s}_i = \min\{\hat{q}_i, 1 - \hat{q}_i\}$

Compute $\hat{s'} = sorted(\hat{s}) : \hat{s'}_1 \leq \hat{s'_2} \leq ... \leq \hat{s'}_N$

Compute $\hat{m} = \min\{2 \leq i \leq N : \hat{s}'_i \geq 1/i\}$

$i'(i) =$ the index of $\hat{s}_i$ in $\hat{s}'$

Compute A as the subset of infrequently observed arms $\{(i, j) : i'(i) \leq \hat{m}, j = \mathbb{1}\{\hat{q}_i \leq \frac{1}{2}\}\}$ with $|A| = \hat{m}$

for $(i, j) \in A$ **do**

  for $t \in 1, \ldots, h/2\hat{m}$ **do**

    Choose action $X_{i,t} = j$

    Observe reward $r_t(X_{i,t} = j)$

  end for

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

end for

Algorithm 1: Simple Regret Algorithm

Theorem 1. *Define* $m = \min\{2 \leq i \leq N : q_i \geq 1/i\}$. *Then the algorithm given in Algorithm 1 satisfies*

$$R_s(h) \in O\left(\sqrt{\frac{m}{h} \log\left(\frac{NT}{m}\right)}\right) .$$

Lemma 2. $\mathrm{P}\left\{|\hat{q}_i - q_i| \geq \sqrt{\frac{6q_i}{h} \log \frac{2}{\delta}}\right\} \leq \delta.$

Proof. Let $Z_t = \mathbb{1}\{X_{i,t} = 1\} \in \{0, 1\}$. Then

$$\hat{q}_i = \frac{2}{h} \sum_{t=1}^{h/2} Z_t .$$

Now $Z_1, \ldots, Z_{h/2}$ is an i.i.d. sequence of Bernoulli random variables with with mean q_i. The result follows from the Chernoff bound. ☐

Lemma 3. *Let* $\delta > 0$ *and* $\hat{s}_i = \min\{\hat{q}_i, 1 - \hat{q}_i\}$ *and define* $\hat{m} = \min\left\{2 \leq i \leq N : \hat{s}_{(i)} \geq 1/i\right\}$. *If*

$$h \geq 24m \log \frac{4N}{\delta} ,$$

then

$$\mathrm{P}\{\hat{m} \leq 2m\} \geq 1 - \delta .$$

Proof. It is easy to see that the worst case occurs when

$$q_i = \begin{cases} 0 & \text{if } i < m \\ \frac{1}{m} & \text{otherwise}. \end{cases}$$

Now by Lemma 2 we have with probability at least $1 - \delta$ that

$$\begin{aligned} (\forall i) \qquad |\hat{q}_i - q_i| &\leq \sqrt{\frac{6q_i}{h} \log \frac{2N}{\delta}} \\ &= \sqrt{\frac{6}{mh} \log \frac{2N}{\delta}}. \end{aligned}$$

By assumption

$$h \geq 24m \log \frac{4N}{\delta}.$$

Therefore with probability at least $1 - \delta$ we have

$$(\forall i) \qquad |\hat{q}_i - q_i| \leq \frac{1}{2m}.$$

Note that for i with $q_i = 0$ it is easy to see that $\hat{q}_i = 0$ (the variance is zero). The above guarantee that for i with $q_i = 1/m$ we have $\hat{q}_i \in [1/(2m), 3/(2m)]$ for all i with probability at least $1 - \delta$, which implies that $\hat{s}_i \geq 1/(2m)$ for all i where $q_i = 1/m$. Therefore $\hat{m} \leq 2m$ with probability at least $1 - \delta$. □

Lemma 4. *Let $\delta > 0$ and $\hat{s}_i = \min\{\hat{q}_i, 1 - \hat{q}_i\}$ and define $\hat{m} = \min\left\{2 \leq i \leq N : \hat{s}_{(i)} \geq 1/i\right\}$. If*

$$h \geq 24m \log \frac{4N}{\delta},$$

then

$$\mathrm{P}\left\{\hat{m} < \frac{2}{3}m\right\} \leq \delta.$$

Proof. To over-estimate m, we must identify as balanced some unbalanced arms (ie those with $i < m$. For all arms with $i < m$, by definition, $q_i < \frac{1}{m}$

