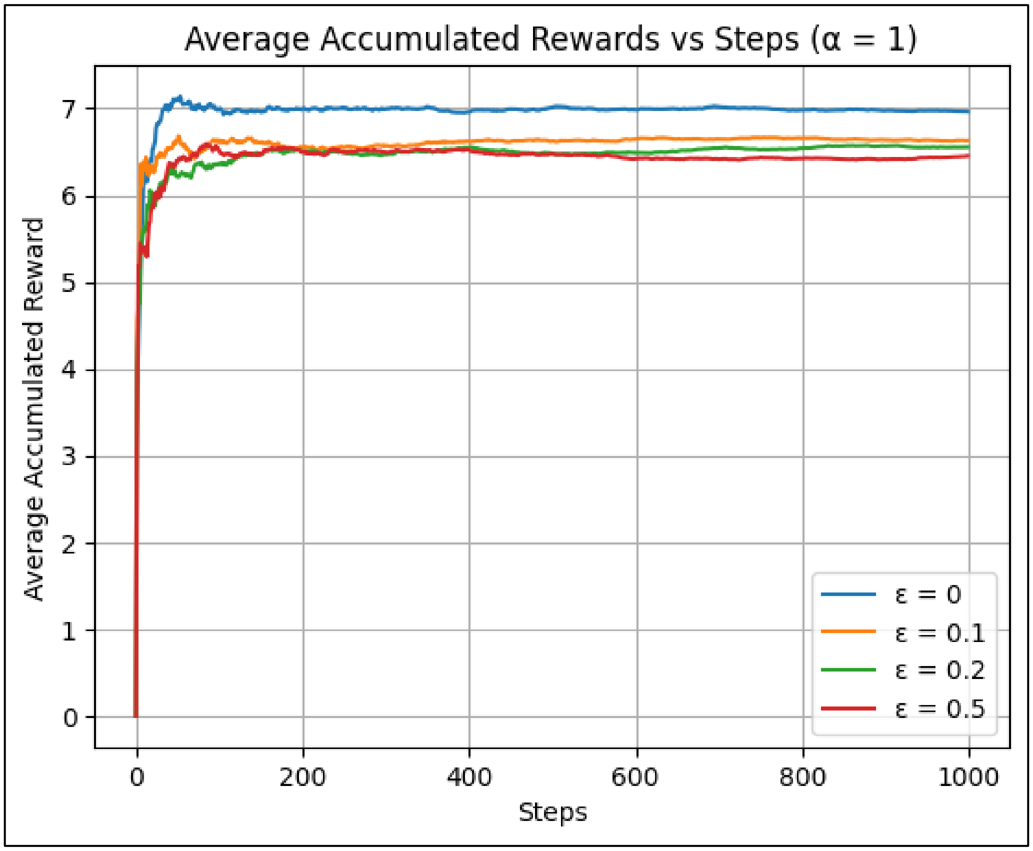
Project 1

Name: Ameya Padwad
