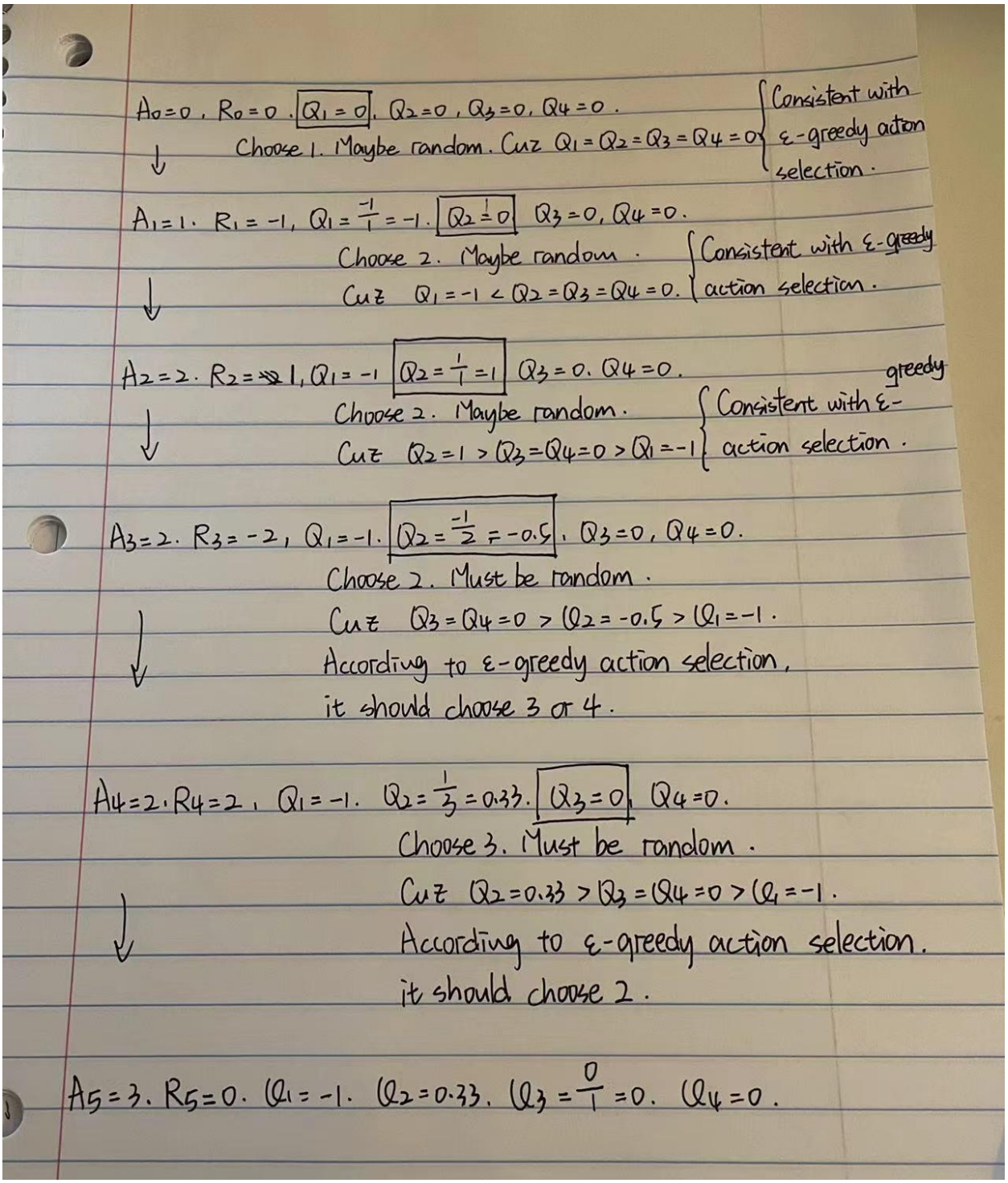
