

Reinforcement Learning

① Problem Setting (DSA5105 convention)

====> MDP $\langle S, A, P, R, \gamma \rangle$

a) State Space : S

$$S_t \in S$$

b) Action Space: A

$$A_t \in \mathcal{A}(S_t)$$

c) Transition - Reward Probability : P

$$\underline{P(S', R | s, a)} = P(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a)$$

\curvearrowright joint distribution for $(S_{t+1}, R_{t+1}) | S_t = s, A_t = a$

Quantity $\Rightarrow p(s' | s, a) = \sum_r p(s', r | s, a) := P(S_{t+1} = s' | S_t = s, A_t = a)$

d) Reward Space : R

$$R_t \in \mathbb{R}$$

e) Discount Rate : γ

====> Policy: $\pi(a|s) := P(A_t = a | S_t = s)$

$$\Leftrightarrow \pi(\cdot | \cdot) : \mathcal{A} \times \mathcal{S} \longrightarrow [0,1]$$

Note: $\pi(\cdot | s) \rightsquigarrow$ probability distribution

$$\Rightarrow A_t \sim \pi(\cdot | S_t)$$

Quantity $\Rightarrow P(\pi)_{ss'} := P^\pi(S_{t+1} = s' | S_t = s)$

$$= \sum_a P^\pi(S_{t+1} = s', A_t = a | S_t = s)$$

$$= \sum_a \pi(a|s) \cdot P(s'|s, a)$$

Quantity

$$\xrightarrow{\textcircled{2}} b(\pi)_s = \underline{E^\pi [R_{t+1} | S_t = s]}$$

$$= \sum_r r \cdot P^\pi(R_{t+1} = r | S_t = s)$$

$$= \sum_r r \cdot \sum_{s'} P^\pi(R_{t+1} = r, S_{t+1} = s' | S_t = s)$$

$$= \sum_{r, s'} r \cdot P^\pi(R_{t+1} = r, S_{t+1} = s' | S_t = s)$$

$$= \sum_{r, s'} r \cdot \sum_a \pi(a|s) \cdot P(s', r | s, a)$$

$$= \sum_a \pi(a|s) \sum_{r, s'} r P(s', r | s, a)$$

$$\xrightarrow{\textcircled{2}} \underbrace{\text{Return: } G_t = \sum_{z=0}^{\infty} \gamma^z R_{t+z+1}}$$

$$\downarrow \quad \text{Our interest } E^\pi [G_0 | S_0 = s]$$

$\xrightarrow{\textcircled{2}} \star \underbrace{\text{DP Framework}}$

① State-Value Function

$$V_\pi(s) := E^\pi [G_t | S_t = s]$$

Relatively trivial

relationship

② Action-Value Function

$$q_{\pi}(s, a) := \mathbb{E}^{\pi}[G_t \mid S_t = s, A_t = a]$$

③ Bellman Equation (Relation)

$$\begin{aligned} 1. \quad V_{\pi}(s) &= \mathbb{E}^{\pi}[G_t \mid S_t = s] \\ &= \mathbb{E}^{\pi}[\mathbb{E}^{\pi}[G_t \mid A_t, S_t = s]] \\ &= \sum_a \pi(a|s) \cdot q_{\pi}(s, a) \end{aligned}$$

$$\begin{aligned} 2. \quad q_{\pi}(s, a) &= \mathbb{E}^{\pi}[G_t \mid S_t = s, A_t = a] \\ &= \mathbb{E}^{\pi}[R_{t+1} + \gamma \underbrace{\sum_{t=0}^{\infty} \gamma^t R_{t+1+t}}_{G_{t+1}} \mid S_t = s, A_t = a] \\ &= \mathbb{E}^{\pi}[R_{t+1} \mid S_t = s, A_t = a] + \gamma \sum_{s'} p(s'|s, a) \cdot V_{\pi}(s') \\ &= \underbrace{\sum_{r, s'} r \cdot p(s', r \mid s, a)}_{r(s, a)} + \gamma \sum_{s'} p(s' \mid s, a) V_{\pi}(s') \end{aligned}$$

