Q2 : There is a spike in the stationary case initially value is 5 for all the actions, as we are exploiting continuously, the agent traverses over all the actions as after taking the action with the highest Q value, the agent gets a small reward(also as step size=0.1, the new Q[a] becomes less than the previous Q[a]), and so in the next iteration it chooses another action and once again gets a small reward and this keeps on continuing till all the actions are chosen. As all the actions are being chosen in the first few steps, one of the chosen actions matches the optimal action and hence we get spikes in the beginning.

In the non stationary case, the optimal action keeps changing at every step, now even though the the agent traverses all the actions in the optimistic case, there’s no guarantee that one of them will definitely be the optimal action, as the q_star is also changing at every step/iteration. As the q_star is non stationary, even exploring does not help much as at every iteration the optimal action changes.

Q3: below

Q4: UCB performs worse than optimistic, greedy and e-greedy in the non stationary case while in the stationary case UCB performs better than e-greedy

NOTES

$\beta_n = \alpha / \bar{o}_n$ where $\bar{o}_n = \bar{o}_{n-1} + \alpha(1-\bar{o}_{n-1})$, $\bar{o}_0 = 0$

$$Q_{n+1} = Q_n + \beta_n (R_n - Q_n)$$
$$= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i$$

Here $\alpha = \beta_n$

$$\Rightarrow Q_{n+1} = (1-\beta_n)^n Q_1 + \sum_{i=1}^{n} \beta_n (1-\beta_n)^{n-i} R_i$$

Now we need to show that $(1-\beta_n)^n = 0$ so that Q_{n+1} is independent of Q_1

$$(1-\beta_n)^n = 1 - {}^nC_1 \beta_n + {}^nC_2 \beta_n^2 + \text{———} (\beta_n)^n$$

$$\rightarrow Q_{n+1} = Q_n + \beta_n (R_n - Q_n)$$

Now $\beta_1 = \alpha / \bar{o}_1$ & $\bar{o}_1 \doteq \alpha$

$$\Rightarrow \boxed{\beta_1 = 1}$$

$$\Rightarrow Q_2 = Q_1 + 1(R_1 - Q_1) = R_1$$

NOTES

Now: similarly

$$Q_3 = Q_2 + \beta_2 (R_2 - Q_2)$$

$$\Rightarrow R_1 + \frac{\alpha}{\bar{O}_2} (R_2 - R_1)$$

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

$$\Rightarrow Q_3 = R_1 + \frac{1}{(2-\alpha)} (R_2 - R_1)$$

$$Q_3 = \cancel{R_1} \frac{(1-\alpha)R_1}{(2-\alpha)} + \frac{R_2}{(2-\alpha)}$$

Similarly $Q_n = Q_{n-1} + \beta_{n-1}(R_{n-1} - Q_{n-1})$ is true

let's show $Q_{n+1} = Q_n + \beta_n(R_n - Q_n)$ is also true by inductⁿ

NOTES

$$Q_{n+1} = Q_{n-1} + \beta_{n-1}(R_{n-1} - Q_{n-1}) + \beta_n(R_n - Q_n)$$

$$Q_{n+1} = \underbrace{\beta_{n-1} R_{n-1} + \beta_n R_n}_{\text{good}} + \underbrace{Q_{n-1}(1 - \beta_{n-1}) - \beta_n Q_n}_{\text{(1) need to change}}$$

Now Q_n

$$\beta_n[Q_{n-1} + \beta_{n-1}(R_{n-1} - Q_{n-1})]$$

$$\Rightarrow (1): \quad Q_{n-1}(1 - \beta_{n-1}) - \beta_n[Q_{n-1} + \beta_{n-1}(R_{n-1} - Q_{n-1})]$$

Now removing $-\beta_n \beta_{n-1} R_{n-1}$ as it is good term we're left with

$$Q_{n-1}(1 - \beta_{n-1}) - \beta_n[Q_{n-1} - \beta_{n-1} Q_{n-1}]$$

$$Q_{n-1}[\underbrace{1 - \beta_{n-1} - \beta_n + \beta_n \beta_{n-1}}_{\text{some const } \alpha}]$$

$$\cancel{\left\{ \beta_n + \beta_{n-1} = \frac{\alpha}{\sigma_n} + \frac{\alpha}{\sigma_{n-1}} \right\}} \text{ No USE}$$

$$\text{Now } Q_{n-1} = Q_{n-2} + \beta_{n-2}[R_{n-2} - Q_{n-2}]$$

$$\Rightarrow Q_{n-2}[1 - \beta_{n-2}](\ \wedge\ )$$

NOTES

Doing this we'll reach:

$Q_1 = R_1$

$\Rightarrow R_1(1-\beta)(\quad)$ ——

terms remaining will be all in terms of R_i & not Q_1 will be present

$\Rightarrow$

Hence Proved