Reinforcement Learning
TP2

Antonin Berthon

November 2018

1 Stochastic Multi-Armed Bandits on Simulated Data

1.1 Linear Bandit Problems

We run the UCB1 and the Thompson Sampling algorithms for two Bernoulli bandit problems with respectively 10 and 100 arms. We also compute the naive strategy that picks the empirically best arm at each iteration, as well as the lower bound $C(p)\log(t)$ from [Lai and Robbins, 1985]. The results are shown in figure 1 and 2. We see that the Thompson algorithm is much better at minimizing the regret than the UCB1 algorithm. Also we see that the lower bound applies asymptotically. Finally, the naive strategy can perform very differently depending on the simulation: either it can find the right arm quickly and so minimizes the regret very well, either it quickly settles for a sub-optimal arm and the regret grows linearly with time (see figure 3).

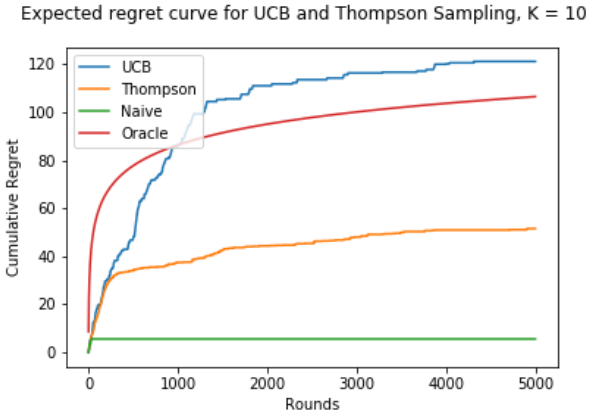

Figure 1: Regret curves associated with the naive strategy and the UCB1 and Thompson Sampling algorithms for a 10 arms Bernoulli bandit problem. Each curve is averaged across 20 iterations.

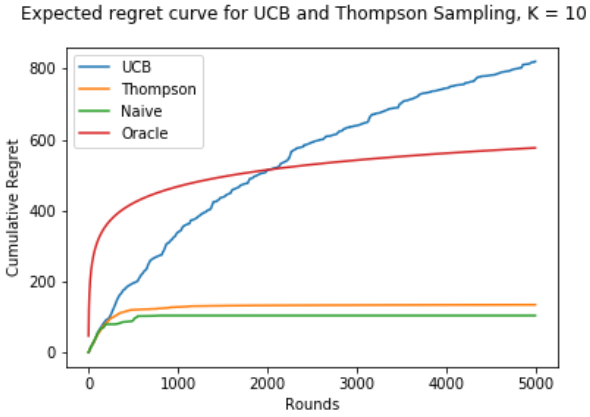

Figure 2: Regret curves associated with the naive strategy and the UCB1 and Thompson Sampling algorithms for a 100 arms Bernoulli bandit problem. Each curve is averaged across 20 iterations.

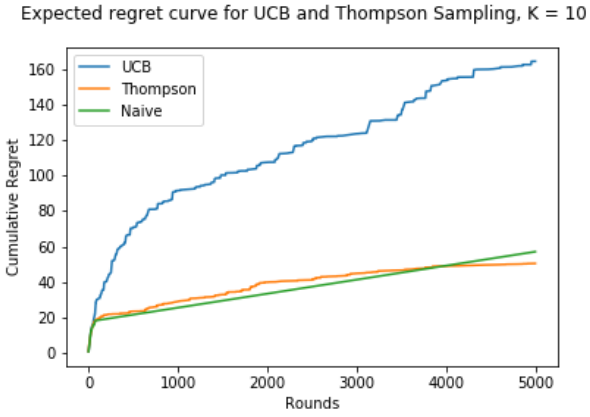

Figure 3: Example of a simulation where the naive strategy settles for a sub-optimal arm.

1.2 Non-parametric Bandits Problems

In order to apply Thompson Sampling to non-parametric bandits problems, we generalize the Thompson Sampling to handle non-binary variables as described in [Agrawal and Goyal, 2012]: with $S_a(t)$ and $N_a(t)$ referring to the number of success and draws of the arm a at time t, we draw K samples $(\theta_a)_{a=1,..,K}$ from the beta distributions $\text{Beta}(S_a+1, N_a-S_a+1)$, play the arm corresponding to the highest θ_a and observe a reward $r(\tilde{t})$. We then perform a Bernoulli trial with parameter $r(\tilde{t})$: if the trial succeeds we count it as a new success for arm a and increment its number of draws, else we only increment the number of draws for a. We plot the regret curve for this adapted Thompson Sampling and for the UCB1 algorithm, which keeps the same form for non-binary variables, in figure 4.

Here we simulate 10 arm bandit problem with 30% of the arms following a Bernoulli distribution, 30% of the arms following a Beta distribution and 40% following an Exponential distribution. Again, we see that Thompson Sampling performs much better than the UCB1 on the cumulative regret criterion. According to [Burnetas and Katehakis, 1996], the notion of complexity holds for a suitable extension of the coefficients of the lower bound $C(p)$.

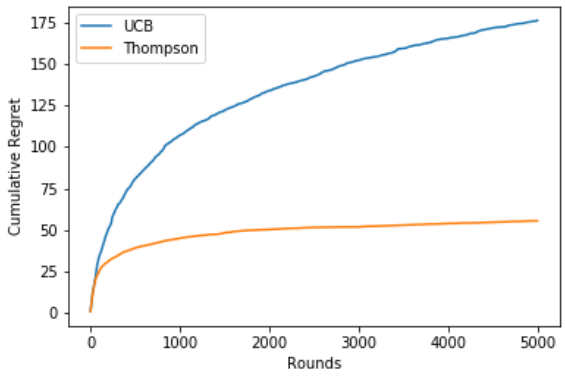

Figure 4: Regret curves associated with the UCB1 and Thompson Sampling algorithms for a 20 arms non-parametric bandit problem. Each curve is average across 10 iterations.

