Surabhi
S. Nath
2016271

RL Homework 1

Theory

__/__/__

Q2 Mysterious spikes

Since Q value is set to a large value of 5 compared to q^* values taken from a normal distribution with 0 mean and variance 1. When an action is selected, the reward will be less, as it is sampled from mean q^* and variance 0.01 due to which the updated Q will reduce, since in $Q'(a) \leftarrow Q(a) + \alpha[r - Q(a)]$, the quantity $r - Q(a)$ will be negative. Since Q' reduces, in the next greedy pick, a different action is selected say a' and again $Q'(a')$ is reduced to lesser than $Q(a')$. Like this, in the first 10 picks / steps, the 10 different actions are explored and all Q' values are lesser than earlier Qs. Now, among all Q', the max value will belong to the optimal action since it would have received highest reward when first picked, hence $(r - Q)$ will be least negative for the optimal action. As a result, in the 11th step, the optimal action will be picked in all runs. Due to this, we see a spike in the graph of steps vs % optimal action around the 11th step.

Observation: Stationary vs Non Stationary

For the stationary case, while the optimistic, greedy approach performs worse than realistic initially, once it has explored, it outperforms the realistic ϵ-greedy eventually. However this cannot work for non stationary cases which may require subsequent need for reexploration due to changing q^* values. ∴ In case of non stationary, the optimistic greedy approach is not as efficient.

Q1 Both graphs of average reward and % optimal action highlight the better performance of constant α as compared to sample average for case of non stationary. This is because the constant α gives more weight to recent rewards which is suitable for nonstationary

Q3 Unbiased constant step size Trick. __/__/____

Step size = $\beta_n = \frac{\alpha}{\bar{o}_n}$

$\bar{o}_n = \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1})$ for $n \geq 0$ with $\bar{o}_n = 0$.

To show: q_n is an exponential recency-weighted average without initial bias.

Proof: Basically we need to show that q_n is independent of q_1 since all bias is included in q_1 values.

q update equation:

$q' = q + \beta_n (R - q)$.

$q_{n+1} = q_n + \frac{\alpha}{\bar{o}_n}(R_n - q_n)$

$q_2 = q_1 + \frac{\alpha}{\bar{o}_{n=1}}(R_1 - q_1)$

$\bar{o}_1 = \bar{o}_0 + (\alpha(1 - \bar{o}_0)) = 0 + \alpha = \alpha$

$\frac{\alpha}{\bar{o}_1} = \frac{\alpha}{\alpha} = (1)$

$\therefore q_2 = q_1 + R - q_1 = R_1$ | q_2 is independent of q_1

$q_3 = q_2 + \frac{\alpha}{\bar{o}_2}(R_2 - q_2)$

$= R_1 + \frac{\alpha}{\bar{o}_2}(R_2 - R_1)$

$\bar{o}_2 = \bar{o}_1 + \alpha(1 - \bar{o}_1) = 2\alpha - \alpha^2 = \alpha(2 - \alpha)$

$q_3 = R_1 + \frac{R_2 - R_1}{2 - \alpha}$ | q_3 is independent of q_1

$q_n = q_{n-1} + \beta[R_{n-1} - q_{n-1}]$

$= q_{n-1}[1 - \beta_{n-1}] + (k) \rightarrow$ indep of q

q_{n-1} : $[1 - \beta_{n-1}]$ indep of q

$q_{n-2} + \beta_{n-2}[R_{n-2} - q_{n-2}] = q_{n-2}$ (m) + (n)

q_{n-2} : q_2

Since q_n depends on q_{n-1} ... which only depends on q_2 which is independent of q_1, q_n is unbiased | q_2

Q4

Comparing UCB, greedy optimistic and realistic ε greedy, for stationary, greedy optimistic starts out as the worst performing method but later outperforms both UCB and realistic methods. Initial exploration causes the low initial average rewards, however later, it performs the best as it greedily chooses the best action each time which is suitable in a stationary setting. Among UCB and realistic ε greedy, UCB performs marginally better. This could be due to adaptive exploration in UCB which increases q value more for the actions not selected in turn encouraging exploitation.

For the non stationary case, UCB performs poorly and in the long run even the optimistic greedy approach fails to perform well due to not exploring at the later times. ε greedy performs similar to UCB and is perhaps easier to extend to non stationary cases as compared to UCB.