CS 4180/5180: Reinforcement Learning and Sequential Decision Making (Fall 2020)– Lawson Wong

Name: [Virender Singh]
Collaborators: [Sunny Shukla, Saurabh Vaidya]

Code Execution To check the working of the code please execute main.py in the src folder and you can check all the plots in the plt folder, to run the code,you can comment rest of the questions and run one at a time to check the results.

Question 1. *1. Exploration vs. exploitation.*
Written: Consider a k armed bandit problem with k = 4 actions, denoted 1, 2, 3, and 4. Consider applying to this problem a bandit algorithm using ϵ-greedy action selection, sample-average action-value estimates, and initial estimates of $Q_1(a) = 0$, for all a. Suppose the initial sequence of actions and rewards is $A_1 = 1$, $R_1 = -1$, $A_2 = 2$, $R_2 = 1$, $A_3 = 2$, $R_3 = -2$, $A_4 = 2$, $R_4 = 2$, $A_5 = 3$, $R_5 = 0$. On some of these time steps the ϵ case may have occurred, causing an action to be selected at random. On which time steps did this definitely occur? On which time steps could this possibly have occurred?

Response:

Time	Action	Reward	$Q_t(1)$	$Q_t(2)$	$Q_t(3)$	$Q_t(4)$
1	1	-1	0	0	0	0
2	2	1	-1	0	0	0
3	2	-2	-1	1	0	0
4	2	2	-1	$\frac{-1}{2}$	0	0
5	3	0	-1	$\frac{1}{3}$	0	0

At time step 1, when we chose action 1, the ϵ case **may have occurred** because no action with higher Q value was present at that time.
At time step 2, when we chose action 2, the ϵ case **may have occurred** because no action with higher Q value was present at that time.
At time step 3, when we chose action 2, there were no actions with better q values so the ϵ case **may have occurred**.
At time step 4, when we chose action 2, there were actions present with better q values like action 3 and 4 with q value $\frac{1}{2}$, so here the ϵ case **definitely occurred**.
Similarly, at time step 5, we chose action 3, but there was action 2 with value $\frac{1}{3}$ and hence the ϵ case **definitely occurred**.

Question 2. . *(RL2e 2.4) Varying step-size weights.*
Written: If the step-size parameters, α, are not constant, then the estimate Q_n is a weighted average of previously received rewards with a weighting different from that given by Equation 2.6. What is the weighting on each prior reward for the general case, analogous to Equation 2.6, in terms of the sequence of step-size parameters?

Response:
Equation 2.6 is derived from Equation 2.5 by taking α to be constant, here $\alpha \in (0, 1)$. However, we have to derive the relation for general case. Let's use α_i for step i and look at equation 2.5.

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

Expanding one step at a time,

$$\begin{aligned}
Q_{n+1} &= (1-\alpha_n)Q_n + \alpha_n R_n & (2)\\
&= (1-\alpha_n)(Q_{n-1} + \alpha_{n-1}(R_{n-1} - Q_{n-1})) + \alpha_n.R_n & (3)\\
&= (1-\alpha_n)(1-\alpha_{n-1}).Q_{n-1} + \alpha_n.R_n + (1-\alpha_n).\alpha_{n-1}.R_{n-1} & (4)\\
&= (1-\alpha_n)(1-\alpha_{n-1})(1-\alpha_{n-2}).Q_{n-2} + \alpha_n.R_n + (1-\alpha_n)\alpha_{n-1}R_{n-1} + (1-\alpha_n)(1-\alpha_{n-1})\alpha_{n-2}R_{n-2} & (5)\\
&= (1-\alpha_n)(1-\alpha_{n-1})(1-\alpha_{n-2})...(1-\alpha 1)Q_1 + \alpha_n R_n + (1-\alpha_{n-1})\alpha_{n-1}R_{n-1} + & (6)\\
(1-\alpha_n)&(1-\alpha_{n-1})\alpha_{n-2}R_{n-2} + ... + (1-\alpha_n)(1-\alpha_{n-1})...(1-\alpha_2)\alpha_1 R_1 & (7)
\end{aligned}$$

Therefore, writing in the formula form:

$$Q_{n+1} = Q_1 \Pi_{i=1}^{n}(1-\alpha_i) + \Sigma_{i=1}^{n} \alpha_i R_i \Pi_{j=i+1}^{n}(1-\alpha_j) \tag{8}$$

Question 3. *(a) Consider the sample-average estimate in Equation 2.1. Is it biased or unbiased? Explain briefly. For the remainder of the question, consider the exponential recency-weighted average estimate in Equation 2.5. Assume that $0 < \alpha < 1$ (i.e., it is strictly less than 1).*
(b) If $Q_1 = 0$, is Q_n for $n > 1$ biased? Explain briefly.
(c) Derive conditions for when Q_n will be unbiased.
(d) Show that Q_n is asymptotically unbiased, i.e., it is an unbiased estimator as $n \Rightarrow \infty$.
(e) Why should we expect that the exponential recency-weighted average will be biased in general?

