CS747 Assignment 1

Manas Vashistha (17D070064)

25th September, 2020

Terms Used

- t - Time step.
- n_a - number of arms.
- r_t - reward at time step t.
- a_t - arm pulled at time step t.
- u_t - number of total pulls just before pulling a_t.
- $\hat{p}_a$ - empirical mean of arm a.
- u_a^t - number of pulls of arm a at time step t.

Assumptions

- All the ties are broken randomly using `np.random.choice(np.where(x == x.max())[0])` .
- In UCB and KL-UCB, for first n_a pulls, the arm corresponding to u_t is pulled.
- In Thompson-sampling-with-hint the first pull is executed using Standard Thompson-Sampling, After this the algorithm is executed for all the following pulls.

Epsilon Greedy

Pull a random arm with probability $epsilon$ (exploration) and, pull the arm with highest empirical mean with probaility (1 - $epsilon$).

UCB

1. Pull arms in order till all the arms are pulled exactly once.
2. Calculate ucb for each arm as $ucb_a^t = \hat{p}_a + \sqrt{\frac{2 \log t}{u_a^t}}$.

3. After that pull the arm for which UCB is maximum.

KL-UCB

1. Pull arms in order till all the arms are pulled exactly once.
2. Search for optimal $q_a \in [\hat{p}_a, 1]$ for each arm a by binary search as follows:
   - For each pull set $Bound = \log(t) + c\log(\log(t))$.
   - Start searching in the interval $[\hat{p}_a, 1]$; Set left$= \hat{p}_a$ and right$= 1$.
   - For each arm set $q_a = \frac{(left + right)}{2}$.
   - Calculate $KL(\hat{p}_a, q_a)$.
   - For a q_a if $KL(\hat{p}_a, q_a) - Bound <= 0$ search for q_a in $[q_a, 1]$ (maximise q_a in left$= q_a$ and right$= 1$) else in $[\hat{p}_a, q_a]$ (find the q_a to satisfy condition left$= \hat{p}_a$ and right$= q_a$).
   - Repeat the above three steps until $abs(KL(\hat{p}_a, q_a) - Bound) < 10^{-4}$ and the interval_width, $(left - right) < 10^{-4}$.
3. Return the q_{max} among q_a for all arms from the above steps.

Thompson-Sampling

Sample means from beta distribution for each arm and whichever arm has highest mean gets pulled at each step.

Thompson Sampling with Hint

Idea-

The provided true means are used to estimate the probability distribution of hints for each of the arms.

Assumption-

First pull is executed using the standard thompson-sampling method. After this the algorithm is executed for all the following pulls.

- Say p_{arg} is the index of the column in $hint$ which has the highest value
  i.e. $p_{arg} = idx(max(hint))$.

Algorithm-

1. Pull arm at first time step using the standard Thompson-Sampling approach.
2. Create an array $mypdf$ of dimensions ($n_a \times n_a$) and set $mypdf[i][j] = \frac{1}{n_a}$ for $0 \leq i, j \in 0, 1, 2, ..., n_a - 1$
3. At step t, if $r_{t-1} == 1$ then multiply $mypdf[a_{t-1}]$ with the given $hint$ else if $r_{t-1} == 0$ multiply $mypdf[a_{t-1}]$ with $(1 - hint)$ and normalize $mypdf[a_{t-1}]$.
4. Pull the arm which has the highest value in the column p_{arg} i.e. $a_t = idx(max(mypdf[:, p_{arg}]))$.
5. Repeat Steps 2-4 for all values of t in $range(T)$.

Experiment with ϵ

Selected values of epsilon: $\epsilon_1 = 0.0003$, $\epsilon_2 = 0.04$, $\epsilon_3 = 0.6$ such that $\epsilon_1 < \epsilon_2 < \epsilon_3$.

