Reinforcement learning
Assignment – 1

Gopisainath Mamindlapalli
Study partner: Harin Kumar

1) Given Q1 = [0 0 0 0], actions : A1 = 1, A2 = 2, A3 = 2, A4 = 2, A5 = 3 and the rewards are: R1 = -1, R2 = 1, R3 = -2, R4 = 2, R5 = 0.

   Using the rewards the following Q values are:

   Q1(A1) = -1/1 = -1 : Q1 = [-1 0 0 0]
   Q2(A2) = 1/1 = 1 : Q2 = [-1 1 0 0]
   Q3(A3) = (1-2)/2 = -1/2 : Q2 = [-1 1/2 0 0]
   Q4(A4) = (1-2+2/3 = 1/3 : Q2 = [-1 1/3 0 0]
   Q5(A5) = 0/1 = 0 : Q2 = [-1 1/3 0 0]

From the above Q values we can say that actions 4,5 are occurred randomly also when there is no action is taken then Q values will be same for all actions (zero in this case) then this event may occur.

2)

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

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

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

3) a) The sample-average estimate in Equation 2.1, which is used to calculate the action value Qn, is demonstrated to be unbiased. In this estimate, Qn is determined by dividing the total rewards from time step 1 to time step n-1 by the number of time steps (n-1). Comparing the predicted value of Qn, designated as E[Qn], with the actual expected reward q of a particular arm is the main focus of bias assessment. The estimate is thought to be unbiased when E[Qn] equals q, and in this case the law of large numbers ensures that it is. E[Qn] converges to q as the number of samples (n) approaches infinity, ensuring that the estimate improves with time and eventually converges to the true value.

   b) The estimate Qn for n > 1 is biassed in the exponential recency-weighted average estimate specified in Equation 2.5 with 0 1 and a starting estimate Q1 = 0. Because is strictly between 0 and 1, it gives the most recent reward Rn less weight than the original estimate Q1 = 0, which leads to this bias. As a result, there is a persistent bias in the estimate as n grows since the estimate Qn fails to converge to the genuine expected reward q and is still strongly influenced by the initial estimate.

c) We must make sure that Qn converges to the genuine expected reward q over time in order to determine the circumstances under which the exponential recency-weighted average estimate Qn will be unbiased. In other words, we're looking for circumstances in which: E[Qn] = q*. Remember that Qn = (1 -)* Qn-1 + * Rn is the update equation for Qn in the exponential recency-weighted average estimate.

To make Qn unbiased,
1. E[Rn] = q: The anticipated benefit at each time step must be equal to the actual anticipated benefit, q. To put it another way, the rewards ought to be a fair estimation of the real rewards.
2. E[Qn-1] = q: The anticipated worth of the earlier estimate Qn-1 ought to be equivalent to the actual anticipated reward q. This requirement makes sure that as n increases, the initial estimate loses its impact. Both of these requirements must be met for Qn to be impartial and to eventually converge to the genuine expected reward q. These requirements guarantee that the rewards and the earlier estimate both offer accurate information about the actual expected reward.

d) The expected value of Qn must converges to the actual expected reward q in order to show that Qn is asymptotically unbiased (i.e., n).

