1)

The estimated value of action a at time step t. We assume that for all a, the initial estimate $Q_1(a)=0$.

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}},$$

$$A_t \doteq \arg\max_a Q_t(a),$$

Q1 = 0

Here's how you calculate the Q-values:

After A1: Q (A1) = − 1/ 1 = −1

After A2: Q(A2) = 1/ 1 = 1

After A3: Q(A2) = (1+(−2))/2 =−0.5

After A4: Q(A2) = (1 + (− 2) + 2)/ 3 = 1/3

Considering the ε-greedy selection procedure and the order in which the rewards are awarded:

A definitive exploration took place at:

Since it's the initial action and all Q-values are the same, the answer is **A1**.

The possible exploration took place at:

A4, indicating that A2 might not have been the avaricious decision if it was selected once more despite a prior negative reward (R3 = -2).

A5, as A3 was the initial selection and might not have been the avaricious option if alternative actions had been evaluated with higher estimated values.

Without knowing the value of ε, we cannot be certain about other steps, but any step where an action is chosen that doesn't have the highest estimated value could involve exploration.

2)

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

Step size may vary with step (e.g., sample average):

$$\alpha_n(a) = \frac{1}{n}$$

To ensure convergence with probability 1
(result from stochastic approximation theory):

$$\sum_{n=1}^{\infty} \alpha_n(a) = \infty \quad \text{and} \quad \sum_{n=1}^{\infty} \alpha_n^2(a) < \infty$$

Large enough to overcome initial bias

Small enough to guarantee convergence

Q.2] Varying step-size weights:

when α_n is non-stationary;

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

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

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

$$= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1} \cdot R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})\left[\alpha_{n-2} R_{n-2} + (1-\alpha_{n-2}) Q_{n-2}\right]$$

$$= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1} R_{n-1} + \cdots + Q_1 \prod_{i=1}^{n} (1-\alpha_i)$$

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

3)

4)

From Figure 2.2, for long-term cumulative reward and probability of selecting the best action: ε = 0 (greedy) may initially lead but could miss the best action if it's not identified early.

ε = 0.1 offers consistent exploration, potentially discovering the true best action but with a lower overall exploitation rate.

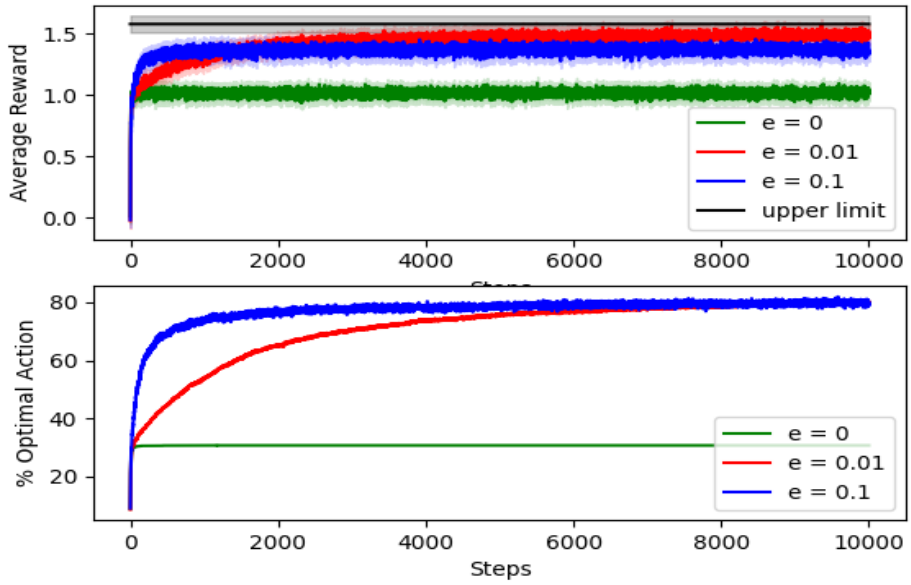
ε = 0.01 provides a balance, likely outperforming ε = 0.1 in cumulative reward due to higher exploitation while still correcting early mistakes through minimal exploration.

Asymptotically: ε = 0's probability of best action choice approaches 100% if correctly identified early.

ε = 0.1's probability is about 91% due to exploration.

ε = 0.01's probability is about 99.1%, balancing exploration and exploitation effectively.

5)



Overall, while the ε = 0 method seems to reach the highest level of optimal action selection as predicted, it does not achieve the highest average reward, which could mean it does not always identify the best action early. The ε = 0.1 and ε = 0.01 methods perform as expected, with ε = 0.01 likely achieving better long-term rewards due to its balance of exploration and exploitation. The actual asymptotic levels will require further analysis beyond the 1000 steps shown to confirm if they align precisely with predictions.

6)

a)

Bias : n Q-value estimates:-

$\because Q_n = \frac{R_1 + R_2 + \ldots + R_{n-1}}{n-1}$; $E(R_n) = q_*$

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

$= \frac{1}{n-1} \times E[R_1 + R_2 + \ldots + R_{n-1}]$

$= \frac{1}{n-1} \times E(R_1) + E(R_2) + \ldots + E(R_{n-1})$

$= \frac{1}{n-1} \times (n-1)\, q_*$

$= q_*$ which is unbiased.

b.] ∵ If $Q_1 = 0$ for $n > 1$;

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

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

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

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

∴ This would be biased only when $\sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} = 1$.

