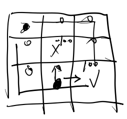
RL Intro

Friday, January 16, 2026 21:57

- Notation



state value

$s \in S$, $a \in A$, $r \in R$

$V(s)$, $q(s, a)$ → action value

$\pi(a|s)$ → policy

3x3

$V(s) = 0.5 \times 100 + 0.5 \times -100 = 0$

$q(s, a = right) = 100$

- MDP $s \xrightarrow[r]{a} s' \xrightarrow{a'} s''$ discount factor

- Bellman equation $V_\pi(s) = E(r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \dots \gamma^{T-t} r_T \mid S = S_t)$

- DP

$= \sum_a \pi(a|s) \sum_{s'} p(s', r|s, a)\,[r + V_\pi(s')]$ → Bellman equation ($q_\pi(s, a)$)

$\pi^* = \arg\max_\pi V_\pi(s)$

$V_\pi(s) < q_\pi(s, a)$

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

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

} Bellman optimality equation

- model-based method { PI, VI } env response: $p(s'|s, a)$, r
- model-free method { MC, TD }

- policy iteration { PE (estimation), PI (improvement) }

• initial π, $V(s)$ for $s \in S$

Loop for $s \in S$:

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

$|V_{old}(s) - V(s)| < \varepsilon$, → $V(s)$ is convergent
}

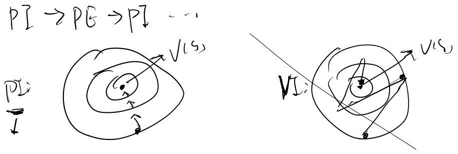
• for $s \in S$:

PI {
$a^* = \arg\max_a q(s, a)$

if $\pi_{a=a^*}(a|s) \neq 1$, → π not optimal → $\pi_{a=a^*}(a|s) = 1$ else $\pi_{a \neq a^*}(a|s) = 0$
}

• if π is not optimal, go to PE again

PE → PI → PE → PI ...



- Value iteration

• initialize π, $V(s)$ for $s \in S$

Loop for $s \in S$:

$V(s) = \arg\max_{a^*} q(s, a)$ → value update

if $\pi_{a=a^*}(a|s) \neq 1$, π is not optimal, → $\pi_{a=a^*}(a|s) = 1$ else $\pi_{a \neq a^*}(a|s) = 0$ → policy update

$|V_{old}(s) - V(s)| < \varepsilon$, → V is convergent

• output π, V

- MC

① × env response $p(s'|s, a)$ [generate experience + avg]

② $S \uparrow$, $V(s)$ cost $\uparrow$,

$p(S_t)$, randomly assign start point (s_0, a_0) — MC with start point

π { exploitation, (greedy action) → $\varepsilon = 0.1$; exploration }

ε-greedy {
$a^* = \arg\max_a q(s, a)$

$\pi_{a=a^*}(a|s) = 1 - \varepsilon + \frac{\varepsilon}{|A(s)|}$

$\pi_{a \neq a^*}(a|s) = \frac{\varepsilon}{|A(s)|}$
}

$s_0 \xrightarrow[r_1]{a_0} s_1 \xrightarrow[r_2]{a_1} s_2 \cdots s_{T-1} \xrightarrow[r_T]{a_{T-1}} s_T$ experience / episode / trajectory → G_T

$q_\pi(s_t, a_t) = E_\pi[R_{t+1} + \gamma R_{t+1} | S=s_t, A=a_t] = E_\pi[G_t | S=s_t, A=a_t]$

version 1

① MC w start port

- initialize π, $q(s,a)$, returns(s,a) empty list for $s \in S$
- for episode k in $1, 2, \cdots, n$:
  - randomly select (S_0, A_0)
  - $\pi \rightarrow S_0, A_0, R_1, S_1, A_1, \cdots R_T, S_T$
  - for t in $T-1, T-2, \cdots, 0$:
    - $G_t^k \rightarrow G_{t-1}^k + \gamma R_t$
    - append G_t^k into returns(S_t, A_t) list
    - $q^k(S_t, A_t) =$ avg returns(S_t, A_t)
  - 给first-visited
  - $a^* = \arg\max_a q(S_t, A_t)$
  - $\pi(a|s) = 1$ and $\pi(a \neq a^*|s) = 0$

② ε-greedy method

由 $\pi \rightarrow S_0, A_0, R_1, \cdots R_T, S_T$

$\pi(a|s) = 1 - \varepsilon + \frac{\varepsilon}{|A(s)|}$ $a = a^*$

and $\pi(a|s) = \frac{\varepsilon}{|A(s)|}$ $a \neq a^*$

- incremental implementation

Same (S_t, A_t) $G_t^1, G_t^2, \cdots, G_t^n$

$q^n(S_t, A_t) = \frac{\sum_{k=1}^{n-1} G_t^k}{n-1} = \frac{\sum_{k=1}^{n-1} w^k G_t^k}{\sum_{k=1}^{n-1} w^k}$

$q^{n+1}(s,a) = \frac{\sum_{k=1}^{n} G_t^k w^k}{\sum_{k=1}^{n} w^k} \rightarrow q^{n+1}(s,a) = \frac{q^n(s,a)\sum_{k=1}^{n-1} w^k + G_t^n w^n}{\sum_{k=1}^{n} w^k} = \frac{q^n \sum_{k=1}^{n-1} w^k + G_t^n w^n + q^n w^n - q^n w^n}{\sum_{k=1}^{n} w^k}$

