Assignment 4 - Reinforcement Learning

Arathy Bastin (03754963)
Varsha Raveendran(03745012)
Sandhanakrishnan Ravichandran (03755546)

July 7, 2022

Exercise 1

(a)

Q(A) = Q(A) + (1/N(A))*[R-Q(A)]
at time t=0:- Q(a) = 0 All action value functions are intialized to zero
at time t=1:- at t=1 we select action 1, Q(1) = 0 + 1*(-1-0) = -1 (rest all remains the same)
at time t=2:- at t=2 we select action 2, Q(2) = 0 + 1*(1-0) = -1 (rest all remains the same)
at time t=3:- at t=3 we select action 1, Q(2) = 1 + (1/2)*(-2-1) = -0.5 (rest all remains the same)
at time t=4:- at t=4 we select action 1, Q(2) = -0.5 + (1/3)*(2+0.5) = 1/3 (rest all remains the same)
at time t=5:- at t=5 we select action 1, Q(3) = 0 + 1*(0-0) = 0 (rest all remains the same)

time instance	Q(1)	Q(2)	Q(3)	Q(4)
0	0	0	0	0
1	-1	0	0	0
2	-1	1	0	0
3	-1	-1/2	0	0
4	-1	1/3	0	0
5	-1	1/3	0	0

(b)

At t=1: $Q(a_k) = 0$ for all k at this point, hence choosing action could have happened as a random choice due to ϵ, or by randomly choosing an action when tie-breaking.

At t=2: Because $Q(a_k) = 0$ for k = 2,3,4, choosing action 2 could have happened as a random choice due to ϵ among actions 2,3,4 or by randomly choosing an action(among actions 2,3,4) when tie-breaking.

At t=3: $Q(a_2) = 1$ at this point, so choosing action 2 is because it has the highest value function and we try to exploit using action 2. So here it is not the ϵ case.

At t=4: $Q(a_k) = 0$ actions (k=3,4) have the highest values but we choose action 2, so choosing action 2 (randomly) has definitely occurred because of ϵ case [exploration].

At t=5: $Q(a_2) = 1/3$ action 2 has the highest value at this point, but we choose action 3 trying to explore so it is definitely a random decision therefore ϵ case has definitely occurred [exploration]

Exercise 2

(a)

Exploitation - exploration dilemma: At every step of decision making, the agent can choose to take an action that will maximize expected reward at current step based on its previous experience. This is exploitation. Alternatively, the agent could choose to take risk and take a different action (one that it has not tried before). This is exploration. Exploitation would optimize the immediate rewards, however the long-term rewards may not be optimal whereas exploration would help in discovering actions that may result in better future rewards, however it is riskier.

In the class MultiArmed Bandit, the probability that the agent explores or exploits is controlled by the variable 'eps'. The agent chooses to explore (chooses a random action) if 'p' (a random number between 0 and 1) is less than eps. Otherwise the agent will exploit, i.e takes an action that gives maximum expected reward at that step.

(b)

The variable 'eps' defines the probability of agent choosing to explore another option instead of being greedy.
Case 1: eps=0 - The variable 'p' is a random value between 0 and 1. The chance that it is less than 0 is very low, thus the agent chooses to exploit more (greedy). However, always exploiting results in lower long-term rewards and sub-optimal policies.

Case 2: eps=0.01 - Here, the probability that the value of p is less than eps increases, thus the chance of agent to explore increases. We can see in the plot that the mean rewards for eps=0.01 is higher than eps=0.

Case 3: eps=0.1 - Here, the probability that the value of p is less than eps increases further, thus the chance of agent to explore is higher. We can see in the plot that the mean rewards for eps=0.1 is larger than the other two cases. This shows that a small chance of exploring actions helps the agent obtain better future rewards than always choosing the maximum reward at the current step.

(c)

(i) With k = 4 and mu ='sequence', the average rewards for each of the k arms (in order) is [0, 1, 2, 3]. The plot is similar to the Figure 1 with the eps=0.1 performing the best. However, the mean reward for each eps value is higher in Figure 2 than Figure 1.