Now by Lemma 2 we have with probability at least $1 - \delta$ that: (this step not quite right - because I haven't said $q_i = 1/m$ as the worst case is that q_i is infintesimally smaller than $1/m$. With smaller q_i the bounds would be slightly tighter but not made up for by the decrease in the mean - ie we are more likely to over-estimate if $q_i = 1/m$ than if $q_i < 1/m$. Need a better way to say this ...). Should probably just combine this with the other side into one lemma (even though we don't need this side for simple regret.

$$\begin{aligned} (\forall i) \qquad |\hat{q}_i - q_i| &\leq \sqrt{\frac{6q_i}{h} \log \frac{2N}{\delta}} \\ &= \sqrt{\frac{6}{mh} \log \frac{2N}{\delta}}. \end{aligned}$$

By assumption

$$h \geq 24m \log \frac{4N}{\delta}.$$

Therefore with probability at least $1-\delta$ we have

$$
\begin{aligned}
(\forall i < m) \qquad |\hat{q}_i - q_i| &\le \frac{1}{2m}\,. \\
\hat{q}_i &< \frac{3}{2}m \\
\implies \hat{m} &> \frac{2}{3}m
\end{aligned}
$$

□

Proof of Theorem 1. for $(i,j) \in A$, the algorithm explicitly selects the action, $X_i = j$, $\frac{h}{2\hat{m}}$ times.

$$\hat{\mu}_{i,j} = \frac{2\hat{m}}{h} \sum_{t=1}^{h/2\hat{m}} r_t(X_i = j)$$

Via Hoeffding's Inequality

$$\mathrm{P}\left\{|\hat{\mu}_{i,j} - \mu_{i,j}| > \epsilon\right\} \le 2\exp -\frac{h\epsilon^2}{\hat{m}}$$

for $(i,j) \notin A$, the algorithm has observed the reward given $X_i = j$ at least $\frac{h}{2\hat{m}}$ times.

$$
\begin{aligned}
(i,j) \notin A \implies &\ \hat{s}_i \ge \frac{1}{\hat{m}} \\
\implies &\sum_{t=1}^{h} \mathbb{1}\{X_i = j\} \ge \frac{h}{\hat{m}}
\end{aligned}
$$

Let $Z_{ij} = \sum_{t=1}^{h/2} \mathbb{1}\{X_i = j\}$ and $t'_1 ... t'_{Z_{ij}} = t : X_{i,t} = j$

$$\hat{\mu}_{i,j} = \frac{1}{Z_{ij}} \sum_{t'=1}^{Z_{ij}} r_{t'}$$

$$
\begin{aligned}
\mathrm{P}\left\{|\hat{\mu}_{i,j} - \mu_{ij}| > \epsilon\right\} =& \sum_{z=1}^{\infty} \mathrm{P}\left\{Z_{ij} = z\right\} \mathrm{P}\left\{\left|\frac{1}{Z_{ij}} \sum_{t'=1}^{Z_{ij}} r_{t'} - \mu_{ij}\right| > \epsilon | Z_{ij} = z\right\} \\
=& \sum_{z=1}^{\infty} \mathrm{P}\left\{Z_{ij} = z\right\} \mathrm{P}\left\{\left|\frac{1}{z} \sum_{t'=1}^{z} r_{t'} - \mu_{ij}\right| > \epsilon\right\} \\
\le& \mathrm{P}\left\{\left|\frac{2\hat{m}}{h} \sum_{t'=1}^{h/2\hat{m}} r_{t'} - \mu_{ij}\right| > \epsilon\right\} \sum_{z=1}^{\infty} \mathrm{P}\left\{Z_{ij} = z\right\} \\
\le& 2\exp -\frac{h\epsilon^2}{\hat{m}}
\end{aligned}
$$

Applying the union bound over all $2N$ actions.

$$\mathrm{P}\left\{\exists(i,j) : |\hat{\mu}_{i,j} - \mu_{i,j}| > \epsilon\right\} \leq 4N \exp -\frac{h\epsilon^2}{\hat{m}}$$
$$\implies \mathrm{P}\left\{\exists(i,j) : |\hat{\mu}_{i,j} - \mu_{i,j}| > \sqrt{\frac{\hat{m}}{h}\log\frac{4N}{\delta}}\right\} \leq \delta$$

Now by Lemma 3,

$$h \geq 24m\log\frac{4N}{\delta} \implies \mathrm{P}\left\{\hat{m} > 2m\right\} \leq \delta$$

