CS 747: Assignment 1

Mohit Madan - 15D070028

September 2, 2019

1. **Inference 1:** It can be seen from the figures that for large ϵ values initially the regret is lower but as the horizon increases the regret of smaller ϵ decreases and comes below large ϵ. Since in the long run less exploration is favourable as the optimal action is decided by then with less uncertainty.
2. **Inference 2:** Regret of epsilon-greedy method for 0.002 and 0.02 is even worse than round robin for small horizons as it may pick up an arm with least reward due to less exploration but in the long run it becomes better because with time it will select the arm with highest mean reward.
3. **Inference 3:** KL-UCB and Thompson Sampling gives the best results in all the three cases Round Robin gives the worst result as it is wasting chances by calling different actions each time
4. **Inference 4:** For instance 2&3 $\epsilon = 0.02$ gives lower regret than $\epsilon = 0.002$. This could possible somehow as it might require more runs for regret of $\epsilon = 0.002$ to get lower than $\epsilon = 0.02$

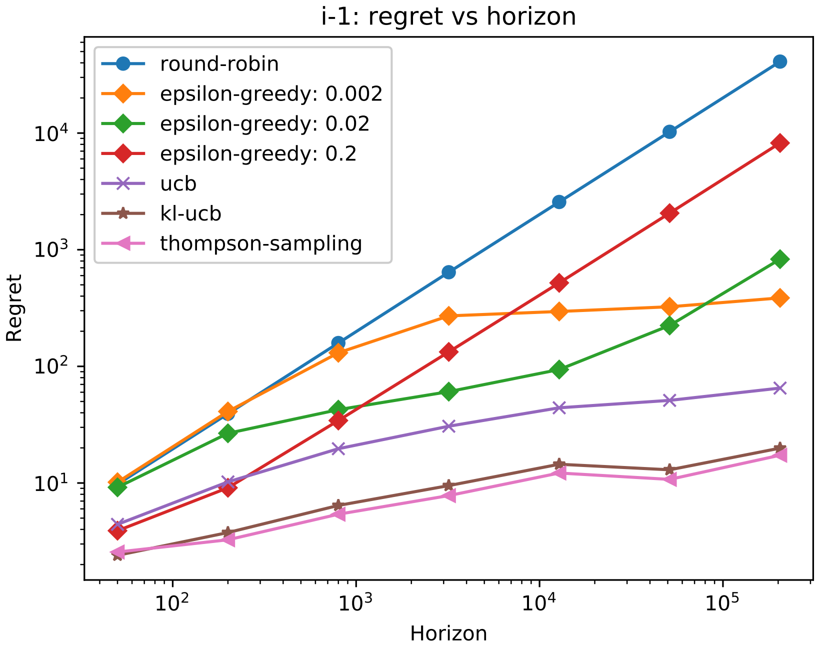

Figure 1: Instance-1 Regret vs Horizon

5. **KL-UCB:** action $= arg\ max_{1 \leq a \leq K}\ max\{q \in \Theta : N[a]d(\frac{S[a]}{N[a]}, q) \leq log(t) + clog(log(t))\}$ Here c is assumed to be 0 for optimal performance.
   Newton Iterations are used to find q since for $p \in [0,1]$, the function $q \mapsto d$ is strictly convex and increasing on the interval [p,1] where $d = p\ log(\frac{p}{q} + (1-p)\ log(\frac{1-p}{1-q})$

$$f(q) = \frac{log(t)}{N[a]} - d(S[a]/N[a], q)$$

$$f'(q) = \frac{q-p}{q(1-q)}$$

   q is iterated successively using $q' = q - \frac{f(q)}{f'(q)}$ until it converges. q is initialized using $p + \delta$ where δ is a very small value. In our case $\delta = 1e-7$ and converges if $f(q)/f'(q) < 1e-9$.

