CS780: Deep Reinforcement Learning

Assignment #1

Name: Divyaksh Shukla
Roll NO.: 231110603

Solution to Problem 1: Multi-armed Bandits

1. Created the environment using Gymnasium library. The environment is a 2-armed bernoulli bandit with some stochasticity. The environment characteristics are taken from a config file which specify the α and β values. I tested out with different values of α and β $[(0,0),(0,1),(1,0),(1,1),(0.5,0.5)]$ and the environment is working as expected. The agent recieves a reward only if it takes an action which takes it correctly in the direction of movement. That is if the the agent moves 'left' and lands in state 1 or moves 'right' and lands in state 2, only then it gets a positive reward.

2. I created a similar environment with 10-arms just like the above. But in this case the environment is not stochastic. The action and state-transitions are deterministic. The expected reward for each action $q_*(s,a)$ is sampled from a standard normal distribution ($\mathcal{N}(0,1)$). The agent then receives a reward from a normal distribution with mean $q_*(s,a)$ and variance 1. The agent has to then learn the optimal action to take in each state, by taking actions and observing the rewards.

3. I created 6 types of bandit agents following different strategies to solve the bandit problem. The agents are:

   (a) Greedy Agent: This agent always takes the action with the highest estimated value. It does not explore the environment.

   (b) Epsilon-Greedy Agent: This agent takes the action with the highest estimated value with probability $1-\epsilon$ and takes a random action with probability ϵ.

   (c) Decaying Epsilon-Greedy Agent: This agent is similar to the epsilon-greedy agent, but the value of ϵ decays "linearly" $\epsilon = max(0, \epsilon_0 - decay\_rate * episode)$ or "exponentially" $\epsilon = \epsilon_0 e^{-decay\_rate * episode}$ with time.

   (d) Softmax Agent: This agent takes actions with probability proportional to the exponential of the estimated value of the action. The agent explores the environment by taking actions by choosing from the below distribution

   $$\pi(a|s) = \frac{e^{Q(s,a)/\tau}}{\sum_b e^{Q(s,b)/\tau}} \tag{1}$$

   (e) UCB Agent: This agent takes actions by choosing the action with the highest upper confidence bound. The upper confidence bound is calculated as $Q(s,a) + c\sqrt{\frac{\ln t}{N(s,a)}}$ where c is a constant and $N(s,a)$ is the number of times the action a has been taken in state s. The action is taken by taking the argmax of the upper confidence bound.

4. Created 50 different bandit problems for 2-armed Bernoulli Bandit with α and β values chosen from uniform distribution $\mathcal{U}(0,1)$. The agents were then tested on these bandit problems. The agents were tested for 1000 episodes and the average reward, average regret and optimal action percentage was calculated. The results are shown in the plots below (Figure 1, 3, 5).

5. Created 50 different bandit problems for 10-armed Gaussain Bandit with $q_*(s,a)$ values chosen from standard normal distribution $\mathcal{N}(0,1)$ and then the agent receives a reward from $\mathcal{N}(q_*(s,a),1) \forall a \in A$. The agents were then tested on these bandit problems for 1000 episodes and the average reward, average regret and optimal action percentage was calculated. The results are shown in the plots below (Figure 2, 4, 6).

6. Created a plot of the average regret of the agents for the 2-armed Bernoulli Bandit.

7. Created a plot of the average regret of the agents for the 10-armed Gaussian Bandit.

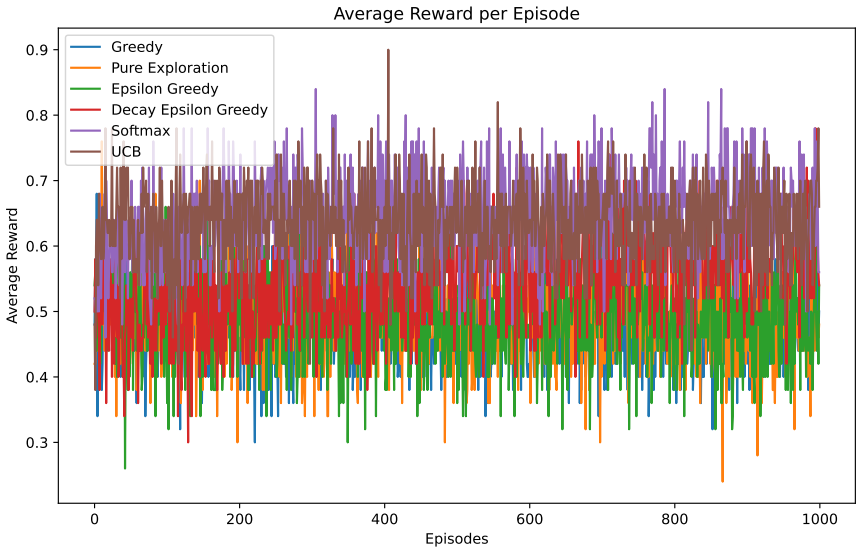

Figure 1: Average reward of 2-armed Bernoulli Bandit

8. This is the same as question 4.
9. Plotting the optimal action percentage of each agent for the 2-armed Bernoulli Bandit.
10. Plotting the optimal action percentage of each agent for the 10-armed Gaussian Bandit.

