Programming Assignment 1
Karan Agarwalla
180050045

T1:

In general, in case of multiple maximal values the first such value is considered.

Epsilon Greedy: We explore with probability ε and exploit with probability 1 - ε. Empirical means of all arms are set to 0 initially.

UCB: All arms are pulled once initially to ensure that the expression of ucb_a^t is well defined.

$$ucb_a^t = \hat{p}_a^t + \sqrt{\frac{2*\ln(t)}{u_a^t}}$$

$\hat{p}_a^t$ is the empirical mean for arm a till time t
u_a^t is the number of times arm a has been pulled till time t

KL-UCB: All arms are pulled once initially to ensure that the expression is well defined.

$$ucb\_kl_a^t = \max\{q \in [\hat{p}_a^t, 1] |\, u_a^t KL(\hat{p}_a^t, q) \leq \ln(t) + cln(\ln(t))\}$$

$\hat{p}_a^t$ is the empirical mean for arm a till time t
u_a^t is the number of times arm a has been pulled till time t

Implementation is done with c = 0 since it provides better results as compared with c>=3. To calculate q binary search with precision = 10^{-3} has been used.

Thompson Sampling: Mean of each arm a is assigned a prior belief in terms of successes (s_a^t) and failures (f_a^t) till time t as Beta($s_a^t + 1, f_a^t + 1$) . Then while selecting an arm, sample **X** is drawn from the distribution of each arm and that arm is pulled for which **X** is maximum. s_a^t *and* f_a^t is initially assigned zero values (In the implementation $s_a^t + 1$ *and* $f_a^t + 1$ is maintained).

T2:

In general, in case of multiple maximal values the first such value is considered.

Thompson Sampling with Hint: We maintain a prior PDF belief of each arm. So, we maintain a two-dimensional matrix *prior* such that $Prior_{ij}^t$ is the probability of the jth arm having the ith mean at time t. Now at an instant t, we draw the arm with the highest probability of the highest mean. Note we have access to a sorted list of means. Then we update the posterior conditioned on the respective reward. Suppose arm **j** is pulled at time t_0. Prior for all other arms remain the same at $t = t_0 + 1$.

p_j: $The\ jth\ mean\ in\ the\ sorted\ list$
$Prior_{ij}^{t}$: $Probability\ of\ jth\ arm\ with\ mean\ p_i\ at\ time\ t$
n: $Number\ of\ bandit\ arms$

We update the belief of **jth arm** as

Reward is 1(Success)

$$Prior_{ij}^{t_0+1} = \frac{\mathrm{Prior}_{\mathrm{ij}}^{\mathrm{t_0}} * \mathrm{p_i}}{\sum_{k=1}^{n} \mathrm{Prior}_{\mathrm{kj}}^{\mathrm{t_0}} * \mathrm{p_k}}$$

Reward is 0(Failure)

$$Prior_{ij}^{t_0+1} = \frac{\mathrm{Prior}_{\mathrm{ij}}^{\mathrm{t_0}} * (1 - \mathrm{p_i})}{\sum_{k=1}^{n} \mathrm{Prior}_{\mathrm{kj}}^{\mathrm{t_0}} * (1 - \mathrm{p_k})}$$

Also,$Prior_{ij}^{0} = \frac{1}{n}$, i.e., we assume a uniform PDF distribution at t = 0.

I also tried to sample values as per the prior pdf and then pulling the arm for which the value drawn is maximum as one does in Thompson Sampling. However, this algorithm does not perform better than Thompson Sampling on all instances and horizon. Then we tweak the algorithm a little bit and pull the arm which has the highest probability of the highest mean. Note in regret we consider the difference with the highest mean. This provides an intuition of minimizing regret.

