



# 强化学习导论

## Introduction to Reinforcement Learning

作者：刘俊宏

组织：哈尔滨工业大学（深圳）



“海棠”系列丛书

# 目录

第一章 导论	1
第二章 贝尔曼公式	2
2.1 状态价值 . . . . .	2
2.2 贝尔曼公式推导 . . . . .	2
2.3 矩阵与向量形式求解 . . . . .	3
2.4 动作价值 . . . . .	4
第三章 贝尔曼最优公式	5
3.1 待定 . . . . .	5

# 第一章 导论

强化学习的目的就是寻找最优策略！

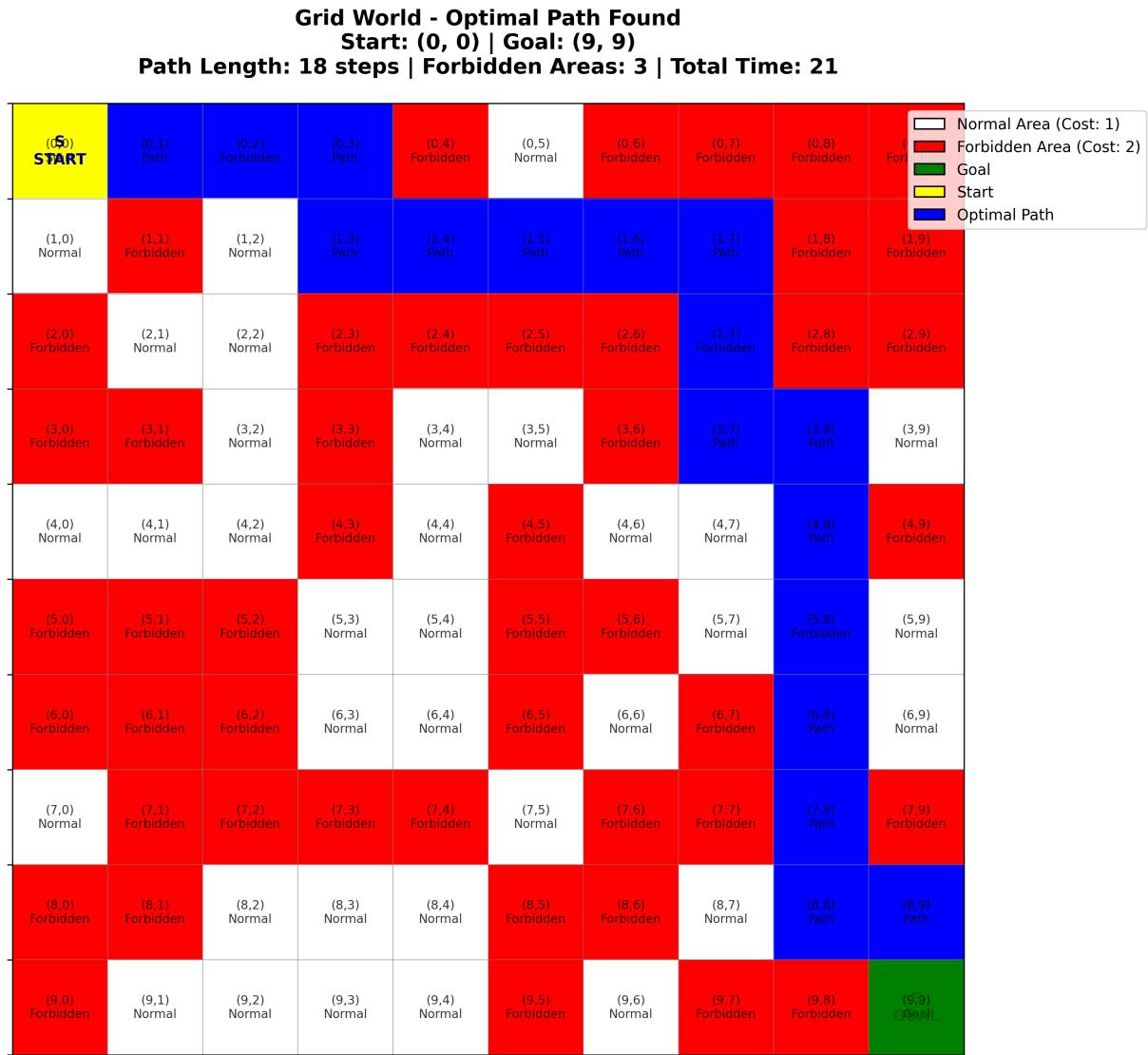


图 1.1: 利用强化学习规划耗时最短的路径

## 第二章 贝尔曼公式

### 内容提要

- 状态价值
- 贝尔曼公式推导

- 矩阵与向量形式求解
- 动作价值

### 2.1 状态价值

我们首先考虑下述多步轨迹：

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

以  $t$  时刻的单步过程为例，系统在时间  $t$  采取动作  $A_t$  从状态  $S_t$  转移到了状态  $S_{t+1}$ ，获得的即时奖励（或动作奖励）为  $R_{t+1}$ 。这里需要明确的一点是，上述过程的行为都是由概率分布决定的：

- $S_t \rightarrow A_t$  是由  $\pi(A_t = a | S_t = s)$  决定的。
- $S_t, A_t \rightarrow R_{t+1}$  是由  $p(R_{t+1} = r | S_t = s, A_t = a)$  决定的。
- $S_t, A_t \rightarrow S_{t+1}$  是由  $p(S_{t+1} = s' | S_t = s, A_t = a)$  决定的。

通过上述轨迹我们可以得到折扣回报（Discount Return）(2.1)，其中  $\gamma$  为折扣率（Discount Rate）。

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \quad (2.1)$$

而状态价值（State Value）(2.2) 所代表的含义就是折扣回报的期望（或者均值）。状态价值是有关状态  $s$  的函数，所以从不同的状态出发得到的状态价值是不同的。此外状态价值与策略  $\pi$  有关，不同的策略可能得到不同的状态价值。当状态价值的值越大，代表的当前策略效果越好。

