状态转移概率：$Pr(S_{t+1}|S_t)$

奖励函数/回报：$R(S_t, S_{t+1})$

衰退系数：$\gamma \in [0,1]$

马尔可夫奖励过程：

$MRP = \{S, Pr, R, \gamma\}$

第t步采取的动作：$a_t \in A$

状态转移概率：$Pr(S_{t+1}|S_t, a_t)$

奖励函数：$R(S_t, a_t, S_{t+1})$

马尔可夫决策过程：

$MDP = \{S, A, Pr, R, \gamma\}$

$$
\begin{aligned}
V_\pi(s) &= \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s] \\
&= \mathbb{E}_{a\sim\pi(s,\cdot)}[\mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s, A_t = a]] \\
&= \textstyle\sum_{a\in A} \pi(s,a) \underline{q_\pi(s,a)}
\end{aligned}
$$

$$
\begin{aligned}
q_\pi(s,a) &= \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s, A_t = a] \\
&= \mathbb{E}_{s'\sim P(\cdot|s,a)}[R(s,a,s') + \gamma \mathbb{E}_\pi[R_{t+2} + \gamma R_{t+3} + \cdots | S_{t+1} = s']] \\
&= \textstyle\sum_{s'\in S} P(s'|s,a)\,[R(s,a,s') + \gamma \underline{V_\pi(s')}]
\end{aligned}
$$

累计回报：$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

价值函数: $V_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$

动作–价值函数: $q_\pi(s,a) = \mathbb{E}_\pi[G_t | S_t = s, A_t = a]$

1.三者本质:都是奖励/回报。
2.均值内涵:概率乘以奖励-加权平均。

价值函数的贝尔曼方程

$$
\begin{aligned}
V_\pi(s) &= \sum_{a\in A} \pi(s,a) \sum_{s'\in S} P(s'|s,a)\,[R(s,a,s') + \gamma V_\pi(s')] \\
&= \mathbb{E}_{a\sim\pi(s,\cdot)} \mathbb{E}_{s'\sim P(\cdot|s,a)}[R(s,a,s') + \gamma \underline{V_\pi(s')}]
\end{aligned}
$$

动作–价值函数的贝尔曼方程

$$
\begin{aligned}
q_\pi(s,a) &= \sum_{s'\in S} P(s'|s,a) \left[R(s,a,s') + \gamma \sum_{a'\in A} \pi(s',a') q_\pi(s',a') \right] \\
&= \mathbb{E}_{s'\sim P(\cdot|s,a)}[R(s,a,s') + \gamma \mathbb{E}_{a'\sim\pi(s',\cdot)}[\underline{q_\pi(s',a')}]]
\end{aligned}
$$

价值函数的贝尔曼方程

$$V_\pi(s) = \sum_{a\in A}\pi(s,a)\sum_{s'\in S}P(s'|s,a)\left[R(s,a,s') + \gamma V_\pi(s')\right]$$

$$= \underline{\mathbb{E}_{a\sim\pi(s,\cdot)}\mathbb{E}_{s'\sim P(\cdot|s,a)}[R(s,a,s')} + \gamma V_\pi(s')]$$

$$R_s = \mathbb{E}_\pi[R_{t+1}|S_t = s] = \sum_{s'\in S} p(s'|s)R(s'|s)$$

$$R_s^a = \mathbb{E}_\pi[R_{t+1}|S_t = s, A_t = a] = \sum_{s'\in S} p(s'|s,a)R(s,a,s')$$

$$R_s = \mathbb{E}_\pi[R_{t+1}|S_t = s] = \sum_{a\in A}\pi(s,a)R_s^a$$

$$\boldsymbol{V(s) = R_s + \gamma\sum_{s'\in S} p(s'|s)V(s')}$$

动作–价值函数的贝尔曼方程

$$q_\pi(s,a) = \sum_{s'\in S} P(s'|s,a)\left[R(s,a,s') + \gamma \sum_{a'\in A} \pi(s',a') q_\pi(s',a')\right]$$
$$= \underline{\mathbb{E}_{s'\sim P(\cdot|s,a)}[R(s,a,s')} + \gamma \underline{\mathbb{E}_{a'\sim\pi(s',\cdot)}[q_\pi(s',a')]]}$$

$$R_s^a = \mathbb{E}_\pi[R_{t+1}|S_t = s, A_t = a] = \sum_{s'\in S} p(s'|s,a) R(s'|s,a)$$

