Reinforcement Learning - Project 1

Fridolin Paiki

February 5, 2024

Part a

Given:

$$Q(a^1) = Q(a^2) = 0$$

α	1	0.9^k	$\frac{1}{1+\ln(1+k)}$	$\frac{1}{k}$
ϵ	0	0.1	0.2	0.5

Table 1: Values of α and ϵ

In this part, we're going to calculate the average cumulative average of rewards using epsilon-greedy approach with initial action values equal to zero ($Q(a^1) = Q(a^2) = 0$) for several combinations of learning rate and epsilon as mentioned in Table 1.

For $\alpha = 1$, as we can see from the graph at Figure 1 and Table 2, the random approach provide a better result in terms of accumulated average of reward which is at $\epsilon = 0.5$. However at the beginning, the generated result from epsilon 0.5 has much bigger fluctuations due to both exploration and exploitation has the same weight. Eventually, this result normalized as more steps taken. Nevertheless, all combination of epsilon values generate the average accumulated rewards that somehow converge in the end as the number of step increases (more than 1000).

From all combination of alpha and epsilon, the random approach (bigger epsilon) provide better results due to more exploration. Also, the smaller the alpha values or the learning rates, it seems to have bigger fluctuations at the earlier time/steps for bigger epsilon values. We can compare the Figure 1, 2, 3, 4 to see the difference, especially when the time/steps is less than 200.

Based on overall performance from all combinations of alpha and epsilon, the best one is when $\alpha = \frac{1}{k}$ or $\alpha = \frac{1}{1+ln(1+k)}$ and $\epsilon = 0.1$, since it provide better balance between exploration and exploitation that provide better results.

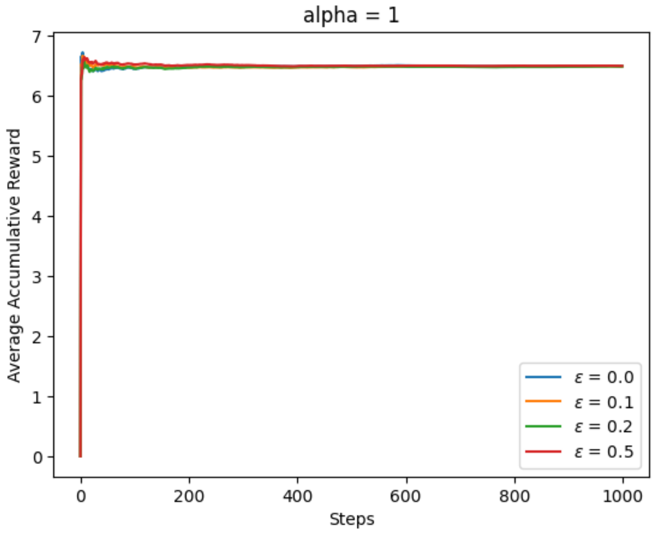

Figure 1: The Average Accumulated Reward for $\alpha = 1$

Epsilon-greedy	Average of action value $Q(a^1)$ of 100 runs	True action value $Q^*(a^1)$	Average of action value $Q(a^2)$ of 100 runs	True action value $Q^*(a^2)$
$\epsilon = 0$ (greedy)	2.5037	5	2.9698	7
$\epsilon = 0.1$	3.5471	5	4.3959	7
$\epsilon = 0.2$	4.1293	5	5.1022	7
$\epsilon = 0.5$ (random)	4.4544	5	5.5029	7

