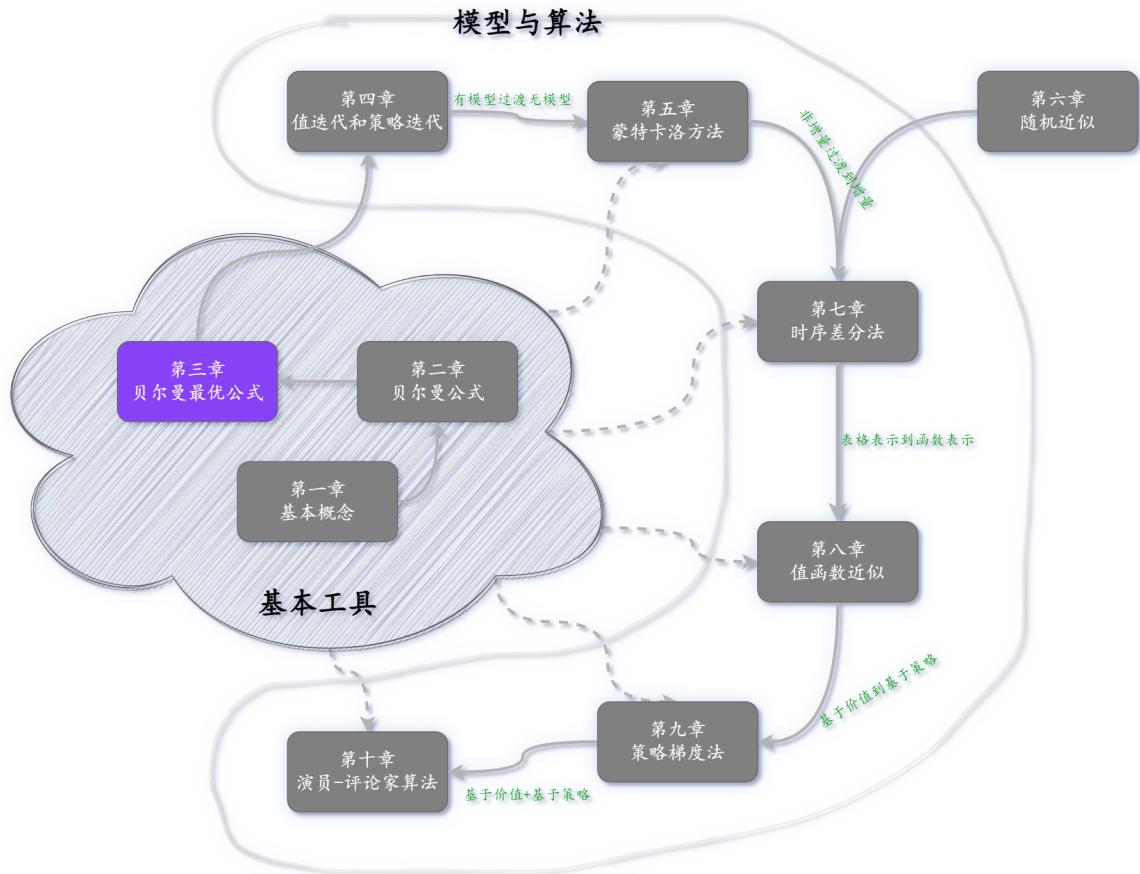


# 第三章 最优状态值和贝尔曼最优方程



强化学习的最终目标是寻求**最优策略（Optimal Policy）**。因此，有必要定义什么是最优策略。在本章中，我们介绍了一个核心概念和一个重要的工具。其核心概念是**最优状态值（Optimal State Values）**，在此基础上我们可以定义最优策略。其中重要的工具是**贝尔曼最优方程（Bellman Optimality Equation）**，从中我们可以求解最优状态值和策略。

前面、现在和以后各章之间的关系如下。前一章（第2章）介绍了任何给定策略的贝尔曼方程。本章介绍贝尔曼最优方程，它是一种特殊的贝尔曼方程，其对应的策略是最优的。下一章（第4章）将介绍一种重要的算法，称为**值迭代（Value Iteration）**，这正是本章所介绍的求解贝尔曼最优方程的算法。

请注意，本章的数学内容比较密集。然而，这是值得的，因为许多基本问题可以得到明确的回答。

## 3.1 范例：如何改进政策？

考虑图3.2所示的策略。在这里，橙色和蓝色的格子分别代表禁区和目标区域。这里的策略并不好，因为它在状态  $s_1$  时选择了  $a_2$ （向右）。我们如何改进现有的策略以获得更好的策略？答案在于状态值和动作值。

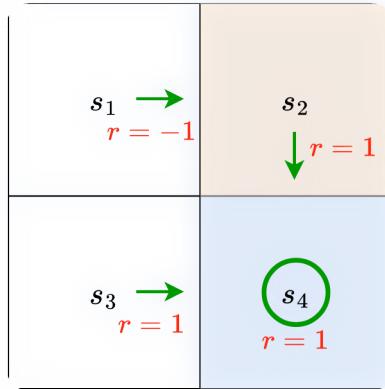


图 3.2

- 直觉：直觉上很清楚，如果在  $s_1$  点选择  $a_3$ （向下）而不是  $a_2$ （向右），策略可以得到改善。这是因为向下移动可以使智能体避免进入禁区。
- 数学：上述直觉可以通过状态值和动作值的计算来实现。

首先，我们计算给定策略的状态值。这个策略的贝尔曼方程是：

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

$$v_\pi(s_1) = +1 + \gamma v_\pi(s_4),$$

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

$$v_\pi(s_4) = +1 + \gamma v_\pi(s_4).$$

设  $\gamma = 0.9$ 。我们可以得到：

$$v_\pi(s_2) = v_\pi(s_3) = v_\pi(s_4) = 10,$$

$$v_\pi(s_1) = 8.$$

其次，我们计算状态  $s_1$  的动作值：

$$\begin{aligned} q_\pi(s_1, a_1) &= -1 + \gamma v_\pi(s_1) = 6.2, \\ q_\pi(s_1, a_2) &= -1 + \gamma v_\pi(s_2) = 8, \\ q_\pi(s_1, a_3) &= 0 + \gamma v_\pi(s_3) = 9, \\ q_\pi(s_1, a_4) &= -1 + \gamma v_\pi(s_1) = 6.2, \\ q_\pi(s_1, a_5) &= 0 + \gamma v_\pi(s_1) = 7.2. \end{aligned}$$

