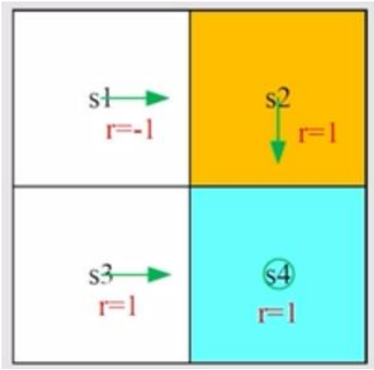
1. 背景



根据贝尔曼公式可以得到:

$$v_\pi(s_1) =- 1 + \gamma v_\pi(s_2)$$
$$v_\pi(s_2) = 1 + \gamma v_\pi(s_4)$$
$$v_\pi(s_3) = 1 + \gamma v_\pi(s_4)$$
$$v_\pi(s_4) = 1 + \gamma v_\pi(s_4)$$

令$\gamma = 0.9$，则可以得到：$v_\pi(s_1) = 8, v_\pi(s_2) = 10, v_\pi(s_3) = 10, v_\pi(s_4) = 10$

求解从s_1出发采取的所有行动得到的 action value，结果为:

$$q_\pi(s_1, a_1) =- 1 + \gamma v_\pi(s_1) = 6.2$$
$$q_\pi(s_1, a_2) =- 1 + \gamma v_\pi(s_2) = 8$$
$$q_\pi(s_1, a_3) = 0 + \gamma v_\pi(s_3) = 9$$
$$q_\pi(s_1, a_4) =- 1 + \gamma v_\pi(s_1) = 6.2$$
$$q_\pi(s_1, a_5) = 0 + \gamma v_\pi(s_1) = 7.2$$

当前的策略为$\pi(a|s_1) = \begin{cases} 1, a = a_2 \\ 0, a \neq a_2 \end{cases}$

直观判断可知该策略并不是最好的，最好的策略应该是

$$\pi_{new}(a|s_1) = \begin{cases} 1, a = a^* \\ 0, a \neq a^* \end{cases}$$

其中$a^* = argmax_a\, q_\pi(s_1, a)$，由上述计算结果可知$a^* = a_3$，故对其进行策略更新，得到$\pi(a|s_1) = \begin{cases} 1, a = a_3 \\ 0, a \neq a_3 \end{cases}$

可以看到，在进行 policy 更新之前，s_2, s_3, s_4的策略都是最优的；若不是这种情况，对s_1进行的策略更新不一定有效，但经过多次迭代更新，最终会接近于最优策略，证明该过程的工具即为贝尔曼最优公式

2. optimal policy：最优策略

若$\exists\pi^*$，有$\forall s\in S,\forall\pi,v_{\pi^*}(s)\geq v_\pi(s)$，则称$\pi^*$为最优策略

- 最优策略是否存在？
- 最优策略是否唯一？
- 最优策略是确定性的还是不确定性的？
- 如何得到最优策略？

3. 贝尔曼最优公式

贝尔曼公式定义：

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

此时策略π是给定的，贝尔曼最优公式定义：

$$\begin{aligned}v(s)&=\max_\pi\sum_a \pi(a|s)\left[\sum_r p(r|s,a)r+\gamma\sum_{s'}p(s'|s,a)v(s')\right]\\&=\max_\pi\sum_a \pi(a|s)q(s,a)\end{aligned}$$

此时$p(r|s,a)$和$p(s'|s,a)$是已知的，策略π是未知的，$v(s)$和$v(s')$是未知的并且是我们的计算目标

矩阵-向量形式，与贝尔曼公式的推导过程相似：

$$v=\max_\pi(r_\pi+\gamma P_\pi v)$$

其中，

$$[r_\pi]_s\triangleq\sum_a \pi(a|s)\sum_r p(r|s,a)r$$

$$[P_\pi]_{s,s'}=p(s'|s)\triangleq\sum_a \pi(a|s)\sum_{s'}p(s'|s,a)$$

4. 求解

贝尔曼最优公式的矩阵-向量形式为：

$$v = \max_{\pi} (r_{\pi} + \gamma P_{\pi} v)$$

在这里我们要求解两个未知量，分别为v和π。那么如何求解带 max 的问题呢？

看以下例子：

$$x = \max_{a} (2x - 1 - a^2)$$

则显然当$a = 0$ 时，x取得最大值，因此可以得到$x = 2x - 1$，即$a = 0, x = 1$

贝尔曼最优公式的定义为：

$$v(s) = \max_{\pi} \sum_{a} \pi(a|s) \left[\sum_{r} p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v(s') \right]$$
$$= \max_{\pi} \sum_{a} \pi(a|s) q(s,a)$$

