Observe then pick best

- 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.
- The goal is to bound the psuedo-regret upto a total number of timesteps T
- Our algorithm will explore by playing uniformly at random for h timesteps and then select the arm with the highest estimated reward for the remaining timesteps $T - h$.
- We assume each arm corresponds to setting one binary input variable X_j to a given value. All the input variables are assumed to be independent causes of the binary reward variable Y. With this structure the probability of a reward is the same under the observation that a variable takes a given configuration as under the action that assigns it. Therefore, for each exploration timestep, we get data on the performance of half of the arms.
- We assume $P(X_j = 1) = \frac{1}{2} \forall j$. With this assumption 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. Relaxing this assumption will require us to have a more targeted exploration phase - as otherwise we do not gain any information about the value of arms that do not occur naturally with reasonable probability.

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:

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

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

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

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

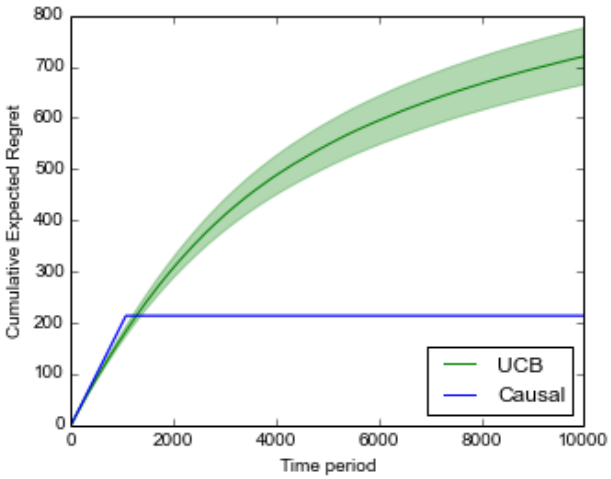
If $T \geq 2$ and $K \geq 2$, the first term dominates and,

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

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 1: 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

1.1 Option 1: Targeted sampling during exploration phase

The key fact we were utilizing to draw conclusions about multiple arms during each timestep of the explore phase is that, given our assumed causal structure, $P(Y|do(X_i = j)) = P(Y|X_i = j)$

If we do some form of targeted sampling, where we say opt to select each action $I = do(X_a = b)$ some specified number of times τ_{ab}, then we can no longer estimate $P(Y|X_i = j)$ simply by from the proportion of successes given $X_i = j$.

$$
\begin{aligned}
P(Y|do(X_i = j)) &= P(Y|X_i = j) && (14)\\
&= \sum_b P(Y|X_i = j, X_a = b)P(X_a = b|X_i = j) && (15)\\
&= \sum_b P(Y|X_i = j, X_a = b)P(X_a = b),\ \forall\, a \in \{1...K\}/i \text{ as } X_a \perp\!\!\!\perp X_i && (16)\\
&= \sum_b P(Y|X_i = j, do(X_a = b))P(X_a = b) && (17)
\end{aligned}
$$

Define $\mu^{ij} = E[Y|X_i = j] = P(Y|X_i = j)$

Let $\hat{\mu}_a^{ij}$ be an estimator for μ^{ij} based on samples where the intervention was on variable a.

$$
\hat{\mu}_a^{ij} = \begin{cases} \frac{1}{q_i(j)} \left(\frac{m_{a,1}^{ij}}{\tau_{a1}} q_a + \frac{m_{a,0}^{ij}}{\tau_{a0}} (1 - q_a) \right) & \text{if } a \neq i \\ \frac{m_{i,j}^{ij}}{\tau_{ij}} & \text{if } a = i \end{cases} \tag{18}
$$

$$
\tag{19}
$$

where

$$
\begin{aligned}
m_{a,b}^{ij} &= \sum_{s \in \{t: I_t = (a,b)\}} Z_{ab,s}^{ij} \text{ and,} && (20)\\
Z_{ab,s}^{ij} &= \mathbb{1}\{X_{i,s} = j, Y_s = 1\} \in \{0, 1\} && (21)
\end{aligned}
$$

TODO check/show this estimator is unbiased

For each arm, specified by the tuple i, j, we now have K estimators $[\hat{\mu}_1^{ij}...\hat{\mu}_K^{ij}]$ which we wish to combine to form a single estimator $\hat{\mu}^{ij}$. We will pool them as a weighted average with weights we can optimize based on the $q's$ so as to minimize the variance of the estimator.

$$
\hat{\mu}^{ij} = \sum_{a=1}^{K} w_a \hat{\mu}_a^{ij} \ \text{, where } \sum_{a=1}^{K} w_a = 1 \tag{22}
$$

