彪哥带你学强化学习

5、彻底理解贝尔曼方程

DEEPLY UNDERSTAND REINFORCEMENT LEARNING

讲师：韩路彪

贝 尔 曼 方 程

意义：可以在后面步骤没走完的情况下计算前面步骤的价值，为时序差分、动态规划等算法提供了理论指导

是什么

1. 用后一步的价值表示前一步的价值
2. 类似递推公式
3. 有多种不同表达形式

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

比如：

Vt+1 -> Qt

Qt+1 -> Vt

Vt+1 -> Vt

Qt+1 -> Qt

Q*t+1 -> Q*t

V*t+1 -> V*t

为什么学

1. 是后面算法的基础
2. 可以加深对模型的理解

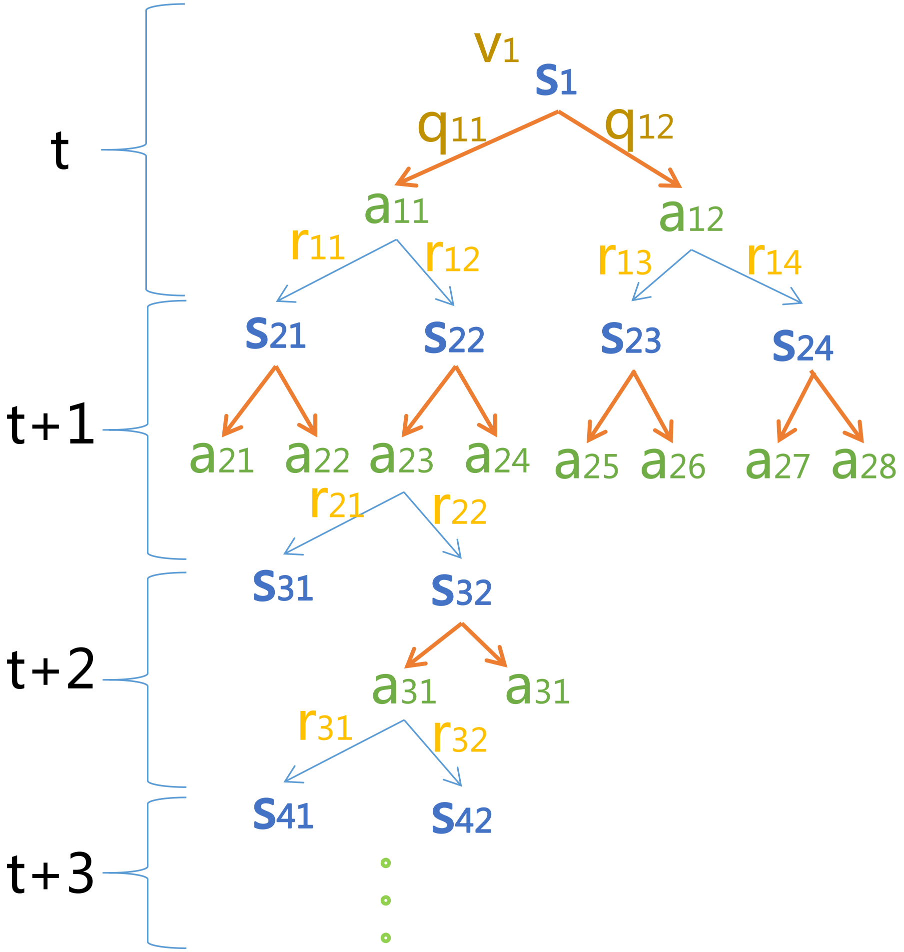
怎么学

以理解含义为主，要会数学公式推导

贝尔曼方程 Vt+1到Qt

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{S_{t+1}\sim p(\cdot|s_t,a_t)}[R_t + \gamma V^{\pi}(S_{t+1})]$$

$$\begin{aligned}
Q^{\pi}(s_t, a_t) &\overset{\triangle}{=} \mathbb{E}^{\pi}(G_t | S_t = s_t, A_t = a_t) \\
&= \mathbb{E}^{\pi}(R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \ldots | S_t = s_t, A_t = a_t) \\
&= \mathbb{E}^{\pi}(R_t + \gamma(R_{t+1} + \gamma R_{t+2} + \ldots) | S_t = s_t, A_t = a_t) \\
&= \mathbb{E}^{\pi}(R_t + \gamma V^{\pi}(S_{t+1}) | S_t = s_t, A_t = a_t) \\
&= \mathbb{E}_{S_{t+1}\sim p(\cdot|s_t,a_t)}[R_t + \gamma V^{\pi}(S_{t+1})]
\end{aligned}$$



贝尔曼方程 Vt+1到Qt

$$Q^\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}\sim p(\cdot|s_t,a_t)}[R_t + \gamma V^\pi(S_{t+1})]$$