Table 2: The Average Action-Values $Q(a^1)$ and $Q(a^2)$ for $\alpha = 1$

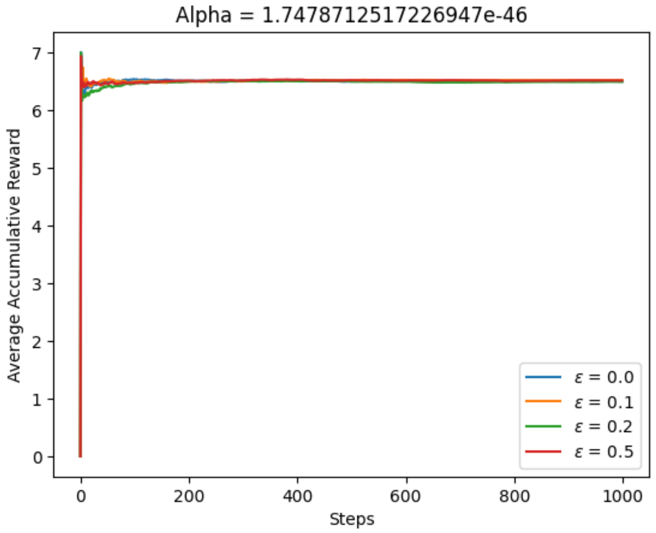

Figure 2: Average Accumulated Reward for $\alpha = 0.9^k$

Epsilon-greedy	Average of action value $Q(a^1)$ of 100 runs	True action value $Q^*(a^1)$	Average of action value $Q(a^2)$ of 100 runs	True action value $Q^*(a^2)$
$\epsilon = 0$ (greedy)	4.5366	5	4.4453	7
$\epsilon = 0.1$	4.4027	5	4.4067	7
$\epsilon = 0.2$	4.2875	5	4.4006	7
$\epsilon = 0.5$ (random)	4.2942	5	4.4478	7

Table 3: The Average Action Values $Q(a^1)$ and $Q(a^2)$ for $\alpha = 0.9^k$

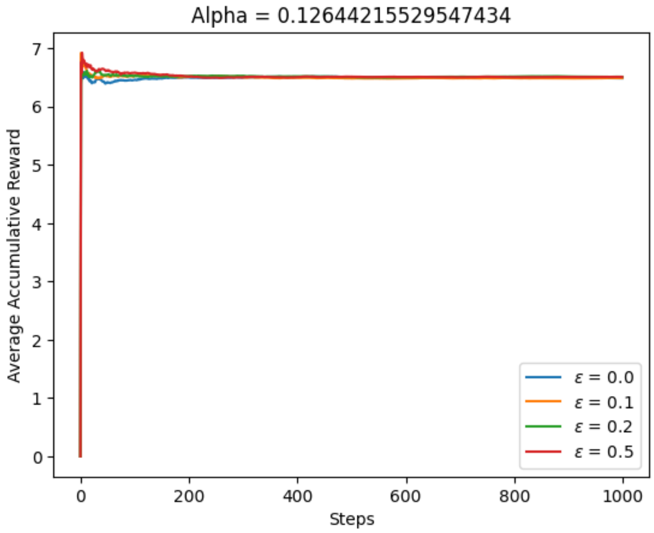

Figure 3: Average Accumulated Reward for $\alpha = \frac{1}{1+\ln{(1+k)}}$

Epsilon-greedy	Average of action value $Q(a^1)$ of 100 runs	True action value $Q^*(a^1)$	Average of action value $Q(a^2)$ of 100 runs	True action value $Q^*(a^2)$
$\epsilon = 0$ (greedy)	4.3437	5	4.4156	7
$\epsilon = 0.1$	4.6899	5	4.5323	7
$\epsilon = 0.2$	4.5877	5	4.4158	7
$\epsilon = 0.5$ (random)	4.3459	5	4.3630	7

Table 4: The Average Action Values $Q(a^1)$ and $Q(a^2)$ for $\alpha = \frac{1}{1+\ln(1+k)}$

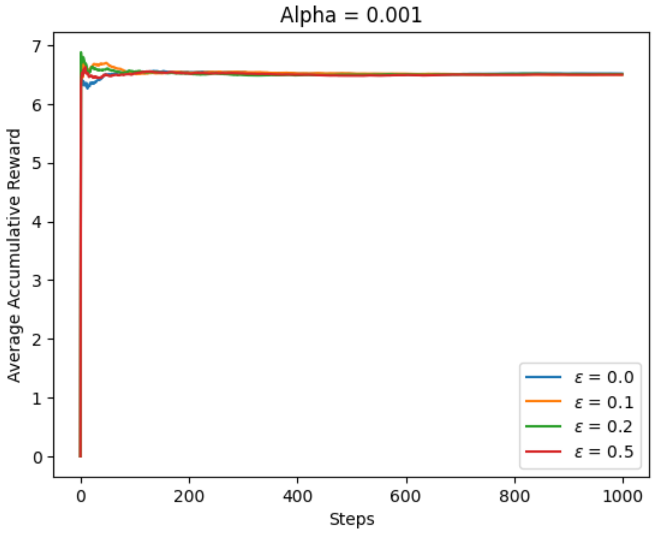

Figure 4: Average Accumulated Reward for $\alpha = \frac{1}{k}$

Epsilon-greedy	Average of action value $Q(a^1)$ of 100 runs	True action value $Q^*(a^1)$	Average of action value $Q(a^2)$ of 100 runs	True action value $Q^*(a^2)$
$\epsilon = 0$ (greedy)	4.4770	5	4.3730	7
$\epsilon = 0.1$	4.5753	5	4.5109	7
$\epsilon = 0.2$	4.1829	5	4.4082	7
$\epsilon = 0.5$ (random)	4.5281	5	4.3541	7

