CS 747: Assignment 1:

Soham Rahul Inamdar(210100149)

September 10, 2023

Objective of the assignment

The objective of this programming assignment was to implement regret minimization algorithms like UCB, KL-UCB and Thompson Sampling Algorithm for some variations of the **multi-armed bandit** setting.

1 Task 1

Task 1 involved the implementation of the UCB, KL-UCB and Thompson Sampling for regret minimization in the standard multi-armed bandit.

1.1 UCB

I implemented the UCB algorithm for a bandit with n arms according to the following expression:

$$ucb_a^t = \hat{p}_a^t + \sqrt{\frac{2\ln t}{u_a^t}}$$

For the first n horizons, we sample each arm once and initialize the empirical mean for all arms. Following this, we choose an arm a such that ucb_a^t is the maximum. u_a^t denotes the number of times arm a has been sampled at time t.

We obtain the following regret vs. horizon plot for the above algorithm:

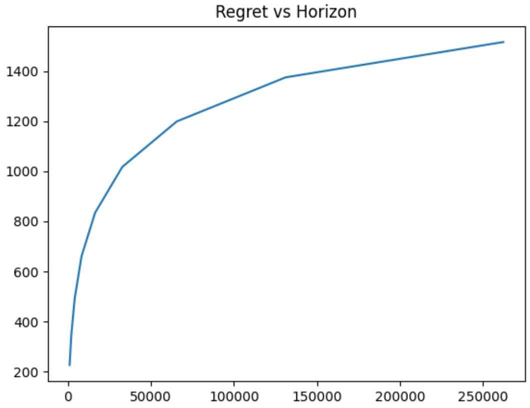

Figure 1: Regret vs. Horizon for UCB

The regret follows a logarithmic trend which is as expected.

1.2 KL-UCB

The KL-UCB algorithm provides us with a tighter confidence bound as compared to the UCB algorithm. We define the following upper bound for this algorithm:

$$ucb - kl_a^t = \max(q \in [\hat{p}_a^t, 1] \text{ s.t. } u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + c\ln(\ln(t)) \tag{1}$$

The KL divergence is defined in the following manner:

$$KL(x, y) = x\ln(\frac{x}{y}) + (1-x)\ln(\frac{1-x}{1-y})$$

where $x, y \in [0, 1)$.
So, essentially we have to find a value q such that

$$KL(\hat{p}_a^t, q) = \frac{\ln(t) + c\ln(\ln(t))}{u_a^t}$$

Here, we can choose the parameter c according to our requirements. We were instructed to choose the value of c as 0.
The required value of q can be calculated using a binary search on the interval $[\hat{p}_a^t, 1]$. This will also ensure that the value we are getting is the maximum value corresponding to equation 1.
So, we calculate q for every arm at an instant of time and sample the arm for which it is maximum.
We obtain the following regret vs. horizon plot for the above algorithm:

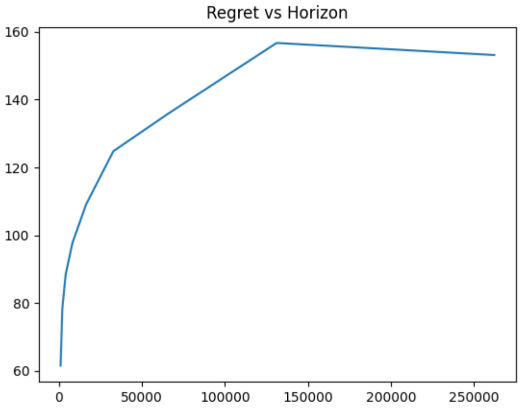

Figure 2: Regret vs Horizon for KL-UCB

The regret follows a logarithmic trend which is as expected. Also, if we compare the plot to Figure 1 we notice that the bound on the regret is much tighter as compared to UCB.

1.3 Thompson Sampling

The Thompson Sampling algorithm makes use of the **Beta Distribution**. The Beta distribution is given by:

$$Beta(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

For each arm we compute

$$x_a^t = Beta(s_a^t + 1, f_a^t + 1)$$

where s_a^t is the number of successful pulls of arm a and f_a^t is the number of failures. Now, we select the arm with the maximum value of x_a^t. We obtain the following regret vs. horizon plot for the above algorithm:

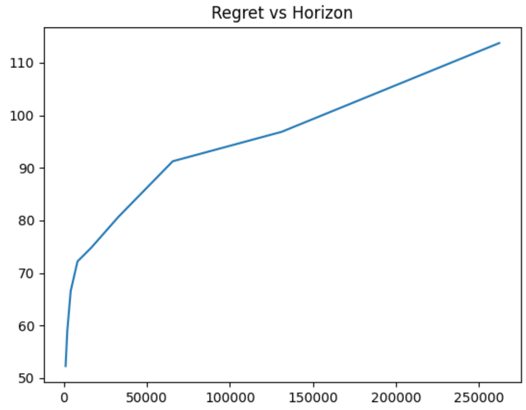

Figure 3: Regret vs Horizon for Thompson Sampling

2 Task 2

Task 2 involved studying the variation of regret with the means of the arms for the UCB and KL-UCB algorithms.

