

## 附录 A 贝尔曼方程

### 定理 A.1. 贝尔曼方程 (将 $Q_\pi$ 表示成 $Q_\pi$ )

假设  $R_t$  是  $S_t$ 、 $A_t$ 、 $S_{t+1}$  的函数。那么

$$Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}} [R_t + \gamma \cdot Q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s_t, A_t = a_t].$$



**证明** 根据回报的定义  $U_t = \sum_{k=t}^n \gamma^{k-t} \cdot R_k$ ，不难验证这个等式：

$$U_t = R_t + \gamma \cdot U_{t+1}.$$

用符号  $S_{t+1} = \{S_{t+1}, S_{t+2}, \dots\}$  和  $A_{t+1} = \{A_{t+1}, A_{t+2}, \dots\}$  表示从  $t+1$  时刻起所有的状态和动作随机变量。根据动作价值函数  $Q_\pi$  的定义，

$$Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}} [U_t \mid S_t = s_t, A_t = a_t].$$

把  $U_t$  替换成  $R_t + \gamma \cdot U_{t+1}$ ，那么

$$\begin{aligned} Q_\pi(s_t, a_t) &= \mathbb{E}_{S_{t+1}, A_{t+1}} [R_t + \gamma \cdot U_{t+1} \mid S_t = s_t, A_t = a_t] \\ &= \mathbb{E}_{S_{t+1}, A_{t+1}} [R_t \mid S_t = s_t, A_t = a_t] + \gamma \cdot \mathbb{E}_{S_{t+1}, A_{t+1}} [U_{t+1} \mid S_t = s_t, A_t = a_t]. \end{aligned} \quad (\text{A.1})$$

假设  $R_t$  是  $S_t$ 、 $A_t$ 、 $S_{t+1}$  的函数。那么，给定  $s_t$  和  $a_t$ ，则  $R_t$  随机性唯一的来源就是  $S_{t+1}$ ，所以

$$\mathbb{E}_{S_{t+1}, A_{t+1}} [R_t \mid S_t = s_t, A_t = a_t] = \mathbb{E}_{S_{t+1}} [R_t \mid S_t = s_t, A_t = a_t]. \quad (\text{A.2})$$

等式 (A.1) 右边  $U_{t+1}$  的期望可以写成

$$\begin{aligned} &\mathbb{E}_{S_{t+1}, A_{t+1}} [U_{t+1} \mid S_t = s_t, A_t = a_t] \\ &= \mathbb{E}_{S_{t+1}, A_{t+1}} [\mathbb{E}_{S_{t+2}, A_{t+2}} [U_{t+1} \mid S_{t+1}, A_{t+1}] \mid S_t = s_t, A_t = a_t] \\ &= \mathbb{E}_{S_{t+1}, A_{t+1}} [Q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s_t, A_t = a_t]. \end{aligned} \quad (\text{A.3})$$

由公式 (A.1)、(A.2)、(A.3) 可得定理。□

### 定理 A.2. 贝尔曼方程 (将 $Q_\pi$ 表示成 $V_\pi$ )

假设  $R_t$  是  $S_t$ 、 $A_t$ 、 $S_{t+1}$  的函数。那么

$$Q_\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}} [R_t + \gamma \cdot V_\pi(S_{t+1}) \mid S_t = s_t, A_t = a_t].$$



**证明** 由于  $V_\pi(S_{t+1}) = \mathbb{E}_{A_{t+1}} [Q(S_{t+1}, A_{t+1})]$ ，由定理 A.1 可得定理 A.2。□

### 定理 A.3. 贝尔曼方程 (将 $V_\pi$ 表示成 $V_\pi$ )

假设  $R_t$  是  $S_t$ 、 $A_t$ 、 $S_{t+1}$  的函数。那么

$$V_\pi(s_t) = \mathbb{E}_{A_t, S_{t+1}} [R_t + \gamma \cdot V_\pi(S_{t+1}) \mid S_t = s_t].$$



**证明** 由于  $V_\pi(S_t) = \mathbb{E}_{A_t}[Q(S_t, A_t)]$ , 由定理 A.2 可得定理 A.3。  $\square$

#### 定理 A.4. 最优贝尔曼方程

假设  $R_t$  是  $S_t$ 、 $A_t$ 、 $S_{t+1}$  的函数。那么

$$Q_\star(s_t, a_t) = \mathbb{E}_{S_t \sim p(\cdot | s_t, a_t)} \left[ R_t + \gamma \cdot \max_{A \in \mathcal{A}} Q_\star(S_{t+1}, A) \mid S_t = s_t, A_t = a_t \right].$$



**证明** 设最优策略函数为  $\pi^\star = \operatorname{argmax}_\pi Q_\pi(s, a)$ ,  $\forall s \in \mathcal{S}, a \in \mathcal{A}$ 。由贝尔曼方程可得:

$$Q_{\pi^\star}(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}} \left[ R_t + \gamma \cdot Q_{\pi^\star}(S_{t+1}, A_{t+1}) \mid S_t = s_t, A_t = a_t \right].$$

根据定义, 最优动作函数是

$$Q_\star(s, a) \triangleq \max_{\pi} Q_\pi(s, a), \quad \forall s \in \mathcal{S}, \quad a \in \mathcal{A}.$$

所以  $Q_{\pi^\star}(s, a)$  就是  $Q_\star(s, a)$ 。于是

$$Q_\star(s_t, a_t) = \mathbb{E}_{S_{t+1}, A_{t+1}} \left[ R_t + \gamma \cdot Q_\star(S_{t+1}, A_{t+1}) \mid S_t = s_t, A_t = a_t \right].$$

因为动作  $A_{t+1} = \operatorname{argmax}_A Q_\star(S_{t+1}, A)$  是状态  $S_{t+1}$  的确定性函数, 所以

$$Q_\star(s_t, a_t) = \mathbb{E}_{S_{t+1}} \left[ R_t + \gamma \cdot \max_{A \in \mathcal{A}} Q_\star(S_{t+1}, A) \mid S_t = s_t, A_t = a_t \right].$$

$\square$