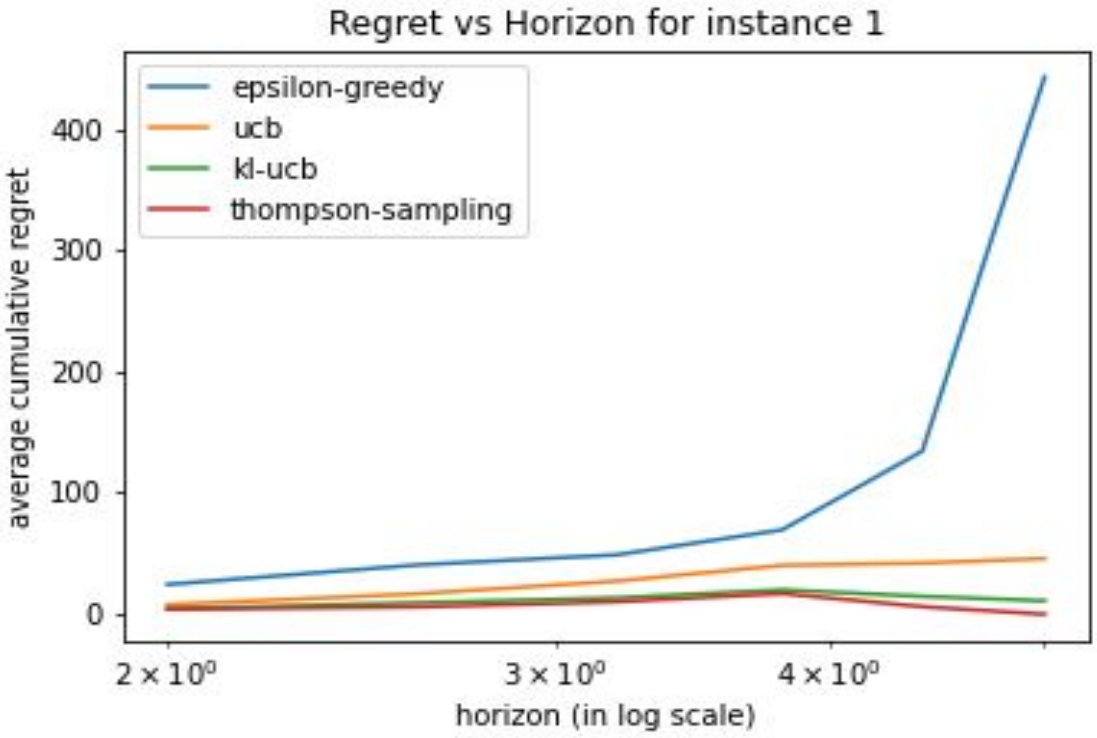
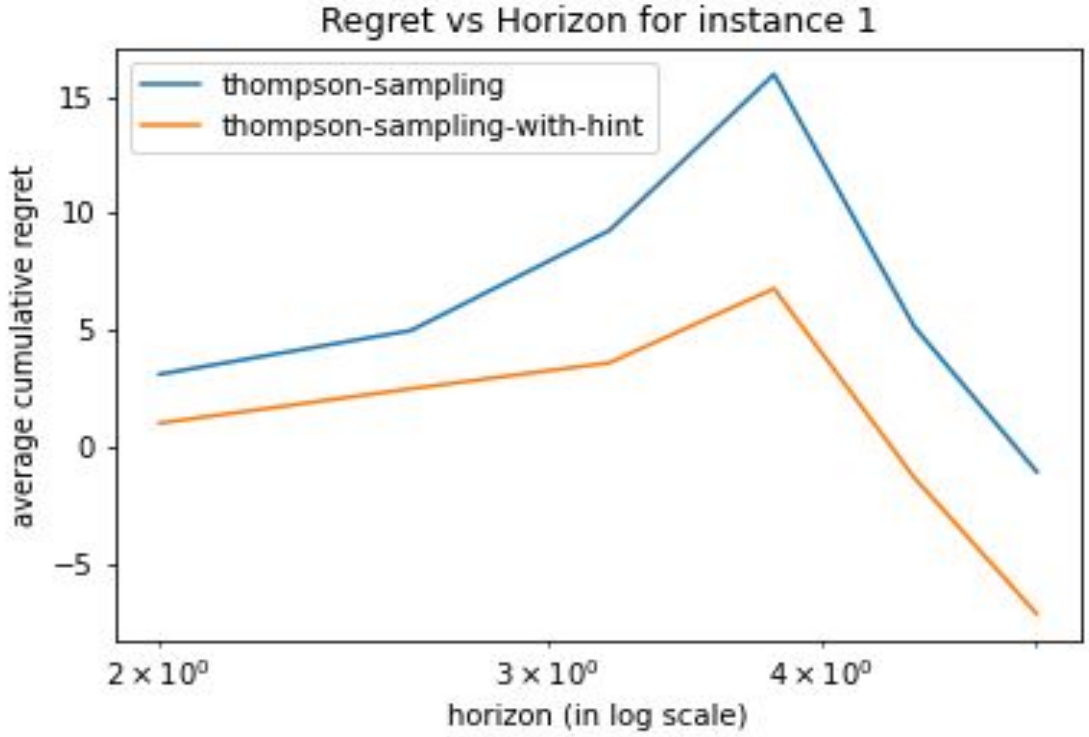
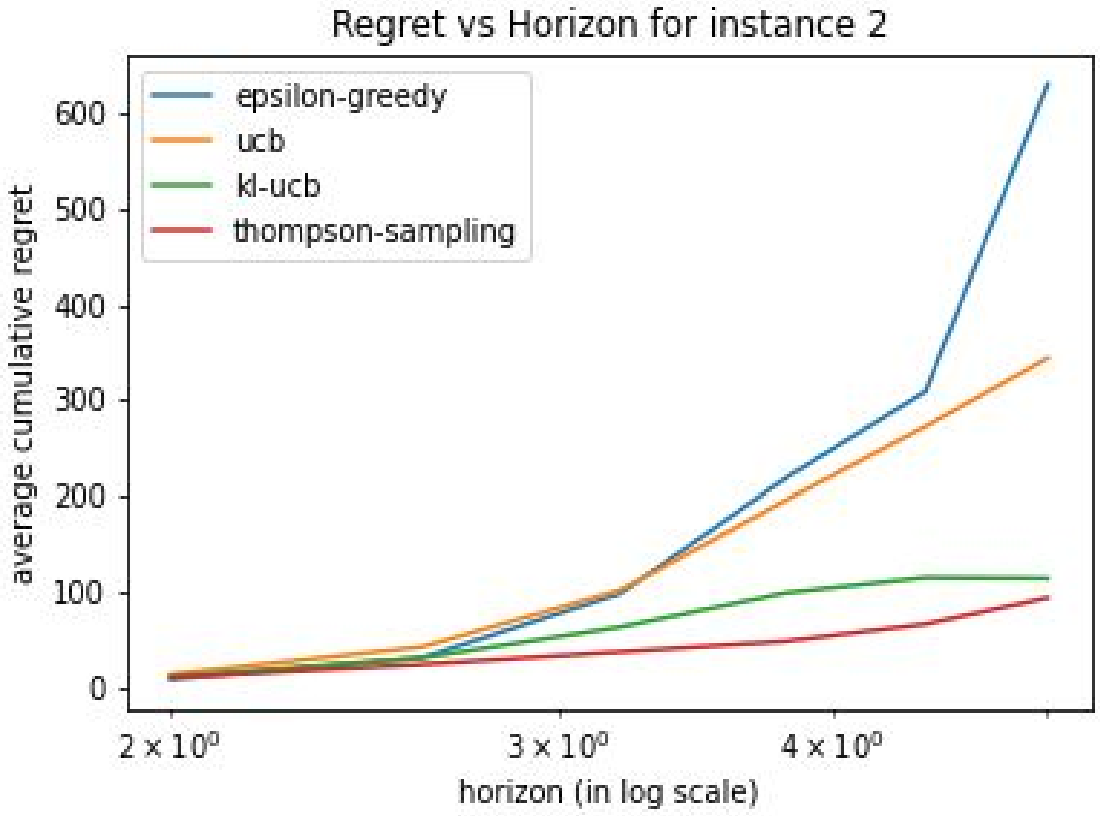
CS747 Assignment - 1 Report

