DSE 314: Reinforcement Learning

Assignment-1

Question 1: Multi-armed bandit problem **(25 marks)**

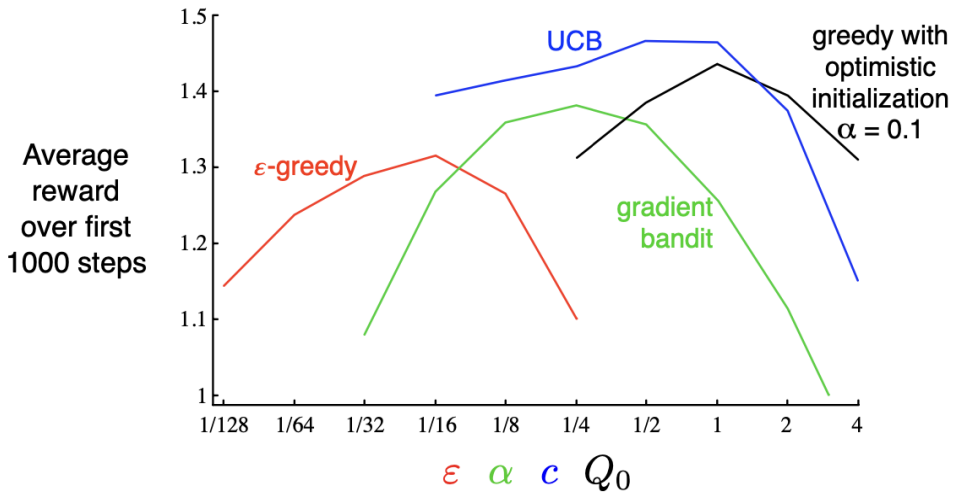
Suppose you have a 5-armed bandit testbed, whose true action values are fixed (stationary distributions). The true reward values of the bandits are q*(a) = [2.5, -3.5, 1.0, 5.0, -2.5]; and the deviation around the mean is given by σ[q*(a)] = [0.33, 1.0, 0.66, 1.98, 1.65]. We are assuming the reward to follow a Gaussian distribution characterized by their given mean and standard deviation.

The (general) multi-armed bandit environment can be found at:
https://github.com/manavmishra96/reinforcement-learning

Implement the following methods in the 5-arm bandit setting:

1. ε-greedy algorithm
2. UCB algorithm
3. Greedy with the optimistic initial value method
4. Gradient bandit algorithm

and plot their respective average reward time evolution graphs (Avg reward vs timestep). Finally, recreate the given plot by changing the respective algorithm parameters:



Question 2: (30 marks)

An undergraduate (not-so-ideal) student at IISER Bhopal has the task of attending classes and eating food during his tenure in college. The student has access to three locations on the campus: hostel (reward: -1), academic building (reward: +3), and canteen (reward: +1), and can either eat food or attend class at a given time.

When the student is at the hostel, (s)he attends classes either by going to the academic building with 50% probability and stays in the hostel with the 50% probability. When hungry, he/she goes to the canteen (from the hostel) with 100% probability. From the academic building, the student attends class where he stays in the academic building with 70% probability or goes to the canteen with 30% probability. When hungry at the academic building, he/she goes to the canteen with 80% probability or stays at the same place with 20% probability. At the canteen, the student has a 60% chance of attending classes by going to the academic building, a 20% chance of attending class by going to the hostel, and a 10% chance of attending from the canteen itself. If hungry, the student stays at the canteen with 100% probability.

Using this information, design a finite MDP by writing down the possible combinations of states, actions, transition probability from one state to another for a given action, and rewards in a tabular form. Also, draw a diagram of the MDP from the information mentioning the probability and rewards.
(Refer to example 3.3 of Chapter 3 in Sutton and Barto: Reinforcement Learning) (5 marks)

- Based on the designed MDP perform value iteration and show the optimal value for each state and the policy obtained. (10 marks)
- Based on the designed MDP perform policy iteration and show the optimal policy. (10 marks)
- Discuss the results obtained from policy iteration and value iteration (5 marks)