Exercise 1: Multi-armed Bandits

Prof: Robert Platt

Date: September 24th, 2021

Name: Guanang Su

Question 1.

1 point. (RL2e 2.2) *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?

A4 and A5 were definitely occur, others (A1, A2 and A3) could possibly have occurred.

i	1	2	3	4	5
A_i	1	2	2	2	3
R_i	-1	1	-2	2	0

$Q(A = 2) = \frac{1-2}{2} = -0.5$

For step 4, $Q3 = Q4 > Q2 > Q1$, so epsilon would definitely occur.

And for step 5, $Q2 > Q3 = Q4 > Q1$, so epsilon would definitely occur.

Question 2.

1 point. (RL2e 2.4) *Varying step-size weights.*
Written: If the step-size parameters, α_n, 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?

$$\begin{aligned}
Q_{n+1} &= Q_n + \alpha\Big[R_n - Q_n\Big] \\
&= \alpha R_n + (1-\alpha)Q_n \\
&= \alpha R_n + (1-\alpha)\left[\alpha R_{n-1} + (1-\alpha)Q_{n-1}\right] \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 Q_{n-1} \\
&= \alpha R_n + (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2 \alpha R_{n-2} + \\
&\qquad \cdots + (1-\alpha)^{n-1}\alpha R_1 + (1-\alpha)^n Q_1 \\
&= (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i.
\end{aligned}$$

$$\begin{aligned}
Q_{n+1} &= Q_n + \alpha_n[R_n - Q_n] \\
&= \alpha_n R_n + (1-\alpha_n) Q_n \\
&= \alpha_n R_n + (1-\alpha_n)(\alpha_{n-1} R_{n-1} + (1-\alpha_{n-1})(\alpha_{n-2} R_{n-2} + (1-\alpha_{n-2}) Q_2)) \\
&= Q_1 \prod_{i=1}^{n} (1-\alpha_i) + \sum_{i=1}^{n} \alpha_i R_i \prod_{j=i+1}^{n} (1-\alpha_j)
\end{aligned}$$

Question 3.

[CS 5180 only.] 2 points. *Bias in Q-value estimates.*
Written: Recall that $Q_n \triangleq \frac{R_1+\ldots+R_{n-1}}{n-1}$ is an estimate of the true expected reward q_* of an arbitrary arm a. We say that an estimate is *biased* if the expected value of the estimate does not match the true value, i.e., $\mathbb{E}[Q_n] \neq q_*$ (otherwise, it is *unbiased*).

(a) Consider the *sample-average* estimate in Equation 2.1. Is it biased or unbiased? Explain briefly.

$$Q_t(a) \doteq \frac{\text{sum of rewards when } a \text{ taken prior to } t}{\text{number of times } a \text{ taken prior to } t} = \frac{\sum_{i=1}^{t-1} R_i \cdot \mathbb{1}_{A_i=a}}{\sum_{i=1}^{t-1} \mathbb{1}_{A_i=a}}$$

In sample-average estimate, sample mean is μ

$$E(Q_n) = \frac{E(R_1)+E(R_2)+\cdots+E(R_{n-1})}{n-1} = \frac{(n-1)q^*}{n-1} = q^*$$

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).

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

(b) If $Q_1 = 0$, is Q_n for $n > 1$ biased? Explain briefly.

It is unbiased.

(c) Derive conditions for when Q_n will be unbiased.

From the equation above, we can see that if the n is small enough like 1, there is no bias since only one sample. Also, if the data is infinite number, the biased would be small enough to regard as 0 since Q1=0, $\sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} = 1$.

(d) Show that Q_n is *asymptotically unbiased*, i.e., it is an unbiased estimator as $n \to \infty$.

