CS 5180 Exercise 1
Harin Kumar Nallaguntla

1. Given the action taken at every timestep and rewards received, the following are the Q values for each action and optimal action that can be taken at every timestep

At t = 0, q0 = [0 0 0 0], optimal actions = [1 2 3 4]
At t = 1, a1 = 1, r1 = -1, q1 = [-1 0 0 0], optimal action = [2 3 4]
At t = 2, a2 = 2, r2 = 1, q2 = [-1 1 0 0], optimal action = 2
At t = 3, a3 = 2, r3 = -2, q3 = [-1 -1/2 0 0], optimal actions = [3 4]
At t = 4, a4 = 2, r4 = 2, q4 = [-1 1/3 0 0], optimal action = 2
At t = 5, a5 = 3, r5 = 0, q5 = [-1 1/3 0 0], optimal action = 2

From the above analysis, it can be said that epsilon case could have possibly occurred on 1, 2 and 3 time steps. It definitely occurred on 4 and 5 time steps.

2. Since the step-size parameters are not stationary at every time step,

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

$$where,$$
$$Q_n = \alpha_{n-1}R_{n-1} + (1 - \alpha_{n-1})Q_{n-1}$$
$$Q_{n-1} = \alpha_{n-2}R_{n-2} + (1 - \alpha_{n-2})Q_{n-2}$$

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

3. a. The sample-average estimate in Equation 2.1, used for estimating the action value Qn, is shown to be unbiased. This estimate calculates Qn as the sum of rewards received from time step 1 to time step n-1 divided by the number of time steps (n-1). The assessment of bias centers on comparing the expected value of Qn, denoted as E[Qn], with the true expected reward $q*$ of a specific arm. When E[Qn] is equal to $q*$, the estimate is considered unbiased, and in this case, it is indeed unbiased due to the law of large numbers. As the number of samples (n) approaches infinity, E[Qn] converges to $q*$, ensuring that over time, the estimate becomes increasingly accurate and ultimately converges to the true value.

b. In the context of the exponential recency-weighted average estimate described in Equation 2.5 with $0 < \alpha < 1$ and an initial estimate Q1 = 0, the estimate Qn for n > 1 is biased. This bias arises because α, being strictly between 0 and 1, assigns less weight to the most recent reward Rn than to the initial estimate Q1 = 0. Consequently, as n increases, the estimate Qn remains significantly influenced by the initial estimate and fails to converge to the true expected reward $q*$, resulting in a persistent bias in the estimate.

c. To derive conditions for when the exponential recency-weighted average estimate Qn will be unbiased, we need to ensure that Qn converges to the true expected reward $q*$ over

time. In other words, we want to find conditions under which: E[Qn] = q*. Recall the update equation for Qn in the exponential recency-weighted average estimate: Qn = (1 - α) * Qn-1 + α * Rn.

Let's examine the expectations of both sides:
E[Qn] = E[(1 - α) * Qn-1 + α * Rn]
E[Qn] = (1 - α) * E[Qn-1] + α * E[Rn]

To make Qn unbiased, we want E[Qn] to be equal to q∗, the true expected reward, which means: E[Qn] = q∗

Now, let's set this condition: q∗ = (1 - α) * E[Qn-1] + α * E[Rn]
To make Qn unbiased, we need the following conditions to hold:

1. E[Rn] = q∗: The expected reward at each time step should be equal to the true expected reward q∗. In other words, the rewards should be an unbiased estimate of the true rewards.
2. E[Qn-1] = q∗: The expected value of the previous estimate Qn-1 should also be equal to the true expected reward q∗. This condition ensures that the influence of the initial estimate vanishes as n becomes larger. If both of these conditions are satisfied, then Qn will be unbiased and will converge to the true expected reward q∗ over time. These conditions ensure that both the rewards and the previous estimate provide unbiased information about the true expected reward.

d. To show that Qn is asymptotically unbiased, we need to demonstrate that as n approaches infinity (i.e., n → ∞), the expected value of Qn converges to the true expected reward q∗. In other words, we want to prove that: lim (n → ∞) E[Qn] = q∗

