形式1：$Q_\pi \leftarrow Q_\pi$

$$U_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \cdots + \gamma^{n-t} R_n$$
$$= \sum_{k=t}^{n} \gamma^{k-t} R_k$$

$$U_{t+1} = R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-t-1} R_n$$
$$= \sum_{k=t+1}^{n} \gamma^{k-t-1} R_k$$

$$\Rightarrow U_t = R_t + \gamma(R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-t-1} R_n)$$
$$= R_t + \gamma U_{t+1}$$

Denote $S_{t+1:} = \{S_{t+1}, S_{t+2}, \cdots\}$ $A_{t+1:} = \{A_{t+1}, A_{t+2}, \cdots\}$

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

替代U_t $$= \mathbb{E}_{S_{t+1:}, A_{t+1:}}[R_t + \gamma U_{t+1} \mid S_t = s_t, A_t = a_t]$$

$$= \mathbb{E}_{S_{t+1:}, A_{t+1:}}[R_t \mid S_t = s_t, A_t = a_t] + \gamma \mathbb{E}_{S_{t+1:}, A_{t+1:}}[U_{t+1} \mid S_t = s_t, A_t = a_t]$$

R_t是S_t, A_t和S_{t+1}的函数（状态转移函数有随机性，因此s_{t+1}也是重要的）

$$\mathbb{E}_{S_{t+1:}, A_{t+1:}}[R_t \mid S_t = s_t, A_t = a_t] = \mathbb{E}_{S_{t+1}}[R_t \mid S_t = s_t, A_t = a_t]$$

而 $U_{t+1} = \sum_{k=t+1}^{n} \gamma^{k-t-1} R_k$ 是 $S_{t+1:} = \{S_{t+1}, S_{t+2}, \cdots S_n\}$

$A_{t+1:} = \{A_{t+1}, A_{t+2}, \cdots A_n\}$ 的函数

$$\mathbb{E}_{S_{t+1:}, A_{t+1:}}[U_{t+1} \mid S_t = s_t, A_t = a_t]$$

$$= \mathbb{E}_{S_{t+1}, A_{t+1}}\left[\mathbb{E}_{S_{t+2:}, A_{t+2:}}[U_{t+1} \mid S_{t+1}, A_{t+1}] \,\middle|\, S_t = s_t, A_t = a_t\right]$$

（$S_{t+2:}, A_{t+2:}$）

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

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

形式2: $Q_\pi \leftarrow V_\pi$

由于 $V_\pi(S_t) = \sum_{a\in\mathcal{A}} \pi(a|s_t)\cdot Q_\pi(s_t,a) = \mathbb{E}_{A_t}[Q(s_t,A_t)]$

同理 $V_\pi(S_{t+1}) = \mathbb{E}_{A_{t+1}}[Q(S_t,A_t)]$

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

$= \mathbb{E}_{S_{t+1}}\left[\mathbb{E}_{A_{t+1}}[Q_\pi(S_{t+1},A_{t+1})]\middle|S_t=s_t,A_t=a_t\right]$

$= \mathbb{E}_{S_{t+1}}[V_\pi(S_{t+1})|S_t=s_t,A_t=a_t]$

$\therefore Q_\pi(s_t,a_t) = \mathbb{E}_{S_{t+1},A_{t+1}}\left[R_t+\gamma V_\pi(S_{t+1})\middle|S_t=s_t,A_t=a_t\right]$

形式3: $V_\pi \leftarrow V_\pi$

由于 $V_\pi(S_t) = \mathbb{E}_{A_t}[Q(S_t,A_t)]$

$V_\pi(S_t) = \mathbb{E}_{A_t}\left[\mathbb{E}_{S_{t+1},A_{t+1}}\left[R_t+\gamma V_\pi(S_{t+1})\middle|S_t=s_t,A_t=a_t\right]\right]$

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

最优贝尔曼方程.

最优策略函数. $\pi^* = \arg\max_\pi Q_\pi(s,a) \quad \forall s,a$

就是最优动作价值函数

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

没什么不同, 只是现在指定了策略.

原先 $Q(s_t,a_t)$ 是关于 $\pi\in\Pi$ 的函数, 现固定 $\pi=\pi^*$

由 π^* 性质, $Q_{\pi^*}(S_{t+1},A_{t+1}) = \max_{A\in\mathcal{A}} Q_{\pi^*}(S_{t+1},A)$

$Q_{\pi^*}(S_t,a_t) = \mathbb{E}_{S_{t+1}}\left[R_t+\gamma\max_{A\in\mathcal{A}} Q_{\pi^*}(S_{t+1},A)\middle|S_t=s_t,A_t=a_t\right]$