

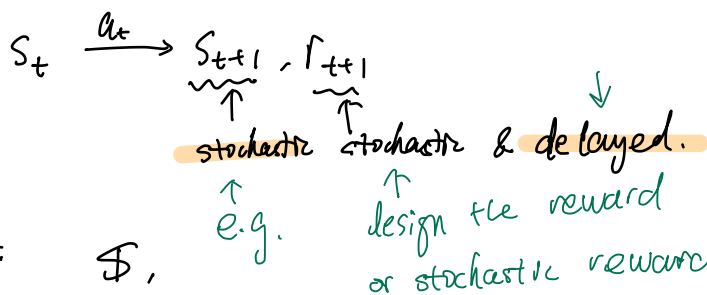
RL: Learn by interacting with the environment

- more close to the nature of learning

(Compared with supervised learning)

- no explicit teacher

- Discover which action yield the most reward by trying it



Formulation:

\mathcal{S} ,
 \mathcal{A} ,

\mathcal{P} , $P(S_{t+1} | S_t, a_t) \leftarrow$ distribution of the next state given the current state & action

$r: \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$

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

Outline:

§ 1. MAB

1). UCB

- Regret Thm

2). Thompson Sampling

3). Optimal-Bayes bandit

§ 2. MDP

1). BE, TD

- SA Thm

TD(λ)

2). Policy Iteration

3). SARSA, Q-learning

- Q-learning convergence Thm

4). Policy ∇

- Policy ∇ Thm

TRPO, PPO

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

§ 3. Optimal control

1). Optimal control \Leftrightarrow HJB.

- Convergence Thm.

2). LQR

- Why Linear control & quadratic value

- How to derive the Riccati Eq.