Response:

(a)

Using equation 2.1, we have
For any action a,

$$Q_t(a) = \frac{\Sigma_{i=1}^{t-1} \mathbb{R}_i . 1_{A_i=a}}{\Sigma_{i=1}^{t-1} 1_{A_i=a}} \tag{9}$$

$$\tag{10}$$

Taking expectation on both the sides,
where, Q_t(a) is the Q value of action a at time t,

$$\begin{aligned}
E[Q_t(a)] &= E[\frac{\Sigma_{i=1}^{t-1} \mathbb{R}_i . 1_{A_i=a}}{\Sigma_{i=1}^{t-1} 1_{A_i=a}}] \\
&= E[\frac{R_1 + R_2 + ... + R_n}{t-1}] \\
&= \text{Assuming the algorithm has converged and takes only action a at each time step} \\
&= \frac{E[R_1 + R_2 + ... + R_n]}{t-1} \\
&= \frac{E[R_1] + E[R_2] + ... + E[R_n]}{t-1} \\
&= \frac{q_*(a) + q_*(a) + ... + q_*(a)}{t-1} \\
&= E[R_i] = q_*(a) \\
&= \frac{t-1}{t-1}(q_*(a)) \\
&= q_*(a)
\end{aligned}$$

(b) Using equation 2.6

$$
\begin{align}
Q_{n+1} &= (1-\alpha)^n Q_1 + \Sigma_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i \tag{11}\\
Q_{n+1} &= \Sigma_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i; \text{Since } Q_1 = 0 \tag{12}\\
E[Q_{n+1}] &= E[\Sigma_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i] \tag{13}\\
E[Q_{n+1}] &= E[\alpha(1-\alpha)^{n-1} R_1 + \alpha(1-\alpha)^{n-2} R_2 + ... + \alpha(1-\alpha)^(n-n) R_n] \tag{14}\\
E[Q_{n+1}] &= \alpha E[(1-\alpha)^{n-1} R_1 + (1-\alpha)^{n-2} R_2 + ... + (1-\alpha)^0 R_n] \tag{15}\\
E[Q_{n+1}] &= \alpha[(1-\alpha)^{n-1} E[R_1] + (1-\alpha)^{n-2} E[R_2] + ... + (1-\alpha)^0 E[n]] \tag{16}\\
E[Q_{n+1}] &= \alpha[(1-\alpha)^{n-1} q_* + (1-\alpha)^{n-2} q_* + ... + 1] \text{since } E[R_i] = q_* \tag{17}\\
&= \text{Taking geometric progression sum for the above series} \tag{18}\\
E[Q_{n+1}] &= \alpha q_* \frac{1-(1-\alpha)^n}{1-(1-\alpha)} \tag{19}\\
E[Q_{n+1}] &= \alpha q_* \frac{1-(1-\alpha)^n}{\alpha} \tag{20}\\
E[Q_{n+1}] &= q_* .1 - (1-\alpha)^n \tag{21}\\
&= \text{hence the expectation depends on the value of } \alpha \text{ and thus} \tag{22}\\
& \tag{23}
\end{align}
$$

Thus $E[Q_{n+1}] \neq q_*$ and hence we can say that Q_n derived in equation 2.5 is biased.

(c)

Again using equation 2.5 we have,

$$Q_{n+1} \doteq Q_n + \alpha[R_n - Q_n] \tag{24}$$

and equation 2.6 is derived from it which states

$$\begin{aligned}
Q_{n+1} &= (1-\alpha)^n Q_1 + \Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i \tag{25}\\
E[Q_{n+1}] &= E[(1-\alpha)^n Q_1 + \Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i] \tag{26}\\
&\text{Since } Q_n \text{is unbiased and hence } E[Q_{n+1}] \text{ is } q_* \tag{27}\\
q_* &= E[(1-\alpha)^n Q_1 + \Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i] \tag{28}\\
q_* &= (1-\alpha)^n E[Q_1] + E[\Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i] \tag{29}\\
q_* &= (1-\alpha)^n Q_1 + E[\Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i] \tag{30}\\
&= E[Q_1] = Q_1 \text{ as initial values are constant} \tag{31}\\
q_* &= (1-\alpha)^n Q_1 + q_*(1-(1-\alpha)^n)\text{from part b} \tag{32}\\
q_* &= (1-\alpha)^n Q_1 + q_* - q_*(1-\alpha)^n \tag{33}\\
q_* &= (1-\alpha)^n (Q_1 - q_*) + q_* \tag{34}\\
0 &= (1-\alpha)^n (Q_1 - q_*) \tag{35}\\
(Q_1 - q_*) &= 0 \tag{36}\\
Q_1 &= q_* \tag{37}\\
& \tag{38}
\end{aligned}$$

Hence Q_n will be unbiased when Q_1or the initial values are equal to q_* which are the true estimates.

(d)

$$Q_{n+1} = (1-\alpha)^n Q_1 + \Sigma_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i \tag{39}$$

$$E[Q_{n+1}] = E[(1-\alpha)^n Q_1 + \Sigma_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i] \tag{40}$$

$$= \text{for n} \Rightarrow \infty \tag{41}$$

$$lim_{n\to\infty} E[Q_{n+1}] = lim_{n\to\infty} E[(1-\alpha)^n Q_1 + \Sigma_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i] \tag{42}$$

$$= lim_{n\to\infty}[(1-\alpha)^n Q_1 + E[\Sigma_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i]] \tag{43}$$

$$= lim_{n\to\infty}(1-\alpha)^n Q_1 + lim_{n\to\infty} E[\Sigma_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i]] \tag{44}$$

$$= lim_{n\to\infty} E[\Sigma_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i]] \tag{45}$$

$$\text{as n} \Rightarrow \infty, (1-\alpha) \Rightarrow 0 \tag{46}$$

$$= lim_{n\to\infty} E[\alpha(1-\alpha)^{n-1} R_1 + \alpha(1-\alpha)^{n-1} R_2 + ... + \alpha(1-\alpha)^{n-(n-1)} R_{n-1} + \alpha(1-\alpha)^0 R_n] \tag{47}$$

$$= lim_{n\to\infty}[\alpha E[(1-\alpha)^{n-1} R_1 + (1-\alpha)^{n-1} R_2 + ... + (1-\alpha)^{n-(n-1)} R_{n-1} + (1-\alpha)^0 R_n]] \tag{48}$$