When n is infinite, $(1-\alpha)^n Q_i$ will decrease to 0,

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

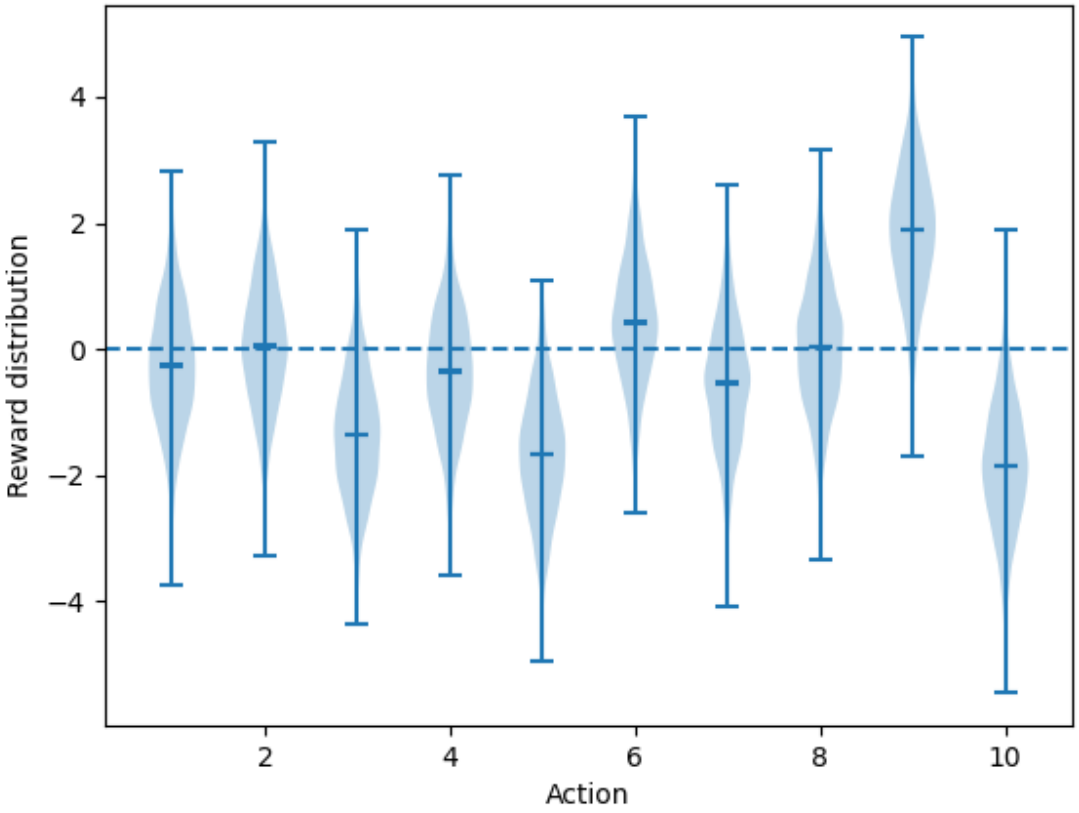
So Qn is asymptotically unbiased.

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

For a non-stationary problem, rewards close to the current state means a closer to the expected result, which help us to find better estimates for current distribution, rather than previous steps.

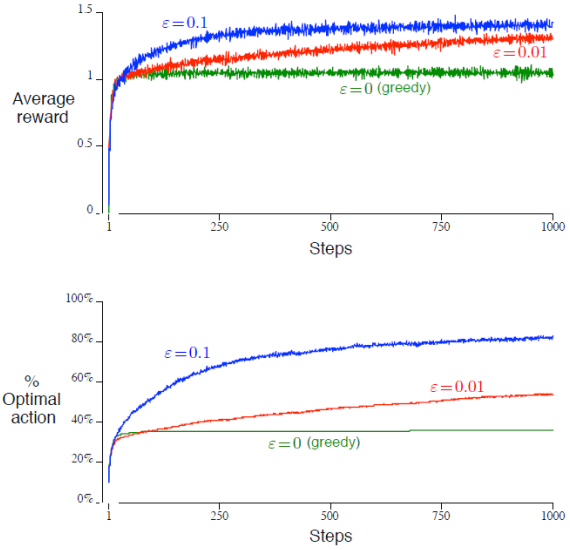
Question 4.

1 point. *Implementing the 10-armed testbed for further experimentation in the remainder of the assignment.*
Code: Implement the 10-armed testbed described in the first paragraph Section 2.3 (p. 28).
Read the description carefully.
Plot: To test that your testbed is working properly, produce a plot similar in style to Figure 2.1 by pulling each arm many times and plotting the distribution of sampled rewards. You can use any type of plot that makes this point effective, e.g., a violin plot, or a scatterplot with some jitter in the horizontal axis to show the sample density more effectively.



Question 5.

1 point. (RL2e 2.3) *Predicting asymptotic behavior in Figure 2.2.*
Written: 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.)



when step size $\to \infty$

for $\varepsilon=0$: the agent will run the action with first positive reward happen, so it will always around 1 for the average reward.

for $\varepsilon=0.01$: the optimal percentage is $= 0.01 \times \frac{1}{10} + (1-0.01) \times 1$
$= 99.1\%$

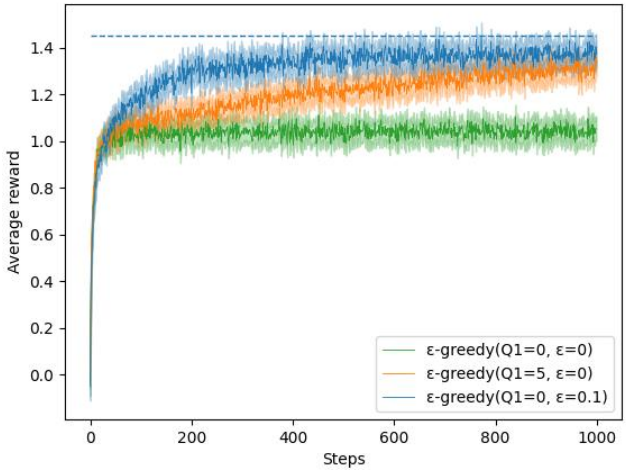
for $\varepsilon=0.1$: the optimal percentage is $= 0.1 \times \frac{1}{10} + (1-0.1) \times 1$
$= 91\%$

