

# 强化学习

何家豪

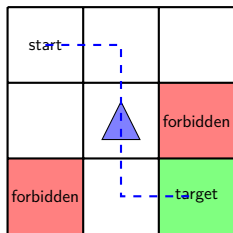
hejiahao@ruc.edu.cn  
中国人民大学信息学院

2025 年 2 月 20 日

# 目录

1. 基本概念

2. 贝尔曼公式



- 格子：可通过/禁止通过/目标，边界

任务：

- 给定一个起始点，找出一条“好”的路径到终点。
- 怎么定义“好”：尽量避免 forbidden 的格子，少走重复的路，尽量不要撞到边界。

# 状态

**状态：**智能体在环境中被描述的状态。在网格世界中，智能体的位置就是他的状态。  
网格世界中一共有 9 个可能的位置，所以智能体也就有 9 个可能的状态： $s_1, s_2, \dots, s_9$ 。

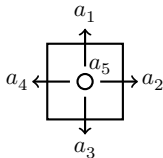
|       |       |       |
|-------|-------|-------|
| $s_1$ | $s_2$ | $s_3$ |
| $s_4$ | $s_5$ | $s_6$ |
| $s_7$ | $s_8$ | $s_9$ |

**状态空间：**所有状态的集合  $\mathcal{S} = \{s_i\}_{i=1}^9$

# 动作

**动作：**对于每个状态，有 5 个可能的动作： $a_1, a_2, \dots, a_5$

- $a_1$ ：向上移动
- $a_2$ ：向右移动
- $a_3$ ：向下移动
- $a_4$ ：向左移动
- $a_5$ ：停留



|       |       |       |
|-------|-------|-------|
| $s_1$ | $s_2$ | $s_3$ |
| $s_4$ | $s_5$ | $s_6$ |
| $s_7$ | $s_8$ | $s_9$ |

**状态的动作空间：**智能体在一个状态可以做的所有动作的集合。 $\mathcal{A}(s_i) = \{a_k\}_{k=1}^5$

**问题：**不同状态的动作空间一样吗？

# 状态转移

|       |       |       |
|-------|-------|-------|
| $s_1$ | $s_2$ | $s_3$ |
| $s_4$ | $s_5$ | $s_6$ |
| $s_7$ | $s_8$ | $s_9$ |

先关注 *forbidden* 的格子：如果在状态  $s_5$  选择了行动  $a_2$  会发生什么？

- 第一种情况：*forbidden* 的格子可以进入，但是会有惩罚。那么，

$$s_5 \xrightarrow{a_2} s_6$$

- 第二种情况：*forbidden* 的格子不能进入，那么，

$$s_5 \xrightarrow{a_2} s_5$$

之后我们基本上考虑的是第一种情况。

# 状态转移

|       |       |       |
|-------|-------|-------|
| $s_1$ | $s_2$ | $s_3$ |
| $s_4$ | $s_5$ | $s_6$ |
| $s_7$ | $s_8$ | $s_9$ |

我们可以用一个表格去形容状态转移。(只能表述确定性的情况)

|       | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|-------|-------|-------|-------|-------|-------|
| $s_1$ | $s_1$ | $s_2$ | $s_4$ | $s_1$ | $s_1$ |
| $s_2$ | $s_2$ | $s_3$ | $s_5$ | $s_1$ | $s_2$ |
| $s_3$ | $s_3$ | $s_3$ | $s_6$ | $s_2$ | $s_3$ |
| $s_4$ | $s_1$ | $s_5$ | $s_7$ | $s_4$ | $s_4$ |
| $s_5$ | $s_6$ | $s_8$ | $s_4$ | $s_2$ | $s_5$ |
| $s_6$ | $s_3$ | $s_6$ | $s_9$ | $s_5$ | $s_6$ |
| $s_7$ | $s_4$ | $s_8$ | $s_7$ | $s_7$ | $s_7$ |
| $s_8$ | $s_5$ | $s_9$ | $s_8$ | $s_7$ | $s_8$ |
| $s_9$ | $s_6$ | $s_9$ | $s_9$ | $s_8$ | $s_9$ |

# 状态转移

|       |       |       |
|-------|-------|-------|
| $s_1$ | $s_2$ | $s_3$ |
| $s_4$ | $s_5$ | $s_6$ |
| $s_7$ | $s_8$ | $s_9$ |

