Stat 548 HW4

Jiahao Yang

June 3, 2018

Problem 0 Certify you read the HW Policies

0.1 List of Collaborators

None

0.2 List of Acknowledgements

None

0.3 Certify that you have read the instructions

I have read and understood these policies.

Problem 1 Logarithmic Regret of UCB

1.1

By the Hoeffding's bound and the union bound, with probability greater than $1-\delta$, we have that for all the arms and for all time steps $K \leq t \leq T$, the confidence bound(ConfBound) is $c\sqrt{\frac{log(t/\delta)}{N_{a,t}}}$.

By the construction of the UCB algorithm, we know that

$$\hat{\mu}_{a,t} + ConfBound \geq \hat{\mu}_{\star} + ConfBound \geq \mu_{\star}$$

Thus,

$$\mu_a \geq \hat{\mu}_{a,t} - ConfBound \geq \mu_{\star} - 2ConfBound,$$

$$\mu_{\star} - \mu_a \leq 2ConfBound = 2c\sqrt{\frac{log(t/\delta)}{N_{a,t}}} \leq 2c\sqrt{\frac{log(T/\delta)}{N_{a,t}}},$$

$$\Delta_a^2 \leq 4c^2 \frac{log(T/\delta)}{N_{a,t}},$$

$$N_{a,t} \leq c_3 \frac{log(T/\delta)}{\Delta_a^2}.$$

For each sub-optimal arm a, let t be the last time that arm a is been pulled up. Then we have $N_{a,T} = N_{a,t} \leq c_3 \frac{log(T/\delta)}{\Delta_a^2}$.

1.2

The regret for arm a is $\mu_\star - \mu_a = \Delta_a$. And from section 1.1, we know that the total number of times that any sub-optimal arm a will be pulled up to time T will be bounded by $c_3\frac{log(T/\delta)}{\Delta_a^2}$.

Thus, with probability greater than $1-\delta$, the total observed regret is bounded as following.

$$T\mu_\star - \sum_{t\leq T}\mu_{a_t} \leq c_3 \sum_{a\neq a_\star}(\frac{log(T/\delta)}{\Delta_a^2}\Delta_a) = c_3 \sum_{a\neq a_\star}\frac{log(T/\delta)}{\Delta_a}.$$

1.3

Choose $\delta = \frac{1}{T^2}$. Note that with probability greater than $1-\frac{1}{T^2}$, our regret is bounded by $c_3\sum_{a\neq a_\star}\frac{log(T/\delta)}{\Delta_a}$. Also if we "fail", the largest regret we can pay is T, and this occurs with probability less than $\frac{1}{T^2}$. Thus the expected regret is,

$$\begin{aligned}
T\mu_\star - E[\sum_{t\leq T}X_t] &\leq (1-\frac{1}{T^2})c_3\sum_{a\neq a_\star}\frac{log(T/(1/T^2))}{\Delta_a} + \frac{1}{T^2}T \\
&\leq c_3\sum_{a\neq a_\star}\frac{3log(T)}{\Delta_a} + \frac{1}{T}\frac{K-1}{max(\Delta_a)} \\
&\leq \max(3c_3, 1)\sum_{a\neq a_\star}\frac{log(T)+\frac{1}{T}}{\Delta_a} \\
&\leq \max(3c_3, 1)\sum_{a\neq a_\star}\frac{2log(T)}{\Delta_a} \\
&\leq c_4\sum_{a\neq a_\star}\frac{log(T)}{\Delta_a}
\end{aligned}$$

1.4

Since $c_4\sum_{a\neq a_\star}\frac{log(T)}{\Delta_a} \leq c_4\sum_{a\neq a_\star}\frac{log(T)}{\Delta_{min}} = c_4\frac{(k-1)log(T)}{\Delta_{min}} \leq c\frac{Klog(T)}{\Delta_{min}}$, we have

$$T\mu_\star - E[\sum_{t\leq T}X_t] \leq c\frac{Klog(T)}{\Delta_{min}}.$$

The UCB algorithm is shown below.

