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.

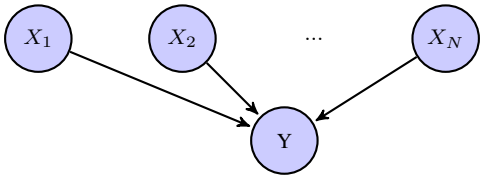
We take a first step towards unifying these approaches by considering a variant of the stochastic multi-armed bandit problem where we have prior knowledge of the causal structure governing the available actions. This structure creates dependencies between the rewards of different arms such that selecting 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, $X_{t,i} \sim \text{Bernoulli}(q_i)$

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.

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

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

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.

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

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

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 at least 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 $m = N$

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

We then have estimates

$$\hat{\mu}_{ij} = \frac{\sum_{t=1}^{h/2} \mathbb{1}\{Y = 1, X_i = 1\}}{\frac{h}{2}q_{ij}} \tag{14}$$

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

For these arms the Hoeffdings gives

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

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

$$P(\hat{\mu}_i - \mu_i > \frac{D}{2}) \leq e^{-hD^2/4m^2} \tag{17}$$

and

$$P(\Delta_{\hat{i^*}} > D) \leq Ke^{-hD^2/4m^2} \tag{18}$$

If we let $D = \sqrt{\frac{4m^2 \log(hK)}{h}}$

$$R_T \leq h + T\left(\sqrt{\frac{4m^2 \log(hK)}{h}} + \frac{1}{h}\right) \tag{19}$$

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

$$R_T \leq 4T^{\frac{2}{3}}m^{2/3}(log(KT))^{\frac{1}{3}} \tag{20}$$

Note, I think the $m^{2/3}$ should be improvable to something close to $m^{1/3}$ is I use Berstein's instead of Hoeffdings to bound the estimator for the arms where $q > 1/m$ but I haven't got the equations to quite work out for that yet.

It should also be possible to generalize to handle the case were the q's are unknown, since we should be able to get a reasonable estimate for them while we are observing. They key will be how the resulting uncertainty in m effects the bounds.

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

$$\hat{q}_i = \frac{2}{h}\sum_{t=1}^{h/2} \mathbb{1}\{X_i = 1\} \tag{21}$$

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:

$$\hat{\mu}_i = \frac{\sum_{t=1}^{h/2} \mathbb{1}\{Y = 1, X_i = 1\}}{\frac{h}{2}\hat{q}_i} = \frac{\sum_{t=1}^{h/2} \mathbb{1}\{Y = 1, X_i = 1\}}{\sum_{t=1}^{h/2} \mathbb{1}\{X_i = 1\}} \tag{22}$$

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

$$\hat{\mu}_i = \frac{2\hat{m}}{h}\sum_{t=1}^{h/2\hat{m}} \mathbb{1}\{Y = 1 | X_i = 1\} \tag{23}$$

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:

$$q_i = \begin{cases} 0 & \text{if } i < m \\ \frac{1}{m} & \text{if } i \geq m \end{cases} \tag{24}$$

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

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

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

Via Bernstein's Inequality:

$$2P(q_i - \hat{q}_i \geq C | \hat{q}_i \leq \frac{1}{2}) \leq 2\exp(-\frac{hC^2}{4q_i}) := \gamma \tag{25}$$

$$\implies P(q_i - \hat{q}_i \geq 2\sqrt{\frac{\log(2/\gamma)}{mh}}) \leq \gamma \tag{26}$$

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

$$\hat{q}_i \geq q_i - C \implies \hat{m}_i \leq \frac{1}{q_i - C} = \frac{1}{\frac{1}{m} - C}, \text{where } C < \frac{1}{m} \tag{27}$$

$$\implies q_i - \hat{q}_i \leq C \implies \hat{m}_i - m \leq \frac{m^2 C}{1 - mC} \tag{28}$$

$$\implies P(\hat{m}_i - m \geq \frac{m^2 C}{1 - mC}) \leq \gamma \tag{29}$$

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 $\varphi \leq m$

$$P(\hat{m}_i - m \geq \varphi) \leq 2\exp(-\frac{h\varphi^2}{4m(\varphi+m)^2}) \tag{30}$$

