附录 A 贝尔曼方程

定理 A.1. 贝尔曼方程（将 Q_π 表示成 Q_π）

假设 R_t 是 S_t、A_t、S_{t+1} 的函数。那么

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

♡

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 $\mathcal{S}_{t+1:} = \{S_{t+1}, S_{t+2}, \cdots\}$ 和 $\mathcal{A}_{t+1:} = \{A_{t+1}, A_{t+2}, \cdots\}$ 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_π 的定义，

$$Q_\pi(s_t, a_t) \;=\; \mathbb{E}_{\mathcal{S}_{t+1:}, \mathcal{A}_{t+1:}}\Big[U_t \,\Big|\, S_t = s_t, A_t = a_t\Big].$$

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

$$\begin{aligned} Q_\pi(s_t, a_t) \;&=\; \mathbb{E}_{\mathcal{S}_{t+1:}, \mathcal{A}_{t+1:}}\Big[R_t + \gamma \cdot U_{t+1} \,\Big|\, S_t = s_t, A_t = a_t\Big] \\ &=\; \mathbb{E}_{\mathcal{S}_{t+1:}, \mathcal{A}_{t+1:}}\Big[R_t \,\Big|\, S_t = s_t, A_t = a_t\Big] \;+\; \gamma \cdot \mathbb{E}_{\mathcal{S}_{t+1:}, \mathcal{A}_{t+1:}}\Big[U_{t+1} \,\Big|\, S_t = s_t, A_t = a_t\Big]. \end{aligned} \tag{A.1}$$

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

$$\mathbb{E}_{\mathcal{S}_{t+1:}, \mathcal{A}_{t+1:}}\Big[R_t \,\Big|\, S_t = s_t, A_t = a_t\Big] \;=\; \mathbb{E}_{S_{t+1}}\Big[R_t \,\Big|\, S_t = s_t, A_t = a_t\Big]. \tag{A.2}$$

等式 (A.1) 右边 U_{t+1} 的期望可以写成

$$\begin{aligned} &\mathbb{E}_{\mathcal{S}_{t+1:}, \mathcal{A}_{t+1:}}\Big[U_{t+1} \,\Big|\, S_t = s_t, A_t = a_t\Big] \\ =\; &\mathbb{E}_{S_{t+1}, A_{t+1}}\Big[\mathbb{E}_{\mathcal{S}_{t+2:}, \mathcal{A}_{t+2:}}\Big[U_{t+1} \,\Big|\, S_{t+1}, A_{t+1}\Big] \,\Big|\, S_t = s_t, A_t = a_t\Big] \\ =\; &\mathbb{E}_{S_{t+1}, A_{t+1}}\Big[Q_\pi(S_{t+1}, A_{t+1}) \,\Big|\, S_t = s_t, A_t = a_t\Big]. \end{aligned} \tag{A.3}$$

由公式 (A.1)、(A.2)、(A.3) 可得定理。 □

定理 A.2. 贝尔曼方程（将 Q_π 表示成 V_π）

假设 R_t 是 S_t、A_t、S_{t+1} 的函数。那么

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

♡

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

定理 A.3. 贝尔曼方程（将 V_π 表示成 V_π）

假设 R_t 是 S_t、A_t、S_{t+1} 的函数。那么

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

♡

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

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

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

♡

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：

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

根据定义，最优动作函数是

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

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

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

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

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

□