Intervention Bandits

Blah blah

April 28, 2015

Abstract

An abstract.

1 Introduction

Useful references are: **?**.

2 Notation

Assume we have a known causal model with binary variables $\boldsymbol{X} = \{X_1..X_K\}$ that independently cause a target variable of interest Y, figure . 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, $(\boldsymbol{X}_t, Y_t)$ This problem can be viewed as a variant of the multi-armed bandit problem.

Let $p \in [0,1]^K$ be a fixed and known vector. In each time-step t:

1. The learner chooses an $I_t \in \{1, \ldots, K\}$ and $J_t \in \{0, 1\}$.
2. Then $X_t \in \{0,1\}^K$ is sampled from a product of Bernoulli distributions, $X_{t,i} \sim \text{Bernoulli}(p_i)$
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 $Y_t \sim \text{Bernoulli}(q(\tilde{X}))$ where $q : \{0,1\}^K \to [0,1]$ is unknown and arbitrary.

The expected reward of taking action i, j is $\mu_{i,j} = \mathbb{E}[q(X)|do(X_i = j)]$. The optimal reward and action are 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

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

3 Estimating $\mu_{i,j}$

The most natural way to estimate $\mu_{i,j}$ is to compute an empirical estimate based on samples when that action was taken. This approach would lead directly to the UCB algorithm with $2K$ actions and a regret bound that depended linearly on K.

In this instance we can significantly outperform the standard approach by exploiting the known causal structure of the problem.

$$
\begin{aligned}
P(Y|do(X_i = j)) &= P(Y|X_i = j) && (1)\\
&= \sum_b P(Y|X_i = j, X_a = b)P(X_a = b|X_i = j) && (2)\\
&= \sum_b P(Y|X_i = j, X_a = b)P(X_a = b),\ \forall a \in \{1...K\}/i \text{ as } X_a \perp\!\!\!\perp X_i && (3)\\
&= \sum_b P(Y|X_i = j, do(X_a = b))P(X_a = b) && (4)
\end{aligned}
$$

3.1 Estimators

Fix some time-step t and $i \in \{1, \dots, K\}$ and $j \in \{0, 1\}$.

Considering equation 4, a natural way to construct an empirical estimator, $\hat{\mu}_a$, for $P(Y|do(X_i = j))$ utilizing the information obtained by intervening on variable a is:

$$
\hat{\mu}_a = \begin{cases} \frac{m_{a,1}}{n_{a,1}} p_a + \frac{m_{a,0}}{n_{a,0}}(1 - p_a) & \text{if } a \neq i \\ \frac{m_{i,j}}{n_{i,j}} & \text{if } a = i \end{cases} \tag{5}
$$

$$
\tag{6}
$$

where $\frac{m_{a,b}}{n_{a,b}}$ is an empirical estimator for $P(Y|X_i = j, do(X_a = b))$,

$$
m_{a,b} = \sum_{s=1}^{t} \mathbb{1}\{X_i = j, I = a, J = b, Y = 1\}_s \tag{7}
$$

$$
n_{a,b} = \sum_{s=1}^{t} \mathbb{1}\{X_i = j, I = a, J = b\}_s \tag{8}
$$

This gives K estimators $\{\hat{\mu_1}...\hat{\mu_K}\}$ to be pooled into a single estimator $\hat{\mu}$.

$$
\hat{\mu} = \sum_{a=1}^{K} w_a \hat{\mu}_a = w_i \frac{m_{i,j}}{n_{i,j}} + \sum_{a \neq i} w_a \left[p_a \frac{m_{a,1}}{n_{a,1}} + (1 - p_a) \frac{m_{a,0}}{n_{a,0}} \right] \tag{9}
$$

$$
\tag{10}
$$

If p is not known p_a could be replaced with its empirical estimate $\hat{p}_a$

$$\hat{p}_a = \frac{\sum_{s=1}^{t} \mathbb{1}\{X_a = 1, I \neq a\}_s}{\sum_{s=1}^{t} \mathbb{1}\{I \neq a\}_s} \tag{11}$$

Now I want to select weights w to minimize the variance in $\hat{\mu}$ and then get some kind of high probability bound for it. However, if I'm not very careful about how I pool things in equation 9, the estimator could get very biased, ie estimators $\hat{\mu_a}$ where I have data only for one state of the variable a are not very informative, so should tend to get low weight, but if we choose to sample the second state based on previous payoffs - then we could end up assigning high weights to those estimators that had fluctuated above the true value...

This complicates getting correct uncertainty bounds for a UCB type algorithm/proof.

Consider, what exactly you mean by not unbiased ...