Instance1	Horizon(T) = 102400
ϵ	Average Regret
0.003	878.74
0.04	825.54
0.6	12273.44

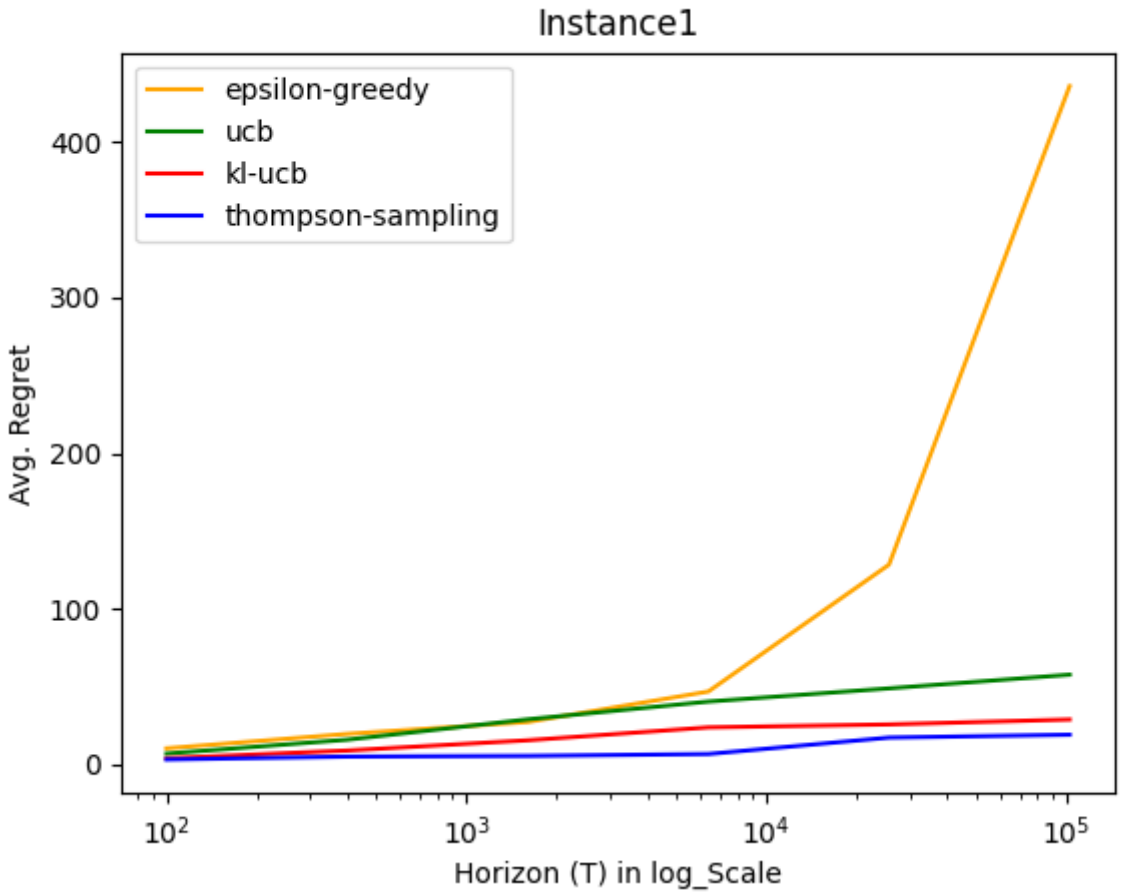
Instance2	Horizon(T) = 102400
ϵ	Average Regret
0.003	6345.2
0.04	982.94
0.6	12252.38

