048843 - HW2

Itay Hubara and Zohar Rimon

April 2020

1 Exercise E1

1.1

The results of the run:

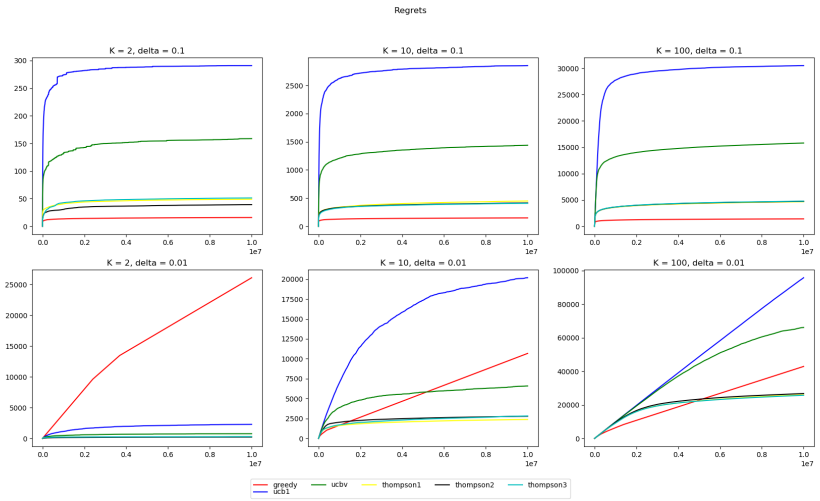

Figure 1: Mean regret vs time

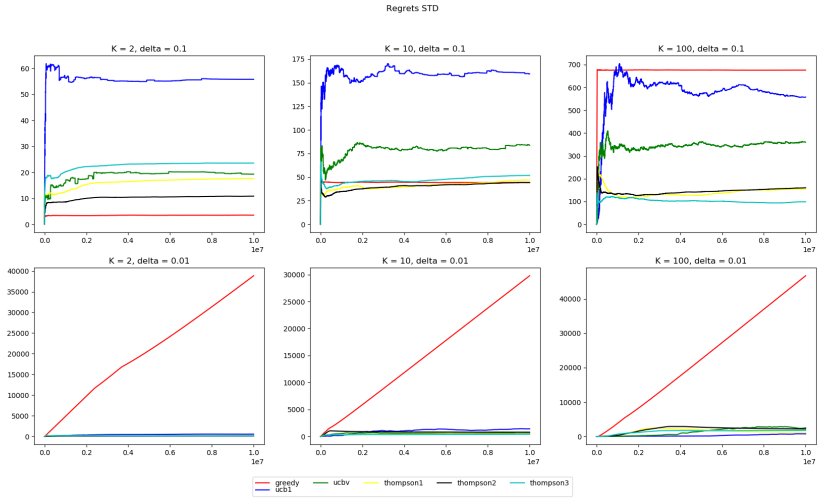

Figure 2: std of the regrets vs time

And the exploration index:

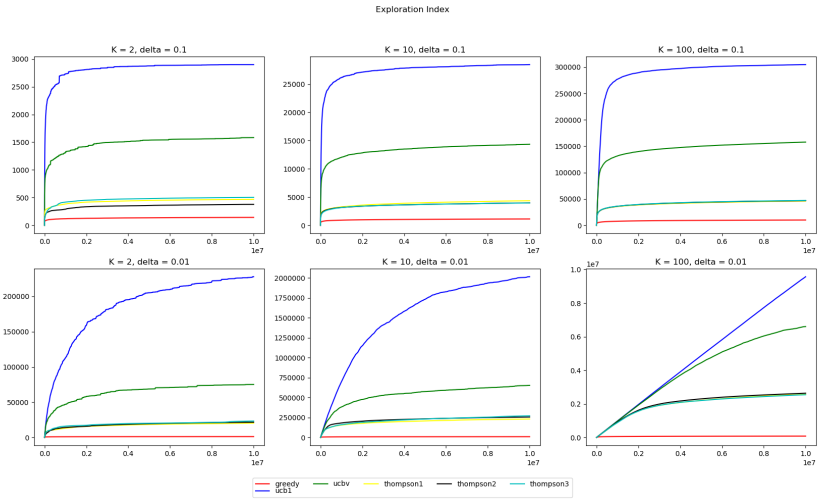

Figure 3: exploration index vs time

From the exploration index, we can understand the behaviour of the algorithms. We can see, for example, that the decaying epsilon greedy works well for the case of large δ. Though, for small deltas the algorithm performs purely due to the early stop of exploration.

