* Exercise 1 : Multi-Armed Bandits *

Q.1) Exploration vs. exploitation.

→ we have k-armed bandit problem where $K = 4$

→ Applying a bandit algorithm using ε-greedy action selection

Sample-average action value estimate & initial estimate

$$Q_1(a) = 0 \quad \forall a \qquad \text{---- (1)}$$

→ Initial sequence of actions & rewards

$A_1 = 1, \; R_1 = -1$
$A_2 = 2, \; R_2 = 1$
$A_3 = 2, \; R_3 = -2$
$A_4 = 2, \; R_4 = 2$
$A_5 = 3, \; R_5 = 0$

→ Q. On what timesteps, ε-case occurred Definitely?

$$Q_t(a) = \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{No. of times } a \text{ taken prior to } t}$$

$$= \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i = a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i = a}}$$

- $A_1 = \text{argmax}_a Q_1(a)$

  ↳ from (i) $Q_1(a) = 0$
  → All actions have same weight
  → break ties randomly with probability $(1-\epsilon)$

  $A_1 = 1$
  $R_1 = -1$

- Now for $t=2$, since we got negative reward from action 1, we continue to exploit initial expected value of '0' for A2

  Here, we only have $Q_2(a=1) = Q_1 + \frac{1}{1}\cdot[R_1 - Q_1]$
  $= -1$

  $\forall a = 2 \text{ to } k \quad Q_2(a) = 0$

  Here, algorithm chose

  $A_2 = 2$
  $R_2 = 1$

  This indicates, action at timestep 2 (A_2) was chosen to exploit with '$1-\epsilon$' chance.

- for $t = 3$

  $Q_3(a=1) = -1$

  $Q_3(a=2) = 1$

  $\forall a = 3 \text{ to } k \quad Q_3(a) = 0$

  ∴ Agent decides to exploit $a=2$ again

$A_3 = 2$
$R_3 = -2$

- for $t = 4$

$$Q_4(a=1) = -1$$

$$Q_4(a=2) = \frac{1 + (-2)}{2} = -0.5$$

$$\forall a = 3 \text{ to } 4 \quad Q_4(a) = 0$$

Here, even if $Q_4(a=2) = -0.5$, agent decides $a=2$ again. This step is definitely exploratory.

$A_4 = 2$
$R_4 = 2$

(since agent could have exploit $a=3$ or $a=4$)

- Now, at $t=5$

$$Q_5(a=1) = -1$$

$$Q_5(a=2) = \frac{1 + (-2) + 2}{3} = \frac{1}{3} = 0.33$$

$$\forall a = 3 \text{ to } 4 \quad Q_5(a) = 0$$

here, agent goes to Definitely explore $a=3$ with ϵ

$A_5 = 3$
$R_5 = 0$

final Q esimates for $t+1 = 6$

$Q_6(a=1) = -1$ $\quad$ $Q_6(a=3) = \frac{0}{1} = 0$

$Q_6(a=2) = 0.33$ $\quad$ $Q_6(a=4) = 0$

Thus, agent definitely explored on $t=4$ & $t=5$ (A_4, A_5)

But, it could have possibly explored on any other timestamp as well.

Q.2) Varying Step-Size weights

→ When step size 'α' is Constant.

Q-value estimate for reward is given by:

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

→ Now when step size is not Constant (following the proof for eq. 2.6)

Consider step size is 'α_n' at each step n.