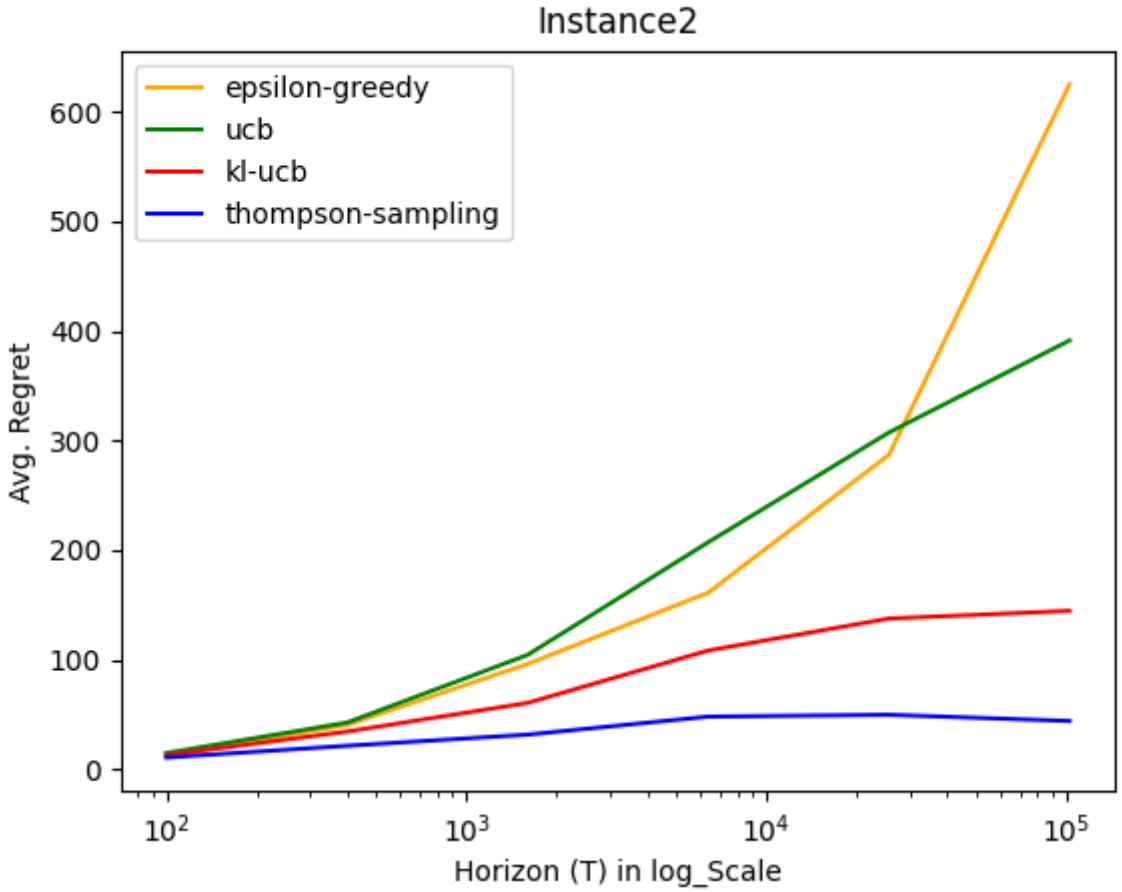
Instance3	Horizon(T) = 102400
ϵ	Average Regret
0.003	8400.06
0.04	1890.7
0.6	25430.64

Clearly, the Regret for ϵ_2 is less than that of ϵ_1 and ϵ_3. This is justified because if we take very low ϵ then there is less chance of exploring and we keep exploiting with the little information we have. Similarly if the ϵ is large then there is less chance of exploit and we do more exploration. Both of these approaches result in a larger value of regret. However if we set our ϵ somewhere in the middle we get a good chance of exploring as well as exploiting and hence achieving smaller regret.