Putting everything together,

$$\hat{\mu}^{ij} = \frac{w_i}{\tau_{ij}} \sum_{s \in \{t: I_t = (i,j)\}} Z^{ij}_{ij,s} + \sum_{a \neq i} \left(\frac{w_a}{q_i(j)} \left[\frac{q_a}{\tau_{a1}} \sum_{s \in \{t: I_t = (a,1)\}} Z^{ij}_{a1,s} + \frac{1 - q_a}{\tau_{a0}} \sum_{s \in \{t: I_t = (a,0)\}} Z^{ij}_{a0,s} \right] \right) \tag{23}$$

We now need to show $E[\hat{\mu}^{ij}] = \mu^{ij}$ and get a high probability bound for their difference. For the latter, lets try and use McDiarmids Inequality.

McDiarmid's Inequality states: If $Z_i \perp\!\!\!\perp Z_j$ and

$$|\phi(Z_1 ... Z_i ... Z_N) - \phi(Z_1 ... Z_i^{'} ... Z_N)| < c_i \ \ \forall i \tag{24}$$

$$P\left(|\phi(\boldsymbol{Z}) - E[\phi(\boldsymbol{Z})]| \geq \epsilon\right) \leq 2 \exp\left(-\frac{2\epsilon^2}{\sum_i c_i^2}\right) \tag{25}$$

$$P\left(|\phi(\boldsymbol{Z}) - E[\phi(\boldsymbol{Z})]| \geq \sqrt{\frac{\sum_i c_i^2}{2} \log \frac{2}{\delta}}\right) \leq \delta \tag{26}$$

For our problem, $\hat{\mu}^{ij} = \phi(\boldsymbol{Z}^{ij})$. All the $Z's$ are independent as they correspond indicator functions over the results at different timesteps.

$$|\phi(...Z^{ij}_{ij,s}...) - \phi(...Z^{'ij}_{ij,s}...)| \leq \frac{w_i}{\tau ij} \quad \leftarrow \tau_{ij} \text{ such } Z\text{'s} \tag{27}$$

$$|\phi(...Z^{ij}_{a1,s}...) - \phi(...Z^{'ij}_{a1,s}...)| \leq \frac{w_a q_a}{q_i(j)\tau_{a1}} \quad \leftarrow \tau_{a1} \text{ such } Z\text{'s for each } a \tag{28}$$

$$|\phi(...Z^{ij}_{a1,s}...) - \phi(...Z^{'ij}_{a1,s}...)| \leq \frac{w_a(1 - q_a)}{q_i(j)\tau_{a0}} \quad \leftarrow \tau_{a0} \text{ such } Z\text{'s for each } a \tag{29}$$

$$\sum_l c_l^2 = \frac{w_i^2}{\tau_{ij}} + \sum_{a \neq i} \frac{w_a^2}{q_i(j)^2} \left(\frac{q_a^2}{\tau_{a1}} + \frac{(1 - q_a)^2}{\tau_{a0}} \right) \tag{30}$$

$$= \sum_{a=1}^{K} w_a^2 f(a), \text{ where} \tag{31}$$

$$f(a) = \begin{cases} \frac{1}{q_i(j)^2} \left(\frac{q_a^2}{\tau_{a1}} + \frac{(1-q_a)^2}{\tau_{a0}} \right) & a \neq i \\ \frac{1}{\tau_{ij}} & a = i \end{cases} \tag{32}$$

We want to select weights to minimize equation 30 so as to achieve as tight a bound as possible in equation 25.

Applying Legrange Multipliers

$$w_a = \frac{1}{f(a) \sum_a \frac{1}{f(a)}} \tag{33}$$

$$\sum_i c_i^2 = \frac{1}{\sum_a \frac{1}{f(a)}} \tag{34}$$

Substitution 34 back into 25

$$P\left(|\hat{\mu}^{ij} - \mu^{ij}| \geq \epsilon\right) \leq 2\exp\left(-2\epsilon^2 \sum_{a=1}^{K} \eta_a^{ij}\right) \tag{35}$$

$$\eta_a^{ij} = \begin{cases} \frac{\tau_{a1}\tau_{a0}q_i(j)^2}{\tau_{a1}(1-q_a)^2 + \tau_{a0}q_a^2} & a \neq i \\ \tau_{ij} & a = i \end{cases} \tag{36}$$

$$\tag{37}$$

Substituting this into 5

$$P(\Delta_{\hat{i^*}} > D) \leq 2\sum_{(i,j)} exp\left(-\frac{D^2}{2}\sum_{a=1}^{K}\eta_a^{ij}\right) \tag{38}$$

$$R_2 \leq (T-h)\left(D + 2\sum_{(i,j)} exp\left(-\frac{D^2}{2}\sum_{a=1}^{K}\eta_a^{ij}\right)\right) \tag{39}$$

