Reinforcement learning

① Deterministic / Non-deterministic Transition

$f(s,a) = s'$ $\qquad$ $f(s,a,s') = P^a_{ss'}$ $\qquad$ $r(s,a) = \sum_{s'} P^a_{ss'}\, \rho(s,a,s')$

② policy → { deterministic $\pi(s) = a$; non-deterministic $\pi(a|s)$ → probability distribution

Key Question: How to make sense that $\pi^*(s) := \arg\max_a Q^*(s,a)$

③ ✡ Main Topic → Idea: determine Best Policy (optimal) $\qquad$ ↳ $Q^*(a,s) = \max_\pi Q^\pi(a,s)$

{ State-Value Function $V^\pi(s) = \mathbb{E}^\pi[R_t \mid S_t = s]$

Action-Value Function $Q^\pi(s,a) = \mathbb{E}^\pi[R_t \mid S_t = s, A_t = a]$

$R_t = r_{t+1} + \gamma r_{t+2} + \cdots$ Reward (long term)

→ The Relationship Between $V^\pi(s)$ & $Q^\pi(s,a)$

① $V^\pi(s) = \mathbb{E}^\pi[R_t \mid S_t = s]$

$= \mathbb{E}^\pi\left[\mathbb{E}^\pi[R_t \mid S_t = s, A_t]\right]$

$= \sum_a \pi(a|s)\, \mathbb{E}^\pi[R_t \mid S_t = s, A_t = a]$

$= \sum_a \pi(a|s) \cdot Q^\pi(s,a)$

→ when $\pi(\cdot)$ is deterministic,

then $V^\pi(s)$ degenerate to $Q^\pi(s, \pi(s))$

i.e., $V^\pi(s) = Q^\pi(s, \pi(s))$

Generally, $V^\pi(s) = \sum_a \pi(a|s) \cdot Q^\pi(s,a)$

② $Q^\pi(s,a) = \mathbb{E}^\pi[R_t \mid S_t = s, A_t = a]$

$= \sum P^a_{ss'} \cdot \mathbb{E}^\pi[\rho(s,a,s') + \gamma R_{t+1} \mid S_{t+1} = s']$

s'

$$= \sum_{s'} P_{ss'}^{a} \left(r(s,a,s') + \gamma V^{\pi}(s') \right)$$

$$= r(s,a) + \gamma \sum_{s'} P_{ss'}^{a} V^{\pi}(s')$$

When $\pi(\cdot)$ is deterministic, $V^{\pi}(s')$ degenerate to $Q^{\pi}(s', \pi(s'))$

$$\Rightarrow Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s'} P_{ss'}^{a} Q^{\pi}(s', \pi(s'))$$

Generally $Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s'} P_{ss'}^{a} V^{\pi}(s')$

These 2 are **Bellman Equation** !

Then, Back to our goal → find **optimal policy $\pi^*(\cdot)$**

Firstly, consider deterministic case $\pi(\cdot)$ ⇒ easy part

$$\rightarrow \begin{cases} V^{\pi}(s) = Q^{\pi}(s, \pi(s)) \\ Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s'} P_{ss'}^{a} V^{\pi}(s') \end{cases}$$

$Q^{\pi}(s, \pi(s))$

$V^{\pi}(s')$

→ Intuitively, if we have $\pi^*(\cdot)$, then we **want**

$$Q^{\pi^*}(s,a) \geq Q^{\pi}(s,a) \quad \text{for } \forall s,a$$

↳ something like 'optimal'

Can we achiere this?

Since $Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s'} P_{ss'}^{a} Q^{\pi}(s', \pi(s'))$

if $\pi^*(\cdot)$ attains the maximum for $\forall (s,a)$ pairs,

then it is necessary that: $F(\pi, s') := Q^{\pi}(s', \pi(s'))$

Necessary Condition to attain the maxima at π^*

$\Leftrightarrow$ consider $\pi(s') = a'$

$\max Q^{\pi}(s', \pi(s'))$ then $\max_{a'} \max_{\pi} Q^{\pi}(s', a')$

π

Actually, what we do is to decompose

$\pi \longrightarrow \begin{cases} \pi(s') \to a' \\ \pi(S \setminus \{s'\}) \to A \end{cases}$

can determine $\pi^*(s') \to a'_0$

do for every s'

$\Downarrow$

$\pi^*: S \longrightarrow A$ optimal exists.