NUID: 002284038

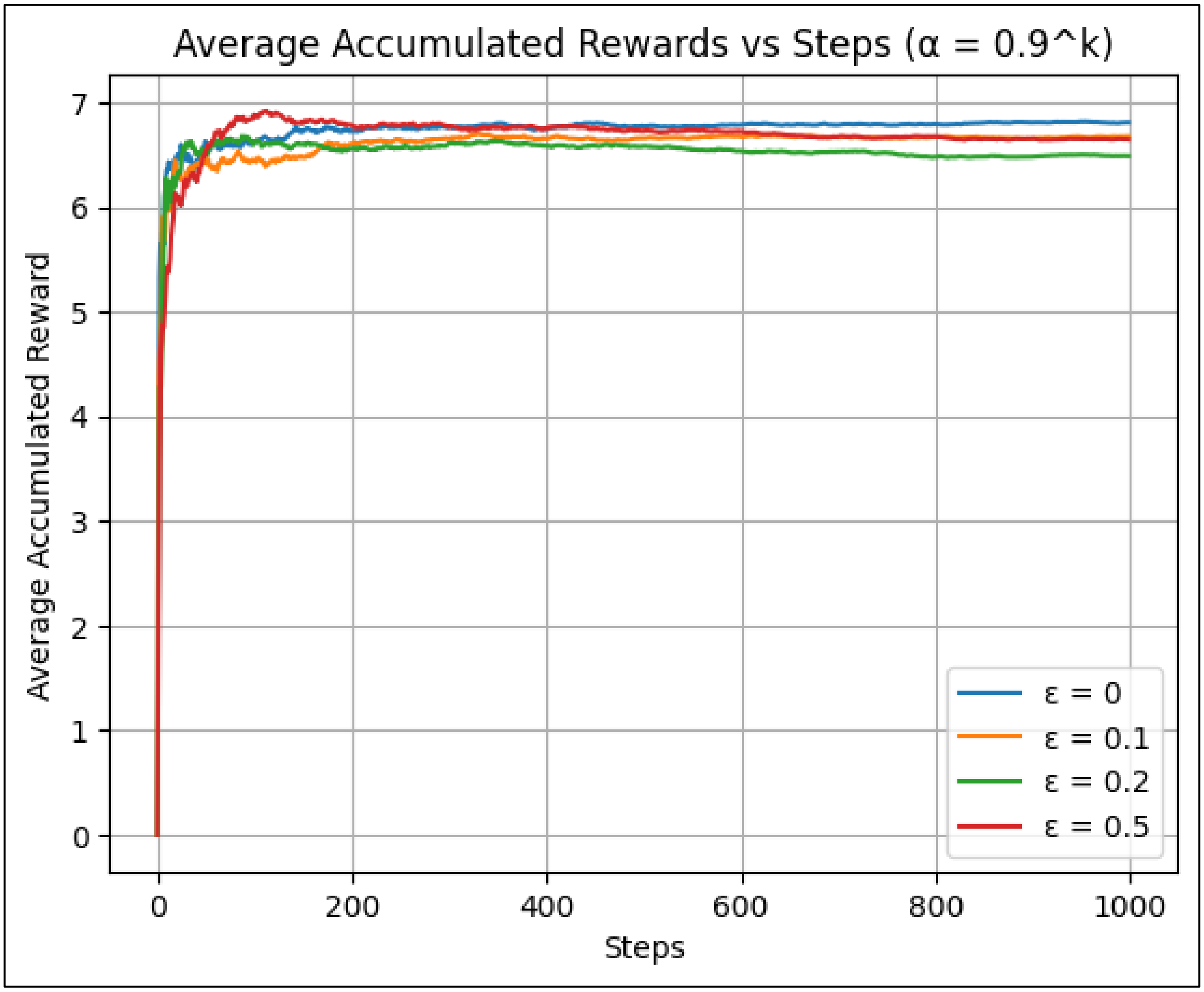
Part a)