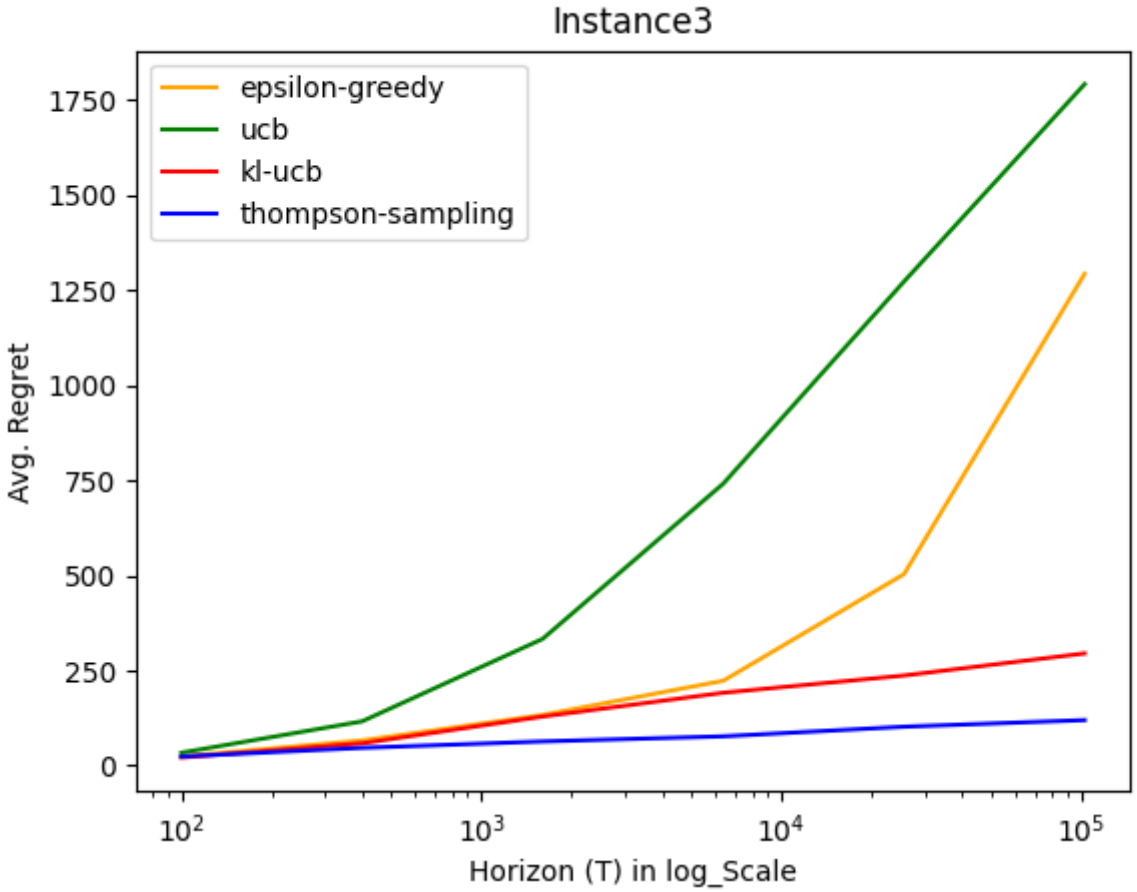
Plots

Plots for Epsilon-Greedy, UCB, KL-UCB and Thompson-Sampling





Continued on Next Page...

