CS6046 Multi-Armed Bandits: Assignment 3 Report

-Akash Reddy A, EE17B001

1 Question 1: Gaussian Explore Then Commit

Pseudocode

Algorithm 1: Explore Then Commit (ETC)

```
Choose $T_0$ rounds for exploration. For 2 arms, set $T_0 = \lceil \frac{4}{\Delta^2}\log(\frac{T\Delta^2}{4})\rceil$ for best regret bound.
Play each arm $i \in [d]$ for $\frac{T_0}{d}$ rounds (exploration)
Estimate $\hat{\mu}_i$ for each arm $i$ as $\hat{\mu}_i = \frac{\sum_{t=1}^{T_0} l_i^t \cdot I(i_t=i)}{\sum_{t=1}^{T_0} I(i_t=i)}$.
For remaining $T - T_0$ rounds, play best arm $\hat{i} = \arg\min_i \hat{\mu}_i$ according to estimated losses (exploitation).
```

1.1 T_0 that Minimises Regret

In the regret analysis of the ETC algorithm, we obtain:

$$R_T \leq \frac{T_0}{d}\Delta + (T - T_0)\Delta e^{-\frac{T_0\Delta^2}{2d}}$$

By substituting $d = 2$, differentiating this bound w.r.t T_0 and equating it to 0, we obtain the analytical value of T_0 that minimises regret.

$$\begin{aligned}\frac{d}{dT_0}(\frac{T_0}{2}\Delta + (T - T_0)\Delta e^{-\frac{T_0\Delta^2}{4}}) &\approx \frac{d}{dT_0}(\frac{T_0}{2}\Delta + T\Delta e^{-\frac{T_0\Delta^2}{4}}) \\ &= \frac{\Delta}{2} + \frac{-\Delta^2}{4}T\Delta e^{-\frac{T_0\Delta^2}{4}} \\ &= \frac{\Delta}{2} + \frac{-T\Delta^3}{4}e^{-\frac{T_0\Delta^2}{4}} \\ &= 0\end{aligned}$$

Thus,

$$\begin{aligned}\frac{T\Delta^3}{4}e^{-\frac{T_0\Delta^2}{4}} &= \frac{\Delta}{2} \\ e^{-\frac{T_0\Delta^2}{4}} &= \frac{4}{T\Delta^2} \\ T_0 &= \frac{4}{\Delta^2}\log\frac{T\Delta^2}{4}\end{aligned}$$

which is about 367 in this case.

1.2 Regret vs T_0 Plot for $\sigma^2 = 1$

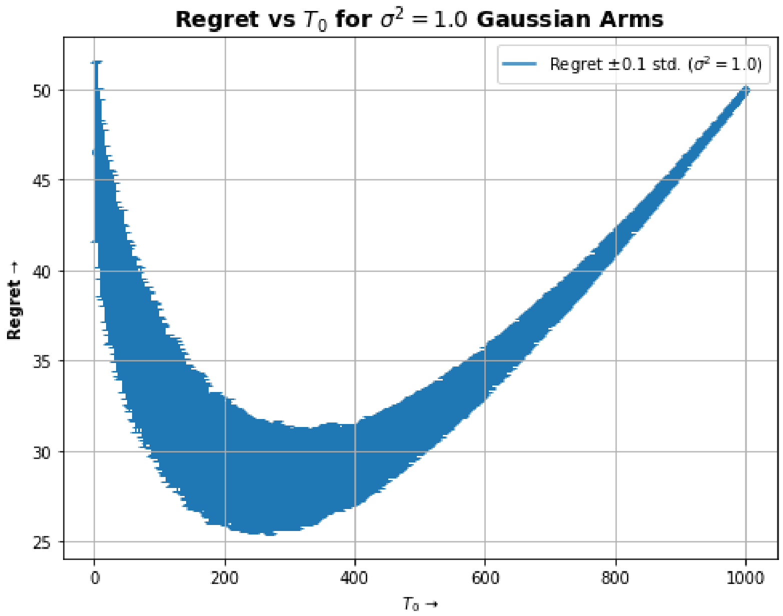

Figure 1: Regret vs T_0 Plot for ETC, with errorbars at *each* value of T_0 as asked over 10000 runs.

