	A0=0, R0=0. Q1=0, Q2=0, Q4=0. (Consistent with
	Choose 1. Maybe random. Cuz Q1 = Q2 = Q3 = Q4 = or &-greedy action
	7. Comments
	$A_{1}=1$ , $R_{1}=-1$ , $Q_{1}=\frac{-1}{1}=-1$ . $Q_{2}=0$ , $Q_{3}=0$ , $Q_{4}=0$ .
	Choose 2. Maybe random. Consistent with &-years
	Cuz $Q_1 = -1 \angle Q_2 = Q_3 = Q_4 = 0$ . (action selection.
	$A_2 = 2 \cdot R_2 = 1 \cdot Q_1 = -1 \cdot Q_2 = \frac{1}{1} = 1 \cdot Q_3 = 0 \cdot Q_4 = 0$ .
	Choose 2. Maybe random. Consistent with &-
	Cut $Q_2=1 > Q_3=Q_4=0 > Q_1=-1$ action selection.
	$A_{3=2}$ . $R_{3=-2}$ , $Q_{1=-1}$ . $Q_{2}=\frac{-1}{2}=-0.5$ , $Q_{3}=0$ , $Q_{4}=0$ .
-181	Choose 2. Must be random.
	Cut Q3=Q4=0 > Q2=-0.5 > Q1=-1.
	According to E-greedy action selection,
	it should choose 3 or 4.
	A4=2, R4=2, Q1=-1. Q2====================================
	Choose 3. Must be random.
	Cut Q2=0.33 > Q3 = Q4=0 > (Q=-1.
	According to e-greedy action selection.
	it should choose 2.
	$5=3$ , $R_{5}=0$ . $Q_{1}=-1$ . $Q_{2}=0.33$ . $Q_{3}=\frac{0}{1}=0$ . $Q_{4}=0$ .
1)	9-7. 16=0. W1-1. W1-0.77. Wg = 1 =0. W4=0.

Thus, it may occurred on step 1, 2 and 3
It definitely occurred on step 4 and 5

When on is non-stationary.
anti = an + dn [Rn - an]
= dnRn + c(-dn)Qn
= dnRn+ (1-dn) [dn-1 Rn-1 + c1-dn-1) Qn-1]
=dnRn+c1-dn)dn-1Rn-1+c1-dn)c1-dn-1) [dn-2Rn-2+
(1-dn-2) an-2]
···
= dnRn+ (1-dn)dn-1Rn-1+ + Q1 + C1-di)
N-1 - N
$= dnRn + \prod_{i=1}^{n-1} \left[ \prod_{j=i+1}^{n} (1-d_j) d_i R_i \right] + Q_i \prod_{j=1}^{n} (1-d_i).$

3	R1+R2+ Rn-1 , E(Rn)=9*
(A)	Equation 2.1: Qn = n-1
	$E[Q_n] = E[\frac{P_1 + R_2 + \cdots R_{n-1}}{n-1}]$
	$= \frac{1}{n-1} \times E   R_1 + R_2 + \dots R_{n-1}  $
	$= \frac{1}{n-1} \times (E[R_1] + E[R_2] + \dots + E[R_{n-1}])$
	$=\frac{1}{n-1} \times (n-1) $ 9.*
	= Q* ·
1	
	Thus, it is unbiased.
رط)	If Q1=0. for n>1:
	$(2n+1 = (1-d)^n Q_1 + \sum_{i=1}^n d_i (1-d)^{n-i} R_i = \sum_{i=1}^n d_i (1-d)^{n-i} R_i$
	$E Q_{n+1}  = E \prod_{i=1}^{n} d(i-d)^{n-i}R_{i}  = \prod_{i=1}^{n} d(i-d)^{n-i}E R_{i} $
	$= \frac{1}{1-1} \alpha(1-\alpha)^{n-1} 0$
	Thus, only when $I =  \alpha(1-d)^{n-1} = 1$ , it is unbiased.
0	

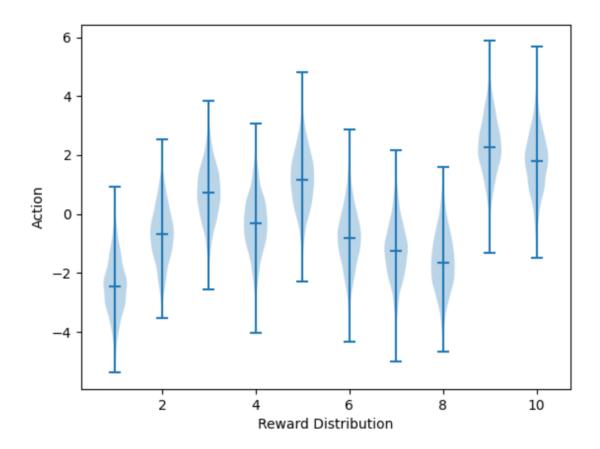
(C)	$Q_{n+1} = (1-a)^n Q_1 + \frac{1}{i=1} d(1-a)^{n-i} R_i$
	When ElQn+1 = 9+, it is unbiased.
	Thus. $E Q_{n+1}  = E c_1-d_1^nQ_1+\prod_{i=1}^nd_i(1-d_i^{n-i}R_i)$
	$= (1-4)^n Q_1 + \sum_{i=1}^n \alpha(1-4)^{n-i} E[R_i]$
	= $(1-a)^n Q_1 + \sum_{j=1}^{n} d(1-a)^{n-j} Q_* = Q_*$ , it is unbiased.
	The condition:  Therefore. $(1-\lambda)^n Q_1 = 0$ , $\Rightarrow Q_1 = 0$ ; $\sum_{i=1}^n \lambda(1-\lambda)^i = 1$ .

