Machine Learning Assignment 5: Reinforcement Learning

Fabian Gobet

13 December 2023

Solutions

Question 1: Performance in the Maze

The robot may not be performing well due to the structure of the reward function. A constant reward of +1 for each time step could be incentivizing the robot to delay reaching the end of the corridor and 'forever' accumulate rewards, that eventually supersede that of reaching the end of corridor, thus leading to aimless wandering instead of goal oriented behaviour.

Question 2: Training Issue or Reward Function Problem

To address this reward issue, the reward function ought to be modified to encourage reaching the end of the corridor more directly.

We could, for example, implement a small negative reward for each time step that doesn't lead to the end of the corridor to incentivize the robot to take the shortest path to the end of the corridor.

The main key change should be that at least non-terminal leading steps don't provide a positive reward, because as long this is the case then the robot will be motivated to take more aimless steps to accumulate rewards.

Question 3: Discounted Return Equation

The discounted return for a non-episodic task can be written as:

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

where G_t is the total discounted return at time t, R_{t+1} is the reward after time t, and γ is the discount factor.

Question 4: Boundedness of Infinite Series for G_t

Notice that:

$$G_t = \sum_{i \geq 0} \gamma^i R_{t+1+i} \leq \sum_{i \geq 0} \gamma^i r_{max}$$

This is an infinite geometric series. As such, its convergence is determined by the value of γ, that is $|\gamma| < 1$.

We know that $\gamma \in [0, 1]$, thus a necessary condition of convergence is $\gamma \in [0, 1)$, i.e. $0 \leq \gamma < 1$.

Question 5: Values of $G_0, G_1, G_2, G_3, G_4, G_5$

If we consider $T = 5$, then we can assume there is no rewards after $t = 5$, hence $G_5 = 0$. Given $\gamma = 0.9$ and rewards $R1 = -1$, $R2 = -0.5$, $R3 = 2.5$, $R4 = 1$, $R5 = 3$, using the formula of question 3 the values are:

$$
\begin{aligned}
G_5 &= 0 \\
G_4 &= R_5 + \gamma G_5 = 3 \\
G_3 &= R_4 + \gamma G_4 = 3.7 \\
G_2 &= R_3 + \gamma G_3 = 5.83 \\
G_1 &= R_2 + \gamma G_2 = 4.747 \\
G_0 &= R_1 + \gamma G_1 = 3.2723
\end{aligned}
$$

Question 6: Effect of Adding a Constant c to Each Reward

If the same constant c is added to all rewards then let G'_t denote the new return function such that:

$$G'_t = R_{t+1} + c + \gamma(R_{t+2} + c) + \cdots = G_t + c(1 + \gamma + \gamma^2 \cdots) \qquad (1)$$

Notice that for a finite set of episodes we would have to consider the passed episodes for the sum of the geometric series, that is:

$$\frac{c(1-\gamma^T)}{1-\gamma} - \frac{c(1-\gamma^{t-1})}{1-\gamma} = \frac{c}{1-\gamma}(\gamma^{t-1} - \gamma^T) \qquad (2)$$

where if $|\gamma| < 1$, then for infinite many episodes we have

$$\lim_{T\to\infty} \frac{c}{1-\gamma}(\gamma^{t-1} - \gamma^T) = \frac{c\gamma^{t-1}}{1-\gamma} \qquad (3)$$

Therefore, from (1), (2) and (3) we have that

$$G'_t = G_t + \begin{cases} \frac{c}{1-\gamma}(\gamma^{t-1} - \gamma^T), & \text{if } t \leq T < \infty \\ \frac{c}{1-\gamma}\gamma^{t-1}, & \text{otherwise} \end{cases}$$

So we can see that adding a constant c to each reward does affect the return function. The same conclusion can be inferred if every reward was added with a different constant c, i.e. that it would affect the return function.

Question 7: Bound of Infinite Series for G_t with Constant Reward

For a constant reward of $+1$ starting at episode t, the bound is:

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

assuming $0 < \gamma < 1$.