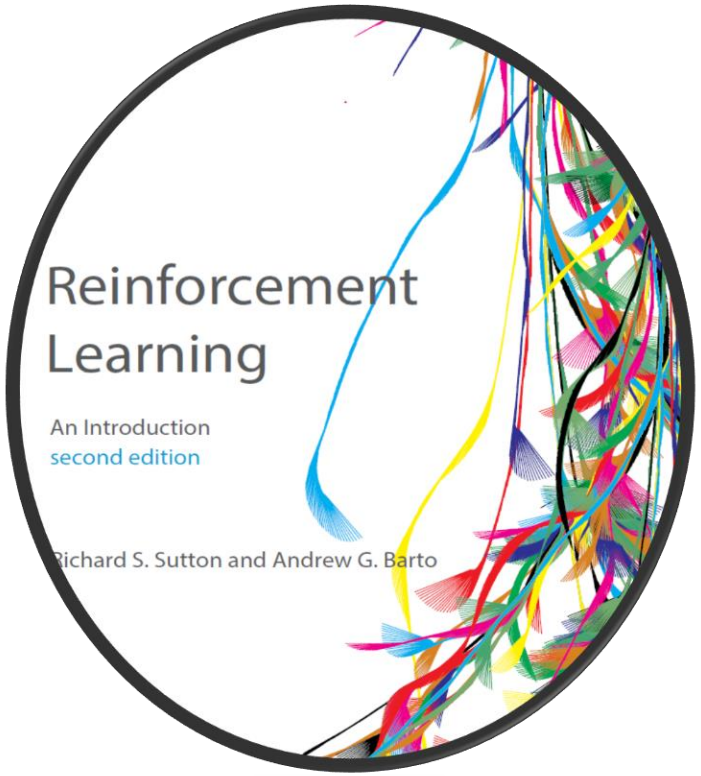



# 第三章：有限马尔可夫决策过程

**何新卫**

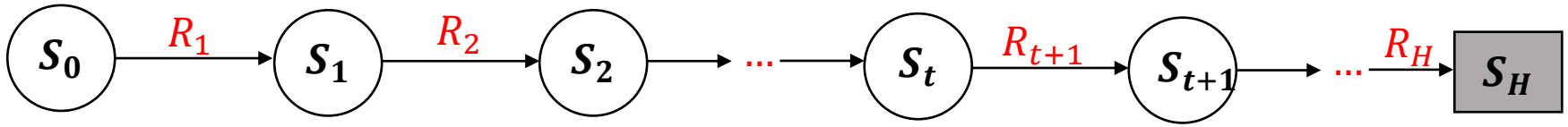
**华中农业大学信息学院**

# 上讲回顾

马尔可夫过程
(Markov process, MP)
$< \boldsymbol{p}, \mathcal{S} >$

马尔可夫奖励过程
(Markov reward process, MRP)
$< \boldsymbol{p}, \mathcal{S}, \mathcal{R} >$

# 回报（Return）




- 智能体与环境交互，有如下几种情形
- **回合制**：交互过程分成很子序列，这些子序列有被称为回合(episodes)， 每个子序列有终止状态。比如，玩了一把王者荣耀，玩了一局超级玛丽，下了一局围棋等等。这些任务也被成为回合制任务（**episodic tasks**），回合制有个特殊状态叫做终止状态，即终止与于$H$， 注意$H$也是一个随机变量，不同回合， $H$也不一样。
- **持续性任务**：任务没有结束，即 $H = \infty$
- $H$：**被称为范围(Horizon)**

# 折扣回报（Return）：回顾

- 智能体在$t$时刻，其目标是最大化期望汇报

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

- **折扣系数**（discount rate）：$0 \leq \gamma \leq 1$超参数，对未来的奖励打折扣
  - ✓**$\gamma$=0**: 只关注即时奖励
  - ✓**$\gamma$=1**：关注长期的累计奖励
  - ✓**$0 < \gamma < 1$**：$\gamma$越大，越关注长期奖励，接近0，关注短期奖励

# $G_t$和$G_{t+1}$计算回报的关系

$$
\begin{aligned}
G_t &\triangleq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots \\
&= R_{t+1} + \gamma (R_{t+2} + \gamma R_{t+3} + \gamma^2 R_{t+4} + \cdots) \\
&= R_{t+1} + \gamma\, G_{t+1}
\end{aligned}
$$