值得注意的是，动作  $a_3$  具有最大的动作值：

$$\forall_{i \neq 3}, q_\pi(s_1, a_3) > q_\pi(s_1, a_i)$$

因此，我们可以更新策略，即在状态  $s_1$  时，选择  $a_3$ 。

这个例子说明，如果我们选择具有最大动作值（Greatest Action Value）的动作去更新，我们可以获得更好的策略。这是许多强化学习算法的基本思想。

这个例子非常简单，因为给定的策略只对状态  $s_1$  不利。如果该策略对其他状态也不好，那么选择具有最大动作值的动作仍然会产生更好的策略吗？此外，是否总是存在最优策略？最优策略是什么样子的？我们将在本章中回答所有这些问题。

## 3.2 最优状态值和最优策略

虽然强化学习的最终目标是获得最优策略，但有必要首先去定义什么是最优策略。该定义基于状态值。特别是，考虑两个给定的策略  $\pi_1$  和  $\pi_2$ 。如果  $\pi_1$  的状态值大于或等于任何状态的  $\pi_2$  的状态值：

$$\forall s \in \mathcal{S}, v_{\pi_1}(s) \geq v_{\pi_2}(s),$$

那么说  $\pi_1$  比  $\pi_2$  好。此外，如果一个策略比所有其他可能的策略都好，那么这个策略就是最优的。这一点在下文中正式说明。

**定义 3.1（最优策略和最优状态值）：**

对于任意  $s \in \mathcal{S}$  和任何其他策略  $\pi$ ，如果  $v_{\pi^*}(s) \geq v_{\pi}(s)$ ，则策略  $\pi^*$  是最优的。 $\pi^*$  的状态值是**最优状态值**。

上述定义表明，与所有其他策略相比，最优策略对于每个状态都具有最大的状态值。这个定义也引出了许多问题：

- ① 存在性：最优策略是否存在？
- ② 唯一性：最优策略是唯一的吗？
- ③ 随机性：最优策略是随机的还是确定性的？
- ④ 算法：如何得到最优策略和最优状态值？

必须明确回答这些基本问题，才能彻底理解最优策略。例如，关于最优策略的存在，如果最优策略不存在，那么我们就不需要费力地设计算法来找到它们。我们将在本章的剩余部分回答所有这些问题。

## 3.3 贝尔曼最优方程

分析最优策略和最优状态值的工具是**贝尔曼最优方程**（Bellman Optimal Equation，简写 BOE）。通过求解这个方程，我们可以获得最优策略和最优状态值。接下来我们将介绍 BOE 的表达式，然后对其进行详细分析。

对于每个  $s \in \mathcal{S}$ ，BOE 的元素表达式为

$$\begin{aligned} v(s) &= \max_{\pi(s) \in \Pi(s)} \sum_{a \in \mathcal{A}} \pi(a | s) \left( \sum_{r \in \mathcal{R}} p(r | s, a)r + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a)v(s') \right) \\ &= \max_{\pi(s) \in \Pi(s)} \sum_{a \in \mathcal{A}} \pi(a | s) q(s, a), \end{aligned} \tag{3.1}$$

其中  $v(s)$ ,  $v(s')$  是待求解的未知变量，从而得到

$$q(s, a) \doteq \sum_{r \in \mathcal{R}} p(r | s, a)r + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a)v(s')$$

这里,  $\pi(s)$  表示状态  $s$  的一个策略, 而  $\Pi(s)$  是  $s$  的所有可能策略的集合。

BOE 是分析最优策略的一个优雅而强大的工具。然而, 理解这个方程式可能不是一件容易的事。例如, 这个方程有两个未知变量  $v(s)$  和  $\pi(a | s)$ 。对于初学者来说, 如何从一个方程中解出两个未知变量可能会令人困惑。此外, BOE 实际上是一个特殊的贝尔曼方程。然而, 看到这一点是不容易的, 因为它的表达式与贝尔曼方程的表达式非常不同。我们还需要回答以下有关 BOE 的基本问题。

- ✓ 存在性: 这个方程有解吗?
- ✓ 唯一性: 解是否唯一?
- ✓ 算法: 如何解这个方程?
- ✓ 最优性: 解与最优策略之间有什么关系?

一旦我们能够回答这些问题, 我们就会清楚地了解最优状态值和最优策略。

### 3.3.1 最大化 BOE 右侧

接下来, 我们将在(3.1)中阐明如何解决 BOE 右侧的最大化问题。乍一看, 如何从一个方程中解出两个未知变量  $v(s)$  和  $\pi(a | s)$ , 可能会让初学者感到困惑。事实上, 这两个未知变量是可以一个接着一个解决的。下面的例子说明了这个想法。

□ **示例3.1:** 考虑两个未知变量  $x, y \in \mathbb{R}$ , 满足

$$x = \max_{y \in \mathbb{R}} (2x - 1 - y^2)$$

第一步是求解方程右侧的  $y$ 。不管  $x$  的值是多少, 我们总是有  $\max_y (2x - 1 - y^2) = 2x - 1$ , 其中当  $y = 0$  时达到最大值。第二步是解  $x$ 。当  $y = 0$  时, 方程变成  $x = 2x - 1$ , 从而  $x = 1$ 。因此,  $y = 0$  和  $x = 1$  是方程的解。

我们现在转向解决 BOE 右侧的最大化问题。(3.1)中的 BOE 可以简明地写成

$$v(s) = \max_{\pi(s) \in \Pi(s)} \sum_{a \in \mathcal{A}} \pi(a | s) q(s, a), \quad s \in \mathcal{S}$$