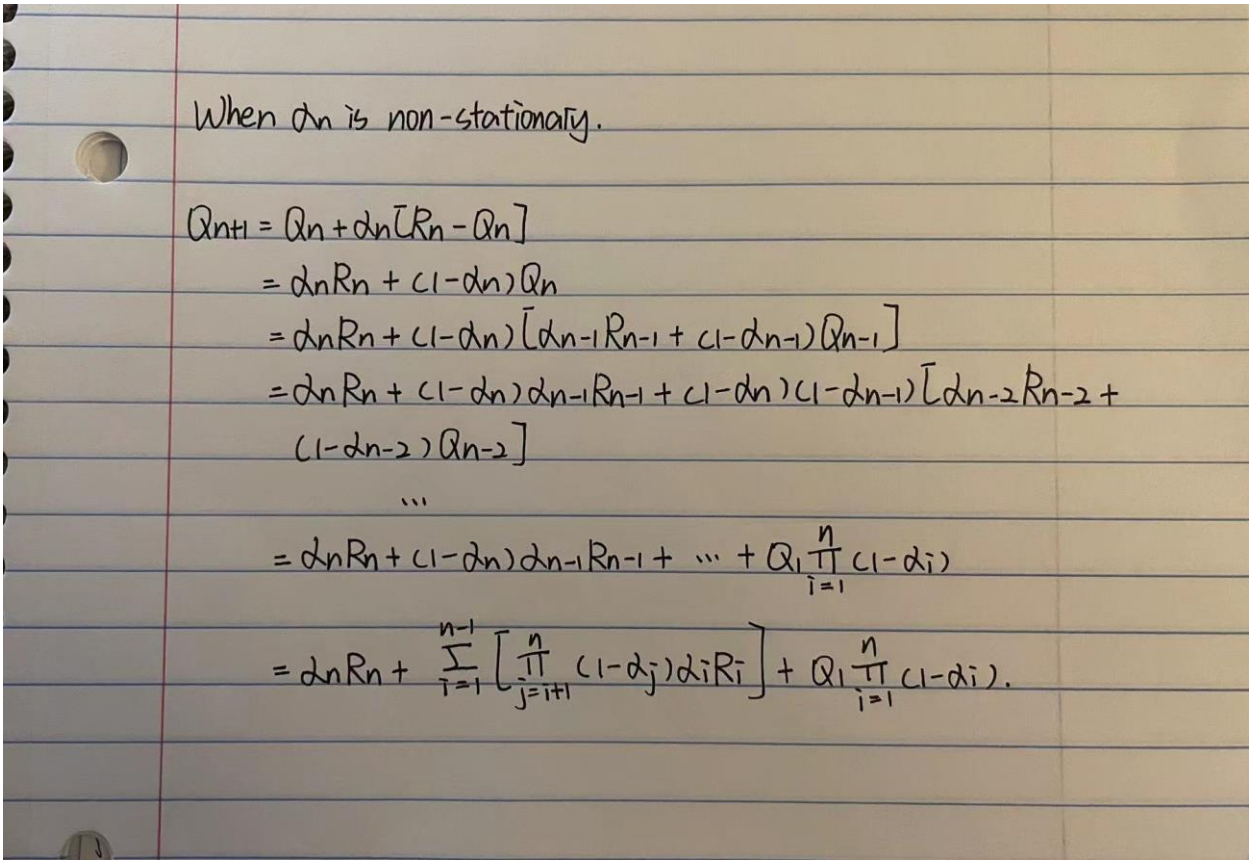
Q1:



Thus, it may occurred on step 1, 2 and 3

It definitely occurred on step 4 and 5

Q2:



When α_n is non-stationary.

$$
\begin{aligned}
Q_{n+1} &= Q_n + \alpha_n[R_n - Q_n] \\
&= \alpha_n R_n + (1-\alpha_n) Q_n \\
&= \alpha_n R_n + (1-\alpha_n)[\alpha_{n-1} R_{n-1} + (1-\alpha_{n-1}) Q_{n-1}] \\
&= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1} R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})[\alpha_{n-2} R_{n-2} + \\
&\quad (1-\alpha_{n-2}) Q_{n-2}] \\
&\quad \cdots \\
&= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1} R_{n-1} + \cdots + Q_1 \prod_{i=1}^{n} (1-\alpha_i) \\
&= \alpha_n R_n + \sum_{i=1}^{n-1} \left[\prod_{j=i+1}^{n} (1-\alpha_j) \alpha_i R_i \right] + Q_1 \prod_{i=1}^{n} (1-\alpha_i).
\end{aligned}
$$

Q3:

(a) Equation 2.1: $Q_n = \frac{R_1 + R_2 + \cdots R_{n-1}}{n-1}$, $E(R_n) = q_*$

$$E|Q_n| = E\left|\frac{R_1 + R_2 + \cdots R_{n-1}}{n-1}\right|$$

$$= \frac{1}{n-1} \times E|R_1 + R_2 + \cdots R_{n-1}|$$

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

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

$$= q_* .$$

Thus, it is unbiased.

(b) If $Q_1 = 0$. for $n > 1$:

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

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

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

Thus, only when $\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$, it is unbiased.

Q3:

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

When $E|Q_{n+1}| = q^*$, it is unbiased.

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

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

$= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} q^* = q^*$, it is unbiased.

The condition:

Therefore. $(1-\alpha)^n Q_1 = 0$, $\rightarrow$ $Q_1 = 0$; $\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$.

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

When $n \to \infty$. since $0 < \alpha < 1$. $(1-\alpha)^n Q_1 \to 0$.

$\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1-(1-\alpha)^{n-1}$ and $(1-\alpha)^{n-1} \to 0$. ~~thus~~

thus. $\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} \to 1$. then it satisfied the condition $\begin{cases} Q_1 = 0 \\ \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1 \end{cases}$

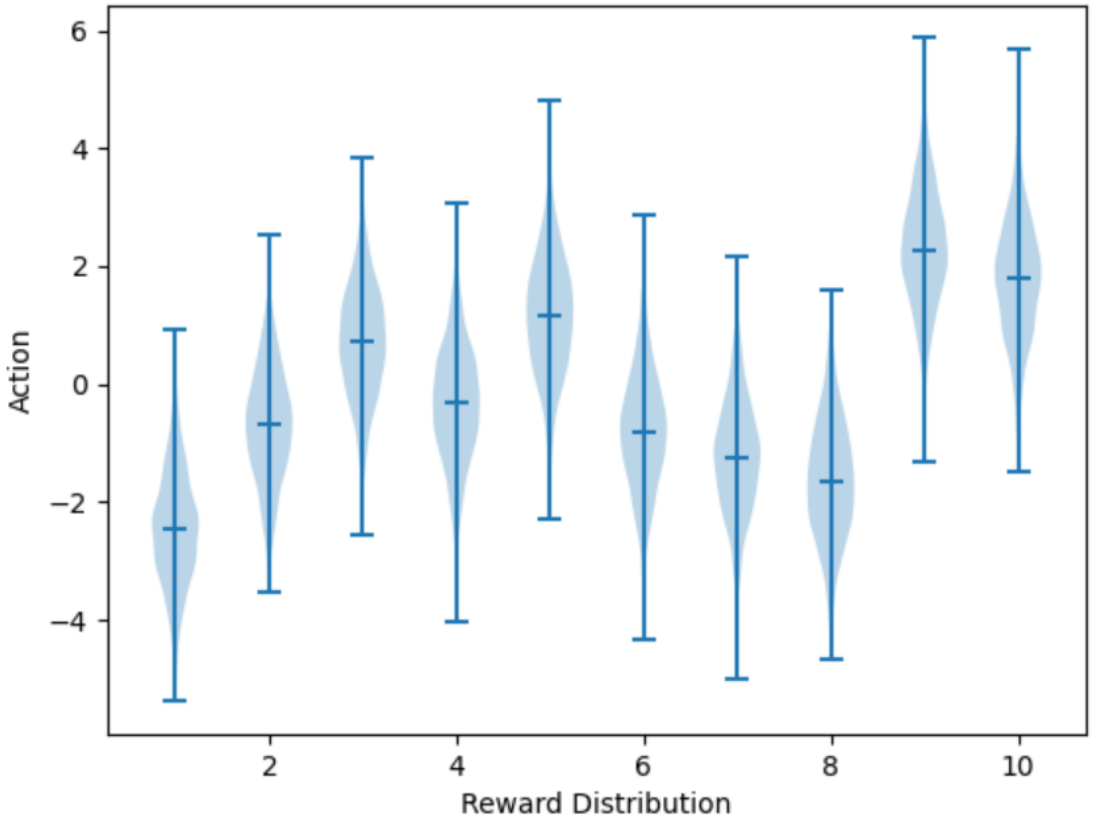
so it is unbiased when $n \to \infty$.