$$= lim_{n\to\infty}[\alpha[(1-\alpha)^{n-1} E[R_1] + (1-\alpha)^{n-1} E[R_2] + ... + (1-\alpha)^{n-(n-1)} E[R_{n-1}] + (1-\alpha)^0 E[R_n]]] \tag{49}$$

$$= lim_{n\to\infty}[\alpha[(1-\alpha)^{n-1} q_* + (1-\alpha)^{n-1} q_* + ... + (1-\alpha)^{n-(n-1)} q_* + (1-\alpha)^0 q_*]] \tag{50}$$

$$= lim_{n\to\infty}[\alpha q_*[(1-\alpha)^{n-1} + (1-\alpha)^{n-1} + ... + (1-\alpha)^{n-(n-1)} + (1-\alpha)^0]] \tag{51}$$

$$= \alpha q_*[\frac{1}{1-(1-\alpha)}] \tag{52}$$

$$\text{as } lim_{n\to\infty} \Sigma_{i=0}^{n} a.r^n = \frac{a}{1-r} \tag{53}$$

$$= q_* \tag{54}$$

Thus for $lim_{n\to\infty}$, Q_n is an unbiased estimator.

(e)

$$
\begin{aligned}
Q_{n+1} &= (1-\alpha)^n Q_1 + \Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i && (55)\\
E[Q_{n+1}] &= E[(1-\alpha)^n Q_1 + \Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i] && (56)\\
&= E[(1-\alpha)^n Q_1] + E[\Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i] && (57)\\
&= (1-\alpha)^n E[Q_1] + E[\Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i] && (58)\\
&= (1-\alpha)^n Q_1 + E[\Sigma_{i=1}^{n} \alpha (1-\alpha)^{n-i} R_i] && (59)\\
&= (1-\alpha)^n Q_1 + q_* [1 - (1-\alpha)^n] && (60)\\
&\text{using results from part b} && (61)\\
&= (1-\alpha)^n Q_1 + q_* - q_* [(1-\alpha)^n] && (62)\\
&= (1-\alpha)^n (Q_1 - q_*) + q_* && (63)\\
& && (64)
\end{aligned}
$$

This shows that $Q_{n+1} \neq q_*$ and hence the exponential recency-weighted average is biased more often than not.

Question 4. *Implement the 10-armed testbed described in the first paragraph Section 2.3 (p. 28).*

Response:

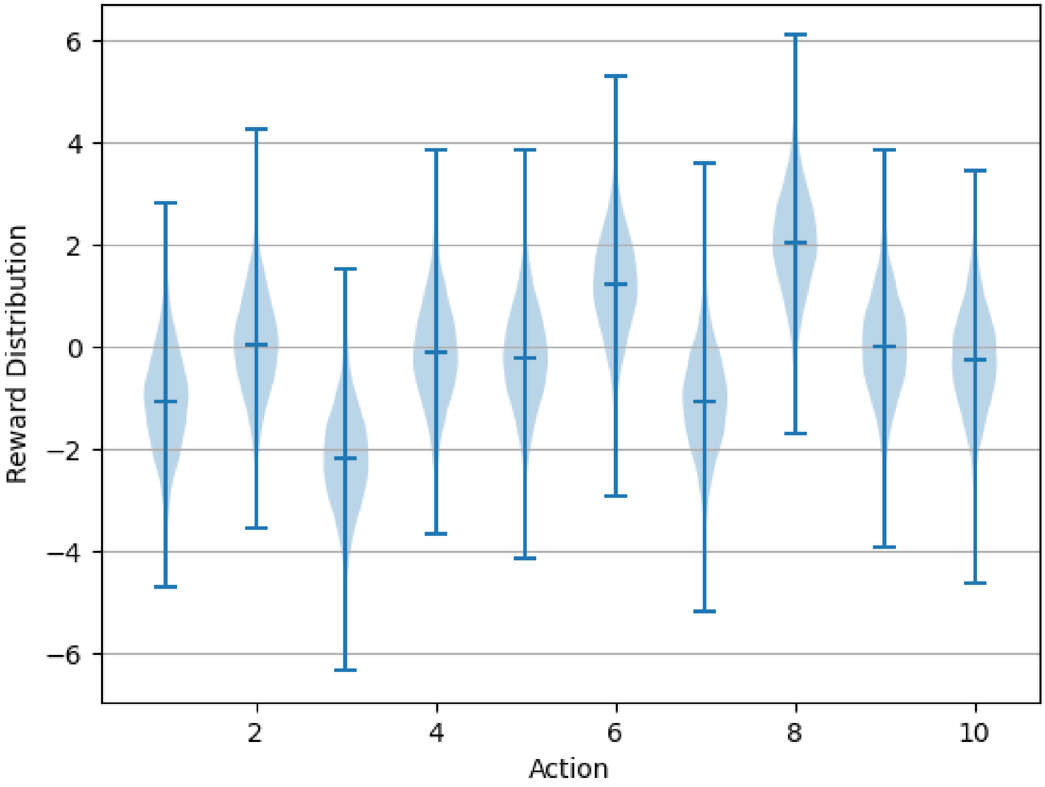

Figure 1: Q4 : Reward distribution for ten bandit arms with mean derived from a gaussian distribution

Question 5. *Predicting asymptotic behavior in Figure 2.2.*
In the comparison shown in Figure 2.2, which method will perform best in the long run in terms of cumulative reward and probability of selecting the best action? How much better will it be? Express your answer quantitatively. (Compute what you expect the asymptotic performances to be for the lower graph, and possibly the upper graph if you want a small mathematical workout.)