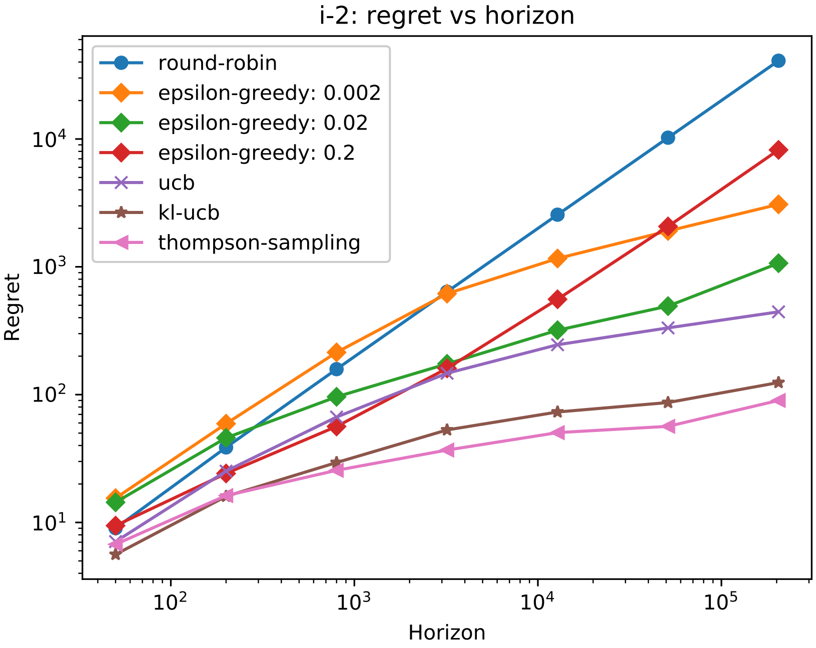

Figure 2: Instance-2 Regret vs Horizon

6. **Thompson Sampling:** Used beta random function at each chance to with inputs (1.0+num_success) and (1.0+num_failures) to find the best action. α and β values were chosen to be 1.0 in the equation $\theta = Beta(S + \alpha, F + \beta)$

7. **Assumption 1:** In UCB $A_t = argmax[Q_t(a) + c\sqrt{\frac{ln\ t}{N_t(a)}}]$, $c = \sqrt{2}$

8. **Assumption 2:** In case of epsilon greedy, while exploring it is randomly selecting the arm from all the arms including the current best arm.

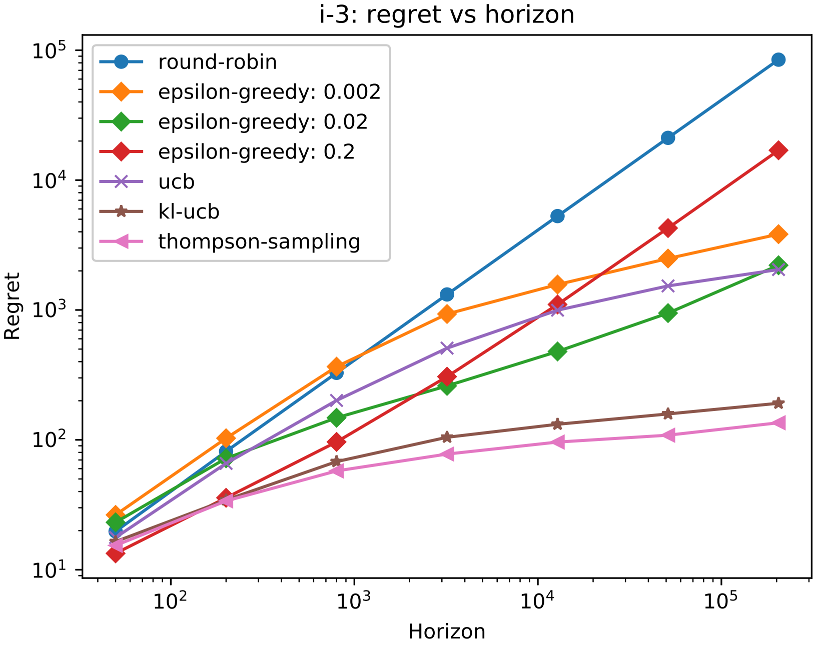

Figure 3: Instance-3 Regret vs Horizon