Reinforcement Learnig: Homework 2

Raphaël Avalos

November 26, 2018

1 Question 1

The following two Bernouili Bandit problem cosnidered are represented in the table bellow. The following plots are an average of 1000 simulations with $\rho = 0.2$.

	Arm 0	Arm 1	Arm 2	Arm 3
First Problem $p =$	0.65	0.5	0.45	0.6
Second Problem $p =$	0.43	0.56	0.51	0.55

Figure 1: MAB 1

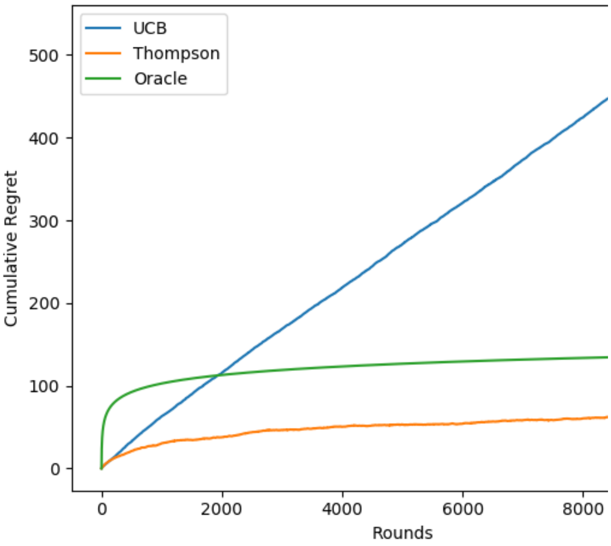
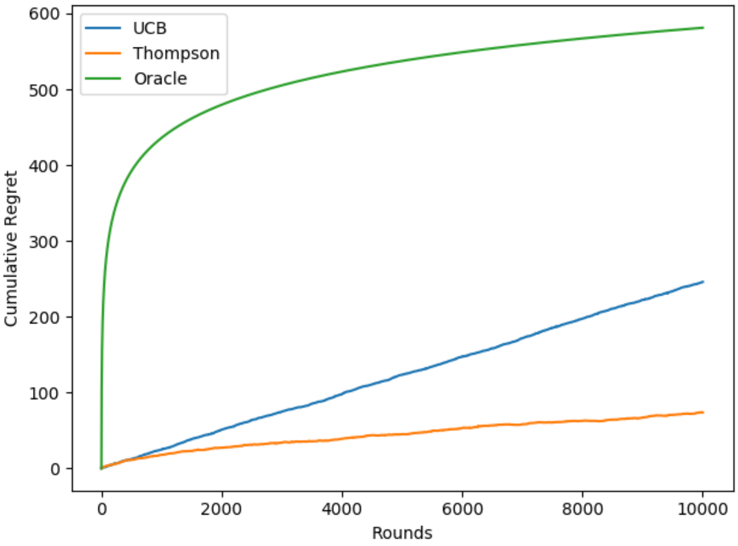