α = 1



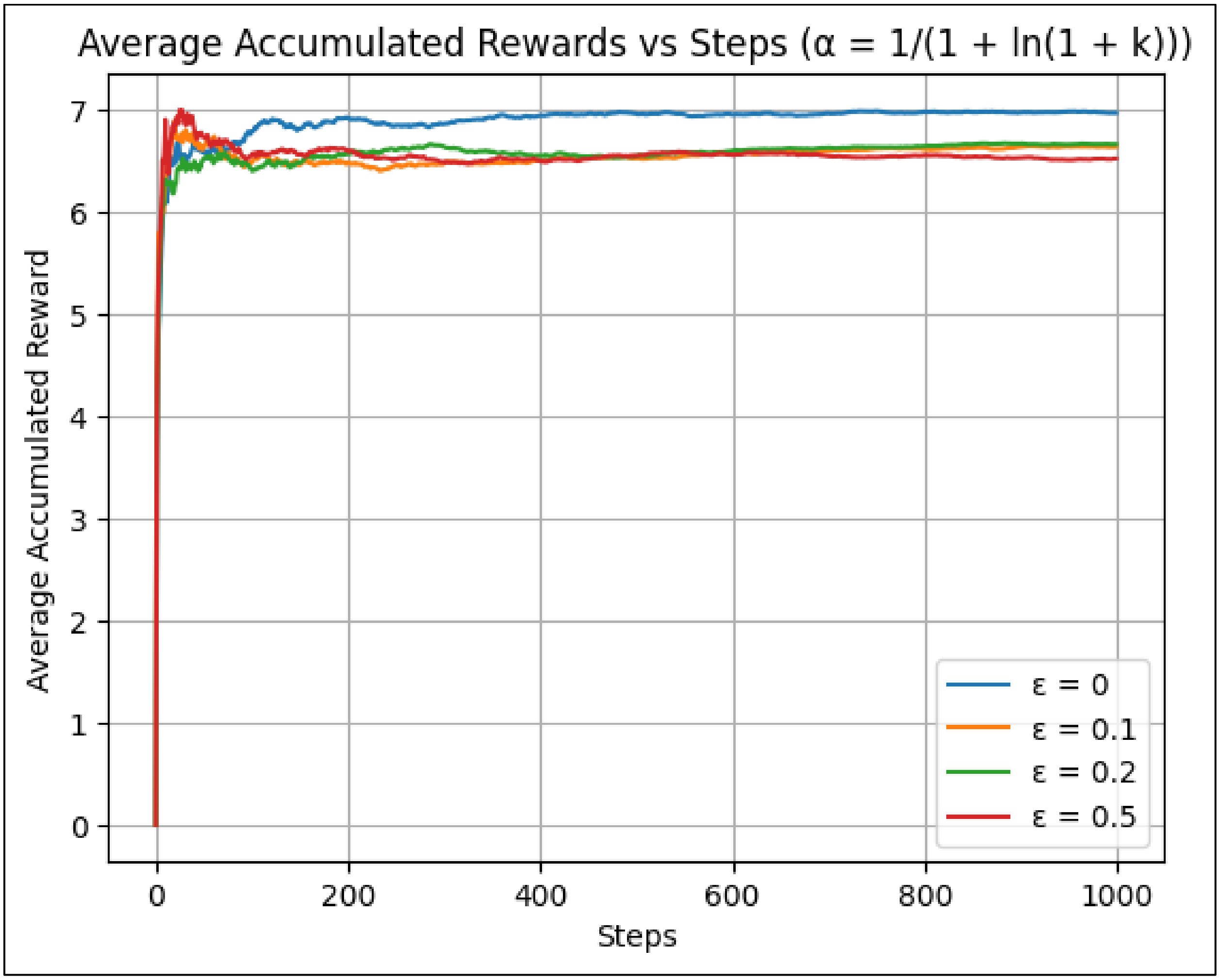
Epsilon-greedy	**Average of action value $Q(a^1)$ of 100 runs**	**True action value $Q(a^1)$**	**Average of action value $Q(a^2)$ of 100 runs**	**True action value $Q(a^2)$**
ε = 0 (greedy)	-57.59	6	4.23	7
ε = 0.1	-7.66	6	0.83	7
ε = 0.2	-6.72	6	2.88	7
ε = 0.5 (random)	-0.56	6	4.52	7

