Definition

- State $s \in S$
- Action $a \in A$
- reward $r \in \mathbb{R}$
- Transition model $Pr(s_{t+1}|s_t, a_t)$
- Reward model $Pr(r|s_{t+1}, s_t, a_t)$
- Discount Factor $\gamma \in [0, 1]$
- Horizon T

Goal is to find a policy $\pi(a|s)$ that maximises the expectation of discounted return.

How RL differs from MDP solutions

- No Transition Model
- No Reward Model

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r_t \right]$$

However, we still solve the MDP problem using RL by interacting with the environment by learning the transition and reward models or directly learning the policy.

Types of RL algorithms

- Model Based- if we try to learn the transition and reward models
- Model Free - here, we don't learn any model dynamics. No transition and reward models. Below are the types of model free RL algorithms
- Value Based- if we try to learn the value function $V(s)$ of the state or value function of state-action pair $Q(s, a)$.
- Policy Based- if we try to learn the policy $\pi(a|s) - \pi(s, a)$ directly.
- Policy Gradient- if we try to learn the policy $\pi(a|s) - \pi(s, a)$ directly using gradient ascent.
- Actor Critic - contains both policy $\pi(a|s) - \pi(s, a)$ and value function $V(s) - Q(s, a)$.