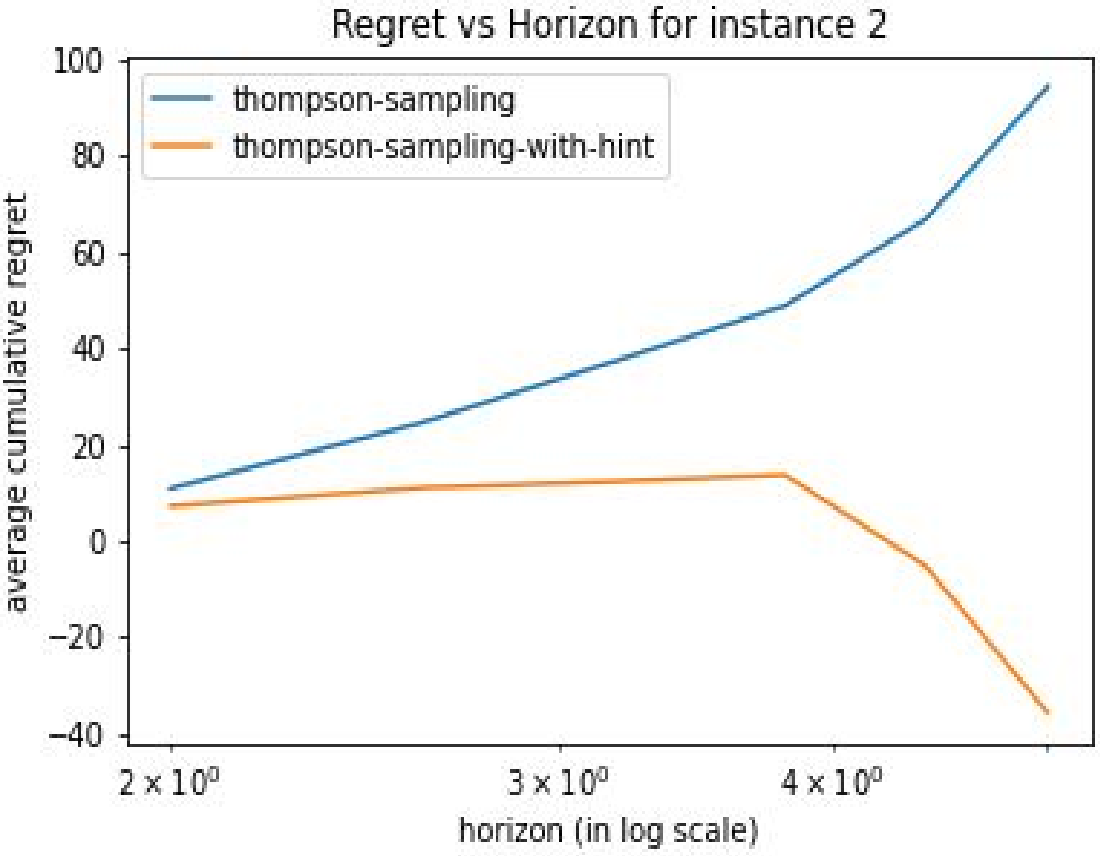
Name - Saurabh Parekh
Roll number - 170100016

