CS6046: Multi-Armed Bandits : Assignment 3

Pragalbh Vashishtha mm19b012

June 2021

Contents

1 Introduction

We shall look at the problem setting:- The two arms are standard Gaussian random variables with meanμ_1= 0 and μ_2= Δ The following algorithms will be implemented in the setting:-

- :Gaussian Explore Then Commit
- Thompson Sampling
- UCB: Upper Confidence Bound

These algorithms are slightly modified to suit the setting. They still retain the spirit of the original algorithm. Also, Variance can also be changed for the arms.

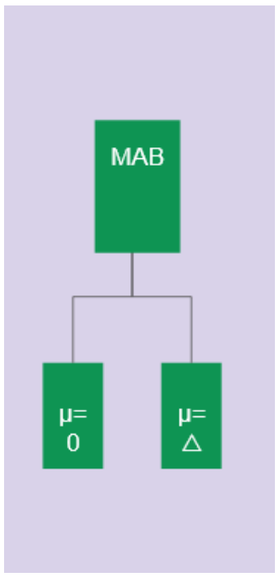


Figure 1: The Problem Setting

2 Gaussian Explore Then Commit

2.1 Insights from plot

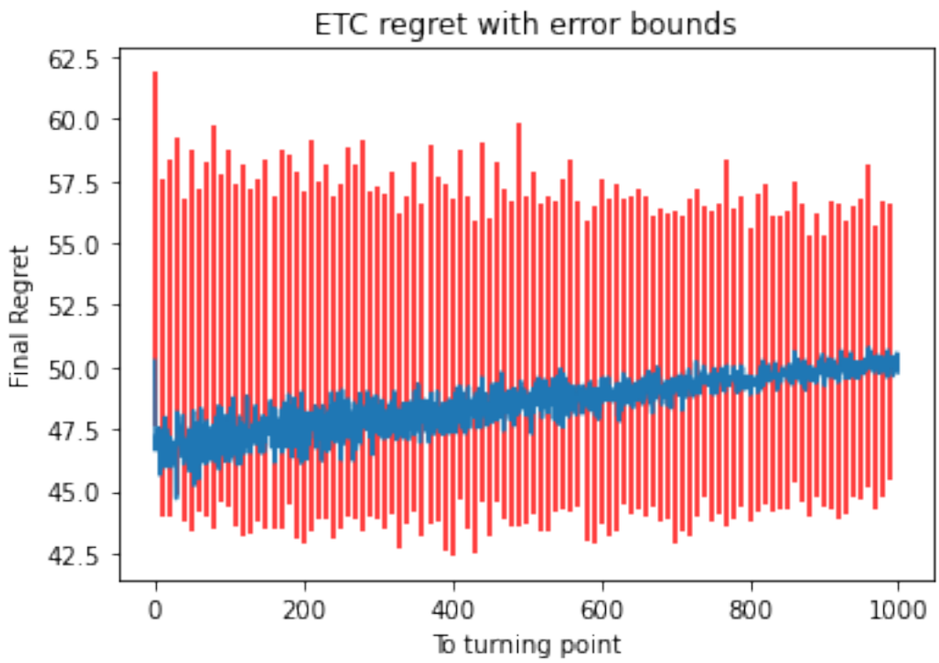


Figure 2: Explore then Commit:- over 10,000 rounds

The error bars here are max value We note that the error bars are much more than the gradation within the mean of the regret graph. Note that although trends may be observed in the regret graph, the error bounds are high enough to offset solid predictions based on said trends. However note that due to the large sample size (10,000 per value of T_0), most of the effect of the high errors are offset.

