Ravina Lad

RL1 : Bandits

Q.1 k-armed bandit → k = 4.

Using ε-greedy action selection algorithm,

Given:

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

Q: On Which timestep ε-case occured → 1) definitely ?
2) May be ?

→ ① $Q_t(a) = \frac{\sum (\text{Rewards when a taken prior to 't'})}{\text{No. of times a taken prior to 't'}} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i = a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i = a}}$

① $A_0 = 0$, $R_0 = 0$ $\boxed{Q_1 = 0}$ $Q_2 = 0$, $Q_3 = 0$, $Q_4 = 0$
choose 1 (Maybe Random) As $Q_1 = Q_2 = Q_3 = Q_4 = 0$.

② $A_1 = 1$ $R_1 = -1$ $Q_1 = \frac{-1}{1} = -1$ $\boxed{Q_2 = 0}$ $Q_3 = 0$, $Q_4 = 0$.

Now for $t = 2$, We got Reward = -ve. from A_1 (action 1) We continue to exploit initial expected val for '0' → A_2

Choose 2, May be Random
$Q_1 = -1 < Q_2 = Q_3 = Q_4 = 0$.
→ consistent with ε-greedy selection.

③ $A_2 = 2$, $R_2 = +1$ $Q_1 = -1$, $Q_2 = \frac{1}{1} = 1$ $Q_3 = 0$, $Q_4 = 0$.

→ Action A_2 selected to exploit
→ Maybe random selection Action 2 again (agent decides to exploit again)
As → $Q_2 = 1 > Q_3 = Q_4 = 0 > Q_1 = -1$

④ $A_3 = 2 \quad R = -2 \quad Q_1 = -1, \quad Q_2 = \frac{1+(-2)}{2} = -0.5$

$Q_3 = 0, Q_4 = 0.$

$Q_3 = Q_4 = 0 > Q_2 = -0.5 > Q_1 = -1$

Now, according to ε-greedy action selection → it should select 3 OR 4.

Here, even though it should select 3/4, agent decides Action 2 again. This step is → definitely Exploratory.

⑤ $A_4 = 2 \quad R_4 = 2 \quad Q_1 = -1, \quad Q_2 = \frac{1-2+2}{3} = \frac{1}{3} = +0.33.$

[Agent could have exploited action 3/4)

$Q_2 = 0.33 > Q_3 = Q_4 = 0 > Q_1 = -1.$

According to ε-greedy action selection Algo – it should select action 2 again

But agent definitely decides to explore 3. (Random).

⑥ $A_5 = 3 \quad R_5 = 0 \quad Q_1 = -1 \quad Q_2 = 0.33 \quad Q_3 = \frac{0}{1} = 0$

$Q_4 = 0.$

Thus, agent definitely explored A_4 & A_5 at $t = 4$ & $t = 5$.

But, it could have explored on (A_1, A_2, A_3) at $t = 1, 2, 3$ timestamps aswell.

Q.2 α_n = step-size parameter = non-constant.

→ When α = const.

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 that step-size isn't const

Let α_n = stepsize at each step n. from eq 2.6.

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

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

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

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

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

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

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

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

Hence, weighting on each of the prior reward is →

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

Q.3 Bias in Q-value estimates:

a) Sample avg. estimate in eqn 2.1 →

We say → estimate is biased if the expected value of the estimate doesn't match the true value.

ie $E[Q_n] \neq q_*$ (otherwise it's unbiased)

$$Q_t(a) = \frac{\text{sum of Rewards taken prior to 't'}}{\text{No. of times 'a' taken}} = \frac{\sum_{i=1}^{t-1} R_i \mathbb{1}_{A_i = a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i = a}} \quad \ldots 2.1$$

$$Q_n = \frac{R_1 + R_2 + R_3 \ldots R_{n-1}}{n-1}$$

$$E[Q_n] = E\left[\frac{R_1 + R_2 + R_3 + \ldots R_{n-1}}{n-1}\right] = \frac{1}{n-1} E[R_1 + R_2 + \ldots R_{n-1}]$$

$$= \frac{1}{n-1}\left(E[R_1] + E[R_2] + \ldots E[R_{n-1}]\right)$$

$$= \frac{1}{n-1}(n-1) q_* = q_* \rightarrow \text{Thus, it is unbiased.}$$

random var → $\mathbb{1}_{\text{predicate}}$ = { 1 → predicate is true; 0 → otherwise }

① Here, when will not select an action 'a' prior to 't'

b) denominator = 0 in eq^n 2.1. → $Q_t(a)$ → some default value.