$$v_\pi(s) = \mathbb{E}[G_t | S_t = s] \quad (2.2)$$

 **笔记** 这里之所以使用折扣回报而不是直接将奖励相加主要基于以下两点：

1. 如果只是简单地将奖励相加，那么在无限长的轨迹中回报就是发散并趋近于无穷的。
2. 从直觉上来讲，当前时刻的奖励对未来的影响是逐渐递减的。

 **笔记** 此外折扣回报与状态价值的区别与联系在于：

- 折扣回报表示的是当系统处在某个状态下，单个动作轨迹的奖励。
- 状态价值评估的是当系统处在某个状态下，所有可能的动作轨迹的期望奖励。

所以说，当系统处在某个状态下并且它的动作轨迹是确定的（也就是只存在一条动作轨迹），那么此时折扣回报与状态价值是等价的。

### 2.2 贝尔曼公式推导

通过公式 (2.1) 我们可以有如下推导：

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

从上述式子可以看出，对于单条轨迹而言，折扣回报  $G_t$  代表了当下时刻的即时奖励  $R_{t+1}$  和未来的折扣奖励  $G_{t+1}$ 。那么同样的，我们也可以对状态价值函数进行拆分：

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

接下来我们分别分析上述两个式子：

$$\begin{aligned} \mathbb{E}[R_{t+1} | S_t = s] &= \sum_a \pi(a|s) \mathbb{E}[R_{t+1} | S_t = s, A_t = a] \\ &= \sum_a \pi(a|s) \sum_r p(r|s, a)r \end{aligned} \tag{2.3}$$

公式 (2.3) 所表示的就是即时奖励的期望。系统在状态  $s$  下根据策略  $\pi$  可能会有多个可能的动作  $a$ ，从状态  $s$  并执行动作  $a$  后又可能有不同概率产生不同的奖励  $r$ 。

$$\begin{aligned} \mathbb{E}[G_{t+1} | S_t = s] &= \sum_{s'} \mathbb{E}[G_{t+1} | S_t = s, S_{t+1} = s'] p(s'|s) \\ &= \sum_{s'} \mathbb{E}[G_{t+1} | S_{t+1} = s'] p(s'|s) \\ &= \sum_{s'} v_\pi(s') p(s'|s) \\ &= \sum_{s'} v_\pi(s') \sum_a p(s'|s, a) \pi(a|s) \end{aligned} \tag{2.4}$$