受示例3.1的启发, 我们可以首先在右侧求解最优  $\pi$ 。如何做到这一点? 下面的示例演示了它的基本思想。

□

□ **示例3.2:** 给定  $q_1, q_2, q_3 \in \mathcal{R}$ , 我们希望求解下面的优化问题去找到  $c_1, c_2, c_3$  的最优解:

$$\begin{aligned} & \max_{c_1, c_2, c_3} \quad \sum_{i=1}^3 c_i q_i = c_1 q_1 + c_2 q_2 + c_3 q_3 \\ & \text{s.t.} \quad c_1 + c_2 + c_3 = 1 \\ & \quad c_1, c_2, c_3 \geq 0 \end{aligned}$$

在不失一般性的情况下, 假设  $q_3 \geq q_1, q_2$ 。那么, 最优解是  $c_3^* = 1, c_1^* = c_2^* = 0$ 。这是因为

$$\forall_{a_1, a_2, a_3}, \quad q_3 = (c_1 + c_2 + c_3) q_3 = c_1 q_3 + c_2 q_3 + c_3 q_3 \geq c_1 q_1 + c_2 q_2 + c_3 q_3$$

受上述例子的启发, 由于  $\sum_a \pi(a | s) = 1$ , 我们得到

$$\sum_{a \in \mathcal{A}} \pi(a | s) q(s, a) \leq \sum_{a \in \mathcal{A}} \pi(a | s) \left( \max_{a \in \mathcal{A}} q(s, a) \right) = \max_{a \in \mathcal{A}} q(s, a),$$

在以下情况，方程取等：

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

这里， $a^* = \underset{a}{\operatorname{argmax}} q(s, a)$ 。总之，最优策略  $\pi(s)$  就是选择  $q(s, a)$  值最大的动作。□

### 3.3.2 BOE 的矩阵 - 向量形式

BOE 指的是为所有状态定义的一组方程。如果我们将这些方程组合在一起，我们可以得到一种简洁的矩阵 - 向量形式，这将在本章中广泛使用。

BOE 的矩阵向量形式是

$$v = \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v), \quad (3.2)$$

其中  $v \in \mathbb{R}^{|\mathcal{S}|}$  和  $\max_{\pi}$  以元素方式执行。 $r_\pi$  和  $P_\pi$  的结构与贝尔曼方程的矩阵向量形式相同：

$$\begin{aligned} [r_\pi]_s &\doteq \sum_{a \in \mathcal{A}} \pi(a \mid s) \sum_{r \in \mathcal{R}} p(r \mid s, a) r, \\ [P_\pi]_{s,s'} &= p(s' \mid s) \doteq \sum_{a \in \mathcal{A}} \pi(a \mid s) p(s' \mid s, a). \end{aligned}$$

由于  $\pi$  的最优值由  $v$  决定，因此 (3.2) 的右侧是  $v$  的函数，表示为：

$$f(v) \doteq \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v).$$

那么，BOE 可以用简洁的形式表示为：

$$v = f(v). \quad (3.3)$$

在本节的剩余部分中，我们将展示如何求解这个非线性方程。

### 3.3.3 收缩映射定理

由于 BOE 可以表示为非线性方程  $v = f(v)$ ，因此我们接下来引入 **收缩映射定理** (Contraction Mapping Theorem)<sup>[1]</sup> 对其进行分析。收缩映射定理是分析一般非线性方程的有力工具。它也称为**不动点定理**。已经了解该定理的读者可以跳过这一部分。否则，建议读者熟悉该定理，因为它是分析 BOE 的关键。

考虑一个函数  $f(x)$ ，其中  $x \in \mathbb{R}^d$  且  $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ 。如果满足

$$f(x^*) = x^*,$$

则称  $x^*$  是**不动点** (Fixed Point)。

也就是说  $x^*$  的映射是它本身。这就是  $x^*$  被称为“不动”的原因。如果对于任何  $x_1, x_2 \in \mathbb{R}^d$ ，存在  $\gamma \in (0, 1)$ ，使得

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

则映射  $f$  是收缩映射（或收缩函数）。在本书中， $\|\cdot\|$  表示向量或矩阵范数。

□示例3.3：我们提供三个示例来演示不动点和收缩映射。

- $x = f(x) = 0.5x, \quad x \in \mathbb{R}$

很容易验证  $x = 0$  是不动点，因为  $0 = 0.5 \times 0$ 。此外， $f(x) = 0.5x$  是收缩映射，因为  $\|0.5x_1 - 0.5x_2\| = 0.5 \|x_1 - x_2\| \leq \gamma \|x_1 - x_2\|$ ，对于任意  $\gamma \in [0.5, 1)$ 。

- $x = f(x) = Ax, \quad \text{其中 } x \in \mathbb{R}^n, \quad A \in \mathbb{R}^{n \times n}, \quad \text{并且 } \|A\| \leq \gamma < 1$

很容 易 验 证  $x = 0$  是 不 动 点 ， 此 外  
 $\|Ax_1 - Ax_2\| = \|A(x_1 - x_2)\| \leq \|A\| \|x_1 - x_2\| \leq \gamma \|x_1 - x_2\|$ 。因此， $f(x) = Ax$  是收缩映射。

- $x = f(x) = 0.5 \sin x, \quad x \in \mathbb{R}$

很容易看出  $x = 0$  是一个不动点。此外，根据中值定理<sup>2</sup> [^3] 可以得出：

$$\left| \frac{\sin x_1 - \sin x_2}{x_1 - x_2} \right| = |0.5 \cos x_3| \leq 0.5, \quad x_3 \in (x_1, x_2)$$

因此  $|0.5 \sin x_1 - 0.5 \sin x_2| \leq 0.5 |x_1 - x_2|$ ，因此  $f(x) = 0.5 \sin x$  是收缩映射。□