Therefore if $h \geq 24m\log\frac{4N}{\delta}$, we have with probability at least $1-2\delta$ that

$$(\forall i,j) \qquad |\hat{\mu}_{i,j} - \mu_{i,j}| \leq \sqrt{\frac{2m}{h}\log\frac{4N}{\delta}} \text{ and } \hat{m} \leq 2m \tag{23}$$

Suppose $h < 24m\log\frac{4N}{\delta}$. Then

$$|\hat{\mu}_{i,j} - \mu_{i,j}| \leq 1 \leq \sqrt{\frac{2m}{h}\log\frac{4N}{\delta}}\,.$$

Therefore

$$|\hat{\mu}_{i,j} - \mu_{i,j}| \leq \sqrt{\frac{24m}{h}\log\frac{4N}{\delta}} \qquad \forall h \text{ with probability at least } 1-2\delta$$

Therefore

$$R_s(h) \leq 2\delta + \sqrt{\frac{24m}{h}\log\frac{4N}{\delta}}$$
$$\leq \frac{8m}{h} + \sqrt{\frac{24m}{h}\log\left(\frac{Nh}{m}\right)}$$

as required. □

1.1.2 Bounding the Bandit Regret

- Use $24Nlog(4N/\delta)$ samples to estimate m
- If $\hat{m} > \lambda = \frac{N^{3/2}}{\sqrt{T}}$ then stop and do UCB with remaining rounds.
- Else use causal explore-exploit and let the exploration time $h = T^{2/3}\hat{m}^{1/3}log(TK)^{1/3}$

There will be 3 cases.

1. $m < \lambda/2 \implies \hat{m} < \lambda$ with prob $1-\delta$

$$R_T \sim O(T^{2/3}m^{1/3}) + O(\sqrt{NT})\delta \implies \text{want } \delta \leq \frac{T^{1/6}}{\sqrt{N}}$$

2. $m > 3\lambda/2 \implies \hat{m} > \lambda$ with prob $1-\delta$

$$R_T \sim O(T^{2/3}N^{1/3})\delta + O(\sqrt{NT}) \implies \text{ want } \delta \leq \frac{N^{1/6}}{T^{1/6}}$$

3. $\lambda/2 < m < 3\lambda/2$, algorithm could end up doing UCB or Explore-Exploit.

$$\begin{aligned} R_T \sim & O(T^{2/3}m^{1/3}) + O(\sqrt{NT}) \\ = & O(\sqrt{NT}) \text{ as } m = O(\frac{N^{3/2}}{\sqrt{T}}) \end{aligned}$$

So if $\delta = \frac{1}{T^{1/6}\sqrt{N}}$ (or $\delta = \frac{1}{T^{1/3}}$) that would be enough concentration around m to choose the correct algorithm ... Is it enough to choose a reasonable total exploration time h? It seems that $\delta = \frac{1}{T^{1/3}}$ works.

Input: T, N
$\delta = \frac{1}{T^{1/3}}$
$T_1 = 24Nlog\,(4N/\delta)$
for $t \in 1, T_1$ **do**
    Do nothing and observe $X_{i,t}$ for $i \in \{1, \dots, N\}$ and r_t
end for
Compute $\hat{q}_i = \frac{2}{h} \sum_{t=1}^{h/2} X_{i,t}$
Compute $\hat{s}_i = \min\{\hat{q}_i, 1 - \hat{q}_i\}$
Compute $\hat{s'} = sorted(\hat{s}) : \hat{s'}_1 \le \hat{s'_2} \le ... \le \hat{s'}_N$
Compute $\hat{m} = \min\{2 \le i \le N : \hat{s}'_i \ge 1/i\}$
if $\hat{m} > \frac{N^{3/2}}{\sqrt{T}}$ **then**
    Switch to the standard UCB algorithm with $K = 2N$ arms.
else
    $h = T^{2/3}\hat{m}^{1/3}log(TK)^{1/3}$
    for $t \in T_1, h/2$ **do**
        Do nothing and observe $X_{i,t}$ for $i \in \{1, \dots, N\}$ and r_t
    end for
    Compute for all $i \in \{1, \dots, N\}$ and $j \in \{0, 1\}$:

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

    Recompute $\hat{q}_i = \frac{2}{h} \sum_{t=1}^{h/2} X_{i,t}$
    Recompute $\hat{s}_i = \min\{\hat{q}_i, 1 - \hat{q}_i\}$
    Recompute $\hat{s'} = sorted(\hat{s}) : \hat{s'}_1 \le \hat{s'_2} \le ... \le \hat{s'}_N$
    Recompute $\hat{m} = \min\{2 \le i \le N : \hat{s}'_i \ge 1/i\}$
    $i'(i) =$ the index of $\hat{s}_i$ in $\hat{s}'$
    Compute A as the subset of infrequently observed arms $\{(i,j) : i'(i) \le \hat{m}, j = \mathbb{1}\{\hat{q}_i \le \frac{1}{2}\}\}$ with $|A| = \hat{m}$
    for $(i, j) \in A$ **do**
        for $t \in 1, \dots, h/2\hat{m}$ **do**
            Choose action $X_{i,t} = j$
            Observe reward $r_t(X_{i,t} = j)$
        end for
        Recompute $\hat{\mu}_{i,j} = \frac{2\hat{m}}{h} \sum_{t=1}^{h/2\hat{m}} r_t(X_{i,t} = j)$
    end for
end if

Algorithm 2: Bandit Regret Algorithm

Since we don't know $\boldsymbol{q}$, and thus m, we can't set the exploration time h in advance based on m as we did for the known $\boldsymbol{q}$ case. Instead we first use $24Nlog(4N/\delta)$ samples to estimate m. If $\hat{m} < \frac{N^{3/2}}{\sqrt{T}}$ we will continue with the causal algorithm and let the total exploration time $h = T^{2/3}\hat{m}^{1/3}log(2TK)^{1/3}$.

Since $m \le N$, by Eq. (23), we have that with probability at least $1 - 2\delta$,

$$(\forall i, j) \qquad |\hat{\mu}_{i,j} - \mu_{i,j}| \le \sqrt{\frac{\hat{m}}{h} \log \frac{4N}{\delta}} \text{ and } \hat{m} \le 2m \tag{24}$$

So the regret in this case is bounded by,

$$
\begin{aligned}
R(T) &= 2\delta E\left[R(T)\right| \text{not } Eq.\ (24)] + (1-2\delta)E[R(T)|Eq.\ (24)] \\
&\leq 2\delta T + \left(h + T\sqrt{\frac{\hat{m}}{h}\log\frac{4N}{\delta}}\right) \\
&\leq 2T^{2/3} + h + T\left(\sqrt{\frac{\hat{m}}{h}\log\left(4NT^{1/3}\right)}\right),\ \text{letting } \delta = \frac{1}{T^{1/3}} \\
&\leq 2T^{2/3} + h + T\left(\sqrt{\frac{\hat{m}}{h}\log\left(2TK\right)}\right) \\
&\leq 2T^{2/3} + 2T^{2/3}(2m)^{1/3} log(2TK)^{1/3} \\
&\leq 6T^{2/3} m^{1/3} log(TK)^{1/3}
\end{aligned}
$$

1.1.3 Older Attempts

We now consider the case where $\boldsymbol{q}$ is not known in advance. Assume as before that we observe for $h/2$ timesteps. From the observations gained in this phase we estimate $\boldsymbol{q}$.

$$
\hat{q}_i = \frac{2}{h}\sum_{t=1}^{h/2} \mathbb{1}\{X_i = 1\} \tag{25}
$$

Let $\bar{q}_i = min(\hat{q}, 1-\hat{q})$ and construct $\bar{\boldsymbol{q}}$ such that $\bar{q}_1 \leq \bar{q}_2 \leq ... \leq \bar{q}_N$. We then estimate m with $\hat{m} = min_i : \bar{q}_i \geq \frac{1}{i}$

We estimate the reward for the apparently common arms, $i \geq \hat{m}$, as:

$$
\hat{\mu}_i = \frac{\sum_{t=1}^{h/2} \mathbb{1}\{Y=1, X_i=1\}}{\frac{h}{2}\hat{q}_i} = \frac{\sum_{t=1}^{h/2} \mathbb{1}\{Y=1, X_i=1\}}{\sum_{t=1}^{h/2} \mathbb{1}\{X_i=1\}} \tag{26}
$$

We play the remaining arms $\frac{h}{2\hat{m}}$ times and estimate their reward as:

$$
\hat{\mu}_i = \frac{2\hat{m}}{h}\sum_{t=1}^{h/2\hat{m}} \mathbb{1}\{Y=1|X_i=1\} \tag{27}
$$

We now consider how errors in the estimation of m effect our estimates of the arm rewards. For the arms with $i \geq \hat{m}$ we know $\hat{q}_i \geq \frac{1}{\hat{m}}$ and thus our estimate is based on at least $\frac{h}{2\hat{m}}$ observations. Similarly we explicitly play the infrequently observed arms, $i < m$, $\frac{h}{2\hat{m}}$ times. Thus our estimates will only be worse than the known $\boldsymbol{q}$ case if $\hat{m} > m$.

