Reinforcement Learning

# Lab Assignment 2
# 190020007-190020021-190020039

## Experimental setup:

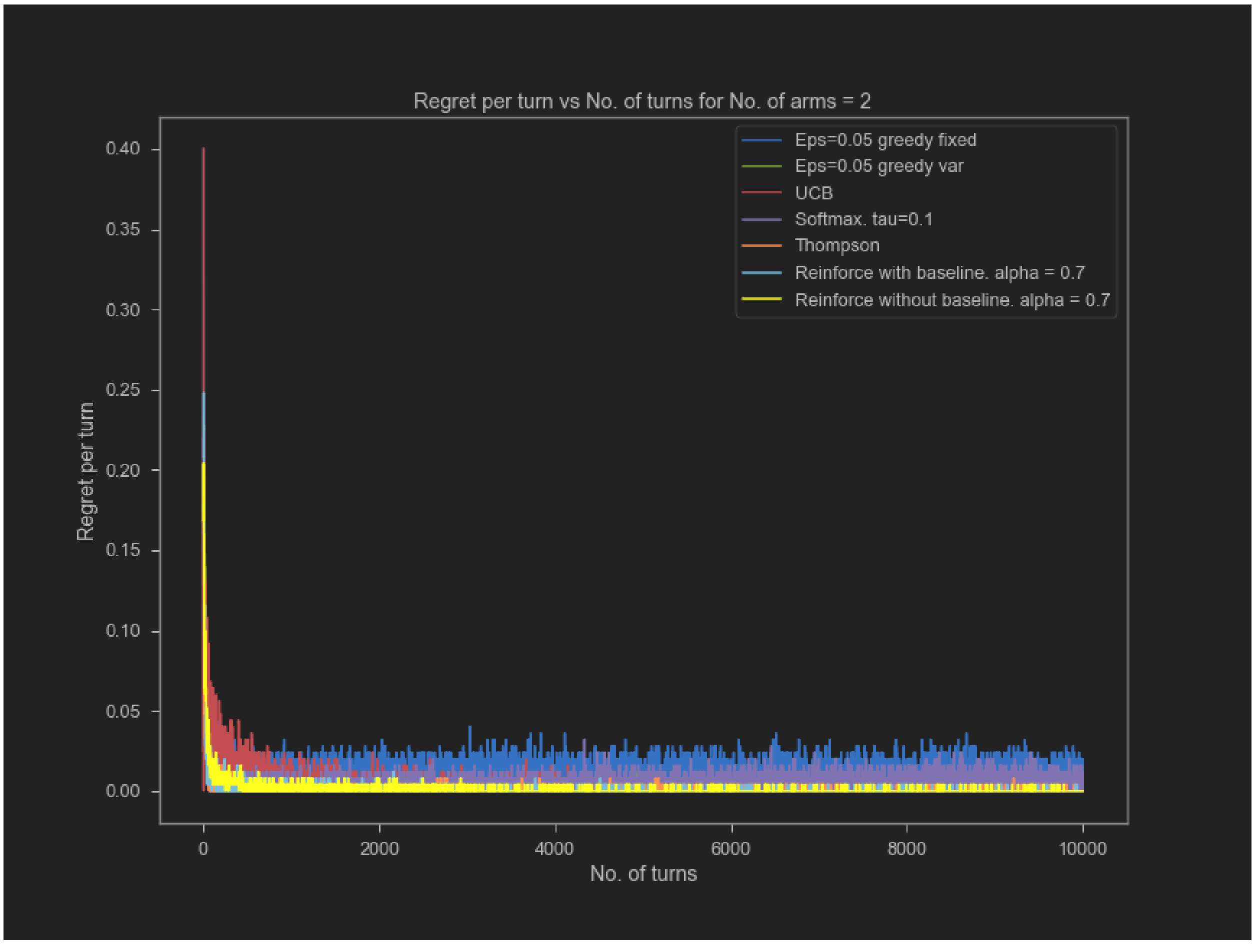
- First of all, for each of the assigned algorithms we test for each number of arms i.e K= 2,5,10 different hyper-parameter values to find out the most optimal performance from the chosen set of hyper-parameters.
- For the same we plot three metrics which are
    - Cumulative regret vs number of turns
    - Regret per turn vs number of turns
    - Percentage of times optimal arm picked vs number of turns
- Next step is to test for a different number of arms for which we have considered 2,5,10.
- Once we have got our optimal choice for different hyper-parameters we again use metrics Cumulative regret vs number of turns, Regret per turn vs number of turns, Percentage of times optimal arm picked vs number of turns to compare different algorithms for the optimal hyper-parameters and for different number of arms(2,5,10).
- These all tests have been done for bernoulli reward distribution.
- Also each algorithm was run for 10,000 steps and then results were averaged for 100 independent runs of the algorithms and graphs were plotted for the average of all the simulations/runs.
- Test bed is similar to one used in the tutorial paper provided.
- We have used cumulative regret also to visualize the order of regrets.
- Our observations are summarized at the end of the report.