The "expectation definition" of regret defined for stochastic bandits in the class as:

$$\begin{aligned}
R_T &= \mathbb{E}[\sum_{t=1}^{T} l_{i_t}^t] - \min_i \mathbb{E}[\sum_{t=1}^{T} l_i^t] \\
&= \mathbb{E}[\sum_{t=1}^{T} l_{i_t}^t] - \sum_{t=1}^{T} \mu^* \\
&= \mathbb{E}[\sum_{t=1}^{T} \sum_{i=1}^{d} (\mu_i - \mu^*) \cdot 1(i_t = i)] \\
&= \sum_{i=1}^{d} \Delta_i \mathbb{E}[n_t(i)] \\
&= \Delta \mathbb{E}[n_t(2)]
\end{aligned}$$

where the last equation is for the 2-arm case and arm 1 is assumed to be optimal, is adopted for this assignment on stochastic bandits. This "expectation definition" of regret is evaluated as $\Delta n_t(2)$ at each timestep t in each run, and the randomness across runs (which gives errorbars) is captured by the variance in the random variable $n_t(2)$. This way of estimating regret gives satisfactory results for all the algorithms, therefore I have stuck with it.

- The value of T_0 minimising the regret is 280-300 on different runs of the program. This is close to our theoretical T_0 value from the first part of the question. Around this point, enough exploration has been done such that the better arm is picked with high probability and no regret is accumulated thereof. For higher values of T_0, despite finding the better arm, the continued exploration increases the regret. (It is not clearly understood what the theoretical "bound" asked for in this question is. No mention of the bounds on cumulative regret over time has been made.)
- The variance of the regret is highest at smaller T_0 values. This is because we have not explored enough, and therefore, either extreme can happen with significant probability - either the good arm gets picked (in the short "exploration" phase) and no regret occurs, or the bad arm gets picked and we get near-maximum regret in that run. Therefore, this region of the curve has very high variance.
- On the other hand, for higher values of T_0, there is more than enough exploration done in order to understand which arm is better (past the point of minimum regret). Therefore, in each run, the variance

in picking of the exploited arm, is reduced (as sufficient exploration has been done). The "additional" exploration beyond the required point incurs regret in pretty much the same way in every run (low variance). Also the remaining timesteps ($T - T_0$) are not as many as for earlier values of T_0, therefore the spread/variance of the accrued extra regret by playing one arm in the exploitation phase is reduced. The regret from the long exploration phase is similar for all runs, and therefore the total regret has lower variance.

