__/__/____

Q 2. Exercise 2.6

In optimistic greedy approach, the agent tends to explore more in the initial steps. This is due to the dissatisfying rewards it receives i.e the actual rewards recieved are much less than the optimistic action value estimates and hence different actions are tried. As optimal action would also be chosen, hence no. of times it is chosen (across several runs) in the initial steps gives spikes.

(along with others)

Q3. The incremental rule for estimated reward Q is given by :

$$Q_{n+1} = Q_n + \underbrace{\alpha}_{\text{step size}} [R_n - Q_n]$$

with step size $\beta_n = \frac{\alpha}{\bar{o}_n}$, we have and $\bar{o}_n = \bar{o}_{n-1} + \alpha(1 - \bar{o}_{n-1})$

$$\begin{aligned}
Q_{n+1} &= Q_n + \beta_n [R_n - Q_n] \\
&= \beta_n R_n - \beta_n Q_n + Q_n \\
&= \beta_n R_n + (1 - \beta_n) Q_n \\
&= \beta_n R_n + (1-\beta_n)(Q_{n-1} + \beta_{n-1}(R_{n-1} - Q_{n-1})) \\
&= \beta_n R_n + (1-\beta_n)\beta_{n-1} R_{n-1} + (1-\beta_n) Q_{n-1} - \beta_{n-1}(1-\beta_n) Q_{n-1} \\
&= \beta_n R_n + (1-\beta_n)\beta_{n-1} R_{n-1} + (1-\beta_n)(1-\beta_{n-1}) Q_{n-1} \\
&\vdots \\
&= \beta_n R_n + (1-\beta_n)(\beta_{n-1} R_{n-1}) \; - - - \; + \left(\prod_{i=1}^{n} (1-\beta_i)\right) Q_1
\end{aligned}$$

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

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

$$\therefore \prod_{i=1}^{n} (1-\beta_i) = 0 \;\; \therefore \text{ Initial bias is eliminated}$$