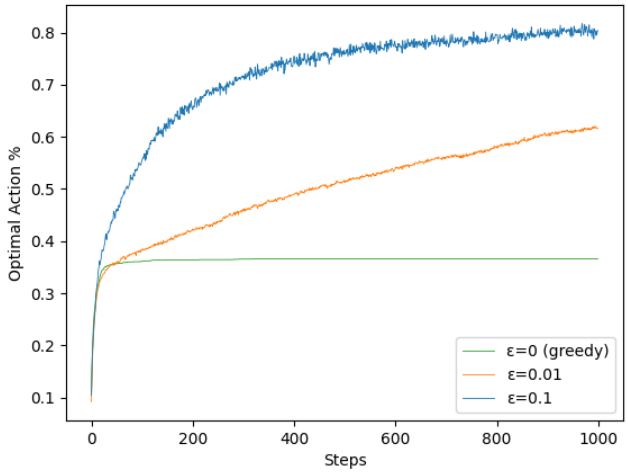
$99.1\% > 91\%$, so in the long run, $\varepsilon=0.01$ will probably selecting the best action.

Question 6.

2 points. *Reproducing Figure 2.2.*

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

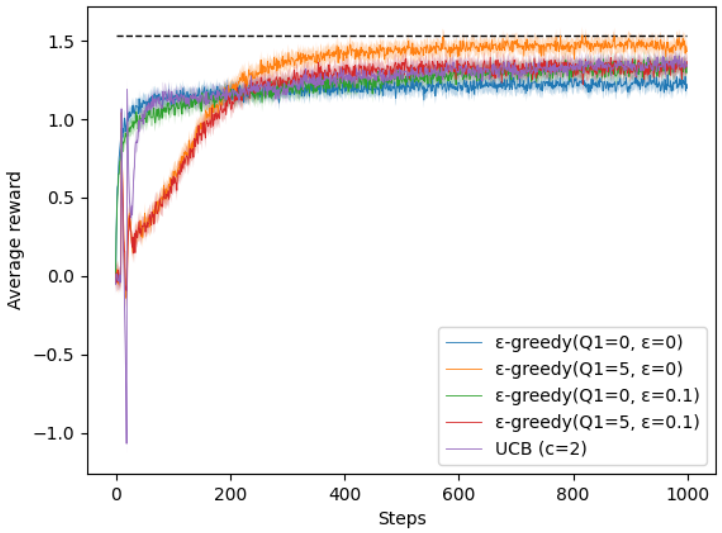


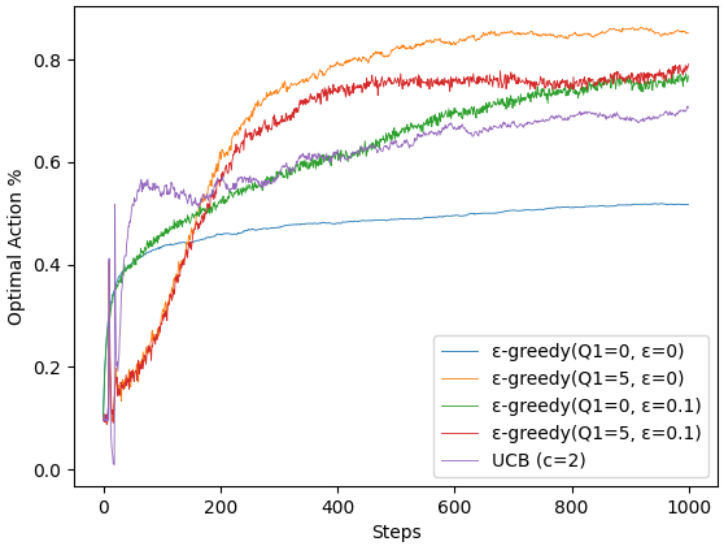


The average will not reach the asymptotic levels predicted, since it is random pick from steps, the upper bound, each greedy will explore in a range below or slightly higher than the bound value. Therefore, the average reward could not reach the asymptotic for the value we picked.

Question 7.

[CS 5180 only.] 2 points. *Reproducing and supplementing Figures 2.3 and 2.4.*
Code: Implement the ε-greedy algorithm with optimistic initial values, and the bandit algorithm with UCB action selection.





Written: 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.

The sharp increase is happened to since the beginning of UCB algorithm, it tries with all arms, which is wildly optimistic to encourage the action-value to explore. After the exploration, the arm will pick the highest mean value and randomly run one of the index with highest value, the reward could have a drop.