For a fixed m the $\boldsymbol{q}$ most likely to lead us to overestimate m is:

Justify this claim.

$$
q_i = \begin{cases} 0 & \text{if } i < m \\ \frac{1}{m} & \text{if } i \geq m \end{cases} \tag{28}
$$

We now bound $P(\hat{m} - m > \varphi)$ for this worst case.

For $i < m$,we have $\bar{q}_i = q_i = 0$. For $i \geq m$

Is this really true, justify

$$\begin{aligned}P(q_i - \bar{q}_i \geq C) &= P(q_i - \hat{q}_i \geq C | \hat{q}_i \leq \frac{1}{2}) P(\hat{q}_i \leq \frac{1}{2}) + P(q_i - (1 - \hat{q}_i) \geq C | \hat{q}_i > \frac{1}{2}) P(\hat{q}_i > \frac{1}{2}) \\ &\leq 2P(q_i - \hat{q}_i \geq C | \hat{q}_i \leq \frac{1}{2})\end{aligned}$$

Via Bernstein's Inequality:

$$2P(q_i - \hat{q}_i \geq C | \hat{q}_i \leq \frac{1}{2}) \leq 2\exp(-\frac{hC^2}{4q_i}) := \gamma \tag{29}$$

$$\implies P(q_i - \hat{q}_i \geq 2\sqrt{\frac{\log(2/\gamma)}{mh}}) \leq \gamma \tag{30}$$

Define $\hat{m}_i = \frac{1}{\hat{q}_i}$.

$$\hat{q}_i \geq q_i - C \implies \hat{m}_i \leq \frac{1}{q_i - C} = \frac{1}{\frac{1}{m} - C}, \text{ where } C < \frac{1}{m} \tag{31}$$

$$\implies q_i - \hat{q}_i \leq C \implies \hat{m}_i - m \leq \frac{m^2 C}{1 - mC} \tag{32}$$

$$\implies P(\hat{m}_i - m \geq \frac{m^2 C}{1 - mC}) \leq \gamma \tag{33}$$

Let $\varphi = \frac{m^2 C}{1 - mC} \implies C = \frac{\varphi}{m(\varphi + m)} \implies \gamma = 2\exp(-\frac{h\varphi^2}{4m(\varphi+m)^2}) \implies \varphi = \frac{2m\sqrt{m\log(2/\gamma)}}{\sqrt{h} - 2\sqrt{m\log(2/\gamma)}}$

Note this implies that if $h \geq 16m\log(2/\gamma)$ then $\varphi \leq m$

$$P(\hat{m}_i - m \geq \varphi) \leq 2\exp(-\frac{h\varphi^2}{4m(\varphi + m)^2}) \tag{34}$$

If $\hat{m}_i - m \geq \varphi$ for at most φ of the variables $i \geq m$, then $\hat{m} - m \leq \varphi$

Let $W_i = \mathbb{1}\{\hat{m}_i - m \geq \varphi\}$, $E[W_i] \leq \gamma$,$V[W_i] \leq \gamma$

$$P(\sum_{i=m}^{N} W_i \geq (N - m)\gamma + \varepsilon) \leq \exp(-\frac{\varepsilon^2}{2(N - m)\gamma + 2\varepsilon/3}) \tag{35}$$

Letting $(N - m)\gamma + \varepsilon = \varphi \implies \varepsilon = \varphi - (N - m)\gamma$

$$P(\hat{m} - m \geq \varphi) = P(\sum_{i=m}^{N} W_i \geq \varphi) \leq \exp(-\frac{(\varphi - (N - m)\gamma)^2}{2(N - m)\gamma + 2(\varphi - (N - m)\gamma)/3}) \tag{36}$$

$$\leq \exp(-\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3}), \text{ where } N\gamma < \varphi \tag{37}$$