Looking at equation 38, let $A = \frac{D^2}{2}$ and

$$c_a = \frac{\tau_{a1}\tau_{a0}}{\tau_{a1}(1-q_a)^2 + \tau_{a0}q_a^2} \tag{40}$$

$$f(\boldsymbol{\tau}) := \frac{1}{2}P(\Delta_{\hat{i^*}} > D) = \sum_{i=1}^{N} e^{-A(\tau_{i1} + q_i^2 \sum_{a\neq i} c_a)} + \sum_{i=1}^{N} e^{-A(\tau_{i0} + (1-q_i)^2 \sum_{a\neq i} c_a)} \tag{41}$$

The goal is to find some assignment $\boldsymbol{\tau}(\boldsymbol{q})$ to minimize this subject to the constraint that $\sum_i \tau_{i1} + \sum_i \tau_{i0} = h$

No luck attempting to do this directly (via Lagrange multipliers ...)

Note, letting $\tau_a = \tau_{a1} + \tau_{a0}$ then c_a is maximized to equal τ_a if $\tau_{a1} = q_a\tau_a$

Consider the case where we have only 2 variables.

$$f(\boldsymbol{\tau}) = e^{-A(\tau_{11} + q_1^2 c_2)} + e^{-A(\tau_{21} + q_2^2 c_1)} + e^{-A(\tau_{10} + (1-q_1)^2 c_2)} + e^{-A(\tau_{20} + (1-q_2)^2 c_1)} \tag{42}$$

If we consider the extreme case where $q_1 = q_2 = 1$, ie the arms $X_1 = 0$ and $X_2 = 0$ never occur naturally.

$$f(\boldsymbol{\tau}) = 2e^{-A(\tau_{11} + \tau_{21})} + e^{-A(\tau_{10})} + e^{-A(\tau_{20})} \tag{43}$$

Which is minimized subject to the constraint $\tau_{11} + \tau_{10} + \tau_{21} + \tau_{20} = h$ if we let $\tau_{10} = \tau_{20} = \frac{Ah - log(2)}{3A}$. Note that as h gets large, we play the arms that do not occur naturally roughly $\frac{1}{3}$ of the time as opposed to $\frac{1}{4}$ if we were to

play uniformly. We only learn anything about the arms that do not naturally occur when we play them directly. For the other arms we learn something all the time (provided we have some level of balance in all the other arms...) so it makes sense to play the arms that don't occur naturally a little bit more to keep the overall bounds tight.

Ok, what happens if we just let $\tau_{ij} = \frac{h}{K}$

$$c_a = \frac{h}{K}\left(\frac{1}{(1-q_a)^2 + q_a^2}\right) \geq \frac{h}{K} \tag{44}$$

$$f \leq \sum_{i=1}^{N}\left(e^{-A(\frac{h}{K} + q_i^2\frac{(K-1)}{K}h)} + e^{-A(\frac{h}{K} + (1-q_i)^2\frac{(K-1)}{K}h)}\right) \tag{45}$$

$$\leq \sum_{i=1}^{N}\left(e^{-Ah} + e^{-A(\frac{h}{K})}\right) \tag{46}$$

Basically worst case, we learn about half the arms fast.

More generally, it would seem by targeted sampling we should be able to learn all the arms with reasonable natural probability fast - and the remaining arms at the rate h/K (or better if we play them more often).

Approaches to finding some better sampling function.

1. Find something motivated by the two variable example and show that it performs better than the uniform sampling approach.
2. Consider some other edge cases

Lets consider the case $q_1 = 1, q_2 ... q_N = \frac{1}{2}$ We will have $\tau_{11}, \tau_{10}, \tau_{ij} = \tau \; \forall i > 1, j$

$$c_a = \begin{cases} \tau_{11} & \text{if } a = 1 \\ 2\tau & \text{otherwise} \end{cases} \tag{47}$$

$$f(\boldsymbol{\tau}) = e^{-A(\tau_{11} + 2(N-1)\tau)} + e^{-A\tau_{10}} + 2(N-1)e^{-A(\tau + \frac{1}{4}(\tau_{11} + 2(N-2)\tau))} \tag{48}$$

$$= e^{-A(\tau_{11} + 2(N-1)\tau)} + e^{-A\tau_{10}} + 2(N-1)e^{-A(\frac{\tau_{11}}{4} + \frac{N\tau}{2})} \tag{49}$$