4 Importance Sampling Approach

Consider down weighting estimators proportional to the number of times we played them

5 Observe then exploit

5.1 Basic idea

Basic idea - observe until the uncertainty on all actions is smaller than ϵ, then exploit (or switch to standard UCB). This is only really going to work if $p_a \sim 0.5$.

Bounds on regret: when observing the regret for each timestep is at most 1 (since the reward is $\in [0, 1]$). When exploiting regret it at most ϵ. If the horizon is n, the cost of exploiting $< \epsilon n$ and the cost of observing is $O(1/\epsilon^2)$ (this comes from how quickly we converge to within ϵ for all arms).

$$R_n = O(n\epsilon + \frac{1}{\epsilon^2}) \tag{12}$$

Differentiating and selecting $\epsilon = (\frac{2}{n})^{1/3}$ to minimize the regret yields:

$$R_n = O(n^{2/3}) \tag{13}$$

For each arm $X_i = j$, we have:

$$\hat{\mu}_{ij} = \frac{m_{ij}}{n_{ij}} \tag{14}$$

where

$$m_{i,j} = \sum_{s=1}^{t} \mathbb{1}\{X_i = j, Y = 1\} \tag{15}$$

$$n_{i,j} = \sum_{s=1}^{t} \mathbb{1}\{X_i = j\} \tag{16}$$

$$\tag{17}$$

and Hoeffding's Inequality gives:

$$P\left(|\hat{\mu}_{ij} - \mu_{ij}| > \sqrt{\frac{1}{2n_{ij}} \log \frac{2}{\delta}}\right) \leq \delta \tag{18}$$

5.2 Alternate Observe and Exploit With Some Decay

Keep track of two separate bounds and always choose the narrower of the two.

5.3 Observe until number of plausibly optional arms $< \alpha$

Define the set of plausibly optimal arms to be those with an upper confidence bound that is higher than the largest lower confidence bound.

$$A_t = \{k : UB(k) \geq max_{k'}(LB(k'))\} \tag{19}$$

We then observe until the size of this set is less than some value α before switching to UCB. This takes into account that we don't need narrow confidence bounds on the arms with very low expected reward. Once we start doing UCB, we keep track of the ucb bound separately and for each arm use whichever bounds are narrower.

5.3.1 Regret during observe phase

Let n be the number of timesteps until the size of the set A_t falls below α or we reach the overall horizon T

$$n = \min(\{t : size(A_t) < \alpha, T\}) \tag{20}$$

Let $Q(k)$ be the expected reward for arm k and μ^* be the expected reward for the best arm.

$$R_n = \max_{i=1\ldots K} E\left[\sum_{t=1}^{n} Y_{it} - \sum_{t=1}^{n} Y_{I_t,t}\right] \tag{21}$$

$$= E[n]\left(\mu^* - E[Q]\right) \tag{22}$$

Without loss of generality, assume the arms are ordered from best to worst.

$$Q(k = 1...K) = [\mu^*, \mu^* - d_2 ... \mu^* - d_k] \tag{23}$$

$$\implies E[Q] = \mu^* - \frac{1}{K}\sum_{k=2}^{K} d_k \tag{24}$$

$$\implies R_n = E[n]\left(\frac{1}{K}\sum_{k=2}^{K} d_k\right) \tag{25}$$

Whenever the following two statements hold only arms with indexes less than α can be in the set A_t:

$$\mu^* - \hat{\mu}^* \leq \frac{d_\alpha}{2} \tag{26}$$

$$\hat{\mu_k} - \mu_k \leq \frac{d_k}{2} \; \forall k \geq \alpha \tag{27}$$

$$\implies \rho(t) \equiv P(\exists k \in A_t : k \geq \alpha) \leq P(\mu^* - \hat{\mu}^* > \frac{d_\alpha}{2}) + \sum_{k=\alpha}^{K} P(\hat{\mu_k} - \mu_k > \frac{d_k}{2}) \tag{28}$$

$$P(n > t) = \begin{cases} 0 & \text{if } t \geq T \\ P(\exists k \in A_t : k \geq \alpha) & \text{otherwise} \end{cases} \tag{29}$$

We now concentrate bounding 28. For random variables $W_{sij} = P(Y|X_i = j)$ drawn from a Bernoulli distribution with mean p_{ij} the Chernoff inequality gives:

$$P(\frac{1}{t}\sum_s W_{sij} - p_{ij} \geq \gamma) \leq \exp(-\frac{t\gamma^2}{2p_{ij}(1-p_{ij}) + 2\gamma/3}) \tag{30}$$

$$\implies P(\frac{1}{t}\sum_s \frac{W_{sij}}{p} - \frac{p_{ij}}{p} \geq \epsilon) = P(\hat{\mu}_k - \mu_k \geq \epsilon) \leq \exp(-\frac{tp\epsilon^2}{2 + 2\epsilon/3}) = \exp(-\frac{t\epsilon^2}{4 + 4\epsilon/3}) \tag{31}$$