Response:

For any ϵ case, we may expect the action to be greedy for $1 - \epsilon$ times out of total times, also additionally when for the ϵ times, it is taking random action, the probability of choosing the greedy action is 0.1 as this is 10 arm bandit. Hence, for ϵ **= 0**, the probability of choosing the greedy action is 1, hence it will select the greedy action 100% of times.
For ϵ **= 0.1**, it will choose the best action 0.9 + 0.1*0.1 = 0.91 i.e. **91%** of times.
For ϵ **= 0.01**, it will choose the best action 0.99 + 0.01*0.1 = 0.991 i.e. **99.1%** of times.
Now, using these value we can see that ϵ = 0.01 will be best, the greedy one although has 100% chances of selecting the best action, due to lack of exploration, it will be stuck in the greedy action which is dependent on the initial set values. The ϵ **= 0.01** might have less reward than the ϵ **= 0.1** case in the beginning but after some time when it has explored enough, it will take the optimal action more times than the ϵ **= 0.1** case and hence will return higher cumulative reward.

Question 6. *Reproducing Figure 2.2. Code: Implement the ϵ-greedy algorithm with incremental updates. Note that in the graph: "All the methods formed their action-value estimates using the sample-average technique (with an initial estimate of 0)." Plot: Reproduce the curves shown in Figure 2.2*

Response:

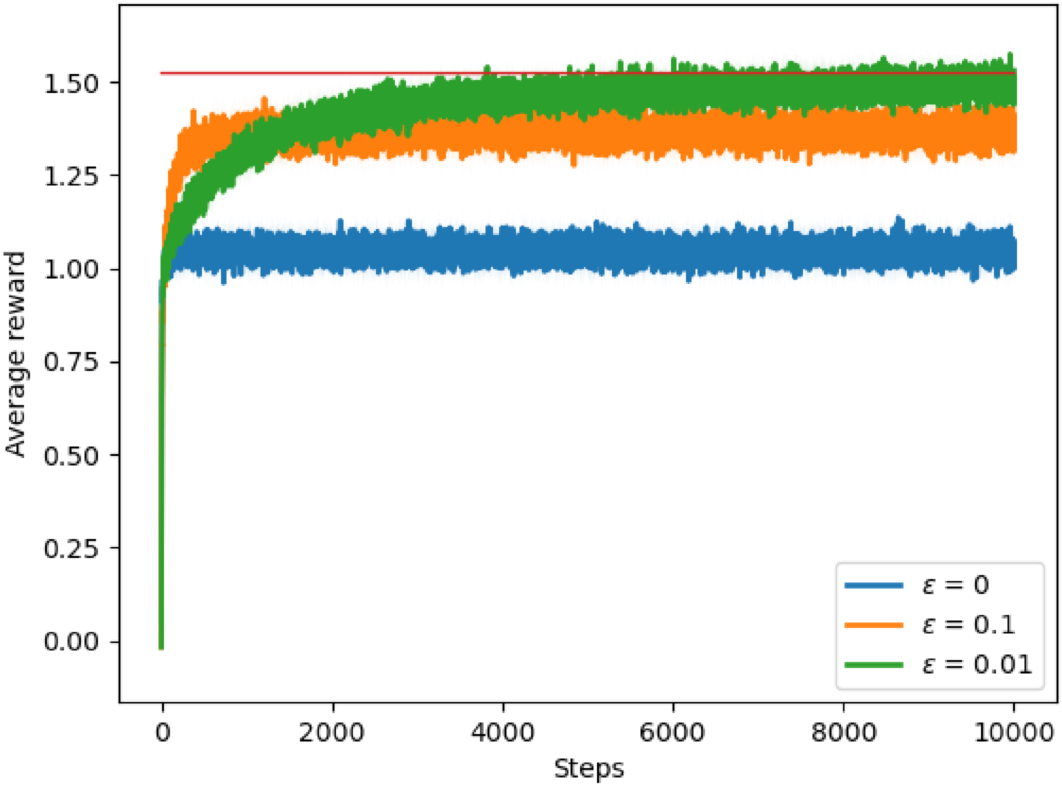

Figure 2: Q6 : Average reward plot for different value of ϵ

Both the curves in Figure 2 and Figure 3 reach the asymptotic levels as we have stated in Q5. Also notice that the curve for $\epsilon = 0.01$ is overpassing curve with $\epsilon = 0.1$ in the long run as we had predicted.