在这里我们假设在网格世界中，则$q(s,a)$共有$q(s,a_1),\dots,q(s,a_5)$，要求得最优的π，使得$v(s)$最大，如何求解，看以下例子：

假设q_1, q_2, q_3是给定的向量，则需找到常量c_1, c_2, c_3，求得以下结果：

$$\max_{c_1,c_2,c_3} c_1 q_1 + c_2 q_2 + c_3 q_3 \,, c_1 + c_2 + c_3 = 1$$

假设$q_3 \geq q_1, q_2$，则最优解为$c_3 = 1, c_1 = \ c_2 = 0$

最后得到结果为：$v(s) = \max_{\pi} \sum_a \pi(a|s) q(s,a) = \max_{a \in A(s)} q(s,a)$

得到的最优策略为：

$$\pi(a|s) = \begin{cases} 1, a = a^* \\ 0, a \neq a^* \end{cases}, a^* = argmax_a \, q(s,a)$$

将贝尔曼最优公式表示为：

$$v = f(v)$$
$$f(v) := \max_{\pi} (r_{\pi} + \gamma P_{\pi} v)$$

■ 不动点

若$\exists x \in X$，有映射关系$f: X \rightarrow X$，满足$f(x) = x$，则 x 为不动点。

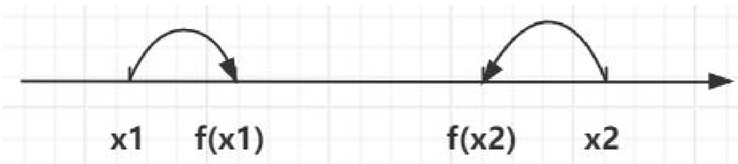
例如：$f(x) = 0.5x$，求解$f(x) = x = 0.5x$得到$x = 0$，即为不动点

■ contraction mapping：压缩映射

若存在映射关系f，使得

$$\|f(x_1) - f(x_2)\| \leq \gamma \|x_1 - x_2\|, \ \gamma \in (0, 1)$$

则该映射关系f被称为压缩映射



映射关系$f(x) = 0.5x$为压缩映射，因为满足：

$$\|0.5x_1 - 0.5x_2\| = 0.5\|x_1 - x_2\| \leq \gamma \|x_1 - x_2\|, \ \gamma \in [0.5, 1)$$

对矩阵-向量形式依然成立：

若有映射关系$f(x) = Ax, x \in R^n, A \in R^{n \times n}, \|A\| \leq \gamma < 1$，则有

$$\|Ax_1 - Ax_2\| = \|A(x_1 - x_2)\| \leq \|A\|\|x_1 - x_2\| \leq \gamma \|x_1 - x_2\|$$

若映射关系f为压缩映射，则其：

（1） 一定存在不动点x^*

（2） 不动点是唯一的

（3） 存在序列$\{x_k\}$且满足$x_{k+1} = f(x_k)$，当$k \rightarrow \infty, x_k \rightarrow x^*$

贝尔曼最优公式：$v = f(v) = \max_{\pi} (r_\pi + \gamma P_\pi v)$

$f(v)$为压缩映射（不证明）：

$$\|f(v_1) - f(v_2)\| \leq \gamma \|v_1 - v_2\|$$

故该解v^*一定存在且唯一，且可通过迭代算法进行求解：

$$v_{k+1} = f(v_k) = \max_{\pi} (r_\pi + \gamma P_\pi v_k)$$

则可求得最优 state value 和最优 policy：

$$v^* = \max_{\pi}(r_\pi + \gamma P_\pi v^*)$$

$$\pi^* = \arg\max_{\pi}(r_\pi + \gamma P_\pi v^*)$$

则：

$$v^* = r_{\pi^*} + \gamma P_{\pi^*} v^*$$

最优策略表示为：

For any $s \in \mathcal{S}$, the deterministic greedy policy

$$\pi^*(a|s) = \begin{cases} 1 & a = a^*(s) \\ 0 & a \neq a^*(s) \end{cases}$$

is an optimal policy solving the BOE. Here,

$$a^*(s) = \arg\max_a q^*(a, s),$$

where $q^*(s,a) := \sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v^*(s')$

5. 哪些因素决定最优策略？

贝尔曼最优公式为：

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

红色部分是已知部分，故决定最优策略的因素有：

- Reward design： r
- System model： $p(s'|s,a), p(r|s,a)$
- Discount rate： γ