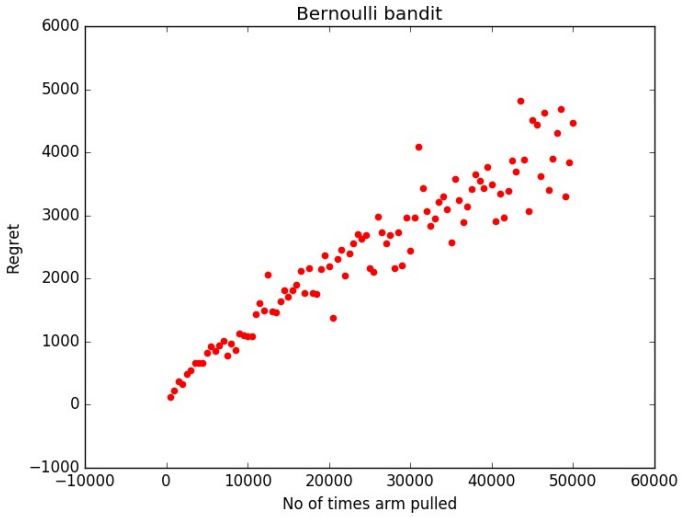
MULTI ARMED BANDITS

May 10, 2020

1 BERNOULLI BANDITS NAIVE ALGORITHM:

In this algorithm, we start with a red ball and a blue ball representing two arms with different means. We pick a ball at random and play the respective arm. If the reward is one, we add a ball of same colour else we continue the process.

To find regret for this process we approximate this to continuos time yules process using independent exponential distributions. The approxiamate regret for this process is $O(T^{\frac{\mu_2}{\mu_1}})$ where second arm mean is less than first arm's mean.



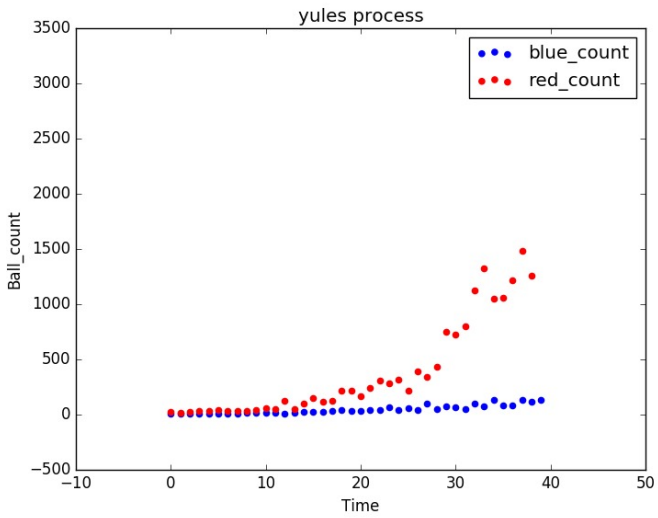
On the x-axis is the number of times both the arms together are pulled. On the y axis is the average regret for 100 episodes for each "n" on x-axis. The means used in the program were 0.6,0.4. The expected regret should be $O(T^{\frac{2}{3}})$

2 YULES PROCESS:

For approximating we regret we used Yules process in the naive algorithm. This process demonstrates growth of number of red and blue balls with time.

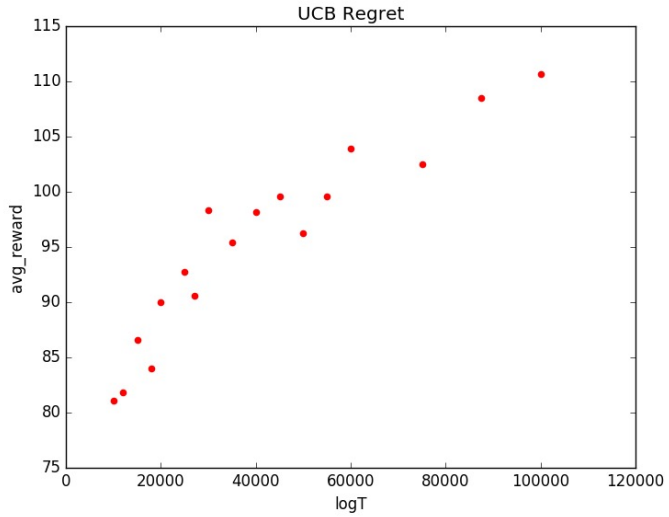
The following graph is plotted with number of blue balls on X axis and no of red balls on Y axis. From the paper, The blue balls grow sub linearly wrt to red balls. $N(B) = O(N(R)^{\frac{\mu_2}{\mu_1}})$.