② upon multiple trials, trial = finite large no. of action selection step.

sample avg. $Q_t(a)$ converges to $q_*(a)$ in limit. ... by law of large numbers.

b) if $Q_1 = 0$ for $n > 1$:

$$Q_{n+1} = Q_n + \alpha(R_n - Q_n) \quad 0 < \alpha < 1 \quad \ldots 2.5.$$

exponential-Recency weighted avg. estimate

→ from eq 2.6,

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

if $Q_1 = 0$ is Q_n for $n > 1$: ?

1) $Q_1 = 0$ & $n = 1$

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

$$Q_2 = \alpha R_1$$

2) $Q_1 = 0$ & $n = 2$

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

3) $n = 3$

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

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

As $n \to \infty$ → By large law of numbers,

$$E[R_i] = q^*$$

→ Let's solves $\alpha \to \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i}$

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

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

$$= 1 - (1-\alpha)^n \neq 1 \quad \text{as } 0 < \alpha < 1$$

$$\therefore Q_{[n+1]} \neq q^*$$

Hence, even if $Q_1 = 0$, our estimate for Q_{n+1} can never reach it's true value.

→ Biased.

c) derive condition for - q_n will be unbiased.

Let'

$$q_{n+1} = (1-\alpha)^n q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i \quad \text{... prev. eq}^n$$

for unbiased q_{n+1}.

we need $E[q_{n+1}] = q^*$

$$E[q_{n+1}] = E[(1-\alpha)^n q_1] + E\left[\sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i\right]$$

$$= (1-\alpha)^n q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} E[R_i]$$

$$= (1-\alpha)^n q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} \cdot q^* = q^*$$

... it is unbiased.

i.e. condition → $q_1 = 0$.

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

d)

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 $n \to \infty$; $0 < \alpha < 1$ $\quad (1-\alpha)^n \to 0$

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

$$= \alpha \left(\frac{1}{1-(1-\alpha)}\right) \quad \text{(infinite sum of G.P.)}$$

$$= 1$$

$$\therefore E[Q_{n+1}] = 0 + \sum_{i=1}^{n} 1 \cdot E[R_i] = q_* \quad *$$

$\therefore$ it is unbiased when $n \to \infty$

c)

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

2) We might obtain large values that still lead to possibility of bias.

3) exponential-recency weighted avg. suffers from initial estimate problem.

4) Unlike sample-avg. estimates, bias won't disappear, due to const value of α, it is always there, even though decreasing over time.

Q.5 Predicting Asymptotic behaviour in fig 2.2

(A) $\varepsilon = 0.01$ → perform better → case: stationary testbed in long run

→ Reason → once it has found optimal action, it will exploit it with 0.99 chance.

→ learning curve can take more time to reach optimal action, as the exploration prob = low.

→ But, cumulative reward = Maximum.

$$\text{Optimal Action} = \left[(-0.01) \times 1 + 0.01 \times \frac{1}{10} \right] \times 100\% = 99.1\%.$$

(B) $\varepsilon = 0.1$ → 10% explore more → it will have better initial estimates. But it will keep on exploiting 90% of time. so in the long run,

cumulative reward (ε[0.1]) < cumulative reward (ε<0.01)