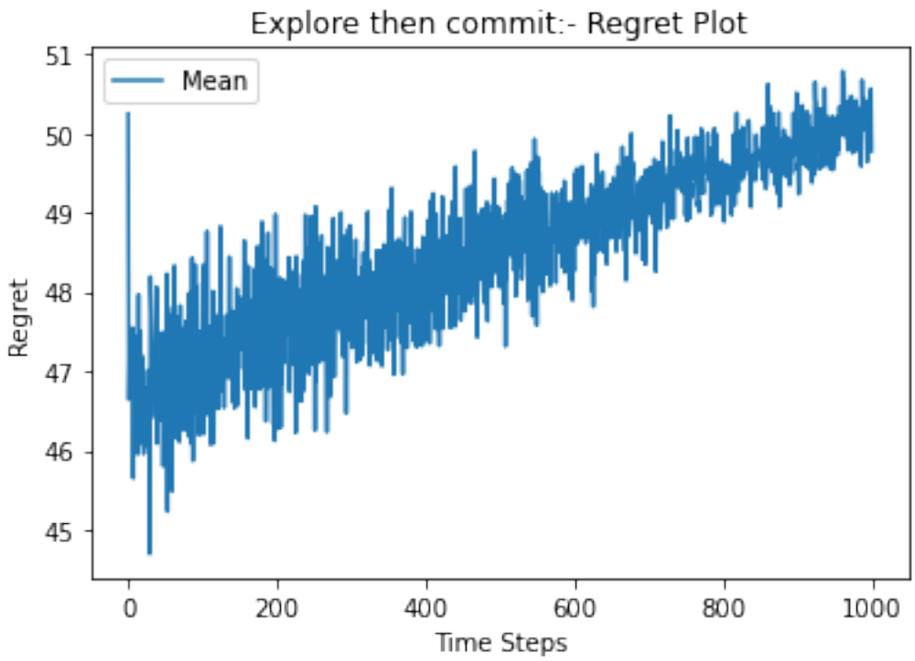


Figure 3: Explore then Commit:- over 10,000 rounds

So here we observe that as the T_0 value increases, The Regret also increases. Note, that the optimal shape is a U-shape, lowest at T_0=367. This is offset by the high error.

2.2 Theoretical bound and proof for optimal T_o

Explore then commit strategy is characterised by T_o, or alternately, by m, where $mK = T_o$ where K is the number of arms.
Thus the algorithm will explore for mK rounds before choosing a single action for the remaining T–mK rounds.

For the sake of the proof, let us consider K arms with rewards Δ_i for arm i. We can write this strategy formally as

$$A_t = \begin{cases} i & if(tmodK) + 1 = i \quad and \quad t \leq mK; \\ argmin_i \quad \hat{\mu}_i x & t > mk \end{cases} \tag{1}$$

where ties in the will be broken in a fixed arbitrary way and $\hat{\mu_i}$ is the average pay-off for arm i up to round t.
Regret can be written as:

$$R_n = \sum_{t=1}^{T} \mathbb{E}[\Delta_{A_t}]. \tag{2}$$

In the first mK rounds ETC is completely deterministic, choosing each action exactly m times. Subsequently ETC chooses a single action that gave the largest average payoff during exploring. Therefore, by splitting the sum we have:

$$R_n = m \sum_{t=1}^{K} \Delta_i + (n - mK) \sum_{t=1}^{K} \Delta_i \mathbb{P}(i = argmax_j \hat{\mu_j}(mK)). \tag{3}$$

We, assume without loss of generality that the optimal arm is i=1 $\mu_1 = max_i \mu_i$. Then,

$$\mathbb{P}(\text{i} =_{\text{j}} \hat{\mu}_{\text{j}}(\text{mK})) \leq \mathbb{P}(\hat{\mu}_{\text{i}}(\text{mK}) – \hat{\mu}_1(\text{mK}) \geq 0) \tag{4}$$

$$= \mathbb{P}(\hat{\mu}_{\text{i}}(\text{mK}) – \mu_{\text{i}} – \hat{\mu}_1(\text{mK}) + \mu_1 \geq \Delta_{\text{i}}) \,. \tag{5}$$

Due to Gaussian Nature of the arms:-

$$\mathbb{P}(\hat{\mu}_{\text{i}}(\text{mK}) – \mu_{\text{i}} – \hat{\mu}_1(\text{mK}) + \mu_1 \geq \Delta_{\text{i}}) \leq \exp\left(-\frac{\text{m}\Delta_{\text{i}}^2}{4}\right) \tag{6}$$

and by straightforward substitution we obtain

$$R_n \leq m \sum_{i=1}^{K} \Delta_i + (n – mK) \sum_{i=1}^{K} \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right) . \tag{7}$$

Choice of m:- If we limit ourselves to K = 2, as in the question, then $\Delta_1 = 0$ and by using $\Delta = \Delta_2$ the above display simplifies to

$$R_n \leq m\Delta + (n – 2m)\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right) . \tag{8}$$

Provided that n is reasonably large this quantity is minimized by

$$m = \frac{4}{\Delta^2} \log\left(\frac{n\Delta^2}{4}\right) \tag{9}$$

and for this choice the regret is bounded by

$$R_n \leq \Delta + \frac{4}{\Delta}\left(1 + \log\left(\frac{n\Delta^2}{4}\right)\right) . \tag{10}$$

Hence we have got our regret bound!

