Q 1)

$Q_1(1) = \dots = Q_1(4) = 0$ step: S

$S=1$ $A_1 = 1, R_1 = -1 \Rightarrow Q_2(1) = -\frac{1}{2}$

$Q_2(2) = \dots = Q_2(4) = 0$

$S=2$ $A_2 = 2, R_2 = 1 \Rightarrow Q_3(1) = -\frac{1}{2}$ $Q_3(3) = Q_3(4) = 0$

$Q_3(2) = \frac{1}{2}$

$S=3$ $A_3 = 2, R_3 = -2$ $\Rightarrow$ $Q_4(1) = -\frac{1}{2}$

select greedy action

$Q_4(2) = -\frac{1}{2}$

$Q_4(3) = Q_4(4) = 0$

$S=4$ $A_4 = 2, R_4 = 2$ $\Rightarrow$ $Q_5(1) = -\frac{1}{2}$

non-greedy action selection

greedy actions are: 3, 4

$Q_5(2) = \frac{1}{2}$

$Q_5(3) = Q_5(4) = 0$

$S=5$ $A_5 = 3, R_5 = 0$ $\Rightarrow$ $Q_6(1) = -\frac{1}{2}$

non-greedy action selection

greed action is: 2

$Q_6(2) = \frac{1}{2}$

$Q_6(3) = Q(4) = 0$

So the ε case occured definitely on Steps: 4, 5

And It could be occured in every Step with $1-\varepsilon$ probability

Q 2)

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

$$= \alpha_n R_n + (1-\alpha_n)\underbrace{Q_n}_{\alpha_{n-1} R_{n-1} + (1-\alpha_{n-1})\underbrace{Q_{n-1}}_{\alpha_{n-2} R_{n-2} + (1-\alpha_{n-2}) Q_{n-2}}}$$

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

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

Q 3)

a. It is unbiased because:

$$Q_n \doteq \frac{R_1 + \dots + R_{n-1}}{n-1} = \frac{\sum_{i=1}^{n-1} R_i}{n-1}$$

$$E(Q_n) = \frac{1}{n-1} \sum_{i=1}^{n-1} E(R_i) = \frac{n-1}{n-1} E(R_i)$$

$$\mathbb{E}(Q_n) = \mathbb{E}(R_i) = q_* = E(R \mid A=a)$$

It would be the same

It could show for all actions

b. Yes, it will be biased towards 0.

c. Q_n will be unbiased if Q1 is initialized to q*

d. As n goes to infinity, the coefficient of the initial estimate Q1: (1-alpha)^n goes to 0, so it becomes unbiased.

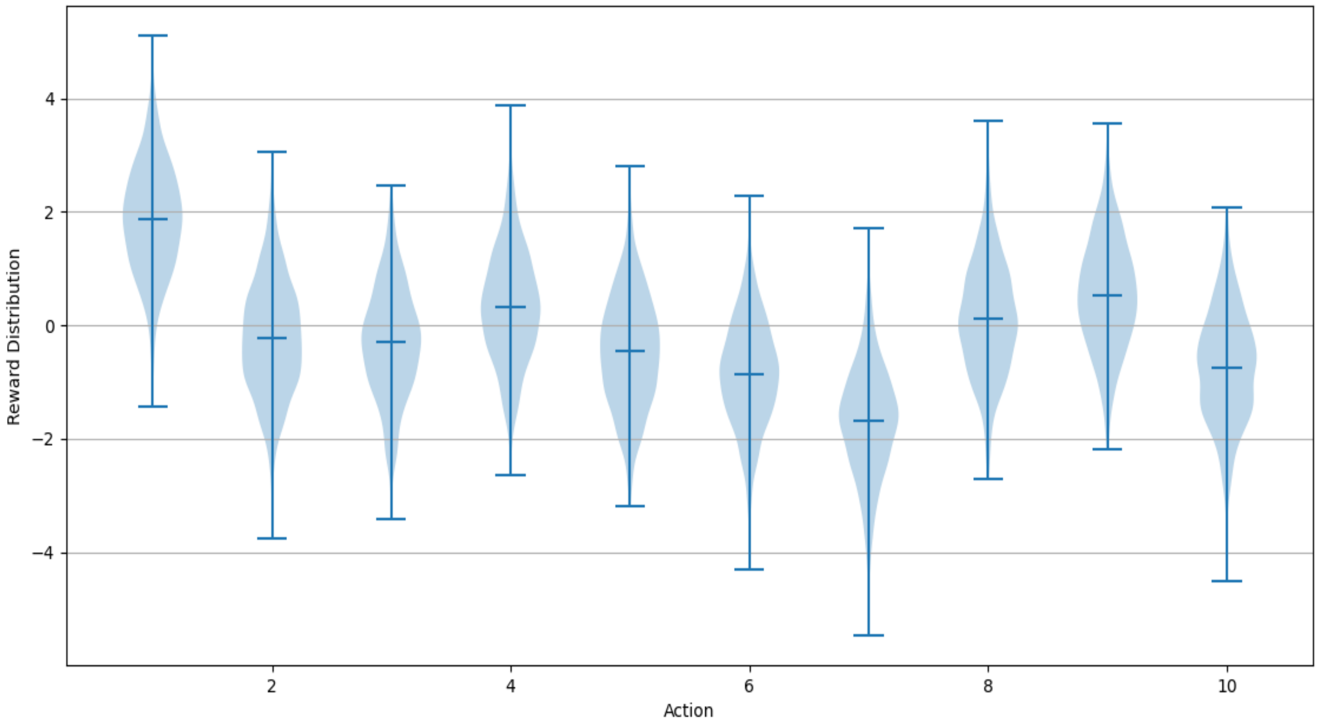
e. Because in practice we deal with episodic tasks, where n is not expected to go toward infinity.