$\Rightarrow$ the existence of optimal policy $\pi^*(\cdot)$

s.t $Q^{\pi^*}(s,a) \geq Q^{\pi}(s,a) \quad \forall s, a, \pi$

Answer: We can achieve this! By picking π^* as follows:

$$\pi^*: s \longmapsto \arg\max_a \left\{ \max_\pi Q^\pi(s,a) \right\}$$

↳ a Naive Method!

↳ Greedy policy → actually is optimal policy

Since we know there exists $Q^*(s,a) \geq Q^\pi(s,a) \quad \forall (s,a,\pi)$

then we turn to check the optimality condition!

$$Q^*(s,a) = \max_\pi Q^\pi(s,a)$$

$$= \max_\pi \left\{ r(s,a) + \gamma \sum_{s'} P^a_{ss'} Q^\pi(s', \pi(s')) \right\}$$

$$= r(s,a) + \gamma \sum_{s'} P^a_{ss'} \max_\pi \left\{ Q^\pi(s', \pi(s')) \right\}$$

decompose →

$$= r(s,a) + \gamma \sum_{s'} P^a_{ss'} \max_{a'} \max_{\tilde{\pi}} \left\{ Q^{\tilde{\pi}}(s', a') \right\}$$

$$= r(s,a) + \gamma \sum_{s'} P^a_{ss'} \boxed{\max_{a'} Q^*(s', a')}$$

↓ determine $\boxed{\pi^*(s') = \arg\max_{a'} Q^*(s', a')}$

from Optimality condition (necessary)

→ For Non-deterministic policy $\pi(a|s)$

Decompose: $\pi: S \to a$, $\tilde{\pi}(s|s,a) \to A$

$$\begin{cases} V^\pi(s) = \sum_a \pi(a|s) \cdot Q^\pi(s,a) \\ Q^\pi(s,a) = r(a,s) + \gamma \sum_{s'} P_{ss'}^a V^\pi(s') \end{cases}$$

$$\Rightarrow V^\pi(s) = \sum_a \pi(a|s) \left(r(a,s) + \gamma \sum_{s'} P_{ss'}^a V^\pi(s') \right)$$

Dynamic Program

⇓

optimality Bellman Equation:

$$\begin{cases} V^*(s) = \max_\pi V^\pi(s) = \max_a Q^*(s,a) = \max_a \left\{ r(a,s) + \gamma \sum_{s'} P_{ss'}^a V^*(s') \right\} \\ Q^*(s,a) = r(a,s) + \gamma \sum_{s'} P_{ss'}^a V^*(s') \end{cases}$$

$$= r(a,s) + \gamma \sum_{s'} P_{ss'}^a \max_{a'} Q^*(s',a')$$

policy $\pi^*(a|s) = \begin{cases} 1 & a \in \arg\max Q^*(s,a) \\ 0, & \text{o/w} \end{cases}$

⇒ Value Iteration:

$$V_{k+1}(s) \leftarrow \max_a \{ r(a,s) + \gamma \sum_{s'} P_{ss'}^a V_k(s') \}$$

⇓ optimal Bellman Equation

$Q_k(\cdot)$ represents under π_{k+1}

KEY: $Q_k(s, \pi_k(s)) \geq V_k(s) = Q_k(s, \pi_{k+1}(s))$

⇒ Policy Iteration

it can be proved: $V_{k+1}(s) \geq V_k(s)$ with greedy policy π_k

$$V_{k+1}(s) \leftarrow \sum_a \pi_k(a|s) \left(r(a,s) + \sum_{s'} P_{ss'}^a V_k(s') \right) \quad \pi_{k+1}$$

$V_{k+1}(s)$ → converge to V_{π_k}

where $\pi_k(a|s)$ = greedy policy for V_k

$$= \begin{cases} 1, & a \in \arg\max_{a'} Q_k(s,a') \\ & \quad = r(s,a') + \gamma \sum_{s'} P_{ss'}^a V_k(s') \\ 0, & \text{else} \end{cases}$$

Last Part Discussion About DP ↔ RL (State-Value $V(s)$ & optimality condition)

Bellman Equ

① $V^{\pi}(s) = \sum_a \pi(a|s) \cdot Q^{\pi}(s,a)$

$= \sum \pi(a|s) \cdot \left[r(a,s) + \sum_{s'} P^{a}_{ss'} V^{\pi}(s') \right]$

☆

If $V^{\pi}(s) \rightarrow$ maxima at state s (optimal)