公式 (2.4) 所表示的就是未来奖励的期望。系统在状态  $s$  下根据策略  $\pi$  可能会有多个可能的动作  $a$ ，执行这些动作可能会导致系统转移到多个不同的状态  $s'$ ，而又可能存在多个不同的动作可以实现系统从状态  $s$  转移到同一个状态  $s'$ ，不同的状态  $s'$  都有相对应的状态价值  $v_\pi(s')$ 。此外，根据马尔可夫链的无记忆性质我们可以得到  $\mathbb{E}[G_{t+1} | S_t = s, S_{t+1} = s'] = \mathbb{E}[G_{t+1} | S_{t+1} = s']$ 。

通过公式 (2.3) 和 (2.4) 我们就可以得到贝尔曼公式 (Bellman Equation) (2.5)，该公式描述了不同状态价值之间的关系，并且其中包括了即时奖励期望和未来奖励期望两部分。

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[G_t | S_t = s] \\ &= \sum_a \pi(a|s) \sum_r p(r|s, a)r + \gamma \sum_{s'} v_\pi(s') \sum_a p(s'|s, a) \pi(a|s) \\ &= \sum_a \pi(a|s) \sum_r p(r|s, a)r + \gamma \sum_{s'} v_\pi(s') p(s'|s, a) \sum_a \pi(a|s) \\ &= \sum_a \pi(a|s) \left[ \sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a) v_\pi(s') \right], \forall s \in S \end{aligned} \tag{2.5}$$

 **笔记** 需要注意的是，贝尔曼公式描述的是不同状态价值之间的关系，所以在强化学习的过程中会同时存在多个表示状态转移关系的贝尔曼公式而非一个，用来完整描述任意状态价值之间的关系。

## 2.3 矩阵与向量形式求解

为了方便表示，我们首先简化贝尔曼公式得到：

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

其中  $r_\pi(s) = \sum_a \pi(a|s) \sum_r p(r|s, a)r$  表示的是即时奖励的期望， $p_\pi(s'|s) = \sum_a \pi(a|s) p(s'|s, a)$  表示的是状态转移的概率。为了方便进行矩阵表达，我们可以进一步改写上述公式。

假设存在状态  $s_i (i = 1, \dots, n)$ , 我们可以写出相对应的贝尔曼公式:

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

那么对于所有的状态, 其矩阵向量对的表达形式为 (2.6):

$$v_\pi = r_\pi + \gamma P_\pi v_\pi \quad (2.6)$$

其中:

- $v_\pi = [v_\pi(s_1), \dots, v_\pi(s_n)]^\top \in \mathbb{R}^n$
- $r_\pi = [r_\pi(s_1), \dots, r_\pi(s_n)]^\top \in \mathbb{R}^n$
- $P_\pi \in \mathbb{R}^{n \times n}$  是状态转移矩阵, 其中  $[P_\pi]_{ij} = p_\pi(s_j|s_i)$

要求解状态价值, 即  $v_\pi$  的值, 我们有两种方法: 一种是直接给出  $v_\pi$  的解析表达式, 一种是通过迭代的方式求近似解。解析解的表达式为  $v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$ , 推导如下:

$$\begin{aligned} v_\pi &= r_\pi + \gamma P_\pi v_\pi \\ v_\pi - \gamma P_\pi v_\pi &= r_\pi \\ (I - \gamma P_\pi)v_\pi &= r_\pi \\ v_\pi &= (I - \gamma P_\pi)^{-1} r_\pi \end{aligned}$$