- 上述公式对于$t < H$均成立。
- 对于$t = H$， 我们定义$G_H = 0$,上述公式也满足。

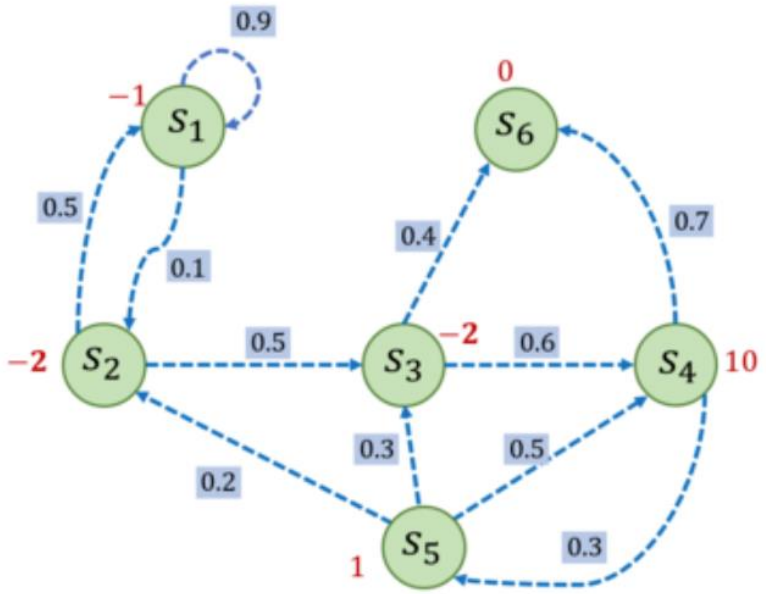
# 例题

- 题目：智能体与环境进行一个回合，获得如下奖励序列（$H = 5$），

$$R_1 = -1, R_2 = 2, R_3 = 6, R_4 = 3, R_5 = 2$$

- 问题：$\gamma = 0.5$，计算$G_0$，$G_1$，$G_2, \ldots, G_5$（提示：从后往前计算）

# 马尔可夫奖励过程（MRP）




状态转移矩阵

$$P = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \end{array} \begin{array}{c} \begin{array}{cccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \end{array} \\ \begin{bmatrix} 0.9 & 0.1 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.6 & 0 & 0.4 \\ 0 & 0 & 0 & 0 & 0.3 & 0.7 \\ 0 & 0.2 & 0.3 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

奖励函数 $\begin{matrix} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 \end{matrix}$

$$R = [-1, -2, -2, 10, 1, 0]$$

马尔可夫奖励过程的一个简单例子：$S = \{s_1, s_2, s_3, s_4, s_5, s_6\}$(注意：注意元素为确定量，非随机变量)

上图又称为马尔可夫链

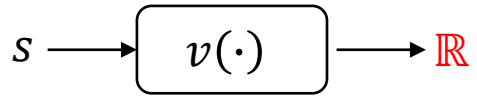
- 问题1：如何估算每个节点的期望回报（或者价值）？

$$G_t \triangleq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \gamma^3 R_{t+4} + \cdots$$(注意：$G_t, R_t$随机变量)

# 马尔可夫奖励过程（MRP）

- 在MRP中，一个状态的期望回报（即从这个状态出发的未来累积奖励的期望）被称为这个状态的**价值**（value）
- 所有状态的价值就组成了价值函数(value function), 写成

$$v(s) \triangleq \mathbb{E}[G_t|S_t = s]$$




- 可以将其展开为如下：

$$\begin{aligned} v(s) &\triangleq \mathbb{E}[G_t|S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma G_{t+1}|S_t = s] \\ &= \mathbb{E}[R_{t+1} + \gamma v(S_{t+1})|S_t = s] \\ &= \mathbb{E}[R_{t+1}|S_t = s] + \gamma \mathbb{E}[v(S_{t+1})|S_t = s] \end{aligned}$$

$v(s) = r(s) + \gamma \sum_{s'} p(s'|s)v(s')$

即时奖励期望正好是奖励函数的输出$r(s)$

可以根据状态$s$出发的转移概率得到

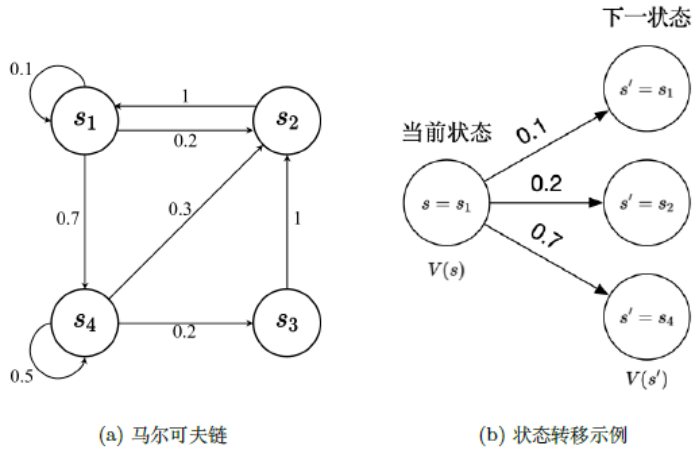
- 上述公式即为MRP的**贝尔曼递归方程（Bellman equation）**