$$r(s, a) := \mathbb{E}[R_{t+1} \mid S_t = s, A_t = a]$$

$$\begin{aligned} &= \sum_{r, s'} r p(s', r \mid s, a) \\ &= \sum_{r, s'} p(s', r \mid s, a) [r + \gamma V_{\pi}(s')] \\ &= b(\pi)_s \quad \boxed{\pi(s) = a} \end{aligned}$$

$$\begin{aligned} 3. \quad V_{\pi}(s) &= \mathbb{E}^{\pi}[R_{t+1} + \gamma G_{t+1} \mid S_t = s] \\ &= \mathbb{E}^{\pi}[R_{t+1} \mid S_t = s] + \gamma \mathbb{E}^{\pi}[G_{t+1} \mid S_t = s] \\ &= b(\pi)_s + \gamma \sum_{s'} p(\pi)_{ss'} \mathbb{E}^{\pi}[G_{t+1} \mid S_{t+1} = s'] \end{aligned}$$

$$= b(\pi)_s + \gamma \sum_{s'} p(\pi)_{ss'} V_{\pi}(s')$$

$$\Rightarrow \boxed{V_{\pi} = b(\pi) + \gamma P(\pi) \cdot V_{\pi}} \rightarrow \boxed{\text{Matrix Form}}$$

When given the policy π , calculate

$$\begin{cases} b(\pi) \in \mathbb{R}^{1 \times 1} \\ P(\pi) \in \mathbb{R}^{1 \times 1} \end{cases}$$

then solve for $V_{\pi} := \mathbb{E}^{\pi}[G_t \mid S_t]$ \Leftrightarrow solve Linear Equation

④ Bellman Optimal Equation \leftarrow Non-trivial relationship for optimality

a) Defn of optimal policy π^*

π^* is optimal policy

$$\Leftrightarrow \forall s \in S, V_{\pi}(s) \leq V_{\pi^*}(s)$$

$\forall \pi \in \text{Policy}$

b) Try to solve for the Optimal Policy π^*

Bellman Equation

1. Policy Improvement [Note: $V_{\pi}(s) = \sum_a \pi(a|s) \cdot q_{\pi}(s,a)$]

given $q_{\pi}(s,a)$, then for any two policy π, π'

$$\text{if } \sum_a \pi'(a|s) q_{\pi}(s,a) \geq \sum_a \pi(a|s) q_{\pi}(s,a) \quad \forall s \in S$$

then we must have $V_{\pi'}(s) \geq V_{\pi}(s) \quad \forall s \in S$

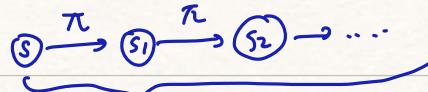
Pf Sketch :

$$\begin{aligned} V_{\pi}(s) &= \sum_a \pi(a|s) q_{\pi}(s,a) \\ &\leq \sum_a \pi'(a|s) q_{\pi}(s,a) \\ &= b(\pi')_s + \gamma \sum_{s'} P(\pi')_{ss'} V_{\pi}(s') \end{aligned}$$

$\Rightarrow \pi' \succcurlyeq \pi$ ★
 $V_{\pi} \leq b(\pi') + \gamma P(\pi') V_{\pi}$
 $\leq b(\pi) + \gamma P(\pi') b(\pi')$
 $\leq \dots \leq (1 - \gamma P(\pi'))^{-1} b(\pi')$
 $= V_{\pi'}$

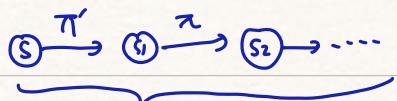
Intuitively:

$$V_{\pi}(s) = \mathbb{E}^{\pi} [G_t | S_t=s]$$

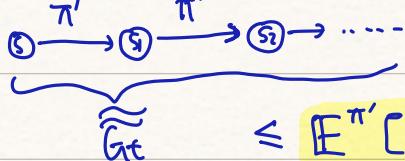


\prod

$\forall s \in S$



\prod



$$\leq \mathbb{E}^{\pi'} [G_t | S_t=s] = V_{\pi'}(s)$$

$$\leq \mathbb{E}^{\pi'} \left[\sum_{t=0}^K \gamma^t R_{t+1} + \gamma^{K+1} \cdot V_{\pi}(s_{K+1}) \mid S_t=s \right]$$