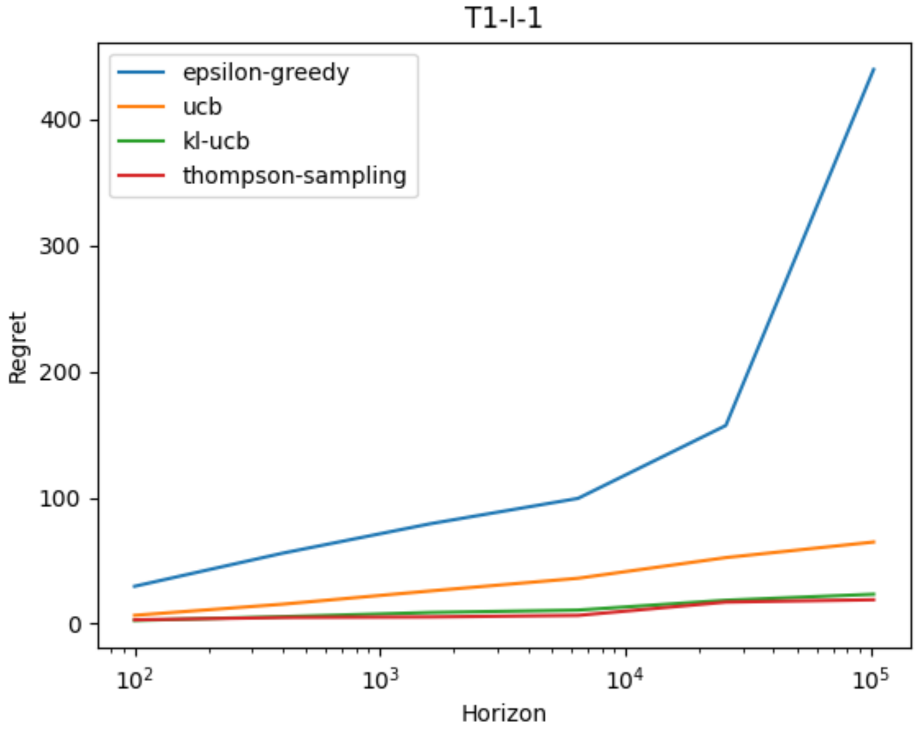
T3:

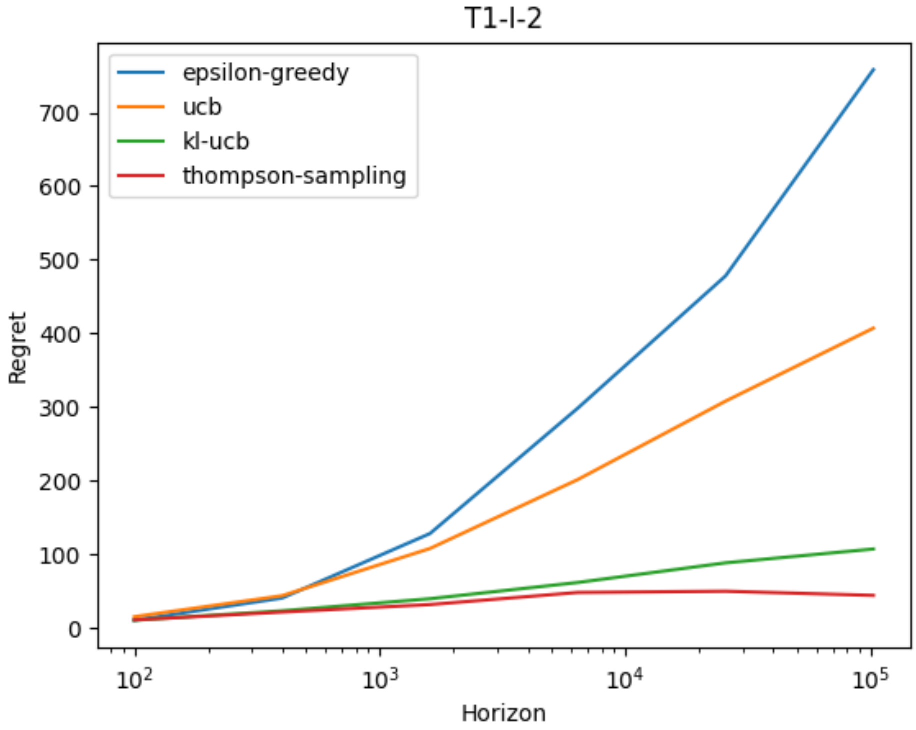
Keeping ε small prevents sufficient exploration whereas keeping ε high prevents sufficient exploitation. The values of ε1 = 0.001, ε2 = 0.01 and ε3 = 0.1 work for all the three given instances. Each of the values mentioned below are averaged over 50 seeds and for horizon 102400.