If $P(\hat{m} - m \geq \varphi) \leq \zeta$.

$$
\begin{aligned}
P(\hat{\mu}_i - \mu_i \geq \epsilon) \leq & P(\hat{\mu}_i - \mu_i \geq \epsilon | \hat{m} - m \leq \varphi) P(\hat{m} - m \leq \varphi) && (38)\\
& + P(\hat{\mu}_i - \mu_i \geq \epsilon | \hat{m} - m \geq \varphi) P(\hat{m} - m \geq \varphi) && (39)\\
\leq & e^{-\frac{h\epsilon^2}{m+\varphi}} + \zeta e^{-\frac{h\epsilon^2}{N}} && (40)
\end{aligned}
$$

Putting it together

$$
\begin{aligned}
P(\hat{\mu}_i - \mu_i \geq \epsilon) &\leq \exp(-\frac{h\epsilon^2}{m+\varphi}) + \exp(-\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3}) \exp(-\frac{h\epsilon^2}{N}) && (41)\\
&\leq \exp(-\frac{h\epsilon^2}{m+\varphi}) + \exp(-\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3}) && (42)\\
&\leq \exp(-\frac{h\epsilon^2}{m+\varphi}) + \exp(-\frac{(\varphi - N\exp(-\frac{h\varphi^2}{4m(\varphi+m)^2}))^2}{2N\exp(-\frac{h\varphi^2}{4m(\varphi+m)^2}) + 2\varphi/3}) && (43)
\end{aligned}
$$

Assume $h > 16m\log(2N)$ and . Let $\varphi = \epsilon\sqrt{h}$. Assume $\epsilon \geq \frac{m}{\sqrt{h}}$ so as to ensure $\varphi > m$

$$
\begin{aligned}
P(\hat{\mu}_i - \mu_i \geq \epsilon) &\leq \exp(-\frac{h\epsilon^2}{m+\epsilon\sqrt{h}}) + \exp(-\frac{(\varphi - N\exp(-\frac{h}{16m}))^2}{2N\exp(-\frac{h}{16m}) + 2\varphi/3}) && (44)\\
&\leq \exp(-\frac{h\epsilon^2}{m+\epsilon\sqrt{h}}) + \exp(-\frac{(\varphi - \varphi/2)^2}{\varphi + 2\varphi/3}) && (45)\\
&= \exp(-\frac{h\epsilon^2}{m+\epsilon\sqrt{h}}) + \exp(-\frac{3\varphi}{20}) && (46)\\
&= \exp(-\frac{h\epsilon^2}{m+\epsilon\sqrt{h}}) + \exp(-\frac{3\epsilon\sqrt{h}}{20}) := \delta && (47)\\
&\leq \begin{cases} 2\exp(-\frac{h\epsilon^2}{m+\epsilon\sqrt{h}}) & \text{if } h < \frac{9m^2}{289\epsilon^2} \\ 2\exp(-\frac{3\epsilon\sqrt{h}}{20}) & \text{otherwise} \end{cases} && (48)
\end{aligned}
$$

$$
\begin{aligned}
\implies \epsilon \leq \begin{cases} \frac{1}{\sqrt{h}}\log(2/\delta) + \sqrt{\frac{m}{h}\log(2/\delta)} & \text{if } h < \frac{9m^2}{289\epsilon^2} \\ \frac{20}{3\sqrt{h}}\log(2/\delta) & \text{otherwise} \end{cases} && (49)\\
&& (50)
\end{aligned}
$$

$$
\implies \epsilon \leq \frac{20}{3\sqrt{h}}\log(2/\delta) + \sqrt{\frac{m}{h}\log(2/\delta)} \tag{51}
$$

Does this fit with the assumption we made about ϵ???

The first term will dominate if

$$\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3} \geq \frac{h\epsilon^2}{m+\varphi} \tag{52}$$

Setting them equal and solving for φ. Roughly,

$$\varphi = \frac{h\epsilon^2}{m+\varphi} \implies \varphi = \frac{1}{2}(\sqrt{4\epsilon^2 h + m^2} - m) \tag{53}$$

$$\implies \epsilon = \sqrt{\frac{\log(1/\delta)(\log(1/\delta)+m)}{h}} \leq \frac{\log(1/\delta)}{\sqrt{h}} + \sqrt{\frac{m}{h}\log(1/\delta)} \tag{54}$$