1.2

We expect that for a relatively simple scenario, where there is no need to encourage exploration of arms that were not chosen frequently the UCB-1 will out perform the UCB-V algorithm. We expect that because the UCB-V algorithm will explore "obviously" sub optimal arms. We test this with a simple scenario of 10 arms with $\delta = 0.4$, thus the optimal arm is easy to find.

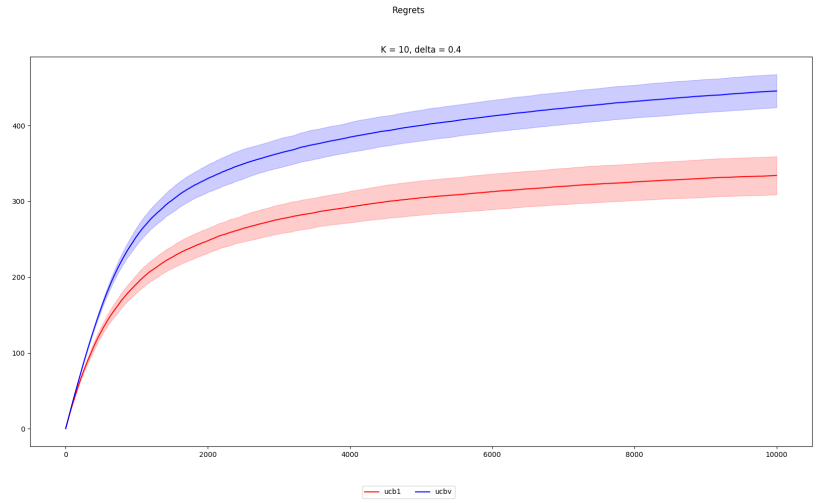

Figure 4: UCB-1 is outperforming UCB-V in this simple case

We ran this scenario 100 times for each algorithm in order to get a good estimate of the mean regret.

2 Exercise T1

For a fix t, $\mathbb{E}[\Delta a_t] = \mathbb{E}[r_t^* - r_t] = r(a_t^*) - \mathbb{E}[r(\hat{a}_t)](1-\epsilon) - \mathbb{E}[r(\bar{a}_t)]\epsilon$. Where $r(a_t^*)$ is the reword of the arm which has the highest empirical reward $\bar{\mu}_t(a) = argmax_a(\sum_{i=1}^{t} r(a_i))$ and $r(\bar{a}_t) = \frac{1}{K}\sum_a(r(a_t))$.

For each arm a, we'll define a reward tape: an $1 \times t$ table in which each cell independently sampled from D_a as done in section 1.3.1. Similarly, without loss of generality, the reward tape encodes rewards as follows: the j-th time a given arm a is chosen by the algorithm, its reward is taken from the j-th cell in this arm's tape. Let $\bar{(v)}_j$ represent the average reward of arm a from first j times that arm a was chosen. Now we can use Hoeffding inequality to derive that

$$\forall j \;\; Pr(|\bar{(v)}_j(a) - \mu(a)| \leq r_t(a)) \geq 1 - \frac{2}{t^4} \quad \text{where} \;\; r(t) = \sqrt{\frac{2log(t)}{n_t(a)}} \qquad (1)$$

Taking a union bound, it follows that (assuming K = arms $\leq t$)

$$Pr(\forall a \forall j \ \ |\bar{(v)}_j(a) - \mu(a)| \leq r_t(a)) \geq 1 - \frac{2}{t^2} \tag{2}$$

Which implies that

$$Pr(\forall a \forall t \ \ |\bar{\mu}_(a) - \mu(a)| \leq r_t(a)) \geq 1 - \frac{2}{t^2} \tag{3}$$

which we'll define to be the clean event.

