Reinforcement Learning Assignment

The Swiss AI Lab IDSIA (USI-SUPSI), Solution: Alan Copa

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)

The robot is getting a +1 reward for each step taken, ignoring if it is a step in the right direction or if it a useful step to reach the final goal in general. This could cause the robot to try to do as many steps as possible since it tries to maximize the reward. Instead to learn a minimum distance walk to the goal it will try to perform the longest path to reach the goal. Even with a +1000 reward for reaching the final goal the robot will walk around without reaching the goal because it finds better to collect many times the +1 reward. Thus the reward function is wrong because it prioritizes step taking and not efficiency in reaching the final goal, and will not ensure that the robot will reach the final goal.

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)

To fix the reward function I would have penalizing reward for taking a step to encourage least step taking, an example could be to have a reward of -1 for each step taken. For this specific case of a straight line maze, I would penalize more moving back by 1 meter (-5 reward), give -2 reward to turning left or right and give -1 reward to go forward by 1 meter. I would still give a final goal reward of +1000 such that that it balances out the negative rewards and incentives the robot to reach the final goal, since it's trying to maximize the cumulative reward (in this case, assuming the robot starts facing the final goal, the maximal cumulative reward should be 0 since the robot is taking 1000 1 meter steps forward (reward -1) to walk 1 km to the final goal (reward +1000)).

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)

$$G_{t+1} = R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \ldots$$

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

Thus:

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

Question 2.2. Assume that the rewards are bounded, i.e., $R_t < r_{\max} \in \mathbb{R}$ for all t. Give a sufficient condition for γ, which assures that the infinite series G_t is bounded. (3 points)

If the rewards are bounded it means that

$$|R_{t+k}| \leq r_{\max}, \quad \forall k \geq 1.$$

thus, we can rewrite the discounted return G_t as:

$$|G_t| \leq r_{\max} \left(1 + \gamma + \gamma^2 + \ldots\right)$$

we recognize the geometric series:

$$\sum_{k=0}^{\infty} \gamma^k = 1 + \gamma + \gamma^2 + \ldots$$

this sum converges by definition to a stable value $\frac{1}{1-\gamma}$ only if $0 \leq \gamma < 1$. Thus, since R_t is a known value, G_t is bounded when $0 \leq \gamma < 1$.

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)

In an episodic task the terminal state has reward $G_5 = 0$ since there are no future rewards.
Following the formula $G_t = R_{t+1} + \gamma G_{t+1}$:
$G_5 = 0$ (Terminal state)
$G_4 = R_5 + \gamma G_5 = 3 + 0.9 \cdot 0 = 3$
$G_3 = R_4 + \gamma G_4 = -5 + 0.9 \cdot 3 = -2.3$
$G_2 = R_3 + \gamma G_3 = 2.5 + 0.9 \cdot -2.3 = 0.43$
$G_1 = R_2 + \gamma G_2 = -1 + 0.9 \cdot 0.43 = -0.613$
$G_0 = R_1 + \gamma G_1 = 1 + 0.9 \cdot -0.613 = 0.4483$

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)

The new discounted return is:

$$G'_t = (R_{t+1} + c) + \gamma(R_{t+2} + c) + \gamma^2(R_{t+3} + c) + \dots$$

rearranging:

$$G'_t = (R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots) + (c + \gamma c + \gamma^2 c + \dots)$$

we identify G_t:

$$G'_t = G_t + c(1 + \gamma + \gamma^2 + \dots)$$

we introduce the converging geometric series from before, that now is finite

$$\sum_{k=0}^{T-t-1} \gamma^k = \frac{1 - \gamma^{T-t}}{1 - \gamma}, \quad \text{for } 0 \leq \gamma < 1.$$

finally:

$$G'_t = G_t + c \cdot \frac{1 - \gamma^{T-t}}{1 - \gamma}.$$

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)

Starting from:

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

we know:

$$R_t = 1, \forall t$$

so the discounted return becomes:

$$G_t = 1 + \gamma + \gamma^2 + \cdots = \sum_{k=0}^{\infty} \gamma^k$$

since G_t is bounded, we have $0 \leq \gamma < 1$, so we can compute the result of the infinite geometric series, thus:

$$\sum_{k=0}^{\infty} \gamma^k = \frac{1}{1-\gamma}$$

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