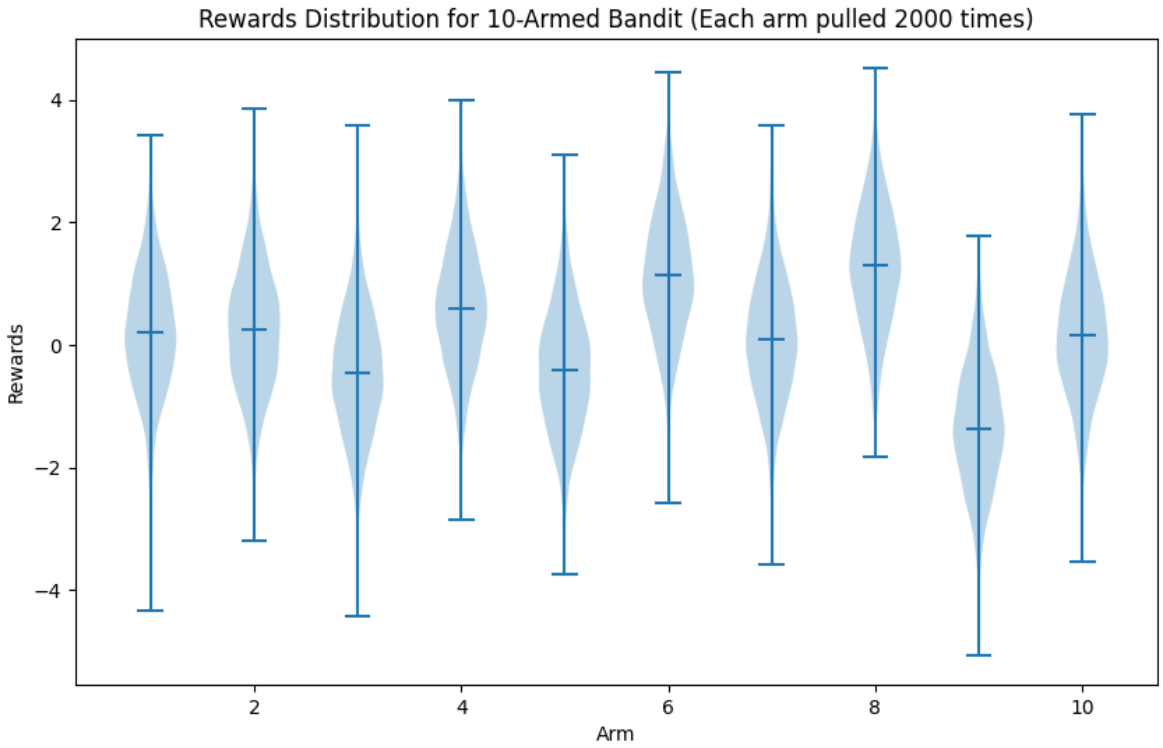
Qn = (1 - α) * Qn-1 + α * Rn
E[Qn] = E[(1 - α) * Qn-1 + α * Rn]
E[Qn] = (1 - α) * E[Qn-1] + α * E[Rn]
lim (n → ∞) E[Qn] = lim (n → ∞) [(1 - α) * E[Qn-1] + α * E[Rn]]
lim (n → ∞) E[Qn] = (1 - α) * lim (n → ∞) E[Qn-1] + α * q∗

Now, consider the limit of E[Qn-1] as n approaches infinity:
lim (n → ∞) E[Qn-1] = lim (n → ∞) E[Qn]

Then, lim (n → ∞) E[Qn] = (1 - α) * lim (n → ∞) E[Qn] + α * q∗
lim (n → ∞) E[Qn] = q∗
Hence Proved.

e. The choice of the weighting parameter α is a critical factor when using the exponential recency-weighted average estimate. While the exponential recency-weighted average itself is not inherently biased, the selection of α is paramount to its performance. If α is set too close to 1, the algorithm may adapt more slowly to changes in the environment, as it would give significant weight to past experiences, potentially missing out on valuable information from recent rewards. On the other hand, if α is chosen to be too small, the algorithm might overly emphasize recent rewards, resulting in estimates with higher variance. Striking the right balance in selecting an appropriate α is essential for the algorithm to effectively navigate various learning scenarios, ensuring both adaptability and accuracy in estimating the expected rewards. This consideration underscores the importance of fine-tuning α to suit the specific characteristics of the problem at hand.

4.