Q3:

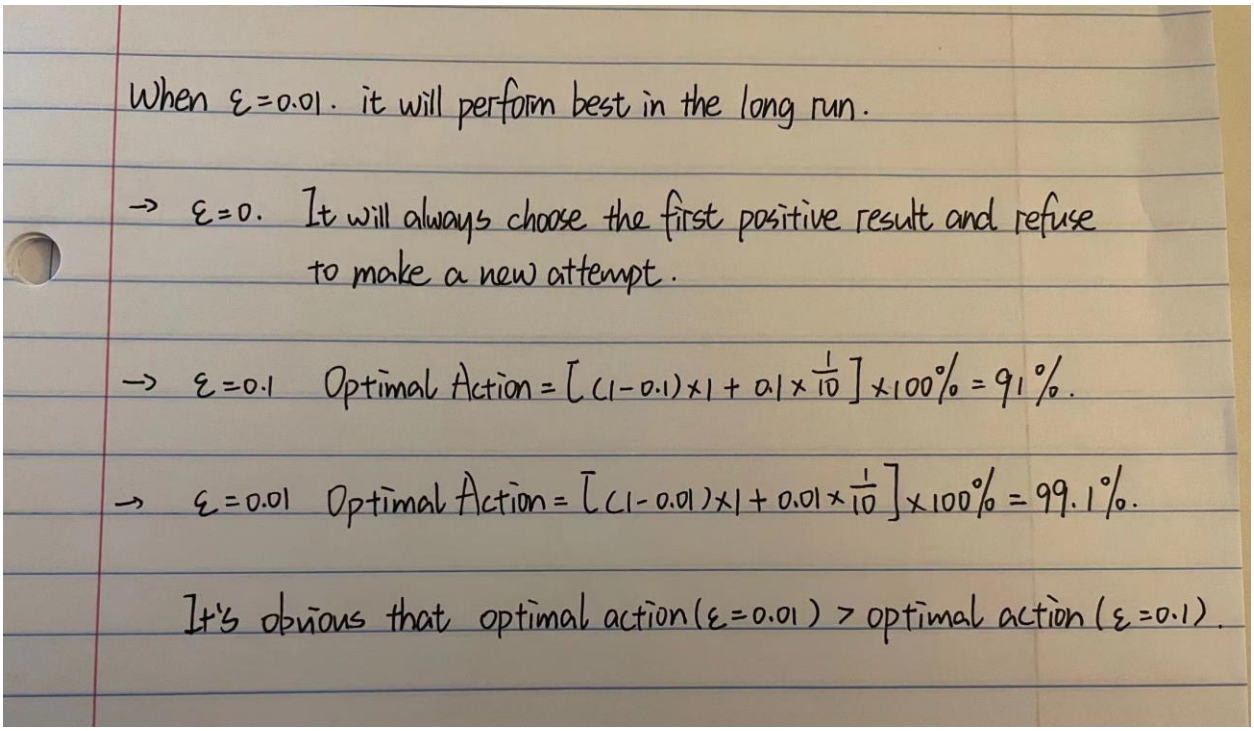
(e) **Why should we expect that the exponential recency-weighted average will be biased in general?**

1. *the actual situation will never be perfectly.*
2. *we can never obtain infinite values, but we can obtain values large enough that this still leads to the possibility of some bias*
3. *The occurrence of bias is more common than the absence of bias*

Q4:



Q5:



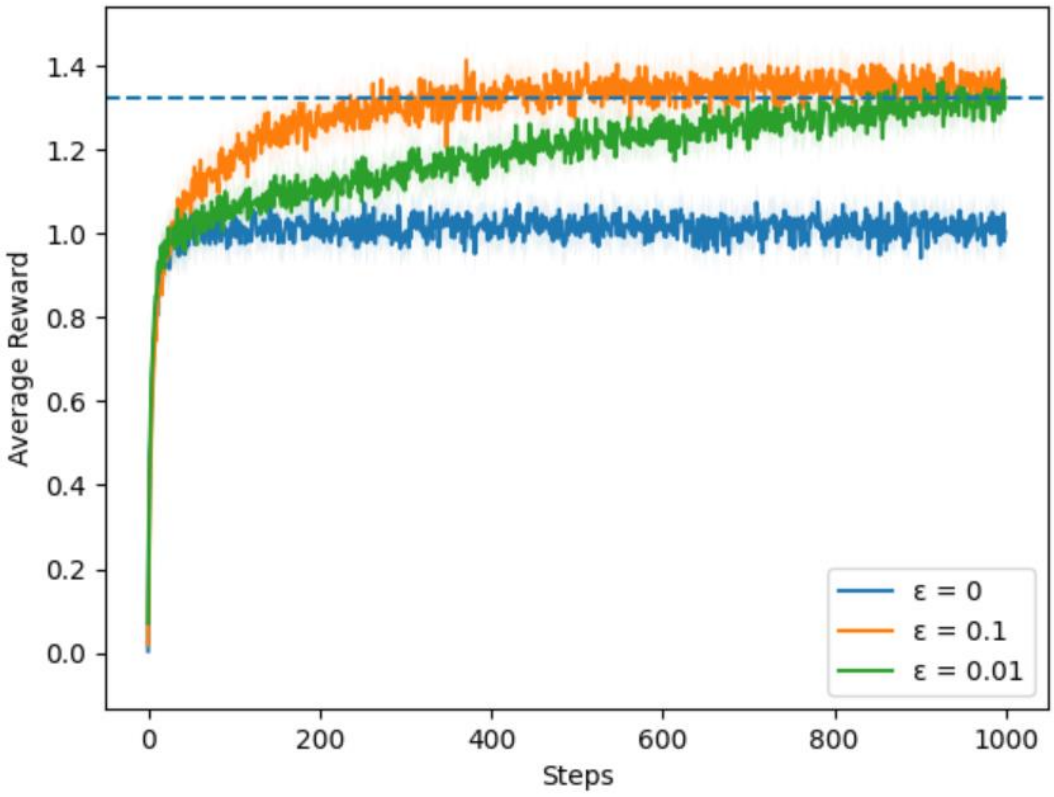
The optimal action(ε = 0.01) is 8.1% better than the optimal action(ε = 0.1)