不动点与收缩性质之间的关系由以下经典定理来表征。

### 定理 3.1 (收缩映射定理)

对于任何具有  $x = f(x)$  形式的方程，其中  $x$  和  $f(x)$  是实向量，如果  $f$  是收缩映射，则以下属性成立：

- ① 存在性：存在满足  $f(x^*) = x^*$  的不动点  $x^*$ 。
- ② 唯一性：不动点  $x^*$  是唯一的。
- ③ 算法：考虑迭代过程

$$x_{k+1} = f(x_k),$$

其中  $k = 0, 1, 2, \dots$ ，然后，对于任何初始猜测  $x_0$ ，当  $k \rightarrow \infty$  时， $x_k \rightarrow x^*$ 。而且，收敛速度呈指数级。

收缩映射定理不仅可以判断非线性方程的解是否存在，而且还提出了求解该方程的数值算法。该定理的证明在框 3.1 中给出。

以下示例演示如何使用收缩映射定理得到的迭代算法来计算方程的不动点。

□ 例 3.4。让我们回顾一下上面提到的例子： $x = 0.5x$ 、 $x = Ax$  和  $x = 0.5 \sin x$ 。虽然已经证明这三个方程的右边都是收缩映射，根据收缩映射定理，它们都有一个唯一的不动点，可以很容易地验证为  $x^* = 0$ 。此外，三个方程的不动点可以通过以下算法迭代求解：

$$x_{k+1} = 0.5x_k,$$

$$x_{k+1} = Ax_k,$$

$$x_{k+1} = 0.5 \sin x_k,$$

对于任意给定的猜测初始值  $x_0$ 。□

### [证明 3.1] 收缩映射定理的证明

- 第 1 部分：我们证明  $\{x_k\}_{k=1}^{\infty}$ ,  $x_k = f(x_{k-1})$  是收敛的。

证明依赖于柯西序列 (Cauchy Sequence)

柯西序列：

$$\forall \varepsilon > 0, \exists N \in \mathbb{N}_+, \text{当 } m, n > N \text{ 时}, \|x_m - x_n\| < \varepsilon$$

直观的解释是，存在一个有限整数  $N$ ，使得  $N$  之后的所有元素彼此足够接近。柯西序列很重要，因为它保证柯西序列收敛。其收敛性将用于证明收缩映射定理。请注意，对于所有  $m, n > N$ ，我们必须有  $\|x_m - x_n\| < \varepsilon$ 。如果我们简单地认为  $x_{n+1} - x_n \rightarrow 0$ ，则不足以声称该序列是柯西序列。例如，对于  $x_n = \sqrt{n}$ ，它  $x_{n+1} - x_n \rightarrow 0$ ，但显然  $x_n = \sqrt{n}$  发散。

接下来我们证明  $\{x_k = f(x_{k-1})\}_{k=1}^{\infty}$  是柯西序列，因此收敛。

- 首先，由于  $f$  是收缩映射，所以我们有

$$\|x_{k+1} - x_k\| = \|f(x_k) - f(x_{k-1})\| \leq \gamma \|x_k - x_{k-1}\|,$$

以此类推，我们有

$$\begin{aligned} \|x_{k+1} - x_k\| &\leq \gamma \|x_k - x_{k-1}\|, \\ &\leq \gamma^2 \|x_{k-1} - x_{k-2}\|, \\ &\vdots \\ &\leq \gamma^k \|x_1 - x_0\|. \end{aligned}$$

由于  $\gamma < 1$ ，我们知道在给定任意  $x_1, x_0$  的情况下，当  $k \rightarrow \infty$  时， $\|x_{k+1} - x_k\|$  以指数方式快速收敛到零。值得注意的是， $\{\|x_{k+1} - x_k\|\}$  的收敛不足以暗示  $\{x_k\}$  的收敛。因此，对于任何  $m > n$ ，我们需要进一步考虑  $\|x_m - x_n\|$ 。尤其，

$$\begin{aligned} \|x_m - x_n\| &\leq \|x_m - x_{m-1} + x_{m-1} - \cdots - x_{n+1} + x_{n+1} - x_n\| \\ &\leq \|x_m - x_{m-1}\| + \cdots + \|x_{n+1} - x_n\| \\ &\leq \gamma^{m-1} \|x_1 - x_0\| + \cdots + \gamma^n \|x_1 - x_0\| \\ &= \gamma^n (\gamma^{m-1-n} + \cdots + 1) \|x_1 - x_0\| \\ &\leq \gamma^n (1 + \cdots + \gamma^{m-1-n} + \gamma^{m-n} + \gamma^{m-n+1} + \cdots) \|x_1 - x_0\| \\ &= \frac{\gamma^n}{1 - \gamma} \|x_1 - x_0\| \end{aligned} \tag{3.4}$$

因此，对于任何  $\varepsilon$ ，我们总能找到  $N$ ，使得  $\|x_m - x_n\| < \varepsilon$  对于所有  $m, n > N$ 。因此，该序列是柯西序列，因此极限表示为  $x^* = \lim_{k \rightarrow \infty} x_k$ 。

- 第 2 部分：我们证明极限  $x^* = \lim_{k \rightarrow \infty} x_k$  是一个不动点。为了做到这一点，由于

$$\| f(x_k) - x_k \| = \| x_{k+1} - x_k \| \leq \gamma^k \| x_1 - x_0 \|,$$

我们知道  $\| f(x_k) - x_k \|$  以指数方式快速收敛到零。因此，我们有  $f(x^*) = x^*$  的极限。

- 第 3 部分：我们证明不动点是唯一的。假设还有另一个不动点  $x'$  满足  $f(x') = x'$ 。然后，

$$\| x' - x^* \| = \| f(x') - f(x^*) \| \leq \gamma \| x' - x^* \|.$$

由于  $\gamma < 1$ ，当且仅当  $\| x' - x^* \| = 0$  时，该不等式成立。因此， $x' = x^*$ 。

- 第 4 部分：我们证明  $x_k$  以指数方式快速收敛到  $x^*$ 。回想一下  $\| x_m - x_n \| \leq \frac{\gamma^n}{1-\gamma} \| x_1 - x_0 \|$ 。由于  $m$  可以任意大，因此我们有：

$$\| x^* - x_n \| = \lim_{m \rightarrow \infty} \| x_m - x_n \| \leq \frac{\gamma^n}{1-\gamma} \| x_1 - x_0 \|.$$