$\alpha = 0.9^k$



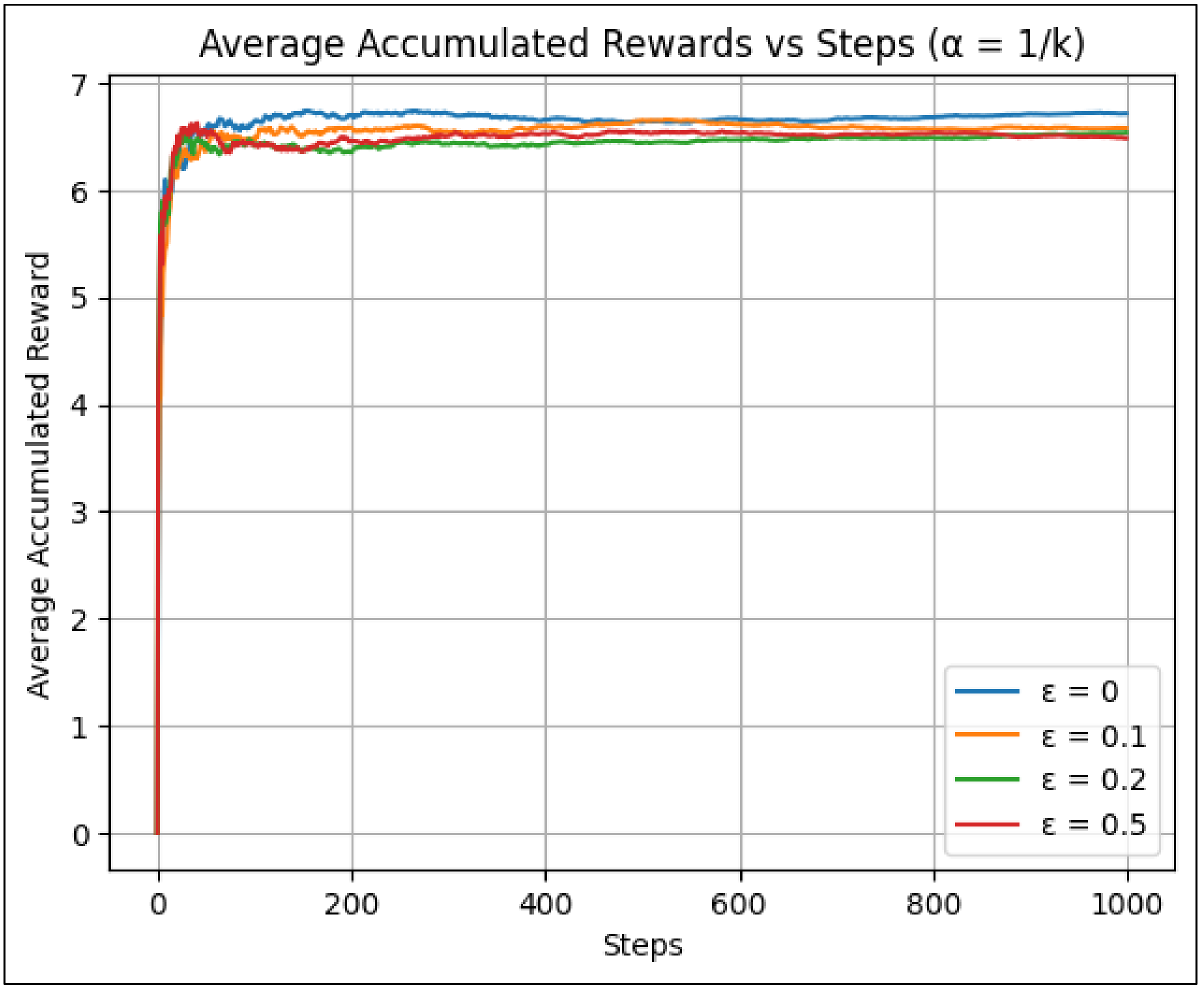
Epsilon-greedy	**Average of action value $Q(a^1)$ of 100 runs**	**True action value $Q(a^1)$**	**Average of action value $Q(a^2)$ of 100 runs**	**True action value $Q(a^2)$**
ε = 0 (greedy)	-17.91	6	0.07	7
ε = 0.1	-1.98	6	3.38	7
ε = 0.2	0.83	6	4.27	7
ε = 0.5 (random)	3.52	6	5.39	7

$\alpha = 1/(1 + \ln(1 + k))$



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 $Q(a^2)$**
ε = 0 (greedy)	-14.02	6	6.85	7
ε = 0.1	2.11	6	5.59	7
ε = 0.2	3.33	6	5.68	7
ε = 0.5 (random)	5.12	6	6.24	7

