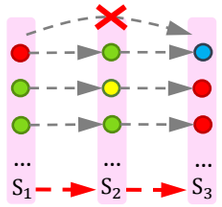
马尔可夫决策过程(MDP)入门篇

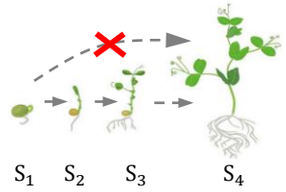
张晨皓2021年 东南大学交通学院 程琳教授研究室

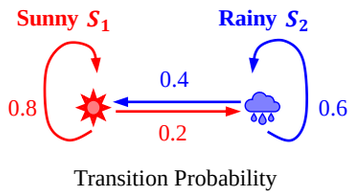
一、马尔可夫过程(State + Probability)

具有**马尔可夫性质**的有限随机状态序列$S_1, S_2, \dots$

$P(S_{t+1}|S_t) = P(S_{t+1}|S_1, \dots, S_t)$ **无后效性**





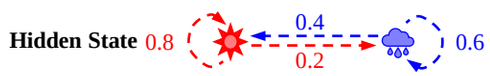


Transition Probability

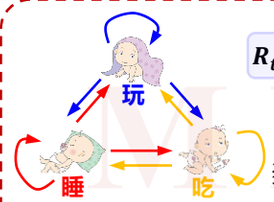
	S_1	S_2
S_1	0.8	0.2
S_2	0.4	0.6

Transition Matrix

状态 + 状态转移概率 = **马尔可夫过程**



马尔可夫过程 + 观测 + 释放概率 = **隐马尔可夫模型**



R_{t+1}:从状态S_t**转换**为S_{t+1}的**奖励**

$R = F(State1 \rightarrow State2)$

奖励与状态转移(转换)相伴生

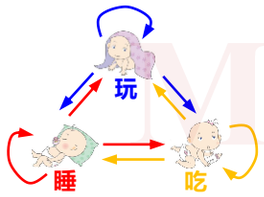
马尔可夫过程 + 奖励 = **马尔可夫奖励过程**

应用

- ✓ 评估问题
  已知 和参数
  求解
- ✓ 解码问题
  已知 和参数
  求解
- ✓ 学习问题
  已知
  已知
  求解 参数

动作产生不确定的状态转移　　确定的状态转移

1.翻身；2.拿起奶瓶；3.蹬腿；4.哭；……



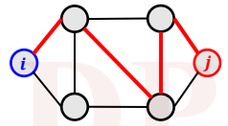
不同的状态，会有不同的动作集合

同一状态下，各动作以概率分布体现

某状态下的某动作，导致不同的状态转移

$P(State1 \rightarrow State2|Action)$

马尔可夫奖励过程 + 动作 = **马尔可夫决策过程**



最短路问题中路段**成本**

即为状态转换**奖励**

动态规划

二、马尔可夫奖励过程(State + Probability + Reward)

即时奖励 R_{t+1}

Reward

从状态S_t转换为S_{t+1}的 即时奖励

上一状态　下一状态

即时奖励 与**状态转移**相伴生

只要发生状态转移，就一定会产生对应的即时奖励

长期奖励 G_t

Return

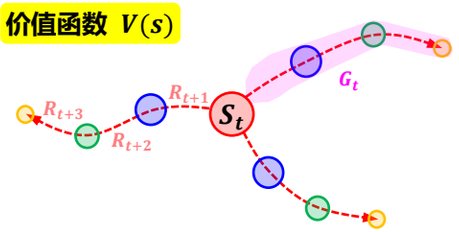
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

γ : 衰减系数,表示即时奖励在当前的折扣值

长期奖励 与**状态转移链**相伴生

不同的状态转移链对应不同的长期奖励

价值函数 $V(s)$



如何找到与**状态**相伴生的量，进而评价状态的质量呢？

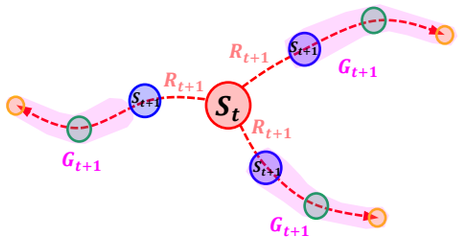
- 存在由状态S_t为起点的**多条**状态转移链
- 每条链都有一个长期折扣奖励G_t
- 则用以S_t为起点的所有状态转移链的G_t的**期望**来作为衡量S_t价值的指标
- $V(S_t) = E[G_t|s = S_t]$ 状态 与 **价值函数** 相伴生