# MRP贝尔曼方程（Bellman equation）矩阵形式及求解

$$\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$$

$$\boldsymbol{v} = [v(s_1), v(s_2), \ldots, v(s_n)]^T$$

$$\boldsymbol{R} = [r(s_1), r(s_2), \ldots, r(s_n)]^T$$




(a) 马尔可夫链　　(b) 状态转移示例

$$\begin{bmatrix} v(s_1) \\ v(s_2) \\ \ldots \\ v(s_n) \end{bmatrix} = \begin{bmatrix} r(s_1) \\ r(s_2) \\ \ldots \\ r(s_n) \end{bmatrix} + \gamma \begin{bmatrix} P(s_1|s_1) & P(s_2|s_1) & \ldots & P(s_n|s_1) \\ P(s_1|s_2) & P(s_2|s_2) & \ldots & P(s_n|s_2) \\ \ldots & \ldots & \ldots & \\ P(s_1|s_n) & P(s_2|s_n) & \ldots & P(s_n|s_n) \end{bmatrix} \begin{bmatrix} v(s_1) \\ v(s_2) \\ \ldots \\ v(s_n) \end{bmatrix}$$

$$\boldsymbol{v} = \boldsymbol{R} + \gamma P \boldsymbol{v}$$
$$(\boldsymbol{I} - \gamma P)\boldsymbol{v} = \boldsymbol{R}$$
$$\boldsymbol{v} = (\boldsymbol{I} - \gamma P)^{-1} \boldsymbol{R}$$

- 可以通过矩阵求逆把$\boldsymbol{v}$的价值直接求出来。但是一个问题是这个矩阵求逆的过程的复杂度是$O(n^3)$。
- 所以当状态非常多的时候，比如从10 个状态到1000 个状态，或者到100 万个状态，当我们有100 万个状态的时候，状态转移矩阵就会是一个100 万乘100 万的矩阵，对这样一个大矩阵求逆是非常困难的。
- 所以这种通过解析解去求解的方法只适用于很小量的马尔可夫奖励过程。

# 蒙特卡洛（MonteCarlo，MC）采样

- 1）从上述状态概率转移矩阵中采样大量轨迹，2）对每一条轨迹上的每个状态，基于回报公式；3）每个状态回报的平均值作为改状态的回报期望值的估计。
- 例如要计算$s_3$的估回报估值，1）可以从$s_3$ 开始基于概率$P$采样大量轨迹估计： $s_3 \rightarrow s_6$ ； $s_3 \rightarrow s_4 \rightarrow s_6; s_3 \rightarrow s_6; \ldots$ 2）采样完后，估算每条轨迹$s_3$的回报，3）最后取平均值作为$s_3$的回报期望估计。（注：需要提前确定$\gamma$数值）

---

1: $i \leftarrow 0, G_t \leftarrow 0$
2: 当 $i \neq N$ 时，执行：
3: 　生成一个回合的轨迹，从状态 $s$ 和时刻 $t$ 开始
4: 　使用生成的轨迹计算回报 $g = \sum_{i=t}^{H-1} \gamma^{i-t} r_i$
5: 　$G_t \leftarrow G_t + g, i \leftarrow i + 1$
6: 结束循环
7: $V_t(s) \leftarrow G_t / N$

---

图 2.6　计算马尔可夫奖励过程价值的蒙特卡洛方法

# 动态规划的方法

- 一直迭代贝尔曼方程，直到价值函数收敛，我们就可以得到某个状态的价值。

1: 对于所有状态 $s \in S, V'(s) \leftarrow 0, V(s) \leftarrow \infty$
2: 当$||V - V'|| > \epsilon$执行
3: $V \leftarrow V'$
4: 对于所有状态 $s \in S, V'(s) = R(s) + \gamma \sum_{s' \in S} P(s'|s)V(s')$
5: 结束循环
6: 返回$V'(s)$,对于所有状态$s \in S$

图 2.7 计算马尔可夫奖励过程价值的动态规划算法

# 总结:

- MP, MRP的异同点

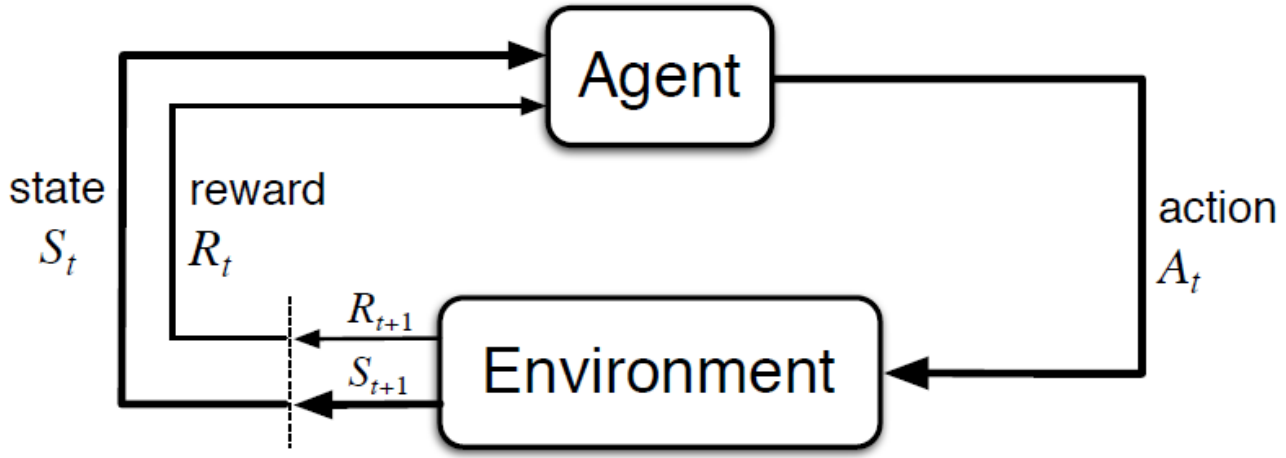
# 马尔可夫决策过程

马尔可夫过程
(Markov process, MP)
$< p, \mathcal{S} >$

马尔可夫奖励过程
(Markov reward process, MRP)
$< p, \mathcal{S}, \mathcal{R} >$

马尔可夫决策过程
(Markov decision process, MDP)
$< p, \mathcal{S}, \mathcal{R}, \mathcal{A} >$