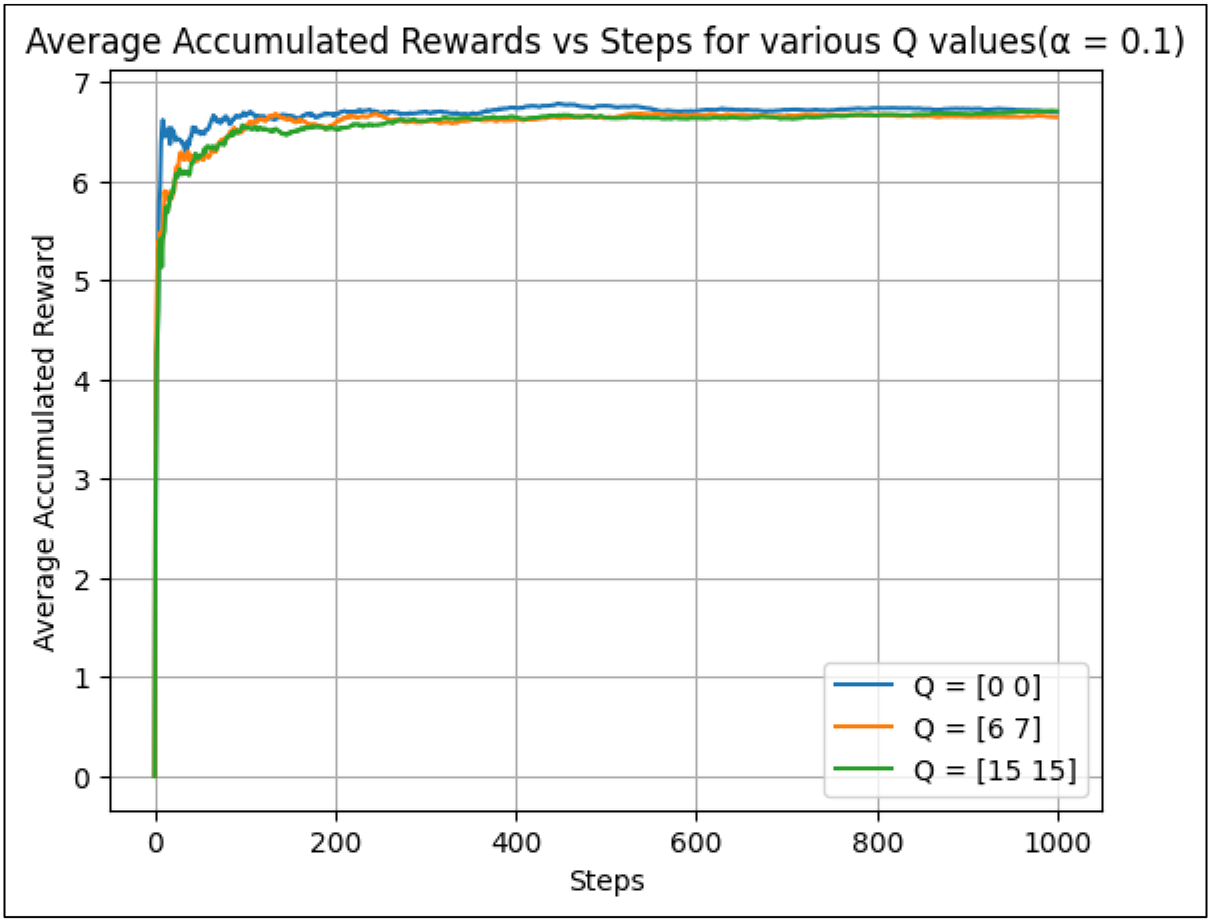
α = 1/k



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 $Q(a^2)$**
ε = 0 (greedy)	-17.15	6	1.71	7
ε = 0.1	0.48	6	3.27	7
ε = 0.2	3.38	6	4.79	7
ε = 0.5 (random)	5.34	6	5.98	7

For all the learning rates, the policy with ε = 0.5 performed the best in terms of getting the closest Q* value. In terms of the average accumulated reward, α = 1 and α = 1/(1 + ln(1 + k)) had the highest rewards. The best ε and α combination was that of α = 1/(1 + ln(1 + k)) and ε = 0.5.

