Reinforcement Learning - Homework 2

Mehdi Boubnan

November 26, 2018

1 Stochastic Multi-Armed Bandits on Simulated Data

1.1 Bernoulli bandit models

For the UCB1 algorithm, we take : $\rho_t = \frac{\rho}{t^\gamma}$ with γ the decay factor.

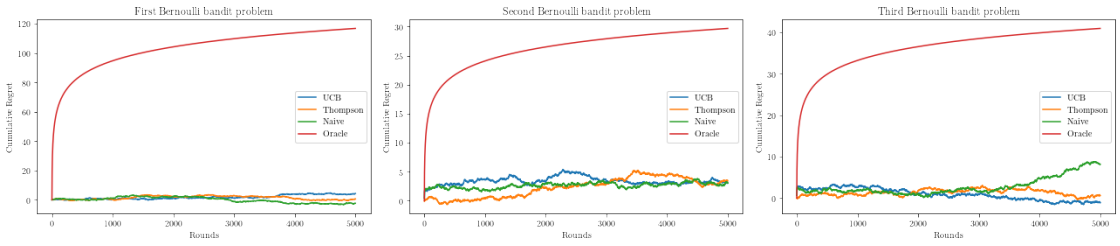
1.1.1 Question 1 :

We'll compare the regret of Thompson Sampling with that of UCB1, for two different Bernoulli bandit problems.

The two Bernoulli bandit problem are the following :

- Four Bernoulli Arms : $\mathcal{B}(0.3)$, $\mathcal{B}(0.25)$, $\mathcal{B}(0.2)$ and $\mathcal{B}(0.1)$
- Four Bernoulli Arms : $\mathcal{B}(0.42)$, $\mathcal{B}(0.2)$, $\mathcal{B}(0.35)$ and $\mathcal{B}(0.15)$

We'll consider a horizon T=5000 for the simulations. We run 50 simulations. We'll also consider the decay $\gamma = 0.5$ and $\rho = 5$.



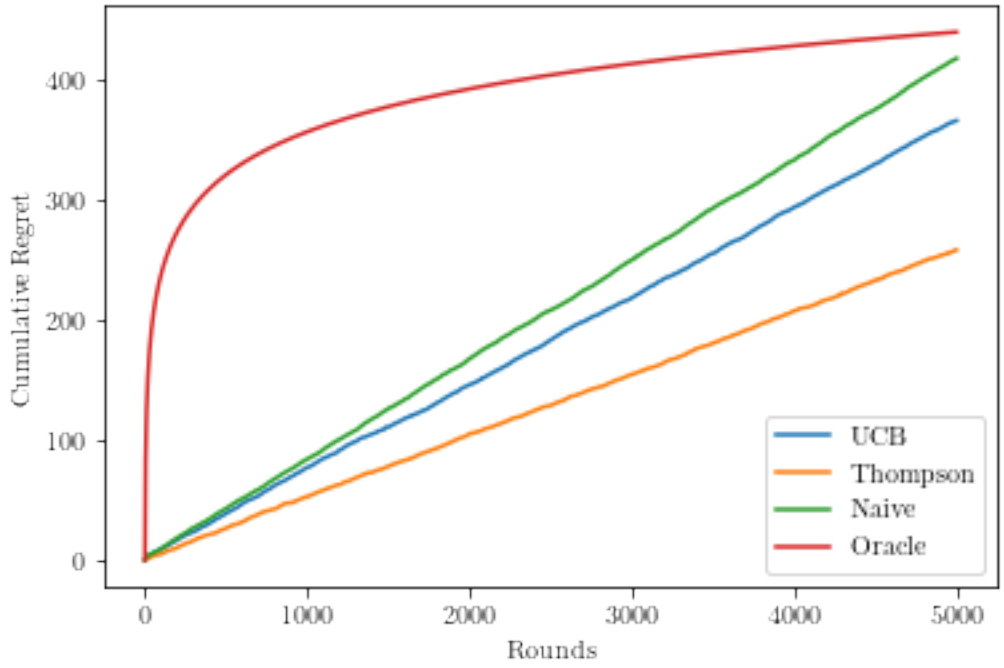
1.2 Non-parametric bandits (bounded rewards)

1.2.1 Question 2 :

We'll build a MultiArmedBandit model with not only Bernoulli arms, and observe the cumulative rewards for the three algorithms. For that purpose we'll use an adaptation of Thompson Sampling to handle non-binary rewards, proposed by Agrawal and Goyal, 2012.

After getting the non-binary reward $\tilde{r}$, we perform a Bernoulli trial with success probability $\tilde{r}$ and observe the output r which will be considered as the binary reward.

We'll consider the following MAB : Four Bernoulli arms, two exponential arms and three beta arms (see code for details)



The notion of complexity still makes sense according to the paper proposed. Indeed, for general nonparametric models, and under appropriate assumptions on the rate of convergence, the notion of complexity still holds as explained in the "Theorem 1".