Where $p_{ij} = P(Y = 1, X_i = j) \leq P(X_i = j) = p = 1/2 \quad \forall (i, j)$

Similarly,

$$P(\mu^* - \hat{\mu}^* \geq \epsilon) \leq exp(-tp^2\epsilon^2) = exp(-\frac{t\epsilon^2}{4}) \tag{32}$$

$$\implies \rho(t) \leq \exp(-\frac{d_\alpha^2 t}{16}) + \sum_{k=\alpha}^{K} exp(-\frac{d_k^2 t}{16(1 + d_k/6)}) \tag{33}$$

$$E[n] = \sum_{t=1}^{T} P(n \geq t) \leq \sum_{t=1}^{T} \rho(t) \tag{34}$$

$$= \min\left(T, \frac{1 - e^{-d_\alpha^2 T/16}}{e^{d_\alpha^2/16} - 1} + \sum_{k=\alpha}^{K} \frac{1 - e^{-d_k^2 T/16(1+d_k/6)}}{e^{d_k^2/16(1+d_k/6)} - 1}\right) \tag{35}$$

Since the regret is $R_n = E[n]\left(\frac{1}{K}\sum_{k=2}^{K} d_k\right)$ the worst case is that just enough of the differences d_k are 0 to make $E[n] = T$ and the remainder are 1. This occurs if $d_\alpha = 0$, In this case the regret is:

$$R_n = T\left(\frac{K - (\alpha + 1)}{K}\right) \tag{36}$$

Clearly, this is not quite the right criteria to decide when to switch ...

What about if I

5.3.2 Regret on observing for h rounds before switching to UCB

How large the regret is if we initialize with a certain number of counts already in place...

Suppose we have O_i observations for arm i. We will now begin doing standard UCB - storing a separate set of bounds, only each time we choose an action we will pick the narrower of the observational/interventional confidence bounds.

$$\hat{\epsilon}_{it} = \min\left(\sqrt{\frac{\alpha \log(t)}{2T_{it}}}, \sqrt{\frac{log(1/\delta)}{2O_i}}\right) \tag{37}$$

$$= \sqrt{\frac{\alpha \log(t)}{2\max(T_{it}, O_i)}} \text{ if we let } \delta = t^{-\alpha} \tag{38}$$

When running UCB, in order to select a non-optimal arm i at timestep t, at least one of the following statements must be true:

1. $\hat{\mu}_{i^*t} + \hat{\epsilon}_{i^*t} \leq \mu_i + \Delta_i$
2. $\hat{\mu}_{it} + \hat{\epsilon}_{it} \geq \mu_i + 2\hat{\epsilon}_{it}$
3. $\Delta_i < 2\hat{\epsilon}_{it}$

Given our expression for the confidence bound in equation 38 we can write statement (3) as:

$$\max(T_{it}, O_i) < \frac{2\alpha \log(t)}{\Delta_i^2} \tag{39}$$

Let:

$$\beta = \frac{2\alpha \log(n)}{\Delta_i^2} \tag{40}$$

If $O_i \geq \beta$, statement (3) is guaranteed to be false for all t and we can bound the expected number of times we select i with how many times we expect the bounds given by (1) and (2) to be true.

For arms i such that $O_i \geq \beta$

$$P(1 \text{ is true}) < t^{1-\alpha} \tag{41}$$
$$P(2 \text{ is true}) < t^{1-\alpha} \tag{42}$$
$$\implies P(I_t = i) < 2t^{1-\alpha} \tag{43}$$
$$\implies R_n(i) \leq \frac{\Delta_i \alpha}{\alpha - 2} \tag{44}$$

For all other arms with $\Delta_i > 0$, the bound is the same as standard UCB

$$R_n(i) \leq \frac{2\alpha \log(n)}{\Delta_i} + \frac{\Delta_i \alpha}{\alpha - 2} \tag{45}$$

Assume we observed for h steps. The expected regret from this period is just $hE[\Delta_i]$. Let $T = h + n$ be the total time.

$$R_T < \sum_{i=1}^{K} \Delta_i \left(\frac{h}{K} + \frac{\alpha}{\alpha - 2} \right) + \sum_{i:\Delta_i > 0} P(O_i \leq \beta) \frac{2\alpha \log(n)}{\Delta_i} \tag{46}$$

Using Hoeffding's Inequality:

$$P(O_i \leq \beta) \leq \begin{cases} 2e^{-2h(\frac{1}{2} - \frac{\beta}{h})^2} & \text{if } h > 2\beta \\ 1 & \text{otherwise} \end{cases} \tag{47}$$

Its tricky - selecting the h to switch at based on the inequality can certainly yield a non-optimal answer. Maybe I should figure this out via simulation?? But some theory to tell me what to simulate would certainly help.