由于  $\gamma < 1$ ，随着  $n \rightarrow \infty$ ，误差以指数方式快速收敛到零。

### 3.3.4 BOE 右侧的收缩特性

接下来我们证明(3.3)中的 BOE 中的  $f(v)$  是收缩映射。因此，可以应用上一小节中介绍的收缩映射定理。

#### 定理 3.2 ( $f(v)$ 的收缩性质)

(3.3) 中 BOE 右侧的函数  $f(v)$  是收缩映射。特别地，对于任意  $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$ ，有

$$\| f(v_1) - f(v_2) \|_\infty \leq \gamma \| v_1 - v_2 \|_\infty,$$

其中  $\gamma \in (0, 1)$  是折扣因子， $\| \cdot \|_\infty$  是最大范数，即向量元素的最大绝对值。

该定理的证明在框 3.2 中给出。这个定理很重要，因为我们可以使用强大的收缩映射定理来分析 BOE。

#### [3.2]：定理3.2的证明

考 虑 任 意 两 个 向 量  $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$ ， 并 假 设  $\pi_1^* \doteq \underset{\pi}{\text{argmax}}(r_\pi + \gamma P_\pi v_1)$  和  $\pi_2^* \doteq \underset{\pi}{\text{argmax}}(r_\pi + \gamma P_\pi v_2)$ 。然 后，

$$f(v_1) = \max_{\pi} (r_\pi + \gamma P_\pi v_1) = r_{\pi_1^*} + \gamma P_{\pi_1^*} v_1 \geq r_{\pi_2^*} + \gamma P_{\pi_2^*} v_1$$

$$f(v_2) = \max_{\pi} (r_\pi + \gamma P_\pi v_2) = r_{\pi_2^*} + \gamma P_{\pi_2^*} v_2 \geq r_{\pi_1^*} + \gamma P_{\pi_1^*} v_2$$

其中  $\geq$  是元素比较。因此，

$$\begin{aligned}
f(v_1) - f(v_2) &= r_{\pi_1^*} + \gamma P_{\pi_1^*} v_1 - (r_{\pi_2^*} + \gamma P_{\pi_2^*} v_2), \\
&\leq r_{\pi_1^*} + \gamma P_{\pi_1^*} v_1 - (r_{\pi_1^*} + \gamma P_{\pi_1^*} v_2), \\
&= \gamma P_{\pi_1^*} (v_1 - v_2).
\end{aligned}$$

类似地，可以证明  $f(v_2) - f(v_1) \leq \gamma P_{\pi_2^*} (v_2 - v_1)$ 。所以，

$$\gamma P_{\pi_2^*} (v_1 - v_2) \leq f(v_1) - f(v_2) \leq \gamma P_{\pi_1^*} (v_1 - v_2).$$

## 定义

$$z \doteq \max \{ |\gamma P_{\pi_2^*} (v_1 - v_2)|, |\gamma P_{\pi_1^*} (v_1 - v_2)| \} \in \mathbb{R}^{|\mathcal{S}|}$$

其中  $\max(\cdot)$ ,  $|\cdot|$  和  $\geq$  都是元素运算符。根据定义,  $z \geq 0$ 。一方面, 我们很容易看出

$$-z \leq \gamma P_{\pi_2^*} (v_1 - v_2) \leq f(v_1) - f(v_2) \leq \gamma P_{\pi_1^*} (v_1 - v_2) \leq z,$$

这意味着

$$|f(v_1) - f(v_2)| \leq z.$$

那么接下来

$$\|f(v_1) - f(v_2)\|_\infty \leq \|z\|_\infty \quad (3.5)$$

其中  $\|\cdot\|_\infty$  是最大范数。

另一方面, 假设  $z_i$  是  $z$  的第  $i$  个条目,  $p_i^T$  和  $q_i^T$  分别是  $P_{\pi_1^*}$  和  $P_{\pi_2^*}$  的第  $i$  行。然后,

$$z_i = \max \{ \gamma |p_i^T (v_1 - v_2)|, \gamma |q_i^T (v_1 - v_2)| \}$$

由于  $p_i$  是一个包含所有非负元素的向量, 并且元素之和等于 1, 因此可以得出:

$$|p_i^T (v_1 - v_2)| \leq p_i^T |v_1 - v_2| \leq \|v_1 - v_2\|_\infty$$

类似地, 我们有  $q_i^T |v_1 - v_2| \leq \|v_1 - v_2\|_\infty$ 。因此,  $z_i \leq \gamma \|v_1 - v_2\|_\infty$ , 因此

$$\|z\|_\infty = \max_i |z_i| \leq \gamma \|v_1 - v_2\|_\infty$$

将这个不等式代入 (3.5) 得出

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

由此得到  $f(v)$  收缩性质的证明。

## 3.4 求解 BOE 的最优策略

有了上一节的准备, 我们就可以求解 BOE 以获得最优状态值  $v^*$  和最优策略  $\pi^*$ 。

- 求解  $v^*$ : 如果  $v^*$  是 BOE 的解, 那么它满足

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

显然,  $v^*$  是不动点, 因为  $v^* = f(v^*)$ 。然后, 收缩映射定理提出了以下结果。

### 定理 3.3: 存在性、唯一性、算法

对于 BOE  $v = f(v) = \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v)$ , 总是存在唯一解  $v^*$ , 可以通过迭代法求解

$$v_{k+1} = f(v_k) = \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v_k), \quad k = 0, 1, 2, \dots$$

