# 预测和控制,马尔可夫性质,马尔可夫过程,马尔可夫奖励过程， 蒙特卡洛法

# 预测与控制

预测（prediction）和控制（control）

| 任务 | 输入 | 输出 |
|---|---|---|
| 预测 | 马尔可夫决策过程<$S,A,P,R,\gamma$>、策略 $\pi$ | 每个状态的价值函数V |
| 控制 | 马尔可夫决策过程<$S,A,P,R,\gamma$> | 最佳策略π、最佳价值函数V |

预测问题是给定一个策略，我们要确定它的价值函数是多少。

而控制问题是在没有策略的前提下，我们要确定最佳的价值函数以及对应的决策方案。

递进关系：

初始化策略->计算当前策略各状态价值(预测)->优化策略->计算新的策略各状态价值->继续优化...

# 马尔可夫性质（Markov property）

$$p\left(X_{t+1}=x_{t+1} \mid X_{0:t}=x_{0:t}\right)=p\left(X_{t+1}=x_{t+1} \mid X_t=x_t\right)$$

其中，$X_{0:t}$ 表示变量集合 $X_0, X_1, \cdots, X_t$，$x_{0:t}$ 为在状态空间中的状态序列 $x_0, x_1, \cdots, x_t$。马尔可夫性质也可以描述为给定当前状态时，将来的状态与过去状态是条件独立的。如果某一个过程满足**马尔可夫性质**，那么未来的转移与过去的是独立的，它只取决于现在。马尔可夫性质是所有马尔可夫过程的基础。

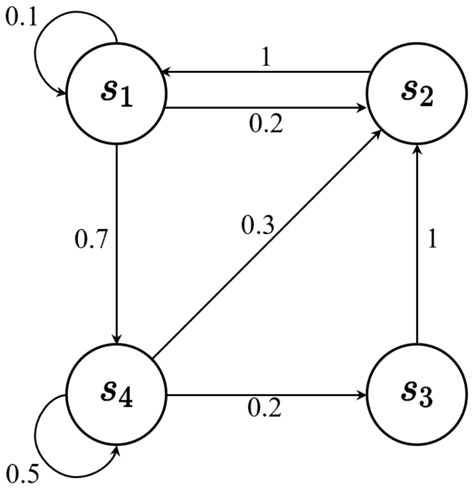
**满足：**

**天气预报**：假设我们只考虑“晴天”和“雨天”两种状态，且明天的天气只依赖于今天的天气，而与昨天及更早的天气无关。例如，如果我们知道今天是晴天，那么明天是晴天的概率就是一定的，不会因为昨天是雨天而改变。

**不满足：**

**股票价格**：股票的价格不仅仅依赖于当前的价格，还可能受到过去价格的影响。
例如，如果一只股票在过去一段时间内持续上涨，那么投资者可能预期这只股票会继续上涨，从而推高股票的价格。
而不是仅仅取决于这只股票昨天有没有涨。

# 马尔可夫链（Markov Chain）离散时间的马尔可夫过程

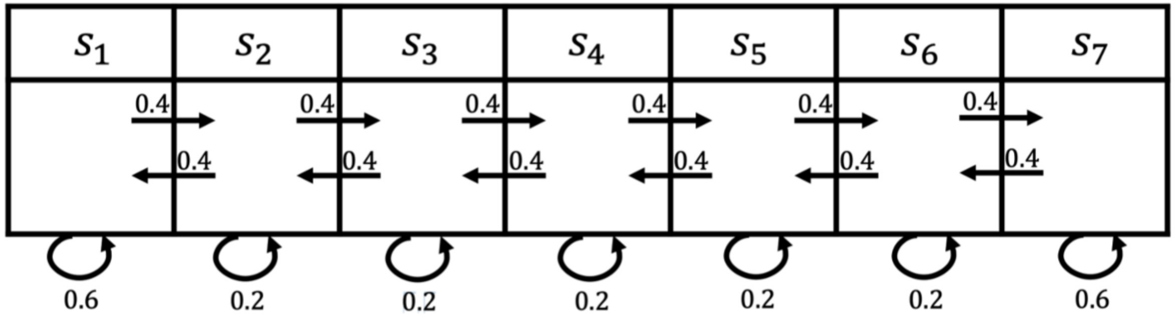



我们可以用**状态转移矩阵**（state transition matrix）$\boldsymbol{P}$ 来描述状态转移 $p\left(s_{t+1}=s' \mid s_t=s\right)$:

$$
\boldsymbol{P}=\left(\begin{array}{cccc}
p\left(s_1 \mid s_1\right) & p\left(s_2 \mid s_1\right) & \ldots & p\left(s_N \mid s_1\right) \\
p\left(s_1 \mid s_2\right) & p\left(s_2 \mid s_2\right) & \ldots & p\left(s_N \mid s_2\right) \\
\vdots & \vdots & \ddots & \vdots \\
p\left(s_1 \mid s_N\right) & p\left(s_2 \mid s_N\right) & \ldots & p\left(s_N \mid s_N\right)
\end{array}\right)
$$

状态转移矩阵类似于条件概率（conditional probability），它表示当我们知道当前我们在状态 $s_t$ 时，到达下面所有状态的概率。所以它的每一行描述的是从一个节点到达所有其他节点的概率。