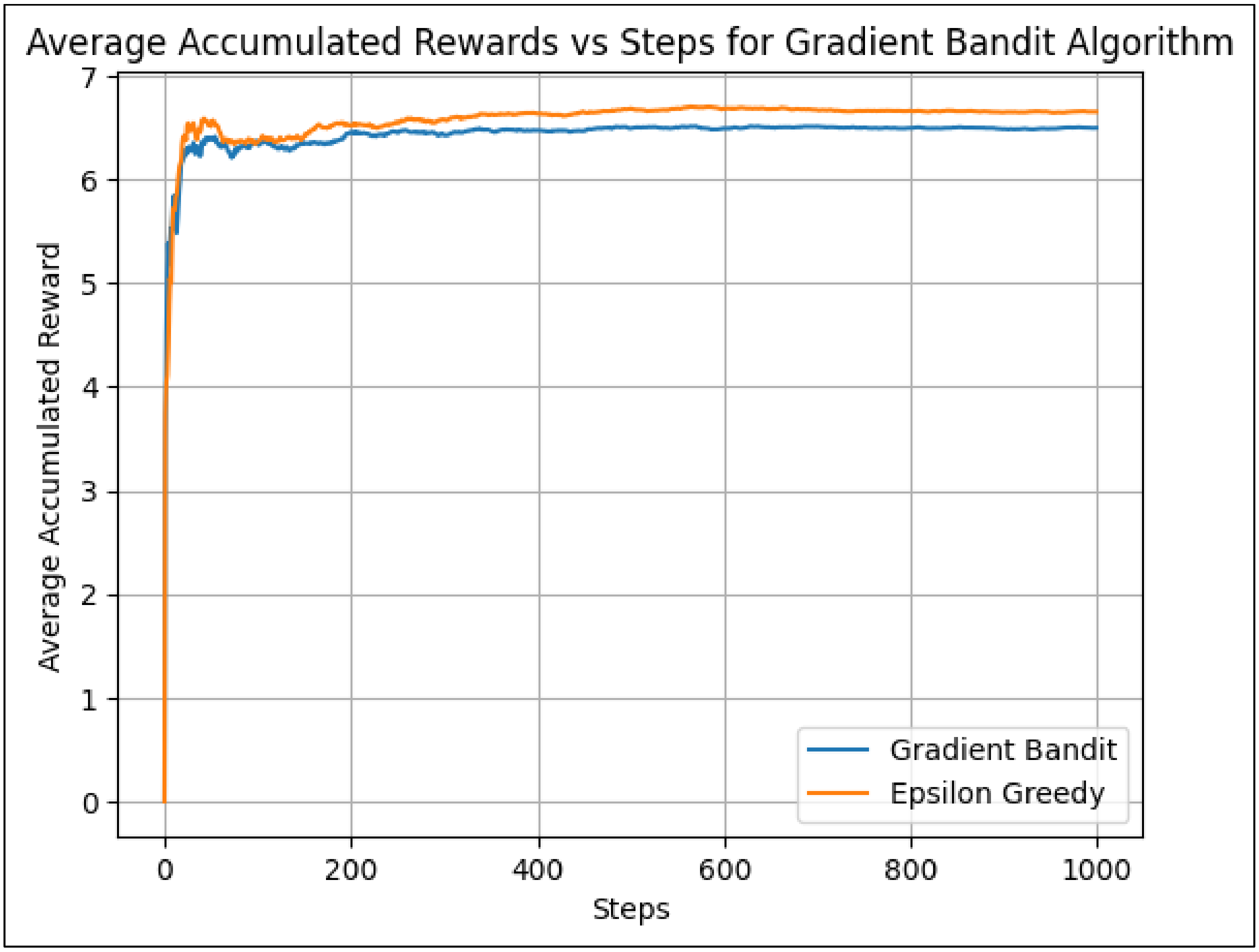
Part b)



Initial Q values	**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]	2.29	6	5.30	7
Q = [6 7]	3.15	6	4.99	7
Q = [15 15]	3.99	6	4.77	7

After 1000 steps, the accumulated average rewards seem to converge despite the different optimistic values for Q. On closer inspection, Q = [0, 0] and Q = [15, 15] seem to have a slight edge over Q = [6, 7]

Part c)



As can be seen from the plot above, epsilon greedy policy with Q = [0, 0], ε = 0.1 and α = 0.1 performs slightly better than gradient bandit policy with H = [0, 0] and α = 0.1. It is also worth noting that the gradient bandit policy performs better than almost all the other combinations of the epsilon greedy policy.