CS 747: Assignment 1

Mohit Madan - 15D070028

September 1, 2019

1. **epsilon-greedy:** 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.

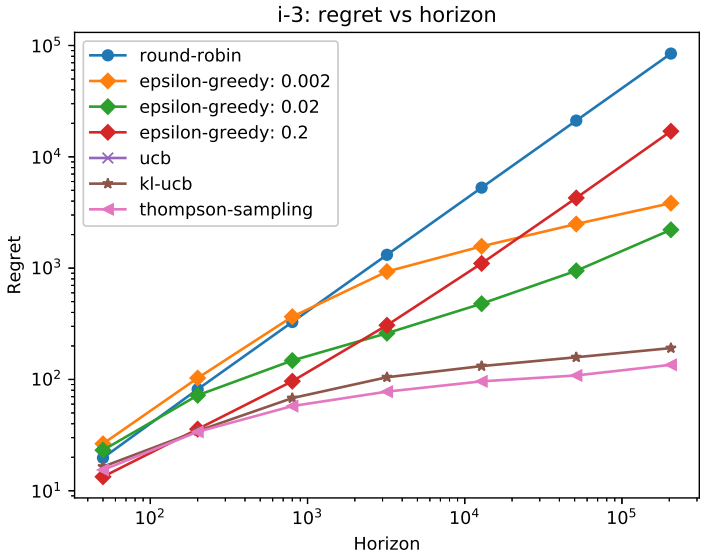

Figure 1: Instance-3 Regret vs Horizon

2. **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 $f(q)/f'(q) < 1e-9$.

3. **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 $\theta = Beta(S + \alpha, F + \beta)$

4. 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

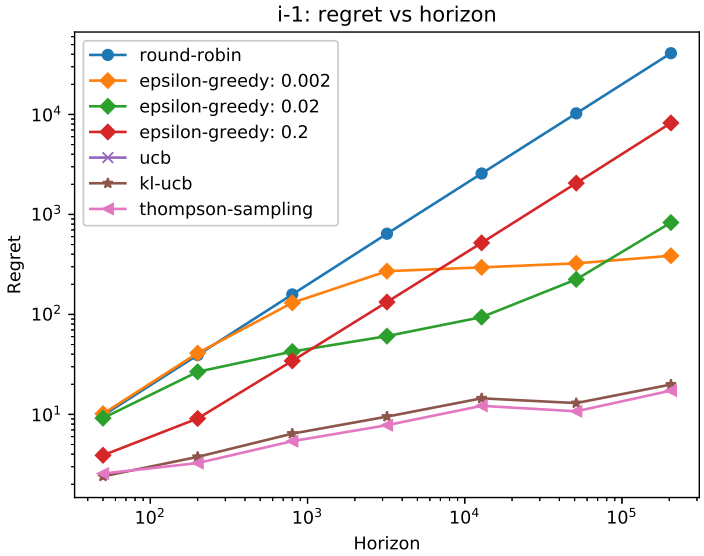

Figure 2: Instance-1 Regret vs Horizon

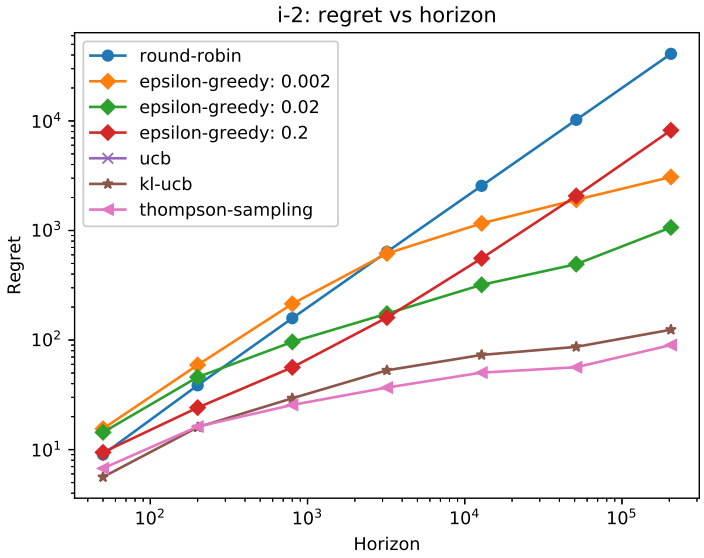

Figure 3: Instance-2 Regret vs Horizon