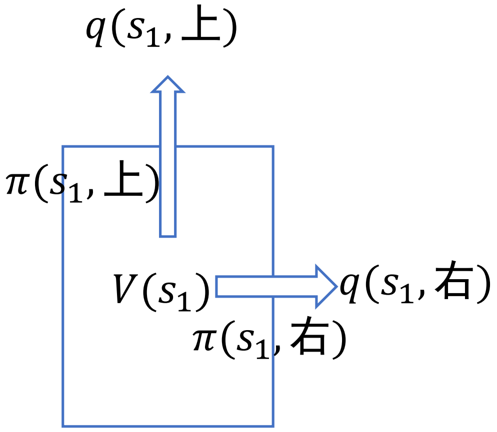
$$\gamma \sum_{s'\in S} P(s'|s,a)\, V_\pi(s')$$

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s'\in S} p(s'|s,a)\, V(s')$$

价值函数与动作-价值函数的关系：以状态s_1的计算为例

$$V_\pi(s) = \sum_{a \in A} \pi(s,a) q_\pi(s,a)$$

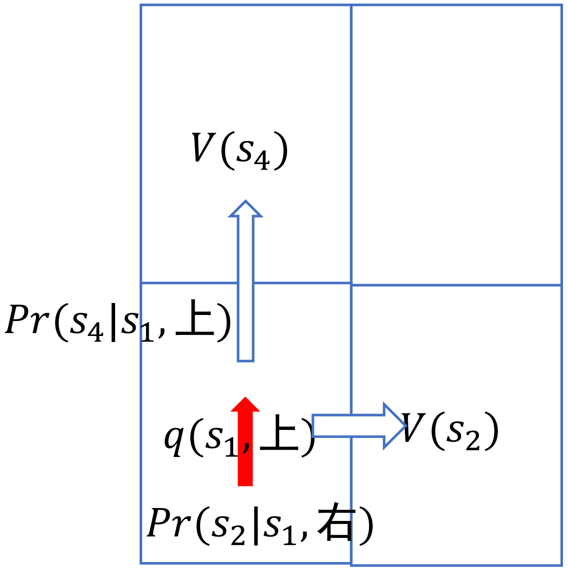
$$V_\pi(s_1) = \pi(s_1,\text{上}) q_\pi(s_1,\text{上}) + \pi(s_1,\text{右}) q_\pi(s_1,\text{右})$$



不同动作下的反馈累加

$$q_\pi(s,a) = \sum_{s' \in S} Pr(s'|s,a)\,[R(s,a,s') + \gamma V_\pi(s')]$$

$$q_\pi(s_1,\text{上}) = Pr(s_4|s_1,\text{上})[R(s_1,\text{上},s_4) + \gamma V_\pi(s_4)]$$



动作确定时状态转移后的反馈结果

Bellman方程的推导

求解V值相当于解方程

The Bellman equation can be expressed concisely using matrices,

$$v = \mathcal{R} + \gamma \mathcal{P} v$$

where v is a column vector with one entry per state

$$\begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}_1 \\ \vdots \\ \mathcal{R}_n \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \dots & \mathcal{P}_{1n} \\ \vdots & & \\ \mathcal{P}_{11} & \dots & \mathcal{P}_{nn} \end{bmatrix} \begin{bmatrix} v(1) \\ \vdots \\ v(n) \end{bmatrix}$$

- 用解线性方程组的方法，只适合小规模的问题；矩阵规模大后,求逆复杂.
- 大规模问题用迭代方法，包括：动态规划、时序差分学习、蒙特卡洛估计等。

- 公式向量形式求解

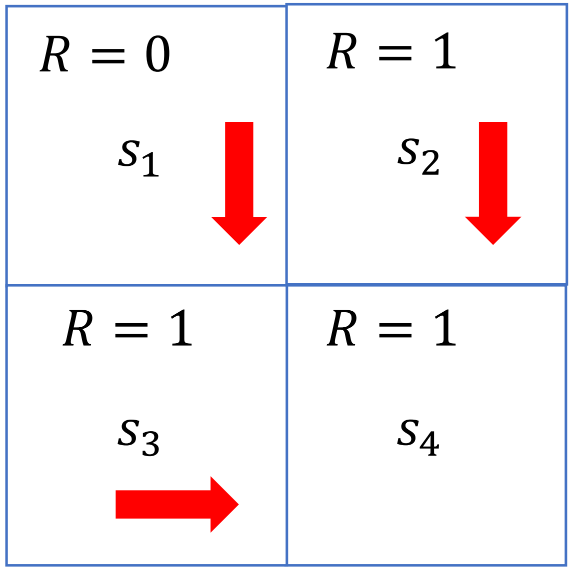
$$V(s) = R_s + \gamma \sum_{s' \in S} p(s'|s) V(s')$$