(ii) With eps=0.1, the agent takes the optimal action (a=3) 88.3% of the time whereas eps=0 takes it only 25% and eps=0.01 takes it 60% of the time. With eps=0, the agent takes a=1 the highest number of times, showing that an initial choice of action without exploration results in low long term reward.

(iii) The plots show that agent with eps=0.1 estimates the rewards better and more accurate than eps=0. The agent with eps=0 estimates only for action with reward=2 and agent with eps=0.01 is in between the other two cases.

(d)

(i) In ϵ-decay, ϵ is dependent on the time step. Hence, as time increases, the probability of exploration decreases. This helps the agent explore more in the beginning, learn optimal estimates for actions and then greedily choose the optimal action in later steps.

(ii) Figure 5 shows that the mean rewards for ϵ-decay are larger than rewards obtained with the ϵ-greedy action selection. In Figure 6, we can observe that with ϵ-decay the agent selects the optimal action at a much higher (94%) percentage of time.

Exercise 3

$Q_{t+1} = Q_t + \alpha(R_t - Q_t)$
Show that this results in Q_{t+1} being a weighted average of past rewards and the initial estimate Q_1

$Q_{t+1} = Q_t + \alpha R_t - \alpha Q_t$
$Q_{t+1} = \alpha R_t + (1-\alpha)Q_t$
$Q_{t+1} = \alpha R_t + (1-\alpha)(\alpha R_{t-1} + (1-\alpha)Q_{t-1})$
$Q_{t+1} = \alpha R_t + (1-\alpha)\alpha R_{t-1} + (1-\alpha)^2 Q_{t-1}$
Similarly,
$Q_{t+1} = \alpha R_t + (1-\alpha)\alpha R_{t-1} + \alpha(1-\alpha)^2 R_{t-2} + ... + \alpha(1-\alpha)^{(t-1)} R_1 + (1-\alpha)^{(t)} Q_1$

$Q_{t+1} = (1-\alpha)^{(t)} Q_1 + \sum_{i=1}^{t} [\alpha(1-\alpha)^{(t-i)} R_i]$

Exercise 4

1. We need states, actions and reward to construct the RL agent. The agent starts using the 3D MRI images and the radiologist annotated subset data. The annotated data would act as the gold standard and the agent explores the solution space for the sub-image. Now the RL agent changes the threshold and the size of the structuring element for each sub-image on an individual basis. By performing each action, for each state-action pair, the agent receives a reward or punishment and updates the Q-matrix. After multiple episodes, the agent would have explored different actions and tries to exploit the most rewarding ones.

2. The problem of landmark detection can be formulated as a MDP(Markov Decision Process). The goal of the agent is to find the anatomical landmark. Input image is the environment with which the agent interacts using a set of actions. Agent interacts with the environment (the image) by a set of six actions in the positive and negative +x , -x, +y ,-y, +z,-z directions.

3. Defining the States: Each state defines a 3D region of interest centered around the landmark in the 3D MRI input image.

4. Defining reward : The reward function is designed such that it encourages the agent to move towards the target plane and learns at the same time. Reward function can be designed with respect to the Euclidean distance D between two points as, R = D (Po, Pt) – D (P1, Pt). P1 is the currently predicted landmark and Pt is the desired target landmark. By checking the difference between the two distances of the previous state and current state, reward is given. If the difference is less than the previous state it means the agent is learning and is awarded with a positive reward. If the distance is more than the previous state then agent is moving away from the landmark and awarded with a negative reward.

5. Q function: Approximated using neural network. The neural network comprises of convolution layers, pooling layers for extracting features from 3D input images. Output from convolutional layers are given as an input to fully connected layers. The last fully connected layer will have an output dimension of 6 because we have 6 actions and will have 6 - Q values. Softmax is used for obtaining the Q values. These values are considered by the agents and rewards are given accordingly. For each time step there will be six Q values and the action will be selected based on highest q value.

6. Terminal State: The state at which the agent has reached the landmark point. It is defined when Q-value of all possible actions approaches zeros (oscillation around zero). Terminal state or the landmark point is defined here as a circular ROI with 1mm radius.