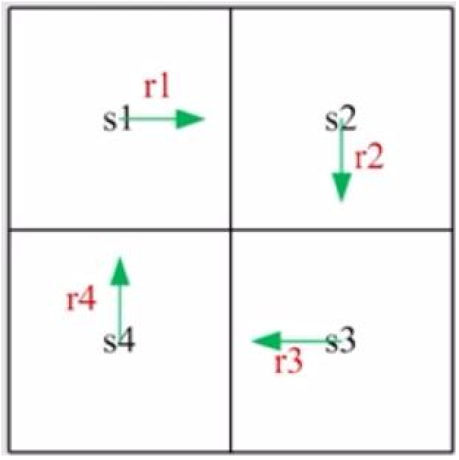
1. return 的计算 & 以不同 state 为起点的 trajectory 之间 return 的关系



以上图为例，用v_i表示以s_i为起点的 trajectory 的 return，得到结果为：

$$
\begin{aligned}
v_1 &= r_1 + \gamma r_2 + \gamma^2 r_3 + \dots \\
v_2 &= r_2 + \gamma r_3 + \gamma^2 r_4 + \dots \\
v_3 &= r_3 + \gamma r_4 + \gamma^2 r_1 + \dots \\
v_4 &= r_4 + \gamma r_1 + \gamma^2 r_2 + \dots
\end{aligned}
$$

由以上结果可知，$v_1 = r_1 + \gamma(r_2 + \gamma r_3 + \gamma^2 r_4 + \dots) = r_1 + \gamma v_2$，其他依此类推，

可以得到：

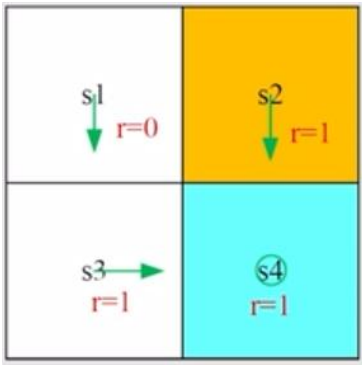
$$
\begin{aligned}
v_1 &= r_1 + \gamma(r_2 + \gamma r_3 + \dots) = r_1 + \gamma v_2 \\
v_2 &= r_2 + \gamma(r_3 + \gamma r_4 + \dots) = r_2 + \gamma v_3 \\
v_3 &= r_3 + \gamma(r_4 + \gamma r_1 + \dots) = r_3 + \gamma v_4 \\
v_4 &= r_4 + \gamma(r_1 + \gamma r_2 + \dots) = r_4 + \gamma v_1
\end{aligned}
$$

写成矩阵-向量形式为：

$$
\underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{\mathbf{v}} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} + \begin{bmatrix} \gamma v_2 \\ \gamma v_3 \\ \gamma v_4 \\ \gamma v_1 \end{bmatrix} = \underbrace{\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}}_{\mathbf{r}} + \gamma \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}}_{\mathbf{P}} \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{\mathbf{v}}
$$

即$v = r + \gamma P v$，推出$v = (I - \gamma P)^{-1} r$

例子:



$$
\begin{aligned}
v_1 &= 0 + \gamma v_3 \\
v_2 &= 1 + \gamma v_4 \\
v_3 &= 1 + \gamma v_4 \\
v_4 &= 1 + \gamma v_4
\end{aligned}
$$

2. state value

考虑以下情形：有一条 trajectory 为

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

其中S_t表示 t 时刻的 state, A_t表示在该 state 采取的 action, R_{t+1}表示在进行 action 后得到的 reward，则该 trajectory 的 discounted return 为:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

