1. $R_0 = 0$ $\quad Q_1(1) = Q_1(2) = Q_1(3) = Q_1(4) = 0$

t_1 $\quad A_1 = 1$ $\quad$ Since $Q_1(a)$ was same, ε-greedy may occured

$R_1 = -1$ $\quad Q_2(1) = -1$ $\quad Q_2(2) = Q_2(3) = Q_2(4) = 0$

t_2 $\quad A_2 = 2$ $\quad$ Since $Q_2(2) = Q_2(3) = Q_2(4)$, ε-greedy may occured

$R_2 = 1$ $\quad Q_3(1) = -1$ $\quad Q_3(2) = 1$, $Q_3(3) = 0$, $Q_4(3) = 0$

t_3 $\quad A_3 = 2$ $\quad$ Since Action 1 was not selected, ε-greedy may occured.

$R_3 = -2$ $\quad Q_4(1) = -1$ $\quad Q_4(2) = 1 + \frac{1}{2}(-2-1) = -\frac{1}{2}$, $Q_4(3) = 0$, $Q_4(4) = 0$

t_4 $\quad A_4 = 2$ $\quad$ Since $Q_4(2) < Q_4(3) = Q_4(4)$, ε-greedy definitely occured

$R_4 = 2$ $\quad Q_5(1) = -1$ $\quad Q_5(2) = -\frac{1}{2} + \frac{1}{3}(2 + \frac{1}{2}) = \frac{1}{3}$, $Q_5(3) = 0$, $Q_5(4) = 0$

t_5 $\quad A_5 = 3$ $\quad$ Since $Q_5(3) < Q_5(2)$, ε-greedy definitely occured.

In sum, the ε-greedy may occured on time step 1, 2, 3

the ε-greedy definitely occured on time step 4, 5

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

$= \alpha_n R_n + (1-\alpha_n) Q_n$

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

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

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

$= \alpha_n R_n + (1-\alpha_n)\alpha_{n-1} R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})[Q_{n-2} + \alpha_{n-2}[R_{n-2} - Q_{n-2}]]$

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

$\cdots$

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

3.

(a) $Q_n = \frac{R_1 + \cdots + R_{n-1}}{n-1} = q^*$ when n is large enough

for E.g. 2.1

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

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

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

$\because E(R_n) = q^*$

$\therefore E(Q_n) = \frac{1}{n-1}(n-1) q^* = q^*$

$$Q_{n+1} = Q_n + \alpha [R_n - Q_n]$$

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

$\because Q_1 = 0$

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

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

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

If $\sum_{i=1}^{n} \alpha (1-\alpha)^{n-i} = 1$ $E(Q_{n+1}) = q^*$ (unbiased)

else $E(Q_{n+1}) \neq q^*$ (biased)

(c)

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

$$E(Q_{n+1}) = E\left[(1-\alpha)^n Q_1\right] + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} q^*$$

$E(Q_{n+1}) = q^*$

If the $E(Q_{n+1})$ is unbiased,

$$E(Q_{n+1}) = q^*$$

$$\therefore (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} q^* = q^*$$

$\therefore$ Conditions:

$$(1-\alpha)^n Q_1 = 0 \longrightarrow \boxed{Q_1 = 0}$$

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

(d) Due to $(1-\alpha)^n + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$

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

when $n \to \infty$

$$(1-\alpha)^n \to 0$$

$$1-(1-\alpha)^n \longrightarrow 1$$

$$\therefore (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} q^* = q^*$$

that is:

$$0 \cdot Q_1 + 1 \cdot q^* = q^*$$

as:

$$q^* = q^*$$

Hence, Q_n is asymptotically

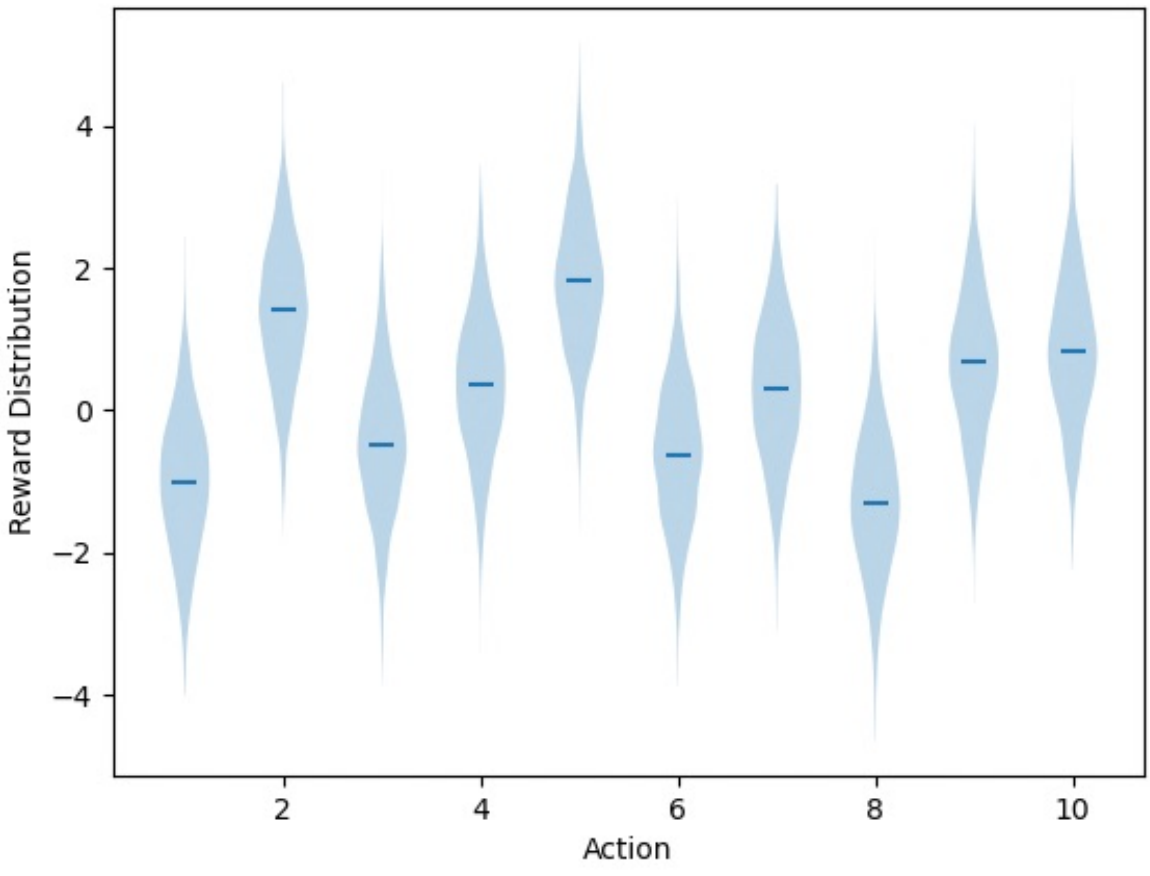
$1-(1-\alpha)^{n-1}$

$E(Q_{n+1}) = q^*$

(e) First, We cannot acquired infinite samples, which means that n is not equal to ∞. As a result, $\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} \neq 1$, so it must have some bias in general. Besides, $Q_n = \frac{R_1 + \cdots + R_{n-1}}{n-1} = q_n^*$ when n is large enough (finite).

But if n is infinite, $Q_n \neq q^*$, which also denote some bias in the equation. All in all, we should expect that the exponential recency-weighted average will be biased in general.

4. Plot



5.

$\varepsilon = 0.01$ will perform best in the long run.

when $\varepsilon = 0.01$,

$$\text{Optimal Action} = 1 - 0.01 + 0.01 \times \frac{1}{10} = 0.991 = 99.1\%$$

when $\varepsilon = 0.1$,

$$\text{Optimal Action} = 1 - 0.1 + 0.1 \times \frac{1}{10} = 0.91 = 91\%$$

when $\varepsilon = 0$,

It will always select the highest reward of current state without any exploration. In the long run, it will lead to low probability of selecting optimal action.

