

Q30 In ϵ -greedy given $\epsilon = 0, 0.1, 0.01$.

0.01 will perform the best of these 3.

PMF of ϵ -greedy is as follows

$$P(A_t = a) = \begin{cases} 1 - \epsilon + \frac{\epsilon}{10} & , a = \operatorname{argmax}_a Q_t(a) \\ \frac{\epsilon}{10} & , a \neq \operatorname{argmax}_a Q_t(a) \end{cases}$$

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

Trivially as $n \rightarrow \infty$ $E[Q_t(a)] = q^*(a) \quad \forall a \in A_t$ all arms chosen as $n \rightarrow \infty$

~~We can say:~~

We can say true at a finite but arbitrarily large n_0

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

$$\begin{aligned} \therefore \text{Long run } P(\text{optimal choice}) &= P(A_t = \operatorname{argmax}_a Q_t(a)) \\ &= 1 - \epsilon + \frac{\epsilon}{10} \\ &\quad \left(> \underbrace{1 - \epsilon} \right) \end{aligned}$$

$$\therefore \text{for } \epsilon = 0.1 \rightarrow P(\text{opt}) = 0.91$$

$$\epsilon = 0.01 \rightarrow P(\text{opt}) = 0.991$$

Note: ~~at~~ $\epsilon = 0$, there no exploration i.e. no PMF

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

\swarrow
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 action 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 joint (arbitrary choice)

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

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

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

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

ii) For constant step size α

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

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

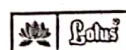
$$= \alpha R_n + (1-\alpha) [\alpha R_{n-1} + (1-\alpha) Q_{n-1}]$$

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

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

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

\therefore as α decreases $(1-\alpha)^n$ increases
 \Rightarrow Greater dependence on Q_1



For no dependence coeffs of Θ , $\Theta_{\text{em}} = 0$

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

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

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

Q6.

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

$N_t(a) = 0$ by defⁿ is considered maximizing reward

\therefore Steps 1-10 are spent exploring

Now at step 11

$$\rightarrow c \sqrt{\frac{\text{Int} \xrightarrow{\text{same}}}{N_t(a)} \xrightarrow{=1}} \rightarrow \text{constant}$$

are equal for all $a \in A_t$

\therefore we greedily choose action w/ highest reward say a_0

\Rightarrow Spike as we went from complete exploratⁿ to pure greedy

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

\therefore as $c \uparrow$ more weight is given to non $Q_t(a)$ term & a_0 will be at a disadvantage as its $N_t = 2$

\therefore As $c \uparrow$ prob $P(N_t(a_0)) \uparrow \therefore$ 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 n_0 , $e = 0.1$ 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 ($\neq 0$) will be greater cumulatively. Since after n_0 , for infinitely time (or much larger than n_0) we are much more likely to lose out on picking the best reward when using a greater e .