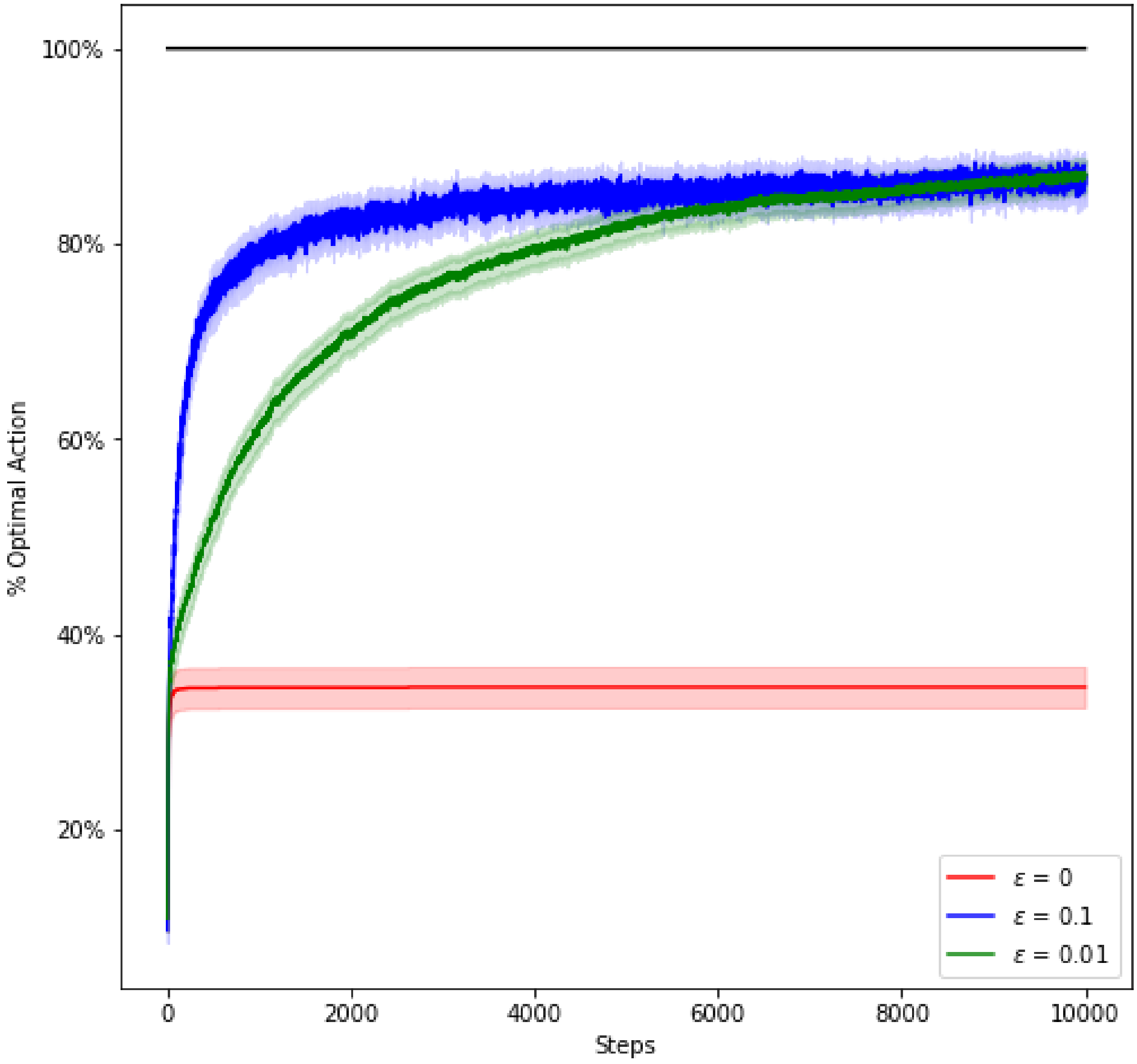

Figure 3: Q6 : Optimal action plot for different value of ϵ

Question 7. *Investigating nonstationary environments. Code: Design and conduct an experiment to demonstrate the difficulties that sample-average methods have for nonstationary problems. Use a modified version of the 10-armed testbed in which all the q(a) start out equal and then take independent random walks (by adding a normally distributed increment with mean 0 and standard deviation 0.01 to all the q(a) on each step).*

Response:

It can be seen from the figure 4 and figure 5 that the constant step size parameter supercedes

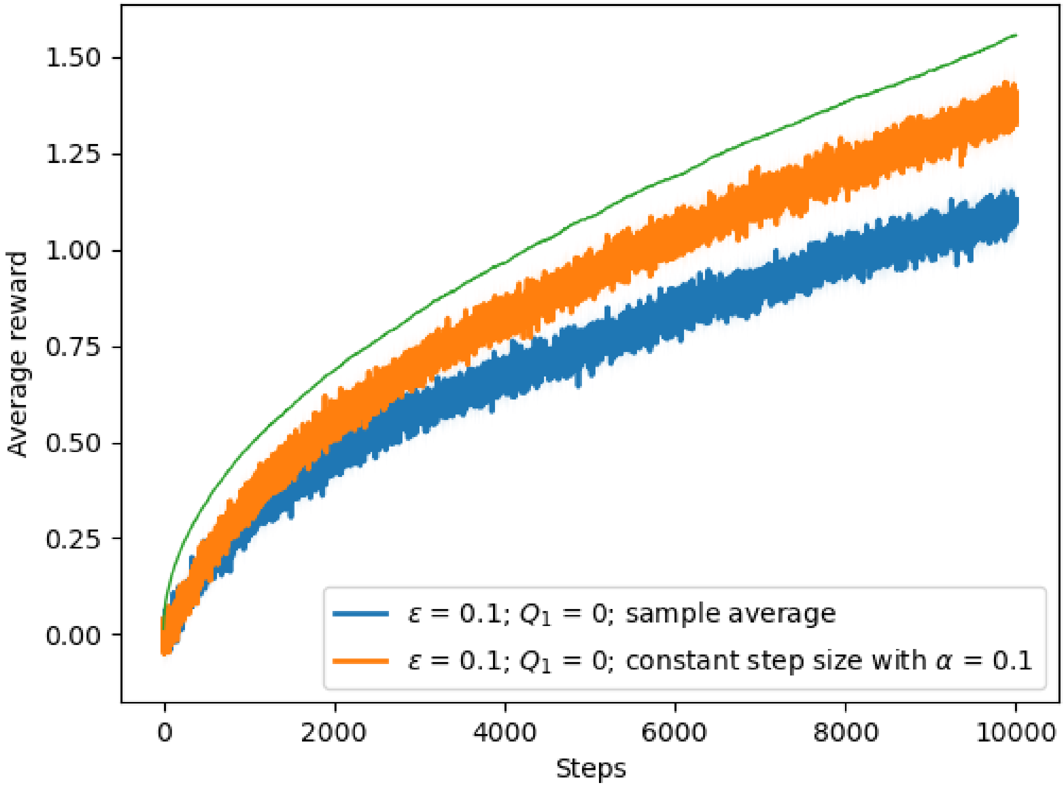

Figure 4: Q7 : Average reward plot where, we test the sample average and constant step size method with ϵ greedy

the sample average method in the long run.