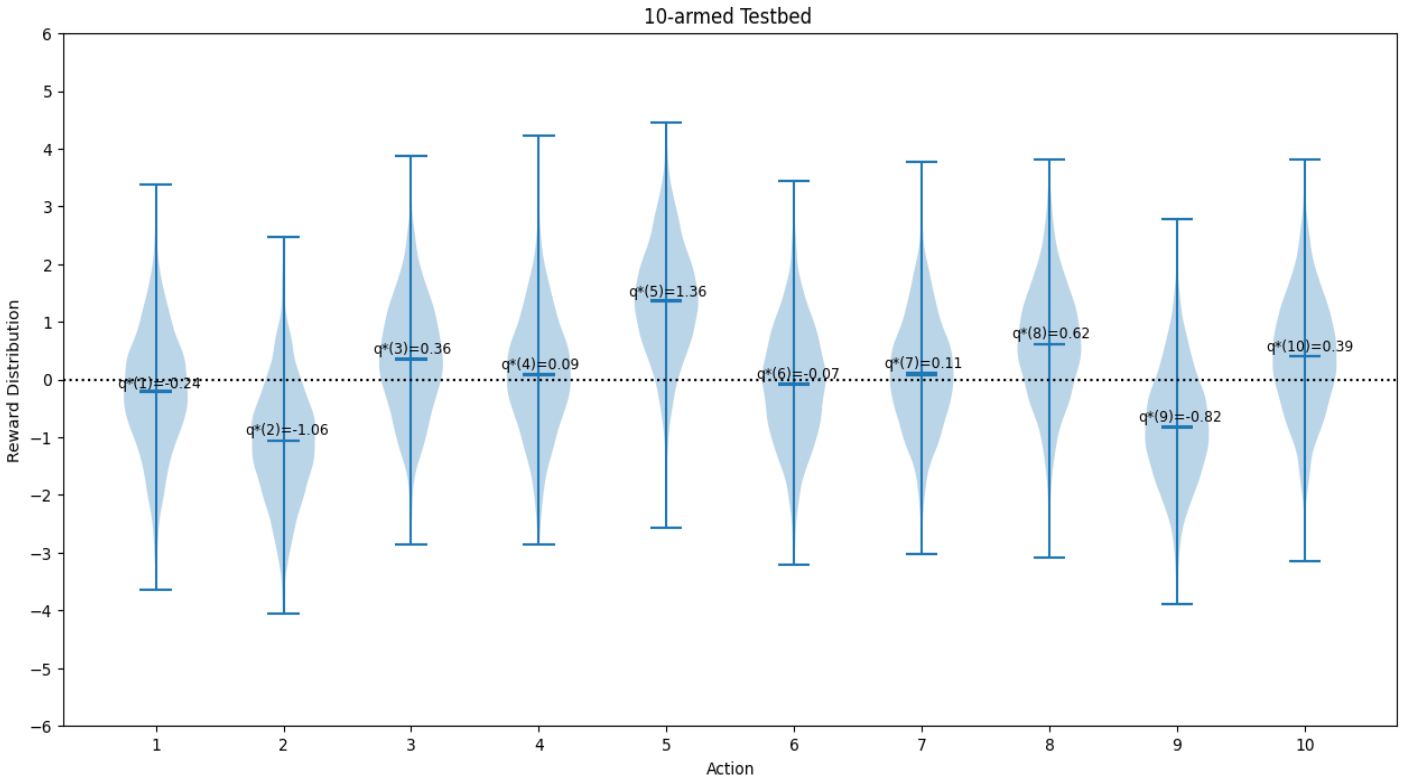
$Q_n = (1 - \alpha) * Q_{n-1} + \alpha * R_n$
$E[Q_n] = E[(1 - \alpha) * Q_{n-1} + \alpha * R_n]$
$E[Q_n] = (1 - \alpha) * E[Q_{n-1}] + \alpha * E[R_n]$
$\lim (n \to \infty) E[Q_n] = \lim (n \to \infty) [(1 - \alpha) * E[Q_{n-1}] + \alpha * E[R_n]]$
$\lim (n \to \infty) E[Q_n] = (1 - \alpha) * \lim (n \to \infty) E[Q_{n-1}] + \alpha * q*$

Now, consider the limit of E[Qn-1] as n approaches infinity:
$\lim (n \to \infty) E[Q_{n-1}] = \lim (n \to \infty) E[Q_n]$

Then, $\lim (n \to \infty) E[Q_n] = (1 - \alpha) * \lim (n \to \infty) E[Q_n] + \alpha * q*$
$\lim (n \to \infty) E[Q_n] = q*$