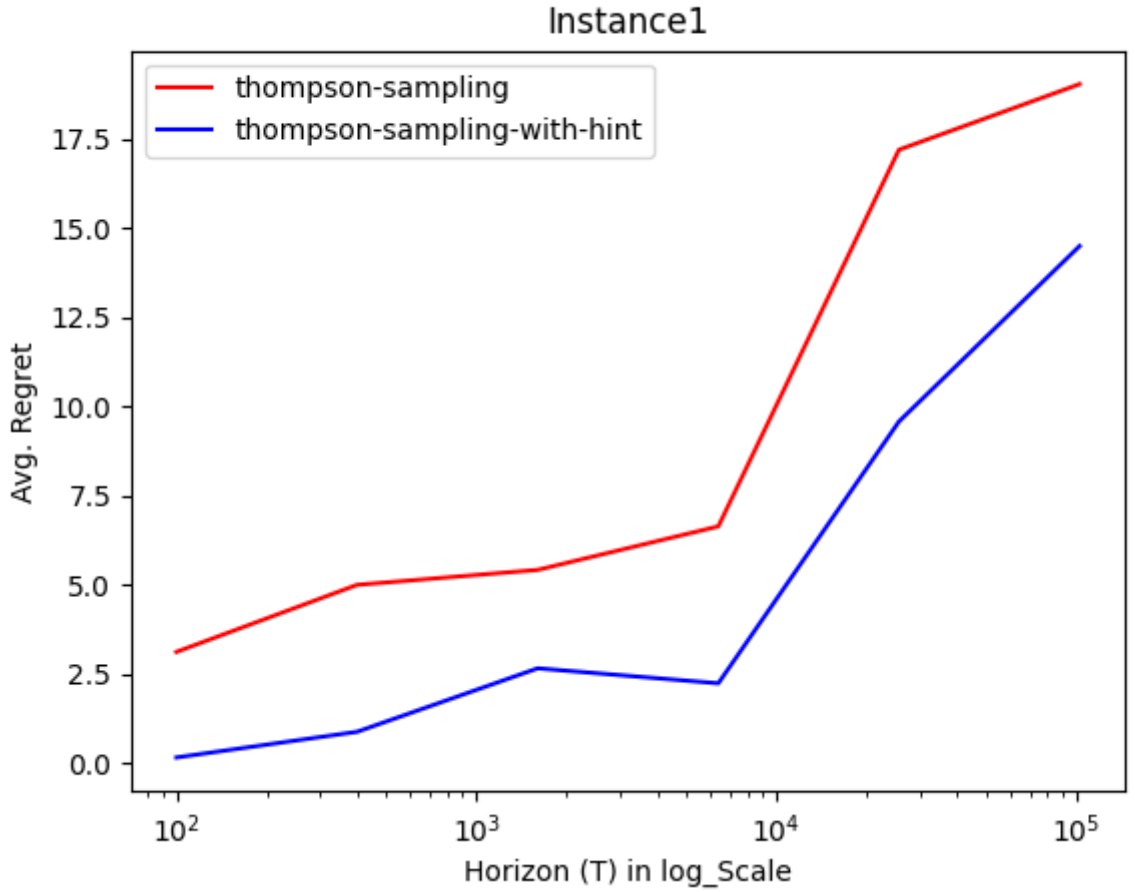


Inferences-

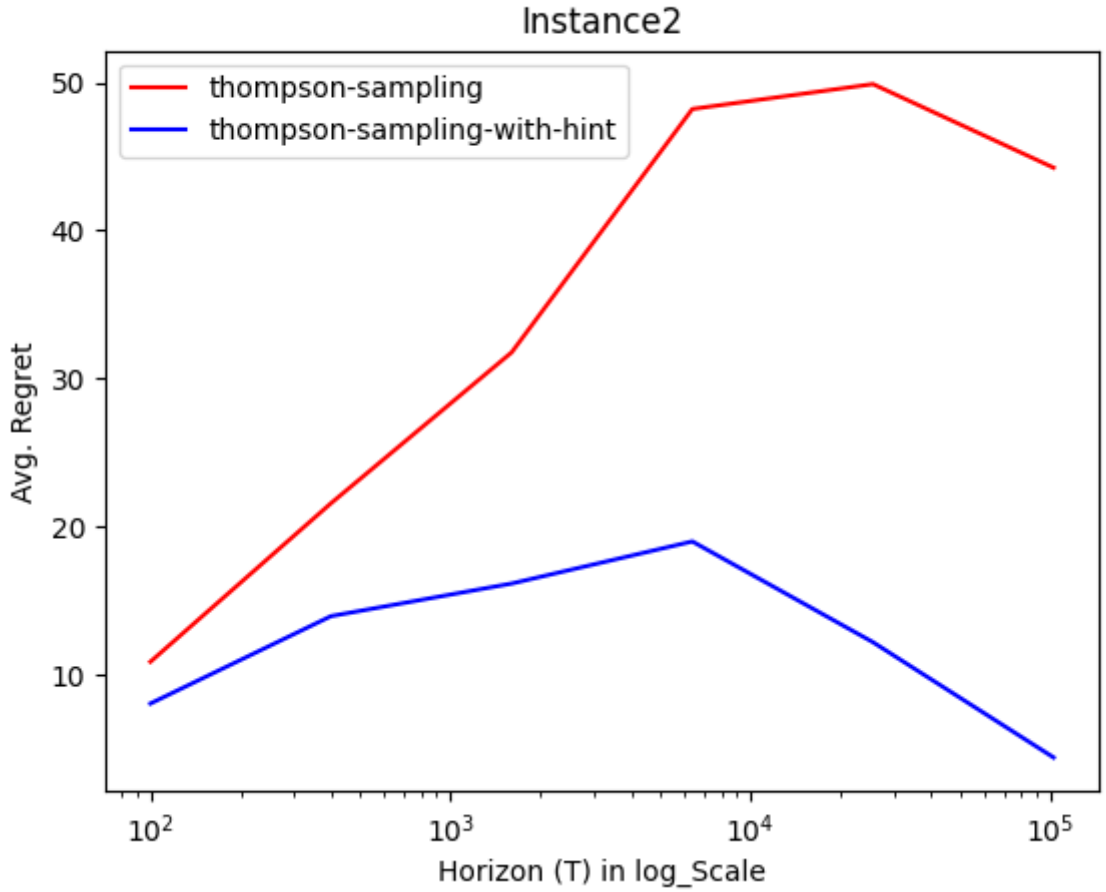
We can draw the following inferences-

- **Instance 1:** Epsilon Greedy performs good for smaller values of horizons but gives very high regret as horizon is increased. UCB gives lower regret than epsilon-greedy but the regret goes up as the horizon is increased. UCB is followed by KL-UCB which gives less regret than UCB. Thompson Sampling achieves the least regret.
- **Instance 2:** The performance of UCB and Epsilon greedy is comparable for lower values of horizon. But as horizon is increased, Epsilon greedy starts performing poorer than UCB. After this, UCB is followed by KL-UCB which gives less regret than UCB continuously. Thompson Sampling achieves the least regret for this instance also.
- **Instance 3:** Here UCB performs bad and gives large values of regret even on smaller horizons. Epsilon greedy retains its behaviour of good performace at smaller horizons and gets poorer as the horizon is increased. Epsilon greedy is followed by KL-UCB which gives less regret. Thompson Sampling achieves the least regret here also.

