Reinforcement Learning **(Date:02-02-2022)**

Lab Assignment 2

Instructor: Dr. Prabhuchandran K J *By:* 211022001, 211022002, 211022005

1 Aim

To compare the performance of different bandits algorithms for Bernoulli and Normal reward distribution.

1.1 ϵ - Greedy Algorithm

In ϵ-greedy method we we choose both arm equally likely with a probability of ϵ and we will choose the arm with maximum expected reward with a probability of $1-\epsilon$. For ϵ_t greedy it chooses $\epsilon = min(1, \frac{cK}{d^2N})$ where c is tunable and K is number of arms, d is difference of mean of best and second best arm. N is number of plays and it increases as we proceed with plays.

1.2 Upper Confidence Bound

This Algorithm picks the arm with maximum value and also balances the exploration-exploitation task by considering less played arms which could give promising returns.

1.3 Thompson Sampling

It starts sampling from $Beta(1,1)$ and after pulls it updates it's beliefs to $Beta(W_{i,t}+1, L_{i,t}+1)$. In Bernoulli case $W_{i,t}$ is number of 1's in $n_{i,t}$ pulls of arm i and $L_{i,t}$ is number of 0's in $n_{i,t}$ pulls of arm i.This samples from the beta distribution and pulls arm with highest sample.

1.4 Reinforce

1.5 Settings

- For $\epsilon-$Greedy $\epsilon = 0.1$, Total plays= $10k$
- Arm winning probabilities For K=2, {0.50,0.57}
- Arm wining probabilities for K=5, {0.50,0.57, 0.32, 0.25,0.2}
- Arm winning probabilities for K=10, {0.50,0.57, 0.32, 0.25,0.2,0.45,0.35, 0.20, 0.31,0.1}
- For ϵ_t greedy, $c = 0.2$. It is observed that more the c value greater is exploration
- For softmax, temperature was set to 0.1.
- As per observation there is a trade-off between ϵ_t and Thompson for the least cumulative regret while softmax gave highest cumulative regret among all in this setting.
- For normal distribution all the bandits have a mean $\mu = \mu_i$ and $\sigma = 1$.
    1. Mean for K=2, μ={0.50,0.57}
    2. Mean for K=5, μ={0.50,0.57, 0.32, 0.25,0.2}
    3. Mean for K=10, μ= {0.50,0.57, 0.32, 0.25,0.2,0.45,0.35, 0.20, 0.31,0.1}

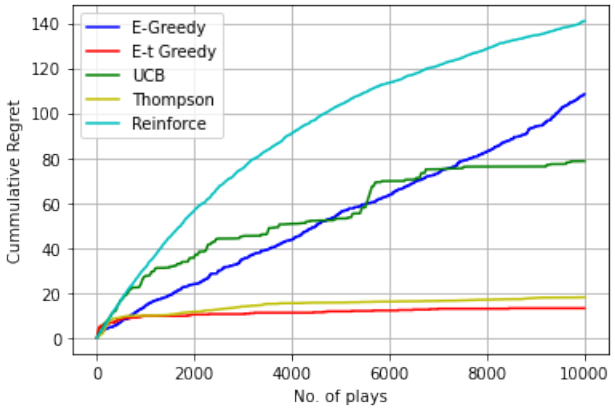
- Mean for K=5, μ={0.50,0.57, 0.32, 0.25,0.2}
- μ= for K=10, {0.50,0.57, 0.32, 0.25,0.2,0.45,0.35, 0.20, 0.31,0.1}
- For n=10 turns.

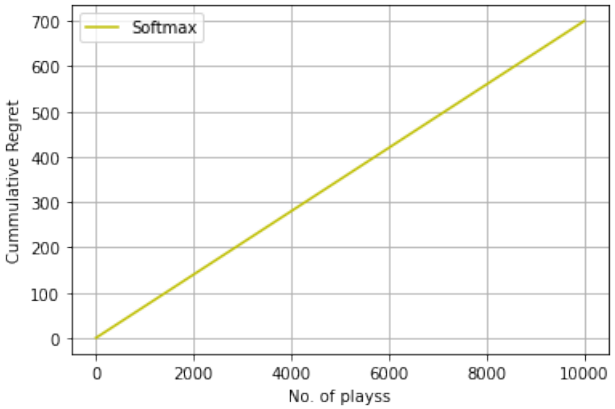
2 Results

Simulation results with the above given parameters:

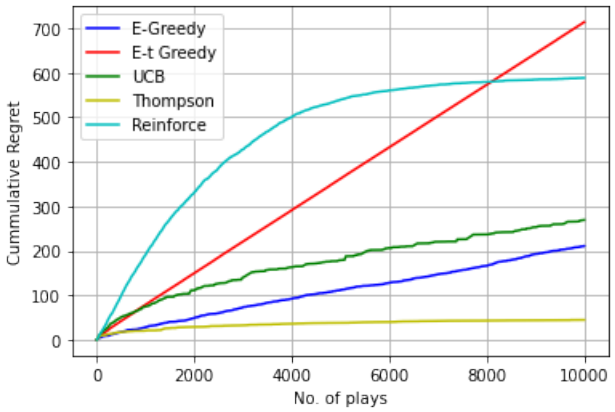
2.1 Bernoulli Distribution

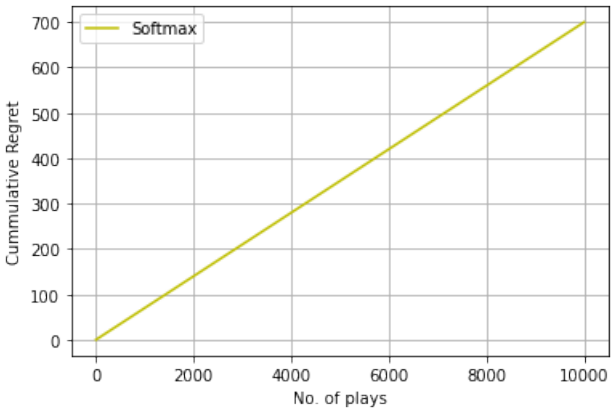
- K=2



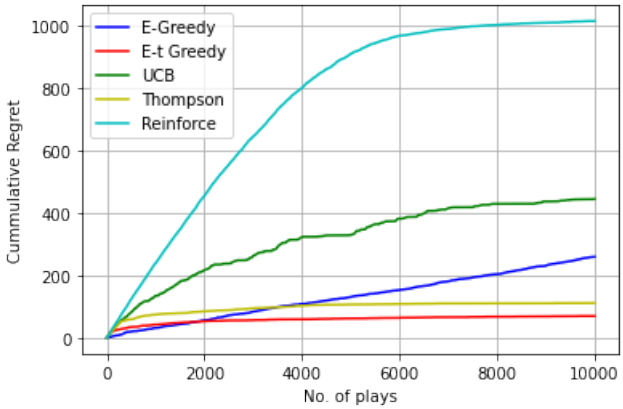


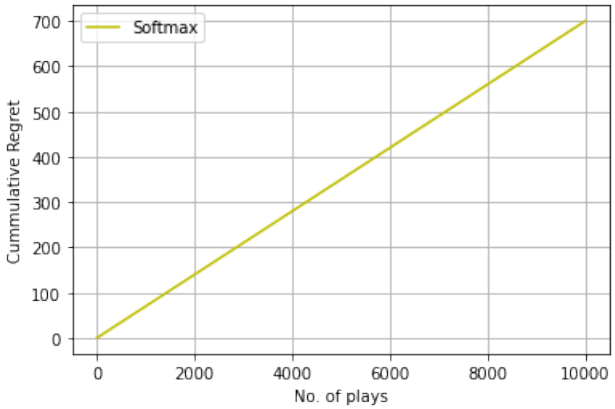
- For K=5





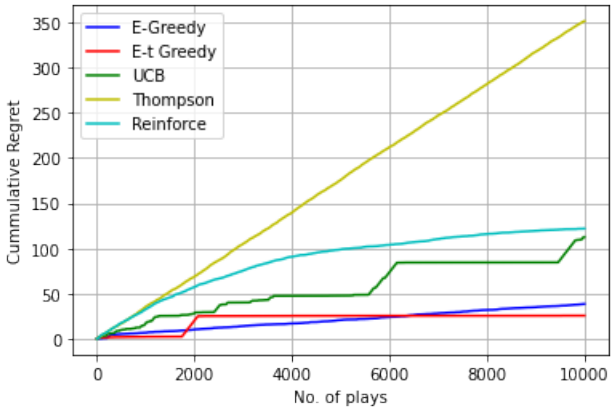
- For K=10

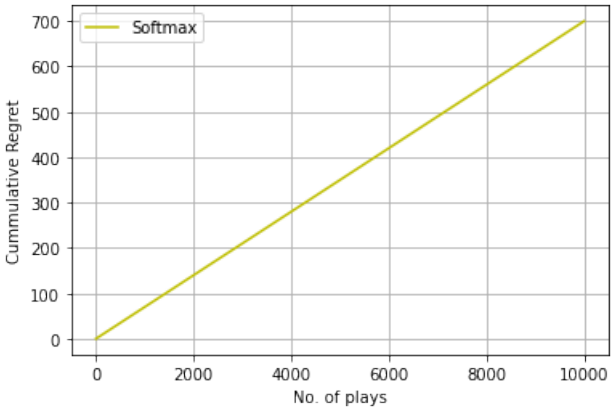




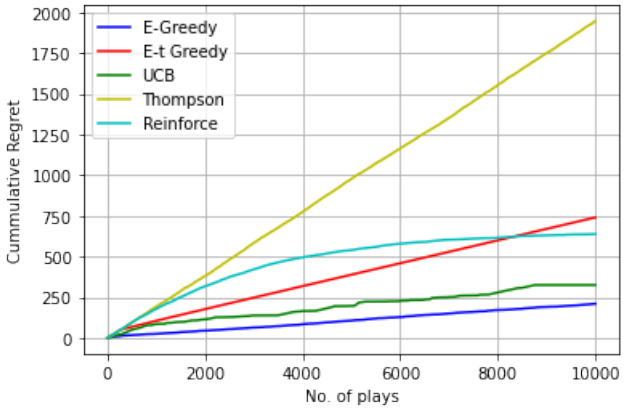