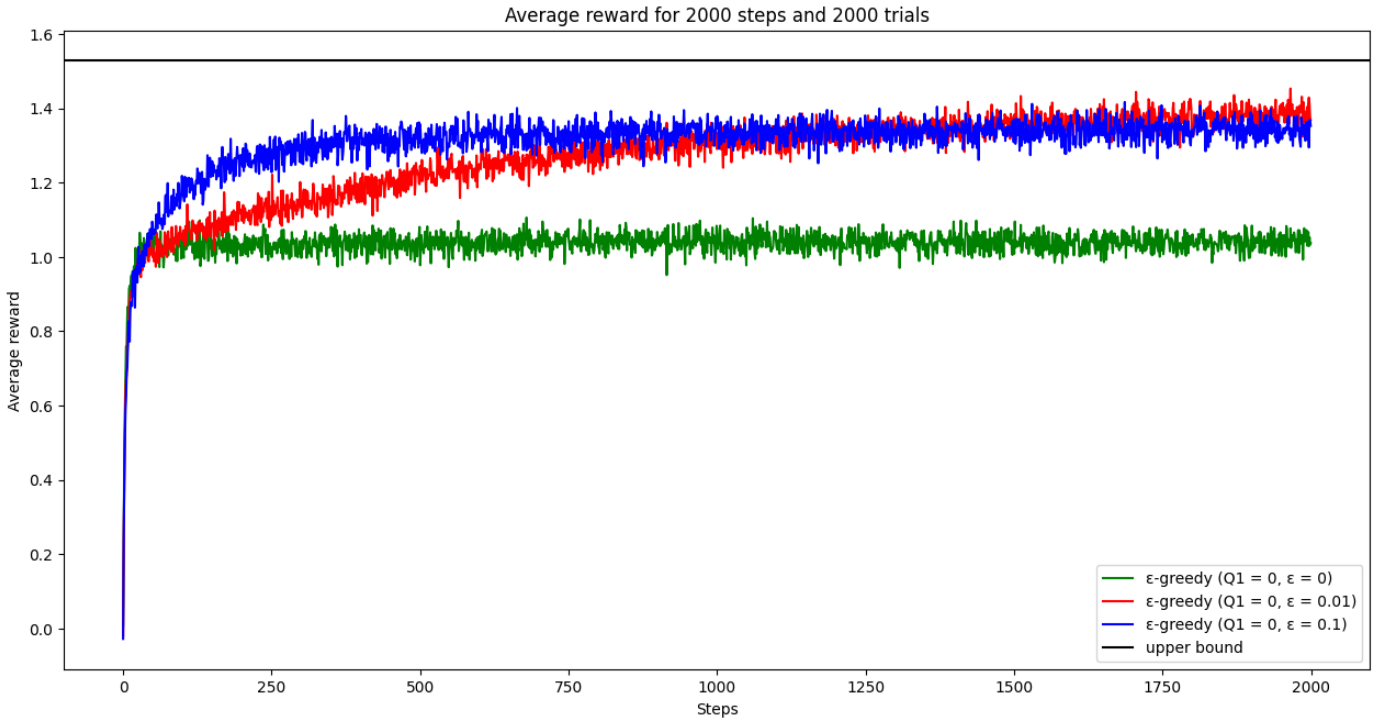
e) When employing the exponential recency-weighted average estimate, the selection of the weighting parameter is crucial. Although the exponential recency-weighted average is not by nature biassed, the choice of is critical to how well it performs. If is set too closely to 1, the algorithm may respond to environmental changes more slowly since it would place a heavy emphasis on prior experiences, thus missing out on important information from more recent rewards.
Alternatively, if is set to be too small, the algorithm might disproportionately emphasise recent rewards, producing estimates with more variance. For the algorithm to successfully navigate different learning contexts, providing both adaptability and accuracy in forecasting the expected rewards, it is crucial to strike the proper balance while choosing a suitable. This point emphasises the significance of tailoring solutions to the particulars of the current issue.

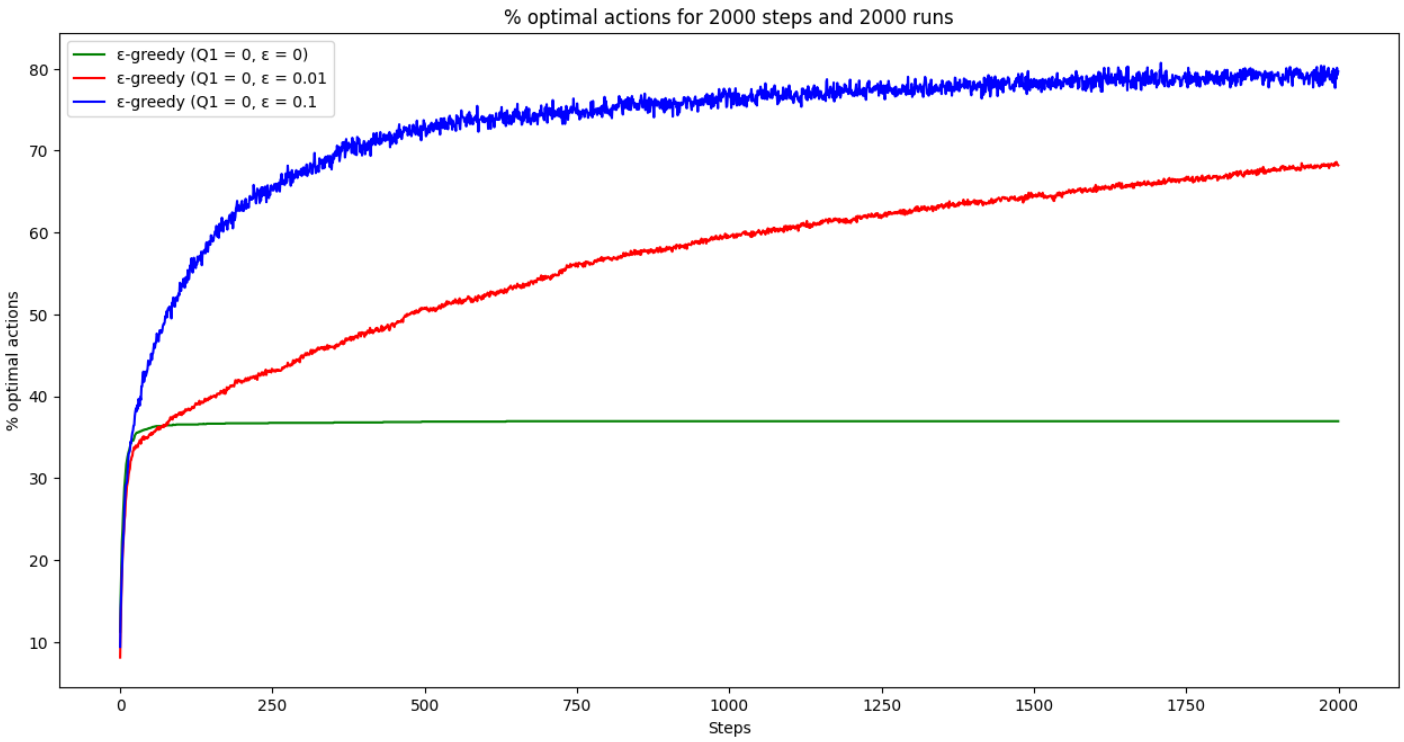
4)