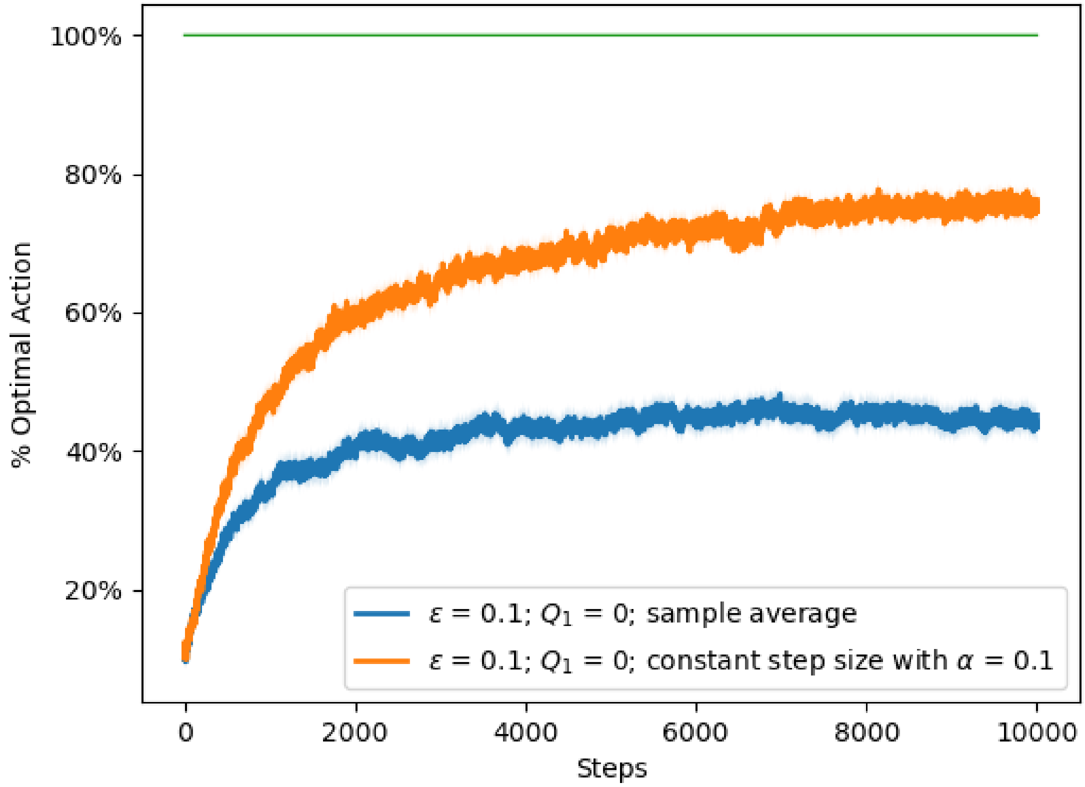

Figure 5: Q7 : Optimal action plot to check the performance of two methods on the number of times they take the optimal action, we test the sample average and constant step size method with ϵ greedy

Question 8. *Reproducing and supplementing Figures 2.3 and 2.4.*
Code: Implement the -greedy algorithm with optimistic initial values, and the bandit algorithm with UCB action selection.

It can be seen from the graph Figure 6 and figure 7that both the curves which are absolute greedy converge together while the other two curves with epsilon 0.1 converge together, this happens irrespective of the different initial values set and thus it can be concluded that in the long run the importance of the initial state reduces and hence the bias reduces. Thus the curves with same epsilon values converge. It can be seen from the above graphs figure 8 and figure 9 that comparing "UCB" and "epsilon greedy with optimistic initial values" that, UCB supercedes the curve with greedy epsilon method in the long run. This happens because by formulation UCB tries to explore more by checking on states that are visited less and it maintains a balance