Solution to Problem 2: MC Estimates and TD Learning

1. Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

2. Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

3. Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

4. Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

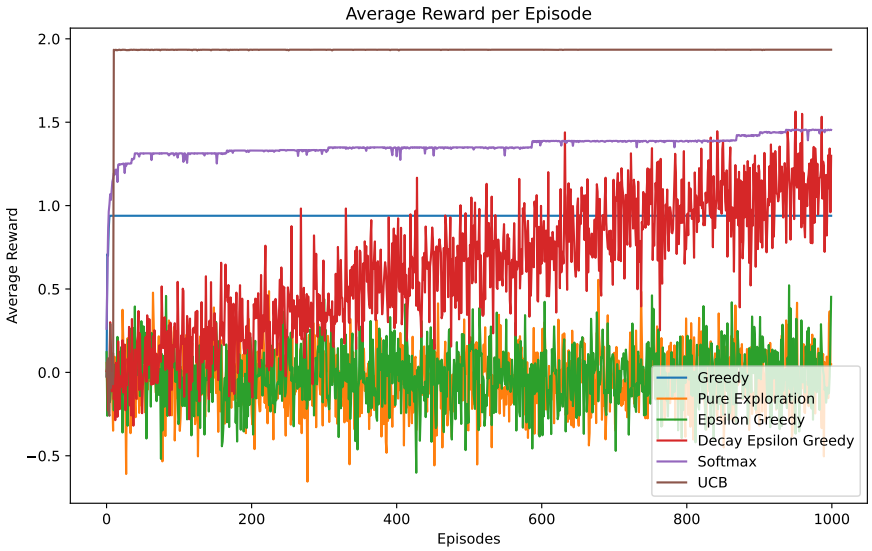

Figure 2: Average reward of 10-armed Gaussian Bandit

5. Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