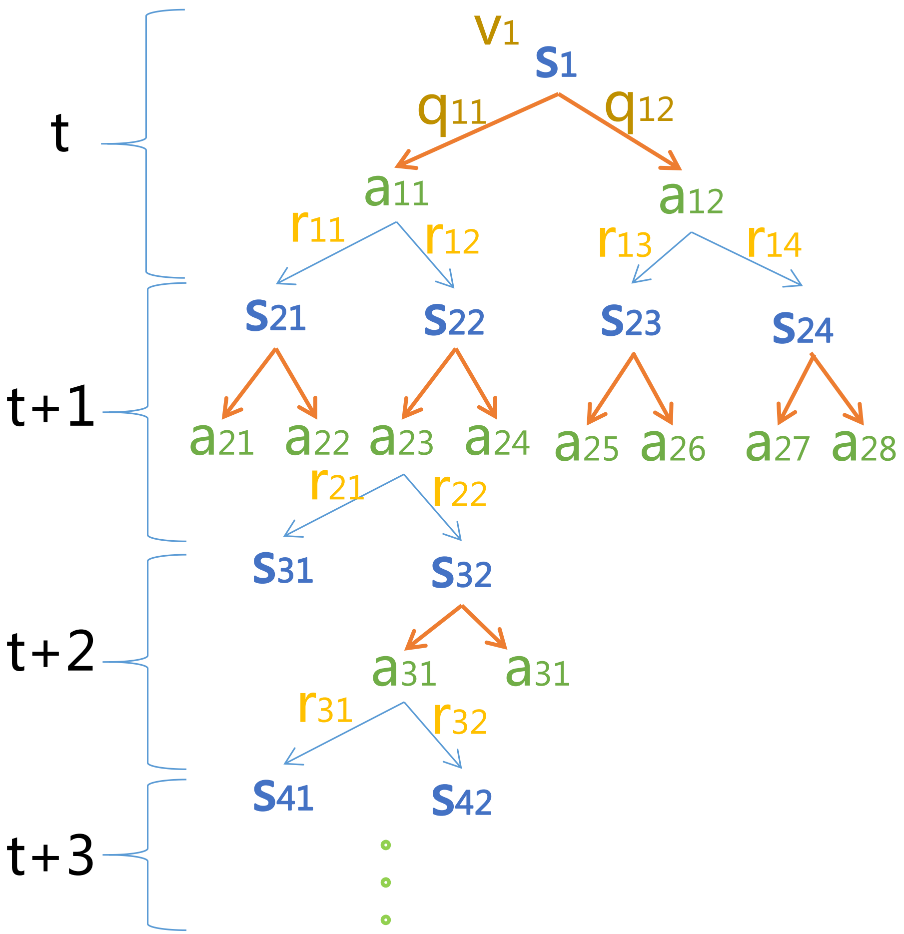
$$Q^\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}\sim p(\cdot|s_t,a_t)}[R(s_t, a_t, S_{t+1}) + \gamma V^\pi(S_{t+1})]$$

$$Q^\pi(s_t, a_t) = \mathbb{E}_{S_{t+1}}[R(s_t, a_t, S_{t+1}) + \gamma V^\pi(S_{t+1})|S_t = s_t, A_t = a_t]$$

$$Q^\pi(s_t, a_t) = \sum_{S_{t+1}\sim p(\cdot|s_t,a_t)} p(S_{t+1}|s_t, a_t)[R(s_t, a_t, S_{t+1}) + \gamma V^\pi(S_{t+1})]$$