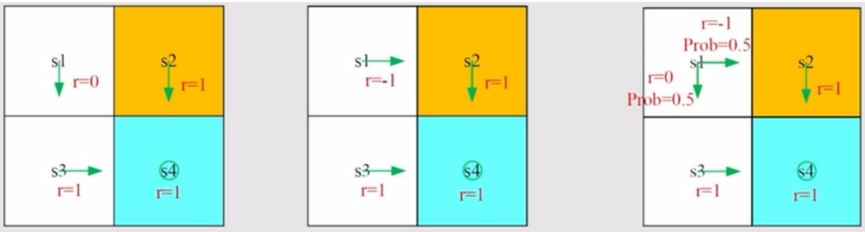
则 state value 定义为G_t的期望值，表示为:

$$v_\pi(s) = E(G_t|S_t = s)$$

表示从状态 s 出发，所有可能的 trajectory 得到的 return 的期望。state value 是状态 s 的函数，且依赖于 policy π，不同的 policy 可能会得到不同的 state value。它代表的是状态 s 的价值

state value 和 return 的区别：return 是一个 trajectory 的结果，而 state value 是多个 trajectory 的结果, 如果从一个 state 出发, 只有一个 trajectory, 那么此时 return 等于 state value；如果从一个 state 出发，有多个 trajectory，那么此时 state value 是这多个 trajectory 得到的 return 的期望。

例如：



Recall the returns obtained from s_1 for the three examples:

$$v_{\pi_1}(s_1) = 0 + \gamma 1 + \gamma^2 1 + \cdots = \gamma(1 + \gamma + \gamma^2 + \dots) = \frac{\gamma}{1-\gamma}$$

$$v_{\pi_2}(s_1) = -1 + \gamma 1 + \gamma^2 1 + \cdots = -1 + \gamma(1 + \gamma + \gamma^2 + \dots) = -1 + \frac{\gamma}{1-\gamma}$$

$$v_{\pi_3}(s_1) = 0.5\left(-1 + \frac{\gamma}{1-\gamma}\right) + 0.5\left(\frac{\gamma}{1-\gamma}\right) = -0.5 + \frac{\gamma}{1-\gamma}$$

在例 1 和例 2 中，从s_1出发均只有一个 trajectory，故此时 state value 等于 return；

在例 3 中，从s_1出发有两个 trajectory，故此时 state value 为这两个 return 的期望

3．Bellman equation：贝尔曼公式的推导

考虑这样一个 trajectory

$$S_t \xrightarrow{A_t} R_{t+1}, S_{t+1} \xrightarrow{A_{t+1}} R_{t+2}, S_{t+2} \xrightarrow{A_{t+2}} R_{t+3}, \dots$$

则该 trajectory 的 discounted return 为

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \ldots) = R_{t+1} + \gamma G_{t+1}$$

则状态S_t的 state value 为

$$\begin{aligned} v_\pi(s) &= E(G_t|S_t = s) = E(R_{t+1} + \gamma G_{t+1}|S_t = s) \\ &= E(R_{t+1}|S_t = s) + \gamma E(G_{t+1}|S_t = s) \end{aligned}$$

第一项表示在状态 s 采取各种 action，得到的所有可能 reward 的期望，则第一项的计算为

$$E(R_{t+1}|S_t = s) = \sum_a \pi(a|s) E(R_{t+1}|S_t = s, A_t = a) = \sum_a \pi(a|s) \sum_r p(r|s,a)\, r$$

其中$\pi(a|s)$表示在状态 s 下采取动作 a 的概率，$E(R_{t+1}|S_t = s, A_t = a)$表示在状态 s 下采取动作 a 得到的 reward 的期望，求得在各种动作下所取得的 reward 的期

望。其中$p(r|s,a)$表示在状态 s 下采取动作 a 得到的 reward 的概率，r表示该 reward，求得在动作 a 下取得的 reward 的期望。

第二项表示在状态 s 出发得到的下一个时刻的 return 的期望，从状态 s 可能到达多个其他状态S'，则第二项的计算为

$$E(G_{t+1}|S_t=s)=\sum_{s'}E(G_{t+1}|S_t=s,S_{t+1}=s')p(s'|s)$$

根据马尔可夫性质，即与历史无关性质，从s到达s'后，再以s'为起点得到的 return 与直接以s'为起点得到的 return 相等，故原式可表示为

$$E(G_{t+1}|S_t=s)=\sum_{s'}E(G_{t+1}|S_{t+1}=s')p(s'|s)$$

则$E(G_{t+1}|S_{t+1}=s')$即为状态s'的 state value，记为$v_\pi(s')$，其中$p(s'|s)$表示从状态s到状态s'的概率，其计算为$p(s'|s)=p(s'|s,a)\pi(a|s)$，其中$\pi(a|s)$表示在状态 s 下采取动作 a 的概率，$p(s'|s,a)$表示在状态 s 下采取动作 a 到达$s'$的概率，故原式可表示为