$= \frac{q^n \sum_{k=1}^{n} w^k + w^n(G_t^n - q^n)}{\sum_{k=1}^{n} w^k} = q^n + \frac{w^n}{\sum_{k=1}^{n} w^k}(G_t^n - q^n)$

α

version 2 (w incremental implementation)

- initialize π, $q(s,a)$, returns(s,a) empty list for $s \in S$ $\theta^{n+1} \leftarrow \theta^n + \alpha \nabla J(\theta)$
- for episode k in $1, 2, \cdots n$
  - $\pi \rightarrow S_0, A_0, R_1, \cdots S_T, A_T$
  - for t in $T-1, T-2, \cdots 1$:
    - $G_t^k \rightarrow G_{t-1}^k + \gamma R_t$
    - $q^{k+1}(S_t, A_t) = q^k + \frac{w^k}{\sum w^k}(G_t^k - q^k(S_t, A_t))$

- TD learning

$R_{t+1} + \gamma q(S_{t+1}, A_{t+1})$

$q(S_t, A_t) \leftarrow q(S_t, A_t) + \alpha[G_t - q(S_t, A_t)]$

$V(S_t) \leftarrow V(S_t) + \alpha[G_t - V(S_t)]$

$\rightarrow R_{t+1} + \gamma V(S_{t+1})$ (TD target)

$V(S_t) \leftarrow V(S_t) + \alpha [R_{t+1} + \gamma V(S_{t+1}) - V(S_t)]$ 1-step learning : TD(0)

- proof:

$V(S_t) = E[G_t | S = S_t] = E[R_{t+1} + \gamma G_{t+1} | S = S_t]$ (TD error)

$= E[R_{t+1} + \gamma V(S_{t+1}) | S = S_t]$ → comb of MC and DP

- generalization of TD error:

$G_t - V(S_t) = R_{t+1} + \gamma G_{t+1} - V(S_t) + \gamma V(S_{t+1}) - \gamma V(S_{t+1})$

$= \delta_t + \gamma (G_{t+1} - V(S_{t+1})) = \delta_t + \gamma \delta_{t+1} + \gamma (G_{t+2} - V(S_{t+2})) = \cdots = \sum_{k=t}^{T} \gamma^{k-t} \delta_k$

where $\delta_t = R_{t+1} + \gamma G_{t+1} - V(S_t)$

- property

TD learning variance ↓, bias ↑ compared to MC

$S_t, S_{t+1}, A_{t+1} R_{t+1}$ → fewer var

- SARSA (state-action-reward-state-action)
  - initialize π, $q(s,a)$ for $s \in S, a \in A$, $q(S_T, A) = 0$
  - for every episode:
    - initialize S_0
    - $\pi(\cdot | S_0) \rightarrow a_0$ (ε-greedy considered)
    - for t in 0, 1, 2 ... T-1:
      - take action $a_t \rightarrow R_{t+1}, S_{t+1}$ (wait until n-step to update $q(S_t, A_t)$)
      - $\pi(\cdot | S_{t+1}) \rightarrow A_{t+1}$
      - $q(S_t, A_t) \leftarrow q(S_t, A_t) + \alpha [R_{t+1} + \gamma q(S_{t+1}, A_{t+1}) - q(S_t, A_t)]$ → ≈ value update
      - $a^* = \arg\max_a q(S_t, a)$
      - update $\pi(a_t | S_t)$ based on ε-greedy → ≈ policy update

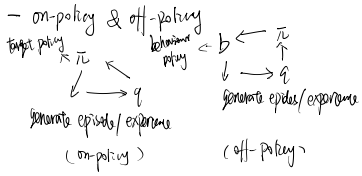
- Q-learning
  - initialize π, $q(s,a)$ for $s \in S, a \in A$, $q(S_T, A) = 0$
  - for every episode:
    - initialize S_0
    - $\pi(\cdot | S_0)$ w greedy → a_0
    - for t in 0, 1, 2 ... T-1:
      - take action $a_t \rightarrow R_{t+1}, S_{t+1}$
      - $\pi(\cdot | S_{t+1}) \rightarrow A_{t+1}$
      - $q(S_t, A_t) \leftarrow q(S_t, A_t) + \alpha [R_{t+1} + \max_a q(S_{t+1}, A_{t+1}) - q(S_t, A_t)]$
      - update $\pi(A_t | S_t)$ based on ε-greedy

- n-step learning

n-step Sarsa

$$q(S_t, A_t) \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n q(S_{t+n}, A_{t+n}) - q(S_t, A_t)$$

- on-policy & off-policy



- importance sampling

$S_0 \xrightarrow[R_1]{A_0} S_1 \xrightarrow[R_2]{A_1} S_2 \cdots S_T$ $\quad (S_0, A_0, R_1, S_1, A_1, \cdots)$

$$\frac{\prod_{t=0}^{T} \pi(A_t|S_t)\, p(S_{t+1}|S_t, A_t)}{\prod_{t=0}^{T} b(A_t|S_t)\, p(S_{t+1}|S_t, A_t)} = \frac{\prod_{t=0}^{T} \pi(A_t|S_t)}{\prod_{t=0}^{T} b(A_t|S_t)} \doteq \rho \quad \text{(importance sampling ratio)}$$