Sumeet Shanbhag

Student ID : 641020714

Reinforcement Learning Assignment #2

1. **Suppose γ=0.8 and we get the following sequence of rewards**
   R_1=−2, R_2=1, R_3=3, R_4=4, R_5=1.0
   Calculate the value of G_0 by using the equation 3.8 (work forward) and 3.9 (work backward) and show they yield the same results

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{3.8}$$

$$\begin{aligned} G_t &\doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\ &= R_{t+1} + \gamma\left(R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots\right) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned} \tag{3.9}$$

Using Equation 3.8 (Work Forward):
For G0: G0 = R1 + γR2 + γ^2R3 + γ^3R4 + γ^4R5
Plugging in the given values:
G0 = -2 + 0.8(1) + 0.8^2(3) + 0.8^3(4) + 0.8^4(1.0) G0
= -2 + 0.8 + 1.92 + 2.048 + 0.4096 G0
= 3.1776

Using Equation 3.9 (Work Backward):
Starting from the end:
G4 = R5
G4 = 1.0
G3 = R4 + γG4 G3 = 4 + 0.8(1.0) G3 = 4.8
G2 = R3 + γG3 G2 = 3 + 0.8(4.8) G2 = 6.84
G1 = R2 + γG2 G1 = 1 + 0.8(6.84) G1 = 6.472
G0 = R1 + γG1 G0 = -2 + 0.8(6.472) G0 = 3.1776
Both methods yield the same result: G0 = 3.1776

2. **Explain how a room temperature control system can be modeled as an MDP? What are the states, actions, rewards, and transitions.**

a) **States:**
The states in this MDP represent the possible temperatures (or temperature ranges) the room can be in. For e.g.,:

- Very cold
- Cold
- Comfortable
- Warm
- Very warm

b) Actions:

The actions represent the decisions the temperature control system can make to influence the room's temperature. Actions might include:

- Increase temperature
- Decrease temperature
- Turn off heating/cooling
- Turn on heating/cooling

c) **Rewards:**

The rewards in this MDP provide feedback on the desirability of the room's temperature after taking an action. The goal is typically to keep the room at a comfortable temperature, so the reward structure might look like:

- Very cold: Negative reward (-10)
- Cold: Slightly negative reward (-5)
- Comfortable: Positive reward (+10)
- Warm: Slightly negative reward (-5)
- Very warm: Negative reward (-10)

d) Transitions:

Transitions represent the probability of moving from one state to another given an action. In the context of a room temperature control system, transitions are influenced by:

- The current temperature
- The action taken
- External factors (e.g., outside temperature, open windows, number of people in the room)

For example, if the room is "Cold" and the action is "Increase temperature," there might be a high probability that the next state will be "Comfortable." However, if there's a snowstorm outside, even with the "Increase temperature" action, the room might still remain "Cold."

3. What is the reward hypothesis in RL?

The reward hypothesis in reinforcement learning suggests that every objective or goal an agent might pursue can be represented as the task of accumulating the highest possible reward over time.

In essence, this idea proposes that instead of hardcoding specific behaviours or tasks for an agent, we can define a reward system. This system assigns values or "rewards" to various outcomes. The agent's main goal then becomes to act in ways that maximize its total reward over a period.

This approach is fundamental to reinforcement learning because it offers a versatile framework for decision-making. By focusing on rewards, we can apply reinforcement learning to a broad spectrum of tasks, from games to real-world applications. The agent learns to make optimal decisions by continually assessing the potential rewards of its actions based on its experiences.

4. **We have an agent in maze-like world. We want the agent to find the goal as soon as possible. We set the reward for reaching the goal equal to +1 With $\gamma\gamma$=1. But we notice that the agent does not always reach the goal as soon as possible. How can we fix this?**

When the reward for reaching the goal is set to +1 and γ =1, the agent does not have an incentive to reach the goal quickly. This is because, regardless of how many steps it takes, the cumulative reward (with a γ of 1) remains the same.

To encourage the agent to find the goal as soon as possible, you can introduce a negative reward for each time step that the agent takes without reaching the goal. This creates an incentive for the agent to minimize the number of steps it takes to reach the goal, as taking more steps would lead to a lower cumulative reward.

a) **Negative Reward for Each Step:** Assign a small negative reward (e.g., -0.01 or -0.1) for each step the agent takes. This way, the longer the agent takes to reach the goal, the lower its total reward will be.

b) **Immediate Reward for the Goal:** Continue to give a reward of +1 when the agent reaches the goal.

With this setup, the agent will be incentivized to reach the goal as quickly as possible to maximize its cumulative reward. The negative reward for each step acts as a "cost" that the agent will try to minimize, pushing it to find the shortest path to the goal.

5. **What is the difference between policy and action?**

a) Policy:

Definition: A policy, denoted as π, defines the behaviour of an agent. It specifies how the agent should act in a given state.

Types:

- **Deterministic Policy:** For a given state, the policy provides a specific action. It's a direct mapping from states to actions: $\pi(s)=a$.
- **Stochastic Policy:** For a given state, the policy provides a probability distribution over actions. It indicates the likelihood of taking each action in that state: $\pi(a|s)$ is the probability of taking action a in state s.