1.3 Using Gaussians with Different Variances

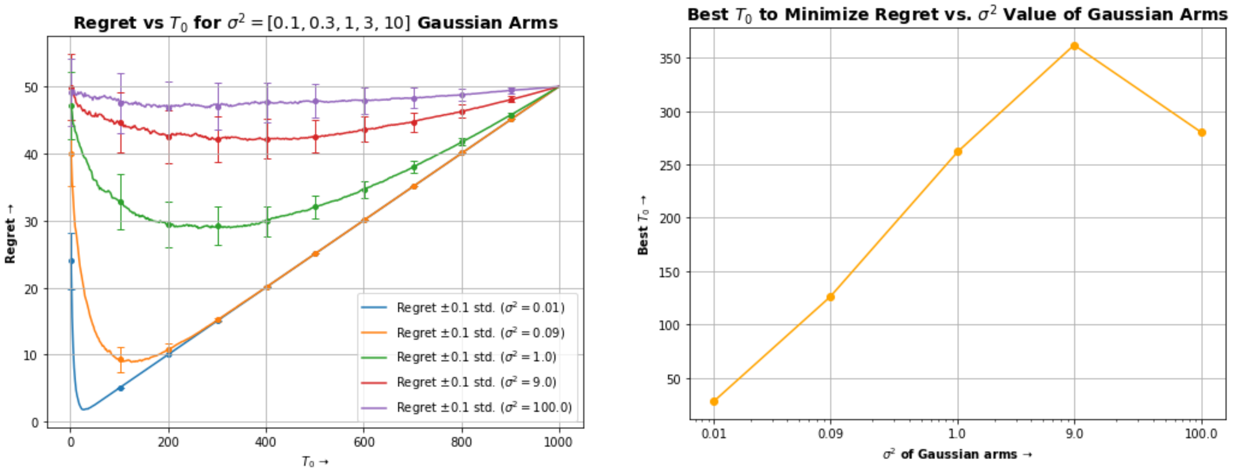

Figure 2: Regret vs T_0 Plot for different σ^2 values, and the corresponding T_0 that minimises regret

The regret analysis that was done in subsection 1.1 assumes that the reward distributions are 1-subgaussian (source). A standard Gaussian distribution is 1-subgaussian, therefore the result held valid. However, when we change the variance of the Gaussian to a σ^2, it becomes σ-subgaussian. In this case the analysis changes as follows:

$$R_T \leq \frac{T_0}{d}\Delta + (T - T_0)\Delta e^{-\frac{T_0\Delta^2}{2d\sigma^2}}$$

Minimising this RHS by differentiating and setting to 0, we get a value of T_0 that is different from the one in 1.1 by a few factors:

$$T_0 = \frac{4\sigma^2}{\Delta^2}\log(\frac{T\Delta^2}{4\sigma^2})$$

This value of T_0 increases with σ^2, and this is what we see in our experiment.

- For values of $\sigma^2 \leq 1$, we see that the value of T_0 is smaller. Also, the same trend that we observed in the $\sigma^2 = 1$ case appears here - the variance in regret is very high until sufficient exploration has been done, as the exploitation could go to either extreme. After sufficient exploration, all the extra exploration tends to increase regret in the same gradual way in all runs (since the exploration is done by picking left and right arms $T_0/2, T_0/2$ times respectively).

- For values of $\sigma^2 \geq 1$, we see that the value of T_0 is larger. The regret vs T_0 curve becomes less skewed. At very high values of σ^2, such as for $\sigma^2 = 100$, the variance in the arm distributions is so large that, by the time sufficient exploration has been done to pick out the best arm with good enough probability, the regret has already built up. Therefore, the regret starts to look very similar across all values of T_0, and randomness factors start determining the best experimental value of T_0. That is what has caused the sudden drop in best T_0 value for $\sigma^2 = 100$ above.

2 Question 2: Thompson vs UCB

Pseudocodes

Algorithm 2: Upper Confidence Bound (UCB)

```
Say there are $d$ arms and UCB is run for $T$ total timesteps
initialise $n_i^0 = 0 \ \forall i \in [d]$ at $t = 0$ (number of picks of arm $i$ until time $t$)
Play each arm once and collect rewards, set them as $\hat{\mu}_i^d \ \forall i \in [d]$ at the end of the $dth$ timestep
for $t = d, d+1, ...T$ do
    $UCB_t(i) = \hat{\mu}_i^{t-1} + \sqrt{\frac{2\log(t)}{n_i^{t-1}}}$
    $i_t = \arg\max_i UCB_t(i)$
    $n_{i_t}^t = n_{i_t}^{t-1} + 1$
    $n_i^t = n_i^{t-1} \ \forall i \neq i_t$
    Play $i_t$ and receive some reward $r_t$
    Update $\hat{\mu}_i^t$ as average of all rewards obtained by pulling $i_t$
endfor
```

Algorithm 3: Thompson Sampling

```
Say there are $d = 2$ arms and Thompson Sampling is run for $T$ total timesteps
initialise parameters of prior used to estimate the parameters of the reward distribution $\forall i \in d$ at $t = 0$
for $t = 1, 2, ...T$ do
    Sample $\theta_i(t) \sim prior_i$ independently $\forall i \in [d]$
    $i_t = \arg\max_i \theta_i(t)$
    Get observation/reward $l_i$ for arm $i$ and calculate posterior parameters using $l_i$ and prior parameters
    Set posterior parameters = new prior parameters
endfor
```

2.1 Adaptations for 2-Armed Standard Gaussian Bandit

1. Both algorithms have been described in the reward setting above. Therefore, they should be changed to the loss setting.