2.1 Part A

In Part A, we had to keep the mean of the first arm constant $p_1 = 0.9$ and vary that of the second arm from 0 to 0.9 in steps of 0.05. The following plot was obtained for the variation of regrets:

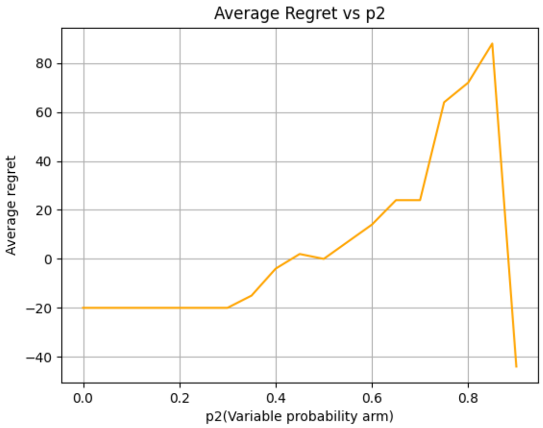

Figure 4: Average Regret vs p_2

As we can see in the plot, the regret increases with p_2 till p_2 reaches the same value as p_1(the optimal arm). This is in accordance with the equation we had studied in class:

$$R_T = \mathcal{O}\left(\sum_{a:p_a \neq p_a^*} \frac{1}{p_a - p_a^*} \log(T)\right)$$

Hence, as $(p_a - p_a^*)$ decreases, the regret accumulated increases. But, when both arms are symmetric, sampling **any one** would produce the same reward and hence, $\mathbb{E}[r_t] = T$ which implies that regret will be **minimal**.

2.2 Part B

In Part B, we had to vary the means of the 2 arms in steps of 0.05 while keeping the difference between them constant(0.1). We obtain the following plots for variation of regret with the smaller mean p_2:

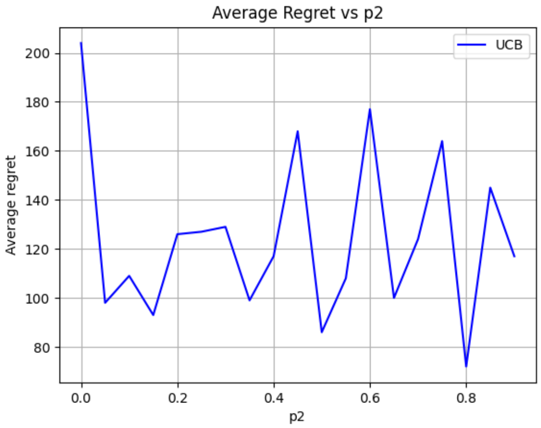

Figure 5: Average Regret vs p_2 for UCB

In the above plot, we can observe that the regret for UCB is lower for values of p_2 less than 0.45 after which it starts to oscillate. This oscillation is attributed to the time-dependent exploration term.

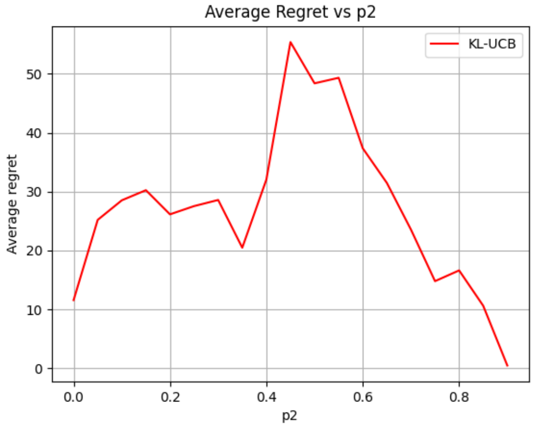

Figure 6: Average Regret vs p_2 for KL-UCB

In the above plot, we see maximum regret at $p_2 = 0.45$ and minimum for $p_2 = 0.9$. The minimum regret can be explained by the fact that $p_1 = 1$ when $p_2 = 0.9$ hence after a few horizons regret will be continuously zero.

3 Task 3

Task 3 involved a multi-armed bandit setting where we have a finite probability of getting a faulty output when we sample one of the arms. The objective was to come up with an algorithm to maximize reward in such a setting. So for an arm with empirical mean p_a, because of fault we will have the new empirical mean

$$p_a^n = \frac{f}{2} + (1 - f)p_a$$

where f is the probability of fault.
Now this is a linear transformation on the original empirical mean and $p_a^n \in (0, 1)$ hence, the **Thompson Sampling** method should work well. Further, the Thomson Sampling method is dependent on the success and failure of each arm and does not depend directly on the empirical means. Hence, I have followed the Thompson sampling method without any modifications.

4 Task 4

In Task 4, we had to deal with 2 bandit instances with equal number of arms at once. According to the problem, we can only decide which arm we want to sample, the bandit instance is chosen uniformly at random and the reward is returned for the chosen instance. The objective was to come up with an algorithm to maximize reward in such a setting. Here, the expected reward at one horizon is:

$$r = \frac{(p_a + p'_a)}{2}$$

This is again a linear transformation on the original means. So, we again use the Thomson Sampling method as it is dependent on the success and failure of each arm and does not depend directly on the empirical means.