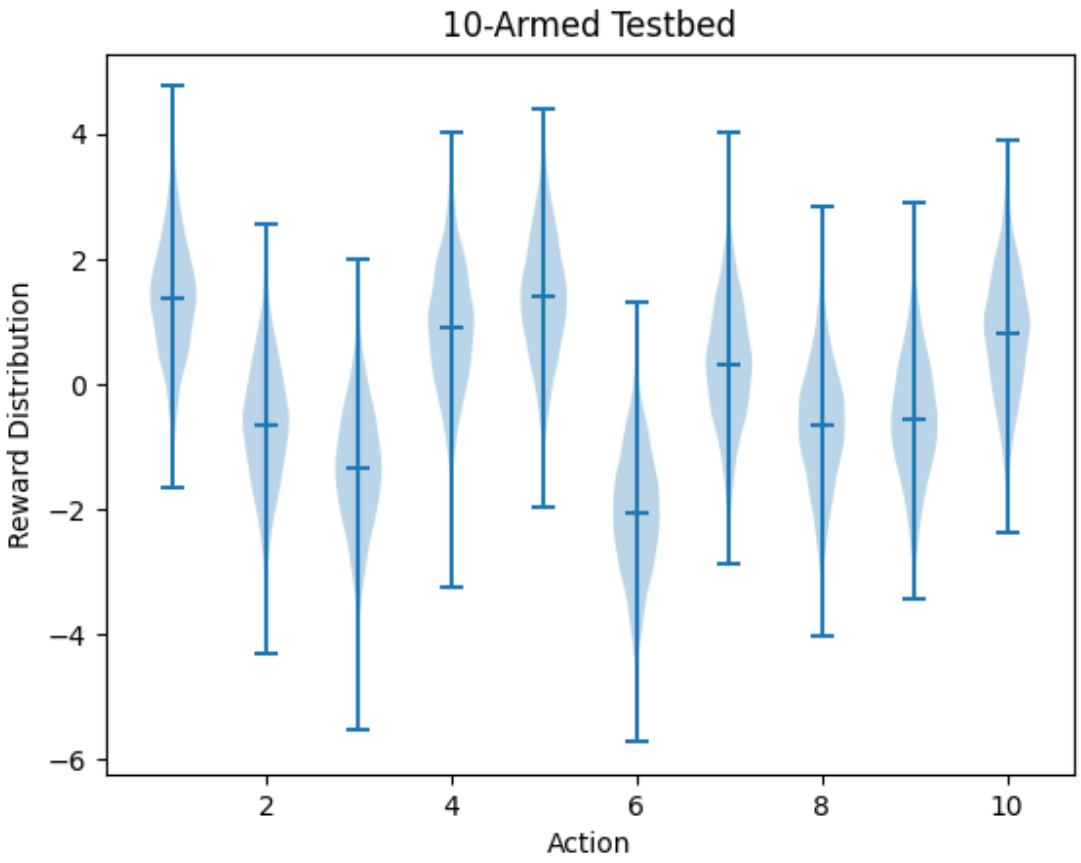
Reinforcement Learning - EXP1
Bandits

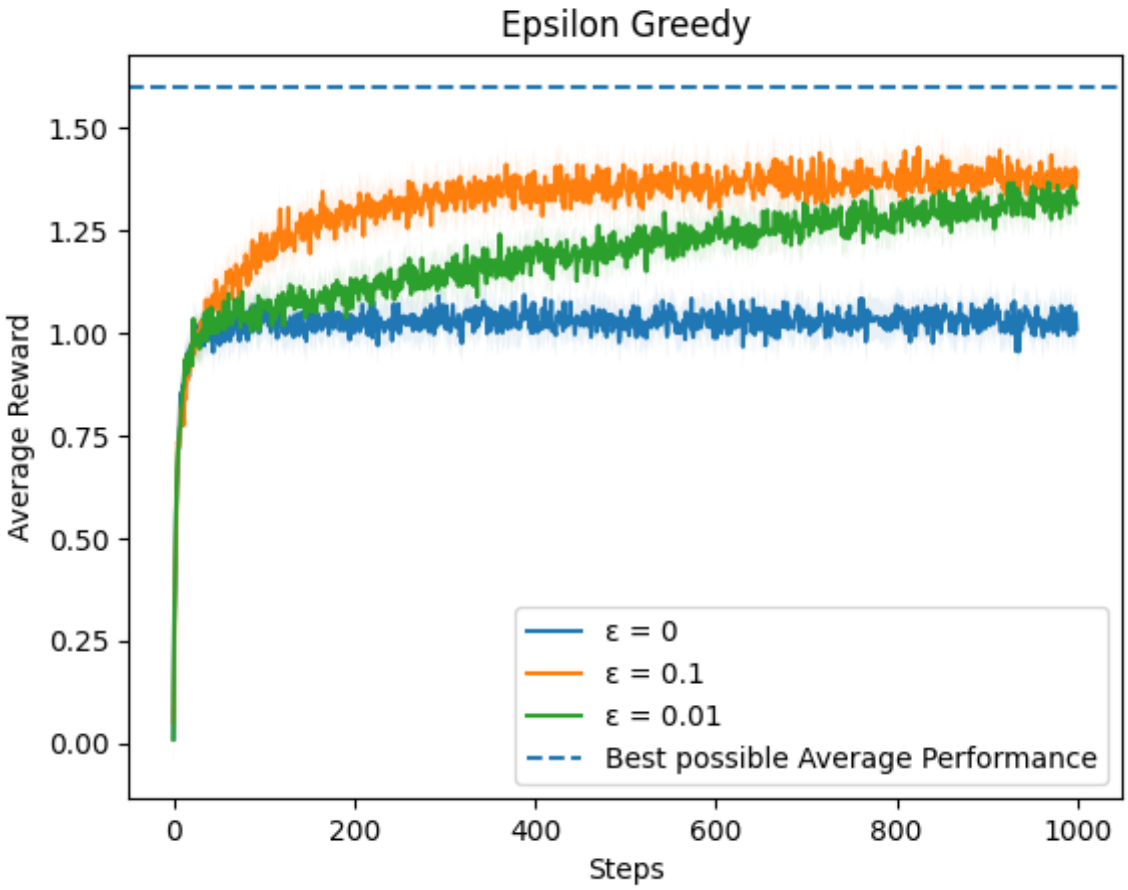
Q4)

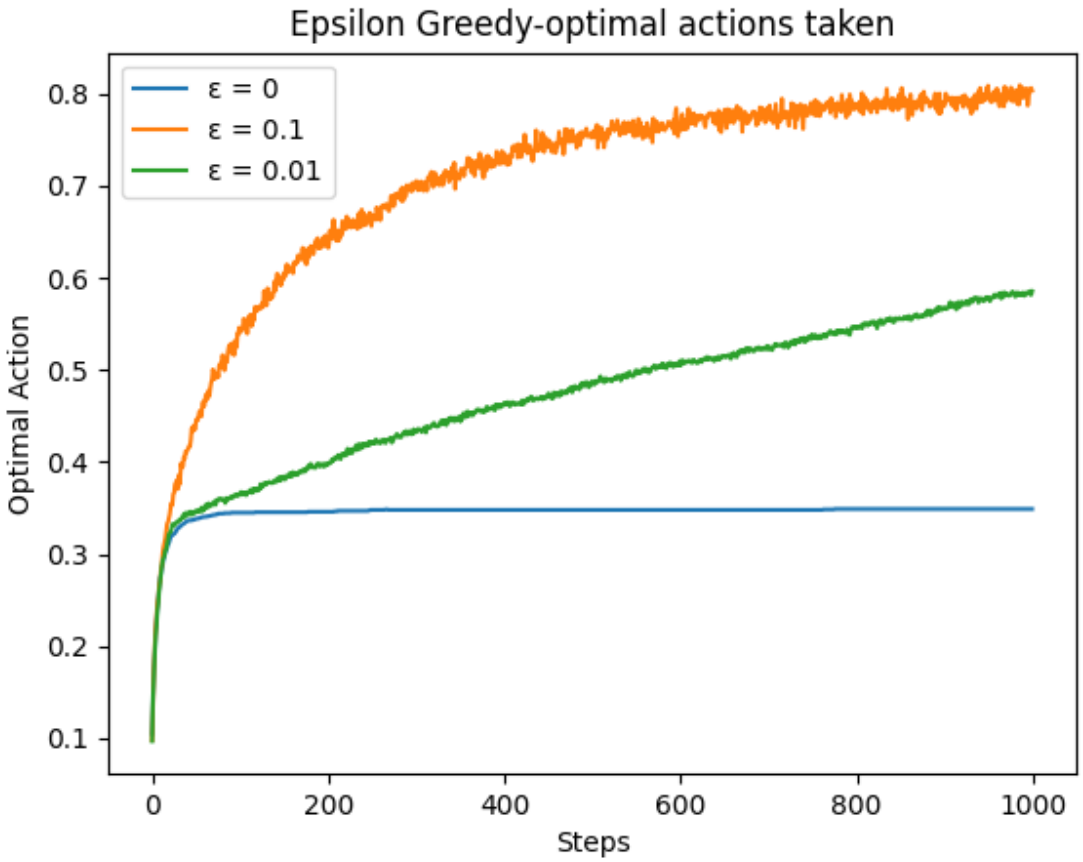