# 智能体与MDP环境交互




- 智能体和MDP环境交互过程如下序列（或者轨迹（trajectory））

$$S_0, A_0, R_1, S_1, A_1, R_2, S_2, A_2, R_3, \ldots.$$

- 注意1： $A_t, R_t, S_t$均为离散随机变量
- 注意2： $R_t, S_t$具有明确显式定义离散概率分布，并取决于上一个状态和动作
- 马尔可夫性质：当给前一个 $t-1$ 时刻的状态和动作，则随机变量$S_t$在$t$时刻取某个特定值$s' \in \mathcal{S}, r \in \mathcal{R}$的概率为：

$$p(s', r \mid s, a) \triangleq \Pr\{S_t = s', R_t = r \mid S_{t-1} = s, A_{t-1} = a\},$$

$$\text{for all } s', s \in \mathcal{S}, \quad r \in \mathcal{R}, \quad a \in \mathcal{A}(s)$$

- $p$ 唯一并完全定义了MDP环境的动态性，是一个四参数的确定性函数

$$p: \mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$$

# 有限MDP

- 有限MDP (inite MDP）特点：

✓状态空间 $\mathcal{S}$包含有限个状态

✓动作空间$\mathcal{A}$包含有限个动作

✓奖励函数$\mathcal{R}$包含有限种奖励数值

# $p$的特点

- $p$ 唯一定义了MDP环境的所有动态特性，满足如下约束

$$\sum_{s' \in \mathcal{S}} \sum_{r \in \mathcal{R}} p(s', r | s, a) = 1, \text{ for all } s \in \mathcal{S},\ a \in \mathcal{A}(s)$$

- 给定$p$，可直接计算状态转移概率，即三参数函数$p: \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$

$$p(s' | s, a) \triangleq \Pr\{S_t = s' | S_{t-1} = s, A_{t-1} = a\} = \sum_{r \in \mathcal{R}} p(s', r | s, a)$$

- 直接计算状态-动作对的期望奖励$r: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$

$$r(s, a) \triangleq \mathbb{E}[R_t | S_{t-1} = s, A_{t-1} = a] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r | s, a)$$

# MDP环境特点

**□MDP环境是抽象且灵活的**

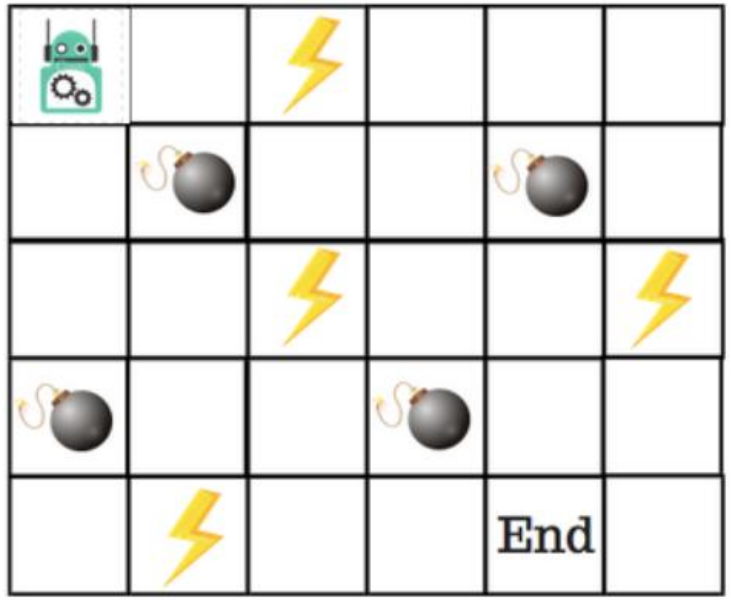
✓时间步(time-step)不一定指固定的实际时间间隔，例如，可以是决策制定和行动执行的任意连续阶段；