5.3.3 General rough thoughts

When observing the confidence interval on all the arms drops approximately with $\sqrt{\frac{1}{n}}$ (provided that the probability of each occurring is the same) - otherwise for each arm it will be weighted by this drop will be weighted by a function of the probability that it occurs. When acting, the interval on a single arm we intervene on drops. What is the informational value of reducing the bounds on various arms (it is higher for arms that are more likely to be good).

5.4 Bounding a weighted sum of unbiased estimators with McDiarmid's Ineqaulity

McDiarmid's Inequality states: If $X_i \perp\!\!\!\perp X_j$ and

$$|\phi(X_1...X_i...X_N) - \phi(X_1...X_i^{'}...X_N)| < c_i \ \ \forall i \tag{48}$$

$$P\left(|\phi(\boldsymbol{X}) - E[\phi(\boldsymbol{X})]| \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_i c_i^2}\right) \tag{49}$$

$$P\left(|\phi(\boldsymbol{X}) - E[\phi(\boldsymbol{X})]| \geq \sqrt{\frac{\sum_i c_i^2}{2}\log\frac{2}{\delta}}\right) \leq \delta \tag{50}$$

In our problem, for each variable V_i and value j we have:

$$\hat{\mu}_a = \tilde{P}(Y|V_i = j) = \begin{cases} \frac{p_a}{n_{a1}}\sum_{s=1}^{n_{a1}} X_{a1,s} + \frac{1-p_a}{n_{a0}}\sum_{s=1}^{n_{a0}} X_{a0,s} & \text{if } a \neq i \\ \sum_{s=1}^{n_i} X_s & \text{if } a = i \end{cases} \tag{51}$$

where $X_{a1,s} \sim P(Y|V_i = j, do(V_a = 1))$ and $X_{a0,s} \sim P(Y|V_i = j, do(V_a = 0))$, both $\in [0, 1]$. We then pool $\{\hat{\mu}_1...\hat{\mu}_K\}$ to get a single estimate.

$$\hat{\mu} = \tilde{P}(Y|V_i = j) = \sum_a w_a\hat{\mu}_a, \text{ where } \sum_a w_a = 1 \tag{52}$$

Let

$$\hat{\mu} = \phi(\boldsymbol{X}) = \sum_{a\neq i} w_a\left[\frac{p_a}{n_{a1}}\sum_{s=1}^{n_{a1}} X_{a1,s} + \frac{1-p_a}{n_{a0}}\sum_{s=1}^{n_{a0}} X_{a0,s}\right] + \frac{w_i}{n_i}\sum_{s=1}^{n_i} X_{i,s} \tag{53}$$

Note: this treats the number of samples n_i,n_{a1}, and n_{a0} as fixed. Even for fixed actions, the latter two are still random variables that depend on $\boldsymbol{p}$ Will this bound still hold?

$$\phi(...X_{\alpha1,s}...) - \phi(...X_{\alpha1,s}^{'}...) \leq w_\alpha\frac{p_\alpha}{n_{\alpha1}} \tag{54}$$

$$\phi(...X_{\alpha0,s}...) - \phi(...X_{\alpha0,s}^{'}...) \leq w_\alpha\frac{1-p_\alpha}{n_{\alpha0}} \tag{55}$$

$$\phi(...X_i...) - \phi(...X_i^{'}...) \leq \frac{w_i}{n_i} \tag{56}$$

$$\sum_i c_i^2 = n_i \left(\frac{w_i}{n_i}\right)^2 + \sum_a \left[n_{a1} \left(w_a \frac{p_a}{n_{a1}} \right)^2 + n_{a0} \left(w_a \frac{1-p_a}{n_{a0}} \right)^2 \right] \tag{57}$$

$$= \frac{w_i^2}{n_i} + \sum_a w_a^2 \left(\frac{p_a^2}{n_{a1}} + \frac{(1-p_a)^2}{n_{a0}} \right) \tag{58}$$

$$= \sum_a w_a^2 f(a), \text{ where} \tag{59}$$

$$f(a) = \begin{cases} \frac{p_a^2}{n_{a1}} + \frac{(1-p_a)^2}{n_{a0}} & a \neq i \\ \frac{1}{n_i} & a = i \end{cases} \tag{60}$$

We want to choose weights $\boldsymbol{w}$ so as to minimize 59 subject to the constraint $\sum_a w_a = 1$.