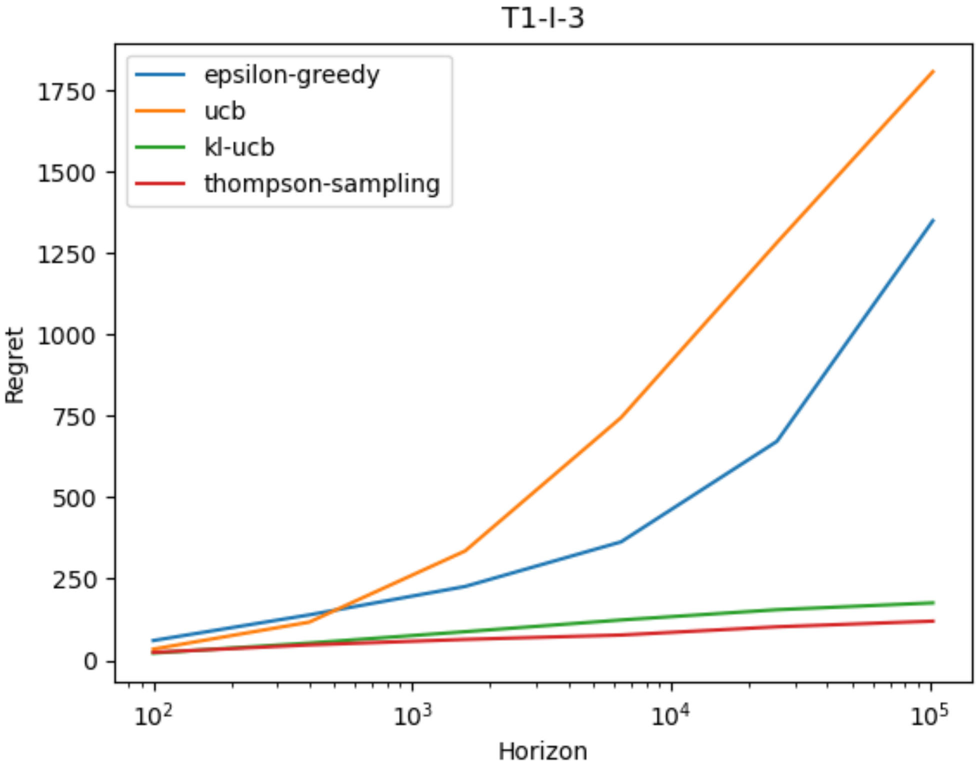
Here are the respective regrets:

	Instance I-1	**Instance I-2**	**Instance I-3**
ε1=0.001	1060.26	3657.6	5095.22
ε2=0.01	339.22	919.9	1221.06
ε3=0.1	2050.84	2098.7	4320.44

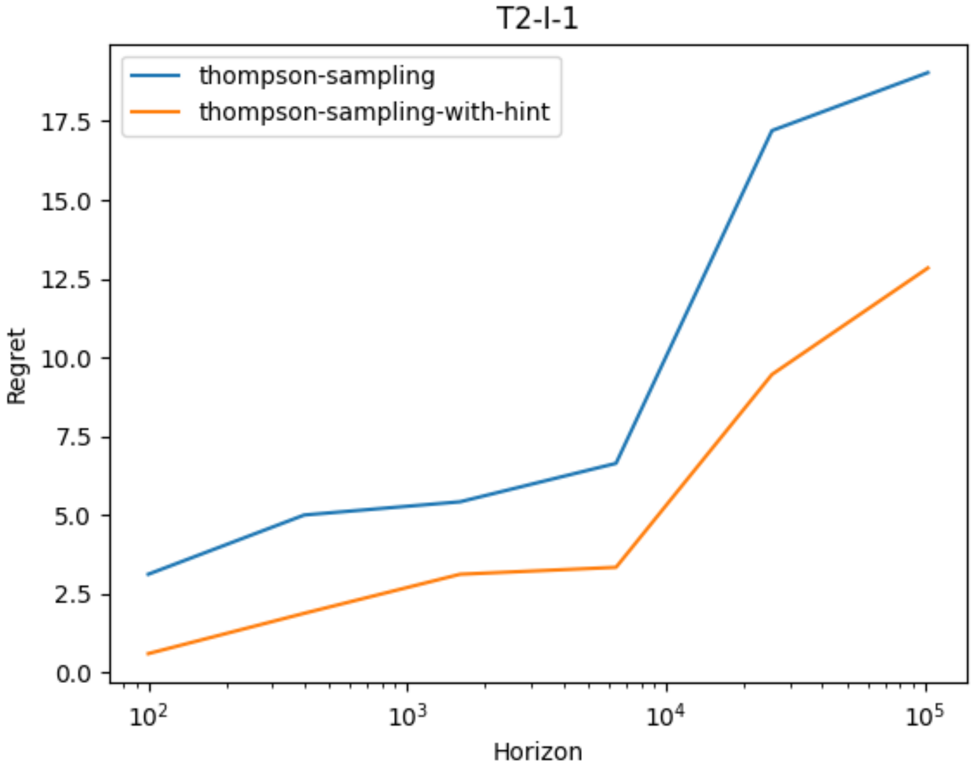
T4:

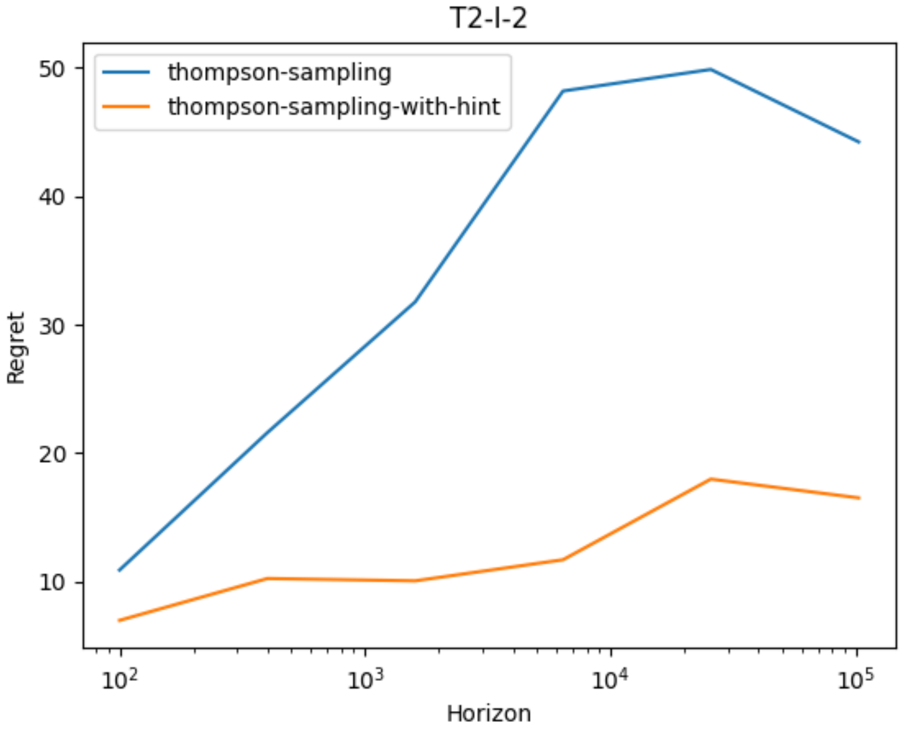

T1-I-1
epsilon-greedy
ucb
kl-ucb
thompson-sampling
Regret
0
100
200
300
400
10^2
10^3
10^4
10^5
Horizon

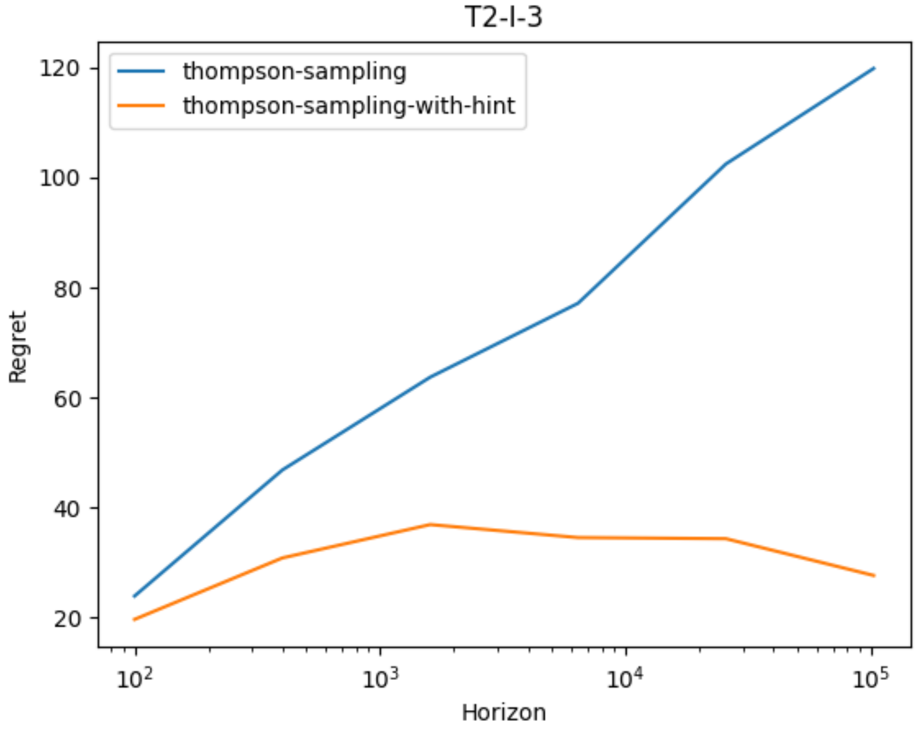

T1-I-2
epsilon-greedy
ucb
kl-ucb
thompson-sampling
Regret
0
100
200
300
400
500
600
700
10^2
10^3
10^4
10^5
Horizon



From the plots of T1, we observe that the regret for Thompson Sampling is the lowest among the other algorithms. Then comes KL-UCB followed by UCB. One notices that that UCB, KL-UCB and Thompson Sampling are almost linear in log(horizon) whereas epsilon-greedy is exponential in log(horizon) or aptly linear in horizon. Now, one sees that UCB performs worse than epsilon-greedy in I-3. However, the slope at the right end indicates otherwise on a larger horizon as epsilon-greedy is exponentially rising whereas UCB is almost linear.







One observes that Thompson-sampling-with-hint performs better than Thompson-sampling as it has more information to base its actions upon.