Reinforcement Learning Recap

① **Problem Setting** (DSA5105 convention)

$\Longrightarrow$ MDP $\langle S, \mathcal{A}, P, R, \gamma \rangle$

a) State Space : S

$S_t \in S$

b) Action Space: $\mathcal{A}$

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

c) Transition - Reward Probability : P

$p(s', r \mid s, a) = \mathbb{P}(S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a)$

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

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

d) Reward Space : $\mathcal{R}$

$R_t \in \mathcal{R}$

e) Discount Rate : γ

$\Longrightarrow$ Policy : $\pi(a \mid s) := \mathbb{P}(A_t = a \mid S_t = s)$

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

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

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

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

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

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

Quantity $\Longrightarrow$ ② $b(\pi)_s = \mathbb{E}^\pi [R_{t+1} \mid S_t = s]$

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

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

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

$$= \sum_{r,s'} r \sum_a \mathbb{P}^\pi (R_{t+1} = r , S_{t+1} = s' , A_t = a \mid S_t = s)$$

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

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

$\Longrightarrow$ Return: $G_t = \sum_{\tau=0}^{\infty} \gamma^\tau R_{t+\tau+1}$

↓

Our interest $\mathbb{E}^\pi [G_0 \mid S_0 = s]$

$\Longrightarrow$ ☆ DP Framework

Relatively trivial

① State-Value Function

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

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)

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

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

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

$= \sum_{r,s'} r\, p(s', r|s,a)$

$= b(\pi)_s$ $\boxed{\pi(s) = a}$

$$= \sum_{r,s'} p(s', r|s,a)\,[r + \gamma V_\pi(s')]$$

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

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

$\Downarrow$

when given the policy π, calculate $\begin{cases} b(\pi) \in \mathbb{R}^{|S|} \\ P(\pi) \in \mathbb{R}^{|S| \times |S|} \end{cases}$

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

④ Bellman Optimal Equation ← Non-trivial relationship for optimality

a) Defn of optimal policy π^*

π^* is optimal policy

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

if $\sum_a \pi'(a|s)\, q_\pi(s,a) \ge \sum_a \pi(a|s)\, q_\pi(s,a) \quad \forall s \in S$

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

$\Rightarrow \pi' \ge \pi$

Pf Sketch:

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

$\mathbb{E}^{\pi'}[R_{t+1} | S_t = s]$

$\star$ $V_\pi \le b(\pi') + \gamma P(\pi') V_\pi$

$\le b(\pi') + \gamma P(\pi') b(\pi') + \gamma^2 P(\pi')^2 V_\pi$

$\le \cdots \le (1 - \gamma P(\pi'))^{-1} b(\pi') = V_{\pi'}$

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

G_t: $s \xrightarrow{\pi} s_1 \xrightarrow{\pi} s_2 \to \cdots$

$\forall s \in S$

$\hat{G}_t$: $s \xrightarrow{\pi'} s_1 \xrightarrow{\pi} s_2 \to \cdots$

$\tilde{G}_t$: $s \xrightarrow{\pi'} s_1 \xrightarrow{\pi'} s_2 \to \cdots$

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

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

2-step transition probability

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

$K \to +\infty$

$V_{\pi'}(s)$

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 $\pi_*(a|s) > 0 \Leftrightarrow a \in \underset{a' \in A}{\text{argmax}}\ q_{\pi^*}(a', s)$

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

Bellman Optimality Equation

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

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

$\mathbb{E}[R_{t+1} | S_t = s, A_t = a]$ $\quad$ $\mathbb{E}^{\pi^*}[R_{t+2} + \cdots | S_t = s, A_t = a]$

3. Uniqueness of Bellman Optimality Equation

implies the sufficiency optimality condition

There exists a unique V_* satisfies Bellman Optimality Equation

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

Pf Sketch : $\mathcal{T}(V_*) := \max_a \left(R_s^a + \gamma P_{ss'}^a \cdot V_* \right)$

$a_1 = \underset{a}{\text{argmax}} \left(R_s^a + \gamma P_{ss'}^a u_1 \right)$

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

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

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

Contraction Mapping theorem

$\Longrightarrow$ Corollary: (Sufficient Bellman Optimality Condition)

① If π is a policy $\Longrightarrow$ $\begin{cases} q_\pi(s,a) \\ v_\pi(s) \end{cases}$

② Either

$$\begin{cases} \pi(a|s) > 0 \Rightarrow a \in \underset{a' \in A}{\text{argmax}}\ q_\pi(a', s) \quad \text{(implies)} \\ v_\pi(s) = \max\limits_{a} \sum\limits_{s',r} p(s',r|s,a)\cdot r + \gamma \sum\limits_{s'} p(s'|s,a)\cdot v_\pi(s') \end{cases}$$

then π is an optimal policy !!!

