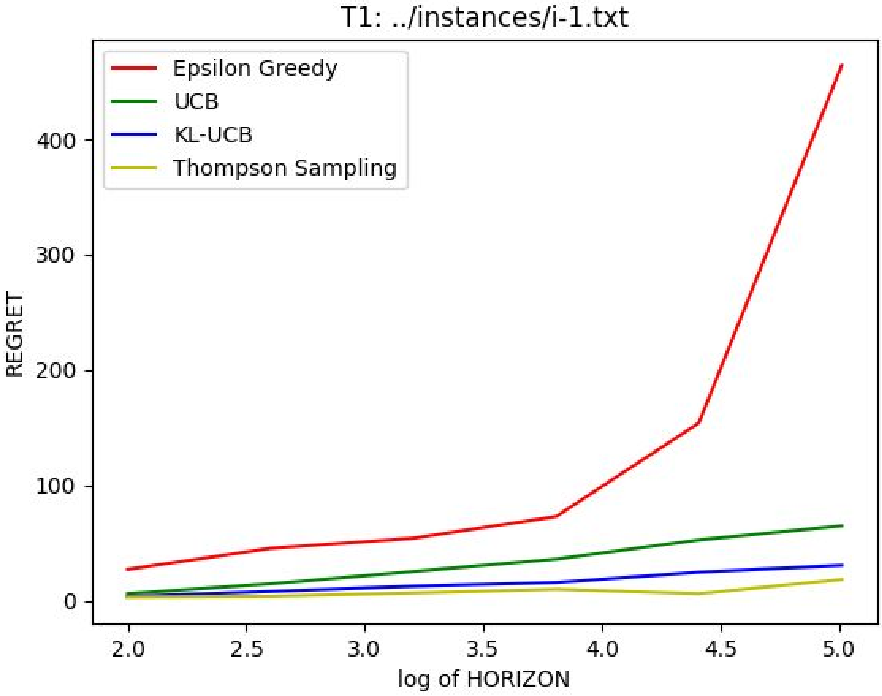
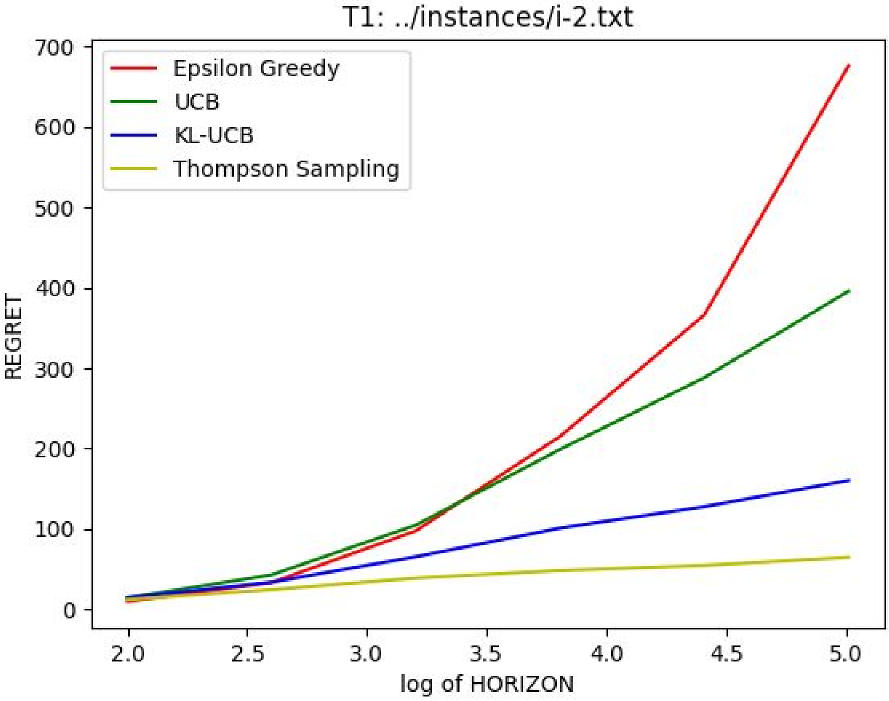
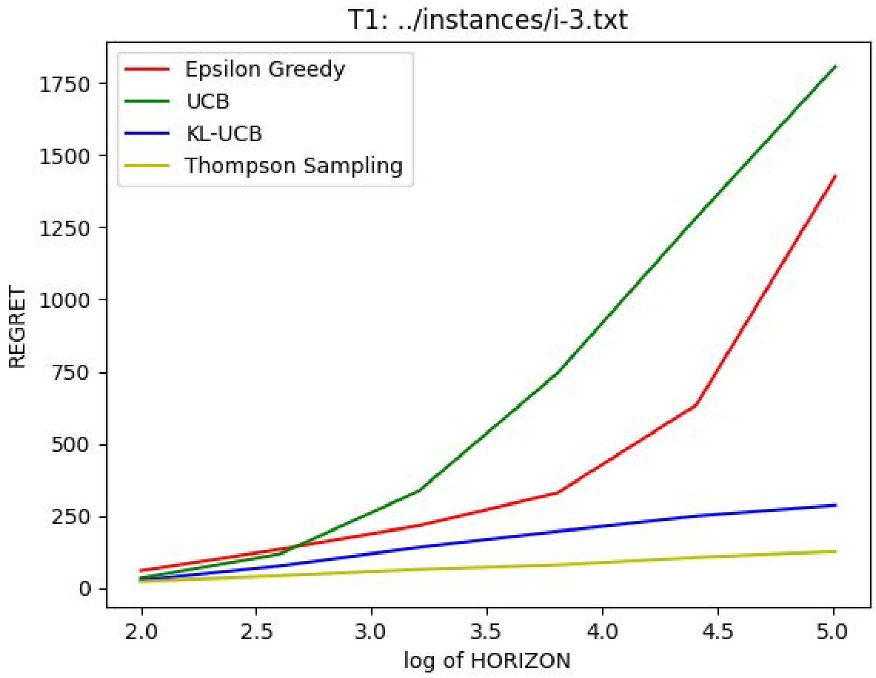
Programming Assignment 1