因为需要求逆, 所以直接求解析解的计算量会比较大。在实际应用中一般使用公式 (2.7) 所展示的迭代方法求近似解:

$$v_{k+1} = r_\pi + \gamma P_\pi v_k \quad (2.7)$$

该方法通过随机初始一个  $v_0$ , 并求出相应的  $v_1$ , 再持续迭代下去直到  $v_k \rightarrow v_\pi$  当  $k \rightarrow \infty$ 。

**证明** 这里我们证明为什么当  $k$  趋近于无穷时,  $v_k$  趋近于  $v_\pi$ 。

首先定义一个误差  $\delta_k = v_k - v_\pi$ , 于是我们只需要证明  $\delta \rightarrow 0$ 。接着我们将  $v_{k+1} = \delta_{k+1} + v_\pi$  和  $v_k = \delta_k + v_\pi$  代入到  $v_{k+1} = r_\pi + \gamma P_\pi v_k$  中, 得到:

$$\begin{aligned} \delta_{k+1} + v_\pi &= r_\pi + \gamma P_\pi(\delta_k + v_\pi) \\ \delta_{k+1} &= -v_\pi + r_\pi + \gamma P_\pi \delta_k + \gamma P_\pi v_\pi \\ \delta_{k+1} &= \gamma P_\pi \delta_k + (-v_\pi + r_\pi + \gamma P_\pi v_\pi) \\ \delta_{k+1} &= \gamma P_\pi \delta_k \end{aligned}$$

接着我们可以得到等式  $\delta_{k+1} = \gamma P_\pi \delta_k = \gamma^2 P_\pi^2 \delta_{k-1} = \dots = \gamma^{k+1} P_\pi^{k+1} \delta_0$ 。因为  $0 \leq P_\pi^{k+1} \leq 1$  且  $\gamma < 1$ , 所以  $\delta_{k+1} = \gamma^{k+1} P_\pi^{k+1} \delta_0 \rightarrow 0$ 。

## 2.4 动作价值

动作价值 (Action Value) 的定义如公式 (2.8) 所示。

$$q_\pi(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a] \quad (2.8)$$

我们回忆状态价值的表达式为:  $v_\pi(s) = \mathbb{E}[G_t | S_t = s] = \sum_a \mathbb{E}[G_t | S_t = s, A_t = a] \pi(a|s)$ , 所以动作价值与状态价值之间的关系可以用下列式子 (2.9) 表示, 该式子展示了如何通过动作价值计算状态价值:

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s, a) \quad (2.9)$$

同时, 我们回顾贝尔曼公式 (2.5):  $v_\pi(s) = \sum_a \pi(a|s) [\sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_\pi(s')]$  可以发现存在动作价值  $q_\pi(s, a)$  的等价关系 (2.10), 该式子展示了如何通过状态价值计算动作价值:

$$q_\pi(s, a) = \sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_\pi(s') \quad (2.10)$$

由此我们可以发现公式 (2.9) 和 (2.10) 就像硬币的正反两面, 提供了状态价值与动作价值相互求解的方法。

# 第三章 贝尔曼最优公式

## 内容提要

□

### 3.1 待定

贝尔曼最优公式 (Bellman Optimality Equation) 的表达式 (3.1) 如下所示:

$$\begin{aligned} v_\pi(s) &= \max_{\pi} \sum_a \pi(a|s) \left[ \sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_\pi(s') \right], \forall s \in S \\ &= \max_{\pi} \sum_a \pi(a|s)q(s, a), \forall s \in S \end{aligned} \quad (3.1)$$

在上述公式中,  $p(r|s, a)$  和  $p(s'|s, a)$  是已知的;  $v_\pi(s)$  和  $v_\pi(s')$  是未知的并需要被求解; 策略  $\pi$  也是需要求解的。

接下来我们将上述式子转化为矩阵向量的形式:  $v = \max_{\pi} (r_\pi + \gamma P_\pi v)$ , 其中  $[r_\pi]_s = \sum_a \pi(a|s) \sum_r p(r|s, a)r$ ,  $[P_\pi]_{s's} = p(s'|s) = \sum_a \pi(a|s) \sum_{s'} p(s'|s, a)$ 。