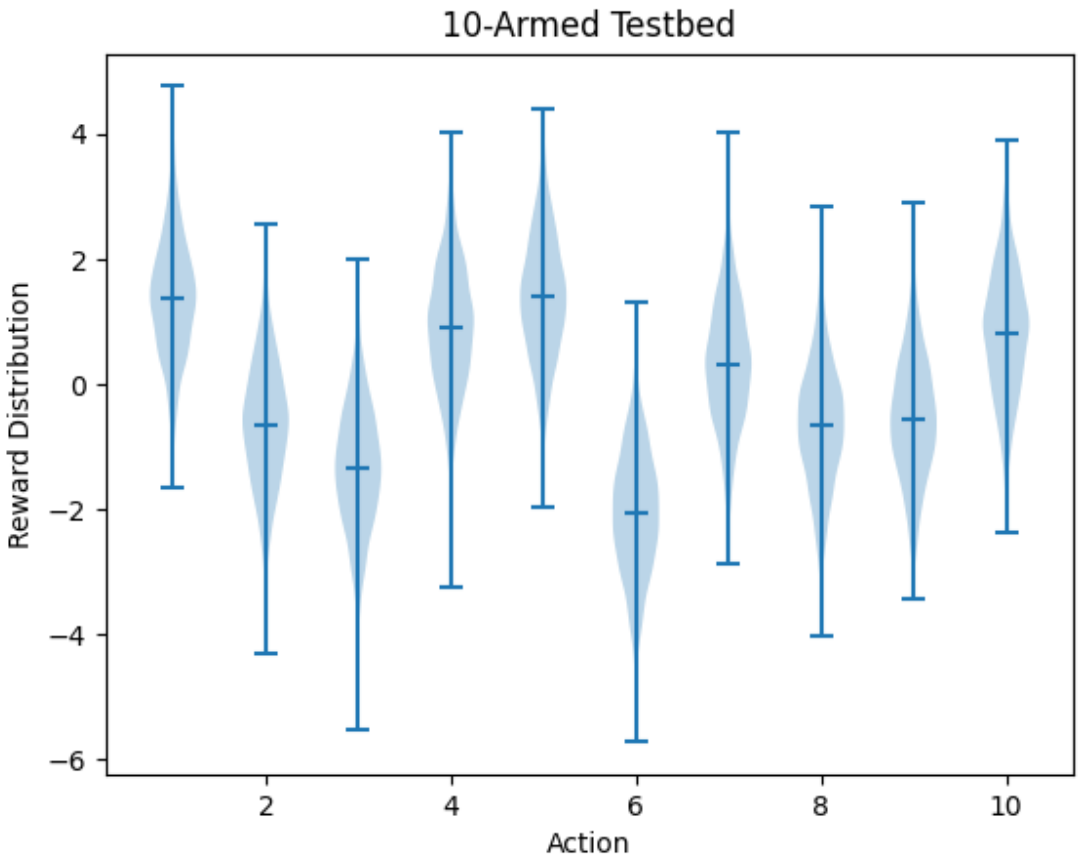
$$\text{optimal action} = \left[0.1 \times 1 + 0.1 \frac{1}{10} \right] \times 100\% = 91\%$$

(C) $\varepsilon = 0$ it will select first +ve result & refuse to make new attempt.

Reinforcement Learning - EXP1
Bandits

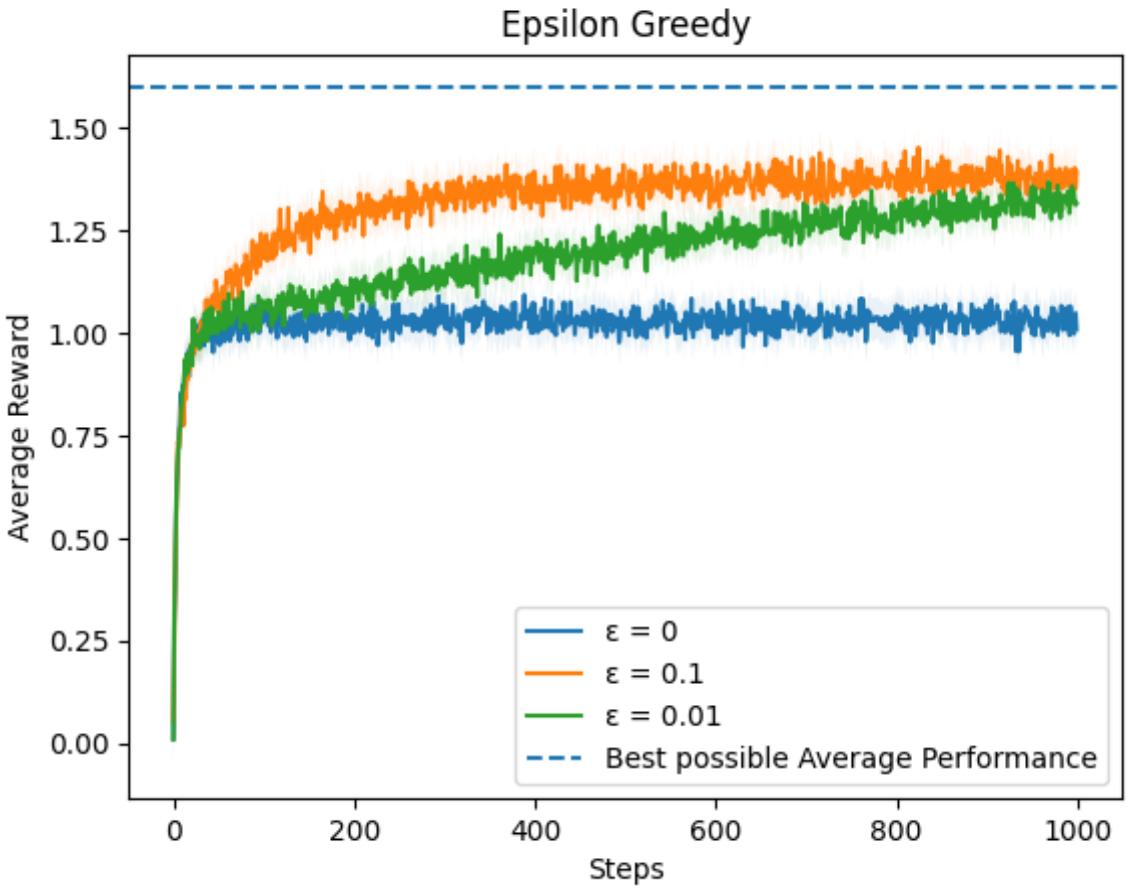
Q4)

