Master M2 MVA 2018/2019
Reinforcement Learning - TP2

Souhaib ATTAIKI

November 20, 2018

1 Stochastic Multi-Armed Bandits on Simulated Data

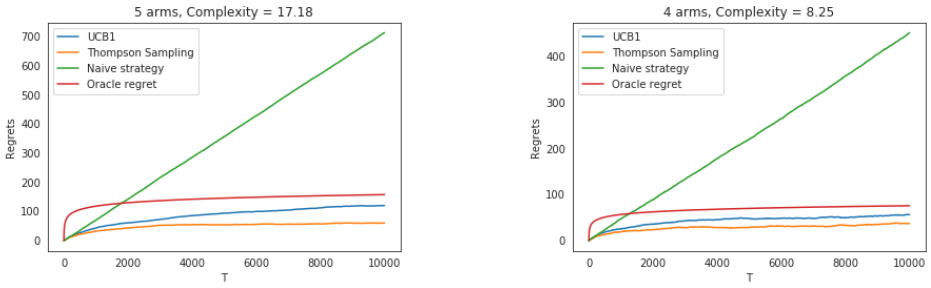
1.1 Bernoulli bandit models

We have created two Bernoulli bandit problems (see *visualisations.pdf*) :

— The first has **5 arms** and a complexity equal to **17.18**

— The second has **4 arms** and a complexity equal to **8.25**

The regret curve of the UCB1, Thomson Sampling, Naive strategy algorithms as well as the regret curve of the "oracle" are presented in the following figures for the two Bernoulli bandit models.



(a) Regret curves, pb1

(b) Regret curves, pb2

FIGURE 1 – Regret curves

It can be seen that the Thomson Sampling algorithm gives better results than UCB1. We also note that Naive strategy has a very poor performance and that its regret curve increases linearly with Time horizon T.

Concerning the lower bound of Lai & Robbins, it can be seen that it is respected from the very first stages of learning.

1.2 Non-parametric bandits (bounded rewards)

Adapted Thomson sampling To use Thomson sampling on non-binary reward arms, we adapt the latter by taking a sample from the arm that returns a reward α such that $0 \leqslant \alpha \leqslant 1$, and the returned reward of the adapted TS is a sample from a Bernoulli distribution that has as a parameter α.

We have implemented a mixed Multi-arm bandit problem, that has 2 ArmBernoulli, 2 ArmExp and 2 ArmBeta with different parameters (*See code*). The regret curves for UCB1 and adapted Thomson Sampling algorithms are presented in the following figure.

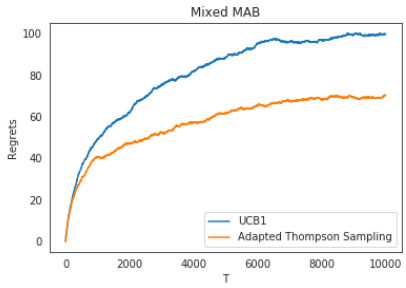


FIGURE 2 – Regret curves

It can be seen that even in this case, Thomson sampling gave a better result than UCB1.

The notion of complexity has no meaning in this case, because we are in a non-parametric case, and it is ambiguous to define the parameters p and p^*.

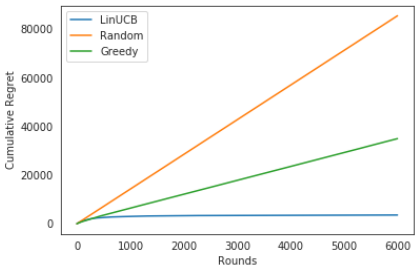
2 Linear Bandit on Real Data

The three algorithms (Linear UCB, random, ϵ-greedy) are implemented in visualizations.ipynb (see generated pdf). The figure 3 shows the different results found.

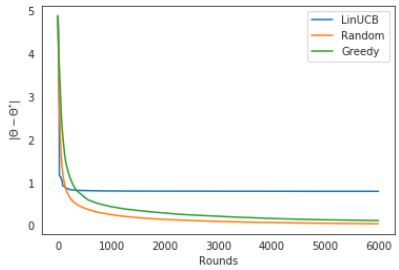
Several tests were performed to determine the optimal parameters for each algorithm, the goal being to minimize $\|\theta - \theta^*\|$, with minimal regret and rapid convergence. We have made the following choices :

- **Linear UCB** : $\lambda = 2$, $\alpha = 40$
- **Random** : $\lambda = 1$
- ϵ**-greedy** : $\lambda = 1$, $\epsilon = 0.4$

We can see that the greedy is the algorithm that approximates θ quickly and accurately in comparison with the other two, this is due to the fact that the arm is chosen randomly, so we have a lot of exploration, however, it generates that the algorithm accumulates a lot of regret due to the non-optimality of the chosen movie.



(a) Expected Cumulative Regret



(b) l_2-norm

FIGURE 3 – Cumulative regret and distance curves

ϵ-greedy has less regret but his approximation of θ is less good, this is explained by the compromise between choosing the right arm (exploitation with probability 1-ϵ) and randomly choosing an arm (exploration with probability ϵ).

Linear UCB is the algorithm with minimal regret, however its approximation of θ is not good enough than the other two, this is due to the fact that he focuses only on exploitation and not on exploration.