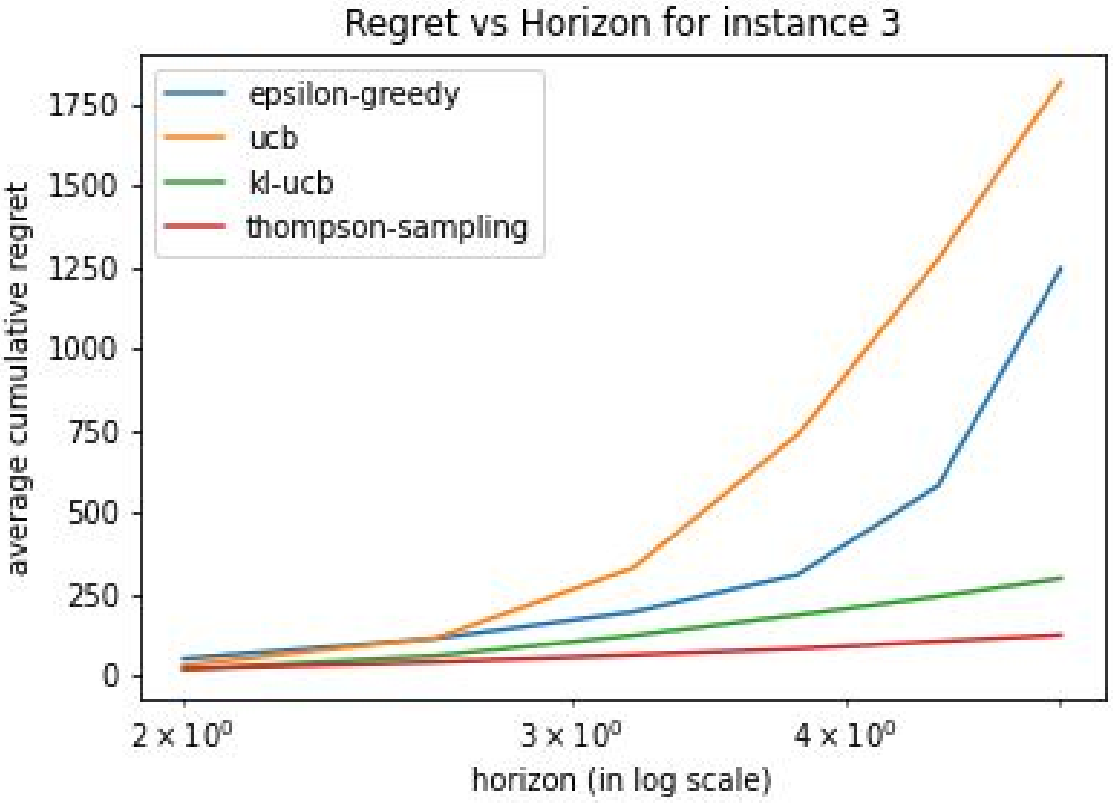
Graphs:

