

Q3.

$$B_n = \frac{\alpha}{\theta_n}, \quad \theta_n = \theta_{n-1} + \alpha(1 - \theta_{n-1})$$

$$\theta_0 = 0, \quad n \geq 1.$$

now, $\theta_n = \theta_{n-1} + \alpha(1 - \theta_{n-1})$

$$= \alpha + (1 - \alpha)\theta_{n-1}$$

$$= \alpha + (1 - \alpha)(\theta_{n-2} + \alpha(1 - \theta_{n-2}))$$

$$= \alpha + \alpha(1 - \alpha) + (1 - \alpha)^2 \theta_{n-2}$$

$$= \sum_{i=1}^n \alpha(1 - \alpha)^{n-i} + (1 - \alpha)^n \theta_0$$

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

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

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

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

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

So, $B_n = \frac{\alpha}{1 - (1 - \alpha)^n} \quad \text{--- (1)}$

Now, $\theta_{n+1} = R_n \cdot B_n + (1 - R_n) \theta_n$

$$\Rightarrow \theta_{n+1} = R_n \cdot \frac{\alpha}{B_n} + \frac{\theta_n - \alpha}{\theta_n} \cdot \theta_n$$

Also,

$$\theta_n - d = \theta_{n-1}(1-d) - d(1)$$

$$R_n \cdot \frac{d}{\theta_n} + \frac{\theta_n - d}{\theta_n} \cdot (R_{n-1} \theta_{n-1} + (1 - \theta_{n-1}) Q_{n-1})$$

$$= R_n \cdot \frac{d}{\theta_n} + \frac{\theta_n - d}{\theta_n} \cdot \frac{d}{\theta_{n-1}} R_{n-1} + \dots + \left(\frac{\theta_n - d}{\theta_n} \dots \frac{\theta_1 - d}{\theta_1} \cdot Q_1 \right)$$

$$= R_n \frac{d}{\theta_n} + \frac{(1-d)\theta_{n-1}}{\theta_n} \cdot \frac{d}{\theta_{n-1}} R_{n-1} + \dots + F, \quad (\text{from (1)})$$

where $F = \frac{\theta_n - d}{\theta_n} \cdot \frac{\theta_{n-1} - d}{\theta_{n-1}} \dots \frac{\theta_1 - d}{\theta_1} \cdot Q_1$

$$= \frac{(1-d)\theta_{n-1}}{\theta_n} \cdot \frac{(1-d)\theta_{n-2}}{\theta_{n-1}} \dots \frac{(1-d)\theta_0}{\theta_1} \cdot Q_1$$

$$= 0, \quad \text{since } \theta_0 = 0. \quad - (III)$$

$$= \sum_{\theta_n} R_i \frac{(1-d)^{n-i}}{\theta_n} + 0 \quad (\text{from (III)})$$

\therefore coefficients of R_i are exponentially decreasing older rewards.

2 New weights for R_i is $\frac{\alpha(1-\alpha)^{n-i}}{1-(1-\alpha)^n}$

now $\bar{R} = \frac{\alpha}{1-(1-\alpha)^n} \sum_{i=1}^n \frac{1}{(1-\alpha)^i}$

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

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

$= \frac{1}{1-(1-\alpha)^n}$ Hence weights sum to 1.

• Checking for convergence: $\sum_{i=1}^n B_i \leq \infty$

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

~~$\sum_{i=1}^n \frac{\alpha}{1-(1-\alpha)^i}$~~

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

for $1 \geq \alpha > 0$, $\sum B_n$

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

$= \alpha \sum$

$$(1-\alpha) \geq (1-\alpha)^n \text{ for } n \geq 1$$

now

$$1 - (1-\alpha)^n \geq 1 - (1-\alpha) \quad \text{--- (iv)}$$

or

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

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

now

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

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

$$\text{so } \sum_{n=1}^{\infty} B_n \leq \infty$$

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

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

U4.

Stationary Optimistic greedy:

1. Spikes are visible ^{after the} ~~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, immediately 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 $\phi^*(a)$ for bandits has not changed much. But after time passes the initial distribution has changed and it no longer has the correct
2. Realistic greedy keeps exploring estimates.
2. Realistic greedy keeps exploring other ~~est~~ actions and eventually gets a better estimate.

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 less. This is better than a realistic greedy
2. ~~An~~ 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 immediately 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.