Q 4)

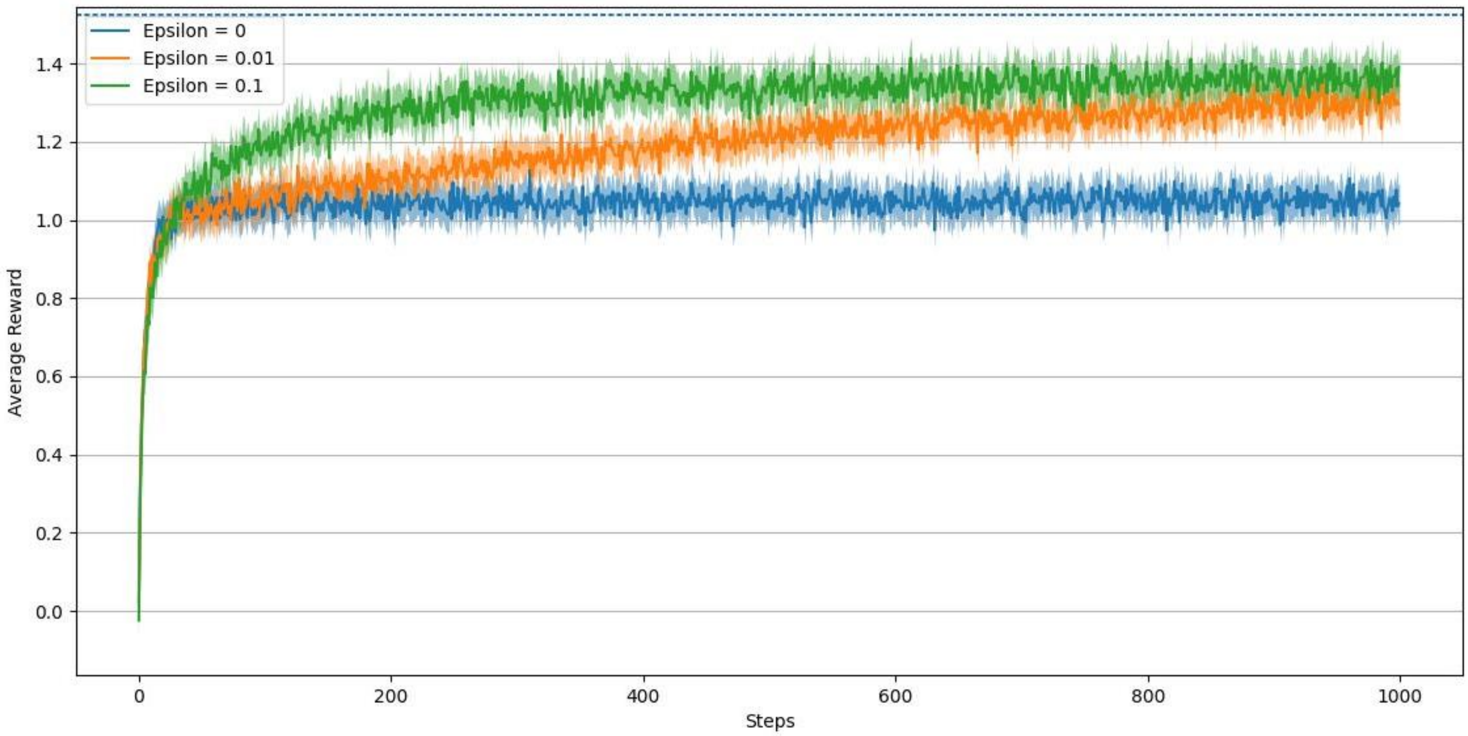
$$Pr(A_t=1) = \frac{e^{H_t(1)}}{e^{H_t(1)} + e^{H_t(2)}} = \frac{1}{1+e^{\frac{H_t(2)}{H_t(1)}}} = \text{sigmoid}\left(-\frac{H_t(2)}{H_t(1)}\right)$$

