马尔科夫决策过程

一、马尔科夫过程

- 马尔科夫性
  - 未来状态的条件概率分布仅依赖于当前状态
- 马尔科夫链
  - 具备马尔可夫性的随机变量序列
- 跟 HMM、CRF 的假设一致

二、马尔科夫奖励过程

- 定义
  - 马尔科夫链加上奖励函数
- 回报和价值函数
  - 回报（return）可以定义为奖励的逐步叠加，假设时刻 *t* 后的奖励序列为 rt+1, rt+2, rt+3 · · · ，则回报为
    - $$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \ldots + \gamma^{T-t-1} r_T$$
    - T 是最终时刻，γ 是折扣因子，越往后得到的奖励，折扣越多。这说明我们更希望得到现有的奖励，对未来的奖励要打折扣
  - 对于马尔可夫奖励过程，状态价值函数被定义成回报的期望
    - $$\begin{aligned} V^t(s) &= \mathbb{E}\left[G_t \mid s_t = s\right] \\ &= \mathbb{E}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots + \gamma^{T-t-1} r_T \mid s_t = s\right] \end{aligned}$$
  - 使用折扣因子的原因
    - 有的马尔科夫链有环，但我们可能有无限奖励
    - 我们无法建立完美的虚拟环境，因此对未来的预估不一定准确
- 贝尔曼方程
  - $$V(s) = \underbrace{R(s)}_{\text{即时奖励}} + \gamma \underbrace{\sum_{s' \in S} p(s' \mid s) V(s')}_{\text{未来奖励的折扣总和}}$$
- 计算马尔科夫奖励过程价值的迭代算法
  - 动态规划
  - 蒙特卡洛采样学习
  - 时序查分学习

三、马尔科夫决策过程

- 相对于马尔科夫奖励过程，马尔科夫决策过程多了决策（动作）
- 这样的话，状态转移不仅取决于当前状态，也取决于当前动作
  - $p\left(s_{t+1} \mid s_t, a_t\right)=p\left(s_{t+1} \mid h_t, a_t\right)$
- 策略定义了在某个状态下，应该采取什么样的动作
  - $\pi(a \mid s)=p\left(a_t=a \mid s_t=s\right)$
-

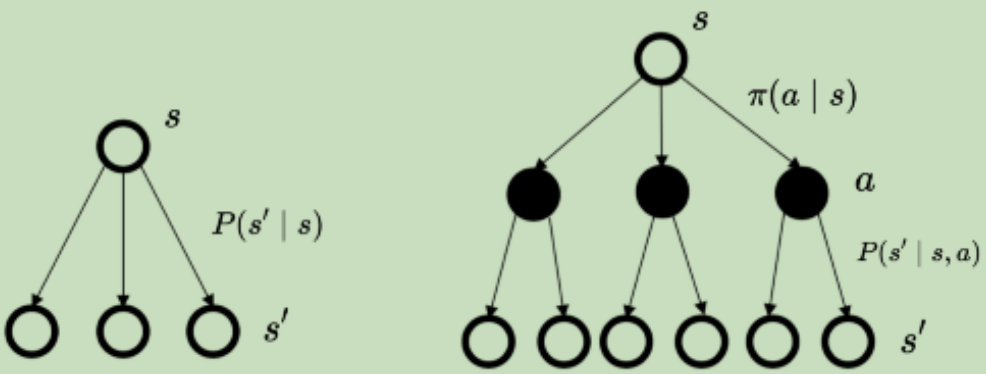

图 2.9　马尔可夫决策过程与马尔可夫过程/马尔可夫奖励过程的状态转移的对比