Table 5: The Average Action Values $Q(a^1)$ and $Q(a^2)$ for $\alpha = \frac{1}{k}$

Part b

For this one, instead of using initial action-values as zero, there will be combination of optimistic initial values, which are $[Q(a^1)\ Q(a^2)] = \{[0\ 0], [5\ 7], [20\ 20]\}$. Based on the result that has been generated, basically the $Q = [5\ 7]$ is the best combination which can be expected since they are exactly like the optimal Q from the reward distributions for each bandit, which is 5 and 7. It can also be seen in the graph if we focus more when the steps is below 200. Given the $\alpha = 0.1$ and $\epsilon = 0.1$, $Q = [5\ 7]$ provide better balance between exploration and exploitation.

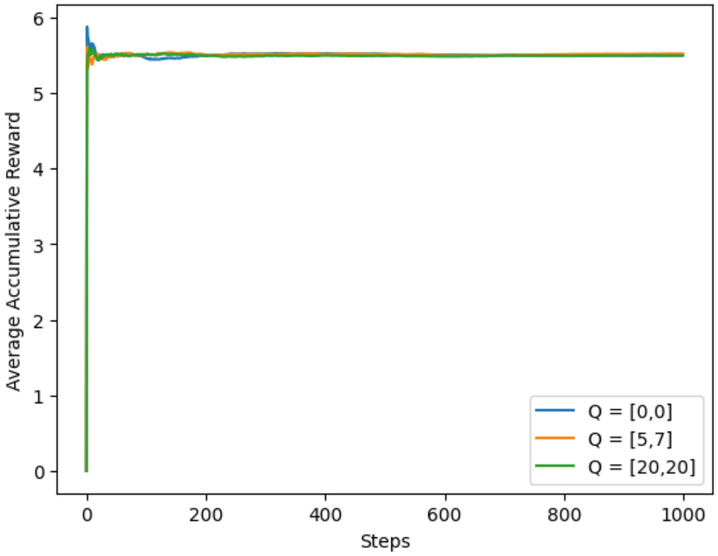

Figure 5: Average Accumulative Reward for Epsilon-Greedy Policy with Multiple Optimistic Initial Values

Epsilon-greedy	Average of action value $Q(a^1)$ of 100 runs	True action value $Q^*(a^1)$	Average of action value $Q(a^2)$ of 100 runs	True action value $Q^*(a^2)$
$Q = [0\ 0]$	4.9392	5	5.5360	7
$Q = [5\ 7]$	4.9849	5	5.7383	7
$Q = [20\ 20]$	5.5517	5	6.0254	7

Table 6: The Average Action Values $Q(a^1)$ and $Q(a^2)$ for $\alpha = \frac{1}{k}$

Part c

This part show the comparison of Epsilon-Greedy policy (with $\alpha = 0.1$, $\epsilon = 0.1$, and $Q(a^1) = Q(a^2) = 0$) to Gradient-Bandit policy (with $H(a^1) = H(a^1) = 0$) in 2-arm bandit case.As expected, the graph at Figure 6 clearly states that Gradient-Bandit has the better results in providing the average accumulative rewards than the Epsilon-Greedy policy. Also the average accumulated rewards in Gradient-Bandit is quickly converge and stabilized.

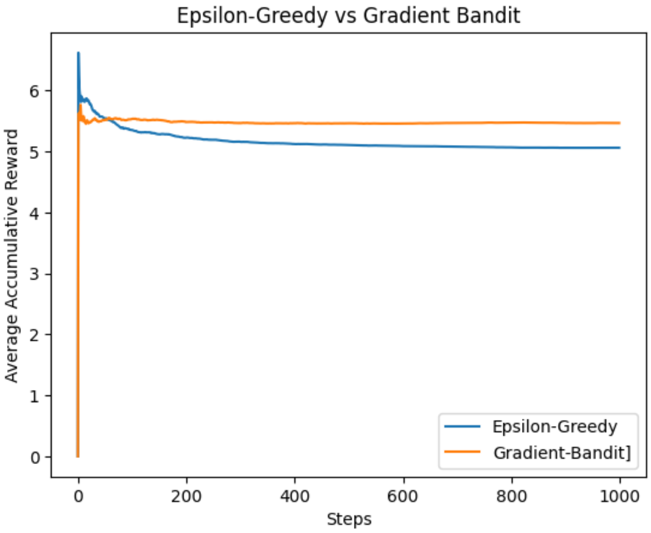

Figure 6: Average Accumulative Reward for Epsilon-Greedy Policy vs Gradient-Bandit Policy