(d)	$Q_{n+1} = (1-d)^{n}Q_{1} + \sum_{i=1}^{N} \alpha(1-d)^{n-i}R_{i}^{-i}$
	When n -> 00. since ocacl. (1-d) Q1->0.
	$\sum_{i=1}^{N} \alpha(1-\alpha)^{i} = 1 - (1-\alpha)^{n-1} \text{ and } (1-\alpha)^{n-1} \to 0. \text{ thus}$ $1 = 1 - (1-\alpha)^{n-1} \text{ and } (1-\alpha)^{n-1} \to 0.$
	thus. $\sum_{i=1}^{n} d(1-d)^{n-i} \rightarrow 1$ . then it satisfied the condition $\begin{cases} 1 & 0 \\ 1 & 1 \end{cases}$
	50 it is unbiased when n > 00.

## (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:



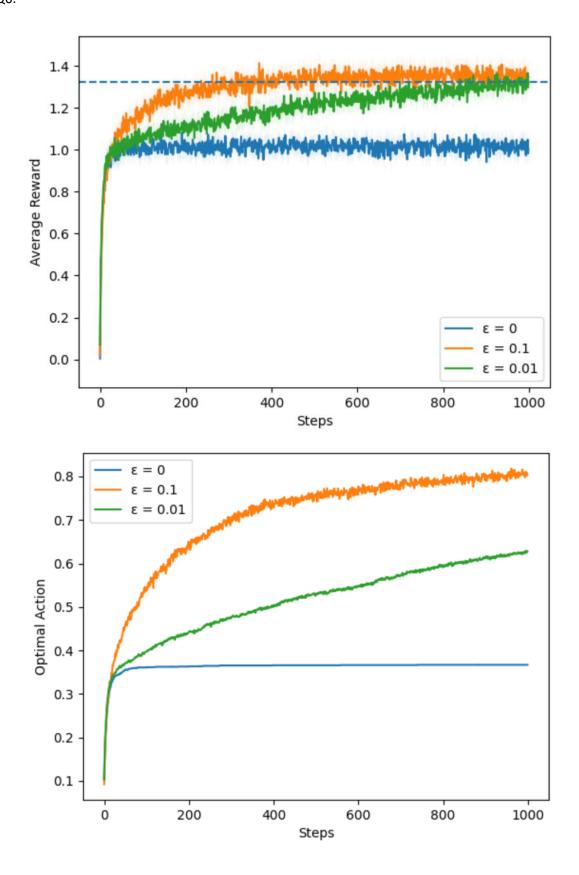
When &=0.01. it will perform best in the long run.
-> &=0. It will always choose the first positive result and refuse to make a new attempt.
-> &=0.1 Optimal Action = [(1-0.1)x1+01x10] x100% = 91%.
-> &=0.01 Optimal Action = [(1-0.01)x +0.01x10]x100%=99.1%.
It's obvious that optimal action ( $\varepsilon = 0.01$ ) > optimal action ( $\varepsilon = 0.1$ ).

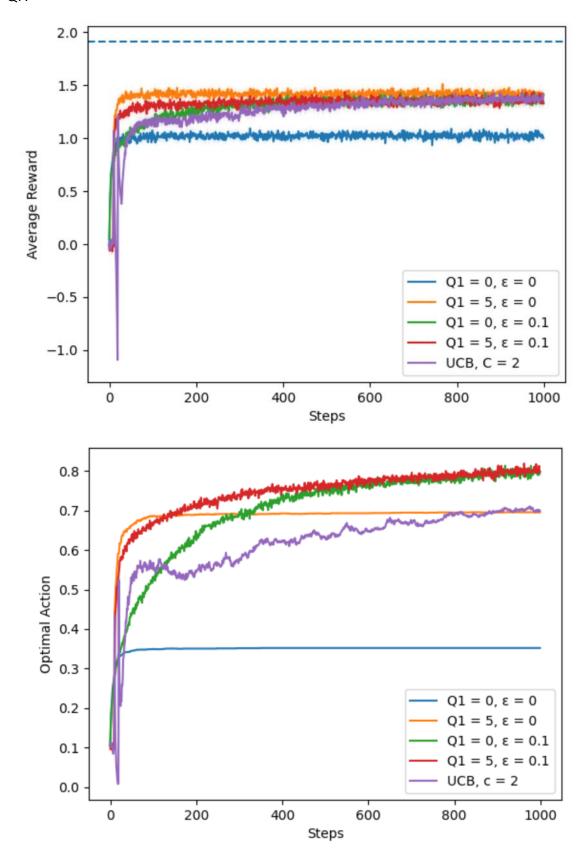
The optimal action(  $\varepsilon$  = 0.01) is 8.1% better than the optimal action(  $\varepsilon$  = 0.1)

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.





## 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.