$$V_\pi(s_i) = R_\pi(s_i) + \gamma \sum_{s_j \in S} p_\pi(s_j | s_i) V_\pi(s_j)$$

$$V_\pi(s_i) = [V_\pi(s_1), V_\pi(s_2), \ldots, V_\pi(s_n)]^T$$

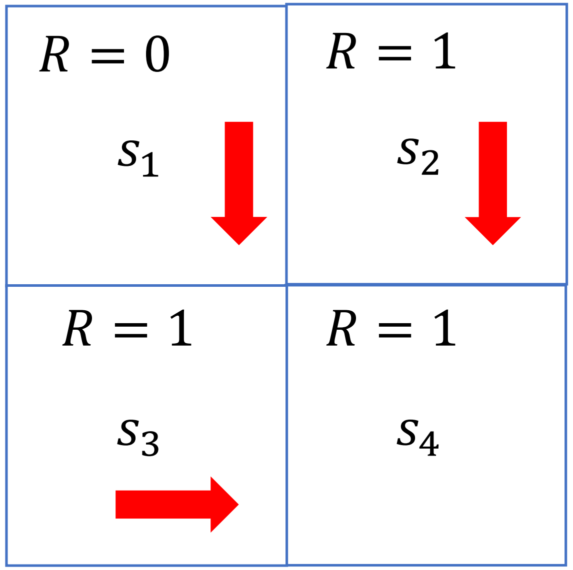
$$R_\pi(s_i) = [R_\pi(s_1), R_\pi(s_2), \ldots, R_\pi(s_n)]^T$$

$$P_\pi(s_j | s_i) = [P_\pi]_{ij}$$



$$\begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix} = \begin{bmatrix} R_\pi(s_1) \\ R_\pi(s_2) \\ R_\pi(s_3) \\ R_\pi(s_4) \end{bmatrix} + \gamma \begin{bmatrix} P_\pi(s_1|s_1) & P_\pi(s_2|s_1) & P_\pi(s_3|s_1) & P_\pi(s_4|s_1) \\ P_\pi(s_1|s_2) & P_\pi(s_2|s_2) & P_\pi(s_3|s_2) & P_\pi(s_4|s_2) \\ P_\pi(s_1|s_3) & P_\pi(s_2|s_3) & P_\pi(s_3|s_3) & P_\pi(s_4|s_3) \\ P_\pi(s_1|s_4) & P_\pi(s_2|s_4) & P_\pi(s_3|s_4) & P_\pi(s_4|s_4) \end{bmatrix} \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix}$$

- 公式向量形式求解



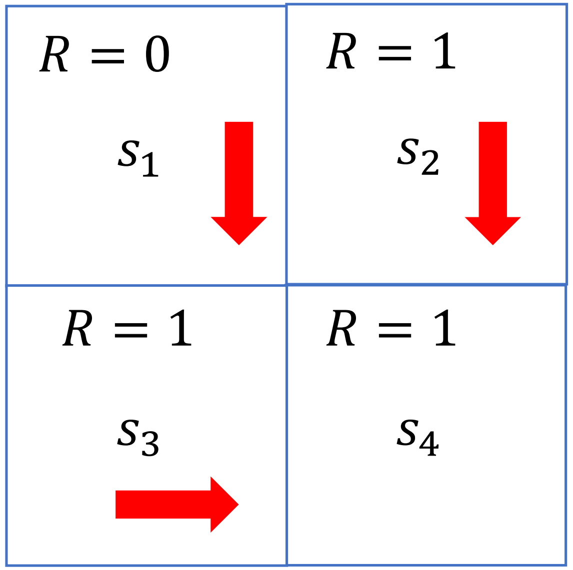
$$\begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix} = \begin{bmatrix} R_\pi(s_1) \\ R_\pi(s_2) \\ R_\pi(s_3) \\ R_\pi(s_4) \end{bmatrix} + \gamma \begin{bmatrix} P_\pi(s_1|s_1) & P_\pi(s_2|s_1) & P_\pi(s_3|s_1) & P_\pi(s_4|s_1) \\ P_\pi(s_1|s_2) & P_\pi(s_2|s_2) & P_\pi(s_3|s_2) & P_\pi(s_4|s_2) \\ P_\pi(s_1|s_3) & P_\pi(s_2|s_3) & P_\pi(s_3|s_3) & P_\pi(s_4|s_3) \\ P_\pi(s_1|s_4) & P_\pi(s_2|s_4) & P_\pi(s_3|s_4) & P_\pi(s_4|s_4) \end{bmatrix} \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix}$$