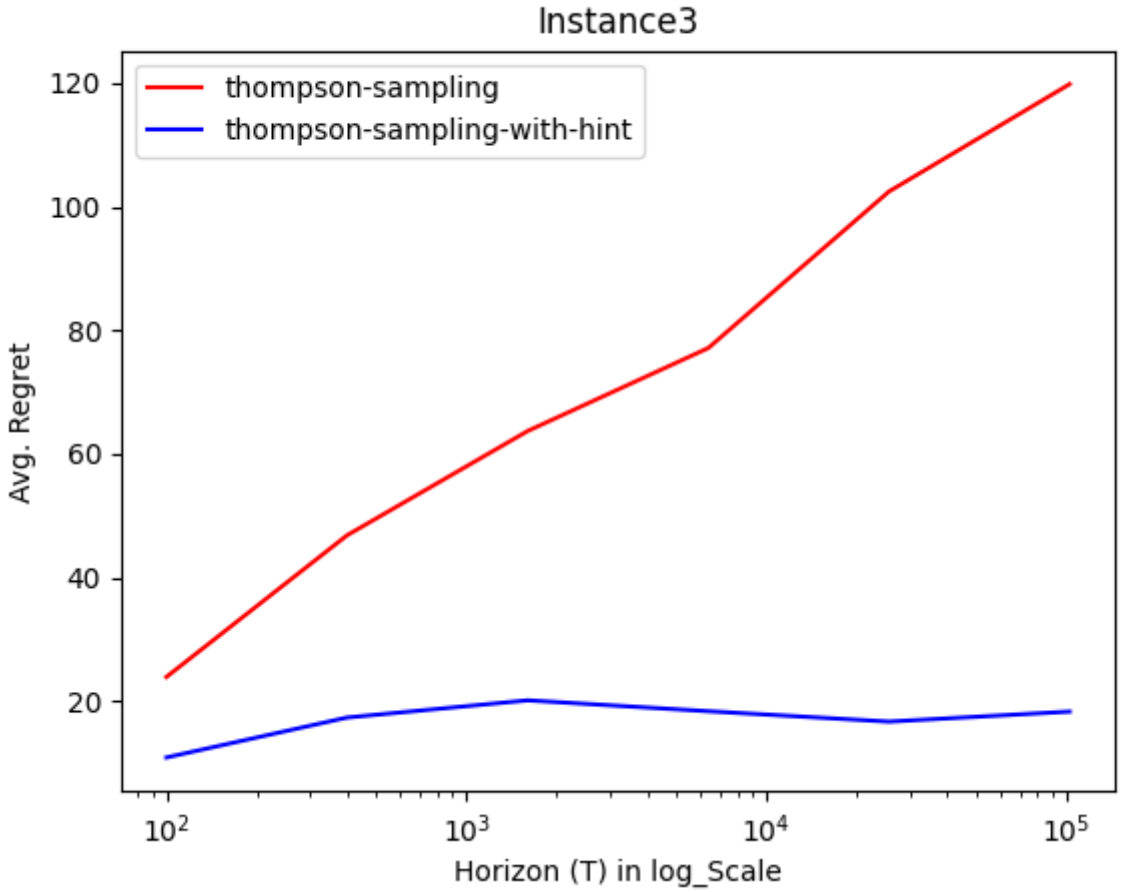
Plots of Thompson-Sampling and Thompson-Sampling-with-hint



Continued on Next Page...



Continued on Next Page...



Inferences-

- Thompson-Sampling-with-hint performs better than normal Thompson-Sampling in all the three instances.

Explanation of Inferences

- Epsilon Greedy plots are drawn for $\epsilon = 0.02$. which seems to perform good at lower horizons as this maintains a balance between exploration and exploitation.
- KL-UCB can be verified to be better than UCB.
- Thompson-Sampling and KL-UCB seem to be the optimal algorithms from the above plots as the horizon is increased.
- An exception of UCB performing worse than Epsilon Greedy can be seen for instance 3. This might be because the Epsilon Greedy algorithm determined the optimal arm during the epsilon exploration and maintained a good balance between exploration and exploitation.