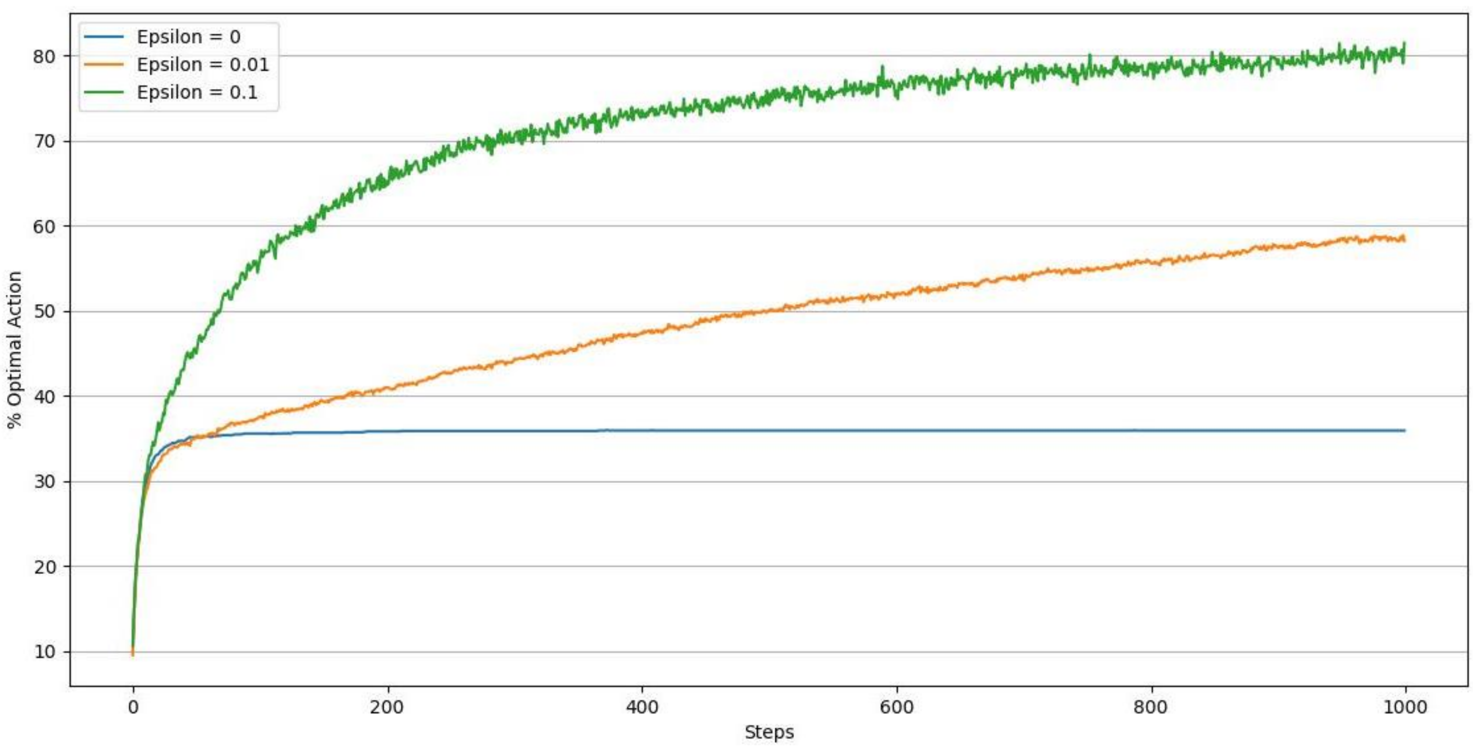
$$Pr(A_t=2) = \frac{e^{H_t(2)}}{e^{H(1)} + e^{H_t(2)}} = \frac{1}{1+e^{\frac{H_t(1)}{H_t(2)}}} = \text{sigmoid}\left(-\frac{H_t(1)}{H_t(2)}\right)$$