Now repeating the above but more exactly ...

$$\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3} \geq \frac{\varphi^2 - 2\varphi N\gamma}{4\varphi/3} \text{ , if } \varphi > 3N\gamma \tag{55}$$

$$= \frac{3}{4}(\varphi - 2N\gamma) \tag{56}$$

$$\frac{3}{4}(\varphi - 2N\gamma) = \frac{h\epsilon^2}{m+\varphi} \implies \varphi = \frac{\sqrt{3}}{6}\sqrt{16\epsilon^2 h + 3m^2 + 12N\gamma(m+N\gamma)} + N\gamma - \frac{m}{2} \tag{57}$$

$$\leq \frac{\sqrt{3}}{2}\sqrt{4\epsilon^2 h + 3mN\gamma} \tag{58}$$

If we let $\varphi = \frac{\sqrt{3}}{6}\sqrt{16\epsilon^2 h + 3m^2 + 12N\gamma(m+N\gamma)} + N\gamma - \frac{m}{2}$

$$P(\hat{\mu}_i - \mu_i \geq \epsilon) \leq 2\exp(-\frac{3}{4}(\varphi - 2N\gamma)) \tag{59}$$

$$\implies \epsilon = \sqrt{\frac{4\log^2(2/\delta) + (3m + 6N\gamma)\log(2/\delta)}{3h}} \tag{60}$$

$$\leq \frac{2\log(2/\delta)}{\sqrt{3h}} + \sqrt{\frac{m+2N\gamma}{h}\log(2/\delta)} \tag{61}$$

Ok - but I still need to ensure there exists a γ such that $N\gamma < \frac{\varphi}{3}$ and $P(\hat{m}_i - m > \varphi) \leq \gamma$

$$P(\hat{m}_i - m > a) \leq \exp(-\frac{ha^2}{4m(m+a)^2}) \tag{62}$$

It feels like I have just gone round and round in circles and pushed the problem to somewhere else ... my expression for φ now contains both ϵ and γ ...

$$\varphi \geq \frac{1}{2}\sqrt{m^2 + 4N\gamma(m+N\gamma)} + N\gamma - \frac{m}{2} \tag{63}$$

$$\Longrightarrow C = \frac{\varphi}{m(\varphi + m)} \geq \frac{\varphi}{\alpha m^2} \text{ if } \varphi < (\alpha - 1)m \tag{64}$$

$$\Longrightarrow P\left(\hat{m}_i - m \geq \alpha m^2 \left(\frac{4\log(1/\gamma)}{3h} + \sqrt{\frac{4\log(1/\gamma)}{mh}}\right)\right) \leq \gamma \tag{65}$$

If we let $\gamma = \frac{e^{-h/m}}{N}$

$$P\left(\hat{m}_i - m \geq \alpha m^2 \left(\frac{4}{3}(\frac{\log N}{h} + \frac{1}{m}) + 2\sqrt{\frac{\log N}{mh} + \frac{1}{m^2}}\right)\right) \leq \gamma \tag{66}$$

$$\Longrightarrow \ P\left(\hat{m}_i - m \geq \alpha m^2 \left(\frac{4}{3}(\frac{2}{m}) + 2\sqrt{\frac{2}{m^2}}\right)\right) \leq \gamma \text{ , if } h > m\log N \tag{67}$$

$$\Longrightarrow \ P\left(\hat{m}_i - m \geq 6\alpha m\right) \leq \gamma \tag{68}$$

This doesn't work. We need $\hat{m}_i - m < (\alpha - 1)m$ in order to get this expression.

Or we could let $\varphi = \frac{\sqrt{3}}{2}\sqrt{4\epsilon^2 h + 3mN\gamma}$

$$P(\hat{\mu}_i - \mu_i \geq \epsilon) \leq 2\exp(-\frac{h\epsilon^2}{m + \frac{\sqrt{3}}{2}\sqrt{4\epsilon^2 h + 3mN\gamma}}) \tag{69}$$

$$\Longrightarrow \ \epsilon = \frac{1}{\sqrt{2}}\sqrt{\frac{3\log^2(2/\delta) + 2m\log(2/\delta)}{h} + \frac{\sqrt{3}}{h^2}\sqrt{h^2\log^2(2/\delta)(3\log^2(2/\delta) + 4m\log(2/\delta) + 3mN\gamma)}} \tag{70}$$

$$\leq \tag{71}$$