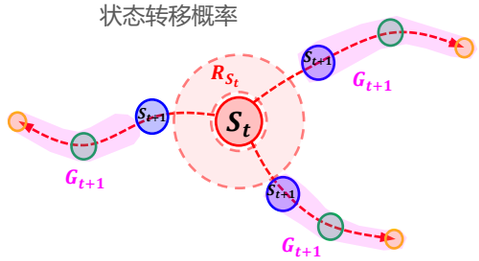
三、马尔可夫决策过程(State + Probability + Reward + Action)

贝尔曼期望方程

$$V(S_t) = \mathrm{E}[R_{t+1} + \gamma V(S_{t+1})|s = S_t] = \underbrace{R_{S_t}}_{\mathrm{E}(R_{t+1})} + \gamma \underbrace{\sum_{S_{t+1} \in S} P_{S_t S_{t+1}} V(S_{t+1})}_{\mathrm{E}(V(S_{t+1}))}$$

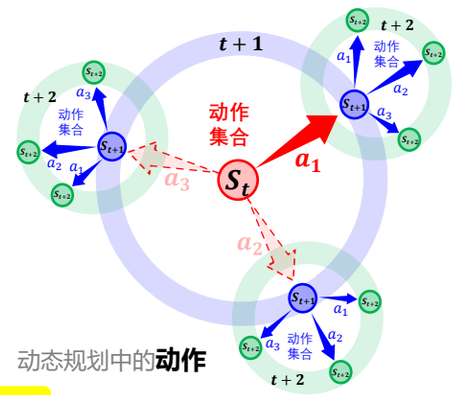
状态转移概率



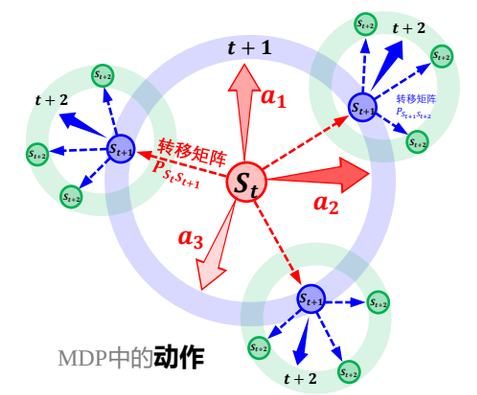


动作a

$a \in A_t$



动态规划中的**动作**

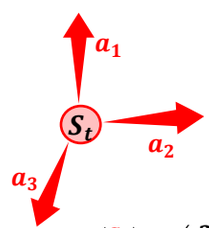


MDP中的**动作**

策略$\pi(a|s)$

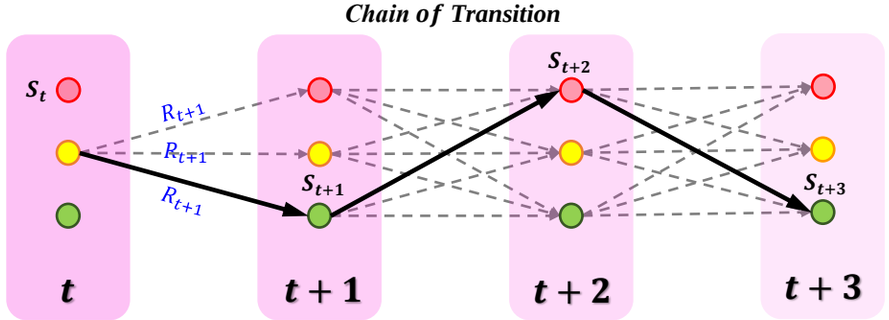
π是给定状态的**动作分布**

$\pi(a|S_t) = P[A_t = a|s = S_t]$



$$\pi(a|S_t) = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} .25 \\ .30 \\ .45 \end{pmatrix}$$

Chain of Transition



G_t R_{S_t} $R_{S_{t+1}}$ $R_{S_{t+2}}$

贝尔曼期望方程

$$V(S_t) = \underbrace{R_{S_t}}_{\text{状态转移概率}} + \gamma \sum_{S_{t+1} \in S} P_{S_t S_{t+1}} V(S_{t+1}) = \underbrace{\sum_{a \in A} \pi(a|s)}_{\text{策略}} [R^a_{S_t} + \gamma \underbrace{\sum_{S_{t+1} \in S} P^a_{S_t S_{t+1}}}_{\text{状态转移概率}} V_\pi(S_{t+1})]$$