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:

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

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

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

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

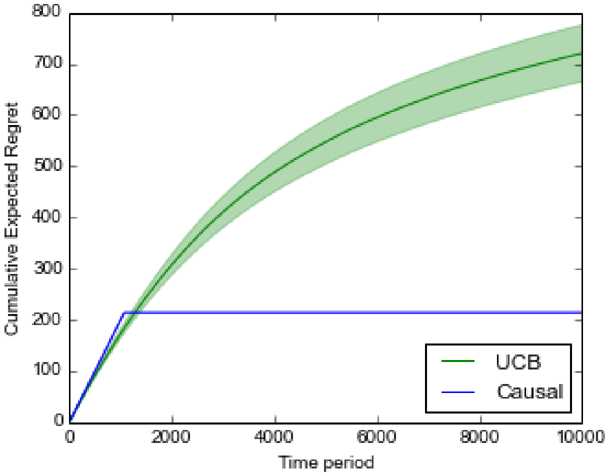
If $T \geq 2$ and $K \geq 2$, the first term dominates and,

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

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

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

$$(19)$$

where

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

For each arm, specified by the tuple i, j, we now have K estimators $[\hat{\mu}^{ij}_1 ... \hat{\mu}^{ij}_K]$ 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.

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

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 the second term

$$\sum_{(i,j)} exp\left(-\frac{D^2}{2}\left(\tau_{ij} + q_{ij}^2 \sum_{a \neq i} \frac{\tau_{a1}\tau_{a0}}{\tau_{a1}(1-q_a)^2+\tau_{a0}q_a^2}\right)\right) \tag{40}$$

Now let $\tau_a = \tau_{a1} + \tau_{a0}$ then $\frac{\tau_{a1}\tau_{a0}}{\tau_{a1}(1-q_a)^2+\tau_{a0}q_a^2}$ is maximized if $\tau_{a1} = q_a\tau_a$

Changing the ratio will only otherwise effect the terms that have where $a = i$. So compare those two terms ...

$$exp(-\frac{D^2}{2}(\tau_{\alpha 1} + q_\alpha g) + exp(-\frac{D^2}{2}(\tau_{\alpha 0} + (1-q_\alpha)g) \tag{41}$$

minimize this subject to the constraint, hmm seems to give and on the face of it different answer.

Consider splitting it up so that we sum first over the terms where $j = 1$ then over the terms where $j = 0$ (may be clearer).

1.2 Option 2: Three stage algorithm

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