Report for Assignment 1 - Bandits

170020016, Drumil Trivedi

T1 - Implementations of Standard Methods (with assumptions)

- Epsilon Greedy
  - I assume that e*horizon is at least the size of the number of arms to allow for uniform exploration across arms least once for each.
  - Ties are broken arbitrarily after epsilon*T exploration steps(numpy implementation takes the first element achieving max)
- UCB
  - I assume that there is sufficient horizon for one round-robin exploration of arms.
  - Secondly, I maintain the mean and number of pulls, which leads to slight floating point differences leading to non-integer rewards, however, I see the difference to be negligible.
  - Implemented the ucb condition i.e. emperical_mean + sqrt(ln(t)/u_t) to take arg max on to choose the arm to pull at time step t.
- KL-UCB
  - I assume that there is sufficient horizon to perform two round-robin exploration fo arms. Two is performed as this provides space in the empirical mean (I would have ideally wanted at least min pulls to be possible have unique values for each arm, (n pulls if <= 2^n arms) however our lower bound 100 for instance 3 clashes with 25 arm for i-3 instance)
  - Implemented the KL-UCB condition to get the min q in the range [p_a,1] s.t. u_t*q < ln(t) + 3* ln(ln(t)) and chose arg max over q to get the arm
  - The search for q is a binary search over the range(p_a,1) broken down into intervals of size 0.001
- Thompson
  - While Thompson doesn't require a round-robin pull I saw that 2 rounds of round-robin helped reduce dependence on the random seed, so I assume the horizon is enough to allow two round-robin (this can be easily scratched out in the code if need be)
  - Fetched samples from beta(success+1,failure+1) for each arm to get a beta_arm vector over all arms over which arg max is taken to find the arm
  - For the values form the beta distributions the max arm is chosen arbitrarily in case of ties (numpy implementation takes the first element achieving max)

T2 - Modified Thompson Using Hint

Implementation-

I utilize the hint as follows. Given that we know that the arms belong to a fixed set of values (the hint) we set a soft distribution for each arm over these hints. We choose the arm with the highest probability of being the max mean (i.e. the soft probability for the last value in the sorted hints is the highest) breaking ties randomly (necessary to be random to prevent the slow start in the beginning). The update of these soft probabilities over the hints for the selected arm is that if the arm resulted in success, it means that it belongs to the higher mean arms, so we multiply its soft probabilities with the hint and normalize it, this skews its soft probabilities in the region of the higher means. However, if it results in failure we multiply it with the (1-hint) i.e. we push it towards lower means. This exploration is done on every time step.

T3 - Choosing Epsilon

instances/i-1.txt
Epsilon = 0.00004
Regret = 819.2
Epsilon = 0.00008
Regret = -1.3
Epsilon = 0.00016
Regret = 819.16
In terms of regret 0.00004>0.00008<0.00016

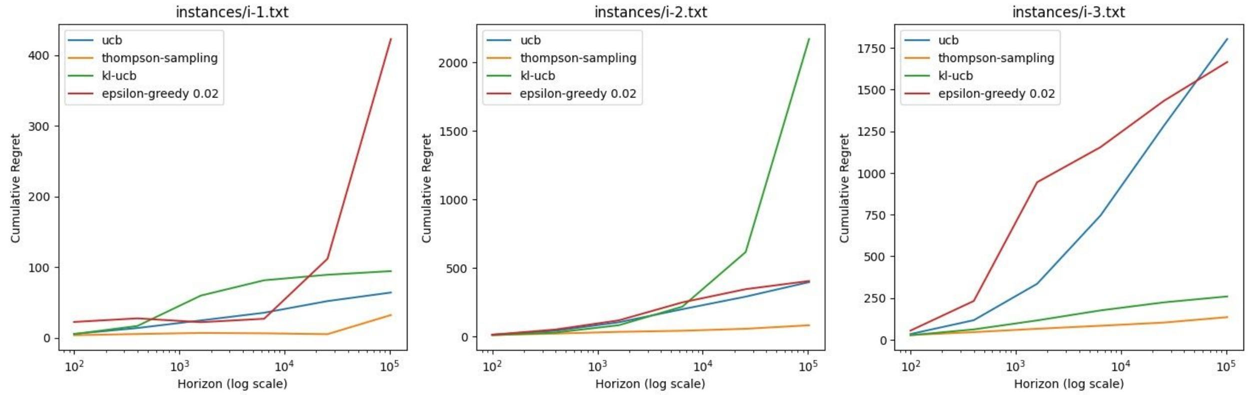
Note - While this example is sufficient to answer the question posed it is doesn't really imply that a particular epsilon_2 is sufficient since there are only two arms we are highly dependant on randomness.

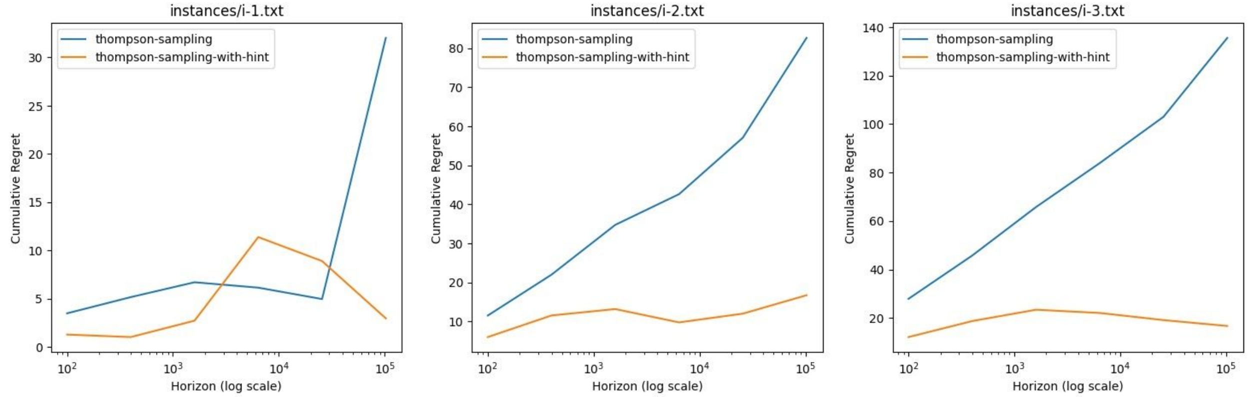
instances/i-2.txt
Epsilon = 0.01
Regret = 207
Epsilon = 0.02
Regret = 9.18
Epsilon = 0.04
Regret = 17.6
In terms of regret 0.01>0.02<0.04

instances/i-3.txt
Epsilon = 0.01
Regret = 248.56
Epsilon = 0.02
Regret = 16.42
Epsilon = 0.04
Regret = 34.26
In terms of regret 0.01>0.02<0.04

Note all the above regrets are scaled down by a factor of 50

T4 - Graphs and Analysis





The graphs almost adhere to our knowledge that epsilon-greedy > ucb > kl-ucb > thompson-sampling > thompson-sampling-with-hint in terms of regret.

The exception being KL-UCB blowing up in instance-2; looking at the individual regret for seeds I saw that the regret is low for several cases. However that few seeds having massive regret (20k for seed 48 10k for 6 others which really push the mean) I believe this is due to randomness and taking more seeds should flatten it out.

The second is the graph of Thompson getting better than Thompson with a hint, that too can be attributed to randomness as for the other two graphs it is indeed following the expectation and since in instance one there are only 2 arms the idea of skew to a side is (multiplying the prior) affected.