Hari Krishna Nama (170050077)

Plots for T1:







X-axis is of base-10 log

As expected Epsilon-Greedy didn't do well in any instance since it's linear regret, whereas others are sub-linear. For fewer arms, KL-UCB and UCB gave similar regrets but as the number of arms increased, the difference between them became very distinct, regret in UCB grew rapidly with the horizon for more arms. Both Thompson Sampling and KL-UCB did very well in all three instances, but thompson sampling has less regret following the explanation from the slides

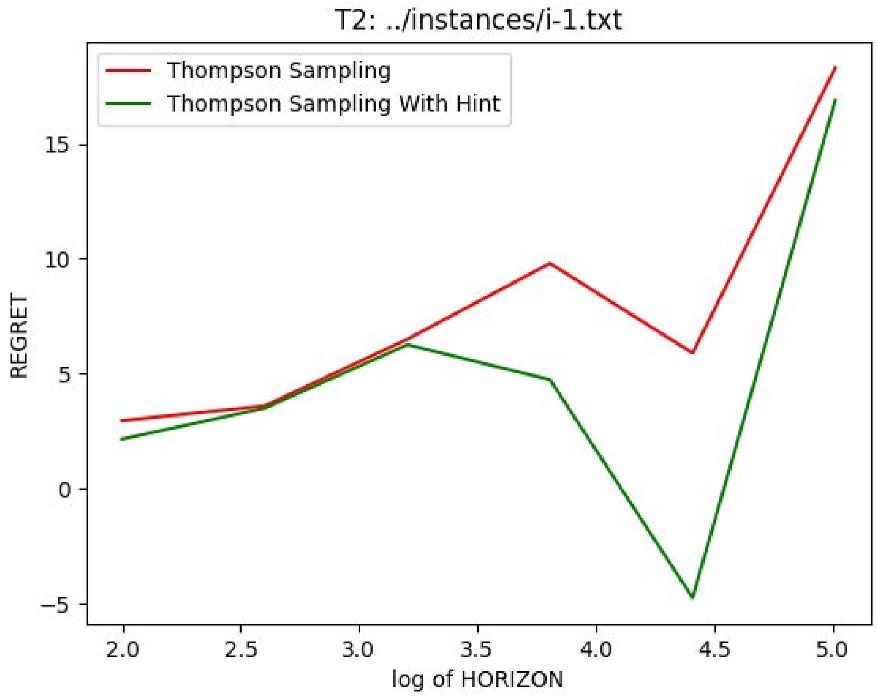
T3:

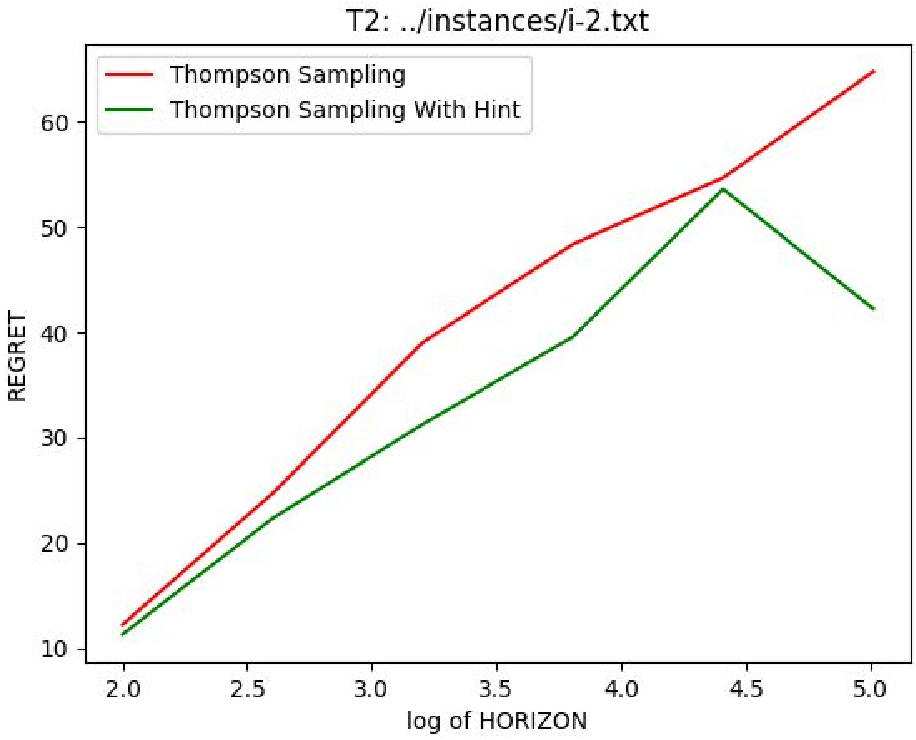
	e1 = 0.001	e2 = 0.01	e3 = 0.1
../instances/i-1.txt	878.1	302.3	2058.8
../instances/i-2.txt	4351.3	933.0	2117.6
../instances/i-3.txt	6424.7	1203.2	4327.3