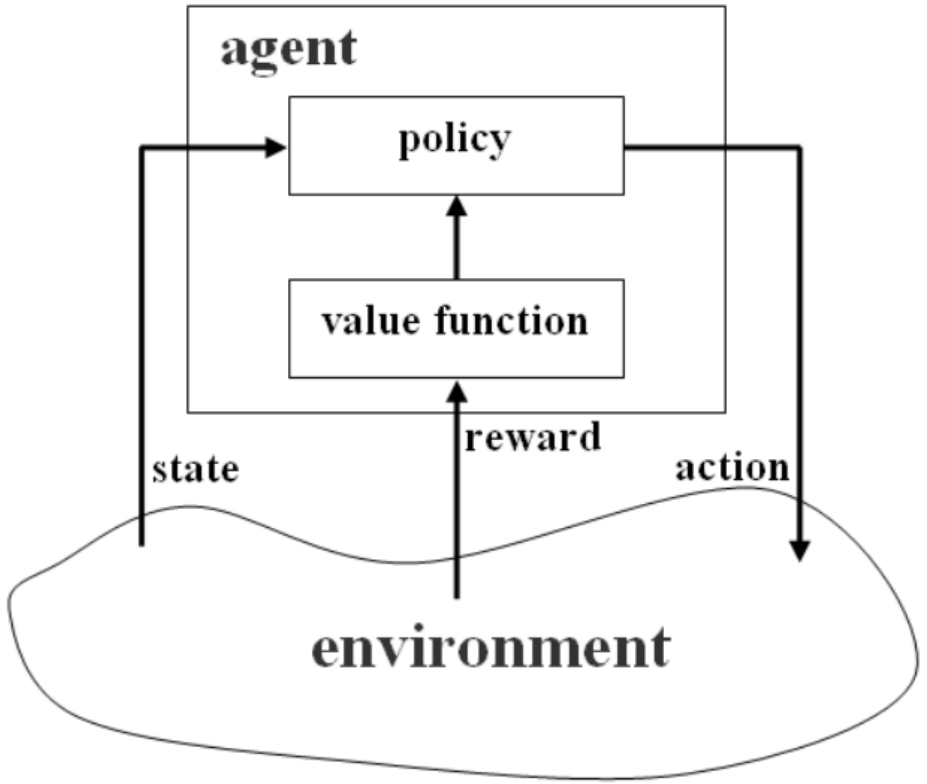
✓动作可以是低级控制（例如施加在机器人手臂电机上的电压）或高级决策选择（例如是否吃午饭）；

✓状态可以采用各种形式，例如低级感知，如直接传感器读数，图像和点云；

✓**给定一个强化学习任务，借助MDP对任务进行形式化描述是理解简化问题的第一步。**

# RL关键任务：价值函数估计



# 价值函数

- 状态价值函数（state value function）：评估智能体在环境中状态$s \in \mathcal{S}$有多好？
- 动作（或者状态-动作）价值函数：智能体在状态$s \in \mathcal{S}$下执行动作$a \in \mathcal{A}(s)$有多好？
- 其实是要评估回报期望

✓状态和动作的回报具有随机性

$$S_t, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, S_{t+2}, A_{t+2}, R_{t+3}, \ldots.$$

$$s, A_t, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, S_{t+2}, A_{t+2}, R_{t+3}, \ldots.$$

$$s, a, R_{t+1}, S_{t+1}, A_{t+1}, R_{t+2}, S_{t+2}, A_{t+2}, R_{t+3}, \ldots.$$

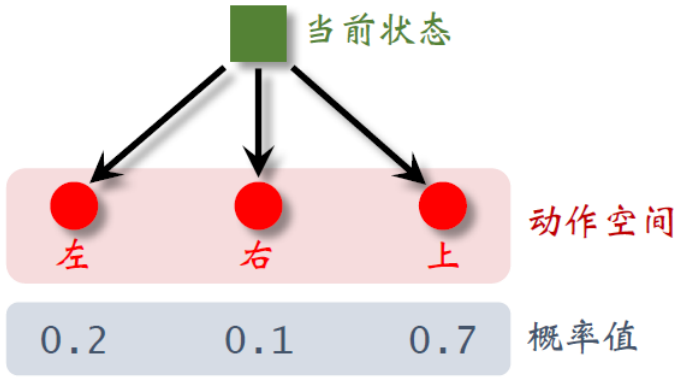


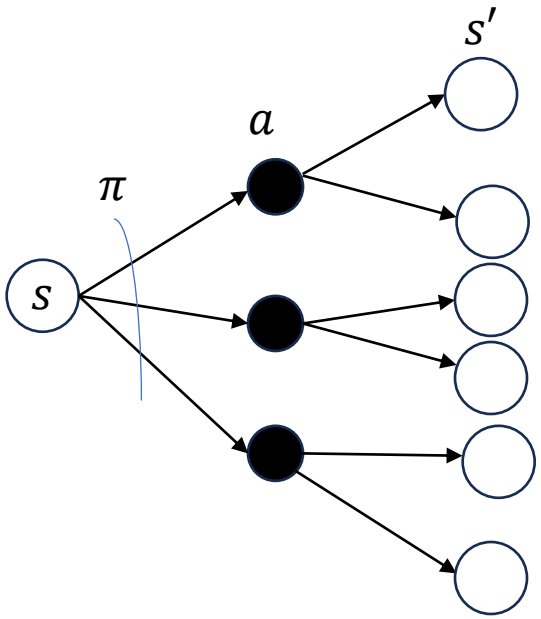

图 3.3: 状态空间是 $\mathcal{A} = \{$左, 中, 右$\}$。把当前状态 $s$ 输入策略函数，策略函数输出三个概率值: 0.2, 0.1, 0.7。所以，对于确定的状态 $s$，智能体执行的动作是不确定的，三个动作都可能被执行。