$K \rightarrow +\infty$
 $V_{\pi'}(s)$

$$\leq \mathbb{E}^{\pi'} [R_{t+1} | S_t=s] + \mathbb{E}^{\pi'} [\gamma R_{t+2} | S_t=s] + \gamma^2 \sum_{s'} P(\pi')_{ss'} V_{\pi}(s')$$

2-step transition probability

Note: if ' $>$ ' appears in the assumption (inequality strictly holds for some s),

then $V_{\pi}(s) < V_{\pi'}(s)$ for some $s \in S$

2. Bellman Optimality Condition

(Necessary)

if π^* is an optimal policy,

then ① $\pi^*(a|s) > 0 \Leftrightarrow a \in \operatorname{argmax}_{a' \in A} q_{\pi^*}(a', s)$

$$② V_{\pi^*}(s) = \sum_a \pi^*(a|s) \cdot q_{\pi^*}(a, s)$$

Bellman Optimality Equation

$$= \max_{a \in A} q_{\pi^*}(a, s)$$

$$V_{\pi^*}(s) = \max_{a \in A} \sum_{r, s'} p(s', r | s, a) \cdot r + \gamma \sum_{s'} p(s' | s, a) \cdot V_{\pi^*}(s')$$

$E[R_{t+1} | S_t=s, A_t=a]$

$E^{\pi^*}[R_{t+2} + \dots | S_t=s, A_t=a]$

implies the sufficiency optimality condition

3. Uniqueness of Bellman Optimality Equation

There exists a unique V^* satisfies Bellman Optimality Equation

$$\text{i.e., } V^*(s) = \max_a \sum_{s', r} p(s', r | s, a) \cdot r + \gamma \sum_{s'} p(s' | s, a) \cdot V^*(s')$$

$$\text{PF Sketch : } T(V^*) := \max_a (R_s^a + \gamma P_{ss'}^a \cdot V^*)$$

$$a_1 = \operatorname{argmax}_a (R_s^a + \gamma P_{ss'}^a \cdot u_1)$$

$$\text{Idea } \Rightarrow T(u_1) - T(u_2) \leq R_s^{a_1} + \gamma P_{ss'}^{a_1} u_1 - R_s^{a_1} - \gamma P_{ss'}^{a_1} u_2$$

$$= \gamma P_{ss'}^{a_1} (u_1 - u_2)$$

$$\Rightarrow \|T(u_1) - T(u_2)\|_\infty \leq \gamma \|u_1 - u_2\|_\infty$$

Contraction Mapping theorem

\implies Corollary: (Sufficient Bellman Optimality Condition)

If $\stackrel{(1)}{\pi}$ is a policy $\implies \begin{cases} q_{\pi}(s, a) \\ V_{\pi}(s) \end{cases}$

Either

$\stackrel{(2)}{\begin{cases} \pi(a|s) > 0 \implies a \in \operatorname{argmax}_{a' \in A} q_{\pi}(a'|s) \\ V_{\pi}(s) = \max_a \sum_{s' \sim r} p(s', r|s, a) \cdot r + \gamma \sum_{s'} p(s'|s, a) \cdot V_{\pi}(s') \end{cases}}$ implies

$$V_{\pi}(s) = \max_a \sum_{s' \sim r} p(s', r|s, a) \cdot r + \gamma \sum_{s'} p(s'|s, a) \cdot V_{\pi}(s')$$

then π is an optimal policy !!!



