Intervention Bandits

Blah blah

November 27, 2014

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. We can run sequential experiments on the system, where at each timestep t we can select a variable on which to intervene and then we 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 $I_t \in \{1, \dots, K\}$ and $J_t \in \{0, 1\}$.
2. Then $X_t \in \{0, 1\}^K$ is sampled from a product of Bernoulli distributions, $X_{t,i} \sim \mathrm{Bernoulli}(p_i)$
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 $Y_t \sim \mathrm{Bernoulli}(q(\tilde{X}))$ 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

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

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 this approach by exploiting the known causal structure of the problem.

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

3.1 Estimators

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

Let $\hat{\mu}_a$ be an empirical estimator for $P(Y|do(X_i=j))$ obtained via marginalization over X_a.

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

where:

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

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

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

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

where:

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

If p is not known, these expression are unchanged except that p_a is replaced with $\hat{p}_a$

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

3.2 Random sampling variant

Basic idea - explore randomly until the uncertainty on all actions is smaller than ϵ, then exploit (or switch to standard UCB). Bounds on regret: when exploring regret is at most 1, when exploiting regret it at most ϵ. If the horizon is n, the cost of exploiting $< \epsilon n$ and the cost of exploring is $O(1/\epsilon^2)$ (this comes from how quickly we converge to within ϵ for all arms).

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

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

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

At each time step we randomly select a variable X_i on which to intervene and set it to 1 with probability p_i.

The estimators from section XX are unbiased. So we can get an estimate of the error in the estimate by

we are pooling estimators - so what is the error in a pooled estimate - is it the same as what Tor wrote before (because he was assuming unbiased estimators? Or is it tighter because we don't need to worry about the the non-fixed nature of ...) I could try a couple of different bounds and see what happens.

3.3 UCB variant

In this case the estimators from section XX 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,