Q 5)



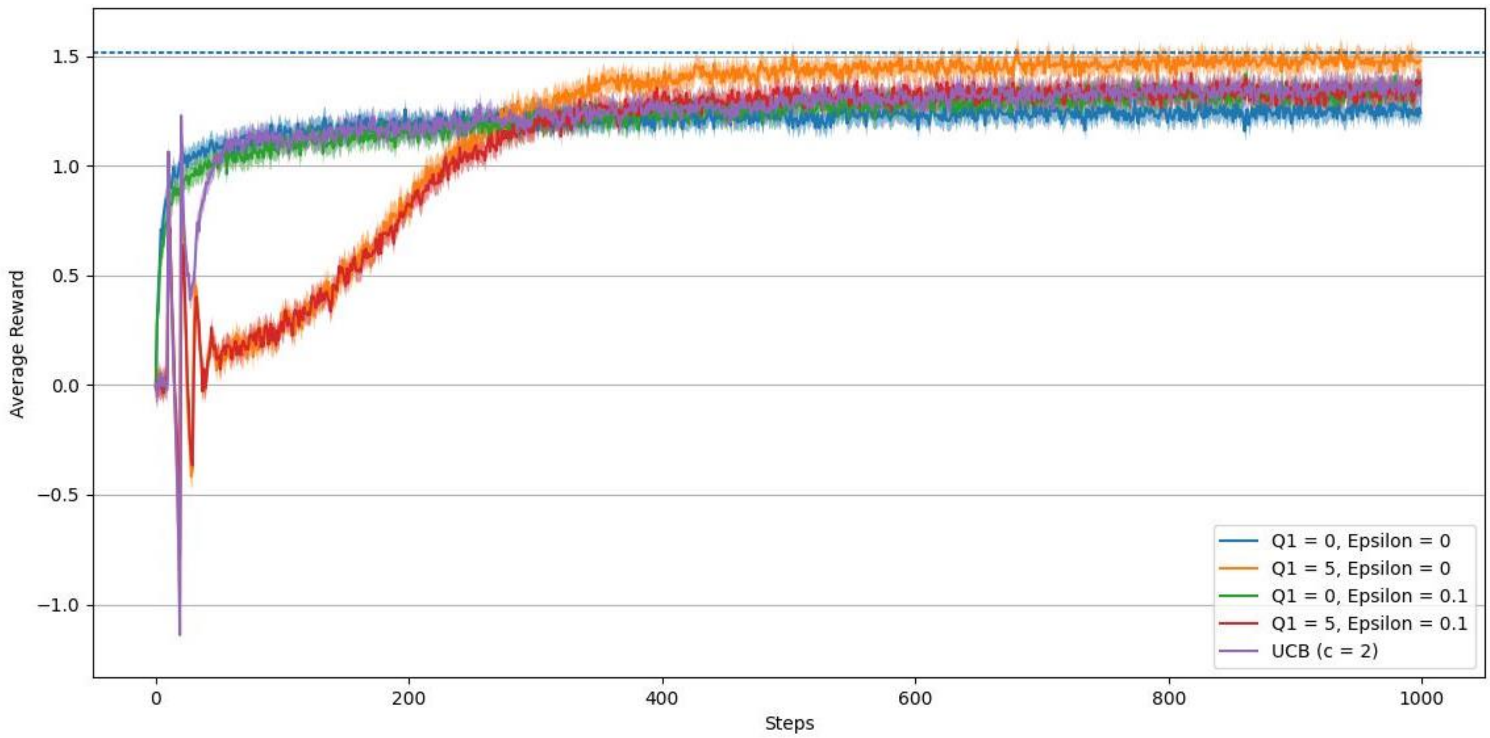
Q 6)