4. Bellman Optimality Condition

Summary

\rightarrow a policy π is optimal policy

$\Leftrightarrow \begin{cases} \pi(a|s) > 0 \implies a \in \operatorname{argmax}_{a' \in A} q_{\pi}(s, a') \\ V_{\pi}(s) = \max_{a \in A} \sum_{s' \sim r} p(s', r|s, a) \cdot r + \sum_{s'} p(s'|s, a) \cdot V_{\pi}(s') \end{cases}$

$$V_{\pi}(s) = \max_{a \in A} \sum_{s' \sim r} p(s', r|s, a) \cdot r + \sum_{s'} p(s'|s, a) \cdot V_{\pi}(s')$$

Logic: $\stackrel{(1)}{\pi}$ optimal $\implies \pi(a|s) > 0$ implies $a \in \operatorname{argmax}_{a' \in A} q_{\pi}(s, a')$

$\implies V_{\pi}(s)$ satisfies Bellman Optimality Equation

$\stackrel{(2)}{\begin{cases} V_{\pi}(s) \text{ satisfies Bellman Optimality Equation} \\ \text{uniqueness solution of B.O.E} \end{cases}} \implies \pi \text{ optimal}$

$\implies \pi(a|s) > 0 \text{ implies } a \in \operatorname{argmax}_{a' \in A} q_{\pi}(s, a')$ $\left. \begin{array}{l} \text{uniqueness solution of B.O.E} \end{array} \right\} \implies \pi \text{ optimal}$

5. Existence of Optimal Deterministic Policy (finite MDP)

\rightarrow For finite MDP, it is sufficient to consider deterministic policy

$$\begin{cases} |A| < \infty \\ |S| < \infty \end{cases}$$

6. Some Notation for Finite MDP

↓

optimal deterministic always exists!

$$1. \begin{cases} V^*(s) := V_{\pi^*}(s) = \max_{\pi} V_{\pi}(s) \\ q^*(s, a) := q_{\pi^*}(s, a) = \max_{\pi} q_{\pi}(s, a) \end{cases}$$

2. if we consider deterministic optimal policy π^*

$$\text{then } \pi^*(s) \in \operatorname{argmax}_{a \in A} q^*(s, a)$$

3. Bellman Optimality Equation

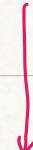
$$\begin{cases} V^*(s) = \max_a q^*(s, a) = \max_a \sum_{s', r} p(s', r | s, a) r + \gamma \sum_{s'} p(s' | s, a) V^*(s') \\ q^*(s, a) = \sum_{s', r} p(s', r | s, a) \cdot r + \gamma \sum_{s'} p(s' | s, a) \cdot \max_{a'} q^*(s', a') \end{cases}$$

② Algorithm to solve for Optimal Value Function

$V^*(s) \quad s \in S$



Optimal Action-Value Function $q^*(s, a)$



Optimal Policy $\pi^*(s) \in \operatorname{argmax}_a q^*(s, a)$

$p(s', r | s, a)$ is a known function!!

Model-based \Rightarrow need information about environment $p(s', r | s, a)$

Model-free \Rightarrow just do the simulation

sampling from the ground-truth $p(s', r|s, a)$

1. Model-based Algorithm

a) Value Iteration

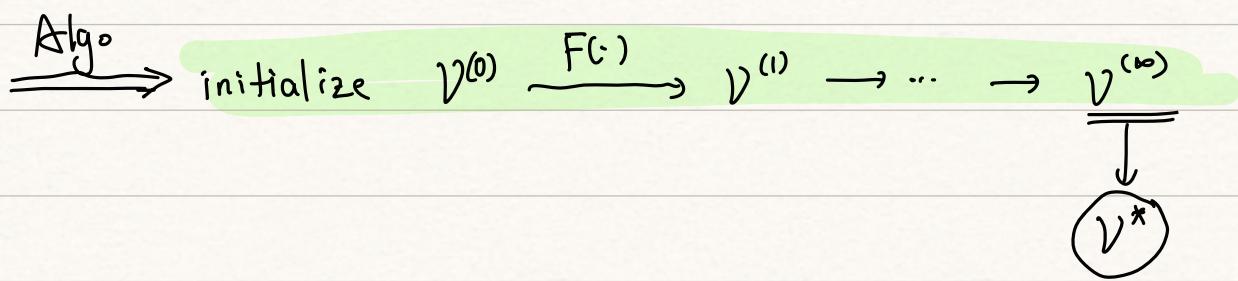
key point $\rightarrow V^*(s) = \max_a \sum_{s',r} p(s',r|s,a)r + \gamma \sum_{s'} p(s'|s,a) \cdot V^*(s')$