Representation: A policy can be represented in various ways, such as a lookup table, a neural network, or other function approximators.

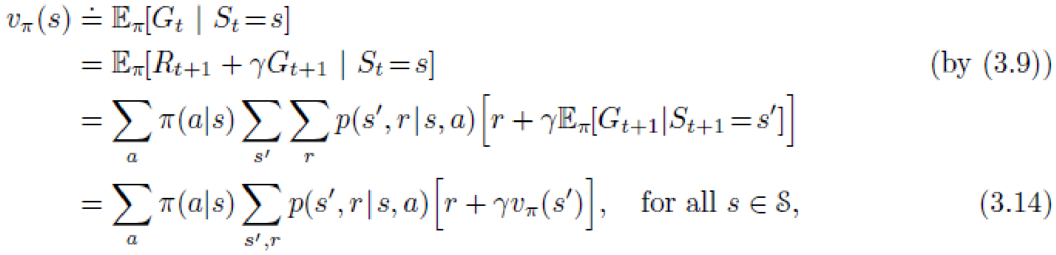
b) Action:

Definition: An action, often denoted as *a*, is a specific decision made by the agent that affects the environment. It's what the agent decides to do at a particular moment.

Space: The set of all possible actions the agent can take is called the action space. For instance, in a game where the agent can move up, down, left, or right, the action space consists of these four actions.

Outcome: Once an action is taken, the environment responds, leading to a new state and a reward. The action, combined with the current state, influences the future state and the immediate reward the agent receives.

6. Exercise 3.14 of the textbook.



$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}_\pi[G_t \mid S_t = s] \\ &= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} \mid S_t = s] && \text{(by (3.9))} \\ &= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r | s, a) \Big[r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s'] \Big] \\ &= \sum_a \pi(a|s) \sum_{s', r} p(s', r | s, a) \Big[r + \gamma v_\pi(s') \Big], \quad \text{for all } s \in \mathcal{S}, && (3.14) \end{aligned}$$

Given the scenario, the agent is in the centre state and has an equal probability of moving to any of the neighbouring states. The value of the centre state is +0.7. γ=0.9

The neighbouring states have values:

- State 1: +2.3
- State 2: +0.4
- State 3: -0.4
- State 4: +0.7

The agent has an equal chance of moving to any of the neighbouring states, so the probability for any action from the centre state is 0.25.

Assuming the immediate rewards for moving to the neighbouring states are 0, the value of the centre state using the Bellman equation is: Value of centre state = 0.25 x (0.9 x (2.3 + 0.4 - 0.4 + 0.7)) Value of centre state = 0.25 x (0.9 x 3.0) Value of centre state = 0.25 x 2.7 Value of centre state = 0.675

The computed value is 0.675, which is close to the given value of 0.7.

7. Exercise 3.17 of the textbook.

For a given state "s" and action "a", the action value is represented as "q(s, a)". This value is calculated based on:

a) The immediate reward received after taking action "a" in state "s".
b) The expected future returns from the next state, which we'll call "s'", after taking action "a'" in that state.

The equation considers all possible next states "s'" and rewards that can be received. It also considers the probability of transitioning to each next state "s'" and the probability of taking each action "a'" in the next state according to the policy.

In mathematical terms, the equation is a summation over all possible next states s' and rewards. For each next state s' and reward, it calculates the immediate reward plus the discounted expected future return from that state. The γ, often represented by γ, determines the present value of future rewards.

Starting from a state-action pair s, a, the agent transitions to a new state "s'" based on the action "a" with a certain probability. In the new state s', the agent has a distribution over possible actions according to the policy. The value of taking each action in state "s'" is given by "q(s', a')".

8. Exercise 3.22 of the textbook.

- If the agent chooses the "left" action, it receives an immediate reward of +1.
- If the agent chooses the "right" action, it receives no reward (or a reward of 0).

1. **When γ=0:**
   At γ =0, the agent only cares about the immediate reward and doesn't consider any future rewards.
   - If the agent chooses left, it gets an immediate reward of +1.
   - If it chooses right, it gets 0.

   Clearly, left is the better choice. So, the optimal policy π* is π_{left}.

2. **When γ=0.9:**
   At γ =0.9, the agent considers future rewards heavily but still gives some weight to the immediate reward.
   - If the agent chooses left, it gives a total reward of 1+0.9(0)=11+0.9(0)=1.
   - If it chooses right, it gives a total reward of 0+0.9(2)=1.80+0.9(2)=1.8.

   With a high discount factor, right is the better choice due to the higher future reward. So, the optimal policy π* is π_{right}.

3. **When γ=0.5:**
   At γ =0.5, the agent balances between immediate and future rewards.
   - If the agent chooses left, it gives a total reward of 1+0.5(0)=11+0.5(0)=1.
   - If it chooses right, it gives a total reward of 0+0.5(2)=10+0.5(2)=1.

   In this case, both actions provide the same total reward. So, both π_{left} and π_{right} are optimal policies.