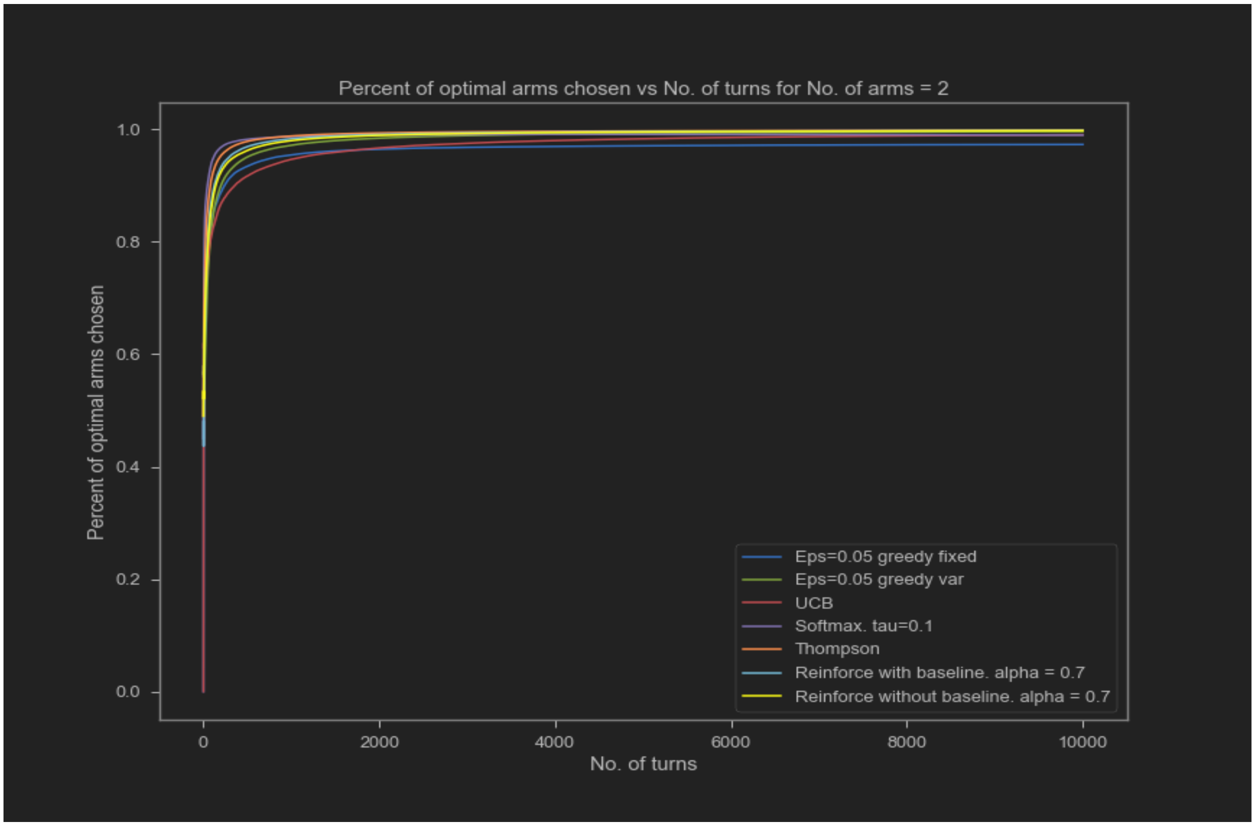
# Comparison of Algorithms:

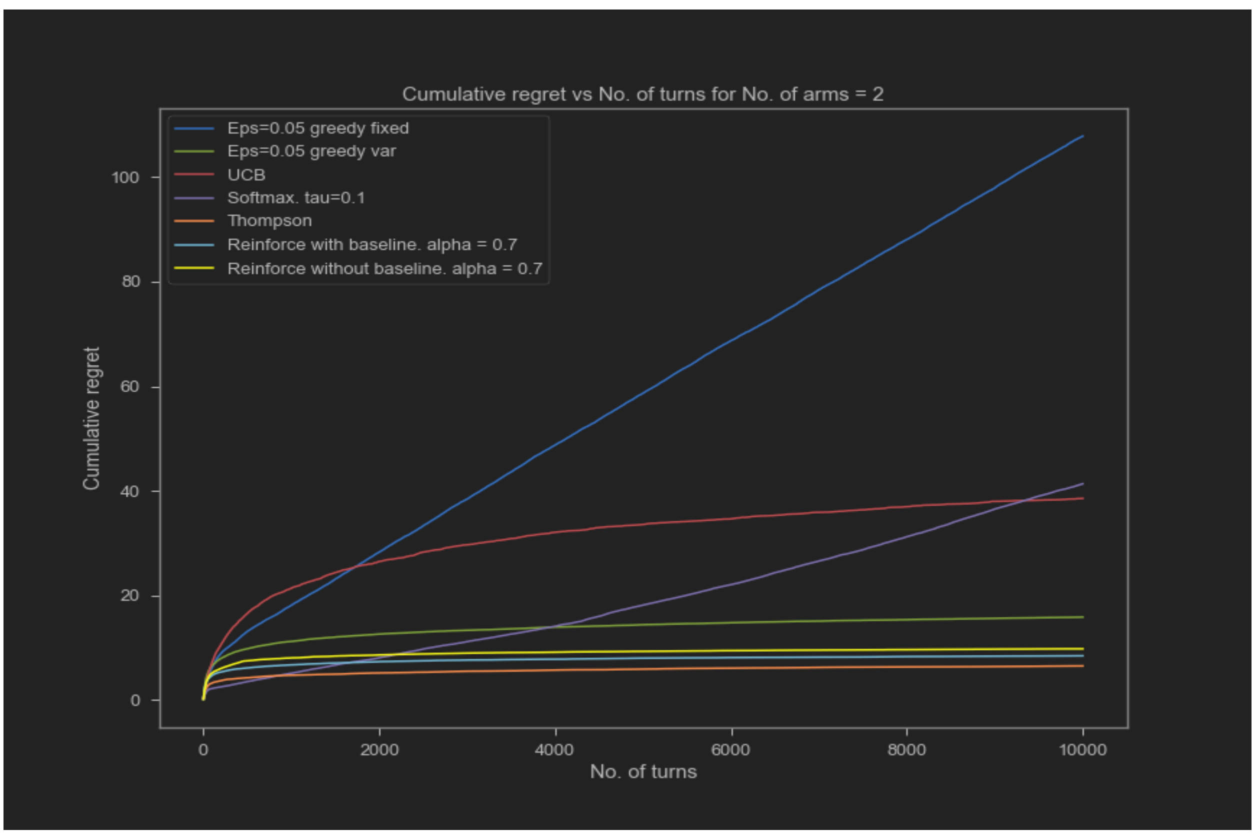
## Setting :

Arms = 2, true mean = [0.4,0.8]