- 回报和未来有关，未来和执行策略有关，具有随机性
- 价值不具有随机性，是一个回报的期望值，且和策略$\pi(a|s)$绑定，有唯一值

# 例题

- MDP环境，假设当前状态为$S_t = s$，动作选择遵循随机策略$\pi$，则根据$\pi$ 和四参数$p(s', r| s, a)$，计算奖励$R_{t+1}$的期望。

$$\mathbb{E}_{\pi}\left[R_{t+1}|S_t = s\right] = \sum_{a} \pi(a|S_t = s) \sum_{s'} \sum_{r} r\, p(s', r| s, a)$$




- 备份图（Background）：
  - 空心圈：状态
  - 实心圈：动作
  - 单向图

# 策略$\pi$下某个状态的回报期望（即状态价值）

- 记录智能体在策略$\pi$下一个状态$s$的价值，记$v_\pi(s)$, 则定义如下:

$$v_\pi(s) \triangleq \mathbb{E}_\pi\left[G_t | S_t = s\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s\right], \text{ for all } s \in \boldsymbol{S}$$

- $\mathbb{E}_\pi[\cdot]$代表智能体遵循某个策略$\pi$期望, $t$ 表示时间步$t$
- $v_\pi(\cdot)$表示智能体在策略$\pi$下状态价值函数

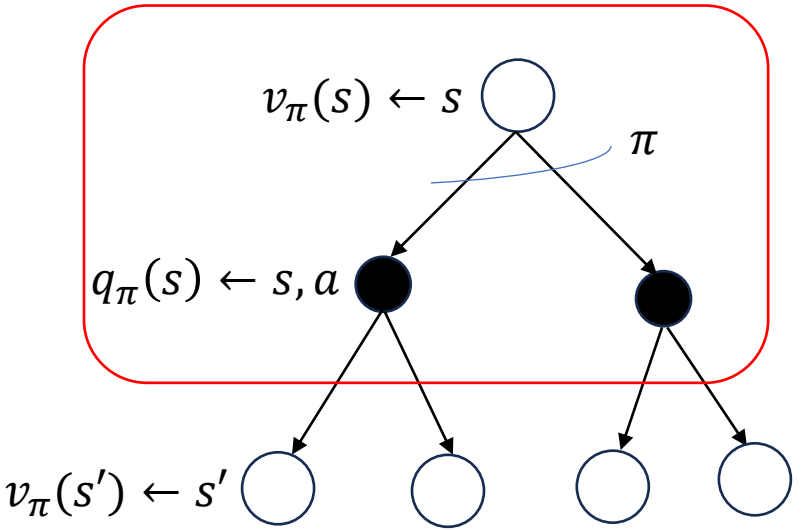
## 状态$s$下执行动作$a \in \mathcal{A}(s)$，而后遵循策略$\pi$的期望回报（动作价值函数）

- 记录智能体遵循策略$\pi$，位于状态$s$执行动作$a$，在之后遵循策略$\pi$的期望回报，记$q_\pi(s,a)$：

$$q_\pi(s,a) \triangleq \mathbb{E}_\pi\left[G_t | S_t = s, A_t = a\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s, A_t = a\right]$$

- $q_\pi(\cdot)$表示智能体在策略$\pi$下**动作（或状态-动作）价值函数**

# $v_\pi$关于$q_\pi$和$\pi$的关系




$$v_\pi(s) = \sum_a \pi(a|s)\, q_\pi(s,a)$$

- 空心圈：状态
- 实心圈：动作

价值信息从一个状态（或状态-动作对）的后继状态（或状态-动作对）转移回它

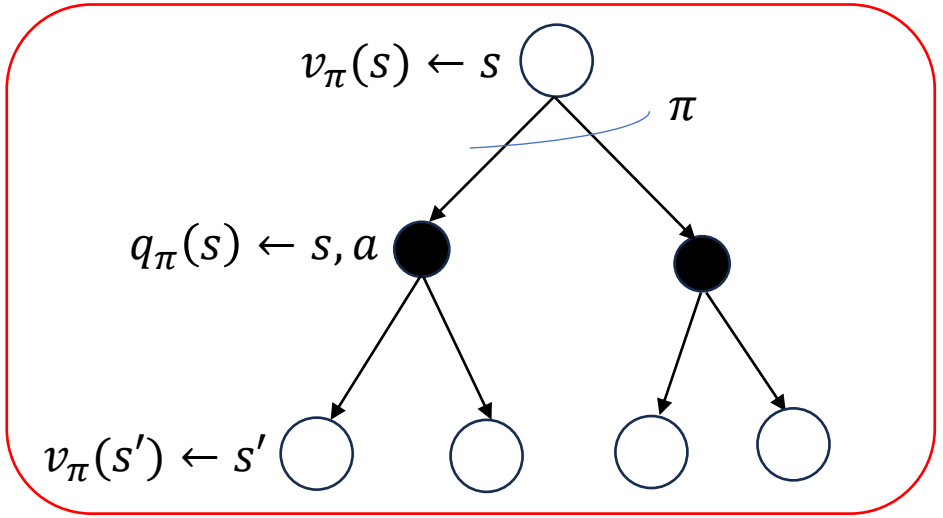
# $v_\pi$的 贝尔曼方程（Bellman equation）

- 对于任意一个策略$\pi$，和任意一个状态$s$，则状态$s$和其后继状态的价值满足如下递归式：

$$
\begin{aligned}
v_\pi(s) &\triangleq \mathbb{E}_\pi\left[G_t | S_t = s\right] \\
&= \mathbb{E}_\pi\left[R_{t+1} + \gamma G_{t+1} | S_t = s\right] \\
&= \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|\, s, a)[r + \gamma v_\pi(s')] \text{ for all } s \in \mathcal{S}
\end{aligned}
$$