from 2.6 updation rule,

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

i.e. $Q_{n+1} = Q_n + \alpha_n [R_n - Q_n]$

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

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

$$= (1-\alpha_n)\left\{Q_{n-1} + \alpha_{n-1}(R_{n-1} - Q_{n-1})\right\} + \alpha_n R_n$$

$$= (1-\alpha_n)\left\{Q_{n-1} + (\alpha_{n-1} \cdot R_{n-1}) - (\alpha_{n-1} \cdot Q_{n-1})\right\} + \alpha_n R_n .$$

$$= (1-\alpha_n)\left\{\alpha_{n-1} \cdot R_{n-1} + (1-\alpha_{n-1})\, Q_{n-1}\right\} + \alpha_n R_n .$$

$$= (1-\alpha_n) \cdot (1-\alpha_{n-1}) \cdot Q_{n-1} + \alpha_{n-1} R_{n-1} (1-\alpha_n) + \alpha_n R_n .$$

$$Q_{n+1} = (1-\alpha_n) \cdot (1-\alpha_{n-1}) \cdot Q_{n-1} + \alpha_n R_n + (1-\alpha_n)\, \alpha_{n-1} R_{n-1}$$

Generalizing .

$$Q_{n+1} = \left(\prod_{i=1}^{n} (1-\alpha_i)\right) Q_1 + \sum_{i=1}^{n}\left[\alpha_i R_i \prod_{j=i}^{n-1} (1-\alpha_j)\right]$$

... for $i=0$ assume $\alpha_0 = 1$

→ Weighting on each of the prior reward is

$$\sum_{i=1}^{n} \alpha_i \prod_{j=1}^{n-1} (1-\alpha_j)$$ for every reward $R_i \in [1, n]$

Q.3) Bias in Q-value estimates:-

we say that an estimate is biased if the expected value of the estimate does not match the true value i.e. $E[Q_n] \neq q_*$ (otherwise, it is unbiased)

a) Sample Average Estimate in equation 2.1

$$Q_t(a) \doteq \frac{\text{Sum of rewards taken prior to 't'}}{\text{number of times 'a' taken prior to t}} \quad \cdots (2.1)$$

$$= \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

random var.

$\mathbb{1}_{\text{predicate}}$ = { 1 if predicate is true; 0, Otherwise

This estimate is **Unbiased!**

→ when we have not selected an action 'a' prior to 't', denominator in 2.1 will be 0. Thus we instead $Q_t(a)$ to some default value (example: 0)

→ Now as we continue our experiment on multiple trials; where each trial consists of large number of action selection step, our sample average $Q_t(a)$ converges to $q_*(a)$ in limit.

where limit is NOT just $t \to \infty$ but denominator $N_t(a) \to \infty$ as well.

∴ (by law of large Numbers)

parts b to e →

Consider exponential-recency weighted average estimate

$$Q_{n+1} \doteq Q_n + \alpha(R_n - Q_n) \qquad 0 < \alpha \leq 1 \quad \ldots (2.5)$$

b) If $Q_1 = 0$ is Q_n for $n > 1$ biased?

→ from eqn. 2.6 we can write eqn. (2.5) as

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

Suppose $Q_1 = 0$ & $n = 1$

$$\therefore Q_2 = \alpha(1-\alpha)^0 R_1$$

$$Q_2 = \alpha R_1$$

similarly for $n = 2$

$$Q_3 = \alpha\{(1-\alpha)R_1 + R_2\}$$

$$Q_3 = \alpha(1-\alpha)R_1 + \alpha R_2$$

for $n = 3$

$$Q_4 = \alpha(1-\alpha)^2 R_1 + \alpha(1-\alpha)^1 R_2 + \alpha R_3$$

& so on.....

need 1

$$\therefore Q_{n+1} = \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} E[R_i]$$

Consider as $n \to \infty$ by LLN,

$E[R_i] = q^*$

$\therefore$ Solving for α

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

$$= \alpha \left[1 + (1-\alpha) + (1-\alpha)^2 + \cdots (1-\alpha)^{n-1} \right]$$

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

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

as $0 < \alpha < 1$

$\therefore Q[n+1] \neq q^*$

Thus even if $Q_1 = 0$ our estimate for Q_{n+1} can never reach its true value.

Thus it is Biased.

c) Derive conditions for when Q_n will be unbiased.