5. Optimal action selection: optimal action is selected with the probability (1 – ε). Non optimal action is selected with probability ε. Optimal action a* may be selected randomly with a probability ε/k, k is 10 here.
Total probability of selecting optimal action: $P(A_i = a^*) = 1 - \varepsilon + \varepsilon/k$
Considering given ε values:
$P(A_i = a^* | \varepsilon=0.01)/P(A_i=a^* | \varepsilon=0.1) = (1-0.01+0.01/10) / (1-0.1+0.1/10) = 0.991/0.91 = 1.09$
ε=0.01 performs better by 1.09 times than ε=0.1

Average reward: Since the rewards are derived from the same distribution with mean of 0, the average reward for random actions will be zero. We consider only average reward for greedy actions.

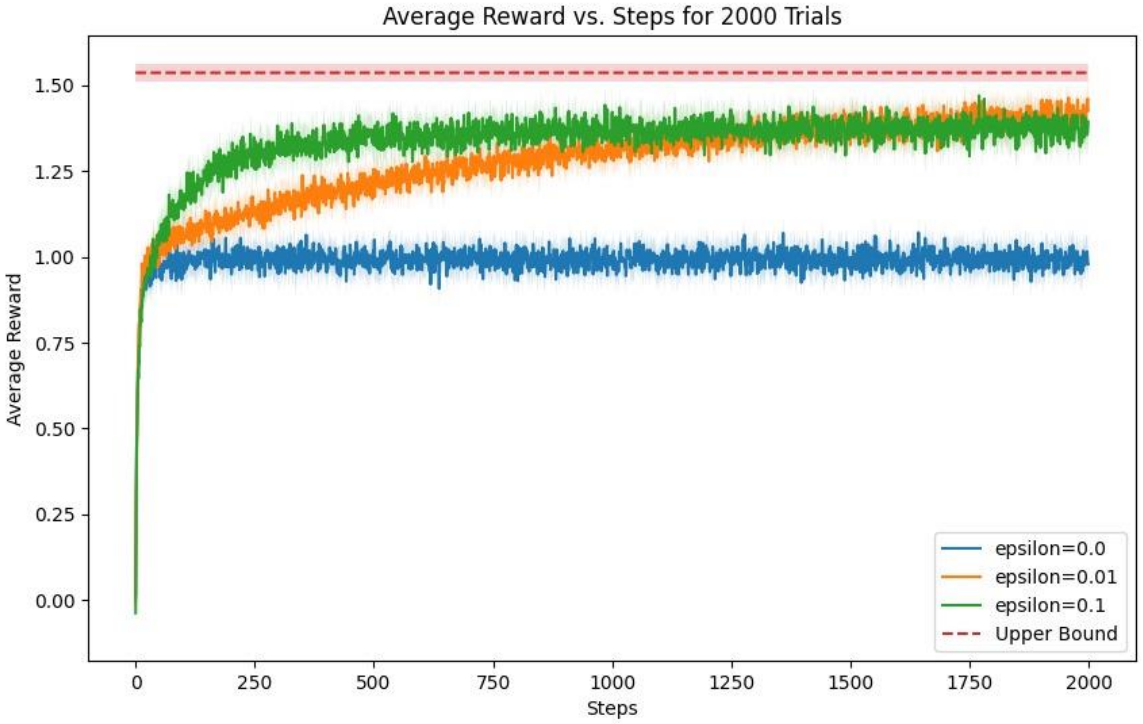
E[R] = q*(a*) * (1 – ε)