状态转移概率：使用概率来描述状态转移。

- 如果我们在状态  $s_1$  选择动作  $a_2$ ，那么我们的下一个状态是  $s_2$ 。
- 数学：

$$P(s_2|s_1, a_2) = 1$$

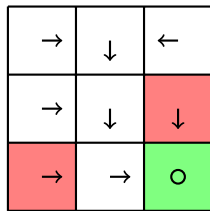
$$P(s_i|s_1, a_2) = 0, \quad \forall i \neq 2$$

这里是**确定性**的情况，还有**不确定性**的情况（比如说会刮风）。

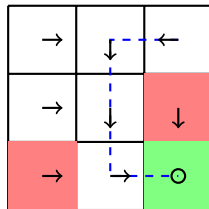
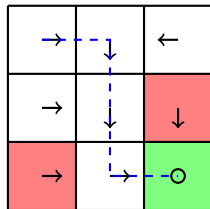
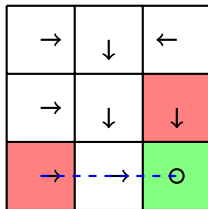


# 策略

**策略**告诉智能体在一个状态下应该选择什么样的动作。  
用箭头表示策略。



根据这个策略，我们可以从不同的出发点开始得到不同的轨迹。



# 策略

|   |   |   |
|---|---|---|
| → | ↓ | ← |
| → | ↓ | ↓ |
| → | → | ○ |

数学表示：用条件概率表示策略。

对于  $s_1$  而言：

$$\pi(a_1|s_1) = 0$$

$$\pi(a_2|s_1) = 1$$

$$\pi(a_3|s_1) = 0$$


$$\pi(a_4|s_1) = 0$$

$$\pi(a_5|s_1) = 0$$

这是一个确定性策略。

# 策略

随机策略：

|   |   |   |
|---|---|---|
|  | ↓ | ← |
| →   | ↓ | ↓ |
| →   | → | ○ |

在这个策略中，对于  $s_1$  而言：

$$\pi(a_1|s_1) = 0$$

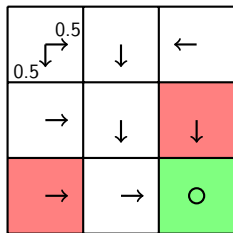
$$\pi(a_2|s_1) = 0.5$$

$$\pi(a_3|s_1) = 0.5$$

$$\pi(a_4|s_1) = 0$$

$$\pi(a_5|s_1) = 0$$

# 策略

|  |  |  |
|--|--|--|
|  |  |  |
|  |  |  |
|  |  |  |

用表格形式表示一个策略。

|       | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ |
|-------|-------|-------|-------|-------|-------|
| $s_1$ | 0     | 0.5   | 0.5   | 0     | 0     |
| $s_2$ | 0     | 0     | 1     | 0     | 0     |
| ...   |       |       | ...   |       |       |
| $s_9$ | 0     | 0     | 0     | 0     | 1     |

这样既可以表示确定性策略，也可以表示随机策略。

奖励是一个智能体做出动作之后得到的一个实数。

- 一个正数的奖励值鼓励智能体做这样的动作；
- 一个负数的奖励值惩罚智能体做这样的动作。

问题：

- 可以让正数代表惩罚，负数代表鼓励吗？
  - 可以。
  - 在这种情况下，奖励被称作代价。
- 奖励值为 0 会怎么样？
  - 只有相对奖励值才有意义，绝对的奖励值没有意义。
  - $r = \{+1, -1\}$  变成  $r = \{+2, 0\}$  并不会对策略有任何影响。

# 奖励

|       |       |       |
|-------|-------|-------|
| $s_1$ | $s_2$ | $s_3$ |
| $s_4$ | $s_5$ | $s_6$ |
| $s_7$ | $s_8$ | $s_9$ |

在网格世界中，我们可以这样设置奖励：

- 如果智能体撞墙了，那么奖励值  $r_{\text{bound}} = -1$ ；
- 如果智能体试图进入一个禁止通过的区域，那么奖励值  $r_{\text{forbidden}} = -1$
- 如果智能体到达终点，那么奖励值  $r_{\text{target}} = +1$
- 其他情况奖励值  $r = 0$

奖励值的存在可以让我们人类引导智能体按照我们想要的方式行动。  
比如上述奖励的设置可以让智能体尽量不要撞墙或者试图进入禁止通过的区域。

# 奖励

|       |       |       |
|-------|-------|-------|
| $s_1$ | $s_2$ | $s_3$ |
| $s_4$ | $s_5$ | $s_6$ |
| $s_7$ | $s_8$ | $s_9$ |

