Reinforcement Learning Assignment

The Swiss AI Lab IDSIA (USI-SUPSI)

Due by 23:59 on the 16th of December, 2024

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.

Use of Large Language Models The problems in the final exam are primarily built through a reformulation of the problems given in this assignment and on the worksheet. While the grading on the final exam and this assignment will be done harshly, a student who can easily answer all the questions on both would be expected to score well on the exam. Thus, while you are not prohibited from using large language models to help answer the questions here, you are advised to ensure you can answer them comfortably without using an LLM (though you may find it useful to use an LLM to help you understand parts of the question). As verbose answers cannot be taken to be a demonstration of your knowledge, if you provide incorrect information in an answer, you will be docked marks in accordance with it. In an extreme case, if you provide two or more answers to a question where one is correct and one is incorrect, you will be marked as though the incorrect answer was your only answer.

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

Question 1

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 1.1. 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
Here, the problem seems that, as the reward funcion function is structure, the robot can just make a "forward-backward" steps infinitely in order to maximize the reward. In that case, it will never reach the end of the corridor

Question 1.2. If there is something wrong with the reward function, how could you fix it? If not, how can you resolve the training issues? (4 points)

ANSWER
A possible solution is simply to replace the +1 for the backward steps with -1, forcing the robot to turn left/right (this operation has a reward value of 0, meaning that it will be performed only when there is an obstacle that makes impossible to move forward (+1) and discourages to go backward (-1).

Question 2

The discounted return for a non-episodic task is defined as $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$, where $\gamma \in [0, 1]$ is the discount factor.

Question 2.1. Rewrite the above equation such that G_{t+1} is on the right-hand side and G_t is **alone** on the left-hand side. (2 points)

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

Question 2.2. Give a sufficient condition for γ, which assures that the infinite series G_t is bounded. (3 points)

ANSWER
If, $R_t < r_{\max} \in \mathbb{R}$ for all t, G_t will be bounded for $\gamma < 1$ since $\gamma^n r_{\max} \to 0$ as much as n increases

Question 2.3. Now consider a task similar to the one described in Question 1 but with an unknown environment and unknown reward function. Let the task be an episodic setting, with the robot running for $T = 5$ time steps before terminating. Suppose $\gamma = 0.9$, and the robot receives the following rewards along the way: $R_1 = 1, R_2 = -1, R_3 = 2.5, R_4 = -5$, and $R_5 = 3$. What are the values for $G_0, G_1, G_2, G_3, G_4, G_5$? Give your answer as a single real number for each of G_0 though G_5 and show your work. (5 points)

ANSWER
$G_t = R_{t+1} + 0.9G_{t+1}$
$G_5 = R_5 = 3$
$G_4 = R_4 + 0.9G_5 = -5 + 0.9 * 3 = -2.3$
$G_3 = R_3 + 0.9G_4 = 2.5 + 0.9 * (-2.5) = 0.43$
$G_2 = R_2 + 0.9G_3 = 1 + 0.9 * 0.43 = -0.61$
$G_1 = R_1 + 0.9G_2 = 1 + 0.9 * (-0.61) = 0.45$
$G_0 = 0.9G_1 = 0.9 * 0.45 = 0.405$

G5 = 0 (terminal state)
G4 = R5 + 0.9*G5 = 3
etc.

Question 2.4. Now consider an 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 close but this only works for infinite horizons
The formula changes as $G_t = (R_{t+1} + c) + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + ... = (old)G_t + c/(1 - \gamma)$. In other words, c adds a contribute as $1/(1\text{-}\gamma)$

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
If we assume that each reward is a constant of +1, we will just get $G_t = 1/(1 - \gamma)$