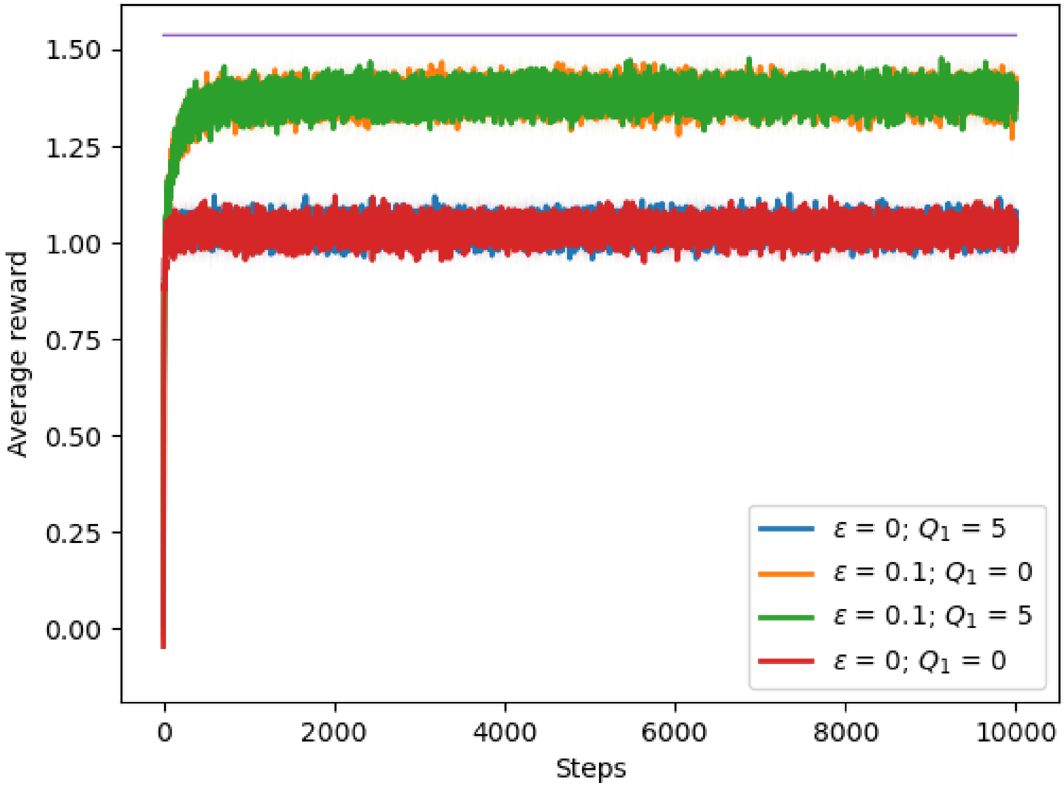

Figure 6: Q8 : Average reward plot for different initial Q and ϵ values

between exploration and exploitation.