The minimum (assuming it exists) should occur at a critical point of:

$$L(w_1 ... w_k, \lambda) = \sum_a w_a^2 f(a) + \lambda((\sum_a w_a) - 1) \tag{61}$$

$$\frac{\partial L}{\partial w_a} = 2 w_a f(a) + \lambda = 0 \tag{62}$$

$$\implies w_a = \frac{-\lambda}{2 f(a)} \tag{63}$$

$$\frac{\partial L}{\partial \lambda} = (\sum_a w_a) - 1 = 0 \tag{64}$$

$$\implies -\frac{\lambda}{2} \sum_a \frac{1}{f(a)} = 1 \implies \lambda = -\frac{2}{\sum_a \frac{1}{f(a)}} \tag{65}$$

$$\implies w_a = \frac{1}{f(a) \sum_a \frac{1}{f(a)}} \tag{66}$$

$$\implies \sum_i c_i^2 = \frac{1}{\sum_a \frac{1}{f(a)}} \tag{67}$$

Substituting 67 into the McDiarmid inequality 50:

$$P\left(|\hat{\mu}^{ij} - \mu^{ij}| > \sqrt{\frac{1}{2 \sum_a \eta_a^{ij}} \log \frac{2}{\delta}} \right) \leq \delta \tag{68}$$

where:

$$\mu^{ij} = P(Y|V_i = j) \tag{69}$$

$$\eta_a^{ij} = \begin{cases} \frac{n_{a1}^{ij} n_{a0}^{ij}}{n_{a1}^{ij}(1-p_a)^2 + n_{a0}^{ij} p_a^2} & a \neq i \\ n_{ij} & a = i \end{cases} \tag{70}$$

$$n_{al}^{ij} = \sum_{s=1}^{t} \mathbb{1}\{do(V_a = l), V_i = j\} \tag{71}$$

$$n_{ij} = \sum_{s=1}^{t} \mathbb{1}\{(do(V_i = j)\} \tag{72}$$

All of the above is to get estimates for $P(Y|V_i = j)$ for some fixed i, j. Suppose we explore for some fixed total number of rounds h, the goal is to minimize the worst confidence bound - that doesn't quite make sense as a goal - we just need to be sure that all actions are worse than our estimate of the best with probability δ. ... anyway

We have control of are the number of times we do each action, n_{al}, which obviously also influences the random variables n_{al}^{ij}, subject to the constraint $\sum n_{al} = h$.

Objective: minimize the expectation of the largest confidence bound after a fixed number of rounds n.

Let $\boldsymbol{n} = [n_{11}, n_{10}, n_{21}, n_{20} ... n_{k1}, n_{k0}]$

$$\tilde{\boldsymbol{n}} = \underset{\{\boldsymbol{n}:\|\boldsymbol{n}\|=h\}}{\arg\min} \max_{i,j} \left(E\left[\sqrt{\frac{1}{2\sum_a \eta_a^{ij}} \log \frac{2}{\delta}} \right] \right) \tag{73}$$

We can bound the first part of η_a^{ij}

$$\eta_a^{ij} = \frac{n_{a1}^{ij} n_{a0}^{ij}}{n_{a1}^{ij}(1-p_a)^2 + n_{a0}^{ij} p_a^2} \leq \frac{n_{a1}^{ij} n_{a0}^{ij}}{\max(n_{a1}^{ij}(1-p_a)^2 + n_{a0}^{ij} p_a^2)} \tag{74}$$

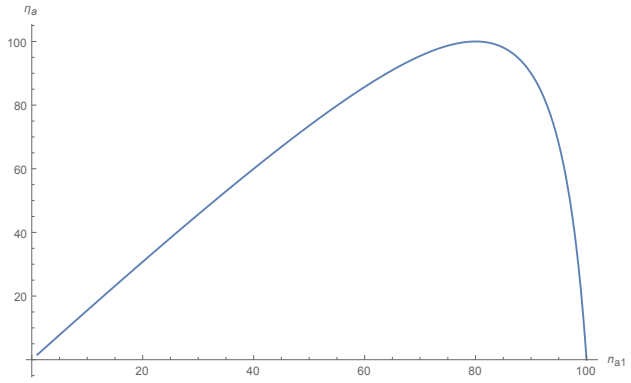
$$= \min\left(\frac{n_{a0}^{ij}}{(1-p_a)^2}, \frac{n_{a1}^{ij}}{p_a^2} \right) \tag{75}$$