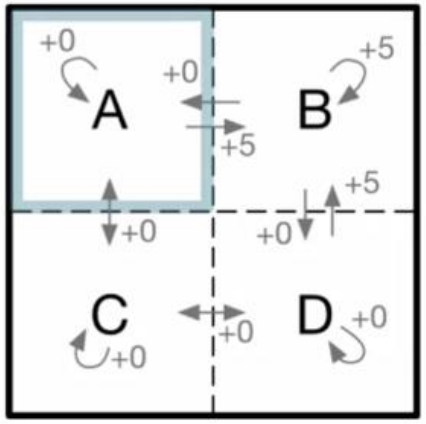
- 这里隐含了动作 $a \in \mathcal{A}(s)$，下一个状态 $s \in \mathcal{S}$，奖励 $r \in \mathcal{R}$

$v_\pi(s)$的备份图：



# 例题：一个简单有限马尔可夫决策过程

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)[r + \gamma v_\pi(s')] \text{ for all } s \in \mathcal{S}$$

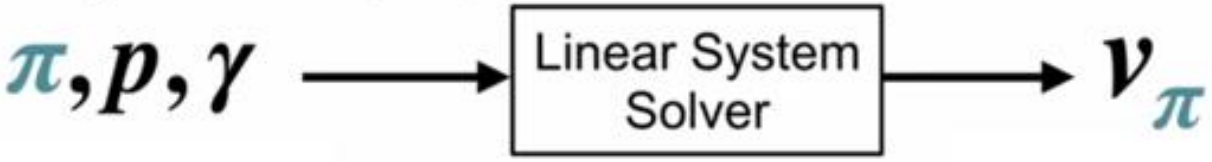




- 状态空间：$\mathcal{S} = \{s_A, s_B, s_C, s_D\}$
- $r$=0.7，奖励如图所示
- 问题：求解状态价值函数（即计算每个状态对应的价值）
- 解法：解决线性方程组即可(n个未知量，n个线性方程组)

$$\begin{cases} v_\pi(s_A) = 0.25[5 + 0.7v_\pi(s_B)] + 0.25[0 + 0.7v_\pi(s_C)] + 0.5[0 + 0.7v_\pi(s_A)] \\ v_\pi(s_B) = 0.5[5 + 0.7v_\pi(s_B)] + 0.25[0 + 0.7v_\pi(s_A)] + 0.25[0 + 0.7v_\pi(s_D)] \\ v_\pi(s_C) = 0.25[0 + 0.7v_\pi(s_A)] + 0.25[0 + 0.7v_\pi(s_D)] + 0.5[0 + 0.7v_\pi(s_C)] \end{cases}$$

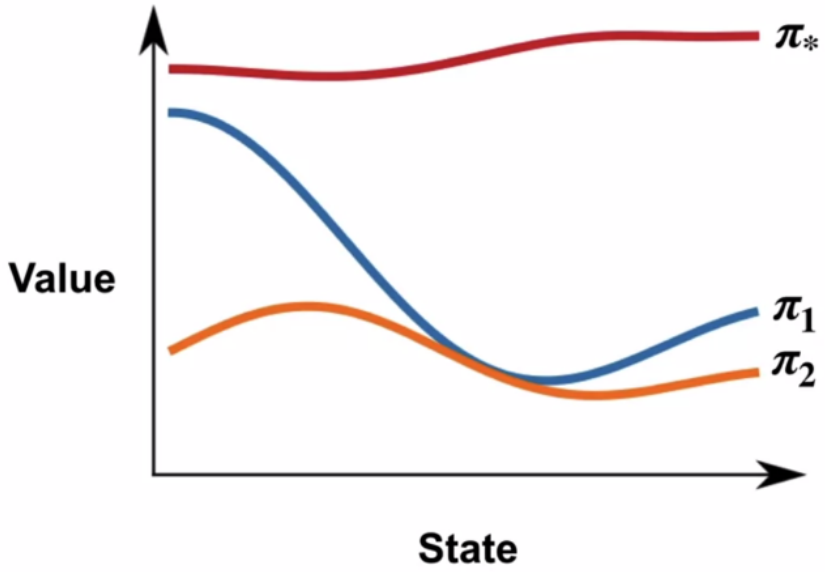
$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a)\left[r + \gamma v_\pi(s')\right]$$




$$\begin{cases} v_\pi(s_C) = 2.2 \\ v_\pi(s_D) = 4.2 \end{cases}$$

# 最优策略和最优价值函数

- 策略$\pi$比策略$\pi'$不差，则对于所有
$$v_{\pi}$$
- 最优策略（$\pi_*$）：至少存在一种
- 最优策略共享同一个最优状态价
$$v_*(s)$$




$$\pi_1 \geq \pi_2$$

- 若存在多个最优策略，则他们共享同一个最优动作价值函数，记为$q_*(s)$：

$$q_*(s,a) \triangleq \max_{\pi} q_{\pi}(s,a), \text{ for all } s \in \mathcal{S}, a \in \mathcal{A}(s)$$