用表格形式表示奖励机制。

|       | $a_1$                  | $a_2$              | $a_3$              | $a_4$              | $a_5$               |
|-------|------------------------|--------------------|--------------------|--------------------|---------------------|
| $s_1$ | $r_{\text{bound}}$     | 0                  | 0                  | $r_{\text{bound}}$ | 0                   |
| $s_2$ | $r_{\text{bound}}$     | 0                  | 0                  | 0                  | 0                   |
| ...   |                        |                    | ...                |                    |                     |
| $s_9$ | $r_{\text{forbidden}}$ | $r_{\text{bound}}$ | $r_{\text{bound}}$ | 0                  | $r_{\text{target}}$ |

只能用来表示确定性的奖励机制。

|       |       |       |
|-------|-------|-------|
| $s_1$ | $s_2$ | $s_3$ |
| $s_4$ | $s_5$ | $s_6$ |
| $s_7$ | $s_8$ | $s_9$ |

数学形式：条件概率

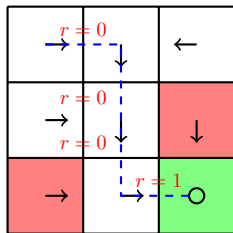
- 在  $s_1$  我们选择动作  $a_1$  奖励会是-1。
- $p(r = -1 | s_1, a_1) = 1$  且  $p(r \neq -1 | s_1, a_1) = 0$

值得注意的是：

在一个确定的状态执行一个确定的动作，得到的奖励并不是确定的。比如努力学习一定会得到奖励，但是具体多少是不一定的。



# 轨迹和回报



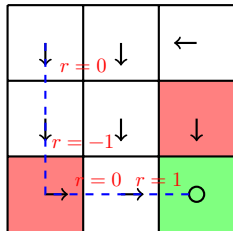
轨迹是一个状态-动作-奖励链。

$$s_1 \xrightarrow[r=0]{a_2} s_2 \xrightarrow[r=0]{a_3} s_5 \xrightarrow[r=0]{a_3} s_8 \xrightarrow[r=1]{a_2} s_9$$

回报是这一条轨迹上所有的奖励之和。

$$\text{return} = 0 + 0 + 0 + 1 = 1$$

# 轨迹和回报



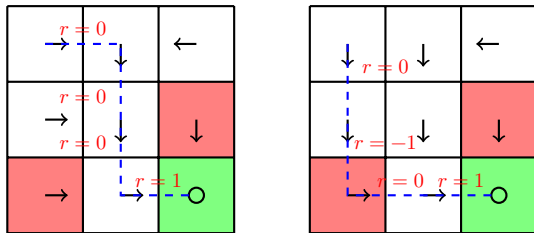
不同的策略会有不同的轨迹

$$s_1 \xrightarrow[r=0]{a_3} s_4 \xrightarrow[r=-1]{a_3} s_7 \xrightarrow[r=0]{a_2} s_8 \xrightarrow[r=1]{a_2} s_9$$

回报是这一条轨迹上所有的奖励之和。

$$\text{return} = 0 + (-1) + 0 + 1 = 1$$

# 轨迹和回报

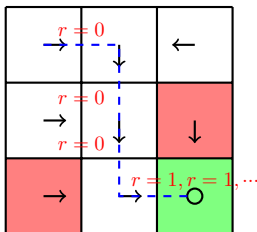


哪个策略更好一点？

- 第一个更好，因为它没有经过禁止通过的区域；
- 第一个更好，因为它的**回报**更高。

回报可以衡量一个策略的好坏。

# 折扣回报



轨迹可能是无限的：

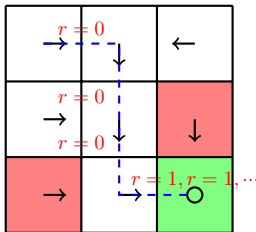
$$s_1 \xrightarrow[r=0]{a_2} s_2 \xrightarrow[r=0]{a_3} s_5 \xrightarrow[r=0]{a_3} s_8 \xrightarrow[r=1]{a_2} s_9 \xrightarrow[r=1]{a_5} s_9 \xrightarrow[r=1]{a_5} s_9 \dots$$

