Reinforcement Learning: Markov-Decision Process

Mitali Meratwal

Part 1

RL elements:

- Agent: Software programs that makes decisions, are the learners, interact with the environment by actions and receive rewards based on these actions.
- Environment: surrounding with which agent interacts (demonstration of the problem to be solved.)
  - In return provides rewards and next state to the agent depending on the action taken.
- State: Position of the agent at a specific time-step in the environment.
- Actions: Any decision we want the agent to learn.
- Rewards: Numerical values that the agent receives on performing some action at some state(s) in the environment, can be positive or negative based its actions.
- Returns: Cumulative reward that the agent wants to maximise.

The Markov property states that the future is independent of the past given the present.

- Transition: Moving from one state to another.
- Transition Probability: The probability that the agent will move from one state to another.

$$P(S_{t+1}|S_t) = P(S_{t+1}|S_1, ..., S_t)$$

State Transition Probability: For Markov State from S[t] to S[t+1] i.e. any other successor state , the state transition probability is given by:

$$P_{ss'} = P(S_{t+1} = s'|S_t = s)$$

Markov Process or Markov Chain:
Memory less random process i.e. a sequence of a random state S[1],S[2],. . . .S[n] with a Markov Property and defined using a set of states(S) and transition probability matrix (P).

Episodic tasks:
Tasks that have a terminal state that is finite states.

Continuous tasks:
Tasks that have no end or no terminal state.

Discount factor (γ):
Determines how much weight is to be given to the immediate and future rewards. $\gamma = 0$ implies more importance to the immediate reward and $\gamma = 1$ implies more importance to future reward. Returns equation for continuous task is as follows:

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

Markov Reward process (MRP): Reward from the current state.
(S,P,R,γ) := (State,Transition Probability matrix, Reward function, Discount factor)

$$R_s = E(R_{t+1}|S_t)$$

Policy: What action to perform in a particular state. Defines a probability distribution over Actions (a $\in$ A) for each state (s $\in$ S). Policy funciton is defined as:

$$\pi(a|s) = P(A_t = a|S_t = s)$$

State Value function: How good it is for the agent to be in a particular state which depends on the action it takes. It is the expected returns starting from state(s) and going to successor states thereafter, with the policy π.

$$v_\pi(s) = E_\pi[G_t|S_t = s]$$

Bellman Equation: to find optimal policies and value function.

$$v(s) = E(R_{t+1} + \gamma v(S_{t+1})|S_t = s)$$
$$v(s) = R_s + \gamma \sum_{s' \in S} P_{ss'} v(s')$$

Markov Decision process: Markov Reward Process with decisions.

Static Value function or Q-function: How good it is for the agent to take action (a) in a state (s) with a policy π.

$$q_\pi(s, a) = E_\pi[G_t|S_t = s, A_t = a]$$