Exploration-Exploitation dilemna

November 26, 2018

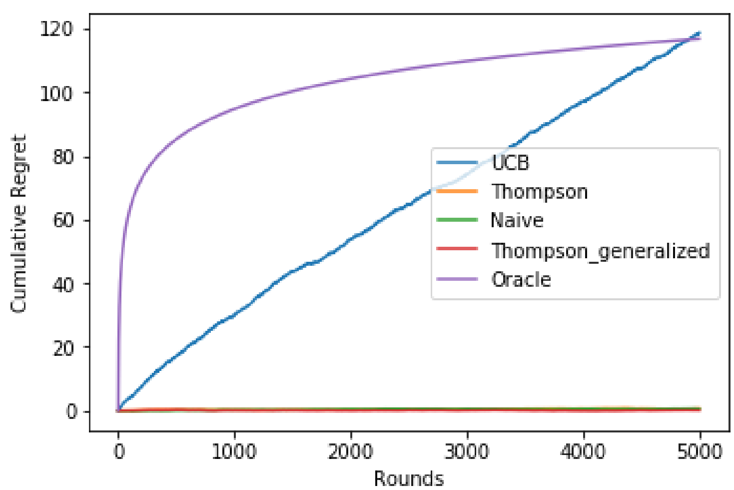
1 Stochastic Multi-Armed Bandits on Simulated Data

1.1 Bernoulli bandit models

We consider the following 4 arms multi-armed bandit model:

$$
\begin{aligned}
r_1 &\sim \mathcal{B}(0.3) \\
r_2 &\sim \mathcal{B}(0.25) \\
r_3 &\sim \mathcal{B}(0.20) \\
r_4 &\sim \mathcal{B}(0.10)
\end{aligned}
$$

We simulate the bandit game using the following algorithms: UCB, Thompson sampling, naive approach, and generalized Thompson, and we plot the cumulative regret



1.2 Non-parametric bandits (bounded rewards)

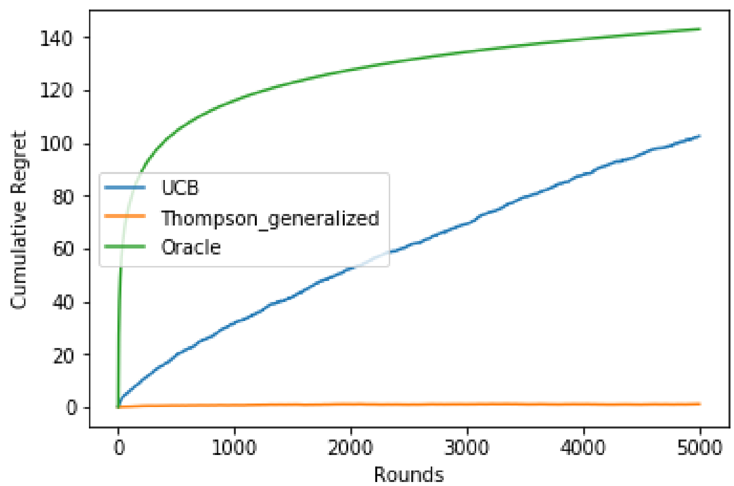
In the generalized Thompson sampling, we choose arm i in the same way we did in Bernoulli bandit problem, however when sampling the arm, this time we get $\tilde{r}_t \in [0, 1]$ instead of $\{0, 1\}$

since We are no longer in a Bernoulli frame, thus we get the reward r_t by sampling Bernoulli distribution $\mathcal{B}(\tilde{r}_t)$.

Here we cosnider the following arms:

$$\begin{aligned} r_1 &\sim \mathcal{B}(0.30) \\ r_2 &\sim \mathcal{B}(0.25) \\ r_3 &\sim Beta(2,5) \\ r_4 &\sim Beta(0.5,0.5) \\ r_5 &\sim \mathcal{E}(1) \\ r_6 &\sim \mathcal{E}(1.5) \end{aligned}$$

according to [Burnetas and Katehakis, 1996], there are no parametric assumptions used in the demonstration of the oracle lower bound, thus the notion of complexity still makes sense.

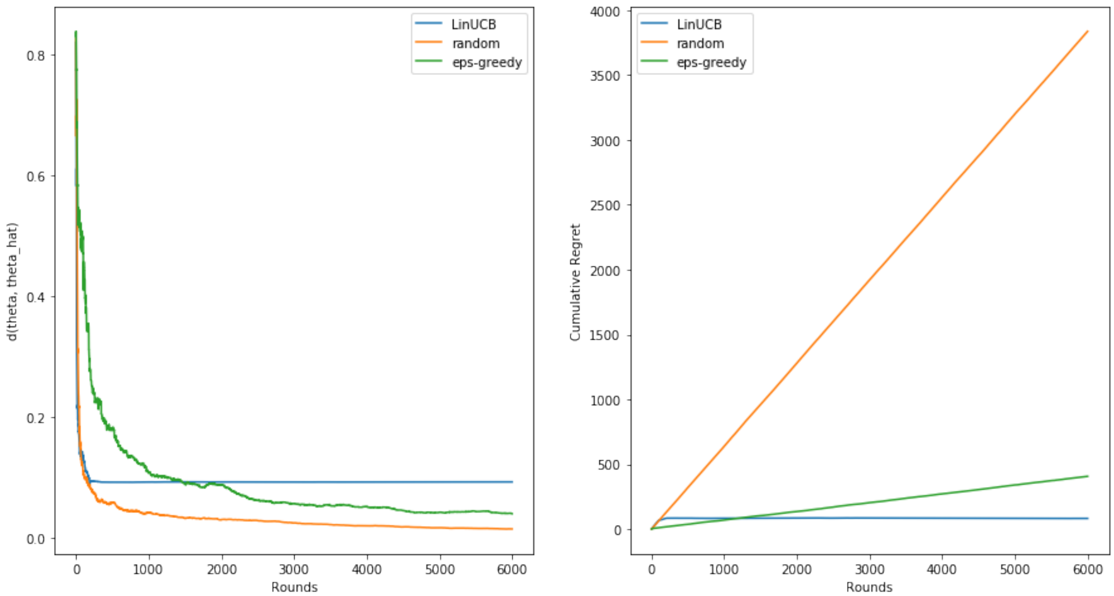


2 Linear Bandit on Real Data

We use $\alpha = 100$ decaying every 10 iterations with a root squared decay, $\lambda = 0.01$ for the linear UCB and $\epsilon = 0.1$ for the $\epsilon-$greedy policy.

the random exploration and $\epsilon-$greedy policy give better estimates of θ than linear UCB, but they suffer more from cumulative regret.

2.1 Toy Model



2.2 Cold-Start Movie Lens Model