给定任何初始猜测  $v_0$ , 当  $k \rightarrow \infty$  时,  $v_k$  的值以指数方式快速收敛到  $v^*$ 。

因为  $f(v)$  是收缩映射, 该定理的证明可以直接从收缩映射定理得出。这个定理很重要, 因为它回答了一些基本问题。

- ✓  $v^*$  的存在性: BOE 的解始终存在。
- ✓  $v^*$  的唯一性: 解  $v^*$  始终是唯一的。
- ✓ 求解  $v^*$  的算法:  $v^*$  的值可以通过定理 3.3 的迭代算法来求解。这种迭代算法叫做**值迭代** (Value Iteration)。
- 求解  $\pi^*$ : 一旦得到  $v^*$  的值, 我们可以通过求解轻松获得  $\pi^*$

$$\pi^* = \operatorname{argmax}_{\pi \in \Pi} (r_\pi + \gamma P_\pi v^*). \quad (3.6)$$

$\pi^*$  的值将在定理 3.5 中给出。将(3.6)代入 BOE

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

因此,  $v^* = v_{\pi^*}$  就是  $\pi^*$  的状态值, 而 BOE 是一个特殊的贝尔曼方程, 其对应的策略是  $\pi^*$ 。

此时, 虽然我们可以求解  $v^*$  和  $\pi^*$ , 但仍不清楚解是否是最优的。以下定理揭示了解的最优性。

### 定理 3.4 ( $v^*$ 和 $\pi^*$ 的最优性)

解  $v^*$  是最优状态值,  $\pi^*$  是最优策略。也就是说, 对于任何策略  $\pi$ , 它认为

$$v^* = v_{\pi^*} \geq v_\pi,$$

其中  $v_\pi$  是  $\pi$  的状态值,  $\geq$  是元素比较。

现在, 我们为什么必须研究 BOE 就很清楚了: 它的解对应于最优状态值和最优策略。上述定理的证明在下面的框中给出。

### [3.3] 定理 3.4 的证明

对于任何策略  $\pi$ , 它认为

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

由于

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

我们有

$$v^* - v^{\pi} \geq (r_{\pi} + \gamma P_{\pi} v^*) - (r_{\pi} + \gamma P_{\pi} v_{\pi}) = \gamma P_{\pi} (v^* - v_{\pi}).$$

重复应用上述不等式可得，我们可以得到

$$v^* - v_{\pi} \geq \lim_{n \rightarrow \infty} \gamma^n P_{\pi}^n (v^* - v_{\pi}) = 0,$$

其中最后一个等式是成立的，因为  $\gamma < 1$  并且  $P_{\pi}^n$  是一个非负矩阵，其所有元素都小于或等于 1（因为  $P_{\pi}^n \mathbf{1} = \mathbf{1}$ ）。因此，对于任何  $\pi$ ,  $v^* \geq v_{\pi \circ}$

接下来我们更仔细地研究 (3.6) 中的  $\pi^*$ 。特别是，以下定理表明始终存在最优的确定性贪婪策略。

### 定理 3.5 (贪婪最优策略)

对于任意  $s \in \mathcal{S}$ , 确定性贪心策略 (Deterministic Greedy Policy)

$$\pi^*(a|s) = \begin{cases} 1, & a = a^*(s), \\ 0, & a \neq a^*(s), \end{cases} \quad (3.7)$$

是解决 BOE 的最优策略。这里

$$a^*(s) = \operatorname{argmax}_a q^*(a, s),$$

其中

$$q^*(s, a) \doteq \sum_{r \in \mathcal{R}} p(r|s, a)r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v^*(s').$$

### [3.4] 定理 3.5 的证明

虽然最优策略的矩阵 - 向量形式为  $\pi^* = \operatorname{argmax}_{\pi} (r_{\pi} + \gamma P_{\pi} v^*)$ , 但其元素形式为

$$\pi^*(s) = \operatorname{argmax}_{\pi \in \Pi} \sum_{a \in \mathcal{A}} \pi(a|s) \underbrace{\left( \sum_{r \in \mathcal{R}} p(r|s, a)r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v^*(s') \right)}_{q^*(s, a)}, \quad s \in \mathcal{S}$$

很明显，如果  $\pi(s)$  选择具有最大  $q^*(s, a)$  的动作，则  $\sum_{a \in \mathcal{A}} \pi(a|s)q^*(s, a)$  最大化。

(3.7) 中的策略被称为贪婪 (Greedy)，因为它寻求具有最大  $q^*(s, a)$  的动作。最后，我们讨论  $\pi^*$  的两个重要属性。

- 最优策略的唯一性:

- 虽然  $v^*$  的值是唯一的, 但  $v^*$  对应的最优策略可能不是唯一的。
- 这一点通过反例很容易地验证。例如, 图3.3所示的两个策略都是最优的。

- 最优策略的随机性:

- 最优策略可以是随机的, 也可以是确定性的, 如图 3.3 所示。
- 根据定理 3.5, 可以肯定总是存在确定性最优策略。

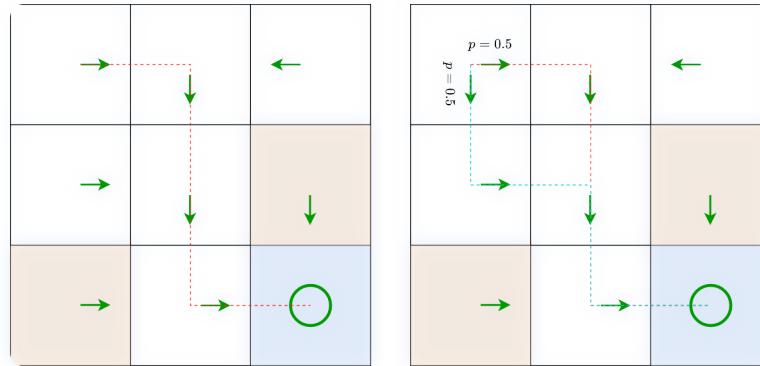


图 3.3

## 3.5 影响最优策略的因素

BOE 是分析最优策略的强大工具。接下来我们用 BOE 来研究哪些因素可以影响最优策略。通过观察 BOE 元素表达式可以很容易回答这个问题:

$$v(s) = \max_{\pi(s) \in \Pi(s)} \sum_{a \in \mathcal{A}} \pi(a|s) \left( \sum_{r \in \mathcal{R}} p(r|s, a)r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v(s') \right), \quad s \in \mathcal{S}$$