4. Bellman Optimality Condition → Summary

→ a policy π is optimal policy

$\Longleftrightarrow$

$$\begin{cases} \pi(a|s) > 0 \Rightarrow a \in \underset{a' \in A}{\text{argmax}}\ q_\pi(s, a') \\ v_\pi(s) = \max\limits_{a \in A} \sum\limits_{s',r} p(s',r|s,a)\cdot r + \sum\limits_{s'} p(s'|s,a)\cdot v_\pi(s') \end{cases}$$

Logic: ① π optimal $\Rightarrow$ $\pi(a|s) > 0$ implies $a \in \text{argmax}\ q_\pi(s,a')$

$\Rightarrow$ $v_\pi(s)$ satisfies Bellman Optimality Equation

② $v_\pi(s)$ satisfies Bellman Optimality Equation $\Rightarrow$ π optimal

uniqueness solution of B.O.E

$\Rightarrow$ $\pi(a|s) > 0$ implies $a \in \text{argmax}\ q_\pi(s,a')$ $\Rightarrow$ π optimal

uniqueness solution of B.O.E

5. Existence of Optimal Deterministic Policy (finite MDP)

→ For finite MDP, it is sufficient to consider deterministic policy

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

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^*

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

↓

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 ⇒ need information about environment $p(s',r|s,a)$

Model-free ⇒ just do the simulation

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

Real-life case!

1. Model-based Algorithm

a) Value Iteration

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

$$\Rightarrow \quad V_* = F(V_*)$$

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

for deterministic policy

$$P(\pi)_{ss'} = \sum_a \pi(a|s) \cdot p(s'|s,a) = p(s'|s, \pi(s))$$

$$b(\pi)_s = \mathbb{E}^{\pi}[R_{t+1} \mid S_t = s]$$

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

$$= \sum_{r,s'} p(s', r|s, \pi(s)) \cdot r$$

Algo → initialize $v^{(0)} \xrightarrow{F(\cdot)} v^{(1)} \longrightarrow \cdots \longrightarrow v^{(\infty)}$

$v^{(\infty)} \downarrow v^*$

b) Policy Iteration

Key point : →

- policy improvement
- finite deterministic policy in finite MDP

$\Downarrow$ ☆

$\longrightarrow$ if $\pi'(s) \in \underset{a}{\text{argmax}}\ q_\pi(s,a)$ for all $s \in S$

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

Algo $\Rightarrow$ initialize $\pi_0 \xrightarrow{E} V_{\pi_0} \xrightarrow{I} \pi_1 \longrightarrow \cdots\cdots$

E: Bellman Equation:

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

evaluation step

$\rightarrow$ might be computationally expensive

I: policy improvement:

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

$\Downarrow$

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

improvement step

2. Model-free algorithm

Generally speaking, there are { MC based / TD based

a) MC-based algo $\Rightarrow$ for policy evaluation without transition probability

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

↑ estimator

$$\frac{1}{N}\sum_{n=1}^{N}\left[\sum_{k=0}^{\infty}\gamma^k R_{k+1}^{(n)}\right]$$

Condition

$$\begin{cases} S_0^{(n)} = s \\ A_0^{(n)} = a \end{cases}$$

comes from the simulator

first-visit MC

$$\begin{bmatrix} A_0^{(1)}, \dots, A_K^{(1)} \\ S_0^{(1)}, \dots, S_K^{(1)} \\ R_1^{(1)}, \dots, R_{K+1}^{(1)} \end{bmatrix} \rightarrow (1)\ MC$$

$$\vdots$$

$$\begin{bmatrix} S_0^{(N)}, \dots, S_K^{(N)} \dots \\ R_1^{(N)}, \dots, R_{K+1}^{(N)} \dots \\ A_0^{(N)}, \dots, A_K^{(N)}, \dots \end{bmatrix} \rightarrow (N)\ MC$$

b) TD-Based Algo ⟶ Q-learning

Limitation of MC based algo → ① computationally intractable

② variance issue when state space is very large

Idea of TD-Based Algo

↓

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

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

Temporal Difference

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

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

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

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

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

off-policy

value iteration