$$E\left[\min\left(\frac{n_{a0}^{ij}}{(1-p_a)^2}, \frac{n_{a1}^{ij}}{p_a^2} \right) \right] \leq \min\left(E\left[\frac{n_{a0}^{ij}}{(1-p_a)^2} \right], E\left[\frac{n_{a1}^{ij}}{p_a^2} \right] \right) \tag{76}$$

$$= P(V_i = j) * \min\left(\frac{n_{a0}}{(1-p_a)^2}, \frac{n_{a1}}{p_a^2} \right) \tag{77}$$

$$\tag{78}$$

Figure 1: The effective number of samples η_a versus n_{a1}, where $p_a = .8$ and the total number of samples, $n_{a1}+n_{a0} = 100$. The effective number of samples is maximized (and equals the total) if we sample each side according to its probability.



5.5 Targeted Sampling

Sample a equal number of points from each arm - until the error on all arms is smaller than some threshold. Then exploit. This should do better than observe-exploit, particularly where the probability of some events occurring without intervention is low.

A more sophisticated version:

5.6 UCB variant

Combine estimators according to:

$$w_a = \frac{n_a}{\sum_{a=1}^{K} n_a} \text{ and } n_{i,j} = \begin{cases} n_{i,j} & \text{if } a = i \\ \frac{1}{2}\min\left\{\frac{n_{a,1}}{p_a}, \frac{n_{a,0}}{1-p_a}\right\} & \text{otherwise} \end{cases}$$

In this case the estimators from section 3.1 are biased - so its a lot harder to prove a regret bound.

Theorem 1. *(Probably False) With probability at least* $1-\delta$ *we have that:* $|\hat{\mu}_t - \mu| \leq \sqrt{\frac{\beta}{\sum_a n_a} \log \frac{1}{\delta}}$*, where* $\beta > 0$ *is some constant.*

Proof. First note that $n_{a,b}$ is a random variable that is bounded by t for all a, b. We use the short-hand $\mu_{i,j}^{a,b} = \mathbb{E}[q(X)|X_i = j, X_a = b]$. Then

$$\mu_{i,j} = p_a \mu_{i,j}^{a,1} + (1-p_a)\mu_{i,j}^{a,0}\,.$$

Now we can apply Hoeffding's bound and the union bound to show that

$$\mathrm{P}\left\{\left|\frac{m_{a,b}}{n_{a,b}} - \mu_{i,j}^{a,b}\right| \geq \sqrt{\frac{1}{2n_{a,b}} \log \frac{4t}{\delta}}\right\} \leq \frac{\delta}{2}\,.$$

Therefore by the union bound

$$\mathrm{P}\left\{\left|p_a\frac{m_{a,1}}{n_{a,1}}+(1-p_a)\frac{m_{a,0}}{n_{a,0}}-\mu_{i,j}\right|\geq p_a\sqrt{\frac{1}{2n_{a,1}}\log\frac{4t}{\delta}}+(1-p_a)\sqrt{\frac{1}{2n_{a,0}}\log\frac{4t}{\delta}}\right\}\leq\delta$$

Now by Jensen's inequality

$$\begin{aligned}p_a\sqrt{\frac{1}{2n_{a,1}}\log\frac{4t}{\delta}}+(1-p_a)\sqrt{\frac{1}{2n_{a,0}}\log\frac{4t}{\delta}}&\leq\sqrt{\left(\frac{p_a}{2n_{a,1}}+\frac{1-p_a}{2n_{a,0}}\right)\log\frac{4t}{\delta}}\\&\leq\sqrt{\max\left\{\frac{p_a}{n_{a,1}},\frac{1-p_a}{n_{a,0}}\right\}\log\frac{4t}{\delta}}\\&=\sqrt{\frac{1}{2n_a}\log\frac{4t}{\delta}}\,.\end{aligned}$$

Similarly,

$$\mathrm{P}\left\{\left|\frac{m_{i,j}}{n_{i,j}}-\mu_{i,j}\right|\geq\sqrt{\frac{1}{2n_a}\log\frac{4t}{\delta}}\right\}\leq\mathrm{P}\left\{\left|\frac{m_{i,j}}{n_{i,j}}-\mu_{i,j}\right|\geq\sqrt{\frac{1}{2n_a}\log\frac{2t}{\delta}}\right\}\leq\delta\,.$$