Plots



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.

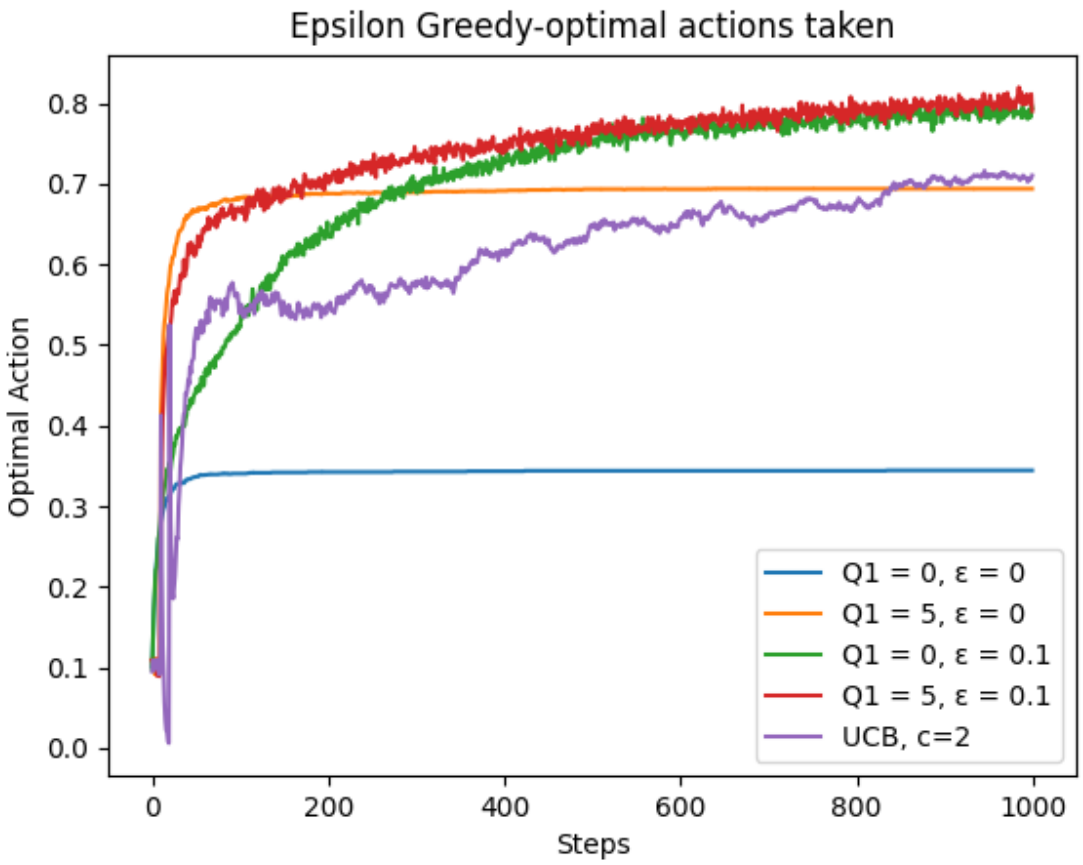
Ans ->

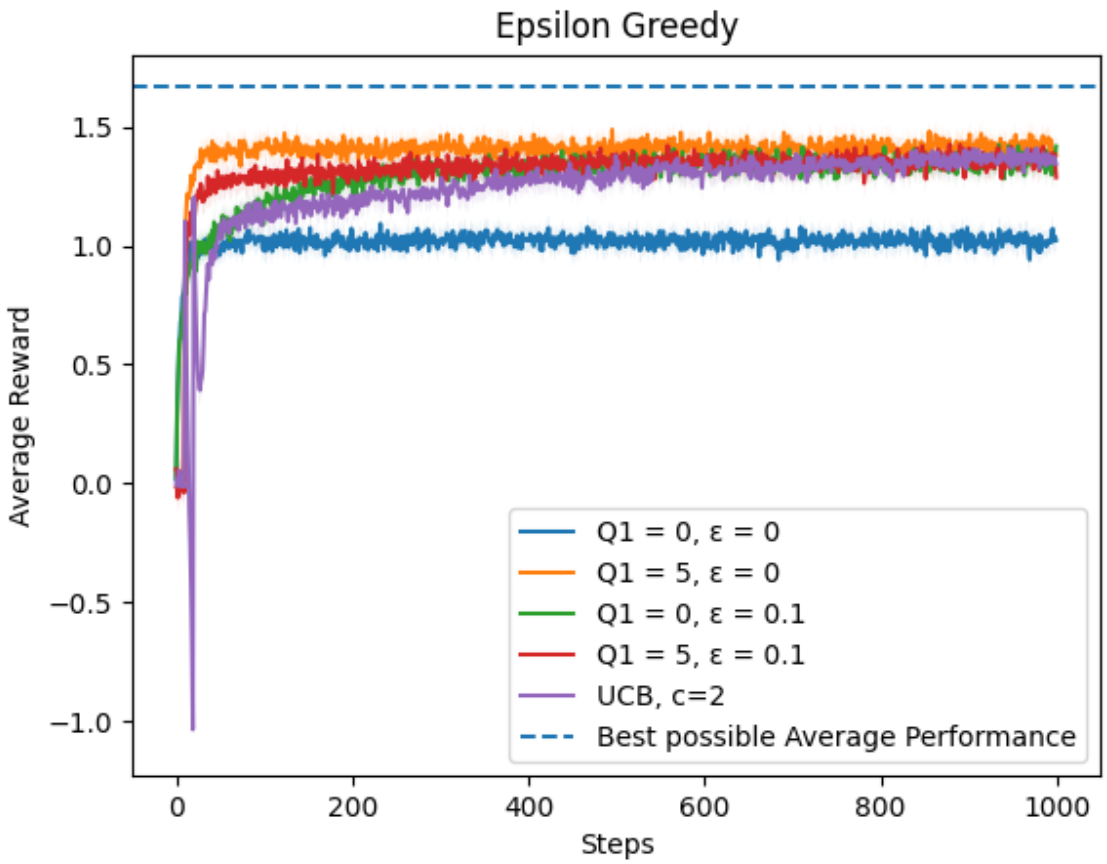
We can observe that - spike is at instance 11, and it drops after that. This happened in both optimistic initialization and UCB. Here, K = 10, we first try all 10 actions, and record the respective rewards. Thus, at t = 11, we will first find the high reward, hence we observe that spike.

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