5) The epsilon greedy method with epsilon value of 0.01 will perform in the long run. The probability of selecting the optimal action is 1- ε + (ε/10).
Therefore, for ε = 0.01 method this value is 0.991
for ε = 0.1 method this value is 0.91
so, it will perform by (0.991/0.91) = 1.089 times better

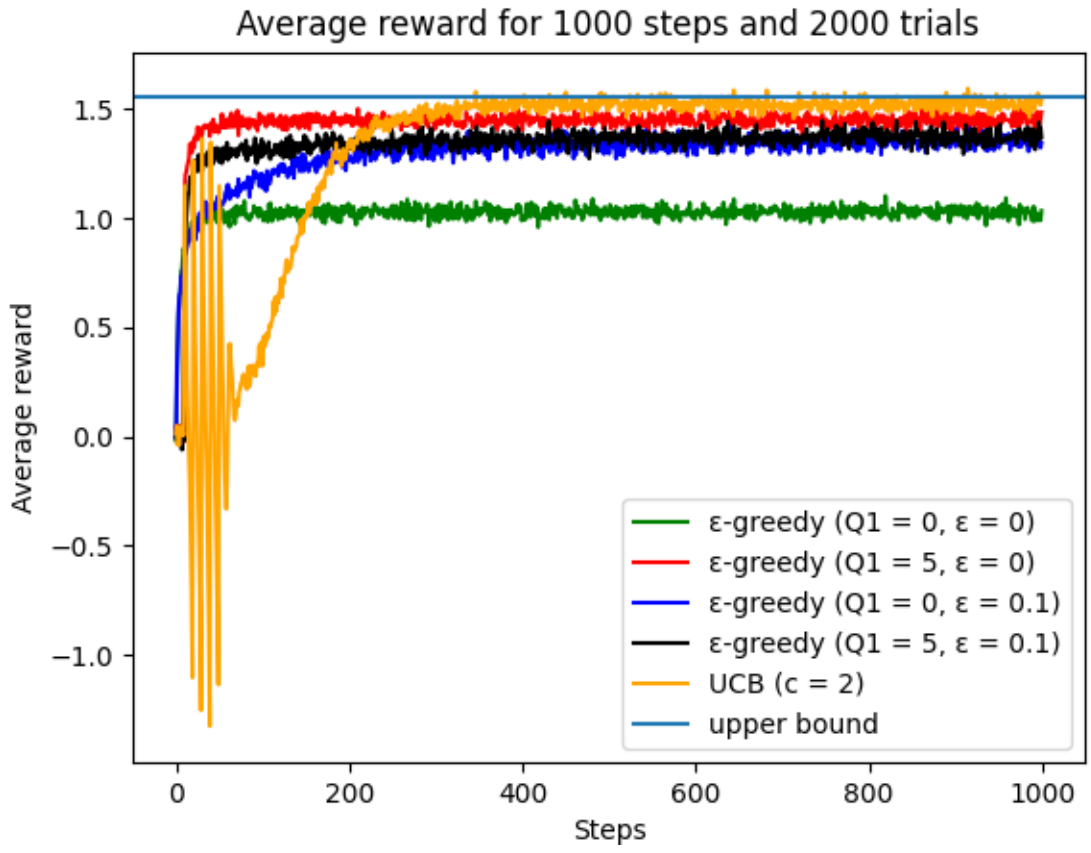
6)

