3. $\beta_n = \dfrac{\alpha}{\bar{o}_n}$ and $\bar{o}_n = \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1})$

$\bar{o}_0 = 0$

$$Q_{n+1} = Q_n + \text{stepsize}\ (R_n - Q_n)$$
$$= Q_n + \beta_n (R_n - Q_n)$$
$$= \beta_n R_n + (1 - \beta_n) Q_n$$
$$= \beta_n R_n + \left(\frac{\bar{o}_n - \alpha}{\bar{o}_n}\right) Q_n$$
$$= \beta_n R_n + \left(\frac{\bar{o}_{n-1} + \alpha - \alpha \bar{o}_{n-1} - \alpha}{\bar{o}_n}\right) Q_n$$
$$= \beta_n R_n + \left(\frac{\bar{o}_{n-1}}{\bar{o}_n}\right) (1-\alpha) Q_n$$

Similarly,

$$Q_n = \beta_{n-1} R_{n-1} + \left(\frac{\bar{o}_{n-2}}{\bar{o}_n}\right)(1-\alpha) Q_{n-1}$$

$$Q_{n+1} = \beta_n R_n + \left(\frac{\bar{o}_{n-1}}{\bar{o}_n}\right)(1-\alpha)\left[\beta_{n-1} R_{n-1} + \left(\frac{\bar{o}_{n-2}}{\bar{o}_n}\right)(1-\alpha) Q_{n-1}\right]$$

$$= \beta_n R_n + \left(\frac{\bar{o}_{n-1}}{\bar{o}_n}\right)(1-\alpha)\beta_{n-1} R_{n-1} + \left(\frac{\bar{o}_{n-2}}{\bar{o}_n}\right)(1-\alpha)^2 Q_{n+}$$

Generalised formula:

$$Q_{n+1} = \sum_{i=0}^{n} \left(\left(\frac{\bar{O}_{n-i}}{\bar{O}_n} \right) (1-\alpha)^i \beta_{n-i} R_{n-i} \right)$$

$$+ \left(\frac{\bar{O}_0}{\bar{O}_n} \right) (1-\alpha)^n Q_1$$

As we know $\bar{O}_0 = 0$,

$$Q_{n+1} = \sum_{i=0}^{n} \left(\frac{\bar{O}_{n-i}}{\bar{O}_n} \right) (1-\alpha)^i \beta_{n-i} R_{n-i}$$

As we don't have any term of Q_1 in the final expression, Q_{n+1} is independant of initial bias (Q_1)

4. Observation:

In the case of non-stationary, ε-greedy performs the best followed by ~~UCB and~~ optimistic and UCB. In the beginning, optimistic performs better as it takes only best possible action without exploring but in the long run, ε-greedy gets a better reward.

In the case of stationary, UCB and optimistic perform better than ε-greedy method. Here, again optimistic performs better in the start. UCB and optimistic are very close in performance but UCB is better (slightly).

2. Observation :

We observe spikes in the initial steps of learning as most of the agents would try to go through all possible actions and receive corresponding rewards. Now based on these rewards, they have a very high chance to pick the optimal action as that action would have given them the best reward.

For non-stationary, optimistic & ε-greedy perform similar at start but ε-greedy overtakes as for a long run exploring with exploiting is better than just exploiting.