number of turns = 10,000, number of simulations =100



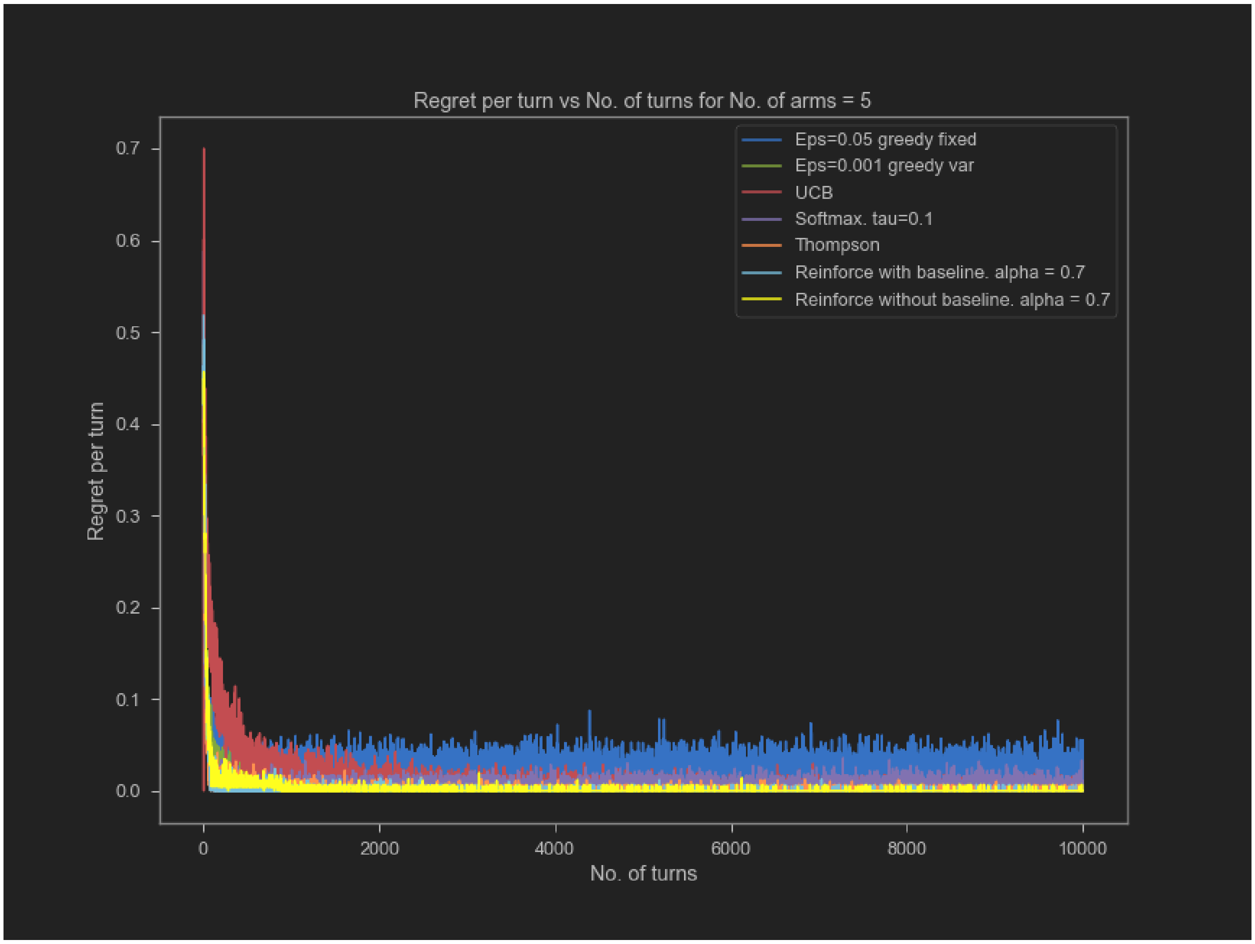

Percent of optimal arms chosen vs No. of turns for No. of arms = 2
Percent of optimal arms chosen
No. of turns
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000
Eps=0.05 greedy fixed
Eps=0.05 greedy var
UCB
Softmax. tau=0.1
Thompson
Reinforce with baseline. alpha = 0.7
Reinforce without baseline. alpha = 0.7

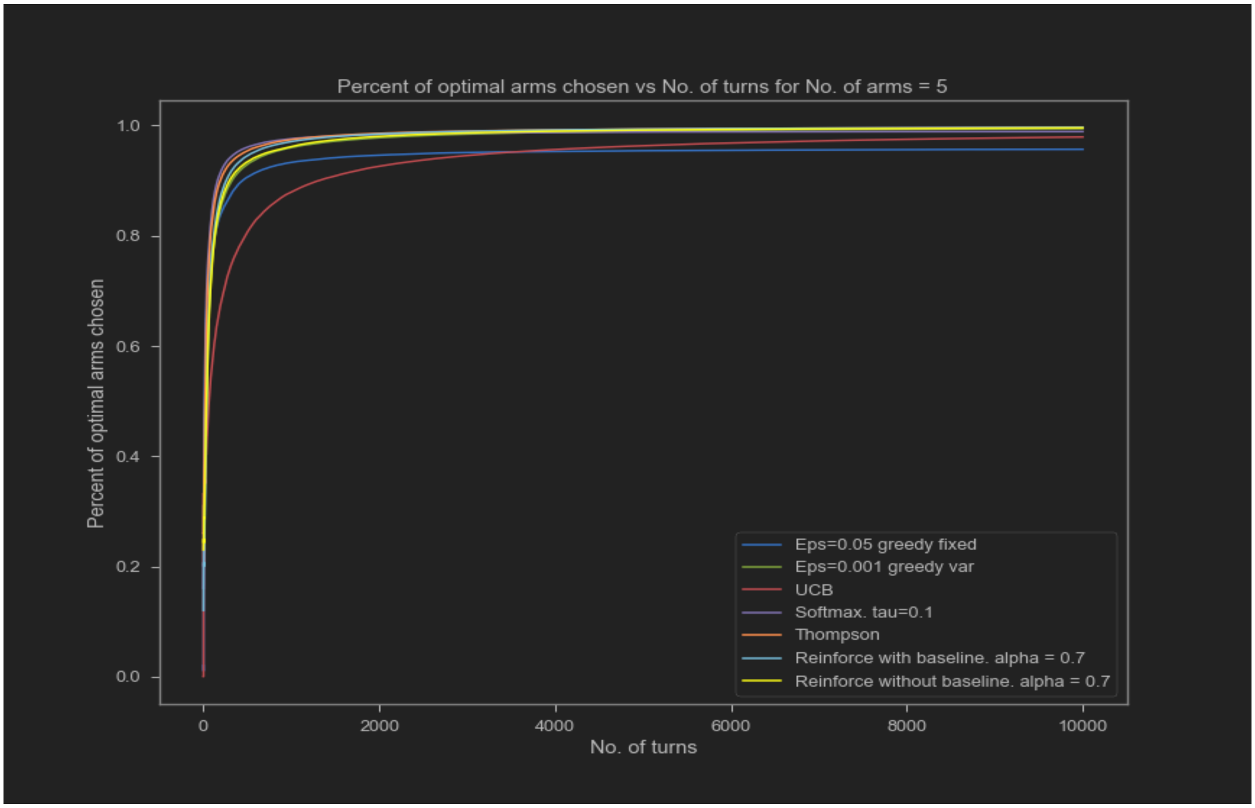

Cumulative regret vs No. of turns for No. of arms = 2
Cumulative regret
No. of turns
100
80
60
40
20
0
0
2000
4000
6000
8000
10000
Eps=0.05 greedy fixed
Eps=0.05 greedy var
UCB
Softmax. tau=0.1
Thompson
Reinforce with baseline. alpha = 0.7
Reinforce without baseline. alpha = 0.7