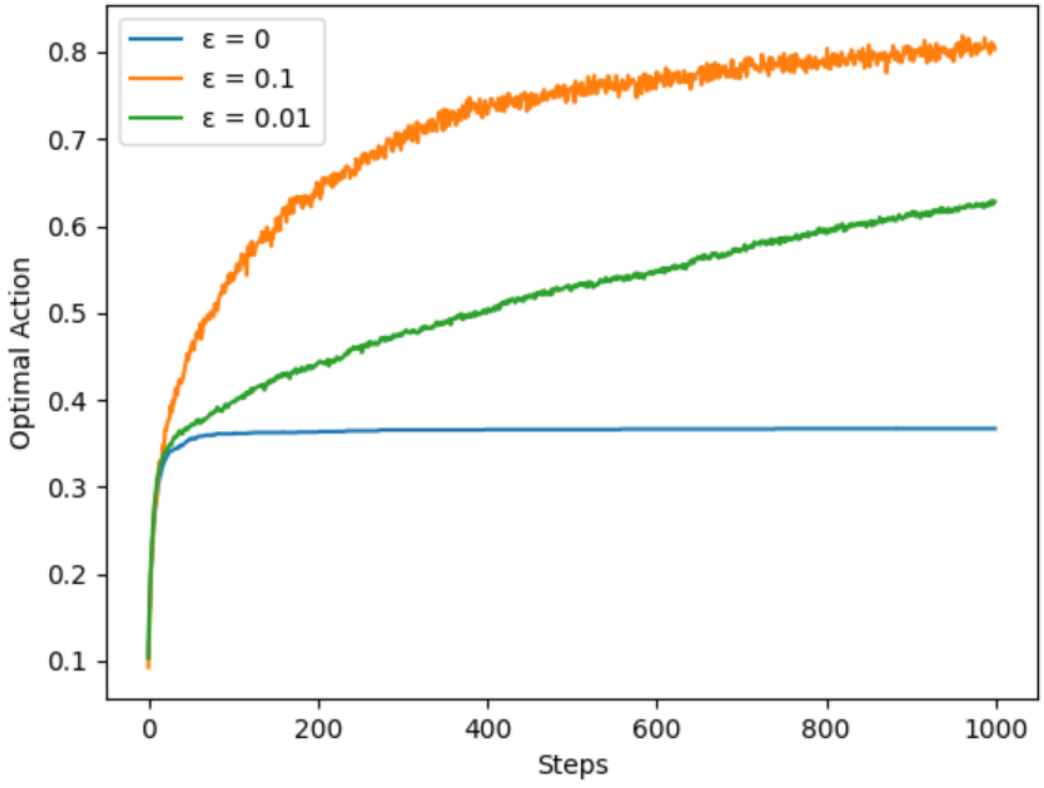
Q6:

Do the averages reach the asymptotic levels predicted in the previous question?

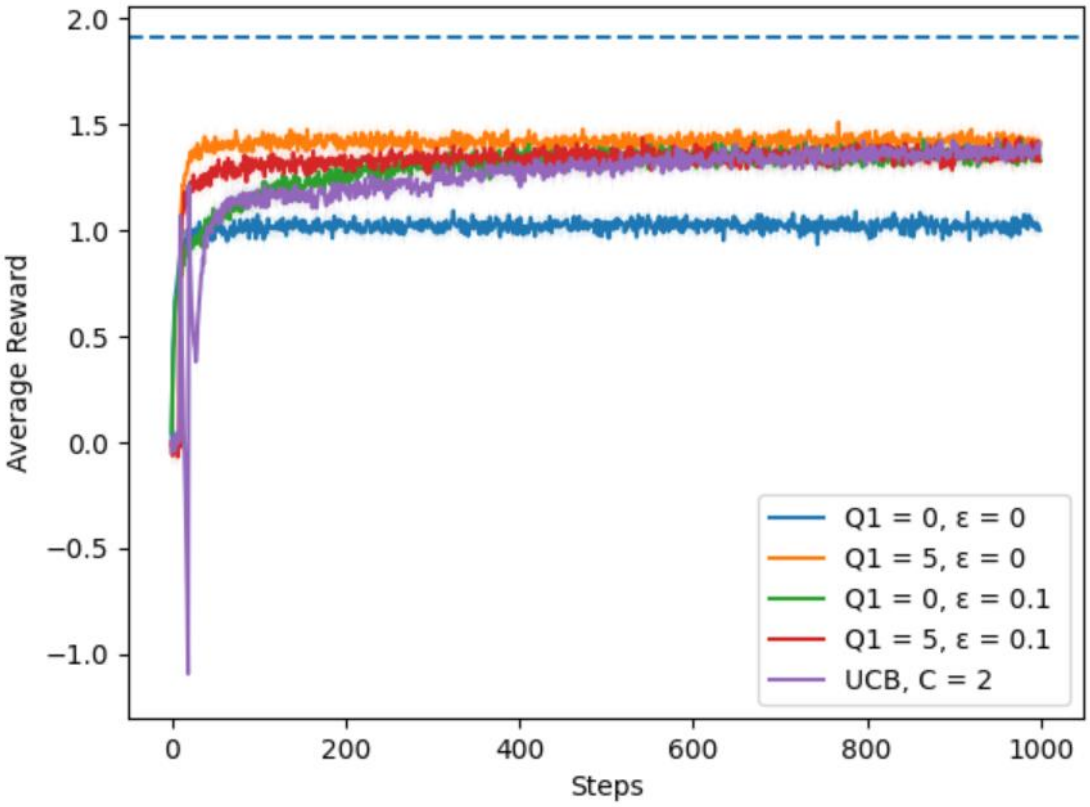
I don't think we have enough steps to get to the asymptotic levels predicted in the previous question. The more steps we have (close to infinity), the better the performance.

