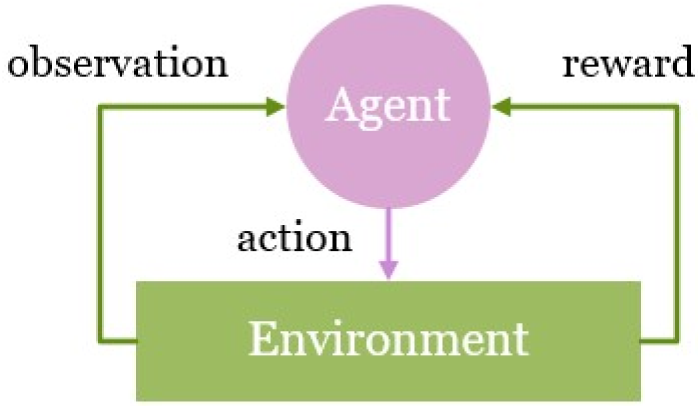
Reinforcement Learning

Reinforcement Learning (RL) is a standard framework to achieve target in Markov Decision Process.

In a MDP $< O, A, P, \gamma, R >$,

- At each time period, the environment is in a state s, and agent in environment receives local observation o based on s.
- Agent takes action a, and receives a local reward from environment r ($R : S \times A \to \mathbb{R}$). Then environment moves to the next state. The process repeats.
- P is the state transition function, $P(s, a, s')$
- γ is the discount factor
- MDP can be represented as $< o_0, a_0, r_1, o_1, a_1, \ldots, o_{t-1}, a_{t-1}, r_t >$.
- In RL process, agent gets a series of sample and improve its policy to get better reward.



We can define the total discounted reward:

$$\mathcal{R} = \sum_{t=0}^{\infty} \gamma^t R_{t+1}$$

Agent's value function:

- Action value function: expected total discounted reward after taking action a in state s

  $$Q^\pi(s, a) = \mathbb{E}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots | s, a\right]$$

  which means,

  $$Q^\pi(s, a) = E\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | s_0 = s, a_0 = a\right]$$

- It is easy to derive value function:

  $$V^\pi(s) = E\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | s_0 = s\right]$$

Agent's policy: $\pi : S \to A$, a map from state to action

- Deterministic policy: $a = \pi(s)$
- Stochastic policy: $\pi(a|s) = \mathbb{P}[a|s]$, the probability of a as the selected action in state s

Now, our goal is to maximize global reward, which means maximize value function.