# IE613: Online Learning - Assignment 3

**Subhadeep Chaudhuri** | 19i190010

## Question 1

**You will implement and compare different algorithms for sampling the arms of a stochastic multi-armed bandit. Each arm provides i.i.d. rewards from a Bernoulli distribution with mean given in the table below. The objective is to minimise the regret. The algorithms you will implement are epsilon-greedy exploration, UCB, KL-UCB, Thompson Sampling.**

| Arm | Arm 1 | Arm 2 | Arm 3 | Arm 4 | Arm 5 |
|---|---|---|---|---|---|
| Mean | 0.4 | 0.3 | 0.5 | 0.2 | 0.1 |

**For epsilon greedy, you can choose** $\epsilon = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.7$

**Answer:** For the epsilon greedy algorithm, we are given multiple choices of $\epsilon$. Since the value of epsilon denotes the probability of a random arm selection in a particular round (and not the best arm in terms of mean observed rewards), the smallest value of $\epsilon$ should give the best result i.e. the lowest regret.

We implement our algorithm for horizon 100000 and average out the results over 50 runs

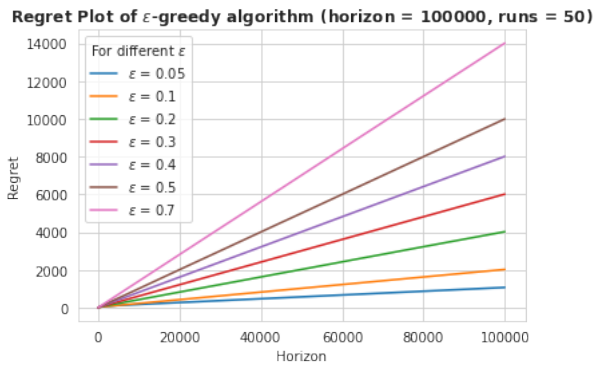


Figure 1: Regret plot for epsilon greedy algorithm for different $\epsilon$

```
For epsilon = 0.05:
Average counts of each arm (L-R: arm1 - arm5): [ 1701.96  1005.78 95295.1    999.12   998.04]
Average regret after algorithm runs till horizon: 1070.3039999999762
--------------------------------------------------------------------------------
For epsilon = 0.1:
Average counts of each arm (L-R: arm1 - arm5): [ 2136.92  2034.72 91820.68  2003.64  2004.04]
Average regret after algorithm runs till horizon: 2023.3440000000048
--------------------------------------------------------------------------------
For epsilon = 0.2:
Average counts of each arm (L-R: arm1 - arm5): [ 4172.28  4040.1  83792.08  4004.22  3991.32]
Average regret after algorithm runs till horizon: 4023.0419999999676
--------------------------------------------------------------------------------
For epsilon = 0.3:
Average counts of each arm (L-R: arm1 - arm5): [ 6106.34  5998.84 75880.08  6021.88  5992.86]
Average regret after algorithm runs till horizon: 6014.110000000007
--------------------------------------------------------------------------------
For epsilon = 0.4:
Average counts of each arm (L-R: arm1 - arm5): [ 8094.62  8018.   67880.36  8016.58  7990.44]
Average regret after algorithm runs till horizon: 8014.212000000011
--------------------------------------------------------------------------------
For epsilon = 0.5:
Average counts of each arm (L-R: arm1 - arm5): [10040.92  9980.14 59984.52 10002.12  9992.3 ]
Average regret after algorithm runs till horizon: 9997.675999999987
--------------------------------------------------------------------------------
For epsilon = 0.7:
Average counts of each arm (L-R: arm1 - arm5): [14023.84 13998.92 43936.18 14024.92 14016.14]
Average regret after algorithm runs till horizon: 14016.100000000011
```

Figure 2: Regret value and average arm pulls observed from epsilon greedy algorithm for different $\epsilon$

**Comments:** From the plot, we observe that the minimum regret is attained by the algorithm for $\epsilon = 0.05$ and the highest regret when $\epsilon = 0.7$. Hence, the results are in line with the expected results, thus validating the accuracy of the algorithm implementation.

Now, we compare the regrets of the 4 algorithms under study - epsilon-greedy exploration, UCB, KL-UCB and Thompson Sampling.

The expected result in terms of regret is as follows:

**Thompson Sampling < KL-UCB < UCB < epsilon-greedy exploration.**

We implement our algorithm for horizon 100000 and average out the results over 50 runs. The obtained results are as follows:

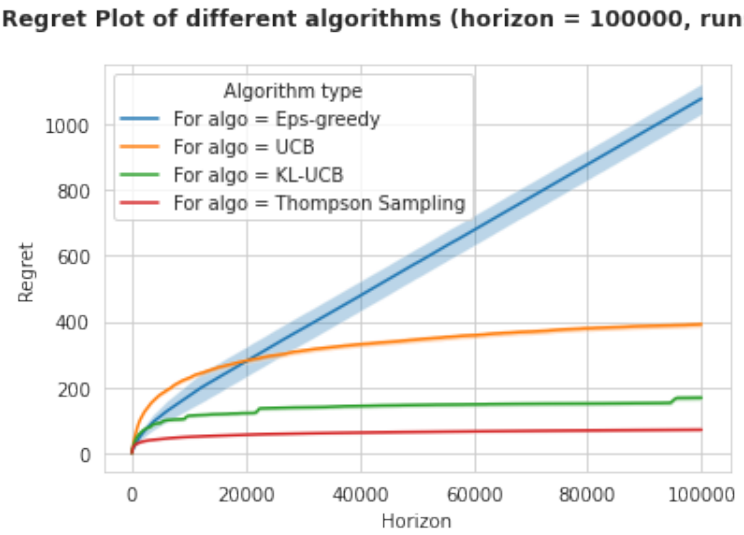


Figure 3: Regret plot for different algorithms

```
For algo type = Eps-greedy:
Average counts of each arm (L-R: arm1 - arm5): [ 1734.68  1019.7  95241.18  1006.24   998.2 ]
Average regret after algorithm runs till horizon: 1078.5599999999895
--------------------------------------------------------------------------------------------
For algo type = UCB:
Average counts of each arm (L-R: arm1 - arm5): [ 1739.28   486.06 97413.9    230.46   130.3 ]
Average regret after algorithm runs till horizon: 392.3979999995687
--------------------------------------------------------------------------------------------
For algo type = KL-UCB:
Average counts of each arm (L-R: arm1 - arm5): [7.966000e+02 2.064000e+02 9.885402e+04 9.460000e+01 4.838000e+01]
Average regret after algorithm runs till horizon: 168.671999999798
--------------------------------------------------------------------------------------------
For algo type = Thompson Sampling:
Average counts of each arm (L-R: arm1 - arm5): [3.216800e+02 9.050000e+01 9.952558e+04 3.862000e+01 2.362000e+01]
Average regret after algorithm runs till horizon: 71.30199999994903
```

Figure 4: Regret value and average arm pulls observed from different algorithms

**Comments:** From the plot, we observe that the minimum regret is attained by the Thompson Sampling algorithm, and the highest regret from the epsilon greedy algorithm (with $\epsilon = 0.05$). The average values of number of pulls of each arm also shows the expected behaviour of the 4 algorithms. Hence, the results are in line with the expected results, thus validating the accuracy of the algorithm implementations.