At each time t:

1) Pull arm :

$$a_t = argmax(\hat{\mu}_{a,t} + c\sqrt{\frac{log(t/\delta)}{N_{a,t}}}) := argmax(\hat{\mu}_{a,t} + ConfBound_{a,t}),$$

where $c \leq 10$ and $\delta = \frac{1}{T^2}$,

2) Observe reward X_i,

3) Update $\mu_{a,t}, N_{a,t}, ConfBound_{a,t}$

Problem 2 Thompson Sampling

2.1

From the question, we know that

$$\mu_1 = 1/6, \mu_2 = 1/2, \mu_3 = 2/3, \mu_4 = 3/4, \mu_5 = 5/6.$$

And it is obvious that if we keep pull the arm with biggest expected reward, the maximum expected reward we can obtain in T steps is $\frac{5T}{6}$.

2.2

The quantities that the algorithm maintains in memory are shown below.

1) the parameters of distribution $Beta(\alpha_{a,t}, \beta_{a,t})$, $\alpha_{a,t}$ and $\beta_{a,t}$.

2) the sample from posterior distribution $Beta(\alpha_{a,t}, \beta_{a,t})$, $\hat{\mu}_{a,t}$.

3) the observed reward, X_t.

4) the number of times we pulled arm a up to time t, $N_{a,t}$.

The updates for the posterior distributions is,

$$Beta(\alpha_{a,t+1}, \beta_{a,t+1}) = \begin{cases} Beta(\alpha_{a,t}, \beta_{a,t}), & a_t \neq a \\ Beta(\alpha_{a,t} + X_t, \beta_{a,t} + 1 - X_t), & a_t = a \end{cases} \tag{1}$$

2.3

The plot is shown below.

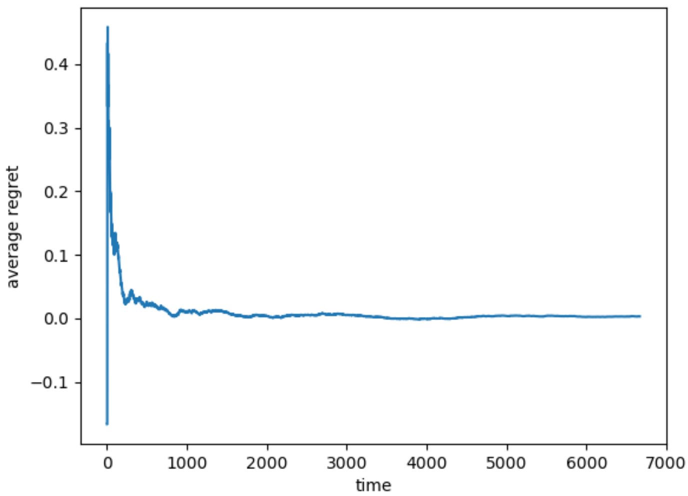

Figure 1: average regret vs time

2.4

The results are shown below.