For $\epsilon_t = t^{-1/3} \cdot (Klogt)^{1/3}$ the expected number of exploration rounds up to round t is on the order of $t^{2/3} \cdot (Klogt)^{1/3}$ thus for each time t the exploration rounds adds to the regret $t^{2/3} \cdot (Klogt)^{1/3}$ while the exploitation rounds adds:

$$O\left(\sqrt{\frac{2log(t)}{n_t(a)}}\right) \cdot t(1 - t^{1/3} \cdot (Klog(t))^{1/3}) \tag{4}$$

Since the number of samples from arm a until round t is lowerbounded by the number of times it was samples for the exploration, $n_t(a) \geq \frac{t^{2/3} \cdot (Klogt)^{1/3}}{K}$. Therefore, the exploitation rounds requires:

$$O\left(\sqrt{\frac{2log(t)}{n_t(a)}}\right) \cdot t(1 - t^{-1/3} \cdot (Klog(t))^{1/3}) =$$
$$O\left(\sqrt{\frac{2Klog(t)}{t^{2/3} \cdot (Klogt)^{1/3})}}\right) \cdot t(1 - t^{-1/3} \cdot (Klog(t))^{1/3}) =$$
$$O\left(t^{-1/3}(Klog(t))^{1/3} \cdot t(1 - t^{-1/3} \cdot (Klog(t))^{1/3})\right) = O\left(t^{2/3}(Klog(t))^{1/3}\right) \tag{5}$$

Thus the clean even adds $O\left(t^{2/3}(Klog(t))^{1/3}\right)$ to the expected regret. Since the bad event is bounded by $t \cdot \frac{1}{t^2}$ we get that:

$$\mathbb{E}[R(T)] = \mathbb{E}[R(T)|cleanevent][cleanevent] + \mathbb{E}[R(T)|badevent][badevent] =$$
$$O\left(t^{2/3}(Klog(t))^{1/3}\right) + t(\frac{1}{t^2}) = O\left(t^{2/3}(Klog(t))^{1/3}\right)$$

3 Exercise T2

Suppose $T_i = 2^i$ and $T_{true} = \sum_{i=0}^{n} T_i$ and the regret of the bandit algorithm is bounded $R(T) \leq O(f(T))$ for a know T and a monotonically increasing function $\mathbb{E}[R(T)]$. Because the length of our interval increases by a power of two, the amount of regret we accumulate also increases by a power of two. So, the sum of our costs is a geometric series, which is dominated by the final term asymptotically. Thus we get that for any regret function of the form $\mathbb{E}[R(T)] = C \cdot T^{\gamma}(log(T))^{\delta}$ with $C, \gamma, \delta > 0$

we get that

$$\begin{aligned}\sum_{m=1}^{log_2(T)} \mathbb{E}[R(2^m)] = \sum_{m=1}^{log_2(T)} C \cdot 2^{m\gamma}(log(2^m))^\delta \leq C \cdot log(T)^\delta \sum_{m=1}^{log_2(T)} 2^{m\gamma} \\ = C \cdot log(T)^\delta \frac{1-(2^\gamma)^{log_2(T)+1}}{1-2^\gamma} = C \cdot T^\gamma (log(T))^\delta \\ = O(R(T))\end{aligned}$$

4 Exercise T3

(a) General form of Bayes estimator:

$$\hat{\theta}_{Bayes} = \arg\max p(\theta|D_n) = \arg\max p_0(\theta)p(D_n|\theta) \tag{6}$$

In our case:

$$p_0(\theta) = \frac{\theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}}{B(\alpha,\beta)} \quad ; \quad p(D_n|\theta) = \prod_{i=1} n\theta^{x_i} \cdot (1-\theta)^{1-x_i} \tag{7}$$

