0. 写在前面

个人学习笔记，如有侵权，请联系792706244@qq.com 删除。

1. 从MDP看强化学习

MDP：马尔科夫决策过程。
马尔科夫决策过程由四元组 (S, A, P, R, γ) 描述， 其中：

$$S = \{s_1, s_2, ...\}：有限状态集$$

$$A = \{a_1, a_2, ...\}：有限动作集$$

$$P：S \times A \times S矩阵：状态转移概率$$

$$R：回报函数$$

$$\gamma：折扣因子$$

马尔科夫决策过程的状态转移概率是包含动作的，即：

$$P_{ss^{'}}^{a} = P(S_{t+1} = s^{'}|S_t = s, A_t = a) \tag{1}$$

强化学习的目标是给定马尔科夫决策过程，寻找最优策略。所谓策略是指状态到动作的映射，策略常用符号 π 表示，它是指给定状态时，动作集上的一个分布，即

$$\pi(a|s) = P(A_t = a|S_t = s) \tag{2}$$

公式(2)的含义是：策略c在每个状态 s 指定一个动作概率。如果给出的策略π是确定性的，那么策略π在每个状态s指定一个确定的动作。
那么在给定一个策略π的时候，我们就可以计算累计回报了。

$$G = R_t + \gamma R_{t+1} + ... = \sum_{k=0}^{+\infty} \gamma^k R_{t+k+1} \tag{3}$$

累计回报G是随机的，跟策略选择π有关， 不是一个确定值，因此无法进行描述，但是其期望是个确定值，可以作为状态函数值的定义。
状态函数定义如下，当智能体系采用策略π的时候， 累计回报服从一个分布，累积回报在状态s处的期望定义为状态值函数：

$$v_\pi(s) = E_\pi[\sum_{k=0}^{+\infty} k^\gamma R_{t+k+1}|S_t = s] \tag{4}$$

注意：状态值函数是与策略π相对应的，这是因为策略π决定了累积回报G的状态分布。

对应的，状态行为值函数如下：

$$q_\pi(s,a) = E_\pi[\sum_{k=0}^{+\infty} k^\gamma R_{t+k+1} | A_t = a, S_t = t] \tag{5}$$

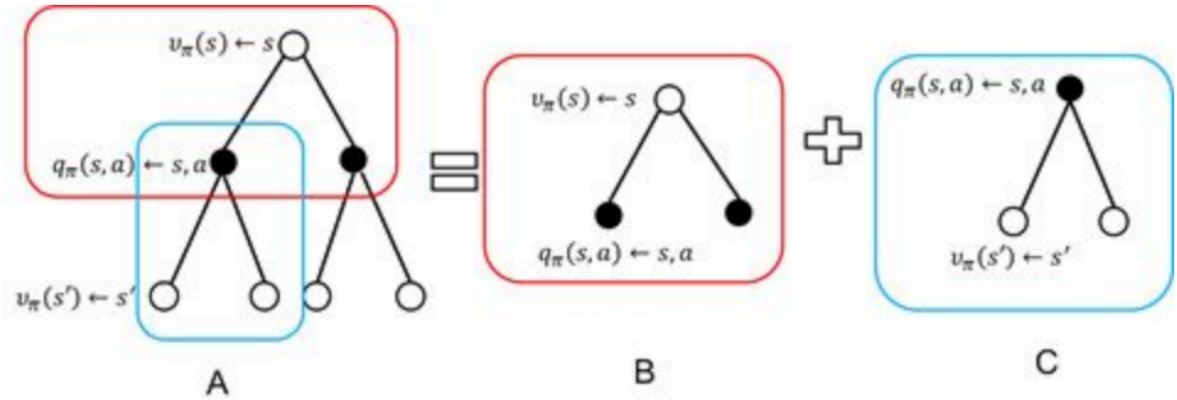
状态值函数的bellman方程如下：

$$\begin{aligned} v_\pi(s_t) &= E_\pi[\sum_{k=0}^{+\infty} k^\gamma R_{t+k+1}] \\ &= E_\pi[R_t + \sum_{k=1}^{+\infty} k^\gamma R_{t+k+1}] \\ &= E_\pi[R_t + \gamma v(s_{t+1})] \end{aligned} \tag{6}$$