2.3 Gaussian arms with different variance

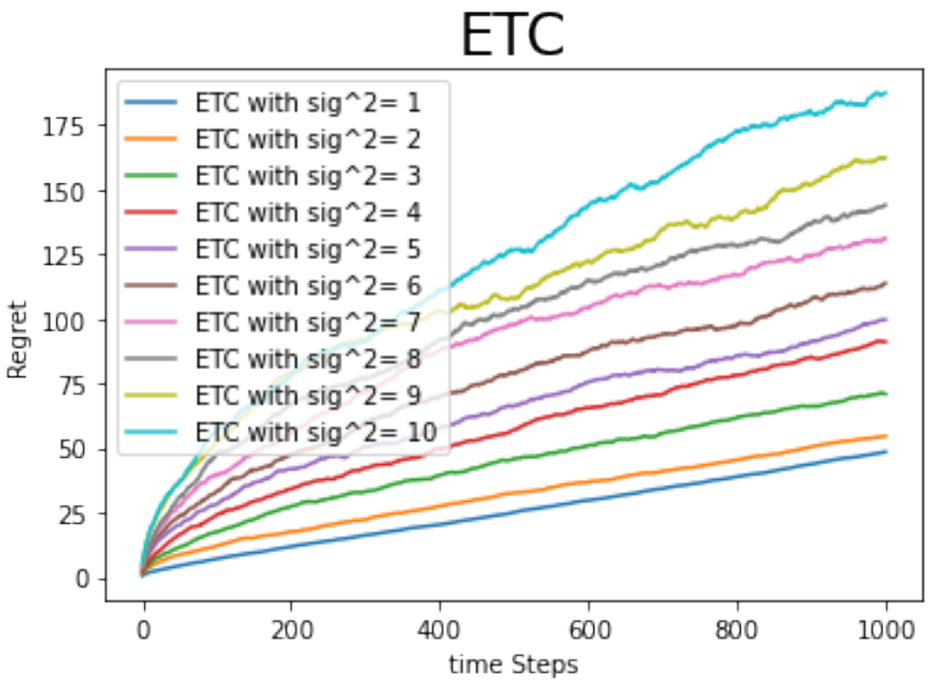


Figure 4: ETC with Varying σ^2

We note that ETC performs best with lower σ^2 values for the same T_o. The trend is as follows:- with increasing variance, an increased regret bound was observed. This has to do with the probability that the arm selection during the learning phase is the worse arm. At high variance, an arm can perform better in the short run, even while being bad in the long run. ETC will stick to that Arm, hence the Regret explodes