2 Linear Bandit on Real Data

2.1 Question 3

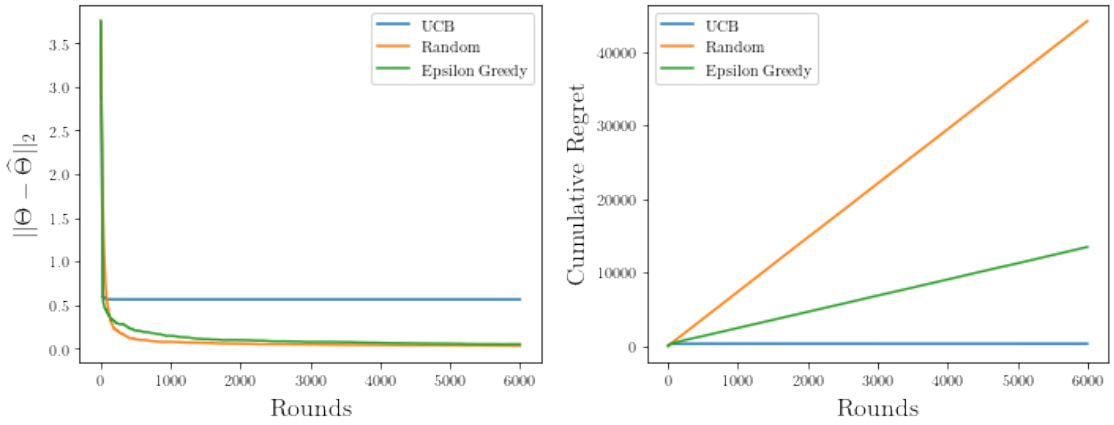
We'll implement the LinUCB algorithm for the provided MovieLens problem, for a horizon **T=6000**, and compare the performance against the random algorithm, and the $\epsilon - greedy$ algorithm. We consider a **decayed uncertainty factor** : $\alpha_t = \frac{\alpha}{t^\gamma}$ with γ the decay factor.

We run **50** simulations and play with the parameters, with a fixed decay $\gamma = 0.5$.

We finally choose the following parameters **(see below justification)** :

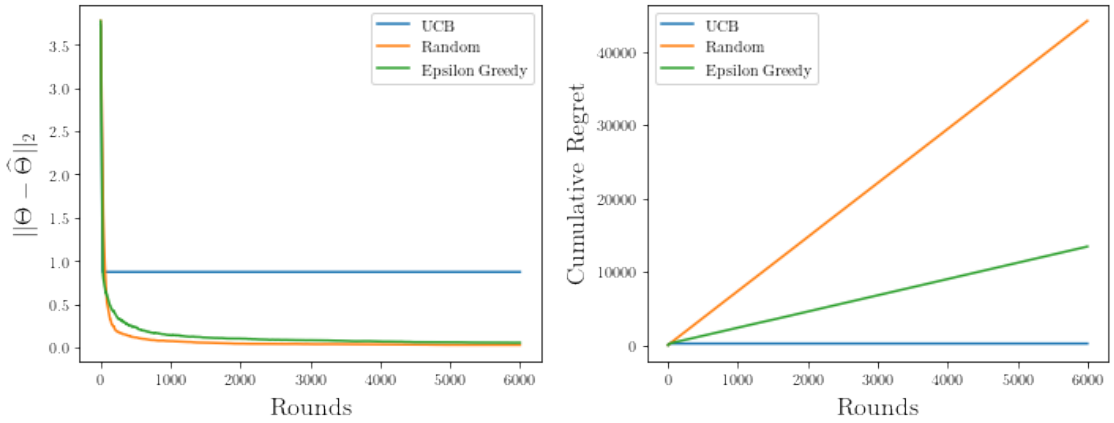
- α **= 70** with $\alpha_t = \frac{\alpha}{\sqrt{t}}$ the decayed uncertainty coefficient
- λ **= 0.1**
- ϵ **= 0.3**

and obtain the following results :



2.1.1 Choice of α

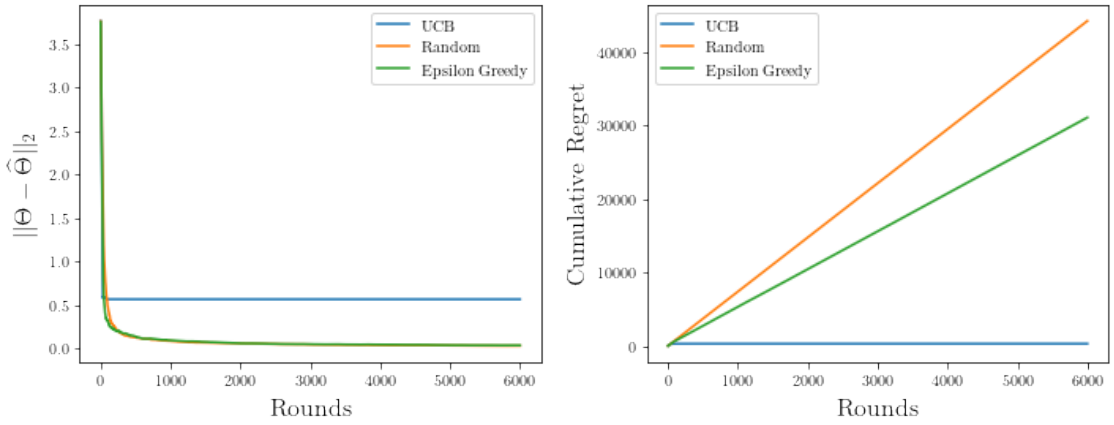
We remark that if we decrease α from 70 to 10 for example, with all the other parameters fixed, we favorise the exploitation over the exploration, and we obtain same cumulative regret, but worse $||\theta - \hat{\theta}||_2$:



We choose α = 70

2.1.2 Choice of ϵ

If we increase ϵ for a value above 0.5 (we favorize exploration over exploitation), we obtain worse cumulative regret for the ϵ greedy as shown below :



We choose $\epsilon = 0.3$

2.1.3 Choice of λ

If we increase λ from 0.1 to 0.8, we obtain similar cumulative regret, but worse $||\theta - \hat{\theta}||_2$: