Lec. 1

State / State space: $\{s_i\}_{i=1}^{n} = S$

Action / Action space: $\{a_1, a_2 \cdots\} = A(s_i)$

State Transition: $p(s'|s,a) = P(S_{t+1}=s' \mid S_t=s, A_t=a)$

$$\sum_{s'\in S} p(s'|s,a) = 1$$

Policy: $\pi(a|s) = P(A_t=a \mid S_t=s)$

Return: 沿着 trajectory 求 Reward 的和

Discounted Return: $G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots$

Lec 2. Bellman Equation:

State value: $V(\pi,s) = V_\pi(s) = E[G_t \mid S_t=s]$

$\because G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} \cdots = G_t = r_{t+1} + \gamma G_{t+1}$

$\therefore V_\pi(s) = E[r_{t+1} + \gamma G_{t+1} \mid S_t=s] = E[r_{t+1} \mid S_t=s] + \gamma E[G_{t+1} \mid S_t=s]$

$E[r_{t+1} \mid S_t=s] = \sum_a \pi(a|s) E[r_{t+1} \mid S_t=s, A_t=a]$

$= \sum_a \pi(a|s) \cdot \sum_r p(r|s,a) \cdot r$

$E[G_{t+1} \mid S_t=s] = \sum_{s'} p(s'|s) \cdot E[G_{t+1} \mid S_{t+1}=s', S_t=s] = \sum_{s'} p(s'|s) \cdot \underbrace{E(G_{t+1} \mid S_{t+1}=s')}_{V_\pi(s')}$

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

$\therefore V_\pi(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s,a) \cdot r + \gamma \sum_{s'} p(s'|s,a) \cdot V_\pi(s')\right]$ (BE 的 Elementwise-form)

矩阵形式 (Matrix-Vector form):

$$V_\pi = \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ \vdots \\ V_\pi(s_n) \end{bmatrix} \quad r_\pi = \begin{bmatrix} r_\pi(s_1) \\ \vdots \\ r_\pi(s_n) \end{bmatrix} = \begin{bmatrix} \vdots \\ \sum_a \pi(a|s_i) \cdot \sum_r p(r|s_i,a) \cdot r \\ \vdots \end{bmatrix}$$

$P_\pi \in R^{n\times n}$: $[P_\pi]_{ij} = p(s_j|s_i) = \sum_a \pi(a|s_i) \cdot p(s_j|s_i,a)$

$\Rightarrow \quad V_\pi = r_\pi + \gamma P_\pi \cdot V_\pi$

$\therefore$ 解析解: $V_\pi = (I - \gamma P_\pi)^{-1} r_\pi$

迭代解: $V_{k+1} = r_\pi + \gamma P_\pi V_k$

收敛性证明: $\begin{cases} ① \ V_{k+1} = r_\pi + \gamma P_\pi V_k \\ ② \ V_\pi = r_\pi + \gamma P_\pi V_\pi \end{cases}$

①-②: $V_{k+1} - V_\pi = \gamma P_\pi (V_k - V_\pi)$ $\quad \delta_{k+1} = \gamma P_\pi \delta_k$

$\therefore \delta_{k+1} = \gamma^{k+1} P_\pi^{k+1} \delta_0$ $\quad \because P_\pi$ 为随机矩阵 $\forall [P_\pi]_{ij} \in [0,1]$

又有 $\gamma < 1$

$\therefore \delta_{k+1} \to 0$

Action Value (动作价值): $q_\pi(s,a) = E[G_t \mid S_t=s, A_t=a]$

$$V_\pi(s) = \sum_a \pi(a|s) \cdot q_\pi(s,a)$$

如果知道所有 Action Value → 求平均得到这个状态的 State Value

如果知道所有状态的 State Value → 能求出所有 Action Value

Lec 3. Bellman Optimal Equation

Optimal Policy: $V_{\pi^*}(s) \geq V_\pi(s) \quad \forall s \in S, \forall \pi$

BE: $V_\pi(s) = \sum_a \pi(a|s) \underbrace{\left[\sum_r p(r|s,a) \cdot r + \gamma \sum_{s'} p(s'|s,a) V_\pi(s')\right]}_{q_\pi(s,a)}$

BOE: $V(s) = \max_\pi \sum_a \pi(a|s) q_\pi(s,a)$ 找一个 π 使这个式子最大

$\Rightarrow$ π 把所有权重分给最大 $q_\pi(s,a)$

$\Rightarrow \max_\pi \sum_a (a|s) q_\pi(s,a) = \max_a q_\pi(s,a)$

$\Rightarrow \quad V(s) = \max_a \left[\sum_r p(r|s,a) \cdot r + \gamma \sum_{s'} p(s'|s,a) \cdot V(s')\right]$

Matrix-Vector Form: $V = \max_\pi (r_\pi + \gamma P_\pi V)$

$f(V) = \max_\pi (r_\pi + \gamma P_\pi V) \qquad V = f(V)$

由压缩映射定理: 可知有条件 $\|f(u) - f(v)\|_\infty \leq \gamma \|u - v\|_\infty \quad \gamma \in [0,1)$

$\therefore$ 存在唯一不动点 $V^* = f(V^*)$ $\quad \therefore$ 算法收敛

值迭代求解 BOE: ①: $q_k(s,a) = \sum_r q(r|s,a) + \gamma \sum_{s'} p(s'|s,a) \cdot V_k(s')$

② $V_{k+1} = \max_a q_k(s,a)$

迭代结束:

$$\pi^*(a|s) = \begin{cases} 1 = \arg\max q^*(s,a) \\ 0 \quad \text{else} \end{cases}$$

奖励函数仿射不变性: 对奖励进行仿射变换 $r' = ar + b \ (a>0)$

最优策略不变

Lec 4. 值迭代 / 策略迭代

值迭代: $V_{k+1} = f(V_k) = \max_\pi (r_\pi + \gamma P_\pi V_k)$

Step 1: $\pi_{k+1} = \arg\max (r_\pi + \gamma P_\pi V_k)$ (PI)

Step 2: $V_{k+1} = r_{\pi_{k+1}} + \gamma P_{\pi_{k+1}} \cdot V_k$ (PE)

策略迭代: $\pi_0 \xrightarrow{PE} V_{\pi_0} \xrightarrow{PI} \pi_1 \xrightarrow{PE} V_{\pi_1} \xrightarrow{PI} \pi_2 \to \cdots$

Step 1. (PE): 计算当前 π_k 对应真实 V_{π_k}

Bellman Equation: $V_{\pi_k} = r_{\pi_k} + \gamma P_{\pi_k} \cdot V_{\pi_k}$

Step 2 (PI): 根据 V_{π_k} 寻找更好策略 π_{k+1}

$$\pi_{k+1} = \arg\max_\pi (r_\pi + \gamma P_\pi V_{\pi_k})$$

PE 采用迭代解 (固定策略下的值迭代)

$$V_{\pi_k}^{(j+1)} = r_{\pi_k} + \gamma P_{\pi_k} \cdot V_{\pi_k}^{(j)}$$

$j \to \infty \quad V_{\pi_k}^{(j)} \to V_{\pi_k}$

可证明: $V_{\pi_{k+1}} \geq V_{\pi_k}$

$\pi_{k+1} = \arg\max_\pi (r_\pi + \gamma P_\pi V_{\pi_k}) \Rightarrow r_{\pi_{k+1}} + \gamma P_{\pi_{k+1}} V_{\pi_k} \geq r_{\pi_k} + \gamma P_{\pi_k} V_{\pi_k} = V_{\pi_k}$ ①

令 $\Delta = V_{\pi_{k+1}} - V_{\pi_k} \Rightarrow \Delta = (r_{\pi_{k+1}} + \gamma P_{\pi_{k+1}} V_{\pi_{k+1}}) - V_{\pi_k}$

$= r_{\pi_{k+1}} + \gamma P_{\pi_{k+1}} V_{\pi_{k+1}} - \gamma P_{\pi_{k+1}} V_{\pi_k} + \gamma P_{\pi_{k+1}} \cdot V_{\pi_k} - V_{\pi_k}$

$= \gamma P_{\pi_{k+1}} (V_{\pi_{k+1}} - V_{\pi_k}) + (r_{\pi_{k+1}} + \gamma P_{\pi_{k+1}} \cdot V_{\pi_k} - V_{\pi_k})$

$\delta = r_{\pi_{k+1}} + \gamma P_{\pi_{k+1}} \cdot V_{\pi_k} - V_k$, 由① $\delta \geq 0$

$\therefore \Delta = \gamma P_{\pi_{k+1}} \Delta + \delta = \gamma(\gamma P_{\pi_{k+1}} \Delta + \delta)\Delta + \delta = \delta + \gamma P_{\pi_{k+1}} \delta + (\gamma P_{\pi_{k+1}})^2 \delta + \cdots$

$\because \delta \geq 0 \quad P \geq 0 \quad \gamma > 0 \quad \therefore \Delta \geq 0 \quad \therefore V_{\pi_{k+1}} \geq V_{\pi_k}$

截断策略迭代:

PE 阶段迭代固定次, 不像值迭代一样 PE 只迭代一次

也不像策略迭代一样 PE 迭代无数次