Lecture 1

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.

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

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

γ: 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:
- γ 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

RL: $\mathcal{S}, \mathcal{A}, P, r, \gamma$ (P: unknown) MDP: $\mathcal{S}, \mathcal{A}, P, r, \gamma$ (P: known)

MAB: $\mathcal{A}, r$

- UCB → Thm 1: upper bd for regret
- Bayes-optimal

LP: $\mathcal{S}, \mathcal{A}, P, r, \gamma$ ($\mathcal{S}, \mathcal{A}$: finite; P: known)

primal ↕ Bellman, Dual ↕ policy ∇, primal-dual → Thm 2 ↑ equivalence.

RL: . TD → Thm 3: convergence

. GTD, policy ∇, AC,

HJB: continuous-time control with known dynamics.

. HJB ⟺ optimal control → Thm 4.