同理，状态行为值函数的bellman方程如下：

$$q_\pi(s_t,a_t) = E_\pi[R_t + \gamma q_\pi(s_{t+1},a_{t+1})|S = s_t, A = a_t] \tag{7}$$

下图展示了状态值函数和行为状态值函数的具体计算过程：



由上图中B可知：

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(a,s) \tag{8}$$

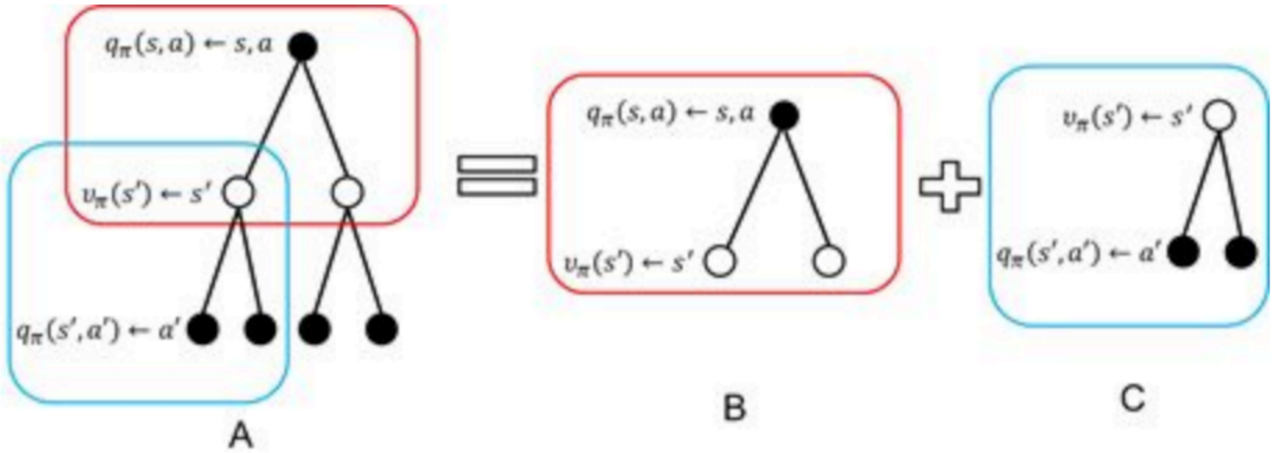
由上图中C可知：

$$q_\pi(a,s) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s') \tag{9}$$

将式（9）代入式（8）得：

$$v_\pi(s) = \sum_{a \in A} \pi(a|s)(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')) \tag{10}$$

下图展示了状态行为值的具体计算过程：



由上图中C可知：

$$v_\pi(s^{'}) = \sum_{a_{'} \in A} \pi(a^{'}|s^{'}) q_\pi(a^{'}, s^{'}) \tag{11}$$

将式（11）代入式（9）可以得：

$$q_\pi(a,s) = R_s^a + \gamma \sum_{s^{'} \in S} P_{ss^{'}}^a \sum_{a^{'} \in A} \pi(a^{'}|s^{'}) q_\pi(a^{'}, s^{'}) \tag{12}$$

定义：最优状态值函数$v^*(s)$ 为在所有策略中值最大的值函数即：

$$v^*(s) = \arg\underset{\pi}{max}\, v_\pi(s) \tag{13}$$

定义：最优行为-状态值函数$q^*(a,s)$ 为在所有的策略中值最大的值函数即：

$$q*(a,s) = \arg\underset{\pi}{max}\, q_\pi(a,s) \tag{14}$$

将式（10）代入式（13）可得最优状态值的贝尔曼方程：

$$v^*(s) = \max_a R_s^a + \gamma \sum_{s^{'} \in S} P_{ss^{'}}^a v_\pi^*(s^{'}) \tag{15}$$

将式（12）代入式（14）可得到最优状态-转移值的贝尔曼方程：

$$q^*(a,s) = \max_a R_s^a + \gamma \sum_{s^{'} \in S} P_{ss^{'}}^a \underset{a_{'}}{max}\, q_\pi^*(s^{'}, a^{'}) \tag{16}$$

若已知最优状态-动作值函数，最优策略可通过直接最大化$q^*(s,a)$决定：

$$\pi^*(a|s) = \begin{cases} 1 & a = \arg max_{a \in A}\, q^*(s,a) \\ 0 & otherwise \end{cases}$$