Report TP2 : The Exploration-Exploitation Dilemma

Plassier Vincent

1 Stochastic Multi-Armed Bandits on Simulated Data

1.1 Bernoulli bandit models

In order to compare the different methods, start by displaying the regression curves obtained with $\rho = 2$ and $\varepsilon = 0.1$:

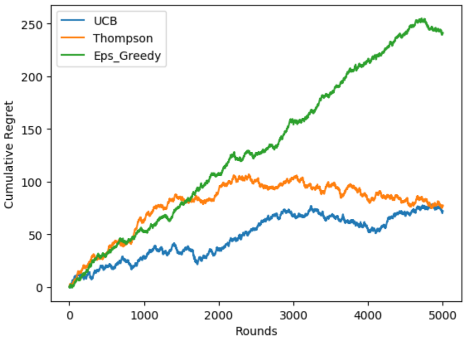
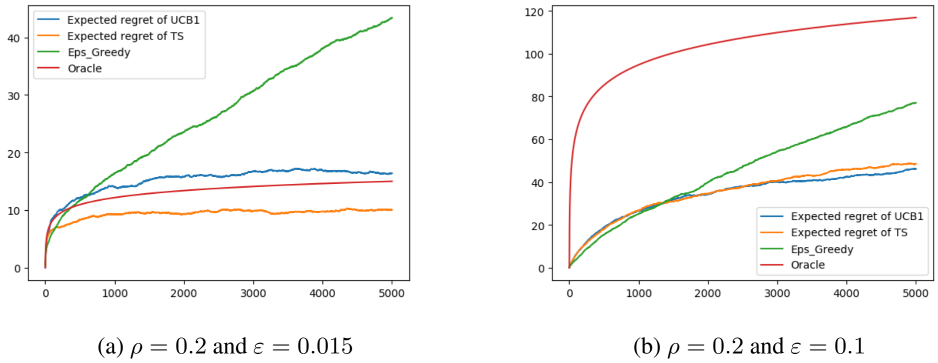


FIGURE 1: Regret obtained with the three methods

We note that the regrets obtained for UCB1 and Thompsom seem slightly better than that obtained by the Eps_Greedy method.

Now after 300 iterations, we display the curves of the expected regret as well as the oracle regret curve for 4 arms of parameters (0.9,0.4,0.2,0.3) in the left and (0.3,0.25,0.2,0.28) in the right :



(a) $\rho = 0.2$ and $\varepsilon = 0.015$

(b) $\rho = 0.2$ and $\varepsilon = 0.1$

FIGURE 2: Regret curves obtained with parametric bandits

This time, we find that Thompson looks slightly better than UCB1 and Eps_Greedy. Furthermore, in the low complex problem, it seems that ε may be choosen lower than in the non-complex case to well perform. In this case, the Thompson curve is under the oracle. This does not contradict the lower bound of Lai and Robbins since it is asymptotically valid. In addition, we note that the expected regret curves of UCB1 and Eps_Greedy strongly depend on the choice of ρ, ε while the Thompson Bayesian algorithm does not require the calibration of a parameter.

1.2 Non-parametric bandits (bounded rewards)

When the arms are not Bernoulli, the posterior distribution of $\beta(S_a(t)+1, N_a(t)-S_a(t)+1)$ would likely be integral. And it seems indispensable to keep the parameters N_a, S_a such that $\beta(S_a(t)+1, N_a(t)-S_a(t)+1)$ is intergal in order to analyse the problem. So we want to artificially come back to Bernouilli problem. One possible way of doing this is to simulate Bernoulli variables from the rewards observed. Indeed, since we supposed that the rewards are bounded in $[0,1]$, at each step t, we can draw a Bernoulli $R_t \sim \mathcal{B}(r_t)$ and we update $S_a(t) = S_a(t)+1$ only if we get $R_t = 1$.

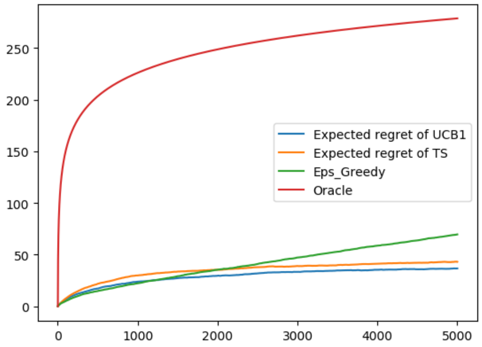


FIGURE 3: Regret curves obtained with non-parametric bandits

This time, the complexity no longer makes sense, as it depended of model's parameters and those aren't define anymore.

2 Linear Bandit on Real Data

Now, consider a horizon $T = 6000$. We have represented on the left curves $\|\hat{\theta}_t - \theta_{\ star}\|_2$ and on the right the expected cumulative regret obtained by the three previous methods :

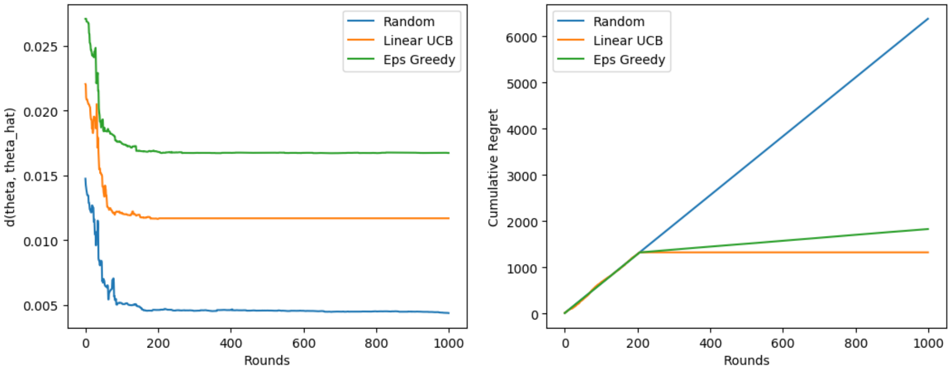


FIGURE 4: $\rho = 0.2$ and $varepsilon = 0.1$

This time, the linear UCB method seems to have a lower expected regret than the others two methods whereas $\|\hat{\theta}_t - \theta_{real}\|_2$ has a higher value.