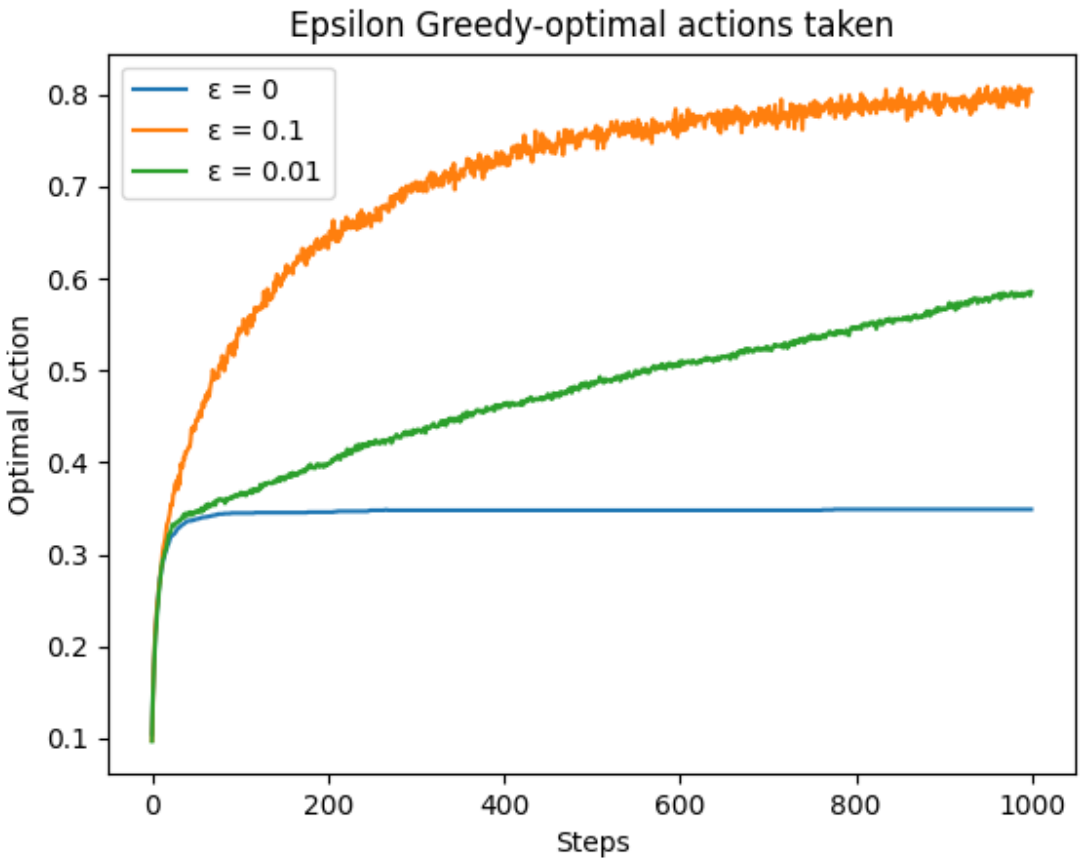
Plots



Q6) Do the averages reach the asymptotic levels predicted in the previous question?
Ans - >
We don’t have enough steps to reach asymptotic levels predicted in the previous question. The more we are closer to infinity, the better the performance.
Here, the % of optimal action taken for epsilon = 0.01 will soon take over epsilon = 0.1 as we increase the number of steps.
As we can also see line trends from below graphs.
We also predicted that - epsilon 0.01 will perform better than 0.1 asymptotically. We can observe this in the average reward plot.

Plots





Q7) : Observe that both optimistic initialization and UCB produce spikes in the very beginning. In lecture, we made a conjecture about the reason these spikes appear. Explain in your own words why the spikes appear(both the sharp increase and sharp decrease). Analyze your experimental data to provide further empirical evidence for your reasoning.

