Divyam Anshumaan
2017147

RL - H.W. 1

Q3.

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

$\bar{o}_0 = 0$, $n \geq 0$.

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

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

$$= \alpha + (1-\alpha)(\bar{o}_{n-2} + \alpha(1-\bar{o}_{n-2}))$$

$$= \alpha + \alpha(1-\alpha) + (1-\alpha)^2 \bar{o}_{n-2}$$

$$\vdots$$

$$= \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} + (1-\alpha)^{n} \bar{o}_0 \cancel{+1-\alpha}$$

$$\cancel{= \alpha\left(\sum 1-\alpha\right)}$$

$$= \alpha(1-\alpha)^n \left(\sum_{i=1}^{n} \frac{1}{(1-\alpha)^i}\right) + (1-\alpha)^n \times 0$$

$$= \alpha(1-\alpha)^n \cdot \frac{1}{1-\alpha} \cdot \frac{\left(\frac{1}{1-\alpha}\right)^n - 1}{\left(\frac{1}{1-\alpha}\right) - 1}$$

$$= \cancel{\alpha} \cdot \cancel{(1-\alpha)^{n-1}} \cdot \frac{\cancel{(1-\alpha)^{n}-1}}{\cancel{\alpha(1-\alpha)^{n-1}}} \qquad 1-(1-\alpha)^n$$

So, $\beta_n = \dfrac{\alpha}{1-(1-\alpha)^n}$ — (1)

Now, $Q_{n+1} = R_n \cdot \beta_n + (1-\beta_n) Q_n$

$\Rightarrow Q_{n+1} = R_n \cdot \dfrac{\alpha}{\bar{o}_n} + \dfrac{\bar{o}_n - \alpha}{\bar{o}_n} \cdot Q_n$

New

$$Q_n - \alpha = Q_{n-1}(1-\alpha) \quad \text{— (11)}$$

$$= R_n \frac{\alpha}{Q_n} + \frac{Q_n - \alpha}{Q_n} \cdot \left(R_{n-1} B_{n-1} + (1 - B_{n-1}) Q_{n-1}\right)$$

$$= R_n \frac{\alpha}{Q_n} + \frac{Q_n - \alpha}{Q_n} \cdot \frac{\alpha}{Q_{n-1}} R_{n-1} + \frac{Q_n - \alpha}{Q_n} \cdot \frac{Q_{n-1} - \alpha}{Q_{n-1}} Q_{n-1}$$

$$\vdots$$

$$= R_n \cdot \frac{\alpha}{Q_n} + \frac{Q_{n-1}(1-\alpha)}{Q_n} \frac{\alpha}{\cancel{Q_{n-1}}} R_{n-1} + \frac{Q(1-\alpha)^2}{Q_n} \alpha R_{n-2} \ldots$$

$$\ldots + F \cdot (*)$$

$$= \alpha \sum \frac{R_i (1-\alpha)^{n-i}}{Q_n} + 0.$$

→ As R_i's are multiplied by $(1-\alpha)^{n-i}$

the weights are exponentially decreases for

* ——— x ———

(from above) $(*)$

$$\frac{Q_n - \alpha}{Q_n} \cdot \frac{Q_{n-1} - \alpha}{Q_{n-1}} \cdot \frac{Q_{n-2} \cdot \alpha}{Q_{n-2}} \cdots \frac{Q_1 - \alpha}{Q_1} \cdot Q_1 = F.$$

$$= \frac{(1-\alpha) \cdot \cancel{Q_{n-1}}}{Q_n} \cdot \frac{(1-\alpha) \cancel{Q_{n-2}}}{\cancel{Q_{n-1}}} \cdots \frac{Q_0}{\cancel{Q_1}} = F = 0$$

as $Q_0 = 0$.