最优状态值和最优策略由以下参数确定:

1. 即时奖励  $r$
2. 折扣因子  $\gamma$
3. 系统模型  $p(s'|s, a)$ ,  $p(r|s, a)$

虽然系统模型是固定的, 但我们接下来讨论当我们改变  $r$  和  $\gamma$  的值时最优策略如何变化。本节中提出的所有最优策略都可以通过定理 3.3 中的算法获得。该算法的实现细节将在第4章中给出。本章主要关注最优策略的基本属性。

### - 基线示例

考虑图 3.4 中的示例, 奖励设置为  $r_{\text{boundary}} = r_{\text{forbidden}} = -1$  和  $r_{\text{target}} = 1$ 。此外, 智能体对于每个移动步骤都会收到  $r_{\text{other}} = 0$  的奖励。折扣因子设置为  $\gamma = 0.9$ 。

有了上述参数, 最优策略和最优状态值如图 3.4(a) 所示。有趣的是, 智能体并不害怕穿过禁区到达目标区域。更具体地说, 从 (行=4, 列=1) 的状态开始, 智能体有两种到达目标区域的选择。第一个选择是避开所有禁区并长途跋涉到达目标区域。第二种选择是穿过禁区。虽然智能体在进入禁区时获得负奖励, 但第二条轨迹的累积奖励大于第一条轨迹。因此, 由于  $\gamma$  值较大, 最优策略是有远见的。

	1	2	3	4	5
1	↓	→	↓	↓	↓
2	↓	↓	↓	↓	↓
3	→	→	↓	↓	↓
4	→	→	○	←	←
5	↓	→	↑	←	←

	1	2	3	4	5
1	5.8	5.6	6.2	6.5	5.8
2	6.5	7.2	8.0	7.2	6.5
3	7.2	8.0	10.0	8.0	7.2
4	8.0	10.0	10.0	10.0	8.0
5	7.2	9.0	10.0	9.0	8.1

图 3.4(a) 基线示例:  $r_{\text{boundary}} = r_{\text{forbidden}} = -1$ ,  $r_{\text{target}} = 1$ ,  $\gamma = 0.9$ 

## - 折扣因子的影响

如果我们将折扣因子从  $\gamma = 0.9$  更改为  $\gamma = 0.5$  并保持其他参数不变，则最优策略变为图 3.4 (b) 所示的策略。有趣的是，智能体现在不敢再冒险了。相反，它会行驶很长的距离到达目标，同时避开所有禁区。这是因为由于  $\gamma$  值相对较小，最优策略变得短视。

	1	2	3	4	5
1	→	→	→	→	↓
2	↑	↑	→	→	↓
3	↑	←	↓	→	↓
4	↑	→	○	←	↓
5	↑	→	↑	←	←

	1	2	3	4	5
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.1
3	0.0	0.0	2.0	0.1	0.1
4	0.0	2.0	2.0	2.0	0.2
5	0.0	1.0	2.0	1.0	0.5

图 3.4(b) 折扣因子改为  $\gamma = 0.5$ ，其他参数与(a)中的相同。

在  $\gamma = 0$  的极端情况下，相应的最优策略如图 3.4(c) 所示。在这种情况下，智能体无法到达目标区域。这是因为每个状态的最优策略都是极其短视的，仅仅选择即时奖励最大的动作，而不是总奖励最大的动作。

	1	2	3	4	5	
1	↓	○	←	↓	↓	0.0
2	↑		↑	○	↑	0.0
3	○	○	↓	○	↓	0.0
4	↓	→	○	←	↑	1.0
5	↑	→	↑	←	←	0.0

	1	2	3	4	5	
1	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	1.0	0.0	0.0	0.0
4	0.0	1.0	1.0	1.0	0.0	0.0
5	0.0	0.0	1.0	0.0	0.0	0.0

图 3.4(c) 折扣因子改为  $\gamma = 0$ , 其他参数与(a)相同

此外, 状态值的空间分布呈现出一种有趣的模式: 靠近目标的状态具有较大的状态值, 而远离目标的状态具有较低的值。从图 3.4 所示的所有示例中都可以观察到这种模式。可以用折扣因子来解释: 如果一个状态必须沿着更长轨迹才能到达目标, 那么它的状态值会更小。

## - 奖励值的影响

如果我们想严格禁止智能体进入任何禁区, 我们可以加大对这类行为的惩罚。例如, 如果  $r_{\text{forbidden}}$  从 -1 变为 -10, 则得到的最优策略可以避开所有禁止区域 (见图 3.4 (d) ) 。

	1	2	3	4	5	
1	→	→	→	→	↓	3.5
2	↑	↑	→	→	↓	3.1
3	↑	←	↓	→	↓	2.8
4	↑	→	○	←	↓	2.5
5	↑	→	↑	←	←	2.3

图 3.4(d)  $r_{\text{forbidden}}$  从 -1 更改为 -10。其他参数与(a)中的相同。

然而, 改变奖励并不总是会导致不同的最优策略。一个重要的事实是, 最优策略对于奖励的仿射变换是不变的。换句话说, 如果我们缩放所有奖励或为所有奖励添加相同的值, 则最优策略保持不变。

### 定理 3.6 (最优策略不变性)

考虑一个马尔可夫决策过程，其中  $v^* \in \mathbb{R}^{|\mathcal{S}|}$  作为满足  $v^* = \max_{\pi \in \Pi}(r_\pi + \gamma P_\pi v^*)$  的最优状态值。如果每个奖励  $r \in \mathbb{R}$  通过仿射变换变为  $\alpha r + \beta$ ，其中  $\alpha, \beta \in \mathbb{R}$  且  $\alpha > 0$ ，则相应的最优状态值  $v'$  也是  $v^*$  的仿射变换：

$$v' = \alpha v^* + \frac{\beta}{1 - \gamma} \mathbf{1}, \quad (3.8)$$

其中  $\gamma \in (0, 1)$  是折扣因子， $\mathbf{1} = [1, \dots, 1]^T$ 。因此，从  $v'$  导出的最优策略对于奖励值的仿射变换是不变的。