→ we can write Q_{n+1} as

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

from previous question we can say that for unbiased Q_{n+1} we need $E[Q_{n+1}] = q^*$

$$\therefore E[Q_{n+1}] = E[(1-\alpha)^n Q_1] + E\left[\sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i\right]$$

Condition 1 :-

$$(1-\alpha)^n E[Q_1] = 0$$

$$\therefore E[Q_1] = 0 \quad \text{i.e. } Q_1 = 0$$

Condition 2 :-

$$E\left[\sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i\right] = q^*$$

... To prove

multiplying & dividing by n.

LHS =

$$E\left[\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} \cdot n \cdot \frac{R_i}{n}\right]$$

$$= n \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} \cdot E\left[\frac{R_i}{n}\right]$$

$E\left[\frac{R_i}{n}\right]$ is nothing but sample average estimate, which converges to q^* by law of large numbers as $n \to \infty$.

∴ we need

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

$$\Rightarrow \alpha\left[1 + (1-\alpha) + (1-\alpha)^2 + \dots + (1-\alpha)^{n-1}\right] = \frac{1}{n}$$

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

$$\Rightarrow n - n(1-\alpha)^n = 1$$

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

⇒ taking log on both sides we get,

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

Necessary condition (2)

$$\therefore \alpha = 1 - e^{\log\left(\frac{n-1}{n}\right)/n}$$

Thus by condition (1) & (2) we can say that Q_n will be unbiased.

---

d) Show that Q_n is asymptotically unbiased!

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

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

as $n \to \infty$ $\forall\, 0 < \alpha < 1$; $(1-\alpha)^n \to 0$

Now,

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

$$= \alpha\left[\frac{1}{1-(1-\alpha)}\right]$$ ... (infinite sum of a geometric progression)

$$= 1$$

$$\therefore E[Q_{n+1}] = 0 + \sum_{i=1}^{\infty} 1 \cdot E[R_i] = q^* \quad \text{... proved}$$

---

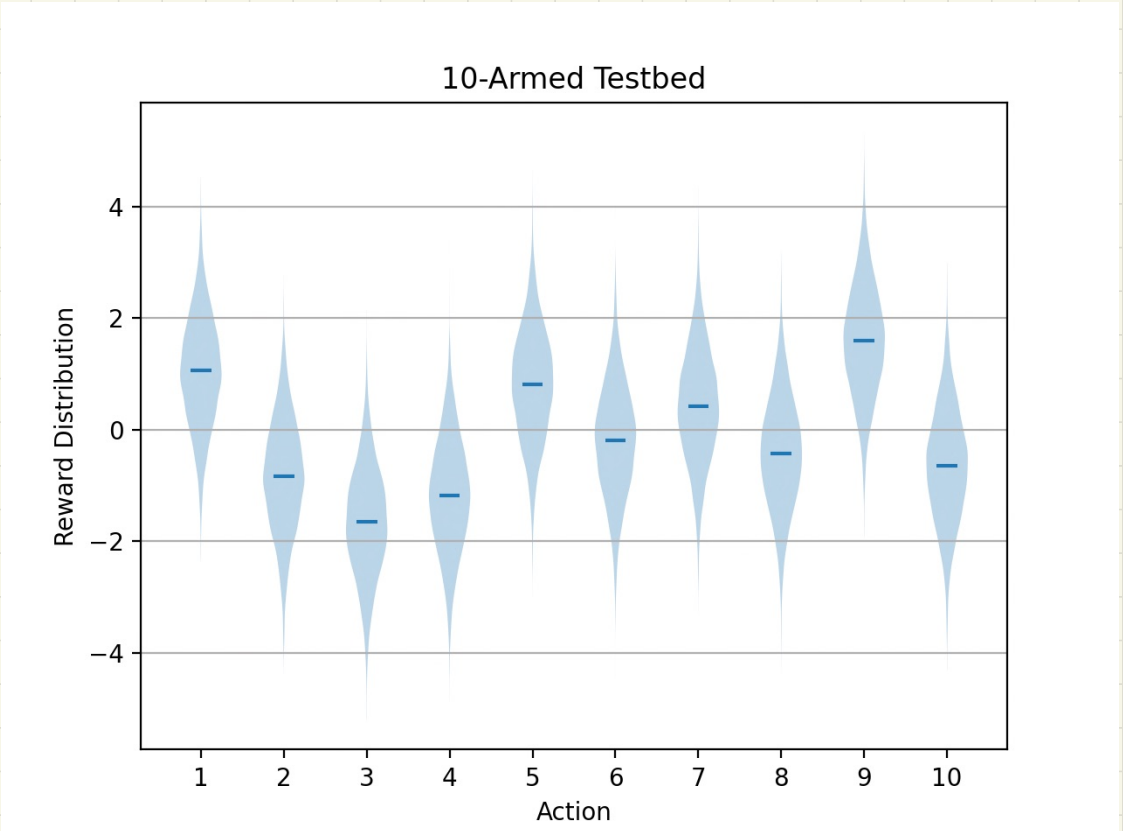
e) Why exponential recency weighted average will be biased in general?