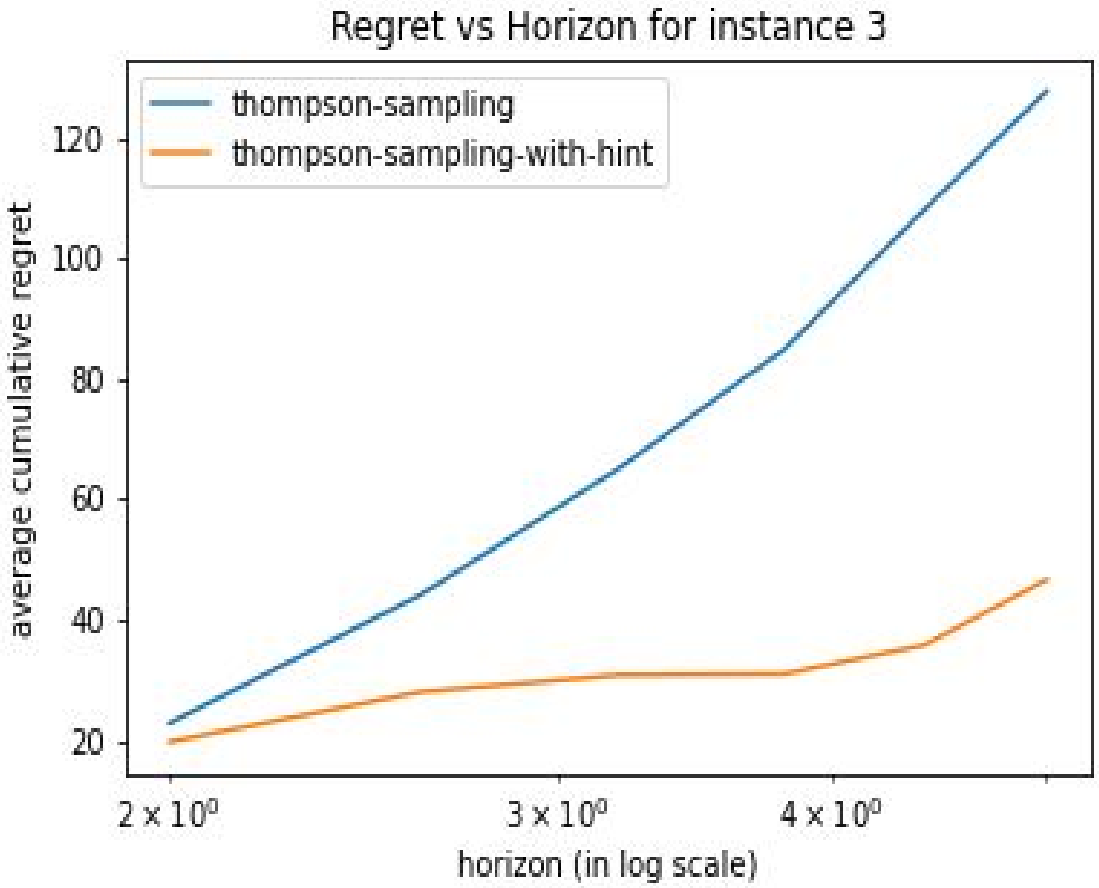





Regret vs Horizon for instance 2
epsilon-greedy
ucb
kl-ucb
thompson-sampling
average cumulative regret
600
500
400
300
200
100
0
2 × 10^0
3 × 10^0
4 × 10^0
horizon (in log scale)


Regret vs Horizon for instance 2
thompson-sampling
thompson-sampling-with-hint
average cumulative regret
100
80
60
40
20
0
−20
−40
2 × 10^0
3 × 10^0
4 × 10^0
horizon (in log scale)