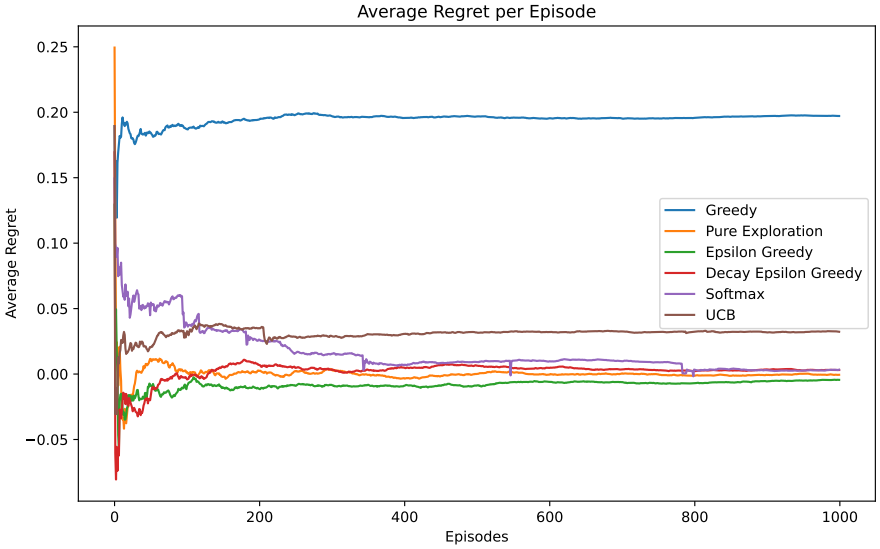

Figure 3: Average regret of 2-armed Bernoulli Bandit

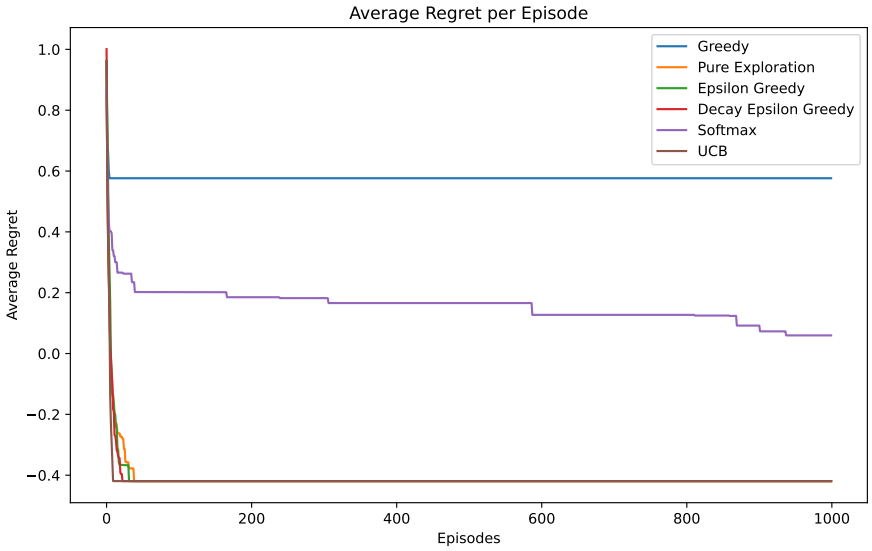

Figure 4: Average regret of 10-armed Gaussian Bandit

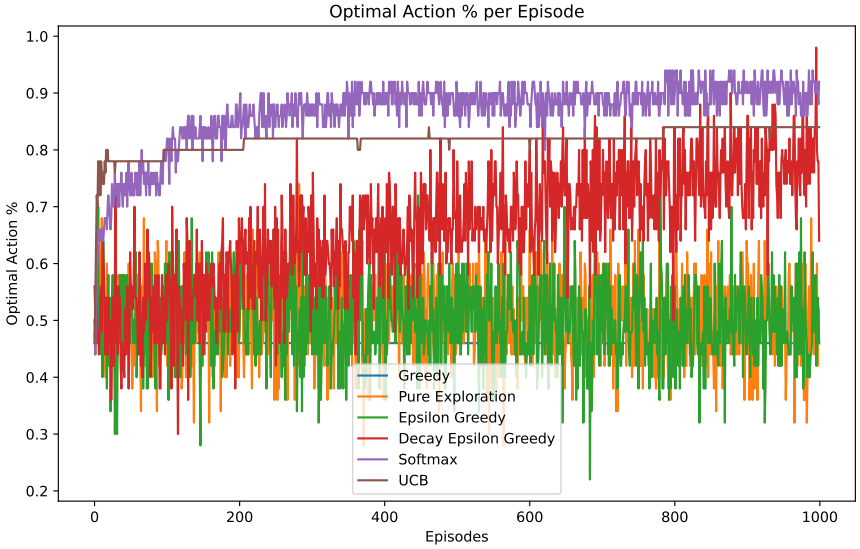

Figure 5: Optimal Action % of 2-armed Bernoulli Bandit

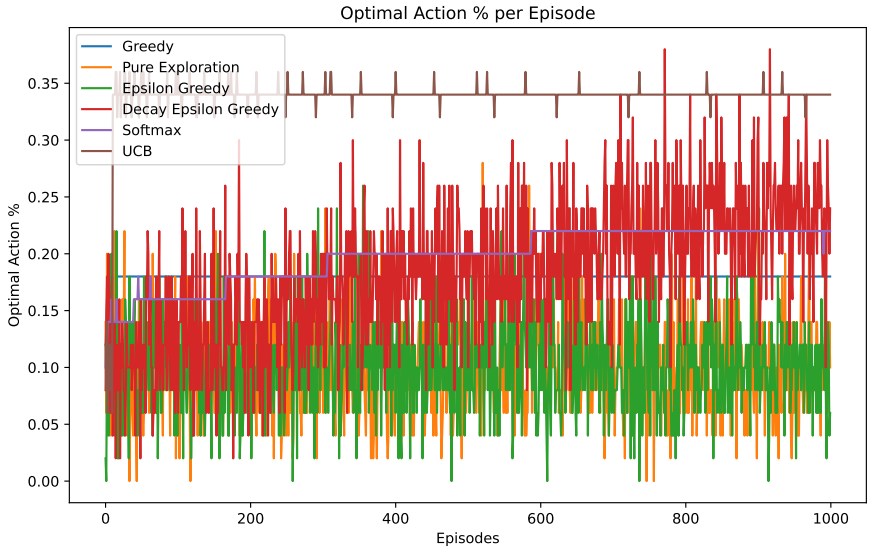

Figure 6: Optimal Action % of 10-armed Gaussian Bandit