# 马尔可夫链例子




从s4开始采样，得到三个轨迹：

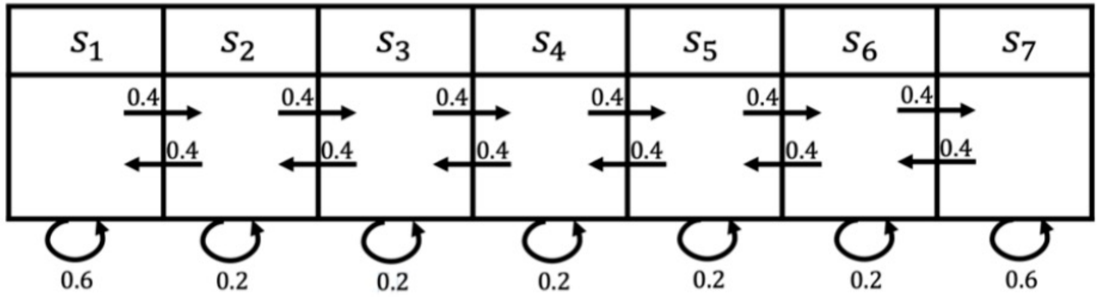
$s_4, s_5, s_6, s_7$

$s_4, s_3, s_2, s_1$

$s_4, s_5, s_6, s_6$

$$\boldsymbol{R} = [5, 0, 0, 0, 0, 0, 10]$$

## +奖励函数 进化！马尔可夫奖励过程：



# 马尔可夫奖励过程（Markov reward process, MRP）

马尔可夫链加上奖励函数（reward function）

奖励函数 $R$ 是一个期望，表示当我们到达某一个状态的时候，可以获得多大的奖励。

折扣因子 $\gamma$

这里我们进一步定义一些概念。**范围（horizon）** 是指一个回合的长度（每个回合最大的时间步数），它是由有限个步数决定的。 **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 $r_{t+1}, r_{t+2}, r_{t+3}, \cdots$，则回报为

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \ldots + \gamma^{T-t-1} r_T$$

其中，$T$是最终时刻，$\gamma$ 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。当我们有了回报之后，就可以定义状态的价值了，就是**状态价值函数（state-value function）**。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，即

$$\begin{aligned} V^t(s) &= \mathbb{E}\left[G_t \mid s_t = s\right] \\ &= \mathbb{E}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots + \gamma^{T-t-1} r_T \mid s_t = s\right] \end{aligned}$$

其中，$G_t$ 是之前定义的**折扣回报（discounted return）**。我们对$G_t$取了一个期望，期望就是从这个状态开始，我们可能获得多大的价值。所以期望也可以看成未来可能获得奖励的当前价值的表现，就是当我们进入某一个状态后，我们现在有多大的价值。

$$\boldsymbol{R} = [5, 0, 0, 0, 0, 0, 10]$$

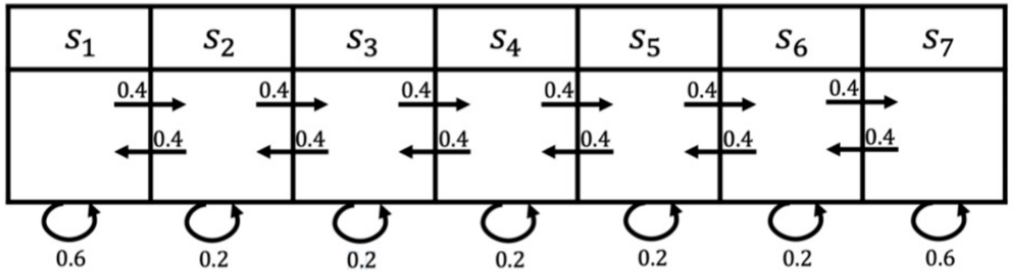



图 2.4 马尔可夫奖励过程的例子

我们对 4 步的回合（$\gamma = 0.5$）来采样回报 $G$。

（1）$s_4, s_5, s_6, s_7$的回报 : $0 + 0.5 \times 0 + 0.25 \times 0 + 0.125 \times 10 = 1.25$

（2）$s_4, s_3, s_2, s_1$的回报 : $0 + 0.5 \times 0 + 0.25 \times 0 + 0.125 \times 5 = 0.625$

（3）$s_4, s_5, s_6, s_6$的回报 : $0 + 0.5 \times 0 + 0.25 \times 0 + 0.125 \times 0 = 0$

$$G_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 r_{t+4} + \ldots + \gamma^{T-t-1} r_T$$

$$\begin{aligned} V^t(s) &= \mathbb{E}\left[G_t \mid s_t = s\right] \\ &= \mathbb{E}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots + \gamma^{T-t-1} r_T \mid s_t = s\right] \end{aligned}$$

蒙特卡洛法：我们可以生成很多轨迹，然后把轨迹都叠加起来。

比如我们可以从 *s*4 开始，采样生成很多轨迹，把这些轨迹的回报都计算出来。

然后将其取平均值作为我们进入 *s*4 的价值。

# 本章小结

- 什么是预测？什么是控制？他们之间的关系是？
- 什么是马尔可夫性质？
- 什么是马尔可夫链？
- 什么是马尔可夫奖励过程？
- 如何计算一段轨迹的回报？
- 如何计算某个状态的价值？

下一章：贝尔曼方程

Credit goes to: EasyRL