RL Assignment 2

Matteo Manias 1822363

November 23, 2023

1 Q-learning and SARSA update

The Q-learning update is given by:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot [r + \gamma \cdot \max(Q(s',a')) - Q(s,a)]$$

the SARSA update is given by:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \cdot [r + \gamma \cdot Q(s',a') - Q(s,a)]$$

the given values of the tabular Q function are:

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

substituting the Q-values, reward = 3, discount factor = 0.9 and learning rate = 0.1 with their respective values, we get:

$$Q(1,2) \leftarrow Q(1,2) + 0.1 \cdot [3 + 0.5 \cdot \max(Q(2,1), Q(2,2)) - Q(1,2)]$$

$$Q(1,2) \leftarrow 2 + 0.1 \cdot [3 + 0.5 \cdot \max(3,4) - 2]$$

$$Q(1,2) \leftarrow 2.3$$

that is the new Q-value for state 1, action 2 is 2.3. The other Q-values remain unchanged.

The SARSA update is given by substituting the same values as above but by choosing the action a' from the policy π.

Given that the action choosen by policy π is the same as the action choosen by the max operator in the q-learning update, the SARSA update will be the same as the q-learning update.

2 Question 2

the n-step error is given by:

$$G_{t:t+n} - V_{t+n-1}(S_t)$$

where $G_{t:t+n}$ is the n-step return, defined as:

$$G_{t:t+n} = R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n})$$

By subtracting $V_{t+n-1}(S_t)$ to the n-step return, we get:

$$G_{t:t+n} - V_{t+n-1}(S_t) = (R_{t+1} + \gamma R_{t+2} + \ldots + \gamma^{n-1} R_{t+n} + \gamma^n V_{t+n-1}(S_{t+n})) - V_{t+n-1}(S_t)$$

This can be written as a sum of TD errors by rearranging the terms as follows:

$$\begin{aligned} &= (R_{t+1} - V_t(S_t)) + \gamma(R_{t+2} - V_{t+1}(S_{t+1})) + \ldots \\ &\quad + \gamma^{n-1}(R_{t+n} - V_{t+n-1}(S_{t+n-1})) + \gamma^n V_{t+n-1}(S_{t+n}) - V_{t+n-1}(S_t) \end{aligned}$$

Given that the value estimates don't change from step to step, we can simplify the expression as follows as some terms cancel out:

$$= \gamma^n V_{t+n-1}(S_{t+n}) - V_{t+n-1}(S_t)$$

by noticing that:

The first term $R_{t+1} - V_t(S_t)$ is the TD error at time step t.

The second term $\gamma(R_{t+2} - V_{t+1}(S_{t+1}))$ involves the TD error at time step $t+1$, where $R_{t+2} - V_{t+1}(S_{t+1})$ is the TD error at time step $t+1$.

And continuing this pattern until the n-th term:

This leaves us with a sum of TD errors:

$$\sum_{k=t}^{t+n-1} \gamma^{k-t} \delta_k$$

where $\delta_k = R_{k+1} - V_k(S_k)$ is the TD error at time step k. This demonstrates that the n-step error can be expressed as a sum of TD errors when the value estimates don't change from step to step.