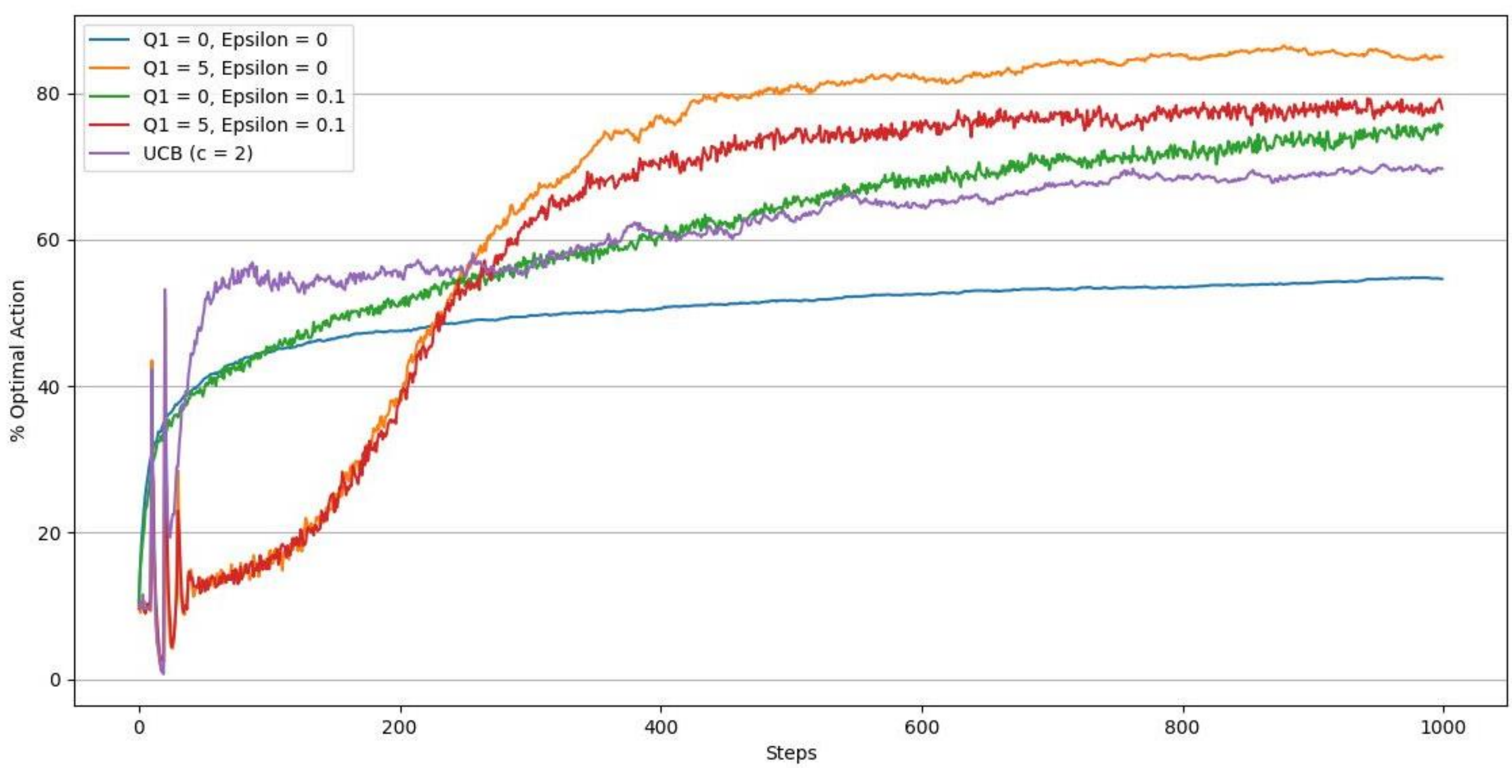


The averages do not reach the asympomatic levels (91% and 99.1%).