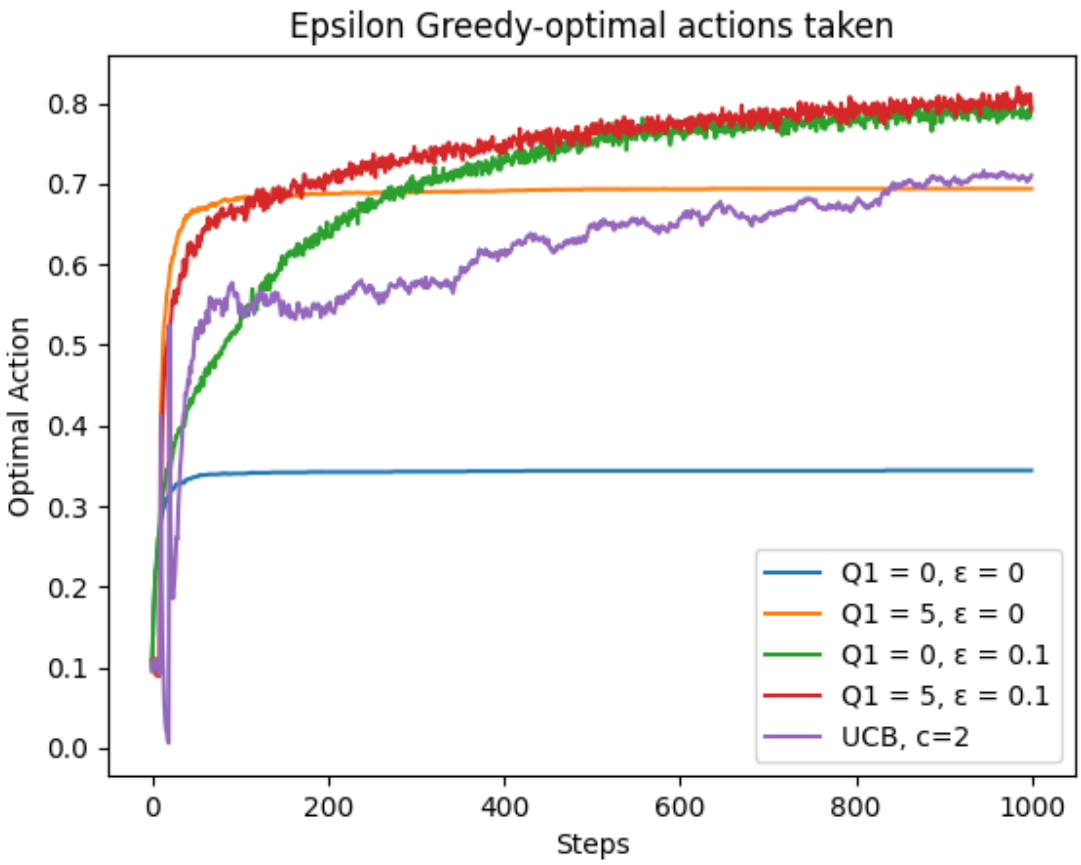
Ans ->

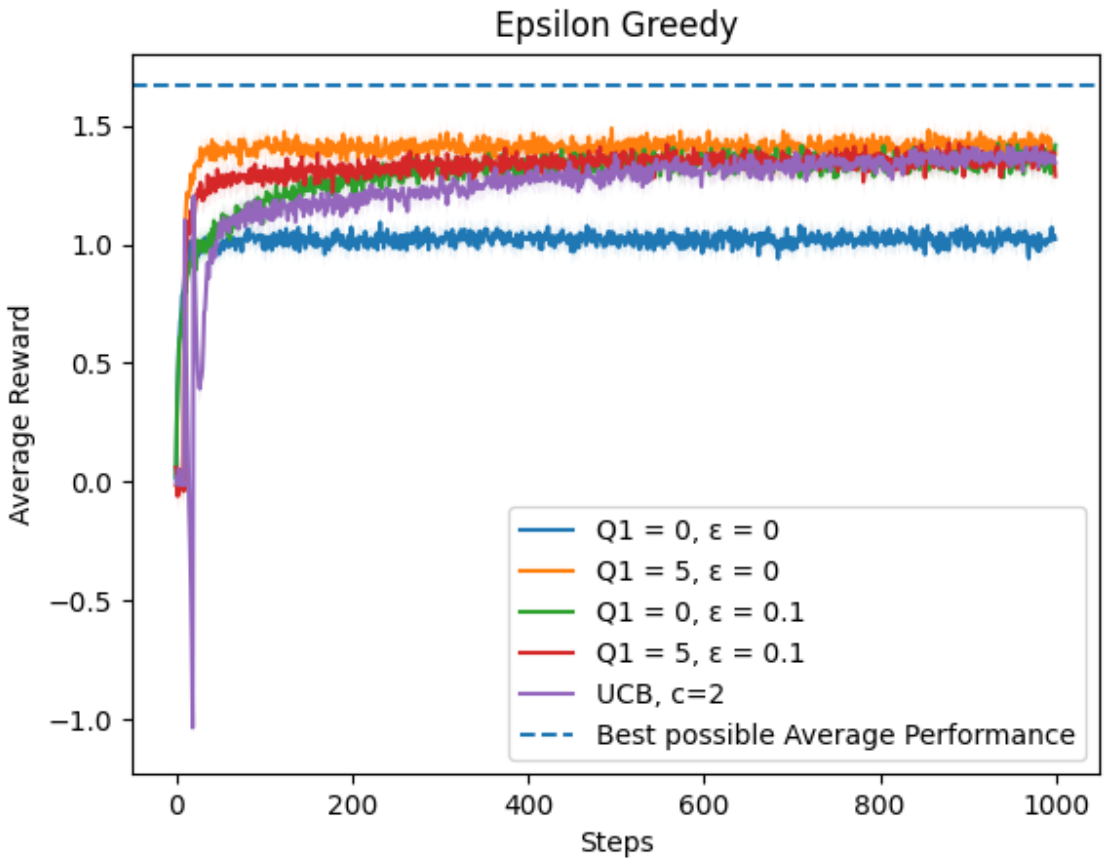
We can observe that - spike is at instance 11, and it drops after that. This happened in both optimistic initialization and UCB. Here, K = 10, we first try all 10 actions, and record the respective rewards. Thus, at t = 11, we will first find the high reward, hence we observe that spike.

A. Optimistic initialization - Q value = 5 initially. Hence, the algo will explore the first 10 steps randomly until we pull all 10 arms. Once all steps have been tried, Q value gets updated and the algorithm makes a better decision and starts exploiting at t = 11.

B. UCB - spikes at 11 only. When Nt(a) = 0 for all actions, while breaking ties randomly, it tries each action. At 11 th step, we can now take the action which is most optimal. But at 12 th step, it would explore more options to search bounds for each reward estimate. And hence, rewards decrease.

Plots


Epsilon Greedy-optimal actions taken
Optimal Action
Steps
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0
200
400
600
800
1000
Q1 = 0, ε = 0
Q1 = 5, ε = 0
Q1 = 0, ε = 0.1
Q1 = 5, ε = 0.1
UCB, c=2


Epsilon Greedy
Average Reward
Steps
1.5
1.0
0.5
0.0
−0.5
−1.0
0
200
400
600
800
1000
Q1 = 0, ε = 0
Q1 = 5, ε = 0
Q1 = 0, ε = 0.1
Q1 = 5, ε = 0.1
UCB, c=2
Best possible Average Performance