Thus the Bayes estimator is:

$$\begin{aligned}\hat{\theta}_{Bayes} &= \arg\max \frac{\theta^{\alpha-1} \cdot (1-\theta)^{\beta-1}}{B(\alpha,\beta)} \prod_{i=1} n\theta^{x_i} \cdot (1-\theta)^{1-x_i} \\ &= \arg\max \left(\theta^{\alpha-1} \cdot (1-\theta)^{\beta-1} \prod_{i=1} n\theta^{x_i} \cdot (1-\theta)^{1-x_i} \right) \\ &= \arg\max \left((\alpha-1)log(\theta) + (\beta-1)\log(1-\theta) + \sum_{i=1}^{n} (1-x_i) log(1-\theta) \right)\end{aligned}$$

By setting the derivation of the above equation to zero we get:

$$\frac{\alpha-1}{\theta} - \frac{\beta-1}{1-\theta} + \sum_{i=1}^{n} \frac{1-x_i}{1-\theta} = 0 \Rightarrow \alpha - 1 + (\alpha+\beta+n-2)\theta - \sum_{i=1}^{n} x_i = 0$$

$$\hat{\theta}_{Bayes} = \frac{S_n + \alpha - 1}{\alpha + \beta + n - 2}$$

where $S_n = \sum_{i=1}^{n} x_i$. Now, lets derive the MLE estimator:

$$\begin{aligned}\hat{\theta}_{MLE} &= \arg\max\left(P(D_n|\theta)\right) \\ &= \arg\max\left(\sum_{i=1}^{n} x_i \log(\theta) + \sum_{i=1}^{n}(1-x_i)log(1-\theta)\right) \\ &= \arg\max\left((\alpha-1)log(\theta) + (\beta-1)\log(1-\theta) + \sum_{i=1}^{n}(1-x_i)log(1-\theta)\right)\end{aligned}$$

and again by setting the derivation to zero we get:

$$\sum_{i=1}^{n}\frac{x_i}{\theta} - \sum_{i=1}^{n}\frac{1-x_i}{1-\theta} = 0 \Longrightarrow \hat{\theta}_{MLE} = \frac{S_n}{n}$$

Thus we get that:

$$\hat{\theta}_{Bayes} = \frac{S_n + \alpha - 1}{\alpha + \beta + n - 2} = \frac{n}{\alpha + \beta + n - 2}\hat{\theta}_{MLE} + \frac{\alpha - 1}{\alpha + \beta + n - 2} \tag{8}$$

As expected $\hat{\theta}_{Bayes}$ converges to $\hat{\theta}_{MLE}$ as n goes to infinity. On the other hand When $n = 0$, thus we have no sample the Bayes estimator is simply the mode of Beta distribution. Finally when $\alpha = \beta = 1$ Beta distribution becomes uniform and thus we get that the estimators are identical.

(b)

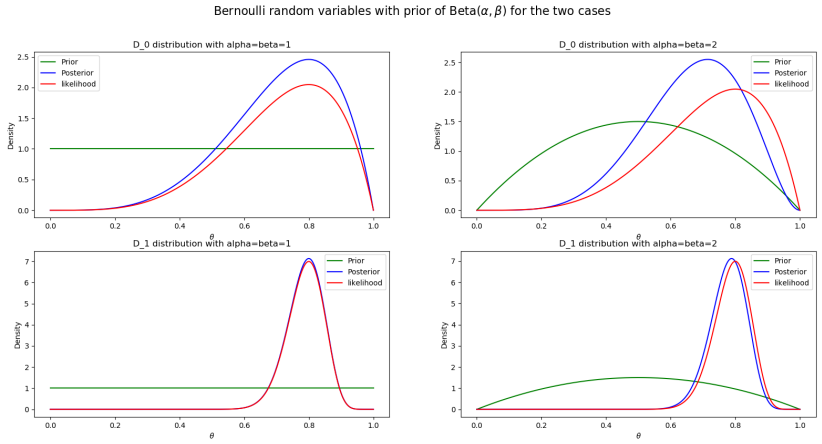

Figure 5: Effect of sample size on the posterior distribution

As can be seen from plots for small sample size the posterior affected from both the likelihood and the prior distributions. As the sample size increase the posterior follow closely to the likelihood.