$$< e^{-A\tau_{10}} + (2(N-1)+1)e^{-A(\frac{\tau_{11}}{4} + \frac{N\tau}{2})} \tag{50}$$

$$= e^{-A\tau_{10}} + (2N-1)e^{-A(\frac{\tau_{11}}{4} + \frac{N\tau}{2})} \tag{51}$$

and the constraint is $\tau_{11} + \tau_{10} + 2(N-1)\tau = h$

If we maximize $(\frac{\tau_{11}}{4} + \frac{N\tau}{2})$ subject to $\tau_{11} + 2(N-1)\tau = (h - \tau_{10})$ we find $\tau = \frac{h-\tau_{10}}{2(N-1)}$ and $\tau_{11} = 0$. Note, as N gets large the sensitivity to the trade off between τ and τ_{11} becomes negligible.

$$\frac{\tau_{11}}{4} + \frac{N\tau}{2} = \begin{cases} \frac{(h-\tau_{10})N}{4(N-1)} & \text{if } \tau_{11} = 0 \\ \frac{(h-\tau_{10})N}{4(N-1)} - \frac{1}{4(N-1)} & \text{if } \tau_{11} = (h - \tau_{10}) \end{cases} \tag{52}$$

$$f(\boldsymbol{\tau}) < e^{-A\tau_{10}} + (2N-1)e^{-A(\frac{1}{4}(h-\tau_{10}))} \tag{53}$$

Which is minimized if

$$\tau_{10} = \frac{h}{5} - \frac{4}{5A} log(\frac{2N-1}{4}) \tag{54}$$

Ok - lets look at the reverse case, where $q_1 = \frac{1}{2}, q_2 ... q_N = 1$

$$\sum_a c_a = \begin{cases} (N-1)\tau_{11} & \text{if } a = 1 \\ 2\tau + (N-2)\tau_{11} & \text{otherwise} \end{cases} \tag{55}$$

The constraint is

$$2\tau + (N-1)\tau_{11} + (N-1)\tau_{10} = h \tag{56}$$

$$f(\boldsymbol{\tau}) = 2e^{-A(\tau + \frac{1}{4}(N-1)\tau_{11})} + (N-1)e^{-A(2\tau + (N-1)\tau_{11})} + (N-1)e^{-A\tau_{10}} \tag{57}$$

Holding τ_{10} constant, the 2nd and 3rd terms are constant. The first is maximized by letting $\tau_{11} = 0, \tau = \frac{h-(N-1)\tau_{10}}{2}$

$$f(\boldsymbol{\tau}) = 2e^{-A\tau} + (N-1)e^{-2A\tau} + (N-1)e^{-A\frac{h-2\tau}{N-1}} \leftarrow \text{ can't minimize analytically} \tag{58}$$

$$\leq (N+1)e^{-A\tau} + (N-1)e^{-A\frac{h-2\tau}{N-1}} \tag{59}$$

which is minimized if

$$\tau = \frac{1}{N+2}\left(h - N + \frac{N}{A}\log(\frac{N(N+1)}{2(N-1)})\right) \tag{60}$$

$$\sim \frac{1}{N+2}\left(h - N + \frac{N}{A}\log(\frac{N}{2})\right) \tag{61}$$

Equation 61 is also the solution that leads to the exponentiated terms in equation 59 being equal - which is should intuitively be optimal since the coefficients in front of them are very similar.

Yet another case study - if $q_1 ... q_N = 1$ So none of the arms $X_i = 0$ occur naturally.

$$c_a = \tau_1 \ \forall a \tag{62}$$

$$\sum_{i \neq a} c_a = (N-1)\tau_1 \ \forall a \tag{63}$$

$$N(\tau_0 + \tau_1) = h \tag{64}$$

$$f(\boldsymbol{\tau}) = Ne^{-AN\tau_1} + Ne^{-A\tau_0} \tag{65}$$

minimized if

$$\tau_0 = \frac{h}{N+1} - \frac{log(N)}{A(N+1)} \tag{66}$$

Note as N get large - this equates to $\tau_0 \sim \frac{h}{N}$ ie almost all the actions are allocated to the arms that do not occur naturally.