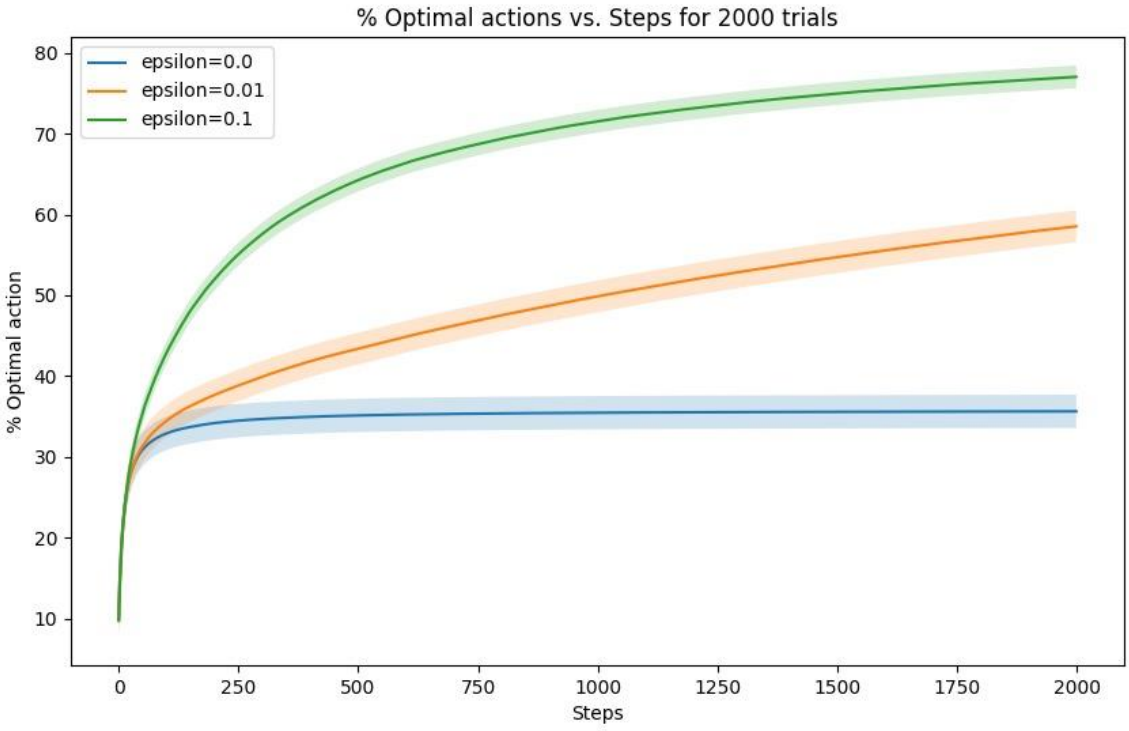
Considering given ε values:

E[R| ε=0.01]/E[R| ε=0.1] = (q*(a*)*(1 – 0.01)) / (q*(a*) * (1 – 0.1)) = 0.99/0.9 = 1.1

ε=0.01 performs better by 1.1 times than ε=0.1

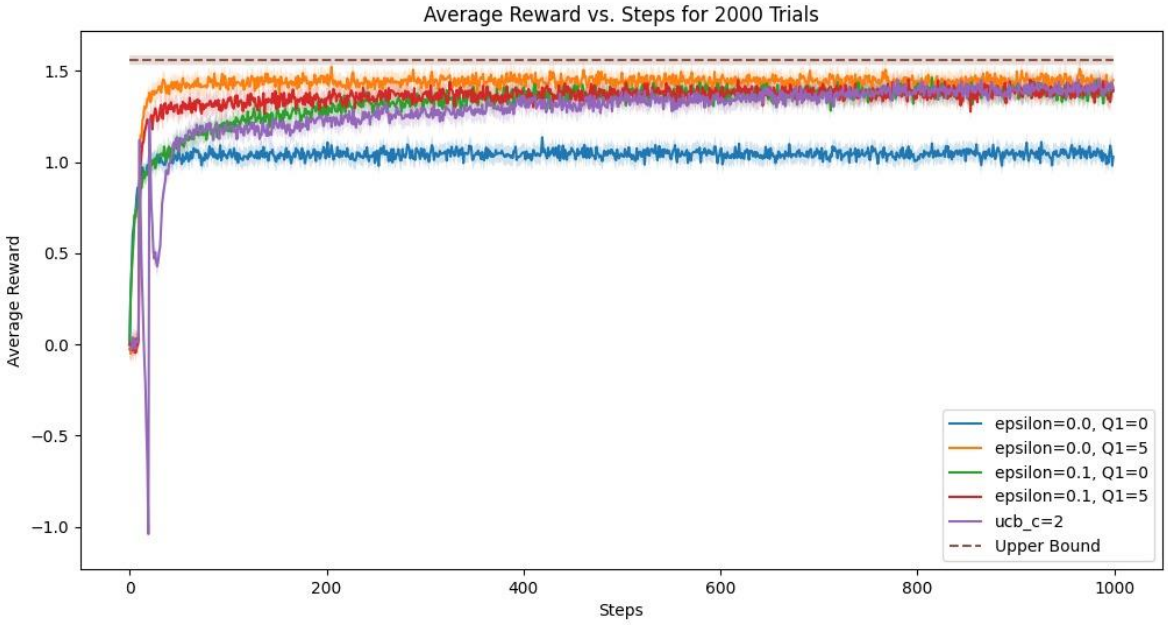
6.