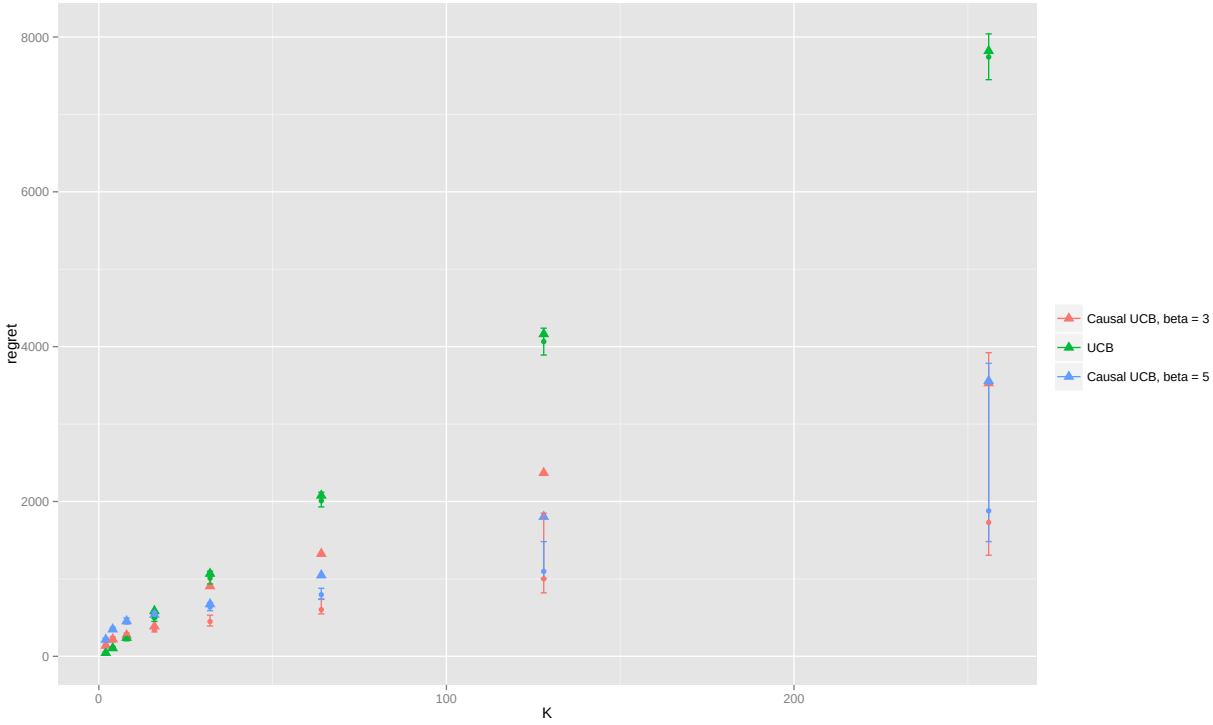
□

6 Algorithm

Algorithm 1 UCB

Input: Number of variables K, vector $p\in[0,1]^K$, horizon n
for $t\in1,\ldots,n$ **do**
  for $i\in1,\ldots,K$ **do**
    for $j\in\{0,1\}$ **do**
      Compute $\tilde{\mu}_{i,j}=\hat{\mu}_{i,j}+\sqrt{\frac{\alpha}{\sum_a n_a}\log n}$
    end for
  end for
  Choose $I_t,J_t=\arg\max_{i,j}\tilde{\mu}_{i,j}$
end for

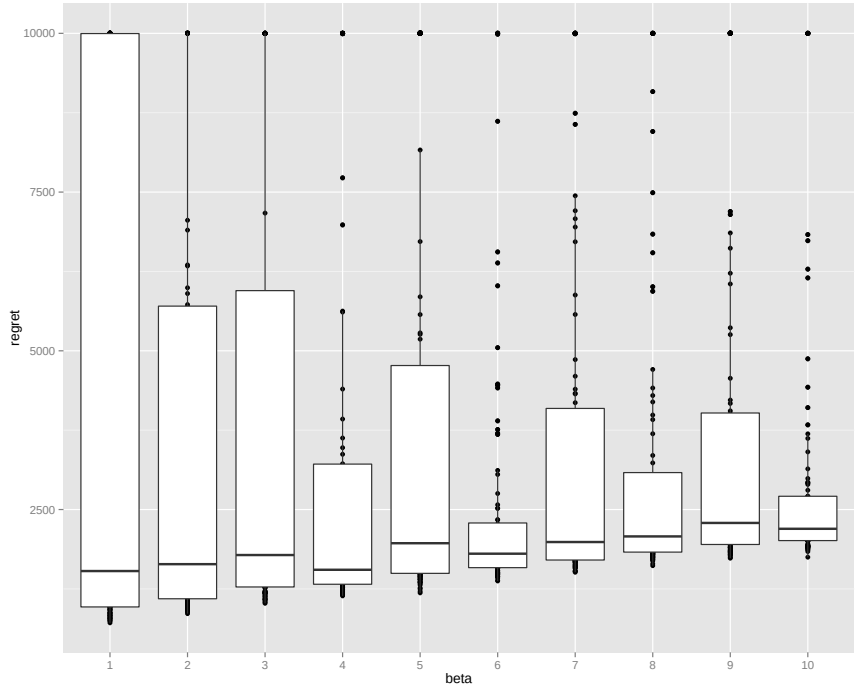
Figure 2: Comparison of the performance of standard UCB versus causal UCB with $\beta = 3$ and $\beta = 5$. 100 simulations were run for each algorithm up to a horizon of 10^5 per value of K. Error bars span the 1st to 3rd quantile of the regret, round points mark the median and triangular points show the mean. For standard UCB the regret increases linearly with the number of arms K. For causal UCB the increase is sub-linear. Increasing β leads to slower convergence but lower variance.