回报是：

$$\text{return} = 0 + 0 + 0 + 1 + 1 + \dots$$

这样的回报是发散的，是属于一种无效的定义。

# 折扣回报



需要引入一种折扣系数  $\gamma \in (0, 1)$ :

$$\begin{aligned}\text{折扣回报} &= 0 + \gamma 0 + \gamma^2 0 + \gamma^3 1 + \gamma^4 1 + \gamma^5 1 + \dots \\ &= \gamma^3 (1 + \gamma + \gamma^2 + \dots) = \gamma^3 \frac{1}{1 - \gamma}\end{aligned}$$

$\gamma$  的作用: 1) 求和会变成一个有限的数; 2) 可以用来平衡短期收益和长期收益。

- 如果  $\gamma$  更接近 0, 那么折扣回报会更依赖于短期收益;
- 如果  $\gamma$  更接近 1, 那么折扣回报会更依赖于长期收益。

# 马尔可夫决策过程 (MDP)

MDP 的关键要素:

- 集合:
  - 状态: 状态集合  $\mathcal{S}$
  - 动作: 动作集合  $\mathcal{A}(s)$  是与状态  $s \in \mathcal{S}$  关联的
  - 奖励: 奖励集合  $\mathcal{R}(s, a)$
- 概率分布 (或被称为系统模型):
  - 状态转移概率: 在状态  $s$ , 采取动作  $a$ , 转移到状态  $s'$  的概率  $p(s'|s, a)$
  - 奖励概率: 在状态  $s$ , 采取动作  $a$ , 得到奖励  $r$  的概率  $p(r|s, a)$
- 策略: 在状态  $s$  采取动作  $a$  的概率  $\pi(a|s)$
- 马尔可夫性质: 无记忆性质

$$p(s_{t+1}|a_t, s_t, \dots, a_0, s_0) = p(s_{t+1}|a_t, s_t)$$

$$p(r_{t+1}|a_t, s_t, \dots, a_0, s_0) = p(r_{t+1}|a_t, s_t)$$

所有本节提到的概念都可以放入马尔可夫决策过程中理解。

本节通过网格世界的例子，阐述了以下概念：

- 状态
- 动作
- 状态转移，状态转移概率  $p(s'|s, a)$
- 奖励，奖励概率  $p(r|s, a)$
- 轨迹，回报，折扣回报
- 马尔可夫决策过程（MDP）。

# 目录

1. 基本概念

2. 贝尔曼公式



# 标注说明

考虑以下单步过程：

$$S_t \xrightarrow{A_t} R_{t+1}, S_{t+1}$$

- $t, t+1$ ：离散时间序列
- $S_t$ ：在时刻  $t$  的状态
- $A_t$ ：在时刻  $t$  的动作
- $R_{t+1}$ ：执行完  $A_t$  之后的奖励
- $S_{t+1}$ ：执行完  $A_t$  之后的状态

注意  $S_t, A_t, R_{t+1}$  都是随机变量。

- $A_t \sim \pi(A_t | S_t = s)$
- $R_{t+1} \sim p(R_{t+1} | S_t = s, A_t = a)$
- $S_{t+1} \sim p(S_{t+1} | S_t = s, A_t = a)$

假设所有的上述概率分布（系统模型）都是已知的。

考虑以下轨迹：

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

其折扣回报为：

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

- $\gamma \in (0, 1)$ ：折扣系数
- $G_t$ ：也是一个随机变量，因为  $R_{t+1}, R_{t+2}, \dots$  都是随机变量

# 状态价值

$G_t$  的数学期望值被称为**状态价值**。

$$v_{\pi}(s) = \mathbb{E}[G_t | S_t = s]$$

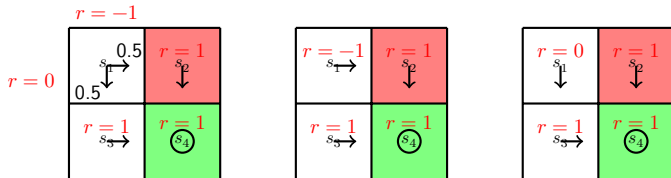
- 它是一个  $s$  的函数。它是当状态处于  $s$  时的条件期望。
- 它是一个  $\pi$  的函数。对于不同的策略，状态价值是不同的。

问题：状态价值和回报的关系是什么？

答：状态价值是从一个状态出发得到的回报的期望值。

# 状态价值

从  $s_1$  出发下面哪一个策略最好?



$$v_{\pi_1}(s_1) = 0.5 \left( -1 + \frac{\gamma}{1-\gamma} \right) + 0.5 \left( \frac{\gamma}{1-\gamma} \right) = -0.5 + \frac{\gamma}{1-\gamma}$$

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

$$v_{\pi_3}(s_1) = 0 + \gamma 1 + \gamma^2 1 + \dots = \frac{\gamma}{1-\gamma}$$

# 贝尔曼公式

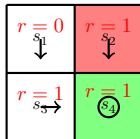
如何计算状态价值  $v_{\pi}(s)$ ?