* Now weights for R_i's $= \dfrac{\alpha(1-\alpha)^{n-i}}{1-(1-\alpha)^n}$

now $\sum \cancel{\alpha} = \dfrac{\alpha}{1-(1-\alpha)^n} \cdot (1-\alpha)^n \sum_{i=1}^{n} \dfrac{1}{(1-\alpha)^i}$

$= \dfrac{\alpha}{1-(1-\alpha)^n} \cdot (1-\alpha)^n \left(\dfrac{1}{(1-\alpha)} \cdot \dfrac{\dfrac{1-(1-\alpha)^n}{(1-\alpha)^n}}{\dfrac{1-1+\alpha}{(1-\alpha)}} \right)$

$= \dfrac{\cancel{\alpha}}{\cancel{1-(1-\alpha)^n}} \cdot (1-\alpha)^{n-1} \cdot \dfrac{\cancel{1-(1-\alpha)^n} \cdot \cancel{(1-\alpha)}}{\cancel{(1-\alpha)^n} \cdot \cancel{\alpha}}$

$= \underline{1}$. Hence weights sum to 1.

---

* Checking for convergence: $\sum_{i=1}^{n} \beta_i = \alpha$

Now, $\beta_n = \dfrac{\alpha}{1-(1-\alpha)^n}$, in general,

~~So $\sum \dfrac{\alpha}{1-(1-\alpha)^n} = $ for $1 > \alpha > 0$,~~

~~$\alpha \; \cancel{\quad} \; \dfrac{1}{\quad}$~~

~~for $1 \geq \alpha \geq 0$, $\sum \beta_n$~~

~~$= \alpha \sum_{n=1}^{\infty} \dfrac{1}{1-(1-\alpha)^n}$~~

~~$= \alpha^2 \sum$~~

Now

$= R_n$

$= R_n$

$= R_n$

$\therefore R$

$= \alpha$

$\rightarrow$

The

θ_{n-}

θ_n

$= ($

$(1-\alpha)\sum(1-\alpha)^n \geq 0$ for $n \geq$

now $1 \geq 1-(1-\alpha)^n \geq 1-(1-\alpha) \geq$ — (iii)

or $\frac{1}{1-(1-\alpha)^n} \geq 1$

~~thus~~ $\cancel{0 \leq \frac{\alpha}{1-(1-\alpha)^n} < 1 \text{ for } 0 \leq \alpha \leq 1}$

now $\frac{\alpha}{1-(1-\alpha)^n} \geq \frac{\alpha}{1}$ (from iii)

and $\sum_{i=1}^{\infty} \frac{\alpha}{1} \leq \sum_{n=1}^{\infty} \frac{\alpha}{1-(1-\alpha)^n}$

so $\sum_{i=1}^{\infty} \beta_i = \infty$

Now checking for $\sum \beta_i^2 < \infty$:

ie, $\sum_{n=1}^{\infty} \frac{\alpha^2}{(1-(1-\alpha)^n)^2}$ $\sum_{n=1}^{\infty} \alpha^2 < \infty$.

Q2.

Stationary Optimistic greedy :

1. Spikes are visible after ~~after~~ ~~in~~ the first pass of the greedy action search. This is :

- Due to an optimistic initial value, which will decrease after it is selected for a given bandit

2. A large spike (40%) ~~optimistic~~ optimal action is selected, immediatly after the first pass since the actual optimal bandit (with the highest expected reward) will have the greatest estimate after the first pass.

3. It initially lags while exploring all bandits but eventually settles as exploration decreases.

Non-Stationary Optimistic greedy :

1. Optimistic greedy initially performs better Because ~~q~~ $q^*(a)$ for bandits has not changed much. But after time passes the initial distribution has changed and it no ~~nor~~ longer has the correct estimate.
~~2. Realistic greedy keeps exploring~~

2. ~~Realistic~~ greedy keeps exploring other ~~est~~ actions and ~~eventually~~ gets a better estimate.

Scanned by CamScanner

Q4.

(Optimal Action)

1. In stationary case, UCB explores all actions first. As $\ln(t)$ is bounded the increments will eventually become negligible and will have explored all actions with lower estimates and more frequent examples explored lesser. This is better than a realistic greedy ~~2.~~ approach that ~~explores~~ keeps exploring suboptimal actions even after a large no. of steps. Optimistic greedy explores actions ~~and then immediately stops~~ due to bias and then immediatly stops exploring due to sample averaging.

- The spike in UCB appears after the first pass over all bandits, as the actual optimal action will have a better estimate and will get selected more often.