2 Linear Bandit on real Data

Choice of parameters

We set $\lambda = 0.1$ for the regularization parameter of the least-square optimization problem of the UCB algorithm.

For the confidence intervals of the UCB algorithm, we chose $\alpha_t = \alpha\sqrt{\log 1 + t}$. The bigger the coefficient α, the bigger the confidence intervals, so the more exploration will be done. This is showed in figure 5 : we see that for an alpha too low, we stick to a sub-optimal theta too early and accumulate a lot of regret. On the other hand, for an alpha too high, we converge to the real theta very well but we keep exploring too much, so the regret still accumulates. In this case a good value for α seems to be $\alpha = 1$.

The same principle goes for the ϵ parameter in the $\epsilon -$ greedy strategy: for ϵ close to 1, we return to the random policy which only does exploration, and as epsilon decreases we do more and more exploitation. We show the distance to θ^* and the final cumulative regret on figure 6 : again we see the trade off between converging quickly to the real θ and exploiting that knowledge. In the following we pick $\epsilon = 0.3$.

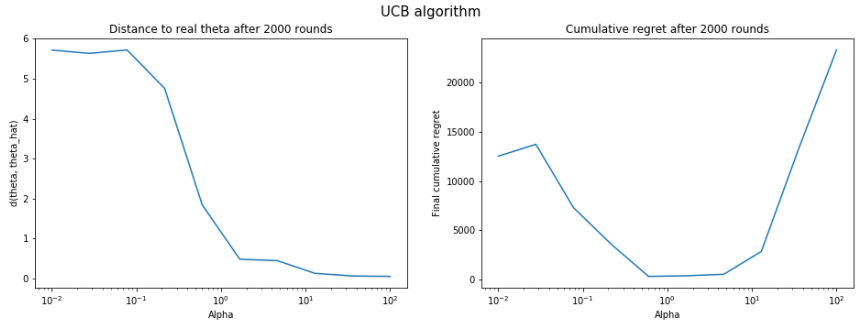

Figure 5: Distance to the real theta and cumulative regret after 2000 rounds average across 10 simulations, with respect to the α coefficient used in the Linear UCB algorithm.

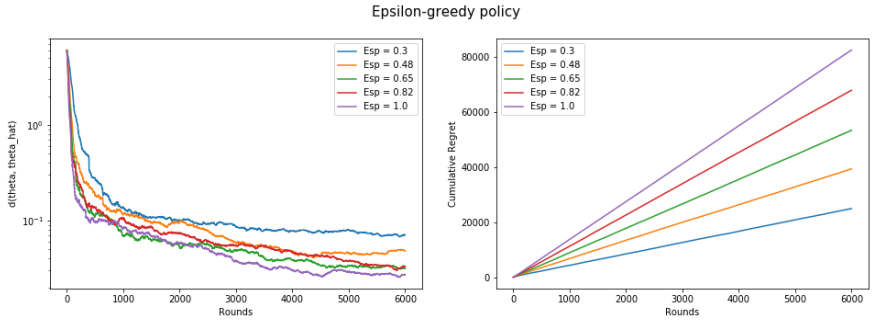

Figure 6: Evolution of the distance to the real theta and the cumulative regret for 2000 rounds averaged across 10 simulations of an $\epsilon-$ greedy policy, for different ϵ coefficients.

2.1 Convergence and performance of the algorithms

Figure 7 represents the L_2 norm of the estimated theta compared to the real theta and the cumulative expected regret for the Linear UCB, the random policy and the $\epsilon-$ greedy strategy, with the parameters described previously ($\epsilon = 3$, $\alpha_t = \sqrt{\log 1 + t}$). We see that the random policy is very quick to converge to θ^*, while the $\epsilon-$ greedy policy takes more time to converge. However, it is much better at minimizing the cumulative regret compared to the random policy which do not exploit its knowledge of θ. The UCB converges close to θ^* very quickly and minimizes the cumulative regret very well. In this case it does not completely converges to θ because it can find the optimal action with its approximation of θ. In a more complex problem where a better knowledge of θ would be required to find the optimal action, the UCB will keep exploring until it has approached θ^* enough.

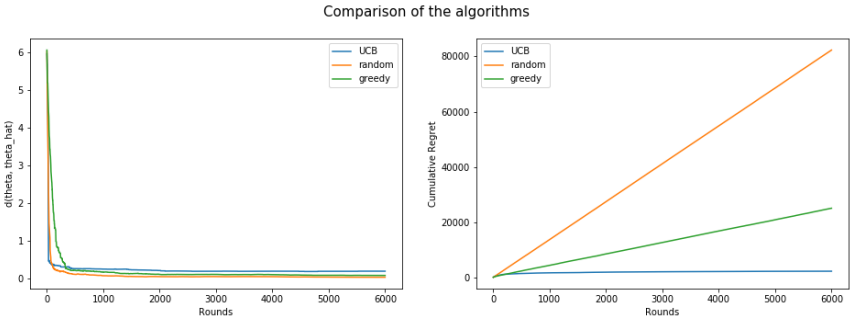

Figure 7: Comparison of the three algorithms using the distance to the real theta and the cumulative regret for 6000 rounds averaged across 10 simulations.