→ Exponential recency weighted average suffers from initial estimate problem

→ Since Q_{n+1} depends on initial action value estimate $Q_1(a)$, it is generally biased

→ Unlike, sample-average estimate, this bias won't disappear due to constant value of α. It is permanent, even though decreasing over time.

Q4: Code Attached
Plot:



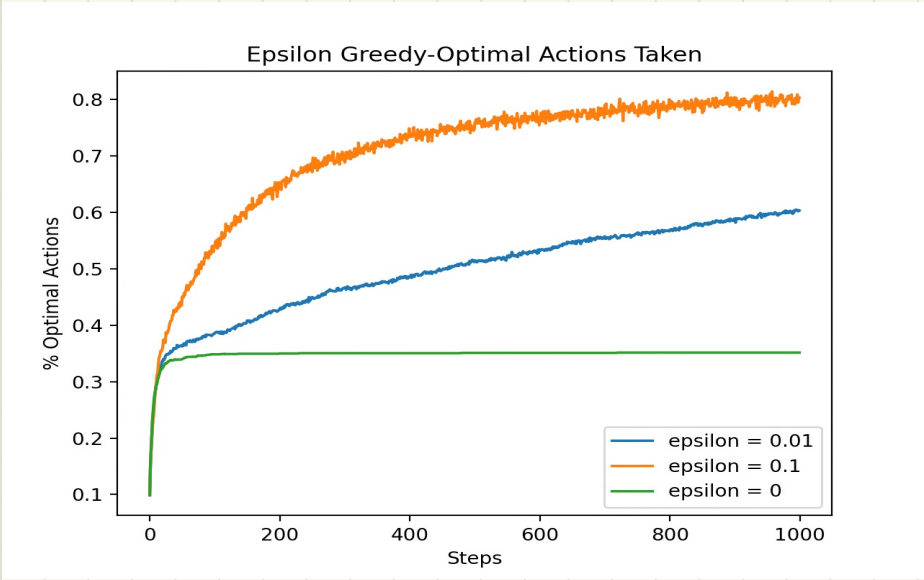
Q5: Predicting Asymptotic Behaviour in 2.2
Ans:
- e = 0.01 will perform better in the case of stationary test bed in the long run. The reason for this behavior is, once it has found optimal action, It'll exploit it with 0.99 chance. The learning curve can take more time to reach optimal action since it has low exploration probability (Epsilon). But once done, the cumulative reward will be maximum out of all three epsilons.