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

∵ when $E|Q_{n+1}| = q*$, which it is unbiased.

& when $E|Q_{n+1}| = E\left|(1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i\right|$

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

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

$$= q*$$

∴ which is unbiased.

∵ But for this the necessary condition is;

$\alpha = (1-\alpha)^n Q_1 = 0 \rightarrow Q_1 = 0$

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

∴ By these conditions we can say it is unbiased.

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

$\because$ when $n \to \infty$; where $0 < \alpha < 1$

& $(1-\alpha)^n Q_1 \to 0$

$\therefore \sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} = 1 - (1-\alpha)^{n-1}$ & $(1-\alpha)^{n-1} \to 0$

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

where the above eq^n. satisfies condition ; $Q_1 = 0$; $\sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} = 1$

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

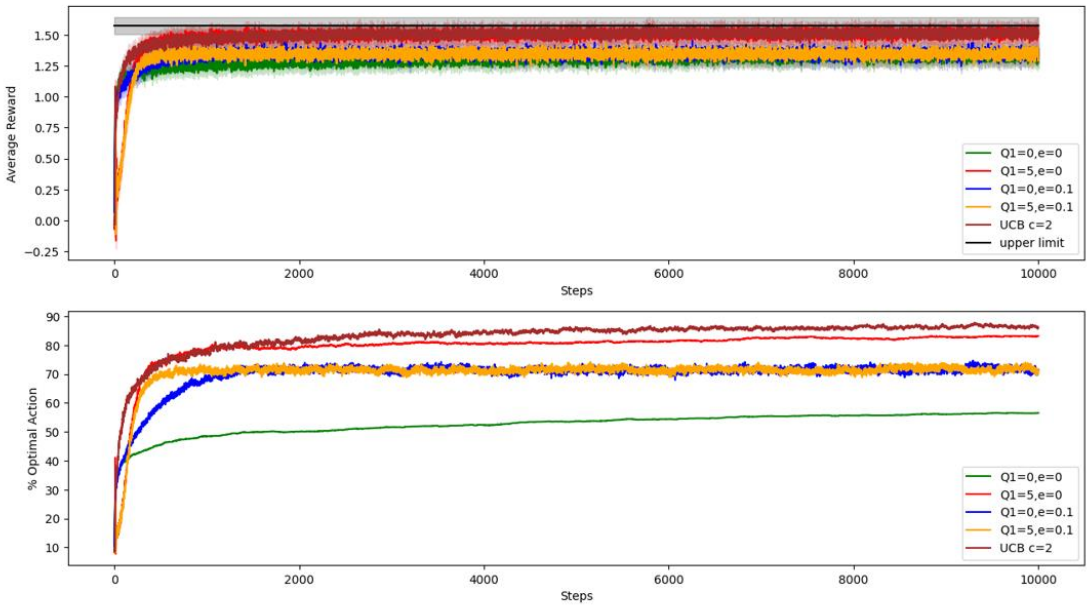
3.e) i) Here, exponential recently weighted average suffers from initial estimate problem where actual situation will never be perfect.

ii) we can never obtain infinite values, but we can have values large enough that there is possibility of some bias.

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

iii) The bias is more common to occur. Unlike sample-average estimate, this bias won't disappear by constant value of α. It would be permanent, even in decreasing over time.

7)



Due to optimistic initialization and UCB with the hyper-parameter setting at c=2 and α=0.1, we get a spike at step 11 which drops down later.

Even if the Q value is initialized 5; it forces the algorithm to explore for the first 10 steps. In this, all the actions are randomly explored, and all 10 arms are at least pulled once. This makes Q=10. All these steps are done then the algorithm makes a better decision & and starts exploiting at the 11th step hence we get the spike in rewards.

Whereas in UCB as well, the spike occurs at step 11. Initially, at N_t(a)=0, the algorithm tries every action at least once while breaking ties randomly. Here all the actions are taken till step 11th but from the 12th step algo. explores more options to find bounds for each reward estimate so the reward decreases afterward. In this, after 10 actions he will adjust to have a more stable performance which leads to another spike.

9)

Written:

Effect of annealing the exploration parameter ϵ time on the performance of the epsilon-greedy agent. Annealing refers to gradually reducing the exploration rate as the agent learns more about the environment. The hypothesis is that as the agent becomes more confident in its knowledge of the action values, it should explore less and exploit more. Experimental Setup: Objective: Compare the performance of a constant ϵ strategy against an annealing ϵ strategy.

Environment: 10-armed bandit problem with rewards drawn from a normal distribution with a mean of 0 and standard deviation of 1. Agents: A standard epsilon-greedy agent with ϵ=0.1. An annealing epsilon-greedy agent with ϵ starting at 0.1 and decreasing by a factor of 0.99 each step until it reaches a minimum of 0.01. Metrics: Average reward and percentage of optimal action taken over time. Trials: 2000 trials for averaging out randomness. Steps: 10000 steps per trial.

Implementation: The annealing epsilon-greedy agent can be implemented by modifying the chosen action method to reduce ϵ after each step according to the annealing schedule.

Results: The results will be presented using two plots: A plot of the average reward over time for both agents. A plot of the percentage of the optimal action taken over time for both agents. The expectation is that the annealing ϵ agent will start by exploring as much as the constant ϵ agent but will exploit more as time goes on. This should lead to higher average rewards and a higher percentage of optimal actions taken as the agent learns more about the environment.