Regret vs Horizon for instance 3
epsilon-greedy
ucb
kl-ucb
thompson-sampling
1750
1500
1250
1000
750
500
250
0
average cumulative regret
2 x 10^0
3 x 10^0
4 x 10^0
horizon (in log scale)


Regret vs Horizon for instance 3
thompson-sampling
thompson-sampling-with-hint
120
100
80
60
40
20
average cumulative regret
2 x 10^0
3 x 10^0
4 x 10^0
horizon (in log scale)

Assumptions and Implementation Details:

1. Epsilon-greedy

- A random arm is sampled with probability **ε** and the arm with highest empirical mean is chosen with probability 1-**ε** for each epoch.
- If 2 or more arms have the highest empirical mean, then the one with the smallest index among those is chosen.

2. UCB

- Initially, round-robin sampling is done once for each arm. If the horizon is less than the number of arms, then the first horizon number of arms are sampled.
- After the initial round-robin sampling, the arm having the highest ucb value is sampled. If there are 2 or more arms with the highest ucb value, then the one with the smallest index among those is chosen.

3. KL-UCB

- Initially, round-robin sampling is done once for each arm. If the horizon is less than the number of arms, then the first horizon number of arms are sampled.
- After the initial round-robin sampling, the arm having the highest kl-ucb value is sampled. Binary search is used for finding the kl-ucb value by taking a precision of 1e-6. Hyperparameter value is chosen as c = 3.

4. Thompson Sampling

- A random sample is taken from the belief beta distribution of each arm and the arm with the highest value is sampled and only it's distribution is updated.
- If there are 2 or more arms with the highest value, then the one with the smallest index among those is chosen.

5. Thompson Sampling with Hint

- A discrete uniform belief distribution is taken initially for each arm.
- A random sample is chosen for each arm at each epoch from the belief distribution and the one with the highest value is sampled.
- If there are 2 or more arms with the highest value, then the one with the smallest index among those is chosen.
- The belief distribution of the sampled arm is updated based on bayesian inference principles. If the reward(evidence) from the sampled arm is 1, then the belief distribution of the sampled arm is updated by multiplying it with the actual means from the hint and dividing it by the normalization factor.
- This process is repeated at every epoch for horizon number of times.

Observations and Interpretations:

1. It was observed that epsilon-greedy performed better than ucb for small horizons.
2. Average cumulative regret for epsilon-greedy can be seen as linear from the graphs whereas it is sublinear for all other algorithms.

3. For instance 3, epsilon-greedy can be seen to be having less average cumulative regret than ucb even for large horizons but it can be noticed from the graph that the slope of the line is higher for epsilon-greedy than ucb. So we can say that regret is still linear for epsilon-greedy whereas it is sublinear for ucb.
4. For all 3 instances, thompson-sampling performs the best followed by kl-ucb and ucb respectively as is evident from the graphs.
5. Thompson-sampling with hint performs much better than thompson-sampling for all 3 instances for all horizons. The reason is that it updates the belief distribution only for discrete number of points rather than a continuous beta distribution. Hence it samples the optimal arm more often and quite early.
6. Kl-ucb takes the highest running time whereas thompson-sampling takes the least amount of time in giving regret.
7. Regrets for thompson-sampling and thompson-sampling with hint decreases at large horizons for instances 1 and 2.
8. It can be concluded from T3 results that the average cumulative regret for epsilon-greedy for large horizons is minimum for the value of **ε** close to 0.02

Results for T3:

1. Instance 1 -> **ε1** = 0.001, **ε2** = 0.02, **ε3** = 0.9
2. Instance 2 -> **ε1** = 0.001, **ε2** = 0.02, **ε3** = 0.9
3. Instance 3 -> **ε1** = 0.001, **ε2** = 0.02, **ε3** = 0.9

- Regret values for instance 1
  1279.46 for **ε** = 0.001
  442.84 for **ε** = 0.02
  2042.28 for **ε** = 0.1
  8183.12 for **ε** = 0.4
  18417.4 for **ε** = 0.9
  20477.76 for **ε** = 1

- Regret values for instance 2
  2579.0 for **ε** = 0.001
  629.8 for **ε** = 0.02
  2076.52 for **ε** = 0.1
  8192.28 for **ε** = 0.4
  18429.8 for **ε** = 0.9
  20505.9 for **ε** = 1

- Regret values for instance 3
  6005.78 for **ε** = 0.001
  1247.26 for **ε** = 0.02
  4356.94 for **ε** = 0.1
  16941.78 for **ε** = 0.4
  38094.84 for **ε** = 0.9
  42308.5 for **ε** = 1