Another case study $q_1 ... q_{N_1} = 1, q_{N_1+1} ... q_N = \frac{1}{2}$ (should encompass all previous examples ...)

$$c_a = \begin{cases} \tau_{11} & \text{if } a \leq N_1 \\ 2\tau & \text{otherwise} \end{cases} \tag{67}$$

$$\sum_{i \neq a} c_a = \begin{cases} (N_1 - 1)\tau_{11} + 2N_2\tau & \text{if } a \leq N \\ N_1\tau_{11} + 2(N_2 - 1)\tau & \text{otherwise} \end{cases} \tag{68}$$

$$N_1\tau_{10} + N_1\tau_{11} + 2N_2\tau = h \tag{69}$$

$$f(\boldsymbol{\tau}) = N_1 e^{-A\tau_{10}} + N_1 e^{-A(N_1\tau_{11} + 2N_2\tau)} + 2N_2 e^{-A(\frac{N_1}{4}\tau_{11} + \frac{N_2+1}{2}\tau)} \tag{70}$$

The 2nd and 3rd terms are minimized by letting $\tau_{11} = 0$ (although the result is only weakly dependnent on the trade off between τ_{11} and τ provided $N_2 >> 2$).

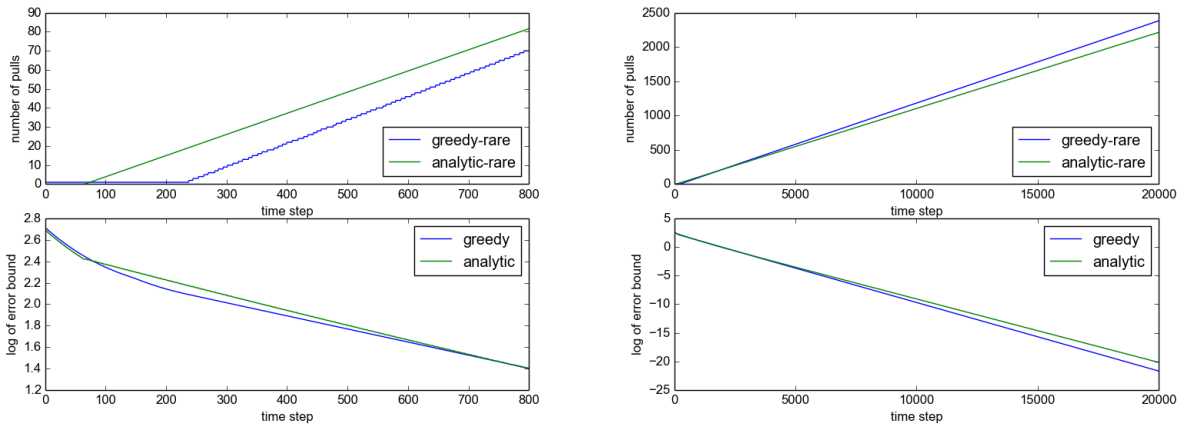
$$f(\boldsymbol{\tau}) = N_1 e^{-A\tau_{10}} + N_1 e^{-A2N_2\tau} + 2N_2 e^{-A\frac{N_2+1}{2}\tau} \tag{71}$$

$$\leq N_1 e^{-A\tau_{10}} + (2N_2 + N_1) e^{-A\frac{N_2+1}{2}\tau} \tag{72}$$

$$= N_1 e^{-A\tau_{10}} + (2N_2 + N_1) e^{-A(\frac{(N_2+1)(h - N_1\tau_{10})}{4N_2})} \tag{73}$$

$$\tau_{10} = \left(\frac{N_2 + 1}{(N_2 + 1)N_1 + 4N_2} \right) h - \frac{4N_2 \log(\frac{(N_2+1)(2N_2+N_2)}{4N_2})}{A((N_2 + 1)N_1 + 4N_2))} \tag{74}$$

It seems like this function can be optimized in a greedy way. That is, start with $\boldsymbol{\tau} = \mathbf{0}$, and in each timestep select the τ_{ij} that leads to the greatest reduction in the error bound. This appears to yield broadly consistent results to the analytic solution.



What about if I repeated one of the above analysis but let the unbalanced q = .75?

What can I say overall about worst/best cases/any symmetries in the problem??? Do these solutions lead to the terms in the exponentials being similar?

Goal is to choose something that will give us a simple bound and will be ok for the worst case. Its still not entirely clear to me what the worst case will be - after all it depends on the way you assign to τ.

What happens if we assume $\sum_{a \neq i} c_a$ is constant ... this is reasonable if N is large

When you look at each term in c_a, its maximized if we play the arms according to their natural probabilities. If the arm is balanced, this will also minimize the sum of the two terms for doing $x_i = 0$ and $x_i = 1$. If the arm is unbalanced, then we have to trade of balancing this sum versus contributing to the sum of c_a. Consider just how bad the sum of c_a can be even if we focus entirely on balancing the two terms ... maybe its already enough.

1.2 Option 2: Some measure of information gain

1.3 Option 3: Three stage algorithm

1. Observe randomly
2. Play low probability arms
3. Pick best and exploit