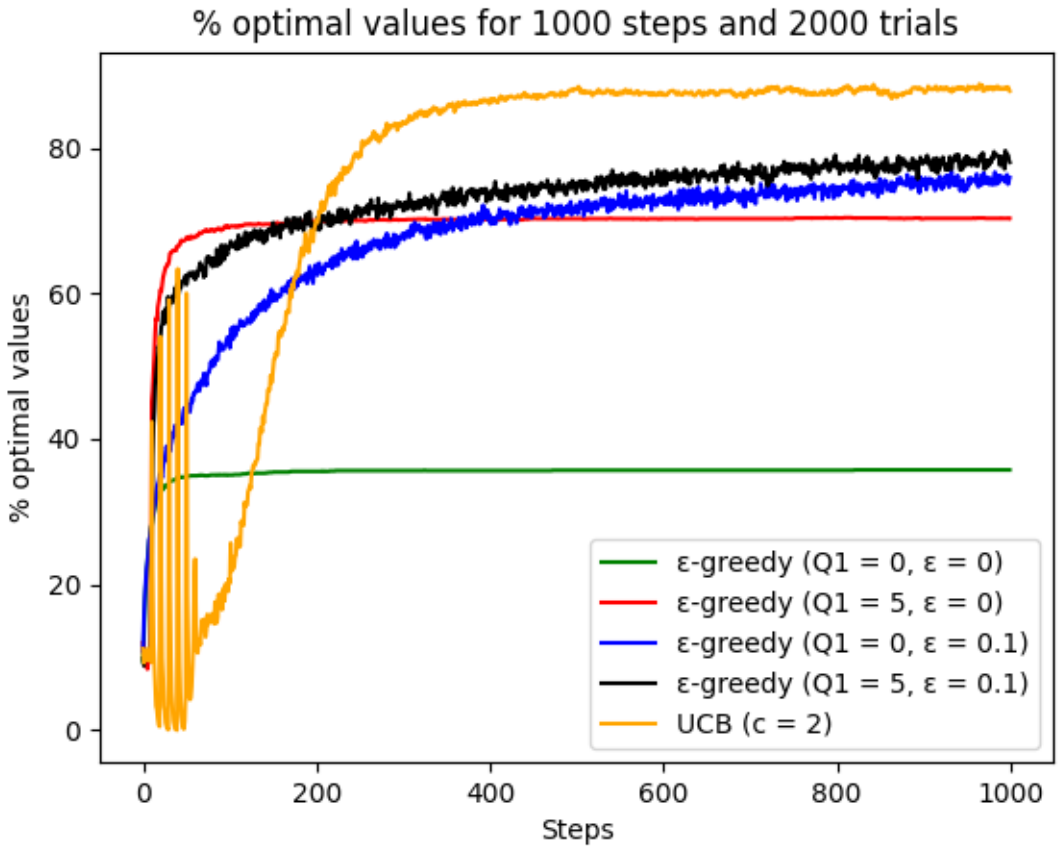




The average reward over steps graph shows that the epsilon=0.01 agent attempts to cross the Upper Bound line. So, as the number of steps tends to infinity, we can claim that it reaches the asymptotic levels.

7)





The initial spikes observed in both optimistic initialization and UCB techniques can be attributable to the deliberate use of high starting estimations for action values (Q-values). As can be observed from the average reward vs steps graph from steps 0 to 100, optimistic initialization with inflated Q-values promotes early exploration because the agent initially favours actions with greater anticipated rewards. As a result, rewards increase quickly at first as different activities are investigated. As can be seen in the average reward vs. steps graph from steps 0 to 150, UCB, on the other hand, promotes exploration by balancing the selection of actions with high uncertainty. This causes an early spike in rewards as uncertain actions with potential high rewards are taken.

The initial spikes observed in both optimistic initialization and UCB techniques can be attributable to the deliberate use of high starting estimations for action values (Q-values). As can be observed from the average reward vs steps graph from steps 0 to 100, optimistic initialization with inflated Q-values promotes early exploration because the agent initially favours actions with greater anticipated rewards. As a result, rewards increase quickly at first as different activities are investigated. As can be seen in the average reward vs. steps graph from steps 0 to 150, UCB, on the other hand, promotes exploration by balancing the selection of actions with high uncertainty. This causes an early spike in rewards as uncertain actions with potential high rewards are taken.