2.2 Normal Distribution

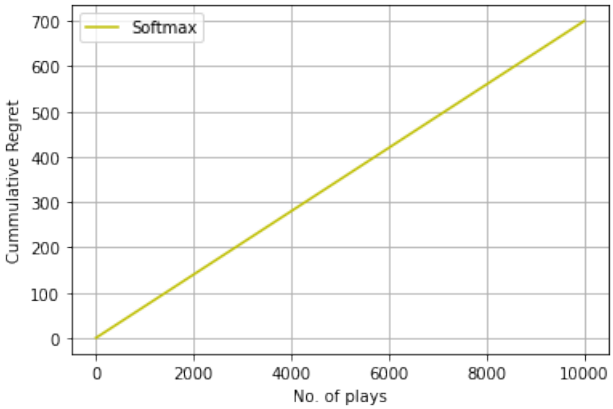
- K=2



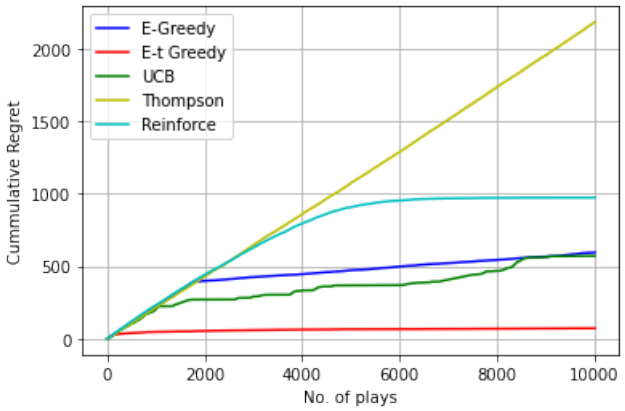


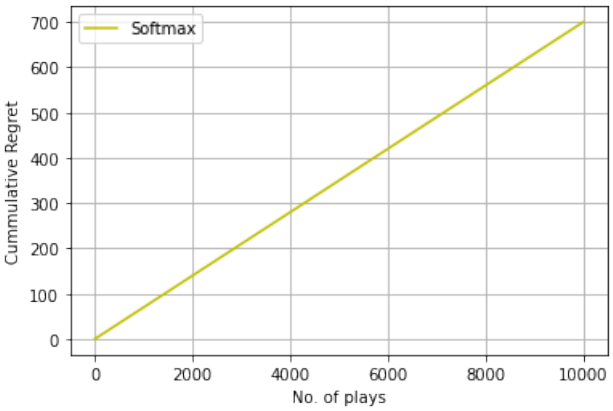
- For K=5





- For K=10





2.3 Observations

- For the Bernoulli Distribution Thomspon sampling gives the consistent performance for all K=2,K=5,K=10.
- For the Normal Distribution ϵ-t Greddy and ϵ-Greedy Performs well.