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

- 以上方程被称作贝尔曼公式。它描述了不同状态价值之间的关系。
- 它由两项组成：一项是即时奖励，一项是长期奖励。
- 贝尔曼公式实际上是一组方程，每一个状态都有它对应的贝尔曼公式。

# 贝尔曼公式

考虑以下策略：



贝尔曼公式：

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

$s_1$  的状态价值：

- $\pi(a = a_3|s_1) = 1$  且  $\pi(a \neq a_3|s_1) = 0$
- $p(s' = s_3|s_1, a_3) = 1$  且  $p(s'|s_1, a \neq a_3) = 0$
- $p(r = 0|s_1, a_3) = 1$  且  $p(r \neq 0|s_1, a_3) = 0$
- $v_{\pi}(s_1) = 0 + v_{\pi}(s_3)$

# 贝尔曼公式

贝尔曼公式：

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

同理可得（取  $\gamma = 0.9$ ）：

$$v_{\pi}(s_1) = 0 + \gamma v_{\pi}(s_3) = \frac{\gamma}{1 - \gamma} = 9$$

$$v_{\pi}(s_2) = 1 + \gamma v_{\pi}(s_4) = \frac{1}{1 - \gamma} = 10$$

$$v_{\pi}(s_3) = 1 + \gamma v_{\pi}(s_3) = \frac{1}{1 - \gamma} = 10$$

$$v_{\pi}(s_4) = 1 + \gamma v_{\pi}(s_4) = \frac{1}{1 - \gamma} = 10$$

从状态价值到动作价值：

- 状态价值：智能体从一个状态出发得到的回报的期望值
- **动作价值**：智能体从一个状态出发执行了一个动作之后得到的回报的期望值

动作价值可以让我们知道在一个状态下采取哪个动作更好。



# 动作价值

定义：

$$q_{\pi}(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a]$$

- $q$  是一个“状态-动作”对的函数。
- $q$  是一个  $\pi$  的函数。对于不同的策略，动作价值是不同的。

它服从条件期望的公式：

$$\mathbb{E}[G_t | S_t = s] = \sum_a \mathbb{E}[G_t | S_t = s, A_t = a] \pi(a|s) \quad (1)$$

所以：

$$v_{\pi}(s) = \sum_a \pi(a|s) q_{\pi}(s, a) \quad (2)$$

# 动作价值

贝尔曼公式：

$$v_{\pi}(s) = \sum_a \pi(a|s) \left[ \sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_{\pi}(s') \right] \quad (3)$$

所以可以发现：

$$q_{\pi}(s, a) = \sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_{\pi}(s') \quad (4)$$

- (2) 式表达了如何用动作价值计算状态价值
- (4) 式表达了如何用状态价值计算动作价值

# 动作价值

贝尔曼公式：

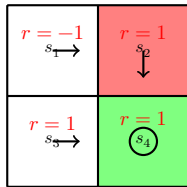
$$v_{\pi}(s) = \sum_a \pi(a|s) \left[ \sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_{\pi}(s') \right] \quad (5)$$

所以可以发现：

$$q_{\pi}(s, a) = \sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_{\pi}(s') \quad (6)$$

- (2) 式表达了如何用动作价值计算状态价值
- (4) 式表达了如何用状态价值计算动作价值

# 动作价值



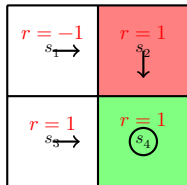
写出  $s_1$  的动作价值:

$$q_{\pi}(s_1, a_2) = -1 + \gamma v_{\pi}(s_2)$$

问题:

- $q_{\pi}(s_1, a_1), q_{\pi}(s_1, a_3), q_{\pi}(s_1, a_4), q_{\pi}(s_1, a_5)$  分别是多少?

# 动作价值



- 动作价值比较关键的原因是我们最关心的其实是要做什么动作。
- 我们可以先把所有状态价值都算出来，然后再算所有的动作价值
- 即使没有模型，我们也可以估计动作价值。

关键概念和结论：

- 状态价值：  $v_{\pi}(s) = \mathbb{E}[G_t | S_t = s]$
- 动作价值：  $q_{\pi}(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a]$
- 贝尔曼公式：

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