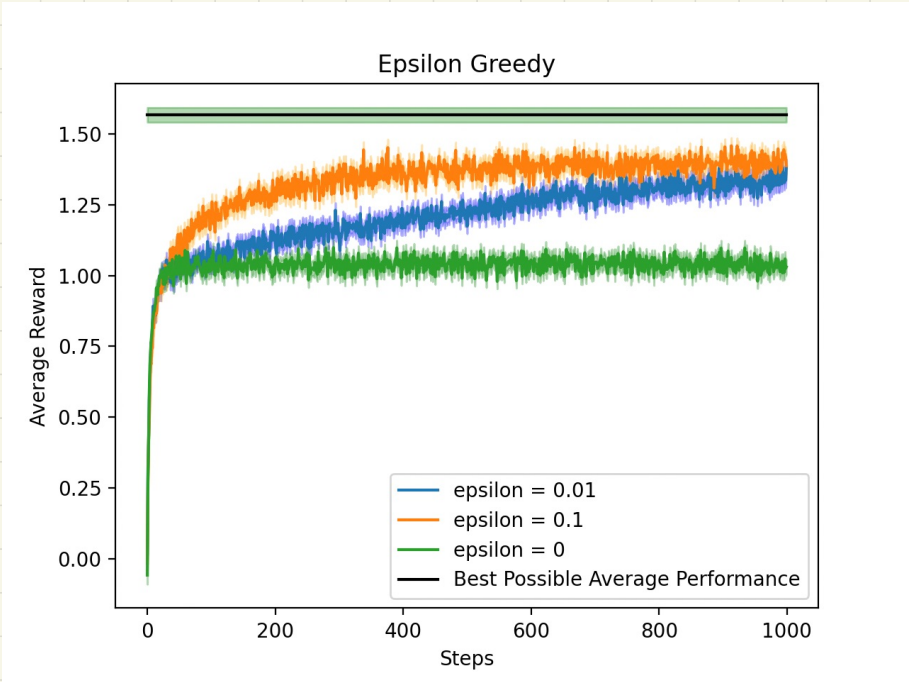
- Comparing with e = 0.1, this epsilon will explore more thus having better initial estimates. But in the end it'll keep on exploiting for just 90% of time. So in the long run, cumulative reward will be less than that of e = 0.01

- Asymptotically, number of times optimal action chosen for e= 0.01 will be; 1/e = 100
similarly for e= 0.1; 1/e = 10

Thus e=0.01 will be better than e = 0.1 for 90% of time

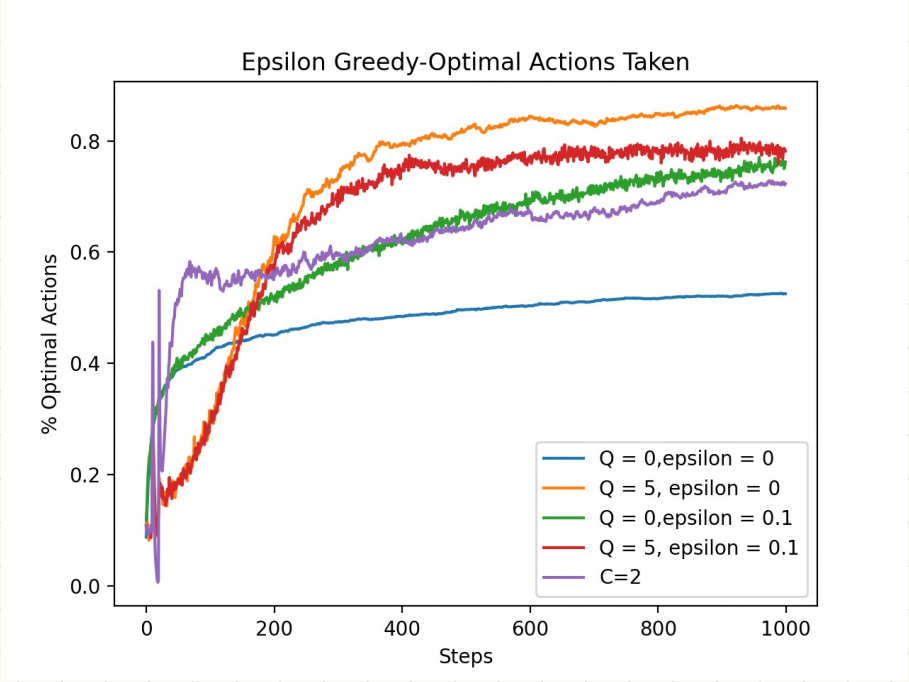
Q6: Reproducing Figure 2.2
Code: Attached
Outputs: Confidence Bands are less visible because of sample size of 2000 at each step