Figure 2: MAB 2



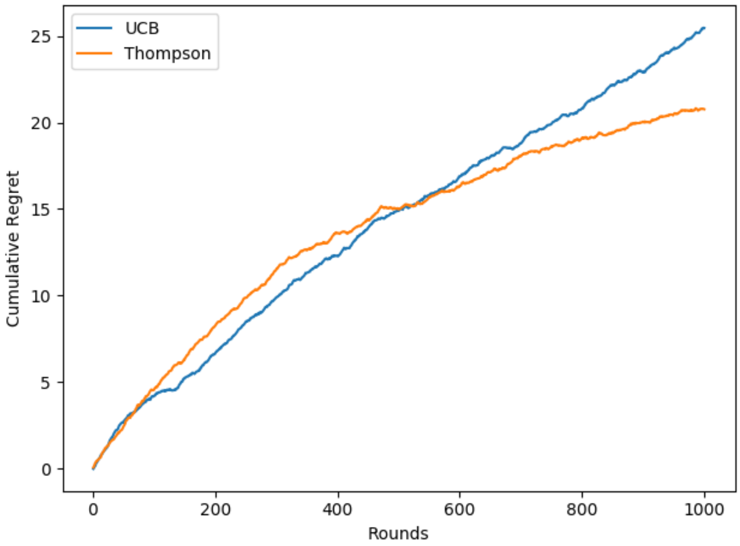
2 Question 2

The Thomson algorithm has been adapted in the following way.

```
arm = np.random.beta(S + 1, F + 1).argmax()
reward = bandits[arm].sample()
draws[t] = arm
N[arm] += 1
    if np.random.random() < reward:
        S[arm] += 1
    else:
        F[arm] += 1
```

	Arm 0	Arm 1	Arm 2	Arm 3
Parameters	$\mathcal{B}(0.7, 0.6)$	$\mathcal{B}(0.5, 0.6)$	$\mathbf{Exp}(0.7)$	$\mathbf{Exp}(0.35)$

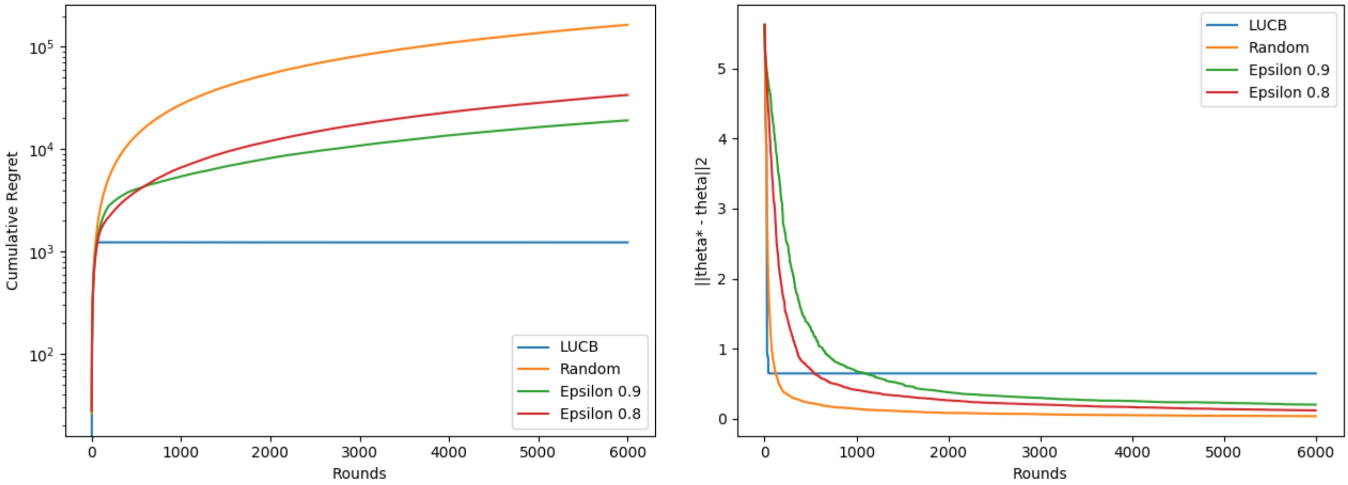
Figure 3: NPM



3 Question 3

In LinearUCB the parameter α affects the exploration. I choose $\alpha_0 = 100$ and $\alpha_{t+1} = max(0, \alpha_t - 1)$ and $\lambda = 1$. The plots are the average over 30 runs with 6000 epochs.
LinearUCB provides the minimal cummulative regret by sacrifacing exploration thereore it doen't compute a great approximation of θ.

Figure 4: LineraUCB vs Random vs Greedy



Cumulative Regret
10^5
10^4
10^3
10^2
0
1000
2000
3000
4000
5000
6000
Rounds
LUCB
Random
Epsilon 0.9
Epsilon 0.8
||theta* - theta||2
5
4
3
2
1
0
0
1000
2000
3000
4000
5000
6000
Rounds
LUCB
Random
Epsilon 0.9
Epsilon 0.8