Assignment 5: Reinforcement Learning

The Swiss AI Lab IDSIA, USI, SUPSI

Name : Krunal Rathod | email : krunal.rathod@usi.ch

December 19, 2023

- **Problem** Suppose a robot is put in a maze with a long corridor. The corridor is 1 kilometre long and 5 meters wide. The available actions to the robot are moving forward 1 meter, moving backward 1 meter, turning left by 90 degrees and turning right by 90 degrees. If the robot moves and hits the wall, then it will stay in its position and orientation. The robot's goal is to escape from this maze by reaching the end of the long corridor.

Question 1. Assume the robot receives a +1 reward signal for each time step taken in the maze and +1000 for reaching the final goal (the end of the long corridor). Then you train the robot for a while, but it seems it still does not perform well at all for navigating to the end of the corridor in the maze. What is happening? Is there something wrong with the reward function? (4 points)
Answer 1. Short answer, yes. The reward structure, in which the robot receives +1 reward at the each steps, makes learning difficult. The lack of intermediate rewards makes it challenging for the robot to find the optimal path and results in slow or inefficient learning. The robot can keep on increasing the reward infinitely by just moving around in the maze which is not leading to the end of the corridor.

Question 2. If there is something wrong with the reward function, how could you fix it? If not, how to resolve the training issues? (4 points)
Answer 2. We can adjust the reward function by assigning +1 or 0 rewards for each step, providing more frequent feedback to guide the robot's learning effectively. We can also introduce small negative rewards for each step taken by robot in maze.

Question 3. The discounted return for a non-episodic task is defined as

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...$$

where $\gamma \in [0, 1]$ is the discount factor. Rewrite the above equation such that G_t is on the lefthand side and G_{t+1} is on the righthand side. (2 points)
Answer 3. To express G_t on the left-hand side and G_{t+1} on the right-hand side, we can rewrite the equation as follows:

$$G_t = R_{t+1} + \gamma \cdot (R_{t+2} + \gamma \cdot R_{t+3} + \ldots)$$

Now, we can observe that the term inside the parentheses are equivalent to G_{t+1}:

$$G_t = R_{t+1} + \gamma \cdot G_{t+1}$$

Question 4. Assume that the rewards are bounded, i.e. $R_t < r_{\max}$ for all t. Give a sufficient condition for γ, which assures that the infinite series for G_t is bounded. (3 points)
Answer 4. For the series to be bounded, the sum should converge, which happens only when the absolute value of the common ratio is less than 1. Therefore, a sufficient condition for γ to assure that the infinite series for G_t is bounded is:

$$\gamma < 1$$

Question 5. Let the task be an episodic setting, and the robot is running for T = 5 time steps. Suppose $\gamma = 0.9$, and the robot receives rewards along the way $R_1 = -1, R_2 = -0.5, R_3 = 2.5, R_4 = 1, R_5 = 3$. What are the values for $G_0, G_1, G_2, G_3, G_4, G_5$? (5 points)
Answer 5. Given the rewards $R_1 = -1, R_2 = -0.5, R_3 = 2.5, R_4 = 1, R_5 = 3$ and $\gamma = 0.9$, to calculate the values

G_0:

$$G_0 = -1 + 0.9 \cdot (-0.5) + (0.9)^2 \cdot 2.5 + (0.9)^3 \cdot 1 + (0.9)^4 \cdot 3 = 3.2723$$

G_1:

$$G_1 = -0.5 + 0.9 \cdot 2.5 + (0.9)^2 \cdot 1 + (0.9)^3 \cdot 3 = 4.747$$

G_2:

$$G_2 = 2.5 + 0.9 \cdot 1 + (0.9)^2 \cdot 3 = 5.83$$

G_3:

$$G_3 = 1 + 0.9 \cdot 3 = 3.7$$

G_4:

$$G_4 = 3$$

G_5:

$$G_5 = 0$$

Question 6. Now consider episodic tasks, and similar to the last question, we add a constant c to each reward, how does it change G_t? (5 points)
Answer 6. Formula for G_t is given by:

$$G_t = R_{t+1} + \gamma \cdot R_{t+2} + \gamma^2 \cdot R_{t+3} + \ldots + \gamma^{T-t-1} \cdot R_T$$

After adding a constant c, the new G_t formula:

$$G_t{}^{new} = (R_{t+1} + c) + \gamma \cdot (R_{t+2} + c) + \gamma^2 \cdot (R_{t+3} + c) + \ldots + \gamma^{T-t-1} \cdot (R_T + c)$$

Simplifying further:

$$G_t{}^{new} = G_t + c \cdot (1 + \gamma + \gamma^2 + \ldots + \gamma^{T-t-1})$$

The sum inside the parentheses is a finite geometric series with the first term 1 and a common ratio γ. Therefore, the modified discounted return $G_t{}^{new}$ is:

$$G_t{}^{new} = G_t + c \cdot \frac{1 - \gamma^{T-t}}{1 - \gamma}$$

Bonus Question. Suppose the infinite series for G_t is bounded, and each reward in the series is a constant of +1. What is a simple formula for this bound? Write it down without using summation. (3 points)

Answer. In our case, each reward is a constant of +1 and the discount factor is γ. So, the bound for the infinite series for G_t can be given by:

$$\frac{1}{1-\gamma}$$