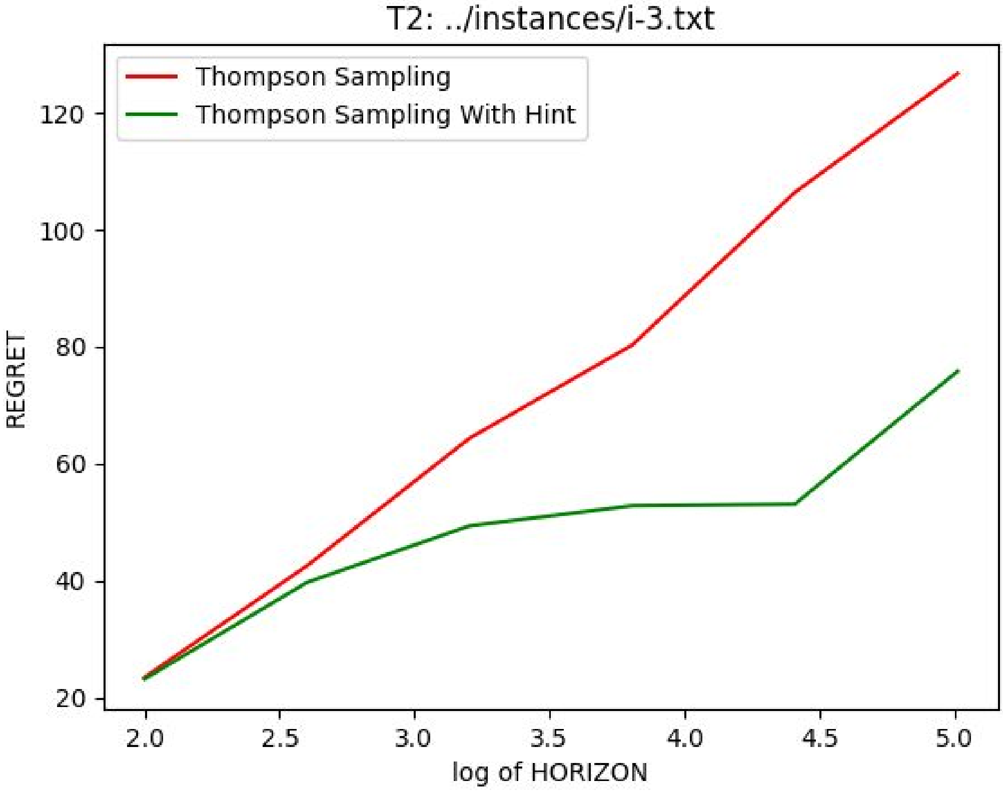
From the above table, we can at all three instances Regret calculated at epsilon = 0.01 is less than regret at epsilon = 0.001 or regret at epsilon = 0.1
Therefore

epsilon1 = 0.001
epsilon2 = 0.01
epsilon3 = 0.1

Plots for T2:


T2: ../instances/i-1.txt
Thompson Sampling
Thompson Sampling With Hint
REGRET
−5
0
5
10
15
2.0
2.5
3.0
3.5
4.0
4.5
5.0
log of HORIZON


T2: ../instances/i-2.txt
Thompson Sampling
Thompson Sampling With Hint
REGRET
10
20
30
40
50
60
2.0
2.5
3.0
3.5
4.0
4.5
5.0
log of HORIZON



X-axis is of base-10 log

For Thompson Sampling With Hint, Hint I assumed is knowing **only** the **highest** mean reward. Only with this assumption, we can create an algorithm which can create low regret in any bandit instance. From the above examples, we can observe the regret for any horizon and any bandit instance is less than the regret obtained from Thompson Sampling. Above algorithm works well with more number of arms in the bandit instance

Algorithm:

- Run the exploration phase for the first epsilon fraction of horizon
  - Use Thompson Sampling to build empirical means in this phase
  - In assignment, exploration phase ran for `max(0.016*horizon, 2*numArms)` rounds
- Then in the exploitation phase:
  - If the empirical mean of any arm is close to the highest mean reward(prior information) then pull that particular arm
  - Else do the Thompson sampling

This algorithm is a hybrid of epsilon greedy and Thompson sampling. Unlike uniformly choosing the arm in the exploration phase, here we use Thompson sampling. In the exploitation phase, if any arm’s empirical mean is close to the highest mean reward(hint information) then pull this arm, else Thompson Sampling

Assumptions for Each Algorithm:

- **Epsilon Greedy:**
    - No first few pulls
    - When multiple arms have same empirical means then the first arm is picked
    - Epsilon is provided in command line argument
- **UCB:**
    - 1 Round robin for the first few pulls
    - When multiple arms have same ucbs then the first arm is picked
    - No algorithmic specific params
- **KL-UCB:**
    - 1 Round robin for the first few pulls
    - When multiple arms have same kl-ucbs then the first arm is picked
    - c is 3
    - max 50 iterations and precision 0.001 while searching appropriate kl-ucb
- **Thompson Sampling:**
    - No first few pulls
    - When multiple arms have same computational samples then the first arm is picked
    - No algorithmic specific params
- **Thompson Sampling With Hint:**
    - Exploration phase for `max(epsilon*horizon, 2*numArms)` rounds
    - When multiple arms have same computational samples or empirical means then the first arm is picked
    - 0.016 is used as epsilon for the exploration phase
    - 0.01 is used for precision while choosing in the exploitation phase