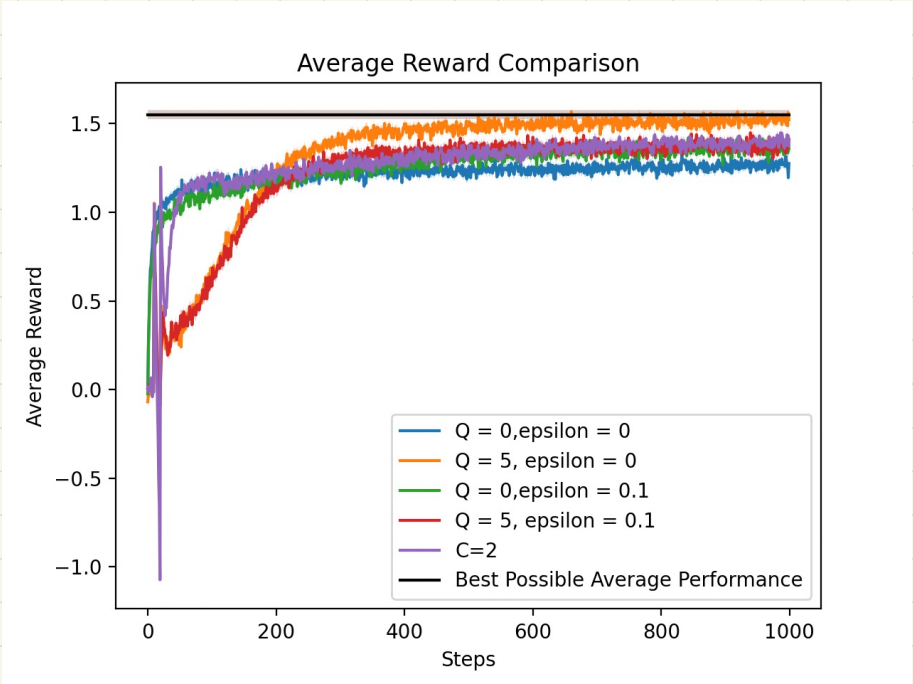




written ans Q6:
In previous Question we predicted e = 0.01 will perform better than e = 0.1 asymptotically. This can be seen from the average reward plot in this question as well.
- Additionally, percentage of optimal action taken for e=0.01 will also take over e = 0.1 as we increase number of steps in each trial. (We can see line trend from the graph)

Q7: code Attached
Plot:





Written Ans Q7:

- As we can see from the graph we get a spike at around step 11 and it drops later. This happens in both optimistic initialization and UCB with hyper parameter setting (c=2, alpha=0.1)

- For optimistic Initialization, since the Q value initialized is 5, this forces algorithm to explore for first ten steps. All the actions are explored randomly until all the 10 arms have pulled once. This updates the Q value in initial 10 steps.
Once all 10 steps have been tried, the Q value gets updated. thus algorithm is able to make better decision and starts exploiting at 11th step, hence the spike in rewards.

- In UCB as well, the spike occurs at step 11. Initially when Nt(a)=0 for all actions, the algorithm tries each action atleast once while breaking ties randomly. On the 11 th step all the actions have been taken and we can now take the action which is most optimal till now. But from 12 th step, algorithm is supposed to explore more options to find bounds for each reward estimate. Thus the reward subsequently decreases.