$$\Rightarrow \underline{V^* = F(V^*)}$$

$$\text{where } F(v) = \max_{\pi \text{ deterministic}} [\gamma P(\pi) \cdot v + b(\pi)]$$

for deterministic policy

$$\left\{ \begin{array}{l} P(\pi)_{ss'} = \sum_a \pi(a|s) \cdot p(s'|s,a) = p(s'|s, \pi(s)) \\ b(\pi)_s = \mathbb{E}^\pi[R_{t+1} | S_t = s] \\ = \sum_a \pi(a|s) \cdot \sum_{r,s'} p(s',r|s,a) \cdot r \\ = \sum_{r,s'} p(s',r|s, \pi(s)) \cdot r \end{array} \right.$$



b) Policy Iteration

Key point



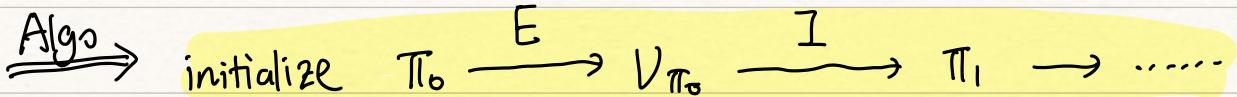
{

- Policy improvement
- finite deterministic policy in finite MDP

↓ *

→ if $\pi'(s) \in \operatorname{argmax}_a q_{\pi}(s, a)$ for all $s \in S$

then $V_{\pi'} \geq V_{\pi}$ (at least as good as)



E: Bellman Equation:

$$V_{\pi_0} = b(\pi_0) + \gamma P(\pi_0) \cdot V_{\pi_0}$$

evaluation step

might be computationally expensive

I: Policy improvement:

$$\pi_1(s) = \operatorname{argmax}_a q_{\pi_0}(s, a) = \operatorname{argmax}_a \left[\sum_{s', r} p(s', r | s, a) r + \gamma \sum_{s'} p(s' | s, a) \cdot V_{\pi_0}(s') \right]$$

$$\pi_1 = \operatorname{argmax}_{\pi \text{ deterministic}} b(\pi) + \gamma P(\pi) \cdot V_{\pi_0}$$