贝尔曼方程 Qt+1到Vt

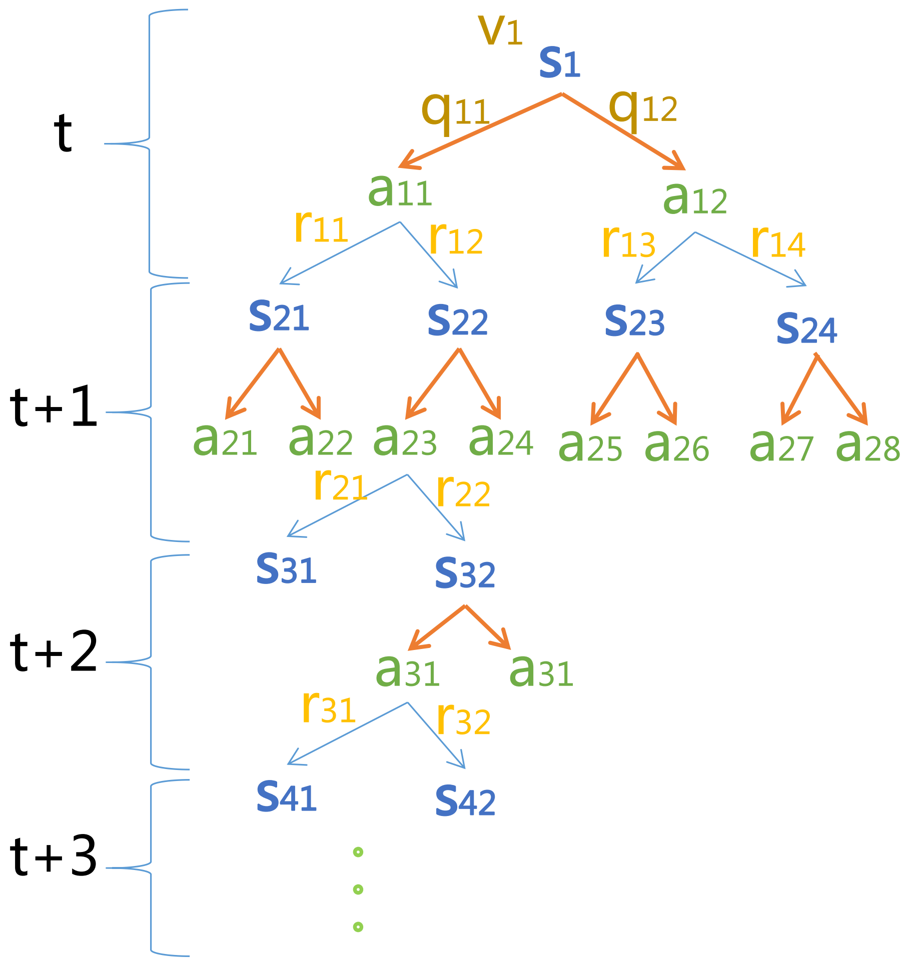
$$V^{\pi}(s_t) = R_t + \gamma \mathbb{E}_{S_{t+1},A_{t+1}}[Q^{\pi}(S_{t+1},A_{t+1})|S_t = s_t]$$



贝尔曼方程 Vt+1到Vt

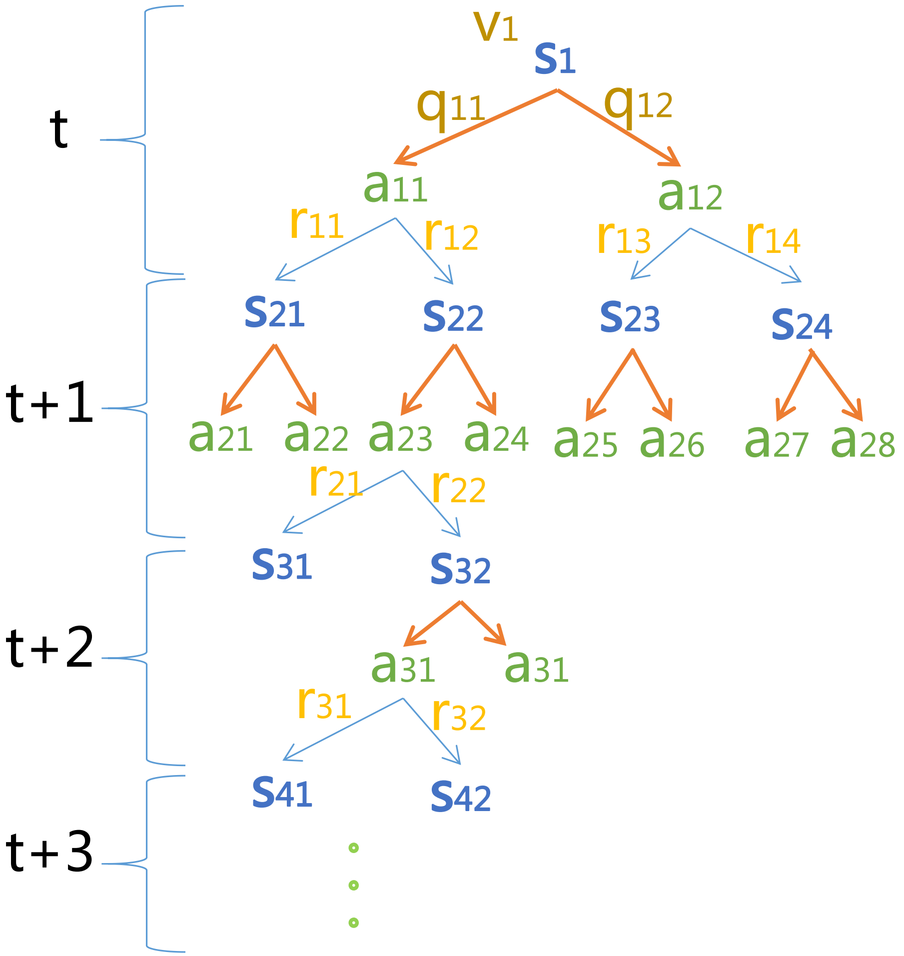
$$V^\pi(s_t) = R_t + \gamma \mathbb{E}_{S_{t+1}}[V^\pi(S_{t+1})|S_t = s_t]$$