   The UCB update rule is currently $i_t = \arg\max_i(\hat{\mu}_i^{t-1} + \sqrt{\frac{2\log(t)}{n_i^{t-1}}})$. It is changed to an "LCB" update rule according to the lower confidence bound of the loss, as: $i_t = \arg\min_i(\hat{\mu}_i^{t-1} - \sqrt{\frac{2\log(t)}{n_i^{t-1}}})$. We wish to find the arm with the minimum loss and minimise cumulative regret.

   Similarly, in Thompson Sampling the priors model the parameters of the *loss* instead of the *reward* distributions. Therefore, the rule by which we pick an arm is changed to: $i_t = \arg\min_i \theta_i(t)$, to try to accumulate minimum loss instead of maximum reward.

2. The Thompson Sampling here is done on Gaussian likelihood functions (or true loss distributions). In Thompson Sampling, it is helpful to use a "conjugate prior" of the likelihood distribution, which produces a posterior distribution of the same family when updated.

   The commonly studied version of Thompson Sampling is for a Bernoulli 2-armed bandit. A beta prior is a suitable "conjugate prior" for this Bernoulli likelihood function. **However, the likelihoods here are Gaussian. Therefore, we will use Gaussian priors to perform Thompson Sampling.** This is a valid conjugate prior only to estimate the **mean** of the likelihood distribution when the variance is known. **Therefore, in our setup, we assume that the variance** (= 1) **of the standard Gaussian arms is known to the algorithm.**

Now, we will speak in terms of precision τ to describe the posterior update rule. Precision $= \tau = 1/\sigma^2$ =1/variance.

In order to estimate the mean loss μ of the likelihood function with known precision τ, of any arm i which was picked at a certain timestep t and a loss l_i was received, we use Gaussian priors with known parameters $\mu_0^{t-1}, \tau_0^{t-1}$. The posterior parameters are updated as (source):

$$\mu_0^t = \frac{\tau_0^{t-1}\mu_0^{t-1} + l_t}{\tau_0^{t-1} + \tau}$$
$$\tau_0^t = \tau_0^{t-1} + \tau$$

These become the new priors for the next timestep. With time, μ_0^t gets closer to the true mean μ_i of the arm, and the variance of the Gaussian prior/posterior that models the mean gets smaller and smaller (precision is increasing each round). Therefore, the Gaussian prior/posterior tends to center around the mean of the arm.

2.2 Regret Plots for $\Delta = 0.1, T = 10000$

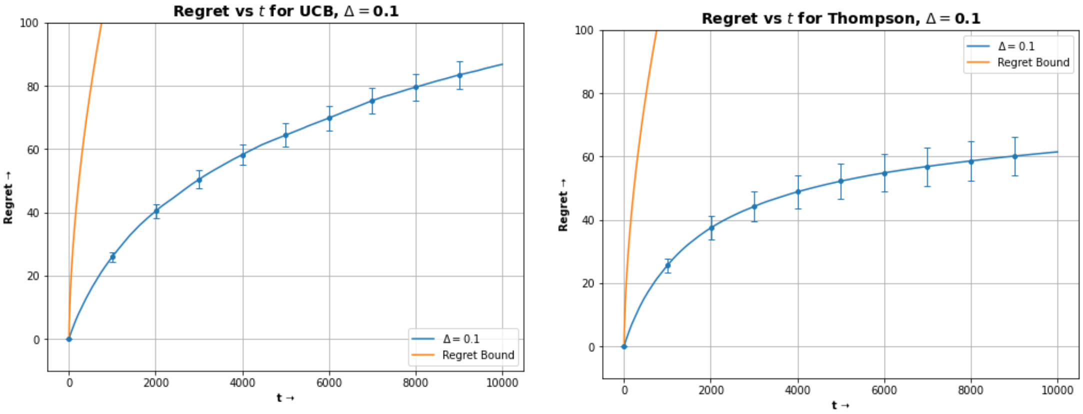

Figure 3: Regret curves and bounds for UCB and Thompson Sampling, $\Delta = 0.1, T = 10000$

