Reinforcement Learning

1. For the following sequence calculate the estimates of the expected reward for all arms.
   i) Using Sample Mean
   ii) Using Exponential Weighted Average (alpha = 0.1)

   Initial Values
   Q1(arm1) = 5
   Q1(arm2) = 8
   Q1(arm3) = -6
   Q1(arm4)= 0

   Sequence
   Action: 2, 3, 4, 4, 1, 2, 3, 3, 1
   Reward: -5, 9, 5, 2, -4, 9, 10, 2, 1

Is the sample mean affected by the choice of initial Q values?
Try to prove mathematically: The dependency of both i) and ii) on the initial Q value.

2. Using epsilon-greedy, generate an episode for 1000 time steps. [Python]

   Action Space: {1,2,3,4}
   Distribution of Rewards Associated with each Arm: {N(0,1), N(0,0.7), N(0, 0.2), N(0.2, 0.5)}
   Use Sample mean as the estimate for e-greedy selection.

   i) epsilon= 0.2
   ii) epsilon= 0.8
   iii) epsilon= 0
   iv) epsilon= 1
   v) Take epsilon to be a function of time, such that it decreases as t increases.

   Plot the Rewards that you get at every time step in all 5 cases.
   What is the average reward for each epsilon?

   ## Here, we are not averaging over runs. If you want to try doing that, then generate a sequence corresponding to an ‘epsilon’ multiple times (say 1000) and then take the average over those 1000 runs for a single time step. [See Next Question]

3. Read section 2.3 (The 10- armed Testbed) and generate figure 2.2 (Both the plots).

4. Repeat the above exercise by considering the variance of the distribution of the rewards associated with each arm to be 4.

5. **Upper Confidence Bound**: Generate figure 2.4 and solve Exercise 2.8.

6. **Bandit Gradient Algorithm:** Generate figure 2.5 (Average over 2000 trials to remove noise).