# CS 747: Programming Assignment 1

Different algorithms were implemented for sampling the arms of the Multi-Armed Bandit. Each arm provides independent rewards from a Bernaulli distribution with a given mean which is different for different arms. The algorithms implemented were round-robin, epsilon-greedy (for 3 values of epsilon), UCB, KL-UCB and Thompson sampling. For each of these implementations the reward was averaged over 50 values of a random seed. This was repeated for 3 different instances of the MAB problem.

## Assumptions

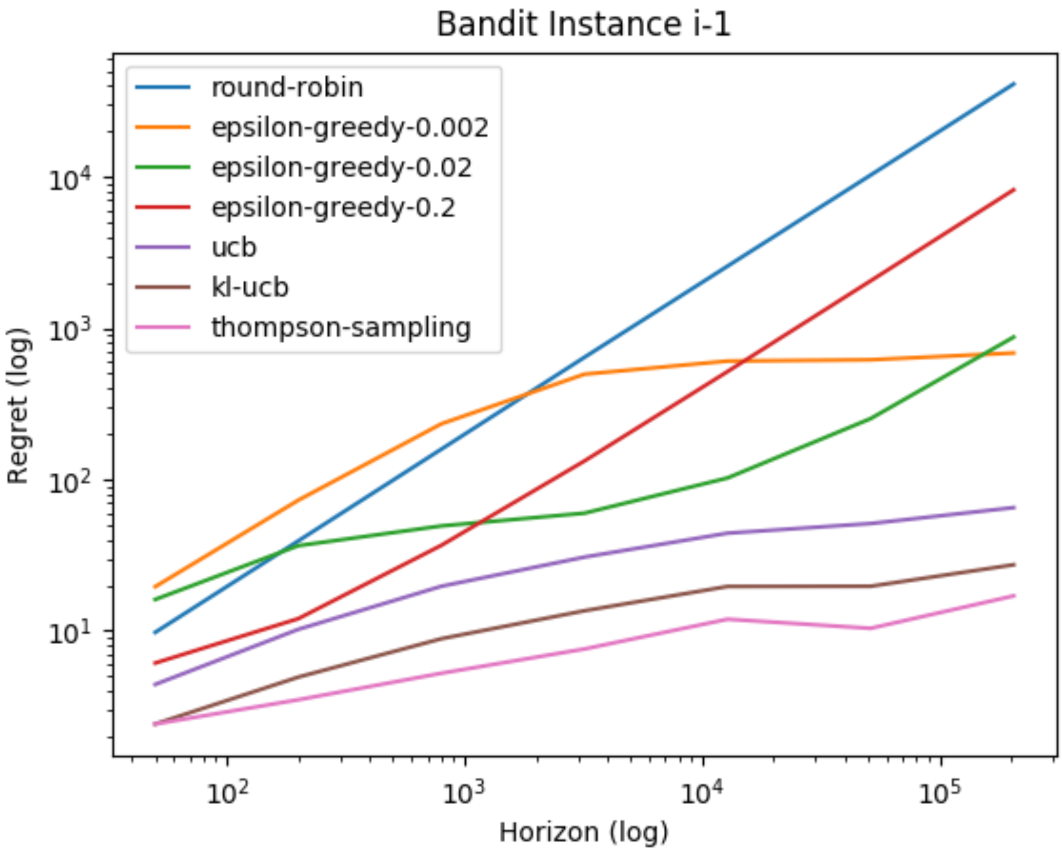
- The precision for the value of q is $10^{-6}$.
- The random seed provided as input(=randomSeed) is used for pulling the arms.
- For randomized algorithms like epsilon-greedy and Thompson Sampling, the seed provided = 2*randomSeed
- In calculating the KL Bernaulli divergence, 0 input is treated as $10^{-10}$ and 1 input is treated as $1-10^{-10}$.

## Results

The regret averaged over 50 values for the random seed is plotted against horizon. Both the average regret and horizon are taken on alogarithmic scale.