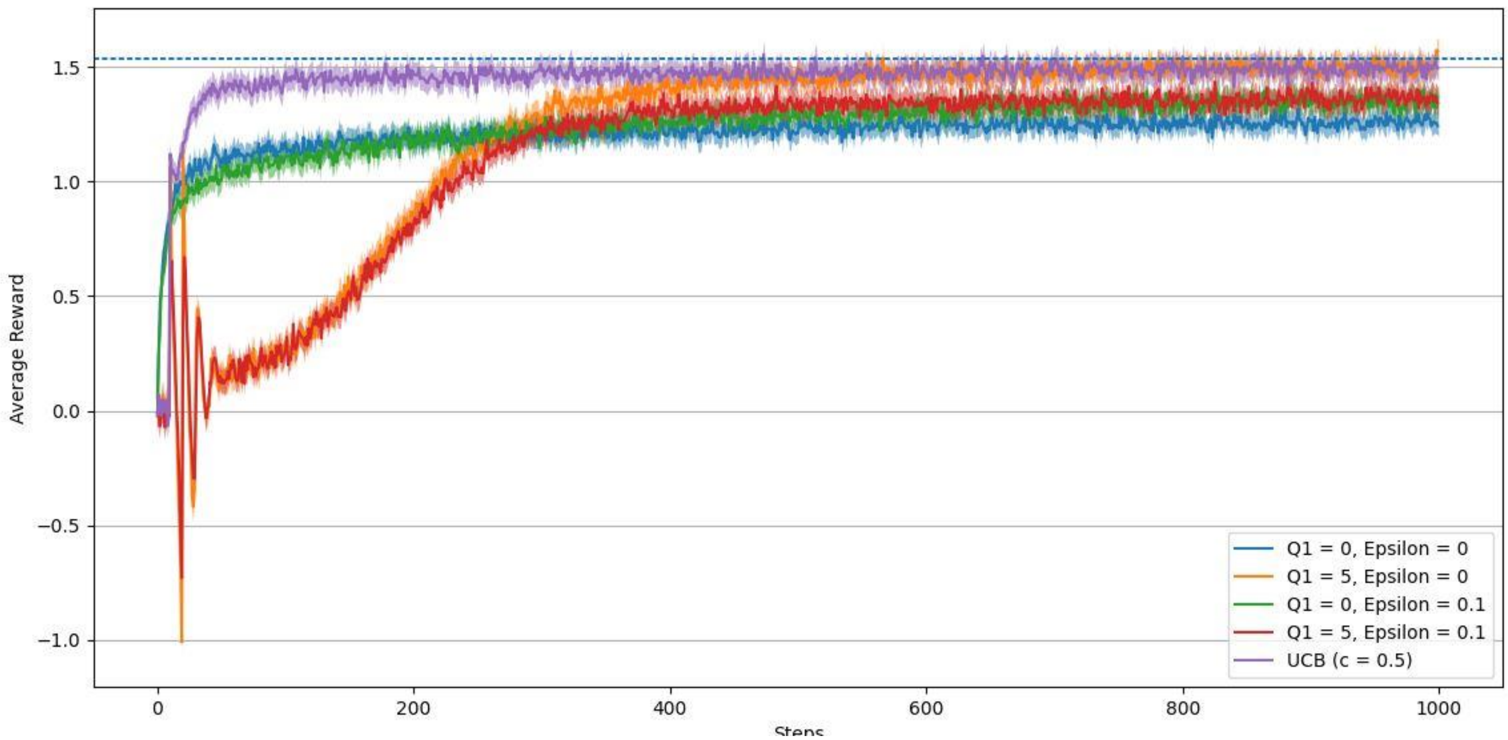
Q 7)