improvement Step

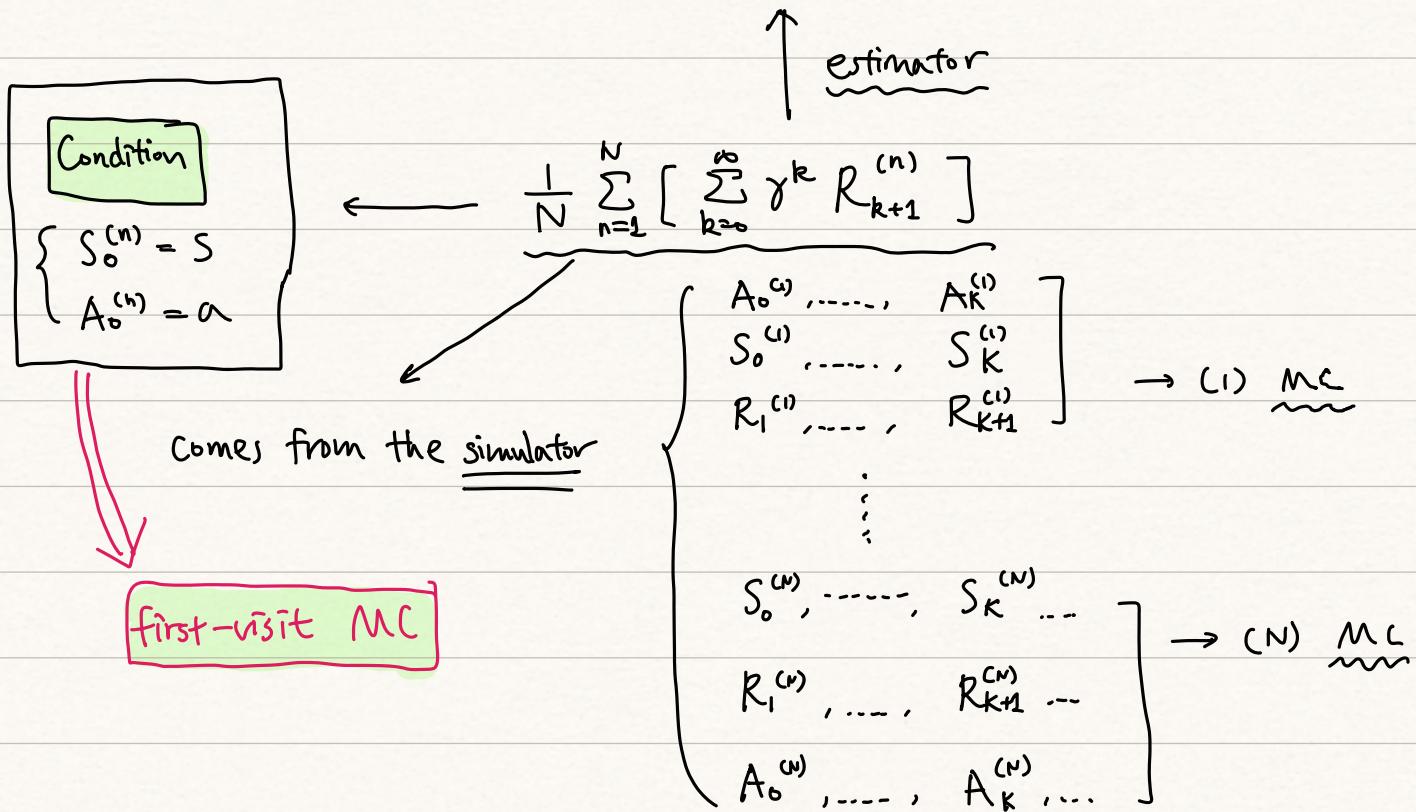
2. Model-free algorithm

Generally speaking, there are

<u>MC based</u>	<u>TD based</u>
-----------------	-----------------

a) MC-based algo ⇒ for policy evaluation without transition probability

$$\text{Idea: } q_{\pi}(s, a) = \mathbb{E}_{\pi} [G_t | S_t = s, A_t = a]$$



b) TD - Based Algo \rightsquigarrow Q-learning

Limitation of MC based algo \rightarrow

- ① Computationally intractable
- ② Variance issue when state space is very large

Idea of TD-Based Algo



$$a_{k+1} = (1-\alpha)a_k + \alpha b \Rightarrow \lim_{k \rightarrow \infty} a_k = b$$

$$= a_k + \alpha(b - a_k)$$

Temporal Difference

Take $V_\pi(s)$ as one example (if our interest is $V_\pi(s)$)

$$\begin{aligned}
 V_\pi(s) &= \mathbb{E}_\pi [G_t | S_t = s] \\
 &= \mathbb{E}_\pi [R_{t+1} + \gamma G_{t+1} | S_t = s] = b(\pi)_s + \sum_{s'} P(\pi)_{ss'} V_\pi(s') \\
 &= \mathbb{E}_\pi [R_{t+1} + \gamma V_\pi(S_{t+1}) | S_t = s] \\
 &\approx r + \gamma V_\pi(s')
 \end{aligned}$$

$$\Rightarrow V_{k+1}(s) = V_k(s) + \alpha [r + \gamma V_k(s') - V_k(s)]$$

Actually, we will focus on Q -function for the convenience of policy improvement!

Q -Learning:

$$Q(s, a) \leftarrow Q(s, a) + \alpha [r + \gamma \max_{a'} Q(s', a') - Q(s, a)]$$

off-policy
value iteration