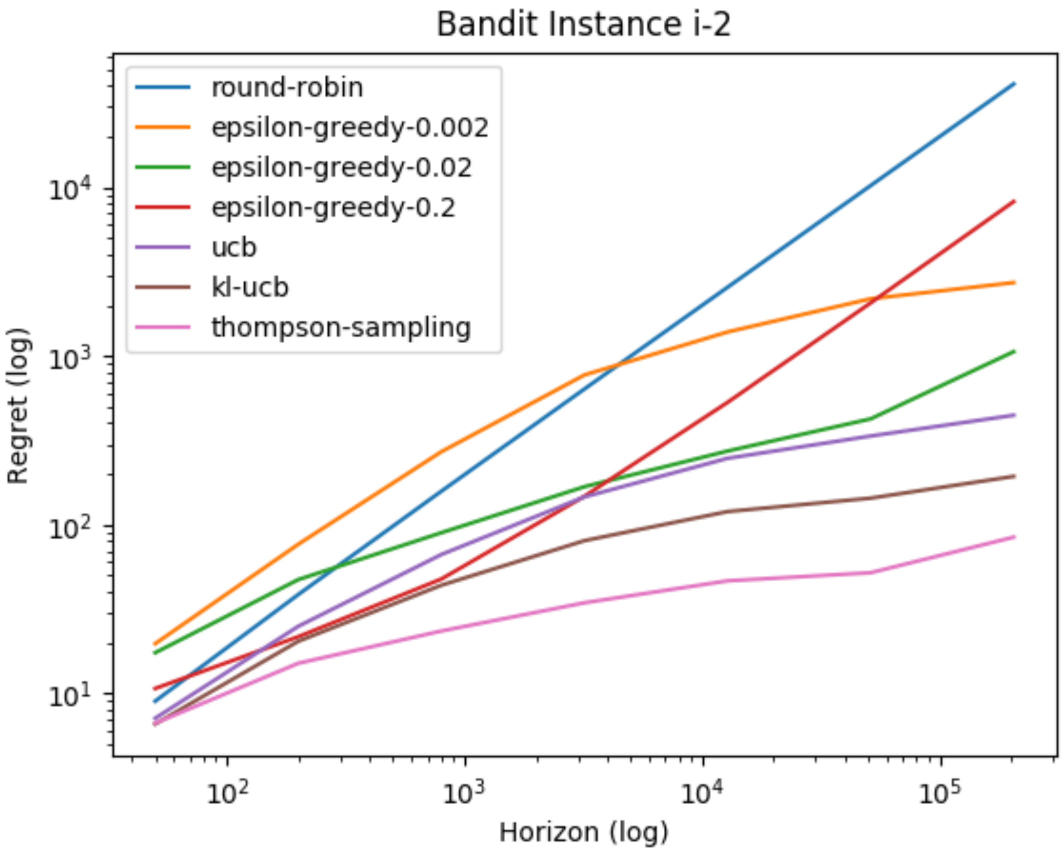
Bandit Instance i-1

- Number of arms = 2



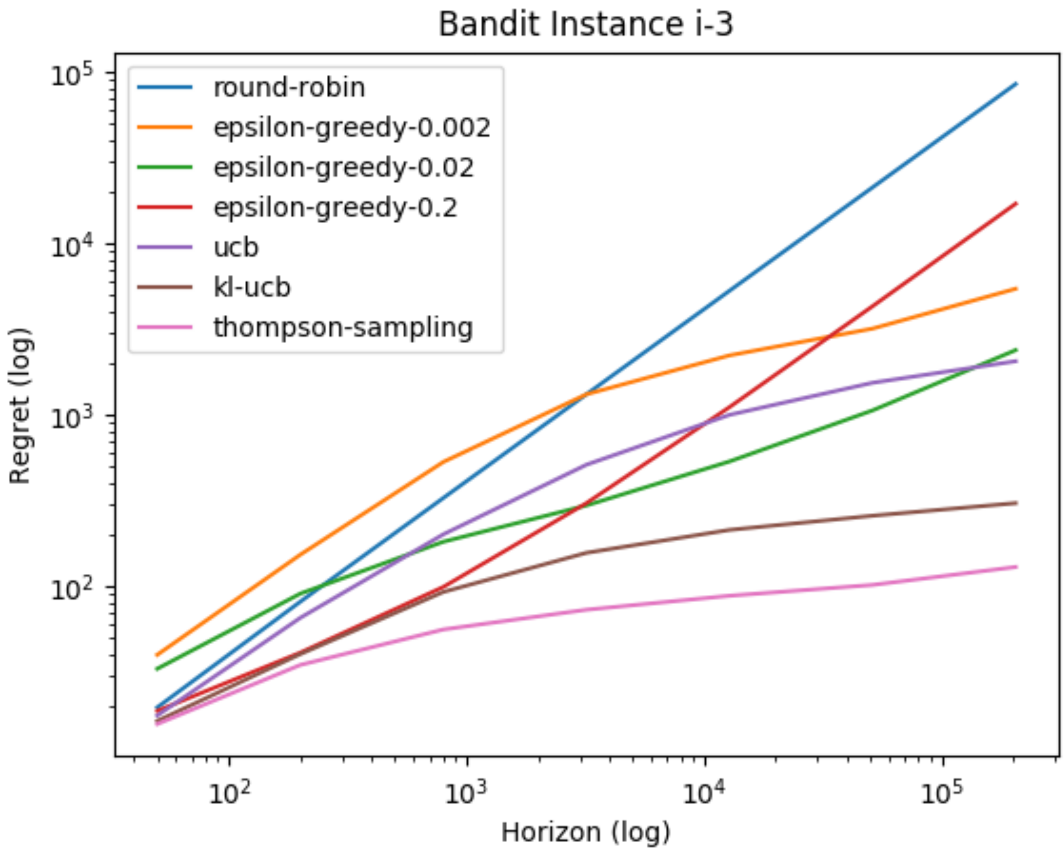
Bandit Instance i-2

- Number of arms = 5



Bandit Instance i-3

- Number of arms = 25




# Observations

---

- Thompson sampling provides the lowest regret on all horizons across all instances, with KL-UCB being the next best.
- Round robin performs the worst for large horizons, and epsilon-greedy with a low probability of exploration performs the worst for small horizons.
- Algorithm with high epsilon performs better progressively with increasing horizon.
- The regret even becomes negative for some runs of the KL-UCB, UCB and Thompson sampling algorithms.