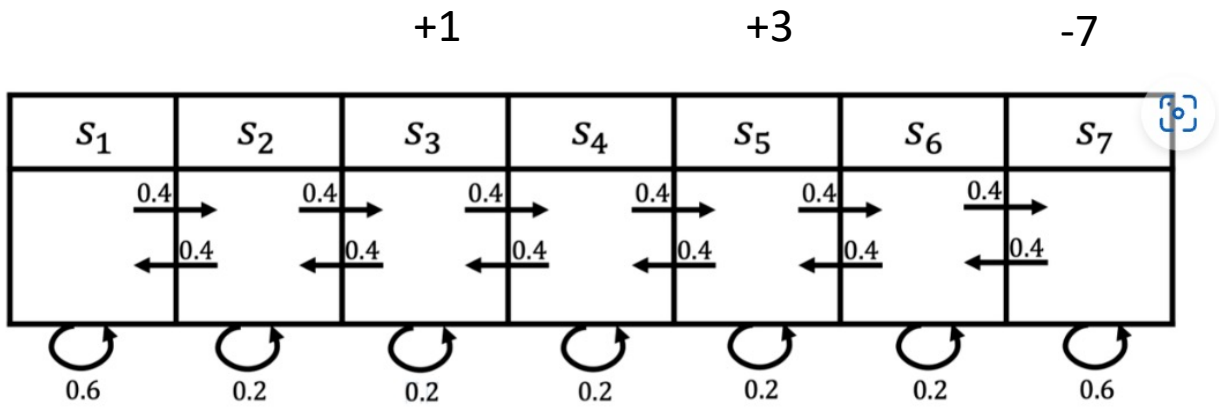
马尔可夫决策过程、策略、动作价值函数Q

马尔可夫决策过程 Markov Decision Process



马尔可夫奖励过程：一艘随波逐流的小船，比如从s2出发，我有0.4概率转移到s3，有0.4概率转移到s1，有0.2概率停留在s2，进入一个状态我可能得到奖励（比如进入s3得到+1），但是我无法控制船向哪里走（没有策略）

马尔可夫奖励过程：一艘人为控制的小船

比如从s2出发，我有三种动作选择：停在s2，向左到s1，或者向右到s3

如果我选择向左转到s1，我可能有0.6概率转移到s1，有0.2概率转移到s3，有0.2概率停留在s2（可能受到水流的影响）

如果我选择向右转到s3，我可能有0.5概率转移到s3，有0.2概率转移到s1，有0.3概率停留在s2

此时，我们具有了决策的能力（马尔可夫决策过程）

马尔可夫决策过程vs马尔可夫奖励过程

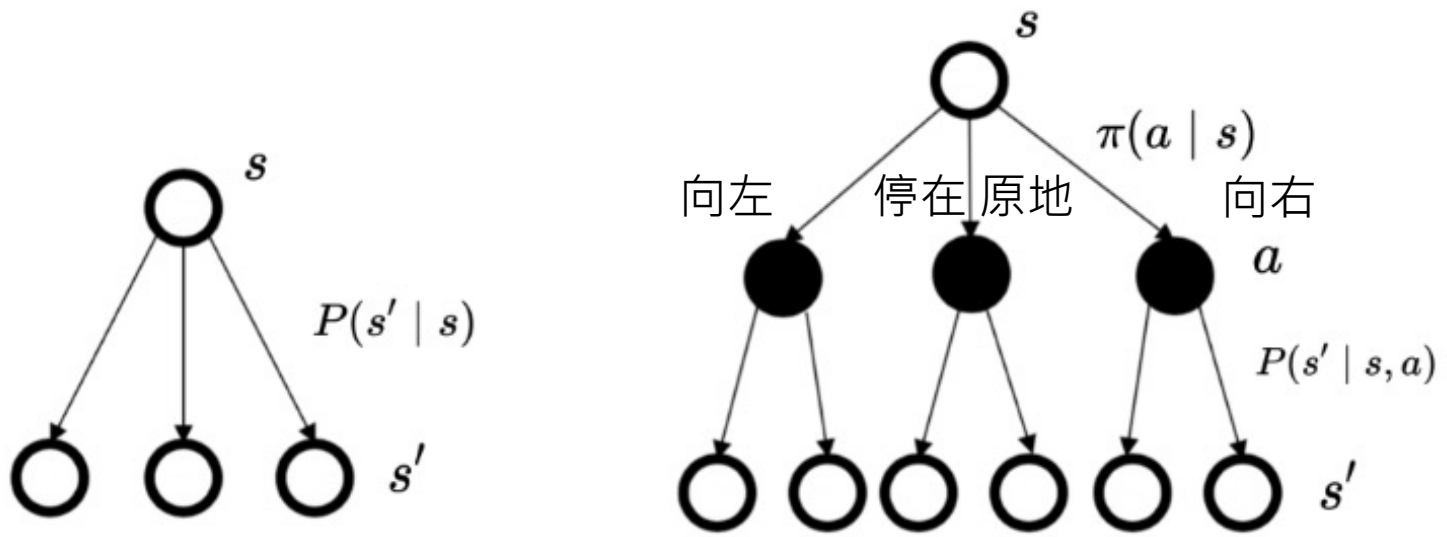


图 2.9 马尔可夫决策过程与马尔可夫过程/马尔可夫奖励过程的状态转移的对比

马尔可夫决策过程的价值函数Q

策略：在某一个状态应该采取什么样的动作

$$\pi(a \mid s) = p\left(a_t = a \mid s_t = s\right)$$

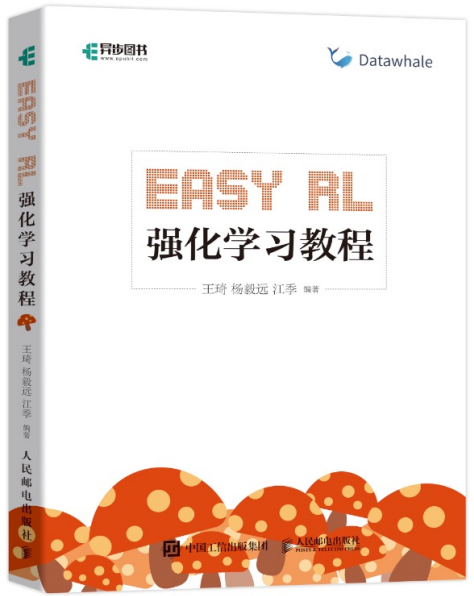
这里我们另外引入了一个 **Q 函数（Q-function）**。Q 函数也被称为**动作价值函数（action-value function）**。Q 函数定义的是在某一个状态采取某一个动作，它有可能得到的回报的一个期望，即

$$Q_\pi(s, a) = \mathbb{E}_\pi\left[G_t \mid s_t = s, a_t = a\right] \tag{2.4}$$

这里的期望其实也是基于策略函数的。所以我们需要对策略函数进行一个加和，然后得到它的价值。 对 Q 函数中的动作进行加和，就可以得到价值函数：

$$V_\pi(s) = \sum_{a \in A} \pi(a \mid s) Q_\pi(s, a) \tag{2.5}$$

本章小结



- 什么是马尔可夫决策过程MDP？
- MDP和马尔可夫奖励过程MRP有什么区别？
- 如何定义马尔可夫决策过程中的价值函数Q和策略？
- 如何将马尔可夫决策过程中的动作状态价值Q转化为状态价值V？

下一章：马尔可夫决策过程的贝尔曼方程

Credit goes to: EasyRL