An Introduction to reinforcement learning

Reinforcement Learnings
Reinforcement learning is learning what to do—how to map situations to actions—so as to maximize a numerical reward signal.

Markov Decision Process

RL problems can be mathematically formulated as a finite Markov Decision Process(MDP). This is one approach to formulate a reinforcement learning problem.
Finite MDPs can be solved by multiple methods: dynamic programming, Monte Carlo method, Temporal difference methods.

- **Agent**: The learner and decision maker is called the agent.
  Ex, a self-driving car, a house cleaning robot, etc.
- **Environment**: Everything outside the agent is called the environment. It is the surroundings the Agent interacts with.
  Ex, road, warehouse, etc.
- **State**: state as a signal conveying to the agent some sense of "how the environment
  is" at a particular time.
  Ex, position/orientation of a robot, climate of a particular day, etc.
- **Action**: It is the decision the Agent takes at a particular time.
  Ex, move forward, lift something, get back to the charging point, etc

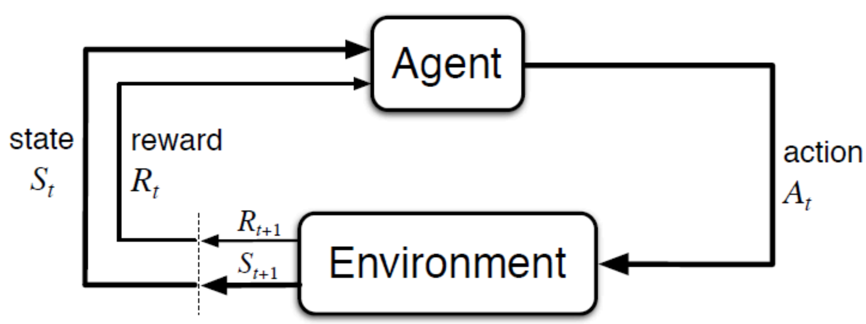

Figure 3.1: The agent–environment interaction in a Markov decision process.

Reward (R_t)

- The numerical signal that the agent receives from the environment at each time step is called the reward.
- Agent's goal is to maximize the total amount of reward it receives. This means maximizing not immediate reward, but cumulative reward in the long run.
- We must provide rewards to it in such a way that in maximizing them the agent will achieve the final goal.

Return (G_t)

It is the total reward that the Agent receives over a long run.

$$G_t \doteq R_{t+1} + R_{t+2} + R_{t+3} + \cdots + R_T,$$

Discounting

The agent tries to select actions so that the sum of the discounted rewards it receives over the future is maximized. In particular, it chooses A_t to maximize the expected discounted return

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

where γ is a parameter, $0 \leq \gamma \leq 1$, called the *discount rate*.

As γ approaches 1, the return objective takes future rewards into account more strongly; the agent becomes more farsighted.

Value functions

functions of states (or of state–action pairs) that estimate how good it is for the agent to be in a given state (or how good it is to perform a given action in a given state).

- State value functions

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \text{ for all } s \in \mathcal{S},$$

- Action value functions

$$q_\pi(s,a) \doteq \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s, A_t = a\right].$$

Policy

a policy is a mapping from states to probabilities of selecting each possible action.
If the agent is following policy π at time t, then $\pi(a|s)$ is the probability that $A_t = a$ if $S_t = s$.