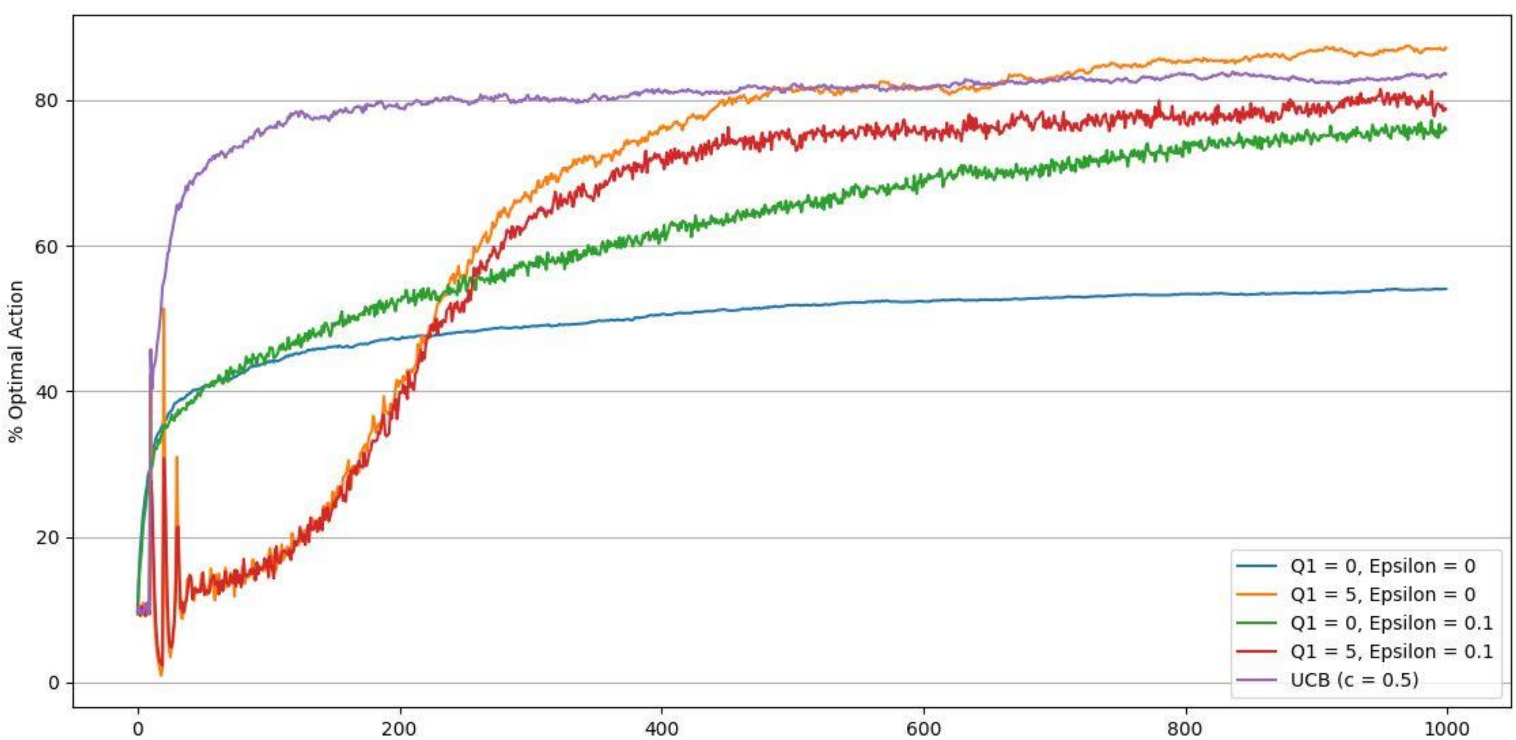


Average Reward
Steps
Q1 = 0, Epsilon = 0
Q1 = 5, Epsilon = 0
Q1 = 0, Epsilon = 0.1
Q1 = 5, Epsilon = 0.1
UCB (c = 2)



% Optimal Action
Steps
Q1 = 0, Epsilon = 0
Q1 = 5, Epsilon = 0
Q1 = 0, Epsilon = 0.1
Q1 = 5, Epsilon = 0.1
UCB (c = 2)





Thes spikes are because actions are selected randomly in early steps. At first steps, large c suppresses the action to select the action greedily. This causes the rewards to be low in the early steps. Once a high reward is reached by some almost random actions, the average reward for that step increases suddenly. But, in the next steps, C again suppresses the agent to keep selecting greedy action, and therefore the reward is lowered. But because the Q for high-reward action has already been updated, there is still a chance of selecting that action. Hence the reward is not as low as before the spike. I have runed another simulation with C=0.5 that shows very weak spikes, two last figures.