写出矩阵，求解?

$$\begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix}$$

方程组求解
矩阵法求解

- 公式向量形式求解



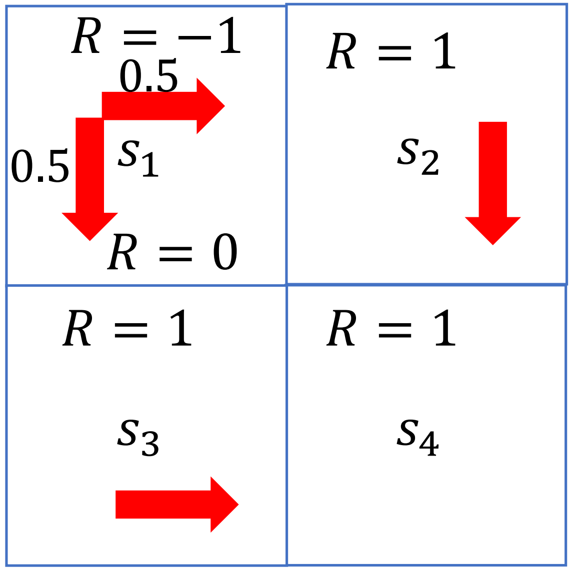
$$\begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix}$$

方程组求解?

$$V_\pi(s_1) = R(s_1, \text{下}, s_3) + \gamma V_\pi(s_3) = 0 + 0.99 \times V_\pi(s_3)$$
$$V_\pi(s_2) = R(s_2, \text{下}, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times V_\pi(s_4)$$
$$V_\pi(s_3) = R(s_3, \text{右}, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times V_\pi(s_4)$$
$$V_\pi(s_4) = R(s_4, *, s_4) + \gamma V_\pi(s_4) = 1 + 0.99 \times V_\pi(s_4)$$

注意：这里R其实是R_{t+1}，即S_t转为S_{t+1}的回报。

- 公式向量形式求解



$$\begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix} = \begin{bmatrix} R_\pi(s_1) \\ R_\pi(s_2) \\ R_\pi(s_3) \\ R_\pi(s_4) \end{bmatrix} + \gamma \begin{bmatrix} P_\pi(s_1|s_1) & P_\pi(s_2|s_1) & P_\pi(s_3|s_1) & P_\pi(s_4|s_1) \\ P_\pi(s_1|s_2) & P_\pi(s_2|s_2) & P_\pi(s_3|s_2) & P_\pi(s_4|s_2) \\ P_\pi(s_1|s_3) & P_\pi(s_2|s_3) & P_\pi(s_3|s_3) & P_\pi(s_4|s_3) \\ P_\pi(s_1|s_4) & P_\pi(s_2|s_4) & P_\pi(s_3|s_4) & P_\pi(s_4|s_4) \end{bmatrix} \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix}$$

求解?

$$R_s = \sum_{s' \in S} p(s'|s) R(s'|s)$$

$$R_\pi(s_i) = \sum_{s_j \in S} p_\pi(s_j|s_i) R_\pi(s_j|s_i)$$

$$\begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix} = \begin{bmatrix} 0.5\times 0 + 0.5\times(-1) \\ 1 \\ 1 \\ 1 \end{bmatrix} + \gamma \begin{bmatrix} 0 & 0.5 & 0.5 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} V_\pi(s_1) \\ V_\pi(s_2) \\ V_\pi(s_3) \\ V_\pi(s_4) \end{bmatrix}$$