- $q_*(s,a)$和$v_*(s)$的关系：

$$q_*(s,a) = \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) | S_t = s, A_t = a]$$

# 寻找最优策略

- 对有限马尔可夫模型，确定性策略的个数是：$|\mathcal{A}|^{|\mathcal{S}|}$
- **穷举法：**我们通过价值函数评估每一个策略，则需要评估$|\mathcal{A}|^{|\mathcal{S}|}$，才能找到搜索到最优策略。
- 穷举法效果不高，我们后面将通过策略迭代或价值代码两种方法求解（第四章内容）。

# 贝尔曼最优方程（Bellman optimality equation）

在最优策略下，**一个状态的价值**必须等于在**该状态下执行最优动作的价值**。

$v_*$ 的贝尔曼最优方程

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a] \\
&= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\
&= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \\
&= \max_a \sum p(s', r \mid s, a)\big[r + \gamma v_*(s')\big].
\end{aligned}
$$

$q_*$ 的贝尔曼最优方程

$$
\begin{aligned}
q_*(s,a) &= \mathbb{E}\Big[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \;\Big|\; S_t = s, A_t = a\Big] \\
&= \sum_{s',r} p(s', r \mid s, a)\Big[r + \gamma \max_{a'} q_*(s', a')\Big].
\end{aligned}
$$

# 贝尔曼最优方程的备份图

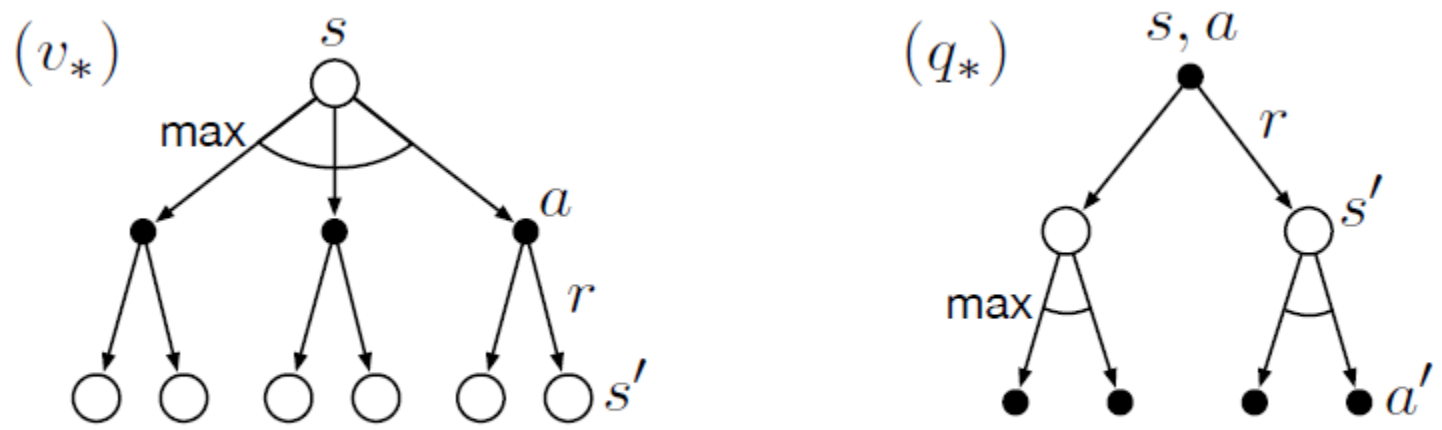



**Figure 3.4:** Backup diagrams for $v_*$ and $q_*$

# 基于最优价值函数的最优策略

$$\pi_*(a|s) = \underset{a \in \mathcal{A}(s)}{\operatorname{argmax}}\, q_*(s,a), \text{ for all } s \in \mathcal{S}$$

$$\begin{bmatrix} q_*(s_1,a_1) & \cdots & q_*(s_1,a_1) \\ \vdots & \ddots & \vdots \\ q_*(s_{|\mathcal{S}|},a_1) & \cdots & q_*(s_{|\mathcal{S}|},a_{|\mathcal{A}|}) \end{bmatrix}$$

最优动作价值函数$q_*$

最优策略提取

按行argmax

$$\pi_*(a|s) = \begin{bmatrix} \pi_*(a|s_1) \\ \pi_*(a|s_2) \\ \dots \\ \pi_*(a|s_{|\mathcal{S}|}) \end{bmatrix} = \begin{bmatrix} a_{j_1} \\ a_{j_2} \\ \dots \\ a_{j_{|\mathcal{S}|}} \end{bmatrix}$$

$$\pi_*(a|s) = \underset{a \in \mathcal{A}(s)}{\operatorname{argmax}} \sum_{s',r} p(s',r|s,a)\,[r + \gamma v_*(s')], \text{ for all } s \in \mathcal{S},$$

# 总结

- MP
- MRP
- MDP
- 奖励期望和回报期望有什么区别
- 状态价值、动作价值
- 贝尔曼方程
- 最优策略