I used *np.random.exponential* library to sample from exponential distributions. The following graph shows evolution of red and blue balls with time:



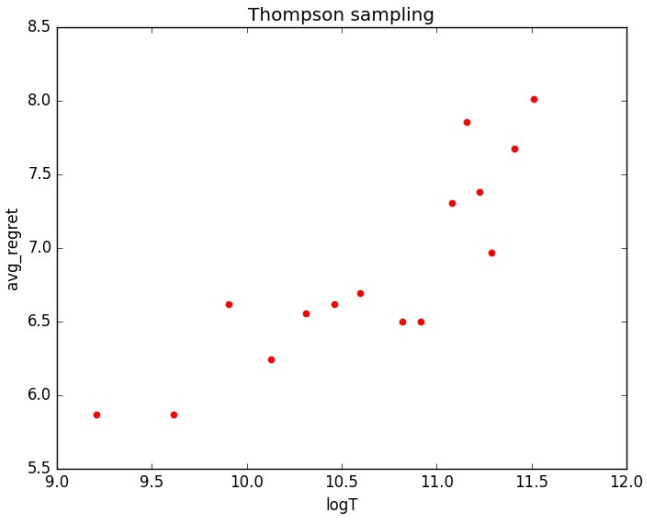
3 UCB SAMPLING:

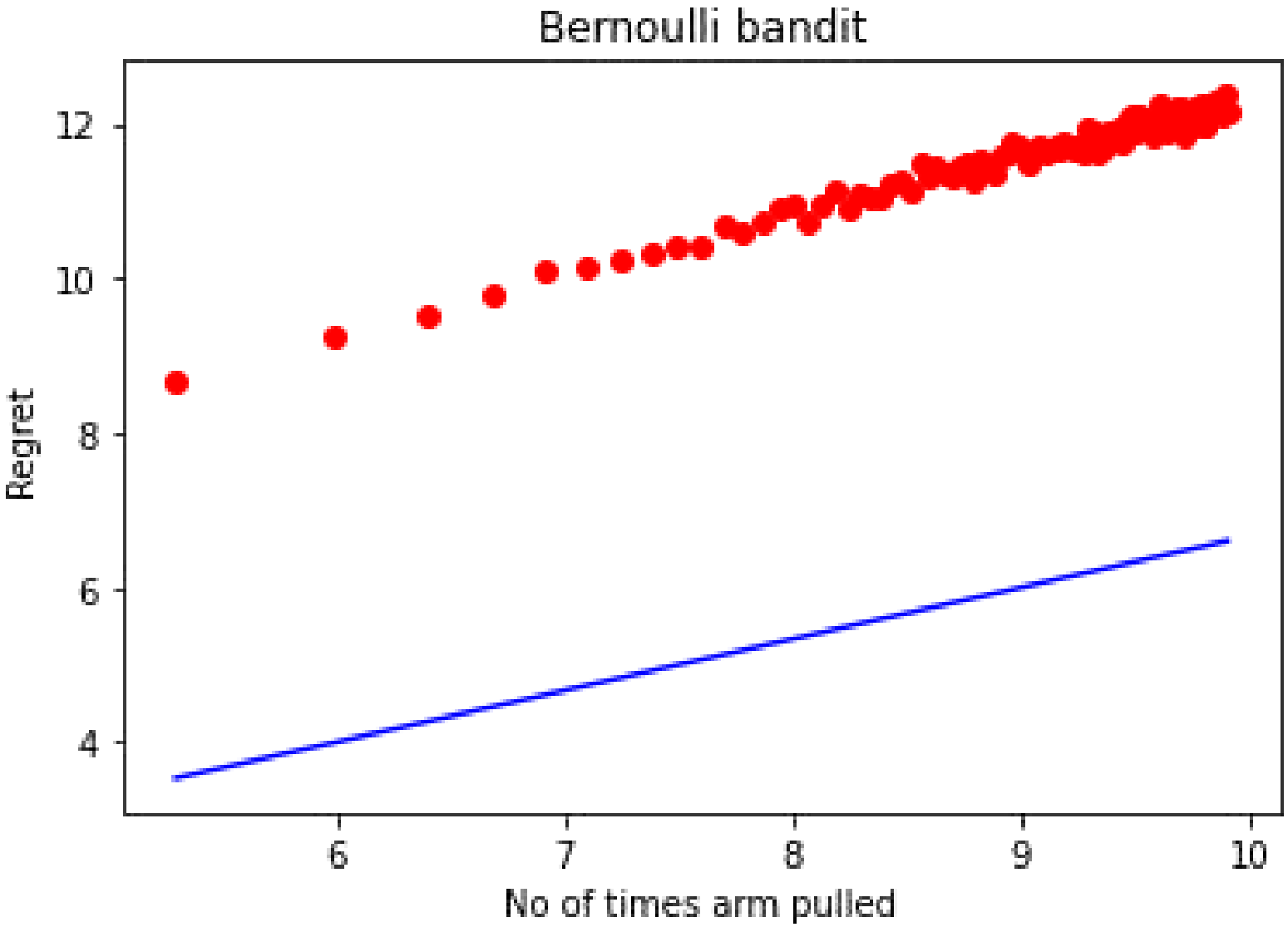
In the UCB implementation, I used 5 arms with means generated using a pseudo random number generator. The value of α used is 2.0 as too high or too low values are problematic as mentioned in paper. The upper bound on regret in UCB is $O(logT)$. The plot should be approximately logarithmic. Each experiment is averaged for 50 episodes for smoothing the plot.