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

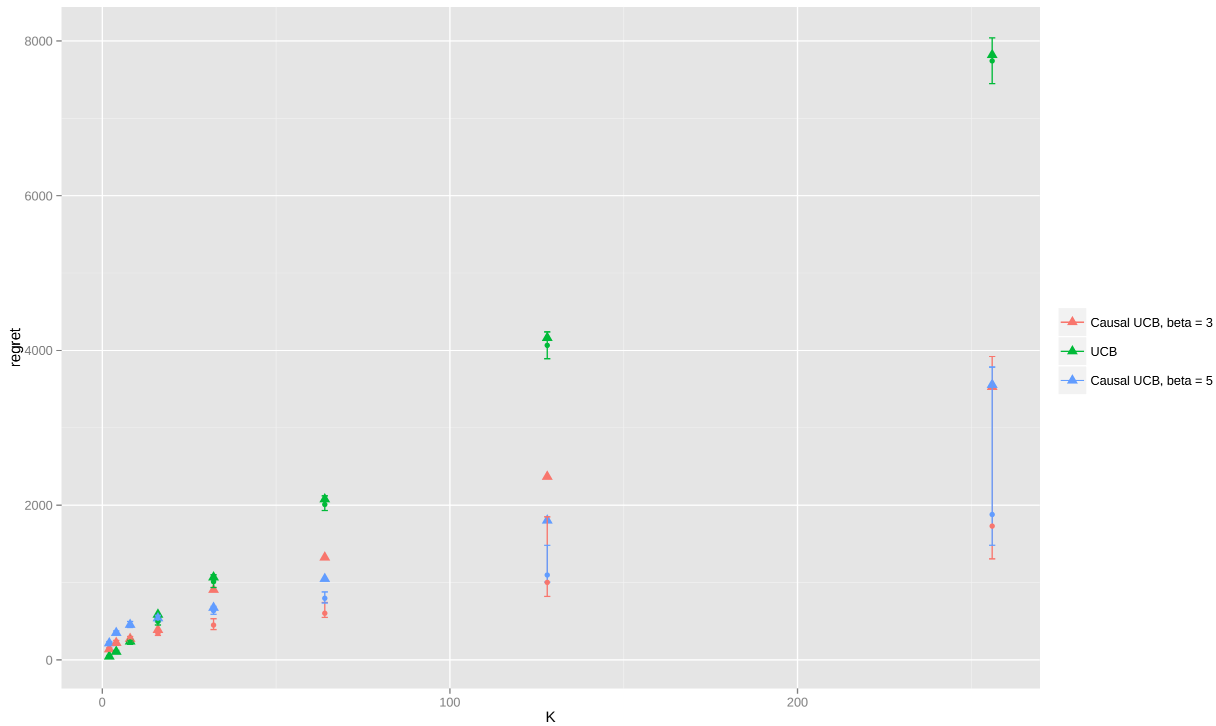
□

4 Algorithm

Algorithm 1 UCB

Input: Number of variables K, vector $p \in [0,1]^K$, horizon n
for $t \in 1, \dots, n$ **do**
    for $i \in 1, \dots, K$ **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

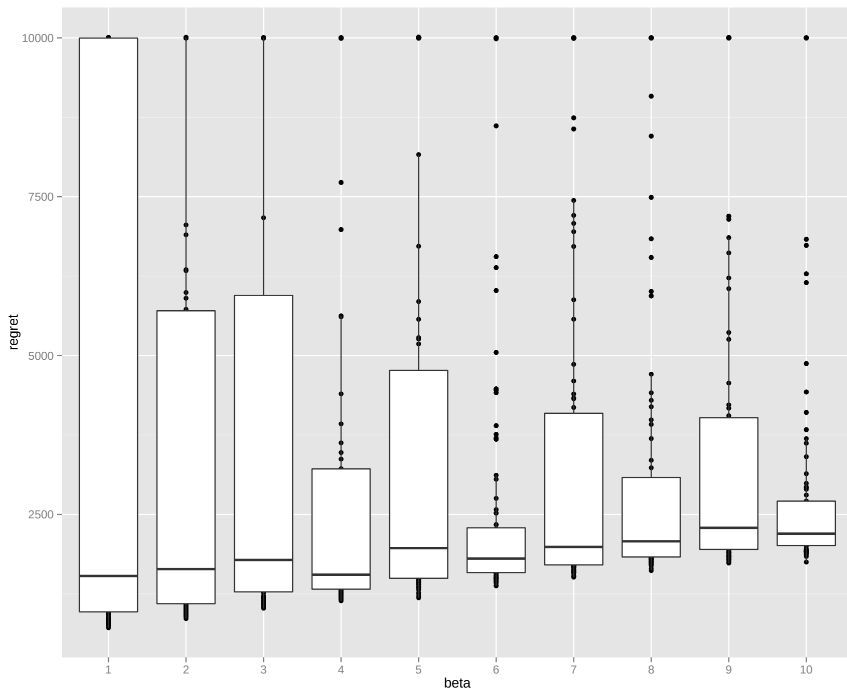
Figure 1: Comparison of the performance of standard UCB versus causal UCB with $\beta = 3$ and $\beta = 5$. 100 simulations were run for each algorithm up to a horizon of 10^5 per value of K. Error bars span the 1st to 3rd quantile of the regret, round points mark the median and triangular points show 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.



5 Theorems

6 Experiments

Figure 2: 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:

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

7 Conclusion

8 Notes

8.1 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:

$$\begin{aligned} & P\left(|\frac{1}{n}\sum_{i=1}^{n} X_i| > \epsilon\right) \leq 2e^{-n\epsilon^2} \\ \implies & P\left(|\frac{1}{n}\sum_{i=1}^{n} X_i| > \sqrt{\frac{\log 4}{n}}\right) \leq \frac{1}{2} \\ \implies & P\left(|\sum_{i=1}^{n} X_i| > \sqrt{n \log 4}\right) \leq \frac{1}{2} \end{aligned} \tag{3}$$

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.

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

Thus Hoeffdings inequality (equation 3) 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)

□