DSA5105 LEC12

Model-based RL $\Longrightarrow$ know the knowledge of $p(s', r \mid s, a)$

Model-free RL $\Longrightarrow$ $p(s', r \mid s, a)$ is unknown

---

① Model-based RL $\Longrightarrow$ work with { Bellman Equa. / Bellman Optimality Equa. } exactly

Recap: a) Bellman Optimality Equation

$$v_*(s) = \max_a q_*(s, a)$$

$$q_*(s, a) = \sum_{s', r} p(s', r \mid s, a)\,(r + \gamma v_{\pi_*}(s'))$$

Value Iteration

Non-linearity → difficulty

iterative method

b) Bellman Equation

$$v_\pi(s) = \sum_a \pi(a \mid s)\, q_\pi(s, a)$$

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

c) Optimal Policy

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

Policy Iteration

d) Policy Improvement

① Value Iteration ⟶ { Bellman Optimality Condition $V_*(s)$; Optimal Policy through $q_*(s,a)$ }

define $F(V)(s) = \max_a \sum_{s',r} p(s',r|s,a)\,(r + \gamma V(s'))$

⟷ Matrix Form

$\{ P(\pi)_{ss'} = p^\pi(s'|s) = p(s'|s,a)$; $b(\pi)_s = \mathbb{E}^\pi[r|s] = \sum_{s',r} p(s',r|s,a)\cdot r$; $a = \pi(s)$ }

$F(v) = \max_\pi \left[b(\pi) + \gamma P(\pi)\cdot v \right]$

where π is deterministic policy

then V_* is optimal value function $\iff V_* = F(V_*)$

$V_{k+1} = F(V_k)$ + Contraction Mapping Theorem

⇓

Convergence (Fixed Point)

② Policy Iteration ⟶ { Bellman Equation ; Policy Improvement }

Finite Algorithm! (for finite MDP)

$\pi \xrightarrow{E} v_\pi \xrightarrow{I} \pi'$

E: $V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\,(r + \gamma V_\pi(s'))$

⟷ Matrix form

$V_\pi = b(\pi) + \gamma P(\pi) V_\pi$

linear Equation Solution

$\pi' = \underset{\tilde{\pi} \in \text{deterministic}}{\text{argmax}} \left[b(\tilde{\pi}) + \gamma P(\tilde{\pi}) \cdot V_\pi \right]$

I: $\pi'(s) = \underset{a}{\text{argmax}}\ q_\pi(s,a)$

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

Model-free Algo

① Monte-Carlo Idea

↳ estimate Expectation from Sample Mean

⇑

guarantee by { LLN, CLT }

$$\mathbb{E}_{x \sim \mu} f(x) = \frac{1}{N} \sum_{i=1}^{N} f(X_i) \qquad X_i \sim \mu \ \text{(i.i.d)}$$

→ When $p(s', r \mid s, a)$ is our known knowledge,

then $V_\pi = b(\pi) + \gamma P(\pi) V_\pi$

→ When $p(s', r \mid s, a)$ is unknown,

then we go back to the defn of $V_\pi(s)$.

$$v_\pi(s) = \mathbb{E}_\pi [G_0 \mid S_0 = s]$$

$\Rightarrow$ **Monte-Carlo Comes in!**

↳ Sample N trajectories (episodes) through Policy π

$$\{ S_t^{(n)}, R_t^{(n)} \} \quad n = 1, 2, \dots, N \qquad t = 1, 2, \dots T$$

and $\boxed{S_0^{(n)} \equiv s}$

$$\Rightarrow v_\pi(s) \approx \frac{1}{N} \sum_{n=1}^{N} G_0^{(n)} \longrightarrow \text{Monte-Carlo Idea}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \sum_{t=0}^{\infty} \gamma^t R_{t+1}^{(n)}$$

$\Rightarrow$ $q_\pi(s,a)$ can be estimated in the same fashion !!

Then, do Policy Improvement $\quad \boxed{\pi'(s) \in \underset{a}{\operatorname{argmax}}\, q_\pi(s,a)}$

Rmk: approximation of improved policy since the q-function is approximated through Monte-Carlo !

GPI $\Longleftarrow$

$\Downarrow$

Generalized Policy Iteration Framework

$\uparrow$

Not solve for v_π exactly

Short-term Monte Carlo Fashion

② Temporal-Difference Idea ⇝ not sample N complete trajectories (Monte-Carlo)

↓

iterate formula: $V(s) \leftarrow (1-\alpha)\, V(s) + \alpha\,(r + V(s'))$

where a sample from policy $\pi \longrightarrow \pi(\cdot|s)$

and (s', r) is achieved by environment simulator

Observation: $V_\pi(s) = \mathbb{E}_\pi [G_0 \mid S_0 = s]$

$= \mathbb{E}_\pi [r + \gamma G_1 \mid S_0 = s]$

Interpretation:

Lemma: $a_{k+1} = (1-\alpha)\, a_k + \alpha b \qquad 0 \leq \alpha < 1$

$\Rightarrow \boxed{\lim_{k\to\infty} a_k = b}$ (converge exponentially)

↓

idea: $V_{k+1} = (1-\alpha) V_k + \alpha\, \mathbb{E}_\pi [G_0 \mid S_0 = s]$

$\Rightarrow \lim_{k\to+\infty} V_k = \mathbb{E}_\pi [G_0 \mid S_0 = s]$ exponentially

↓

General Idea

$\mathbb{E}_\pi [G_0 \mid S_0 = s] \approx G_0$ "one sample from environement simulator"

Illustration

$= \mathbb{E}_\pi [r + \gamma G_1 \mid S_0 = s]$

$\approx r + \gamma\, \mathbb{E}_\pi [G_1 \mid S_0 = s] = r + \gamma \sum_{s'} P(\pi)_{ss'}\, \mathbb{E}_\pi [G_1 \mid S_1 = s']$

$\approx r + \gamma\, \mathbb{E}_\pi [G_1 \mid S_0 = s, S_1 = s']$ → one-term estimator

$= r + \gamma\, V_\pi(s')$

Q-learning Framework ← TD idea (off-policy)

$$q(s,a) \leftarrow (1-\alpha)\, q(s,a) + \alpha \left[r + \gamma \max_{a'} q(s', a') \right]$$

determine which (s,a) pair should be updated

- $a \rightarrow$ sample from $\pi(\cdot \mid s)$ ⇝ (s', r) (simulator)
- $a' \rightarrow$ sample from greedy policy