CS780: Deep Reinforcement Learning

Assignment #1

Name: Somya Gupta
Roll NO.: 211049
Google Colab Link (as tiny URL link): 4

https://colab.research.google.com/drive/1kYIYp0D3Xvmv8EPcld6wN2W1e1jUvudB?usp=sharing

Solution to Problem 1: Multi-armed Bandits

1. Running the 2-Armed Bandit environment on different sets of α and β: $(\alpha, \beta) = (0,0), (1,0), (0,1), (1,1), (0.5, 0.5)$. Expected average reward will be greater in cases with higher α and β.If α ¿ β the terminal state will most probably be 1,and vice versa.
   Average Rewards:[0.0, 0.65, 0.48250000000000004, 1.0241250000000002, 0.45120625]
   States[(0,0):[1, 2, 2, 2, 2], (1,0):[1, 1, 1, 1, 1], (0,1):[2, 2, 2, 2, 2], (1,1):[1, 1, 2, 2, 1], (0.5,0.5):[2, 2, 1, 2, 1]]

2. Running the Multi-Armed Bandit environment of different values of σ: σ=[0.5, 1, 2]
   Average Rewards:[0.2301904915626692, 0.315995115442413283, 0.9190924295390748]
   The average rewards increase as the standard deviation of the Gaussian distribution used to sample rewards increases, which aligns with expectations

3. a.

Table 1: Actions and Rewards

Action	Reward
0	0
0	1
0	1
0	0
0	0
0	1
0	1
0	0
0	0
0	0

Table 2: Q_estimates

Action=0	Action=1
0	0
0.5	0
0.66666667	0
0.5	0
0.4	0
0.5	0
0.57142857	0
0.5	0
0.44444444	0
0.4	0

b.

Table 3: Actions and Rewards

Action	Reward
1	1
1	1
0	0
1	0
0	1
1	1
1	1
0	1
0	1
0	1

Table 4: Q_estimates

Action=0	Action=1
0	1
0	1
0	1
0	0.66666667
0.5	0.66666667
0.5	0.75
0.5	0.8
0.66666667	0.8
0.75	0.8
0.8	0.8

c.

Table 5: Actions and Rewards

Action	Reward
1	1
1	0
0	1
1	1
1	1
1	1
0	0
1	0
1	1
0	0

Table 6: Q_estimates

Action=0	Action=1
0.	1.
0.	0.5
1.	0.5
1.	0.66666667
1.	0.75
1.	0.8
0.5	0.8
0.5	0.66666667
0.5	0.71428571
0.33333333	0.71428571

4. Running all the agents for 2-Armed BAndit the graph generally oscillates between 0 nd 1.Hence an average value of 0.5.
   Agent1:Pure Exploitation
   Agent2:Pure Exploration
   Agent3:Greedy epilson
   Agent4:Decaying Greedy epilson
   Agent5:Softmax
   Agent6:UCB

Figure 1: 2-Armed Bandit

5. Running all the agents for Multi-Armed Bandit the graph generally oscillates between -5 and 5.Hence the average value is 0 which aligns with the mean of reward we would get from the gaussian distributions given .
   Agent1:Pure Exploitation
   Agent2:Pure Exploration
   Agent3:Greedy epilson
   Agent4:Decaying Greedy epilson
   Agent5:Softmax

Agent6:UCB

Figure 2: Multi-Armed Bandit

6. Agent1:Pure Exploitation
   Agent2:Pure Exploration
   Agent3:Greedy epilson
   Agent4:Decaying Greedy epilson
   Agent5:Softmax
   Agent6:UCB

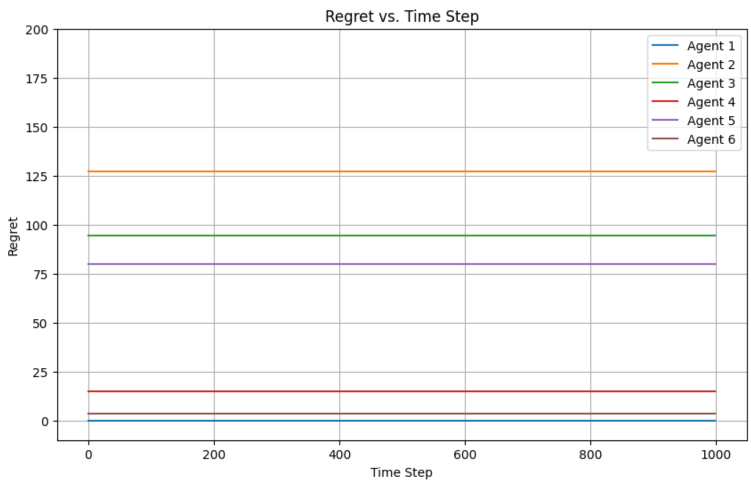

Figure 3: 2-Armed Bandit

7. Agent1:Pure Exploitation
   Agent2:Pure Exploration
   Agent3:Greedy epilson

Agent4:Decaying Greedy epilson
Agent5:Softmax
Agent6:UCB

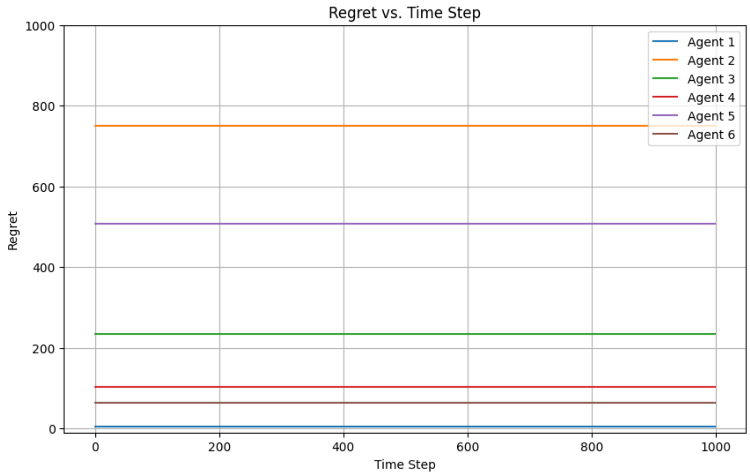

Figure 4: Multi-Armed Bandit

Solution to Problem 2: MC Estimates and TD Learning