MDP

Markov Assumption

In reinforcement learning, the Markov assumption states that the future state of the environment depends only on the current state and the current action, and not on the past history of states and actions. In other words, the past history of the agent's interactions is irrelevant for predicting the future, as long as the current state is known.

Markov Process (or Markov Chain)

- A Markov Process is a tuple (S, P) where:
    - S is a finite set of states.
    - P is a state transition probability matrix.

State Transition Probability

- The probability of transitioning from state s to state s' is denoted as P(s', s).
- This probability represents the likelihood of moving from one state to another in a single step.

States Transition Matrix

- The state transition matrix P is a square matrix where each element P(s', s) represents the probability of transitioning from state s to state s'.
- The matrix dimensions are equal to the number of states in the Markov Process.

Markov Reward Process (MRP)

is an extension of a Markov Process that incorporates rewards.

Components of an MRP

- **State Space (S):** A finite set of possible states.
- **Transition Probability Matrix (P):** A matrix that defines the probability of transitioning from one state to another.
- **Reward Function (R):** A function that maps each state to a reward.

Markov Decision Process (MDP)

MDP is a mathematical framework used to model sequential decision-making problems where the future state depends only on the present state and the current action.

Components of an MDP

- An MDP is defined as a tuple (S, A, P, R, γ), where:
    - **S:** A finite set of states.
    - **A:** A finite set of actions.
    - **P:** A state transition probability matrix, where $P(s'|s, a)$ represents the probability of transitioning from state s to state s' by taking action a.

  - **R:** A reward function, where R(s, a) represents the reward received for taking action a in state s.
  - **γ:** A discount factor between 0 and 1 that determines how much future rewards are valued compared to immediate rewards.

Cumulative Reward

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Breakdown of the terms

- **G_t:** The discounted cumulative reward from time step t onwards.
- **R_{t+1}, R_{t+2}, R_{t+3}, ...:** The rewards received at future time steps t+1, t+2, t+3, etc.
- **γ:** The discount factor, a value between 0 and 1 that determines how much future rewards are valued compared to immediate rewards.
- **Σ:** The summation symbol, indicating that we sum up all the discounted rewards from time step t+1 to infinity.