RL: Learn by interacting with the environment

- more close to the nature of learning (compared with supervised learning)

- no explicit teacher

- Discover which action yield the most reward by trying it

$S_t \xrightarrow{a_t} S_{t+1}, r_{t+1}$

$S_{t+1}$: stochastic; $r_{t+1}$: stochastic & delayed.

e.g. design the reward or stochastic reward

Formulation: $\mathbb{S}$,

$\mathbb{A}$,

$\mathbb{P}$, $p(S_{t+1} \mid S_t, a_t)$ ⇐ distribution of the next state given the current state & action

$r: \mathbb{S} \times \mathbb{A} \to \mathbb{R}$

$\gamma$: discount factor $\gamma \in (0, 1]$

Goal: $\max \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r_i \mid S_0 = s\right]$

Difficulty:
- $\gamma$ close to 1, take more future reward into account
- The optimal action may not be the one who has the highest immediate reward
- After I apply an action, do not know what the next state is.
- the reward could be a random variable

Outline:

§1. MAB

1). UCB
- Regret Thm

2). Thompson Sampling

3) Optimal - Bayes bandit

§2. MDP

1) BE, TD
- SA Thm

TD($\lambda$)

2). Policy Iteration

3). SARSA, Q-learning
- Q-learning convergence Thm

4). Policy $\nabla$
- Policy $\nabla$ Thm

TRPO, PPO

5). Optimization, Double sampling, Primal-dual, BFT

§3. Optimal control

1). Optimal control $\Leftrightarrow$ HJB.
- Convergence Thm.

2) LQR
- Why Linear control & quadratic value
- How to derive the Ricatti Eq.