- UCB and Thompson Sampling are both known to have the same theoretical optimal regret bound $O(\sqrt{Td\log d})$, and both the regret curves are are shown to be well below the regret bound.
- However, Thompson Sampling accrues lesser regret in the $T = 10000$ window. This is expected, as it has been shown in empirical comparisons of UCB and Thompson Sampling that the latter tends to accrue lesser regret in finite time horizons and may have better constant dependencies. However, both algorithms tend to even out as $T \longrightarrow$ inf.
- The Thompson Sampling regret curve has slightly more variance than the UCB regret curve. This could probably be explained by the fact that while UCB picks an arm only optimistically (only maximising the upper confidence bound around the current mean estimate, thereby going in only one direction around the mean estimate), Thompson Sampling samples arms from a distribution centered around the mean in both directions. This could lead to more variance in the losses and hence, regret curves obtained.

2.3 Regret Plots for Varying $\Delta \in [0.1, 0.2, ..., 1.0]$

- Both UCB and Thompson Sampling perform increasingly well as Δ increases.
- This is expected, as the gap between the means of the two arms increases as Δ increases.
- In the case of UCB, this means that the Lower Confidence Bound of the higher-loss estimate is less and less likely to be below that of the lower-loss estimate, even if the higher-loss estimate has been picked fewer times up until any given timestep.

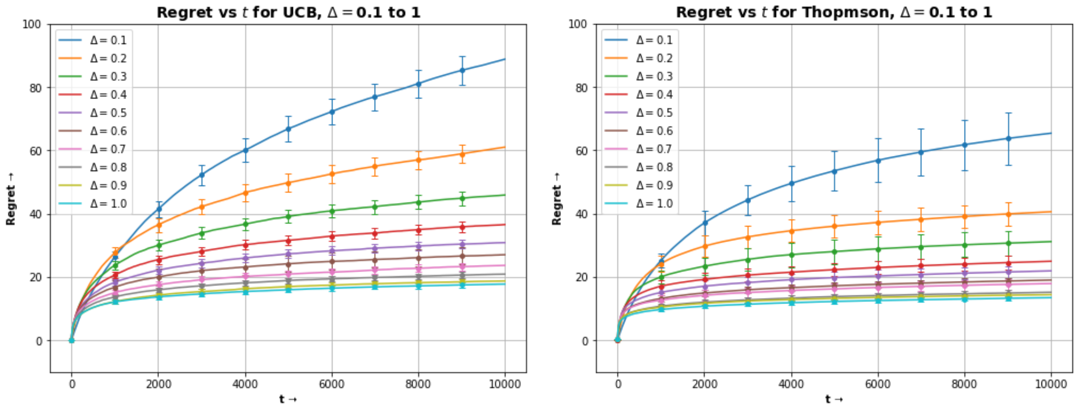

Figure 4: Regret curves for UCB and Thompson Sampling, $\Delta \in [0.1, 0.2, ..., 1.0], T = 10000$

- In the case of Thompson Sampling, the increased size of this gap means that the mean of the Posterior Gaussian distribution of the higher-loss arm, tends to move farther away from that of the lower-loss arm, because of the increased probability of getting a higher loss from this arm, and subsequently the update rule $\mu_0^t = \frac{\tau_0^{t-1}\mu_0^{t-1}+l_t}{\tau_0^{t-1}+\tau}$. When the updated mean of this posterior Gaussian now moves away, and the variance reduces with rounds, the probability of the θ_t of this worse arm being lesser than that of the best arm reduces. Therefore, the best arm is picked more and more, and the accumulated regret reduces as well.

* *