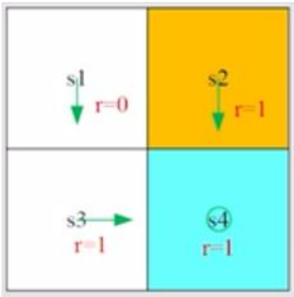
$$E(G_{t+1}|S_t=s)=\sum_{s'}v_\pi(s')\sum_a p(s'|s,a)\pi(a|s)$$

所以可以得到状态 s 的 state value 的计算为

$$\begin{aligned}
v_\pi(s) &= \mathbb{E}[R_{t+1}|S_t=s]+\gamma\mathbb{E}[G_{t+1}|S_t=s],\\
&= \underbrace{\sum_a \pi(a|s)\sum_r p(r|s,a)r}_{\text{mean of immediate rewards}} + \underbrace{\gamma\sum_a \pi(a|s)\sum_{s'} p(s'|s,a)v_\pi(s')}_{\text{mean of future rewards}},\\
&= \sum_a \pi(a|s)\left[\sum_r p(r|s,a)r+\gamma\sum_{s'} p(s'|s,a)v_\pi(s')\right], \quad \forall s\in\mathcal{S}.
\end{aligned}$$

这就是贝尔曼公式，它表示了不同 state 之间的 state value 的关系，它对状态空间中所有状态均成立

通过联立状态空间中所有求该 state 的 state value，进行方程组求解，代入已知条件，即可得到所有 state 的 state value，例子 1：



- $\pi(a = a_3|s_1) = 1$ and $\pi(a \neq a_3|s_1) = 0$.
- $p(s' = s_3|s_1, a_3) = 1$ and $p(s' \neq s_3|s_1, a_3) = 0$.
- $p(r = 0|s_1, a_3) = 1$ and $p(r \neq 0|s_1, a_3) = 0$.

则针对s_1的贝尔曼公式为：

$$\begin{aligned}
v_\pi(s_1) = \pi(a = a_3|s_1)\Bigg[&p(r = 0|s_1, a_3) * 0 + \sum_r p(r \neq 0|s_1, a_3) * r \\
&+ \gamma\left[p(s' = s_3|s_1, a_3) * v_\pi(s_3) + \sum_{s'} p(s' \neq s_3|s_1, a_3) * v_\pi(s')\right]\Bigg] \\
&+ \pi(a \neq a_3|s_1) * \dots \\
= 1 * [1 * 0 + 0 + \gamma * [1 * v_\pi(s_3) + 0]] + 0 = 0 + \gamma v_\pi(s_3)
\end{aligned}$$

从图中也可直观看出：$v_\pi(s_1) = 0 + \gamma v_\pi(s_3)$

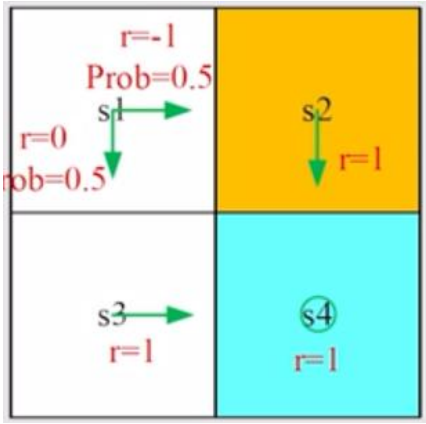
以此类推可以得出：

$$\begin{aligned}
v_\pi(s_2) &= 1 + \gamma v_\pi(s_4) \\
v_\pi(s_3) &= 1 + \gamma v_\pi(s_4) \\
v_\pi(s_4) &= 1 + \gamma v_\pi(s_4)
\end{aligned}$$