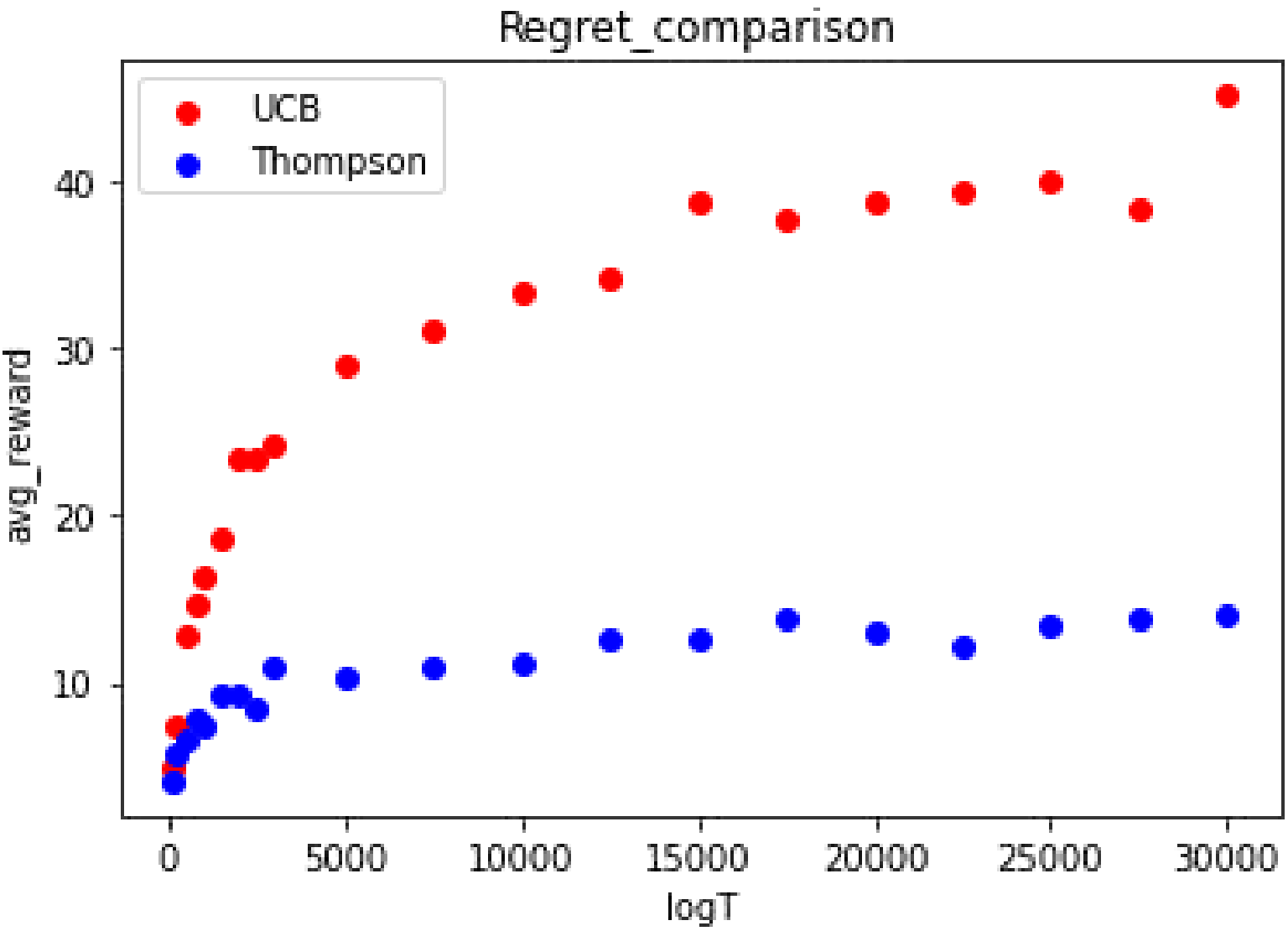


4 THOMSPSON SAMPLING:

To generate samples from beta distribution I used random.beta() module from Numpy library. I used a simple model only with two means. Both the means are generated from a psuedo random generator. The estimated regret is $\leq O(logT)$. Hence the regret the plotted against *log* axis in T. Each experiment is repeated for 25 epsiodes for smoothing the plot. The estimated plot should be a straight line.




Bernoulli bandit
Regret
12
10
8
6
4
6
7
8
9
10
No of times arm pulled


Regret_comparison
UCB
Thompson
avg_reward
10
20
30
40
0
5000
10000
15000
20000
25000
30000
logT