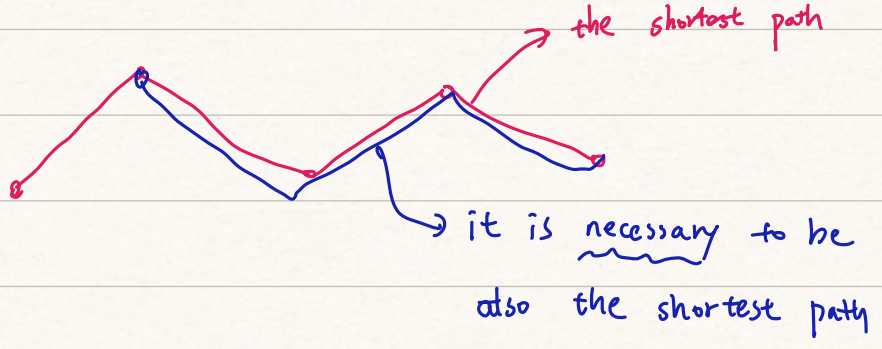
then for the SUBSEQUENT $V^{\pi}(s')$,

it still needs to be optimal

(necessary condition)

↓

DP



② $Q^{\pi}(s,a) = r(s,a) + \gamma \sum_{s'} P^{a}_{ss'} \underbrace{Q^{\pi}(s', \pi(s'))}$

↓

subsequently optimal

Slide Standford

1. $V^{\pi}(s) = \sum_a \pi(a|s) \cdot Q^{\pi}(s,a)$

$= \sum_a \pi(a|s) \sum_{s'} P^{a}_{ss'} \left(\rho(s,a,s') + \gamma V^{\pi}(s') \right)$

$= \sum_a \pi(a|s) \left(R^{a}_{s} + \gamma \sum_{s'} P^{a}_{ss'} V^{\pi}(s') \right)$

$= R_{\pi}(s) + \gamma \sum_{s'} P_{\pi}(s,s') V^{\pi}(s')$

→ 2. Define Bellman Operator $B^\pi(\underline{V}) = R_\pi + \gamma P_\pi \underline{V}$

From (1.), we know $B^\pi(\underline{V}_\pi) = R_\pi + \gamma P_\pi \cdot \underline{V}_\pi$

↓ Bellman State-Value Equation ⇓

$\underline{V}_\pi$ is fixed point of operator B_π !

注：下考慮 Optimal

we define $V^*(s) = \max_\pi V_\pi(s)$

$Q^*(s,a) = \max_\pi Q_\pi(s,a)$

the existence of Optima Policy ⟷ Greedy Policy

$G: \underline{V} \longrightarrow$ greedy policy ($G(\underline{V})$ is a policy)

$G(\underline{V})(s) := \arg\max_a \{ R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \underline{V}(s') \}$

Construction !!!!

→ 考慮 policy: $\pi^*: S \longrightarrow A$

$s \longmapsto \arg\max_a \{ \max_\pi Q^\pi(s,a) \}$ $Q^*(s,a)$

Punchline:

the existence of optimal policy

Therefore, $Q^*(s, \pi^*(s)) = \max_a Q^*(s,a) = V^*(s)$] Optimal Bellman Equation

$Q^*(s', \pi^*(s'))$

$Q^*(s,a) = R(s,a) + \sum_{s'} P_{ss'}^a V^*(s') \cdots$

⇒ π^* can yield optimal state-value function

⇒ $\pi^*(\cdot)$ can lead to $V_{\pi^*} \equiv V^*$ for all $s \in S$

also, $Q_{\pi^*}(s,a) = Q^*(s,a)$ for all

---

3. $V_\pi(s) = R_\pi(s) + \gamma \sum_{s'} P_\pi(s,s') V_\pi(s')$

→ Define $V^*(s) = \max_\pi V_\pi(s)$ → do not have corresponding policy explicitly

$V_\pi(s) = \sum_a \pi(a|s) Q_\pi(s,a) = \max_\pi \{ R_\pi(s) + \gamma \sum_{s'} P_\pi(s,s') V_\pi(s') \}$

$\Rightarrow V_*(s) = \max_\pi \{ \sum_a \pi(a|s) Q_\pi(s,a) \}$

$\leq Q_*(s,a) = Q_{\pi^*}(s,a) = \max_\pi \{ \sum_a \pi(a|s) (R_s^a + \gamma \sum_{s'} P_{ss'}^a V_\pi(s')) \}$