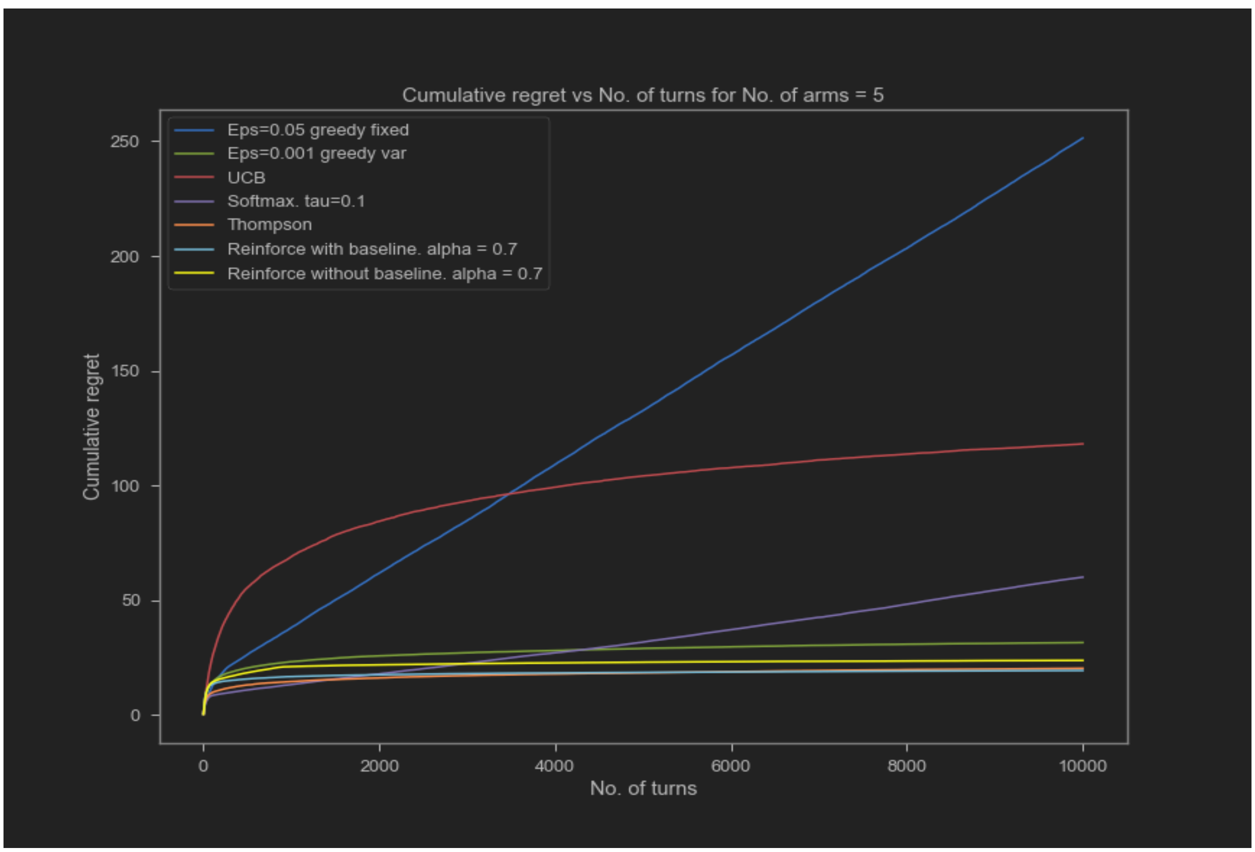
## Setting -

Arms = 5, true mean = [0.2,0.3,0.8,0.25,0.1]