7 Theorems

8 Experiments

Figure 3: The distribution of regret varies with the β parameter in the bound in the estimator. As beta increases, the mean regret increases but the variance decreases. The plot shows the results of running 100 independent bandits, with $K = 256$ and $\epsilon = 0.1$, up to a horizon $h = 10^5$ for each value of β.



Simululations to compare the performance of standard UCB with our modified algorithm. For each number of arms, 100 bandits of each type were created and run upto to a horizon of 1000 timesteps. The mean regret and its standard error from these simulations is plotted in figure **??** The true data was generated from a model where:

$$p = [0.5]^K$$
$$q(\boldsymbol{X}) = \begin{cases} 0.5 & \text{if } X_1 = 0 \\ 0.6 & \text{otherwise} \end{cases}$$

9 Conclusion

10 Notes

10.1 Why Hoeffdings bound doesn't hold if n is a random variable

Let $\{Z_1, Z_2 ... Z_n\} \sim Bernoulli(\frac{1}{2})$ and $X_i = 2Z_i - 1 \implies X_i \in \{-1, 1\}$ and $E[X_i] = 0$.

For a fixed n, Hoeffdings inequality says:

$$P\left(|\frac{1}{n}\sum_{i=1}^{n} X_i| > \epsilon\right) \leq 2e^{-n\epsilon^2} \tag{79}$$

$$\implies P\left(|\frac{1}{n}\sum_{i=1}^{n} X_i| > \sqrt{\frac{\log 4}{n}}\right) \leq \frac{1}{2} \tag{80}$$

$$\implies P\left(|\sum_{i=1}^{n} X_i| > \sqrt{n \log 4}\right) \leq \frac{1}{2} \tag{81}$$

If n is allowed to be dependent on the sequence of values sampled, this inequality no longer holds.

Proof. Choose n based on the sequence of samples seen so far such that:

$$n = min\{n : n > 4 \text{ and } \sum_{i=1}^{n} X_i > \sqrt{n \log\log n}\}$$

By the law of iterated logarithms this quantity is finite with probability ~ 1.

$$\implies P\left(|\sum_{i=1}^{n} X_i| > \sqrt{n \log\log n}\right) \sim 1$$

$$\implies P\left(|\sum_{i=1}^{n} X_i| > \sqrt{n \log 4}\right) \sim 1, \text{ Since } \log\log n > \log 4 \ \ \forall n > 4$$

Thus Hoeffdings inequality (equation 79) does not hold in general if n is not independent of the samples $\{X_i\}$ (The bound does not work if you decide to stop sampling as soon as you reach a point where random walk fluctuations take you outside it)

□

10.2 Importance weighted estimators

Now Hoeffdings bound gives us $P(\hat{\mu}_i - \mu_i > \frac{D}{2}|n_i)$

$$P(\Delta_{\hat{i^*}} > D) \leq K \sum_{\boldsymbol{n}} P(\hat{\mu}_i - \mu_i > \frac{D}{2}|n_i) P(\boldsymbol{n}) \tag{82}$$

$$= K \sum_{\boldsymbol{n}} e^{-n_i D^2/2} P(\boldsymbol{n}) \tag{83}$$

Where $\boldsymbol{n} = [n_1 ... n_K]$ is a vector of the number of observations we have for each arm and $P(\boldsymbol{n}) \sim multinomial(\boldsymbol{p}, h)$ To approximate this, we break the space of $\boldsymbol{n}$ up into those vectors where $n_i > h/4 \forall i$ and those where $\exists i$ such that $n_i < h/4$.