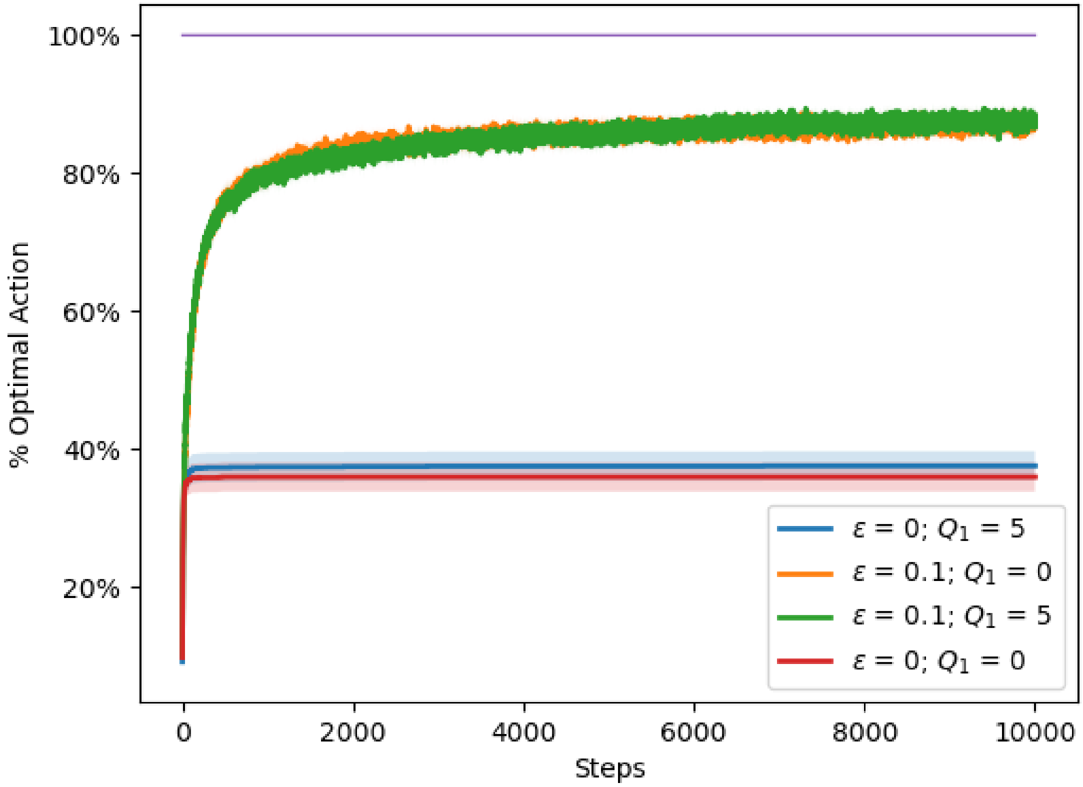

Figure 7: Q8 : Optimal action plot for different initial Q and ϵ values

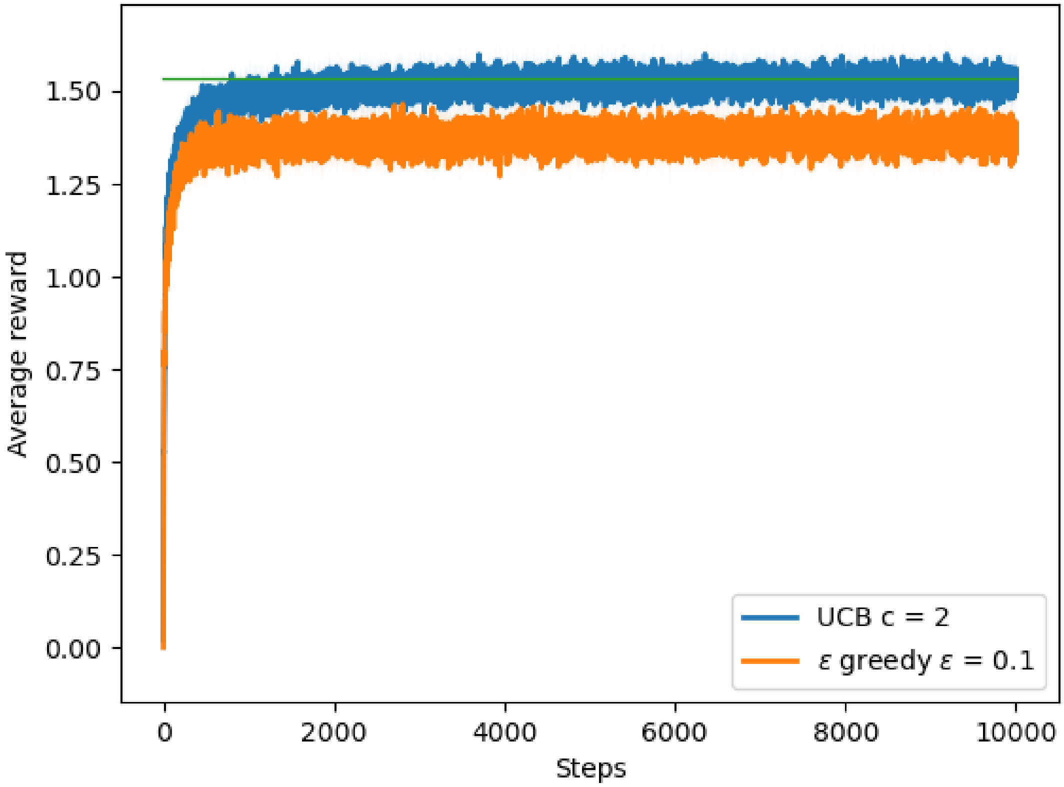

Figure 8: Q8 : Average reward plot for comparing UCB and ϵ greedy

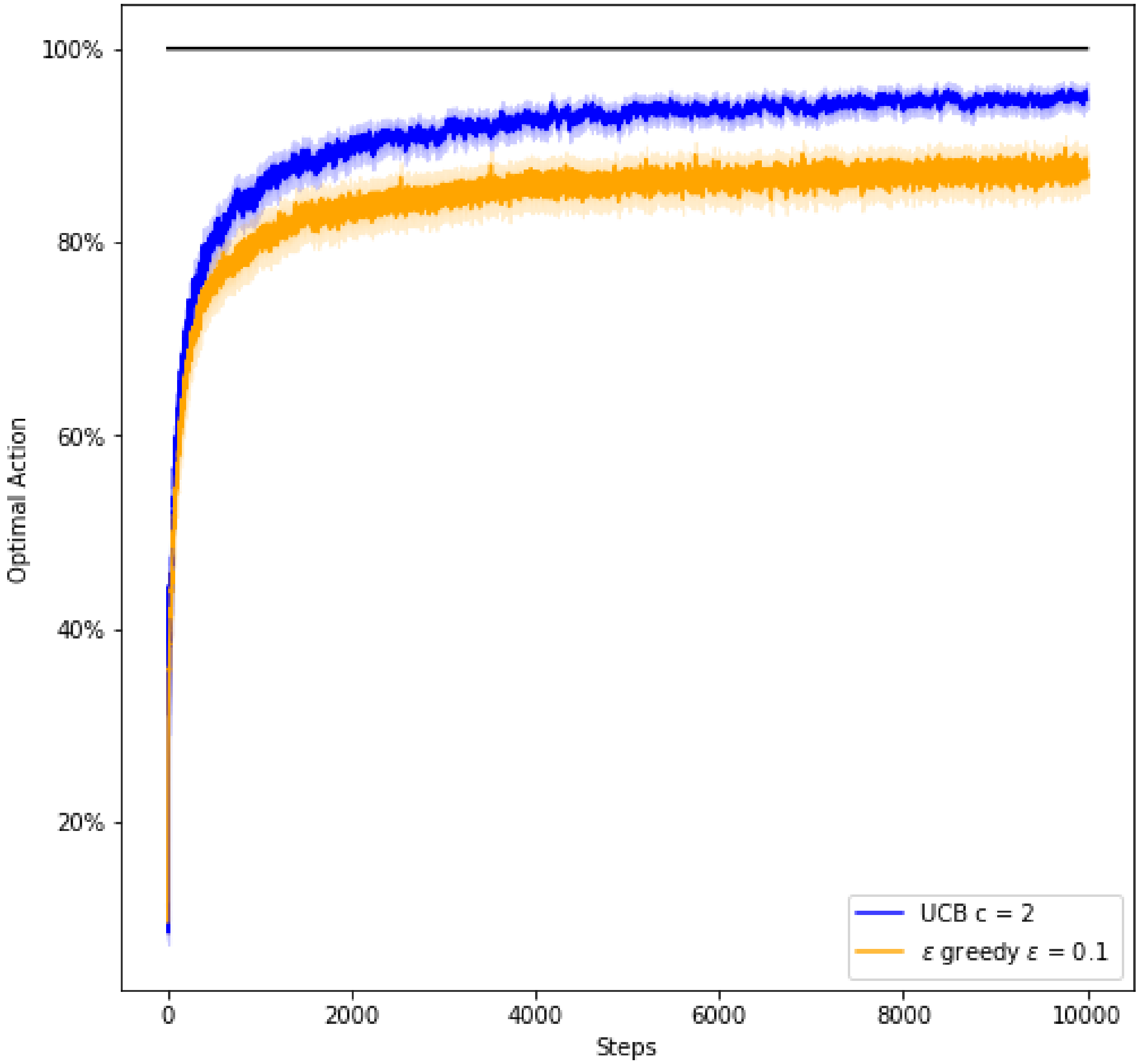

Figure 9: Q8 : Optimal action plot for comparing UCB and ϵ greedy