number of turns = 10,000,number of simulations = 100



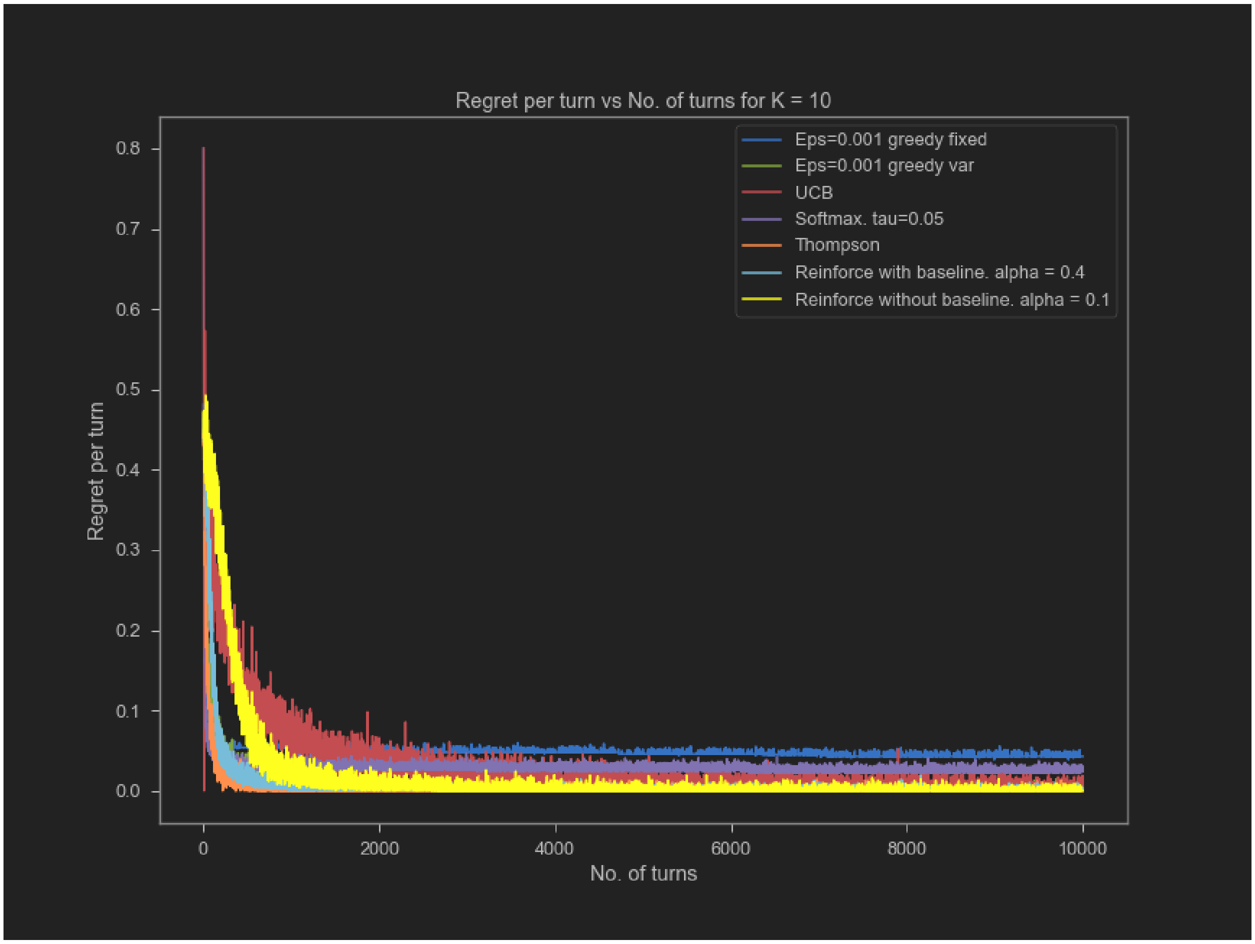

Percent of optimal arms chosen vs No. of turns for No. of arms = 5
Percent of optimal arms chosen
No. of turns
1.0
0.8
0.6
0.4
0.2
0.0
0
2000
4000
6000
8000
10000
Eps=0.05 greedy fixed
Eps=0.001 greedy var
UCB
Softmax. tau=0.1
Thompson
Reinforce with baseline. alpha = 0.7
Reinforce without baseline. alpha = 0.7