联立方程组求解得到：

$$\begin{aligned}
v_\pi(s_1) &= \frac{\gamma}{1 - \gamma} \\
v_\pi(s_2) &= \frac{1}{1 - \gamma} \\
v_\pi(s_3) &= \frac{1}{1 - \gamma} \\
v_\pi(s_4) &= \frac{1}{1 - \gamma}
\end{aligned}$$

例子 2:



从图中可以直观看出:

$$v_\pi(s_1) = 0.5 * [0 + \gamma v_\pi(s_3)] + 0.5 * [-1 + \gamma v_\pi(s_2)] \\ = 0.5\gamma * [v_\pi(s_2) + v_\pi(s_3)] - 0.5$$

$$v_\pi(s_2) = 1 + \gamma v_\pi(s_4)$$
$$v_\pi(s_3) = 1 + \gamma v_\pi(s_4)$$
$$v_\pi(s_4) = 1 + \gamma v_\pi(s_4)$$

联立方程组得到:

$$v_\pi(s_1) = \frac{\gamma}{1-\gamma} - 0.5$$
$$v_\pi(s_2) = \frac{1}{1-\gamma}$$
$$v_\pi(s_3) = \frac{1}{1-\gamma}$$
$$v_\pi(s_4) = \frac{1}{1-\gamma}$$

4. 贝尔曼公式的矩阵-向量形式

已知贝尔曼公式为:

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

将$\sum_a \pi(a|s) \sum_r p(r|s,a) r$记作$r_\pi(s)$，表示在状态 s 下采取所有可能的动作，得到的 reward 的期望，记为 immediate reward

将$\sum_a \pi(a|s)p(s'|s,a)$记作$p_\pi(s'|s)$，表示从状态 s 采取任何动作到达状态s'的概率，

则状态 s 的 state value 可以表示为：

$$v_\pi(s) = r_\pi(s) + \gamma \sum_{s'} p_\pi(s'|s) v_\pi(s')$$

将所有状态使用$s_i (i = 1, \ldots, n)$进行表示，则贝尔曼公式表示为：

$$v_\pi(s_i) = r_\pi(s_i) + \gamma \sum_{s_j} p_\pi(s_j|s_i) v_\pi(s_j)$$

使用向量-矩阵来表示这些变量：

$$v_\pi = [v_\pi(s_1), \ldots, v_\pi(s_n)]^T \in R^n$$
$$r_\pi = [r_\pi(s_1), \ldots, r_\pi(s_n)]^T \in R^n$$
$$P_\pi \in R^{n \times n}, [P_\pi]_{ij} = p_\pi(s_j|s_i)$$

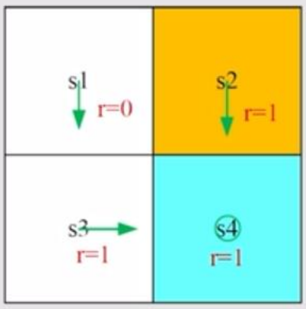
其中$p_\pi(s_j|s_i)$表示从s_i到s_j的概率，则贝尔曼公式表示为：

$$v_\pi = r_\pi + \gamma P_\pi v_\pi$$

如果 n 为 4，则该式子可以表示为：

$$\underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi} = \underbrace{\begin{bmatrix} r_\pi(s_1) \\ r_\pi(s_2) \\ r_\pi(s_3) \\ r_\pi(s_4) \end{bmatrix}}_{r_\pi} + \gamma \underbrace{\begin{bmatrix} p_\pi(s_1|s_1) & p_\pi(s_2|s_1) & p_\pi(s_3|s_1) & p_\pi(s_4|s_1) \\ p_\pi(s_1|s_2) & p_\pi(s_2|s_2) & p_\pi(s_3|s_2) & p_\pi(s_4|s_2) \\ p_\pi(s_1|s_3) & p_\pi(s_2|s_3) & p_\pi(s_3|s_3) & p_\pi(s_4|s_3) \\ p_\pi(s_1|s_4) & p_\pi(s_2|s_4) & p_\pi(s_3|s_4) & p_\pi(s_4|s_4) \end{bmatrix}}_{P_\pi} \underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi}$$