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

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

$$P(\sum_{i=m}^{N} W_i \geq (N-m)\gamma + \varepsilon) \leq \exp(-\frac{\varepsilon^2}{2(N-m)\gamma + 2\varepsilon/3}) \tag{31}$$

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

$$P(\hat{m} - m \geq \varphi) = P(\sum_{i=m}^{N} W_i \geq \varphi) \leq \exp(-\frac{(\varphi - (N-m)\gamma)^2}{2(N-m)\gamma + 2(\varphi - (N-m)\gamma)/3}) \tag{32}$$

$$\leq \exp(-\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3}), \text{where } N\gamma < \varphi \tag{33}$$

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

$$P(\hat{\mu}_i - \mu_i \geq \epsilon) \leq P(\hat{\mu}_i - \mu_i \geq \epsilon | \hat{m} - m \leq \varphi) P(\hat{m} - m \leq \varphi) \tag{34}$$

$$+ P(\hat{\mu}_i - \mu_i \geq \epsilon | \hat{m} - m \geq \varphi) P(\hat{m} - m \geq \varphi) \tag{35}$$

$$\leq e^{-\frac{h\epsilon^2}{m+\varphi}} + \zeta e^{-\frac{h\epsilon^2}{N}} \tag{36}$$

Putting it together

$$P(\hat{\mu}_i - \mu_i \geq \epsilon) \leq \exp(-\frac{h\epsilon^2}{m+\varphi}) + \exp(-\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3}) \exp(-\frac{h\epsilon^2}{N}) \tag{37}$$

$$\leq \exp(-\frac{h\epsilon^2}{m+\varphi}) + \exp(-\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3}) \tag{38}$$

$$\leq \exp(-\frac{h\epsilon^2}{m+\varphi}) + \exp(-\frac{(\varphi - N\exp(-\frac{h\varphi^2}{4m(\varphi+m)^2}))^2}{2N\exp(-\frac{h\varphi^2}{4m(\varphi+m)^2}) + 2\varphi/3}) \tag{39}$$

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

$$P(\hat{\mu}_i - \mu_i \geq \epsilon) \leq \exp(-\frac{h\epsilon^2}{m + \epsilon\sqrt{h}}) + \exp(-\frac{(\varphi - N\exp(-\frac{h}{16m}))^2}{2N\exp(-\frac{h}{16m}) + 2\varphi/3}) \tag{40}$$

$$\leq \exp(-\frac{h\epsilon^2}{m + \epsilon\sqrt{h}}) + \exp(-\frac{(\varphi - \varphi/2)^2}{\varphi + 2\varphi/3}) \tag{41}$$

$$= \exp(-\frac{h\epsilon^2}{m + \epsilon\sqrt{h}}) + \exp(-\frac{3\varphi}{20}) \tag{42}$$

$$= \exp(-\frac{h\epsilon^2}{m + \epsilon\sqrt{h}}) + \exp(-\frac{3\epsilon\sqrt{h}}{20}) := \delta \tag{43}$$

$$\leq \begin{cases} 2\exp(-\frac{h\epsilon^2}{m+\epsilon\sqrt{h}}) & \text{if } h < \frac{9m^2}{289\epsilon^2} \\ 2\exp(-\frac{3\epsilon\sqrt{h}}{20}) & \text{otherwise} \end{cases} \tag{44}$$

$$\implies \epsilon \leq \begin{cases} \frac{1}{\sqrt{h}}\log(2/\delta) + \sqrt{\frac{m}{h}\log(2/\delta)} & \text{if } h < \frac{9m^2}{289\epsilon^2} \\ \frac{20}{3\sqrt{h}}\log(2/\delta) & \text{otherwise} \end{cases} \tag{45}$$

$$\tag{46}$$

$$\implies \epsilon \leq \frac{20}{3\sqrt{h}}\log(2/\delta) + \sqrt{\frac{m}{h}\log(2/\delta)} \tag{47}$$

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

The first term will dominate if

$$\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3} \geq \frac{h\epsilon^2}{m + \varphi} \tag{48}$$