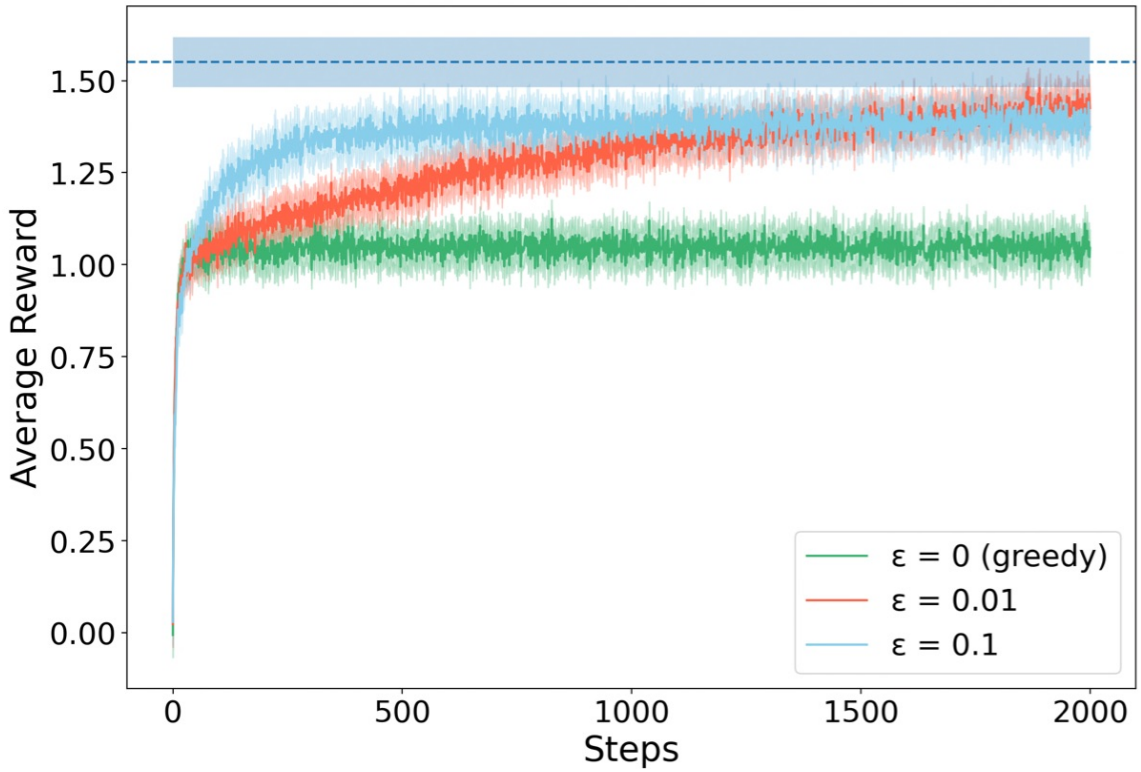
quantitatively:

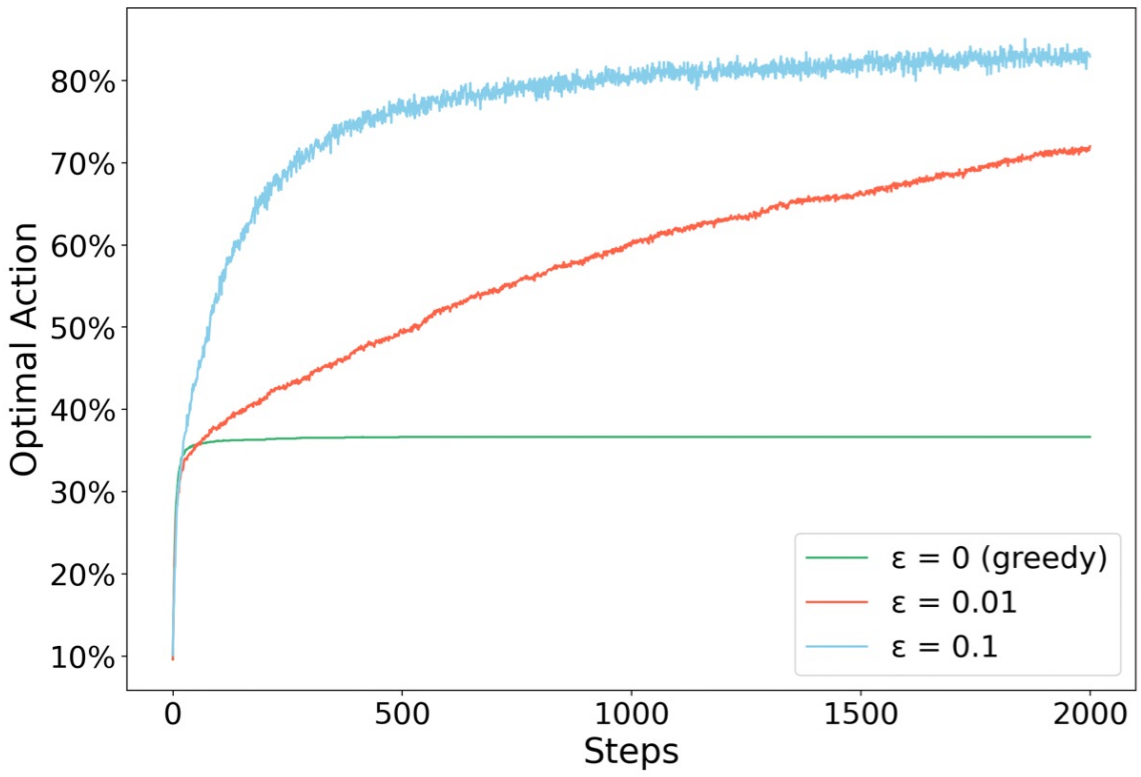
$\varepsilon = 0.01$ is better than $\varepsilon = 0.1$ about 8.1 %

