(Q)3. In ε-greedy given $\varepsilon = 0, 0.1, 0.01$.

0.01 will perform the best of these 3.

PMF of ε-greedy is as follows

$$P(A_t = a) = \begin{cases} 1 - \varepsilon + \frac{\varepsilon}{10} & , a = \underset{a}{\text{argmax}}\ Q_t(a) \\ \frac{\varepsilon}{10} & , a \neq \underset{a}{\text{argmax}}\ Q_t(a) \end{cases}$$

Long run implies over expectation i.e. $n \to \infty$

Trivially as $n \to \infty$ $E[Q_t(a)] = q^*(a)$ $\forall a \in A_t$ | all arms chosen ∞ no. of times

~~We can say~~

We can say true at a finite but sufficiently large n_0

$E[Q_{n_0}(a)] = q^*(a)$ & from this point onwards the greedy selection is optimal one.

$\therefore$ long run $P(\text{optimal choice}) = P(A_t = \underset{a}{\text{argmax}}\ Q_t(a))$

$= 1 - \varepsilon + \frac{\varepsilon}{10}$

$\left(> 1 - \varepsilon \right)$

$\therefore$ for $\varepsilon = 0.1 \to P(\text{opt}) = 0.91$

$\varepsilon = 0.01 \to P(\text{opt}) = 0.991$

Note: ~~If~~ $\varepsilon = 0$, there no exploration i.e. no PMF

$\therefore$ Over Exp ε should tend to 0 but $\varepsilon \neq 0$

i.e. conditioned on the fact we have explored enough to have guaranteed which arm is optimal

Q4

i) $Q_t(a)$ = sum of rewards ~~before~~ when a taken prior to t / number of times a taken prior to t

$\therefore Q_2(a) = R_1(a)$

$Q_3(a) = \frac{R_2(a) + R_1(a)}{2}$

$\vdots$

$Q_n(a) = \frac{\sum_{i=1}^{n-1} R_i(a)}{n-1}$

$Q_1(a)$ simply allows us to make a first (arbitrary choice)

$$Q_{n+1} = \frac{1}{n} \sum_{i=1}^{n} R_i = \frac{1}{n}\left(R_n + (n-1) Q_n\right)$$

$$Q_n = \frac{1}{(n-1)}\left(R_{n-1} + (n-2) Q_{n-1}\right)$$

$\vdots$

$$Q_2 = \frac{1}{1}\left(R_1 + (2-2) Q_1\right) = 1 \times R_1 + 0 \cdot Q_1 = R_1$$

$\therefore$ For Q_n or sample mean Q_1 is in not a dependency

ii) For const step size α

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

$= \alpha R_n + (1-\alpha) Q_n$

$= \alpha R_n + (1-\alpha)[\alpha R_{n-1} + (1-\alpha) Q_{n-1}]$

$= \alpha R_n + \alpha(1-\alpha) R_{n-1} + (1-\alpha)^2 Q_{n-2}$

$= \alpha R_n + \alpha(1-\alpha) R_{n-1} + \alpha(1-\alpha)^2 R_{n-2} \cdots (1-\alpha)^n Q_1$

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

$\therefore$ as α decreases $(1-\alpha)^n$ increases

$\Rightarrow$ Greater dependence on Q_1

For no dependence coeff of Θ_i Term $= 0$

$\therefore (1-\alpha)^n = 0$

Since α has to be real $\therefore \underline{\alpha = 1}$

Hence $\Theta_{n+1} = R_n$ trivially

Q6.

Until step 10 atleast one $N_t(a) = 0$ for $a \in A_t$

$N_t(a) = 0$ by def^n is considered maximizing reward

∴ Steps 1-10 are spent exploring

Now at step 11 → constant

$$\rightarrow c \sqrt{\frac{\ln t}{N_t(a)}} \quad \text{(ln t → same)}$$

are equal for all $a \in A_t$

$N_t(a) = 1$

∴ we greedily choose action w/ highest reward say a_0

→ ⇒ Spike as we went from complete explorat^n to pure greedy

Now at step 12 $N_t(a_0) = 2$ $N_t(a) = 1$ $\forall a \in A_t / \{a_0\}$

∴ as $c \uparrow$ more weightage is given to non $Q_t(a)$ term & a_0 will be at a disadvantage as its $N_t \uparrow$

∴ As $c \uparrow$ ~~pro~~ $P(\text{Not } a_0) \uparrow$ ∴ we have not picked best rewarding arm so far

Q3. Contd.

Since we are more likely to pick the optimal reward in the long run. So even though until a finite but arbitrarily large n0, e = 0.1 will be greater than e = 0.01. Over the long run or infinite steps i.e. expectation e = 0.01 will overtake and be greater than e = 0.1.

Thus in long run 0.01 or smaller e (!=0) will be greater cumulatively. Since after n0, for infinitely time (or much larger than n0) we are much more likely to lose out on picking the best reward when using a greater e.