$\Rightarrow V_*(s) \geq Q_{\pi^*}(s,a) = Q_*(s,a) = \max_\pi Q_\pi(s,a) = \max_\pi \{ R_s^a + \gamma \sum_{s'} P_{ss'}^a V_*(s') \}$

3. Bellman Optimality Equation

Punchline

① $V^*(s) = \max_\pi V_\pi(s)$

4. Define Bellman Optimal Operator

$B_*[\underline{V}(s)] = \max_a \{ R_s^a + \gamma \sum_{s'} P_{ss'}^a \underline{V}(s') \}$

$\rightarrow \underline{V}^*(s) := \max_\pi \underline{V}_\pi(s) \in$ fixed point of $B_*(\cdot)$

i.e. $B_*(V_*(s)) = V_*(s)$

$Q^*(s,a) = \max_\pi Q_\pi(s,a)$

② $V^*(s) = \max_{a \in A} Q^*(s,a)$

$V^*(s) = \max_\pi V_\pi(s) = \max_\pi Q_\pi(s,a)$

$Q^*(s,a) = R(s,a) + \sum_{s'} P_{ss'}^a V^*(s')$

$\Rightarrow$ ③ $V^*(s) = \max_{a \in A} \{ R(s,a) + \sum_{s'} P_{ss'}^a V^*(s') \}$

④ $Q^*(s,a) = R(s,a) + \sum_{s'} P_{ss'}^a \max_{a' \in A} \{ Q^*(s',a') \}$

★ the maxima π must have the property that $\pi(a|s) = 1$ for some a

解释 $\{V(s)\} \rightarrow \{Q(s,a)\}$ from def

$\Rightarrow$ greedy algorithm (policy) always yielding a better policy ($\pi(a|s)=1$)

$V_\pi = B_\pi(V_\pi) \leq B_*(V_\pi)$

$= B_{G(V_\pi)}(V_\pi)$

$:= B_{\hat{\pi}}(V_\pi)$

$:= B_{\hat{\pi}}^N(V_\pi)$

$\approx V_{\hat{\pi}}$

$\rightarrow$ Define Bellman Optimal Operator:

$B^*(\underline{V}) := \max_{a \in A} \{ R(s,a) + \sum_{s'} P_{ss'}^a \underline{V}(s') \}$

③ $\Rightarrow \underline{V}^*$ ∈ the fixed point of B^* operator

i.e., $B^*(\underline{V}^*) = \underline{V}^*$

---

4. Greedy policy for Value Function $\underline{V}$

$G(\underline{V})(s) := \operatorname{argmax}_a \{ R_s^a + \gamma \sum_{s'} P_{ss'}^a \underline{V}(s') \}$

相当于筛选

$\Rightarrow B_{G(\underline{V})} \underline{V} = B_* \underline{V}$ from definition

($B_{G(\underline{V})}$: policy operator; B_*: optimal operator)

and $G(\underline{V})$ is greedy policy for value function $\underline{V}$

Key point:

① $B_\pi(V_\pi) = V_\pi$

② $B_*(V_*) = V_*$

③ $B_{G(\underline{V})}(\underline{V}) = B_*(\underline{V})$

→ γ-contraction $\Rightarrow$ fixed point

$\Rightarrow \lim_{N \to \infty} B_\pi^N(V_0) = V_\pi$

→ Recall **Policy Iteration**

→ ① Policy Evaluation

given policy π_k → calculate V_{π_k}

- Naive → $V_{\pi_k} = R_{\pi_k} + P_{\pi_k} \cdot V_{\pi_k}$ → equation
- Fixed point ⇝ $\lim_{N\to\infty} B^N_{\pi_k}(V_0) = V_{\pi_k}$

$B^N_{\pi_k}$ → linear operator

→ $R_{\pi_k} + P_{\pi_k}(\cdot)$

→ ② Policy improvement → Greedy Policy

$\pi_{k+1} = G(V_{\pi_k})$ → greedy policy

theoretically, we have

$$V_{\pi_k} = B_{V_{\pi_k}}(V_{\pi_k}) \leq B_*(V_{\pi_k}) = B_{G(V_{\pi_k})}(V_{\pi_k})$$

$$= B_{\pi_{k+1}}(V_{\pi_k})$$

Therefore, considering $V_{\pi_{k+1}} = \lim_{N\to\infty} B^N_{\pi_{k+1}}(\pi_k)$

$$\geqslant V_{\pi_k} !$$

从 operator 角度分析！ (妙)！