3 Thompson Vs UCB

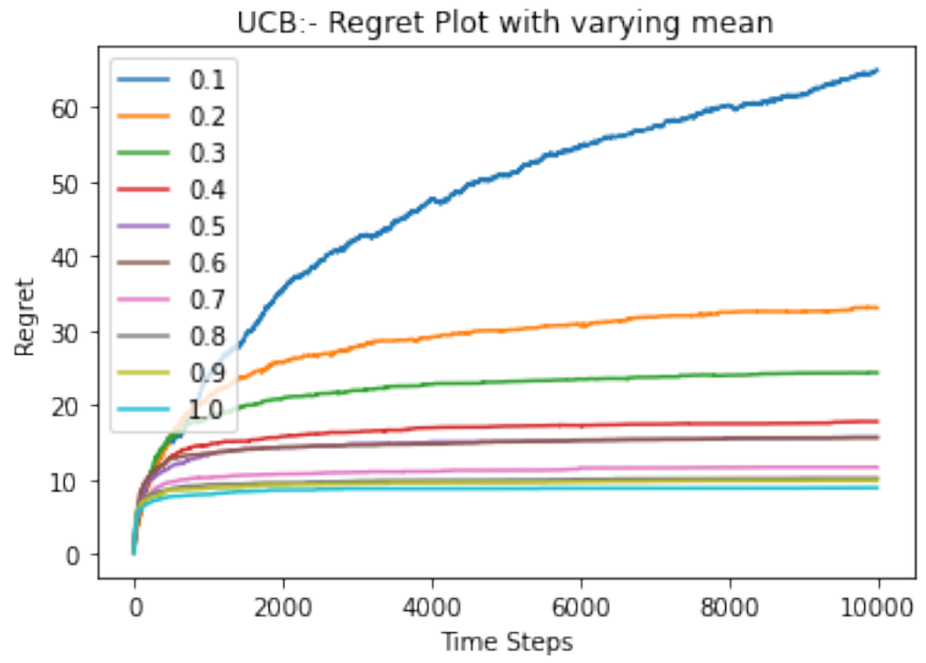


Figure 5: UCB with varying Means, averaged over 100 iterations

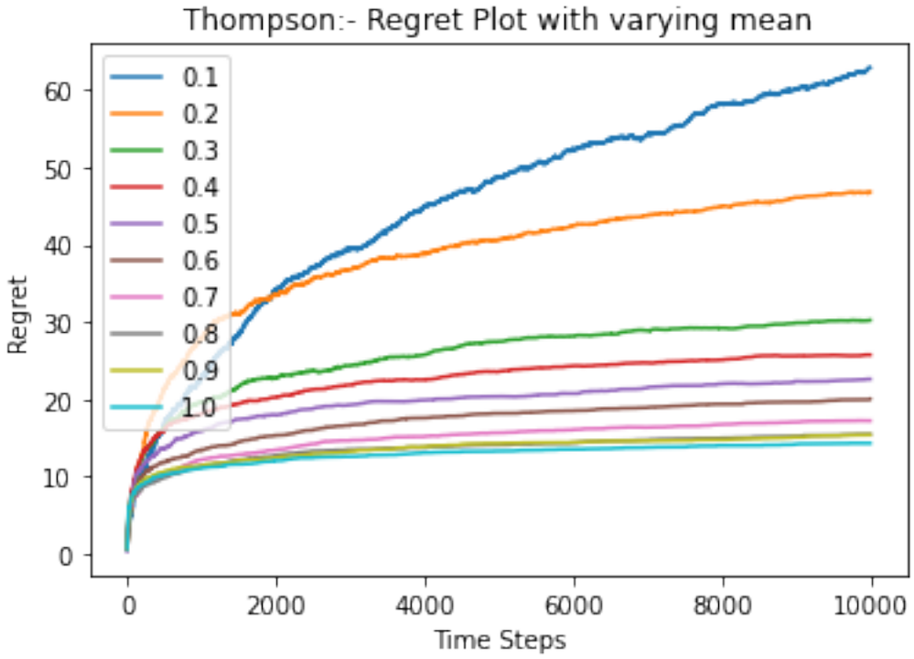
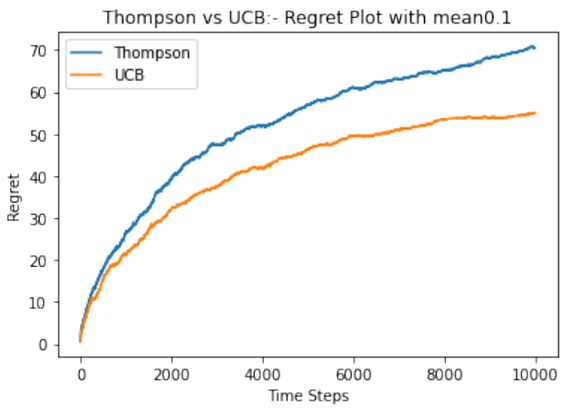
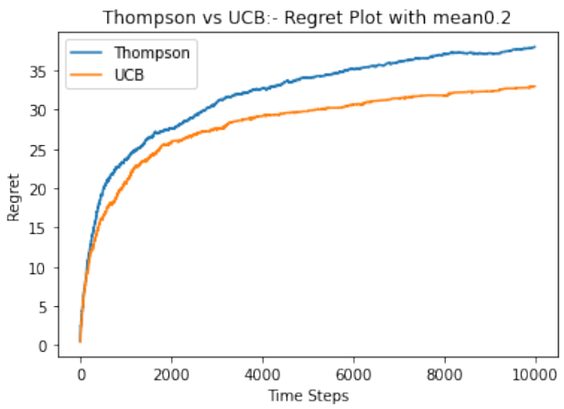


Figure 6: Thompson Sampling with varying Means,averaged over 100 iterations

Both UCB and Thompson show a logarithmic regret bound, as expected. With increasing Mean, the algorithms UCB and Thompson perform better as it is easier to distinguish. Let us now compare the UCB algorithm with the Thompson Sampling Algorithm