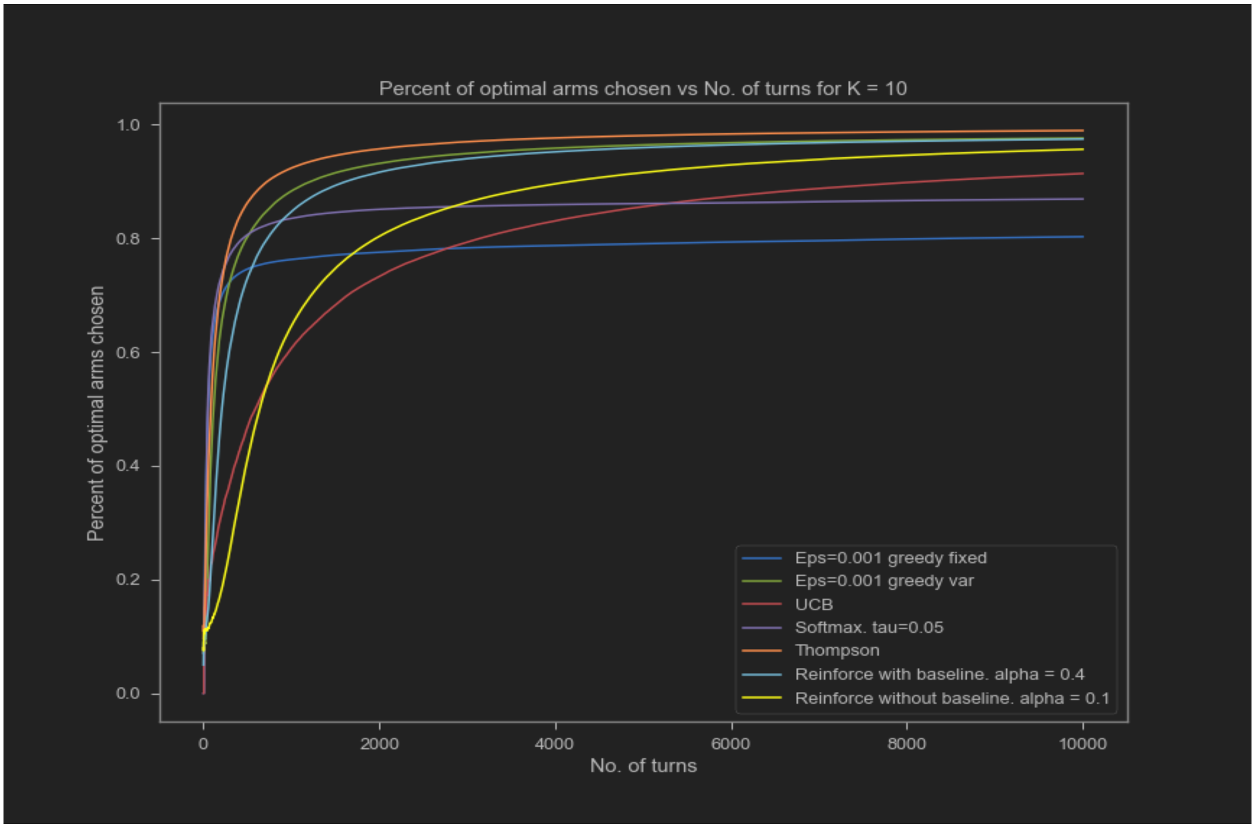

Cumulative regret vs No. of turns for No. of arms = 5
Cumulative regret
No. of turns
250
200
150
100
50
0
0
2000
4000
6000
8000
10000
Eps=0.05 greedy fixed
Eps=0.001 greedy var
UCB
Softmax. tau=0.1
Thompson
Reinforce with baseline. alpha = 0.7
Reinforce without baseline. alpha = 0.7

## Setting -

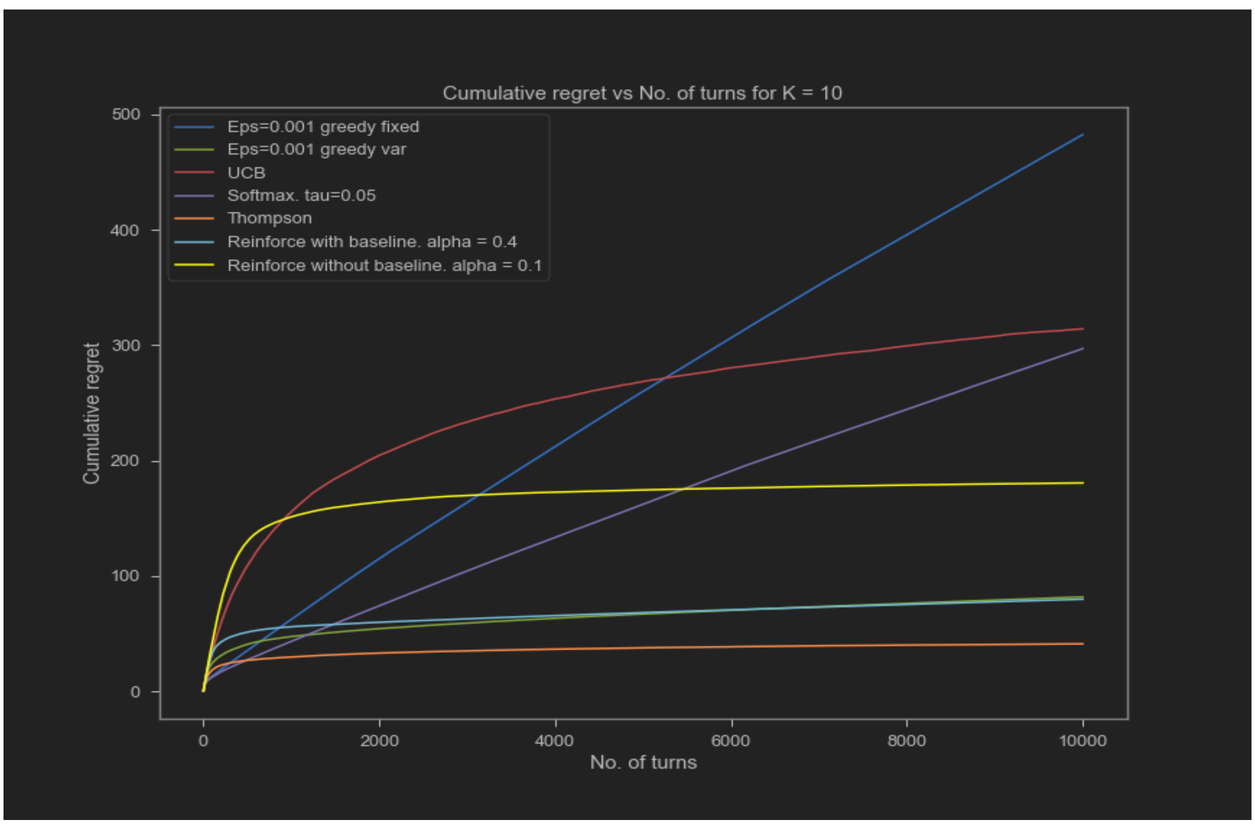
Arms = 10, true mean = [0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.9,0.45,0.35]