6. Written

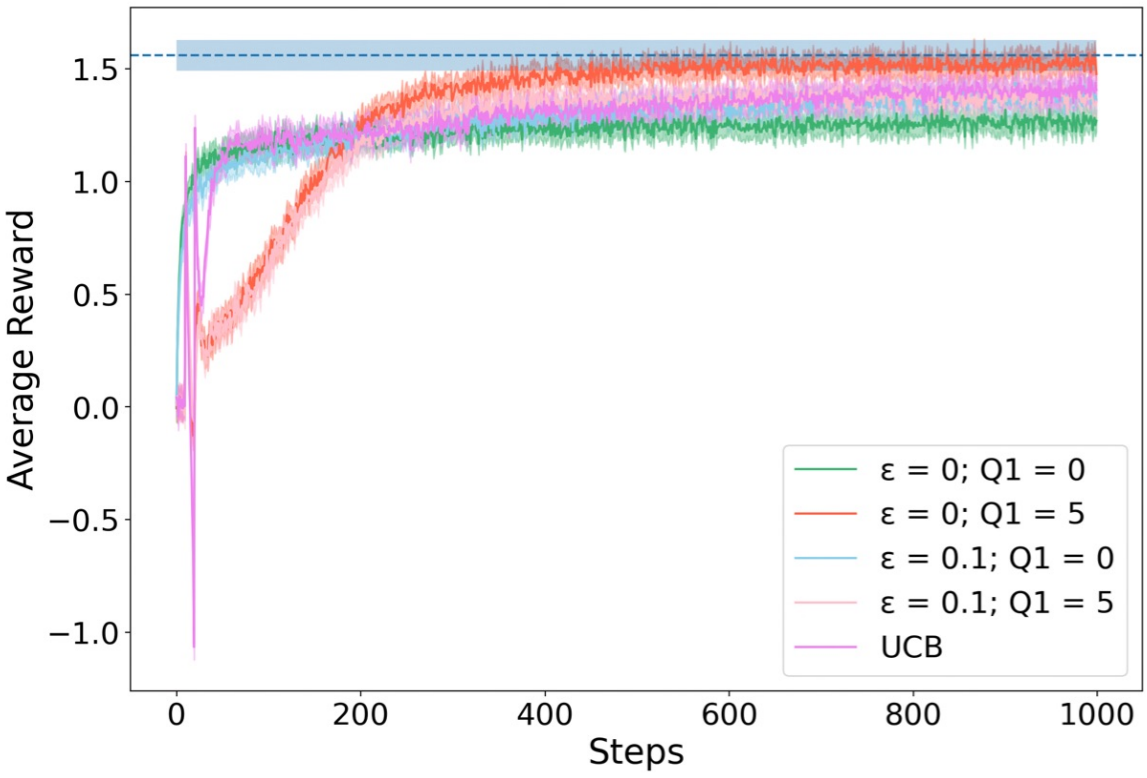
From my observation (as shown below), the averages don't reach the asymptotic level. I think the reason is that the steps are still not enough. If we have much more steps, even approaching infinity, the performance and the average reward will reach the asymptotic level.