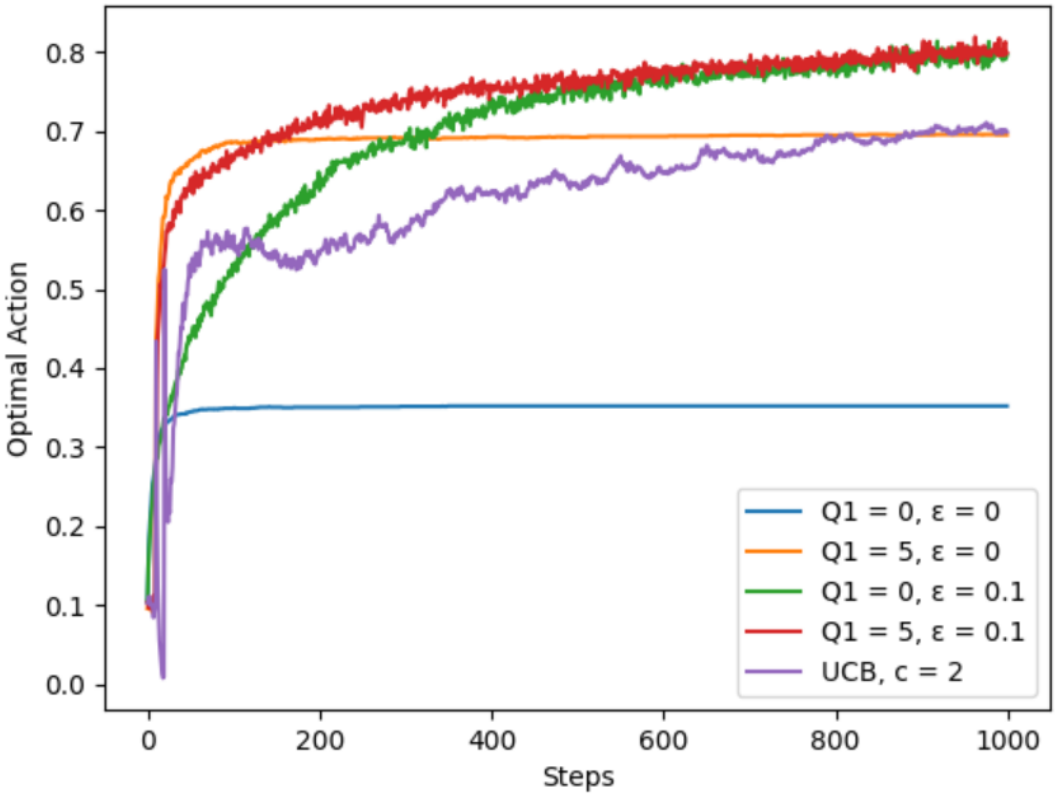
Q6:


Average Reward
Steps
ε = 0
ε = 0.1
ε = 0.01


Optimal Action
Steps
ε = 0
ε = 0.1
ε = 0.01

Q7:


Average Reward
Steps
Q1 = 0, ε = 0
Q1 = 5, ε = 0
Q1 = 0, ε = 0.1
Q1 = 5, ε = 0.1
UCB, C = 2


Optimal Action
Steps
Q1 = 0, ε = 0
Q1 = 5, ε = 0
Q1 = 0, ε = 0.1
Q1 = 5, ε = 0.1
UCB, c = 2

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.

1. I believe that the optimistic initialization products spike is due to the fact that a high Q value was given at the beginning and that a relatively stable result could not be obtained immediately after further exploration.

2. For the UCB, I think it is due to the fact that it first needs to try all the actions to find the one that is likely to obtain a high reward, thus generating the spike, and after 10 actions he will adjust in order to obtain a more stable performance, and that products another spike.