Setting them equal and solving for φ. Roughly,

$$\varphi = \frac{h\epsilon^2}{m+\varphi} \implies \varphi = \frac{1}{2}(\sqrt{4\epsilon^2 h + m^2} - m) \tag{49}$$

$$\implies \epsilon = \sqrt{\frac{\log(1/\delta)(\log(1/\delta) + m)}{h}} \leq \frac{\log(1/\delta)}{\sqrt{h}} + \sqrt{\frac{m}{h}\log(1/\delta)} \tag{50}$$

Now repeating the above but more exactly ...

$$\frac{(\varphi - N\gamma)^2}{2N\gamma + 2\varphi/3} \geq \frac{\varphi^2 - 2\varphi N\gamma}{4\varphi/3} \text{ , if } \varphi > 3N\gamma \tag{51}$$

$$= \frac{3}{4}(\varphi - 2N\gamma) \tag{52}$$

$$\frac{3}{4}(\varphi - 2N\gamma) = \frac{h\epsilon^2}{m+\varphi} \implies \varphi = \frac{\sqrt{3}}{6}\sqrt{16\epsilon^2 h + 3m^2 + 12N\gamma(m + N\gamma)} + N\gamma - \frac{m}{2} \tag{53}$$

$$\leq \frac{\sqrt{3}}{2}\sqrt{4\epsilon^2 h + 3mN\gamma} \tag{54}$$

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

$$P(\hat{\mu}_i - \mu_i \geq \epsilon) \leq 2\exp(-\frac{3}{4}(\varphi - 2N\gamma)) \tag{55}$$

$$\implies \epsilon = \sqrt{\frac{4\log^2(2/\delta) + (3m + 6N\gamma)\log(2/\delta)}{3h}} \tag{56}$$

$$\leq \frac{2\log(2/\delta)}{\sqrt{3h}} + \sqrt{\frac{m + 2N\gamma}{h}\log(2/\delta)} \tag{57}$$

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

$$P(\hat{m}_i - m > a) \leq \exp(-\frac{ha^2}{4m(m+a)^2}) \tag{58}$$

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 ...

$$\varphi \geq \frac{1}{2}\sqrt{m^2 + 4N\gamma(m + N\gamma)} + N\gamma - \frac{m}{2} \tag{59}$$

$$\implies C = \frac{\varphi}{m(\varphi + m)} \geq \frac{\varphi}{\alpha m^2} \text{ if } \varphi < (\alpha - 1)m \tag{60}$$

$$\implies P\left(\hat{m}_i - m \geq \alpha m^2 \left(\frac{4\log(1/\gamma)}{3h} + \sqrt{\frac{4\log(1/\gamma)}{mh}}\right)\right) \leq \gamma \tag{61}$$

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

$$P\left(\hat{m}_i - m \geq \alpha m^2 \left(\frac{4}{3}(\frac{\log N}{h} + \frac{1}{m}) + 2\sqrt{\frac{\log N}{mh} + \frac{1}{m^2}}\right)\right) \leq \gamma \tag{62}$$

$$\implies P\left(\hat{m}_i - m \geq \alpha m^2 \left(\frac{4}{3}(\frac{2}{m}) + 2\sqrt{\frac{2}{m^2}}\right)\right) \leq \gamma \text{ , if } h > m\log N \tag{63}$$

$$\implies P\left(\hat{m}_i - m \geq 6\alpha m\right) \leq \gamma \tag{64}$$

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

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

$$P(\hat{\mu}_i - \mu_i \geq \epsilon) \leq 2\exp(-\frac{h\epsilon^2}{m + \frac{\sqrt{3}}{2}\sqrt{4\epsilon^2 h + 3mN\gamma}}) \tag{65}$$

$$\implies \epsilon = \frac{1}{\sqrt{2}}\sqrt{\frac{3\log^2(2/\delta) + 2m\log(2/\delta)}{h} + \frac{\sqrt{3}}{h^2}\sqrt{h^2\log^2(2/\delta)(3\log^2(2/\delta) + 4m\log(2/\delta) + 3mN\gamma)}} \tag{66}$$

$$\leq \tag{67}$$

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.