Quiz

## Markov Decision Process

1. For a **markov process** with the following transition matrix,

|  |  |  |  |
| --- | --- | --- | --- |
|  | s1 | s2 | s3 |
| s1 | 0.1 | 0.2 | 0.3 |
| s2 | 0.7 | 0.5 | 0.4 |
| s3 | 0.6 | 0.8 | 0.9 |

What are the values of the following transition probabilities: **P(St+1=s1 | St=s3),** **P(St+1=s2 | St=s3) and P(St+1=s2 | St=s1, St-1=s1)**?

1. **P(St+1=s1 | St=s3)** = 0.6
2. **P(St+1=s2 | St=s3) =** 0.8
3. **P(St+1=s2| St=s1, St-1=s1) =** 0.2
4. **In an MDP, on which of the following does the state St+1 depend on? (Choose one or more that apply)**
   1. **St**
   2. **St-1**
   3. **At**
   4. **At-1**
   5. **Rt**
   6. **Rt+1**
5. **What is the difference between “one step reward" and “long term return”?**

|  |  |
| --- | --- |
| **One step reward** | **Long term Return** |
| It is the immediate feedback or reinforcement that the agent receives after taking a single action in a given state. | Considers the sum of rewards the agent receives over an entire sequence of actions and states. |
| It focuses only on the immediate consequence of the agent's action in the current state without considering the future consequences. | It involves considering not only the immediate reward but also the expected future rewards, often discounted over time to give more importance to immediate rewards. |

1. **What happens when the discount factor (𝛄) is set to 0 and when it is set to 1?**  
   The discount factor, 𝛾, is a real value ∈ [0, 1], cares for the rewards agent achieved in the past, present, and future. In different words, it relates the rewards to the time domain.
2. If 𝛾 = 0, the agent cares for his first reward only.
3. If 𝛾 = 1, the agent cares for all future rewards.
4. **For episodic tasks, should the discount factor always be set to 0? Why or why not, give an example.**

No, the discount factor in episodic tasks should not always be set to 0. The discount factor affects how much weight it gives to future rewards in the value function. A discount factor γ = 0 will result in state/action values representing the immediate reward, while a higher discount factor γ = 0.9 will result in the values representing the cumulative discounted future reward an agent expects to receive (behaving under a given policy).

The convergence is influenced by the discount factor depending on whether it’s a continual task or an episodic one. In a continual one, γ must be between [0, 1), whereas an episodic one it can be between [0, 1.]

**Example:** For example, consider a board game where the goal is to reach a certain state to win. If the discount factor is set to 0, the agent would only focus on winning in the current move and wouldn't consider the future moves that might lead to victory. On the other hand, setting a reasonable discount factor enables the agent to consider the importance of future moves and choose actions that contribute to the overall goal.

In many applications it is typically seen to be arbitrarily set to 0.9**.**

1. **Is long term return Gt a random variable? Explain.**

Yes, the long-term return Gt in reinforcement learning is indeed a random variable.

In reinforcement learning, the goal of an agent is to maximize the cumulative reward it receives in the long run. This cumulative reward is defined as the return, denoted Gt. The simplest way to express the return Gt is the sum of all future rewards the agent will receive.

The randomness in Gt comes from the stochastic nature of the rewards Rt and the states St, which are influenced by the agent's actions and the environment's responses.

Even though the return at time t is a sum of an infinite number of terms, the return is actually finite as long as the reward is nonzero and constant, and the discount factor γ < 1.