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



贝尔曼方程 Qt+1到Qt

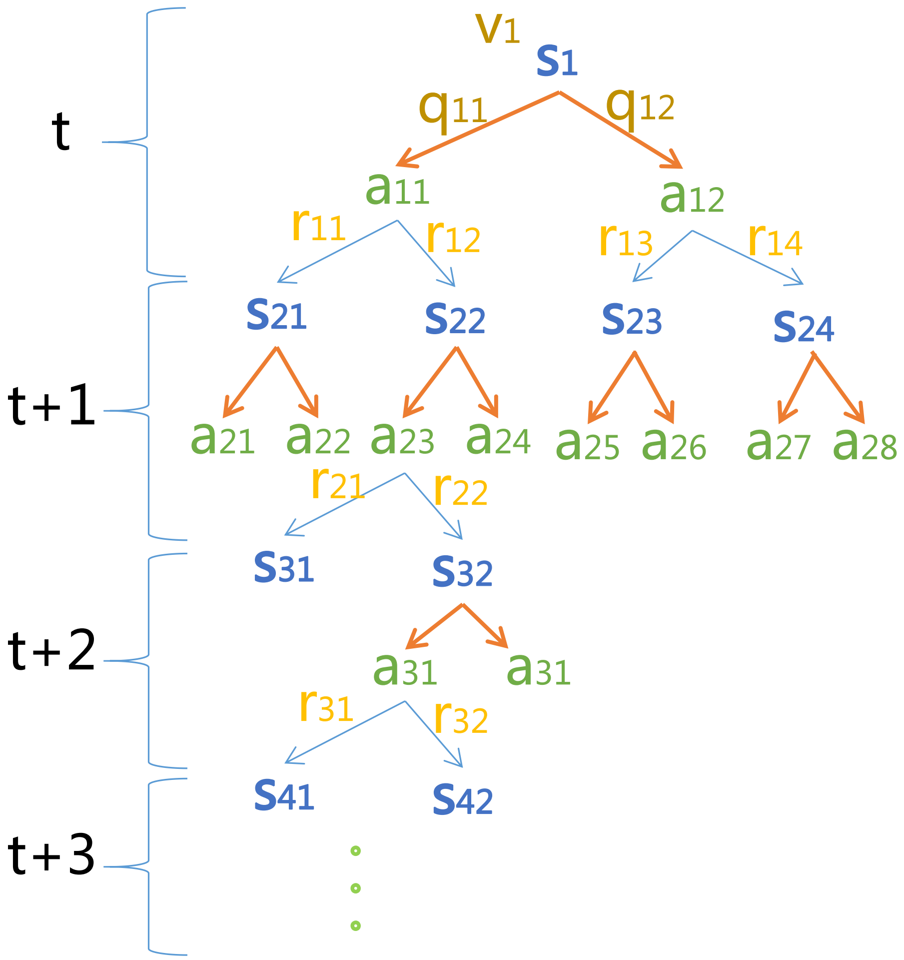
$$Q^{\pi}(s_t, a_t) = \mathbb{E}^{\pi}_{S_{t+1}, A_{t+1}}[R(s_t, a_t, S_{t+1}) + \gamma Q^{\pi}(S_{t+1}, A_{t+1}) | S_t = s_t, A_t = a_t]$$



贝尔曼最优方程 Q*t+1到Q*t

$$Q^{\star}(s_t, a_t) = \mathbb{E}_{S_{t+1} \sim p(\cdot|s_t,a_t)}\left[R(s_t, a_t, S_{t+1}) + \gamma \max_{A_{t+1} \in \mathcal{A}} Q^{\star}(S_{t+1}, A_{t+1})\right]$$

$$Q^{\star}(s_t, a_t) = \sum_{S_{t+1} \sim p(\cdot|s_t,a_t)} p(S_{t+1}|s_t, a_t)[R(s_t, a_t, S_{t+1}) + \gamma \max_{A_{t+1} \in \mathcal{A}} Q^{\star}(S_{t+1}, A_{t+1})]$$



贝尔曼最优方程 V*t+1到V*t

$$V^\star(s_t) = \max_{A_t \in \mathcal{A}}\Big(\sum_{S_{t+1} \sim p(\cdot|s_t, A_t)} p(S_{t+1}|s_t, A_t)[R(s_t, A_t, S_{t+1}) + \gamma V^\star(S_{t+1})]\Big)$$