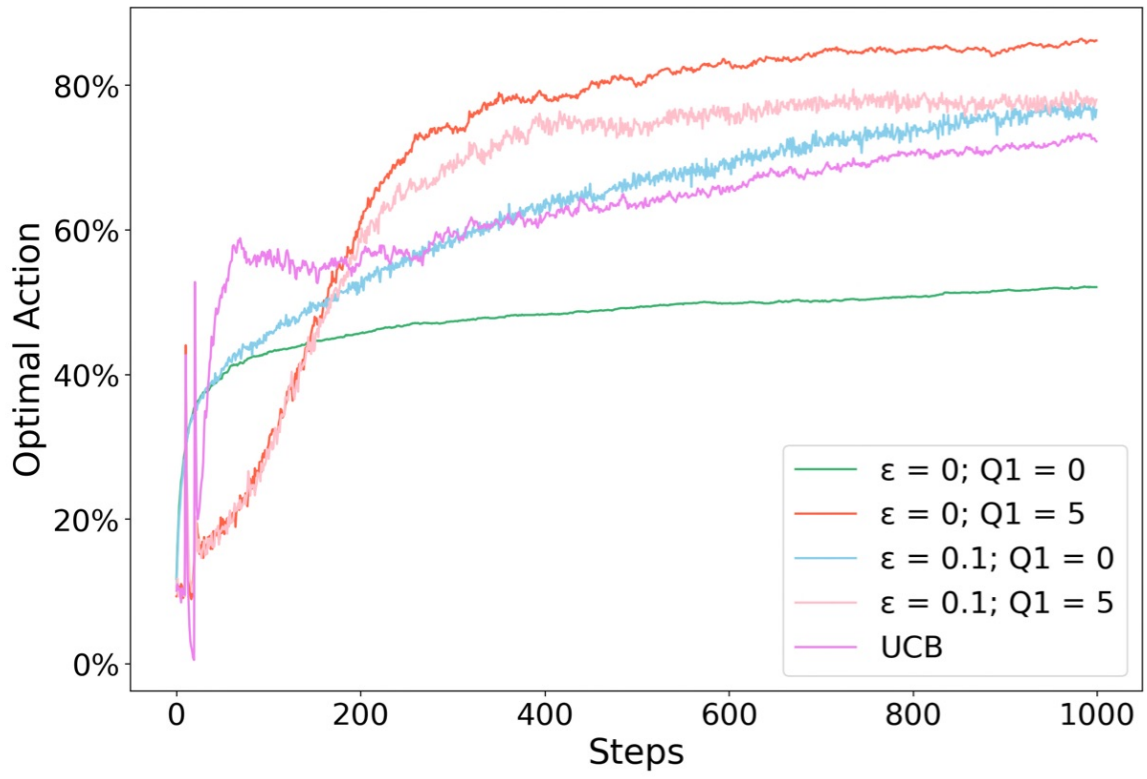
6. Plot


Average Reward
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
2000
Steps
ε = 0 (greedy)
ε = 0.01
ε = 0.1


Optimal Action
80%
70%
60%
50%
40%
30%
20%
10%
0
500
1000
1500
2000
Steps
ε = 0 (greedy)
ε = 0.01
ε = 0.1

7. Plot


Average Reward
1.5
1.0
0.5
0.0
−0.5
−1.0
0
200
400
600
800
1000
Steps
ε = 0; Q1 = 0
ε = 0; Q1 = 5
ε = 0.1; Q1 = 0
ε = 0.1; Q1 = 5
UCB


Optimal Action
80%
60%
40%
20%
0%
0
200
400
600
800
1000
Steps
ε = 0; Q1 = 0
ε = 0; Q1 = 5
ε = 0.1; Q1 = 0
ε = 0.1; Q1 = 5
UCB

7. Written.

I think the reason of spikes is because the exploration is not enough. Also, the past experience cannot support the agent to select the optimal actions. With the continuous exploration process, the agent can gradually select the optimal actions. During this process, the average rewards are increased/decreased rapidly. That's why the spikes appear.

(So the agent will randomly select actions at begining.)

From the UCB in my experimental data, we could find the spikes (both the sharp increase and decrease). The reason is that the UCB first needs to try all actions. Some actions have high reward, some are not, so that's why the spikes appear. The inner property of the UCB (explore all actions at the begining) leads to the spikes.

(Since the rewards are initialized randomly according to the normal distribution.)