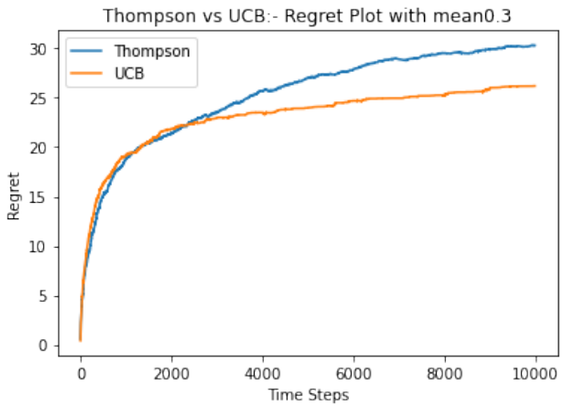


(a) Thompson vs UCB @Δ =0.1

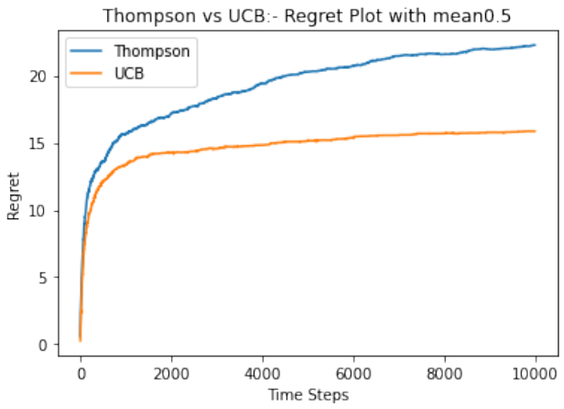


(b) Thompson vs UCB @Δ =0.2

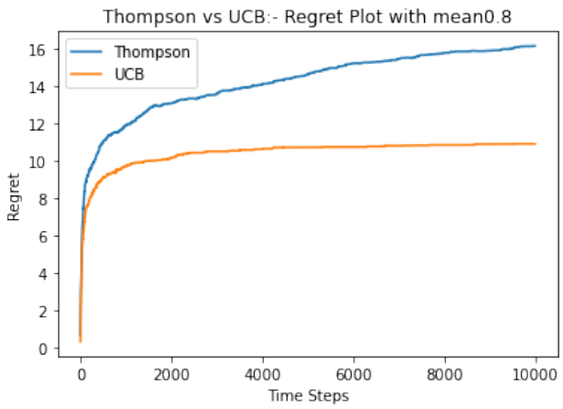
Figure 7: Thompson vs UCB



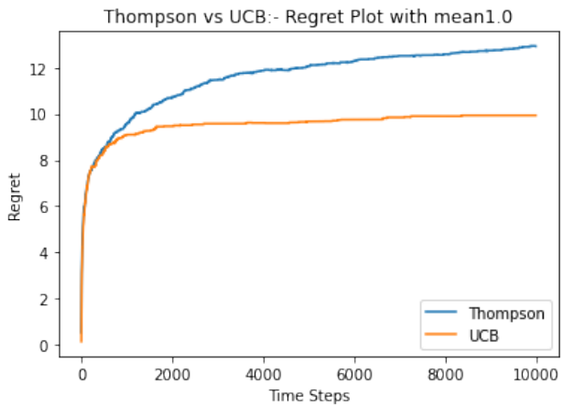
(a) Thompson vs UCB @Δ =0.1



(b) Thompson vs UCB



(a) Thompson vs UCB



(b) Thompson vs UCB

Here we note that UCB outperforms Thompson, albiet marginally. The Regret of the curves decrease with increasing mean, as expected. This is because it is easier for the arms to be distinguished if the mean payoff of arms differ greatly.

4 Conclusion

- The empirical behaviour of the Explore then Commit algorithm was observed over 10,000 iterations for all values of T_0.
- The effect of variance was observed in the ETC algorithm
- Behaviour of the UCB and Thompson Algorithm were shown
- Behaviour ofchange in Δ was observed on the Thopmson and UCB algorithms

5 Refernces

- CS6046 material
- https://papers.nips.cc/paper/2016/file/ef575e8837d065a1683c022d2077d342-Paper.pdf
- https://banditalgs.com/2016/09/14/first-steps-explore-then-commit/

Thank You