RBE 595 — Reinforcement Learning
Week #3 Assignment

Arjan Gupta

Problem 1

Suppose $\gamma = 0.8$ 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.

Answer

Work Forward

From the the book, the *discounted return* (equation 3.8), G_t, is defined as,

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

Plugging in the values from this problem, we get,

$$\begin{aligned} G_0 &= R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 R_5 \\ &= -2 + 0.8 \cdot 1 + 0.8^2 \cdot 3 + 0.8^3 \cdot 4 + 0.8^4 \cdot 1 \\ &= -2 + 0.8 + 0.64 \cdot 3 + 0.512 \cdot 4 + 0.4096 \\ &= 3.1776 \end{aligned}$$

Work Backward

From the book, the "recursive" representation of *discounted return* (equation 3.9), G_t, is defined as,

$$G_t \doteq R_{t+1} + \gamma G_{t+1} \tag{3.9}$$

Plugging in the values from this problem, we get,

$$\begin{aligned} G_0 &= R_1 + \gamma G_1 \\ &= -2 + 0.8 \cdot G_1 \end{aligned}$$

Where we apply 3.8 to G_1,

$$\begin{aligned} G_1 &= R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 \\ &= 1 + 0.8 \cdot 3 + 0.8^2 \cdot 4 + 0.8^3 \cdot 1 \\ &= 6.472 \end{aligned}$$

Therefore,

$$\begin{aligned} G_0 &= -2 + 0.8 \cdot G_1 \\ &= -2 + 0.8 \cdot 6.472 \\ &= 3.1776 \end{aligned}$$

Conclusion

We see that both methods yield the same result, $G_0 = 3.1776$.

Problem 2

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

Answer

A room temperature control system can be modeled as an MDP as follows.

States: The states are the different temperatures that the room can be in.

Actions: The actions are the different actions that the system can take to change the temperature of the room.

Rewards: The rewards are the different rewards that the system can receive for taking an action.

Transitions: The transitions are the different transitions that the system can make from one state to another.

Problem 3

What is the reward hypothesis in RL?

Answer

The book states the *reward hypothesis* as follows,

> That all of what we mean by goals and purposes can be well thought of as the maximization of the expected value of the cumulative sum of a received scalar signal (called reward).

Here is a simple break-down of what the reward hypothesis means:

- In RL, we talk about goals and purposes, which is to find best way to solve a problem.
- Any solution to a complex problem can be broken down into a series of steps, and each step can have a value associated with it.
- We design this 'value' associated with each step as a scalar signal which is received from the environment. This scalar signal is called the *reward*.
- Therefore, our ultimate goal is to maximize the expected value of the cumulative sum of these rewards.

Problem 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 $\gamma = 1$. But we notice that the agent does not always reach the goal as soon as possible. How can we fix this?

Answer

TODO

Problem 5

What is the difference between policy and action?

Answer

TODO

Problem 6

(Exercise 3.14) Write prompt

Answer

TODO

Problem 7

(Exercise 3.17) Write prompt

Answer

TODO

Problem 8

(Exercise 3.22) Write prompt

Answer

TODO