# Observe then pick best

- Avi's point - of all the space of possible functions $f(X_1...X_N)$, most of them will have very similar marginals- so limited? opportunity for regret? Does this mean we effectively only care about a particular subset of the functions where the marginals are significantly different. We could assume we are in that subspace (if we are not then all actions are equally good anyway)
- How do assumptions/priors about the function $f(X_1...X_N)$ flow through to information about the marginals?
- 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 $\mu_1$.
- 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 $\Delta$.

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

## 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 $\Delta$. 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 $\mu_i$ too much, $\hat{\mu}_i - \mu_i < \frac{D}{2}$, and haven't under-estimated $\mu_1$ 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 $\mu_i$, 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

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

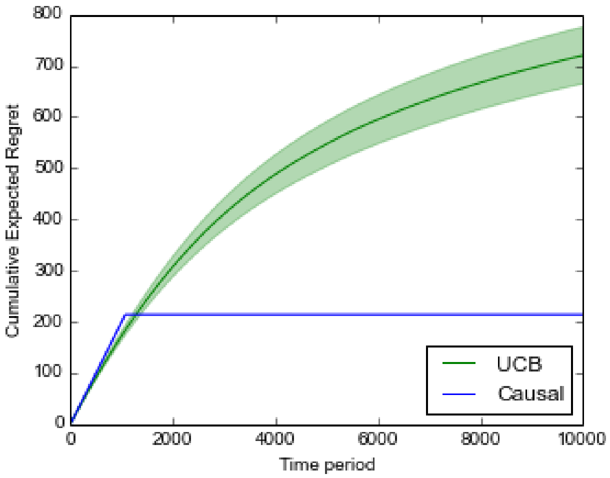
$$R_T \leq 4T^{\frac{2}{3}}m^{2/3}(log(KT))^{\frac{1}{3}} \tag{20}$$

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.

**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.


Cumulative Expected Regret
Time period
UCB
Causal