Reinforcement Learning LEC11 DSA5105

Recap:

State Space of Agent

Reward

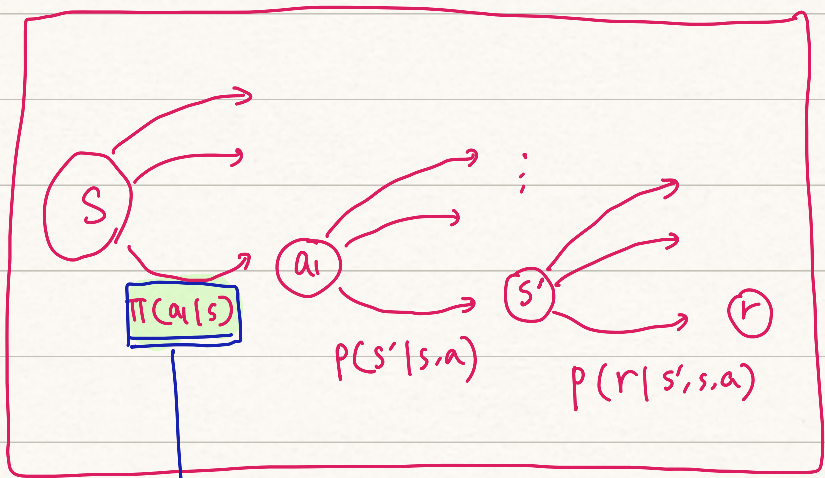
MDP $\longrightarrow$ $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$

Action Space

discount rate

$p(s', r \mid s, a)$

Diagram



Policy (Stochastic)

DP $\longrightarrow$ RL Model-Based Model

$\rightarrow$ $\boxed{\text{State-value Function}}$ $\quad V_\pi(s) = \mathbb{E}^\pi[G_t \mid S_t = s]$

$$= \mathbb{E}^\pi\left[\sum_{\tau=0}^{\infty} \gamma^\tau R_{t+\tau+1} \mid S_t = s\right]$$

We consider a time-homogeneous MDP

$\boxed{\text{action-value Function}}$ $\quad q_\pi(s,a) = \mathbb{E}^\pi[G_t \mid S_t = s, A_t = a]$

$$= \mathbb{E}^\pi\left[\sum_{\tau=0}^{\infty} \gamma^\tau R_{t+\tau+1} \mid S_t = s\right]$$

$\rightarrow$ Goal: derive recursive relation with $q_\pi(s,a)$ / $v_\pi(s)$

$$\boxed{\text{Bellman Equation}}$$

(Bellman Equation) ①

Derivation:

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

$$= \mathbb{E}^\pi[R_{t+1} \mid S_t = s] + \gamma \mathbb{E}^\pi\left[\sum_{\tau=0}^{\infty} \gamma^\tau R_{t+\tau+2} \mid S_t = s\right]$$

$$= b(\pi)_s + \gamma \sum_{s'} P(\pi)_{ss'} \cdot v_\pi(s')$$

$$\begin{cases} b(\pi)_s := \mathbb{E}^\pi[R_{t+1} \mid S_t = s] = \sum_a \pi(a|s) \cdot \sum_{r,s'} p(s',r|s,a) \cdot r \\ P(\pi)_{ss'} = \mathbb{P}^\pi(S_{t+1} = s' \mid S_t = s) \\ \quad = \sum_a \pi(a|s)\, \mathbb{P}(S_{t+1} = s' \mid S_t = s, A_t = a) \\ \quad = \sum_a \pi(a|s) \cdot \sum_r p(s',r|s,a) \end{cases}$$

give the first-step transition probability

$\boxed{\text{easy to compute}}$

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

$$= \sum_{s',r} r \cdot p(s',r|s,a) + \gamma \mathbb{E}^\pi\left[\sum_{\tau=0}^{\infty} \gamma^\tau R_{t+\tau+2} \mid s,a\right]$$

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

$$= \sum_{s',r} p(s',r|s,a)\left(r + \gamma v_\pi(s')\right)$$

Rmk: $$v_\pi = b(\pi) + \gamma \cdot P(\pi) \cdot v_\pi$$

$$\Rightarrow v_\pi = (I - \gamma P(\pi))^{-1} b(\pi)$$

$\Longrightarrow$ idea: when we have $\begin{cases} P(s', r \mid s, a) \\ \text{policy } \pi(a \mid s) \end{cases}$, then we can

compute for $V_\pi(s)$ & $q_\pi(s,a)$ $\forall s \in S$ and $\forall a \in A$

Policy Comparison

(POSET)

partial order

① not necessary happen

Property

② $\begin{cases} a \geq a \\ a \geq b \quad b \leq a \Rightarrow a = b \\ a \geq b \quad b \geq c \Rightarrow a \geq c \end{cases}$

Defn : $\boxed{\pi' \geq \pi}$

$\Leftrightarrow V_{\pi'}(s) \geq V_\pi(s) \quad \forall s \in S$

A trivial question : optimal policy exists or not !

uniques or not !

Bellman Optimality Introduced

Outline: ① Policy Improvement

$\longrightarrow$ if $\sum_a \pi'(a|s) \cdot q_\pi(s,a) \geq \sum_a \pi(a|s)\, q_\pi(s,a)$

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

Rmk: ① Give us the motivation of policy iteration

② Necessary Condition for Optimal Policy

③ a natural choice to improve policy is:

$$\pi'(a|s) = \begin{cases} 1 & a = \arg\max_{a \in A} q_\pi(s,a) \\ 0 & o/w \end{cases}$$

deterministic policy

(necessary part)

② Bellman Optimality Condition

$\pi^*(a|s)$ is an optimal policy

$\Rightarrow$ ① if $\pi^*(a|s) > 0$

then $a \in \arg\max_{a'} q_{\pi^*}(a', s)$

$\Rightarrow$ ② $v_{\pi^*}(s) = \max_a q_{\pi^*}(a, s)$

$$= \max_a \sum_{s',r} p(s',r|s,a)\left(r + \gamma v_{\pi^*}(s')\right)$$

$$q_{\pi^*}(s,a) = \sum_{s',r} p(s',r|s,a)\left(r + \gamma v_{\pi^*}(s')\right)$$

$$= \sum_{s',r} p(s',r|s,a)\left(r + \gamma \max_{a'} q_{\pi^*}(s',a')\right)$$

requires Contraction Mapping Thm to show the uniqueness of Bellman Optimality Equation

(Sufficient Part)

given a policy $\pi^*(a|s) \longrightarrow \begin{cases} v_{\pi^*}(s) \\ q_{\pi^*}(s,a) \end{cases}$

$\longrightarrow$ if $v_{\pi^*}(s) = \max_a \sum_{s',r} p(s',r|s,a)\left(r + \gamma v_{\pi^*}(s')\right)$

then π^* is the optimal policy

$\longrightarrow$ if $\pi^*(a|s) > 0 \Rightarrow a \in \arg\max_{a'} q_{\pi^*}(a'|s)$

then π^* is the optimal policy