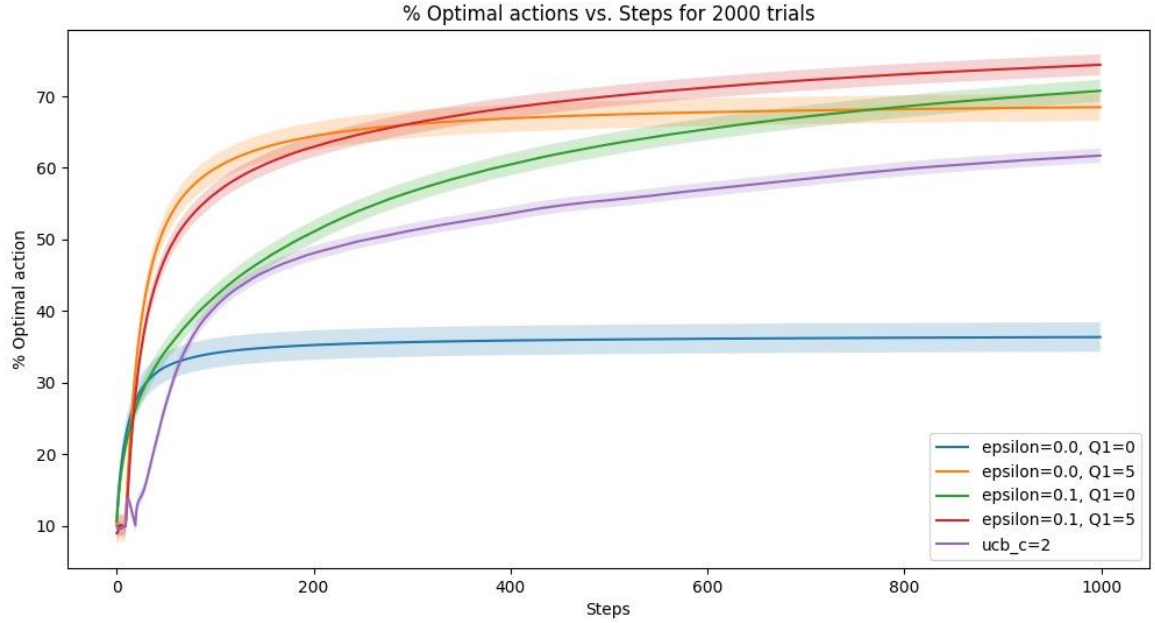




As you can see from Average reward vs steps graph, the epsilon=0.01 agent tries to reach the Upper Bound line. So, we can say it reaches the asymptotic levels as the number of steps tend to infinity

7.





The initial spikes observed in both optimistic initialization and UCB methods can be attributed to the deliberate setting of high initial estimates for action values (Q-values). Optimistic initialization with inflated Q-values encourages early exploration as the agent initially favors actions with higher estimated rewards (as seen from average reward vs steps graph from steps 0 to 100). This results in a surge in rewards at the beginning as various actions are explored. On the other hand, UCB encourages exploration by balancing the selection of actions with high uncertainty, leading to an initial spike in rewards as uncertain actions with potential high rewards are chosen (as seen from steps 0 to 150 in average

reward vs steps graph). These initial spikes reflect the agents' early exploration strategies, which diminish as they gather more data and refine their estimates.

The subsequent sharp decreases observed in both cases can be explained by the convergence of the exploration-exploitation trade-off. As the agents accumulate more data and update their Q-value estimates based on actual rewards, these estimates begin to converge toward the true expected action values. Actions that were initially favored due to optimistic estimates or high uncertainty start to reveal their actual, often lower, rewards, leading to a decline in their selection frequency. Conversely, actions initially underrated become preferred as their true value becomes evident. This convergence towards more accurate estimates results in a decline in the initially high rewards and a shift towards a sharper focus on exploiting optimal actions. Both methods undergo a transition from exploration to exploitation, leading to the observed sharp decrease in rewards following the initial spikes.