RL Assignment-1

Sanchit Malhotra
2016264

A2) (a) At the starting of optimistic greedy, we constantly get disappointed and explore for some initials turns. This happen for all 2000 individual 10-armed bandits. This causes accumulation of majority taking good or bad step together. This causes spikes in early stages.

(b) As we can infer from the graph optimistic greedy is not good method for non-stationary methods. Because after some time this method is the same as greedy which causes no exploration even when the optimal might have changed in non-stationary problem.

A3) As mentioned in the question, taking step size

$$\beta_n = \frac{\alpha}{\bar{o}_n} \quad \text{where} \quad \bar{o}_n = \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1}) \text{ for } n \geq 0 \text{ with } \bar{o}_0 = 0 \quad \text{---(1)}$$

has no initial bias & is exponentially waited

Proof:- $$Q_{n+1} = Q_n + \beta_n [R_n - Q_n]$$

$$Q_{n+1} = Q_n [1 - \beta_n] + \beta_n R_n$$

$$Q_{n+1} = \frac{1}{\bar{o}_n} \left[(\alpha \bar{o}_n - \alpha) Q_n + \alpha R_n \right] \quad \text{---(2)}$$

simplifying eq (1)

$$\bar{o}_n = \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1})$$

$$\boxed{\bar{o}_n - \alpha = \bar{o}_{n-1}(1 - \alpha)}$$

putting this in eq (2).

$$Q_{n+1} = \frac{1}{\bar{o}_n}\left[\bar{o}_{n-1}(1-\alpha)Q_n + \alpha R_n\right] \quad \text{— (3)}$$

putting value of Q_n to eq (3)

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

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

$$Q_{n+1} = \frac{1}{\bar{o}_n}\left[\bar{o}_{n-3}(1-\alpha)^3 Q_{n-2} + \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2\alpha R_{n-2}\right]$$

$$\vdots$$

$$Q_{n+1} = \frac{1}{\bar{o}_n}\left[\bar{o}_0(1-\alpha)^n Q_1 + \alpha R_n + \sum_{i=1}^{n-1}(1-\alpha)^i \alpha R_{n-i}\right]$$

as $\bar{o}_0 = 0$.

we get

$$Q_{n+1} = \frac{1}{\bar{o}_n}\left[\sum_{i=0}^{n-1}(1-\alpha)^i R_{n-i}\right]$$

which is an exponential recency weighted average & as Q_n does not depend Q_1 it is without bias

A4) In stationary, for 1000 iterations UCB and optimistic perform almost equally. But UCB performs slightly better than ε-greedy because of better method of exploration

In Non-stationary, UCB perform much better than optimistic. Since After some time optimistic becomes normal greedy but information gained changes with time hence it is of no-use after some iterations.

UCB performs slightly better than ε-greedy because of better method of exploration.