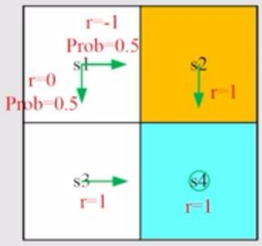
例 1：



For this specific example:

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

例 2：



For this specific example:

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

5. 贝尔曼公式的求解

policy evaluation：给定某种 policy，求解所有 state value 的过程

■ 直接法：通过公式$v_\pi = r_\pi + \gamma P_\pi v_\pi$，得到$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$

■ 迭代法（存疑）：$v_{k+1} = r_\pi + \gamma P_\pi v_k$

给定初始的v_0，得到$v_1, v_2, \ldots, v_k$，当$k \to \infty$时，$v_k \to v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$。通过这种方式我们可以得到一个 state value 序列$\{v_0, v_1, v_2, \ldots\}$

直观感觉 policy evaluation：

	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

	1	2	3	4	5
1	3.5	3.9	4.3	4.8	5.3
2	3.1	3.5	4.8	5.3	5.9
3	2.8	2.5	10.0	5.9	6.6
4	2.5	10.0	10.0	10.0	7.3
5	2.3	9.0	10.0	9.0	8.1

	1	2	3	4	5
1	-6.6	-7.3	-8.1	-9.0	-10.0
2	-8.5	-8.3	-8.1	-9.0	-10.0
3	-7.5	-8.3	-8.1	-9.0	-10.0
4	-7.5	-7.2	-9.1	-9.0	-10.0
5	-7.6	-7.3	-8.1	-9.0	-10.0

	1	2	3	4	5
1	0.0	0.0	0.0	-10.0	-10.0
2	-9.0	-10.0	-0.4	-0.5	-10.0
3	-10.0	-0.5	0.5	-0.5	0.0
4	0.0	-1.0	-0.5	-0.5	-10.0
5	0.0	0.0	0.0	0.0	0.0

6. action value

action value 和 state value 的比较：state value 是从某个状态出发，得到的所有 return 的期望；action value 是从某个状态出发，采取某种行动后（包括采取当前行动得到的 reward）得到的所有 return 的期望

action value 的定义：

$$q_\pi(s,a) = E(G_t|S_t = s, A_t = a)$$

而存在以下关系：

$$E(G_t|S_t = s) = \sum_a E(G_t|S_t = s, A_t = a)\pi(a|s)$$

即从状态 s 出发得到所有 return 的期望等于从状态 s 出发采取所有可能的行动得到的 return 的期望的期望，即存在以下关系，并由该公式可以得出由 action value 可以得到 state value：

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

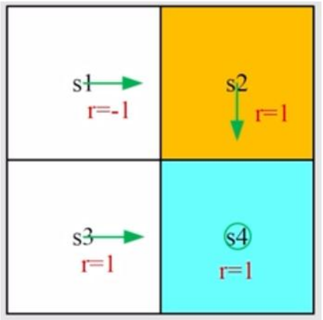
又有贝尔曼公式：

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

则可以得到，并由该公式可以得出由 state value 可以计算出 action value：

$$q_\pi(s,a) = \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v_\pi(s')$$

例子：



求$q_\pi(s_1,a_2)$，$q_\pi(s_1,a_1)$，$q_\pi(s_1,a_3)$，$q_\pi(s_1,a_4)$，$q_\pi(s_1,a_5)$?

$$q_\pi(s_1, a_2) = -1 + \gamma v_\pi(s_2)$$
$$q_\pi(s_1, a_1) = -1 + \gamma v_\pi(s_1)$$
$$q_\pi(s_1, a_3) = 0 + \gamma v_\pi(s_3)$$
$$q_\pi(s_1, a_4) = -1 + \gamma v_\pi(s_1)$$
$$q_\pi(s_1, a_5) = 0 + \gamma v_\pi(s_1)$$