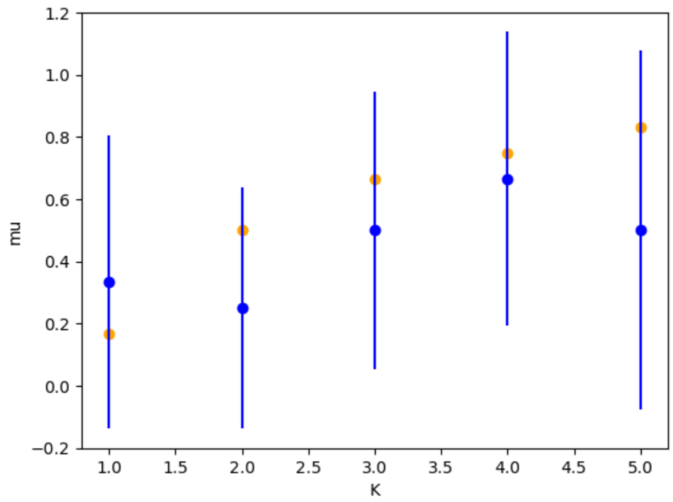

Figure 2: mu vs arm when T = 6

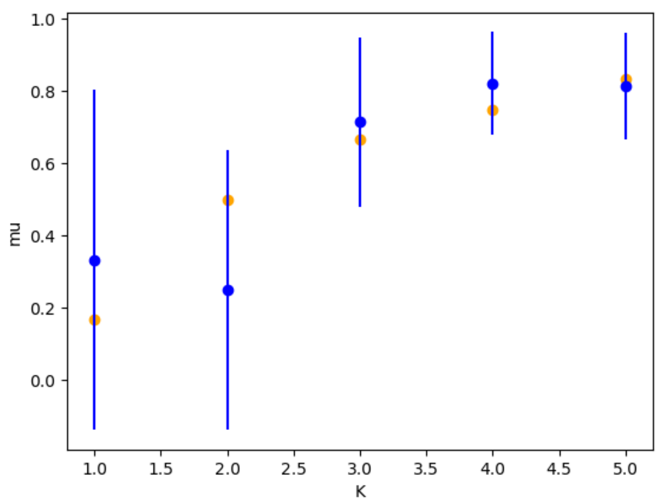

Figure 3: mu vs arm when T = 66

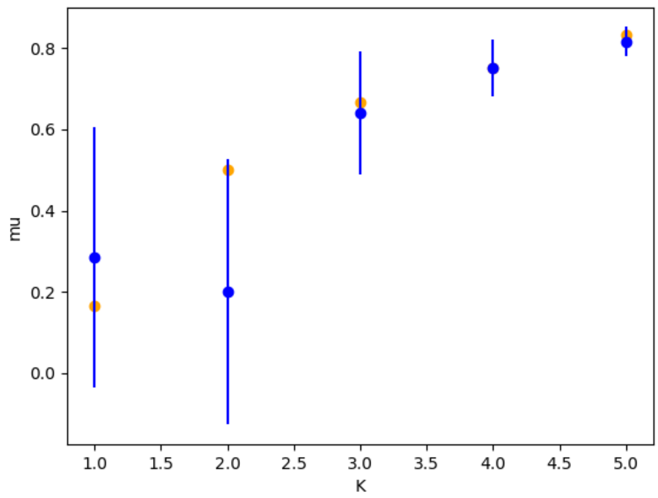

Figure 4: mu vs arm when T = 666

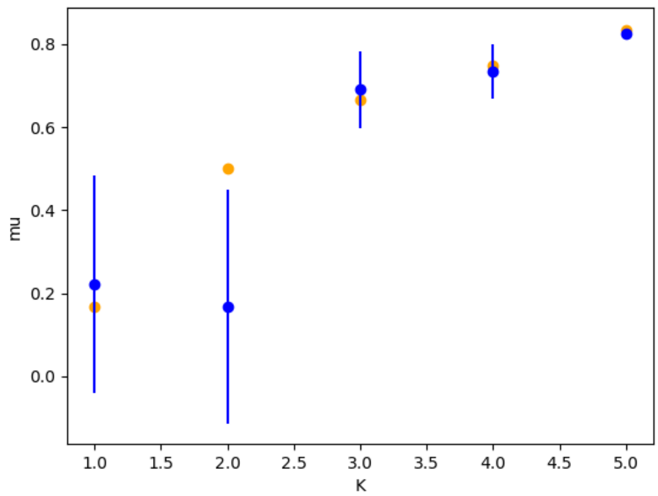

Figure 5: mu vs arm when T = 6666

2.5

The result is shown below.

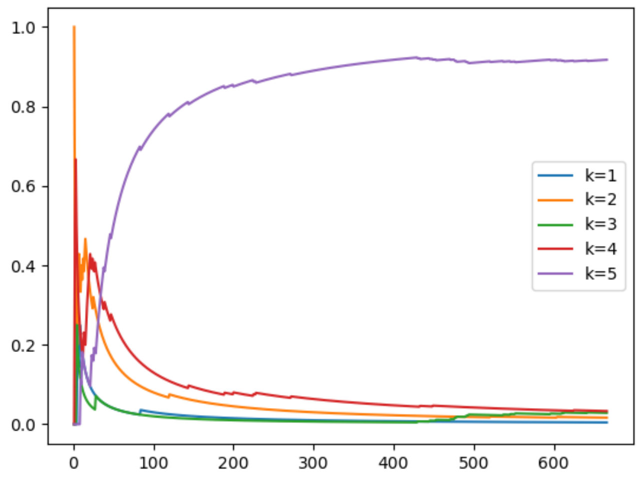

Figure 6: mu vs arm when T = 6666

2.6

From screen print of my code when I choose $T = 6666$ is

```
the first time achieve 0.95 and stays 10 steps: 3529
```

Figure 7: screen print when T = 6666

The first time achieve 0.95 and stays 10 steps is 3529.