Assignment 5: Reinforcement Learning

The Swiss AI Lab IDSIA (USI-SUPSI)

Due by 23:59 on the 22nd of December, 2023

Submission Instructions Please submit your answers in LaTeX (e.g. `http://overleaf.com`) as a single PDF file. Name the file `firstname.lastname.pdf` with `firstname` replaced by your first name and `lastname` replaced by your last name, then upload it to the iCorsi website before the deadline. Incorrectly formatted submissions and late submissions will receive a grade of 0. Keep your answers brief in line with the number of points allocated for each question. Note that there are a total of 23 points and up to 3 bonus points in this assignment, with a maximum score of 23/23.

Collaboration Policy We encourage you to ask questions or to discuss exercises with other students. However, under no circumstances should you share your answer with other students or look at any other students' answers. If two submissions or any answers therein are deemed to be too similar by the responsible TA, or plagiarism, or cheating is deemed likely to have occurred, all students who are believed to be involved will be penalized. Penalties can include receiving a grade of 0 for the course, irrespective of any previously assigned grades. Note that the above will be judged solely by the instructor and according to a balance of probabilities and not according to the principle of beyond reasonable doubt.

For questions on this assignment, you can contact the responsible TA at `dylan.ashley@idsia.ch`

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)

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)

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

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

where $\gamma \in [0, 1]$ is the discount factor. Rewrite the above equation such that G_t is on the left-hand side and G_{t+1} is on the right-hand side. (2 points)

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)

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)

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)

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)