number of turns = 10,000,number of simulations = 100





Percent of optimal arms chosen vs No. of turns for K = 10
Percent of optimal arms chosen
No. of turns
0.0
0.2
0.4
0.6
0.8
1.0
0
2000
4000
6000
8000
10000
Eps=0.001 greedy fixed
Eps=0.001 greedy var
UCB
Softmax. tau=0.05
Thompson
Reinforce with baseline. alpha = 0.4
Reinforce without baseline. alpha = 0.1

Cumulative regret vs No. of turns for K = 10
Cumulative regret
No. of turns
0
100
200
300
400
500
0
2000
4000
6000
8000
10000
Eps=0.001 greedy fixed
Eps=0.001 greedy var
UCB
Softmax. tau=0.05
Thompson
Reinforce with baseline. alpha = 0.4
Reinforce without baseline. alpha = 0.1

# Observations:

Arms = 2:

| Algorithm | Total regret |
|---|---|
| Epsilon Greedy FIxed (eps = 0.05) | 107.828 |
| Epsilon Greedy Variable (eps = 0.05) | 15.880 |
| Softmax (tau = 0.1) | 41.372 |
| UCB | 38.572 |
| Thompson Sampling | 6.548 |
| REINFORCE (with Baseline) | 8.484 |
| REINFORCE(without Baseline) | 9.824 |

Arms = 5:

| Algorithm | Total regret |
|---|---|
| Epsilon Greedy FIxed (eps = 0.05) | 251.393 |
| Epsilon Greedy Variable (eps = 0.001) | 31.633 |
| Softmax(tau = 0.1) | 60.125 |
| UCB | 118.183 |
| Thompson Sampling | 20.415 |
| REINFORCE (with Baseline) | 19.446 |
| REINFORCE(without Baseline) | 23.867 |

Arms = 10:

| Algorithm | Total regret |
|---|---|
| Epsilon Greedy FIxed (eps = 0.001) | 482.601 |
| Epsilon Greedy Variable (eps = 0.001) | 81.874 |
| Softmax(tau = 0.1) | 297.061 |
| UCB | 314.221 |
| Thompson Sampling | 41.209 |
| REINFORCE (with Baseline) | 79.878 |
| REINFORCE(without Baseline) | 180.664 |

**Bernoulli Reward Distribution:**

- **Increase in number of arms:**

  Thompson > REINFORCE(with baseline > REINFORCE (with Baseline) > UCB > Softmax > Epsilon Greedy Variable > Epsilon Greedy Fixed

  This is the order we observe for mostly all arm settings.

  The point to be noted is simple algorithms like softmax and eps-greedy (variable) for arms K = 10 perform at par with policy based algorithms like REINFORCE.

- **Change in true mean value of arms:**

  When means are well separated:

  Algorithms are able to find optimal arm much quicker and also regrets are less

  When means are far from each other:

  Algorithms struggle to find optimal arm and average percent times optimal arm picked till T round decreases.

  Thompson and REINFORCE were not affected much by variation in true mean values.

- **Regret:**

  Logarithmic Regret : Thompson, UCB, REINFORCE(with baseline),REINFORCE(without baseline),Epsilon greedy variable

Linear Regret : Epsilon greedy Fixed, Softmax

Softmax with high values tau parameters perform linear whereas for lower tau values perform logarithmic kind.

These match from what we had seen in the lectures.