### [3.5] 定理 3.6 的证明

对于任何策略  $\pi$ ，定义  $r\pi = [\dots, r_\pi(s), \dots]^T$ ，其中

$$r_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{r \in \mathcal{R}} p(r|s, a)r, \quad s \in \mathcal{S}.$$

如果  $r \rightarrow \alpha r + \beta$ ，则  $r_\pi(s) \rightarrow \alpha r_\pi(s) + \beta$ ，因此  $r_\pi \rightarrow \alpha r_\pi + \beta \mathbf{1}$ ，其中  $\mathbf{1} = [1, \dots, 1]^T$ 。在这种情况下，BOE 就变成

$$v' = \max_{\pi \in \Pi} (\alpha r_\pi + \beta \mathbf{1} + \gamma P_\pi v'). \quad (3.9)$$

接下来，我们通过证明  $v' = \alpha v^* + c \mathbf{1}$  且  $c = \frac{\beta}{1 - \gamma}$  是 (3.9) 的解来求解 (3.9) 中的新 BOE。特别地，将  $v' = \alpha v^* + c \mathbf{1}$  代入 (3.9) 给出

$$\alpha v^* + c \mathbf{1} = \max_{\pi \in \Pi} (\alpha r_\pi + \beta \mathbf{1} + \gamma P_\pi(\alpha v^* + c \mathbf{1})) = \max_{\pi \in \Pi} (\alpha r_\pi + \beta \mathbf{1} + \alpha \gamma P_\pi v^* + c \gamma \mathbf{1}),$$

其中最后一个等式是由于  $P_\pi \mathbf{1} = \mathbf{1}$ 。上式可以重新组织为

$$\alpha v^* = \max_{\pi \in \Pi} (\alpha r_\pi + \alpha \gamma P_\pi v^*) + \beta \mathbf{1} + c \gamma \mathbf{1} - c \mathbf{1},$$

这相当于

$$\beta \mathbf{1} + c \gamma \mathbf{1} - c \mathbf{1} = 0.$$

由于  $c = \frac{\beta}{1 - \gamma}$ ，上式有效，因此  $v' = \alpha v^* + c \mathbf{1}$  是 (3.9) 的解。由于 (3.9) 是 BOE，所以  $v'$  也是唯一解。最后，由于  $v'$  是  $v^*$  的仿射变换，因此动作值之间的相对关系保持不变。因此，从  $v'$  导出的贪婪最优策略与从  $v^*$  导出的贪婪最优策略相同： $\text{argmax}_{\pi \in \Pi}(r_\pi + \gamma P_\pi v')$  与  $\text{argmax}_\pi(r_\pi + \gamma P_\pi v^*)$  相同。

读者可以参考 [2](#) 进一步讨论在什么条件下修改奖励值可以保持最优策略。

## - 避免走无意义的弯路

在奖励设置中，智能体每移动一步都会收到  $r_{\text{other}} = 0$  的奖励（除非进入禁区或目标区域或试图超出边界）。既然零是奖励不是惩罚，那么最优策略会不会在达到目标之前走一些无意义的弯路呢？我们是否应该将  $r_{\text{other}}$  设置为负数以鼓励智能体尽快达到目标？

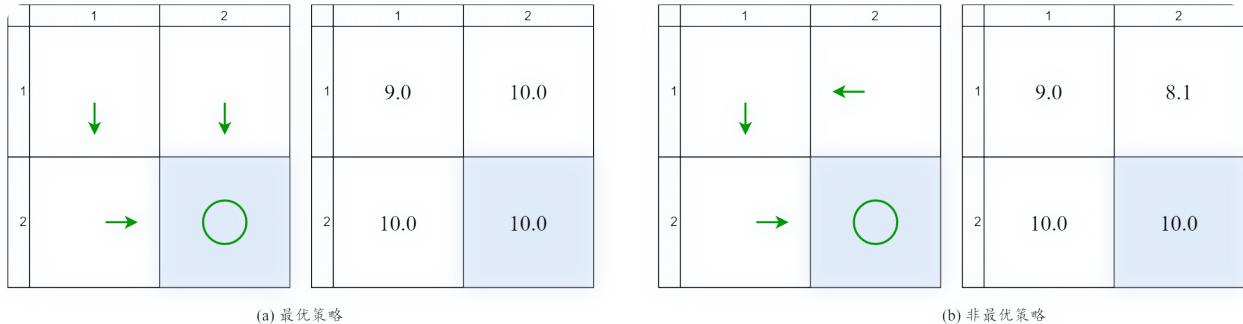


图 3.5

考虑图 3.5 中的示例，其中右下角的单元格是要到达的目标区域。这里的两个策略除了状态  $s_2$  之外都是相同的。根据图 3.5(a) 中的策略，智能体在  $s_2$  处向下移动，最终的轨迹为  $s_2 \rightarrow s_4$ 。根据图 3.5(b) 中的策略，智能体向左移动，得到的轨迹为  $s_2 \rightarrow s_1 \rightarrow s_3 \rightarrow s_4$ 。

值得注意的是，第二项策略在到达目标区域之前走了一些弯路。如果我们只考虑即时奖励，走这条弯路并没有什么关系，因为不会获得负面的即时奖励。然而，如果我们考虑到折扣回报，那么这个弯路就很重要了。特别是，对于第一个策略，折扣回报为

$$\text{return} = 1 + \gamma 1 + \gamma^2 1 + \dots = \frac{1}{1 - \gamma} = 10.$$

作为比较，第二个策略的折扣回报为

$$\text{return} = 0 + \gamma 0 + \gamma^2 1 + \dots = \frac{\gamma^2}{1 - \gamma} = 8.1.$$

显然，轨迹越短，回报越大。因此，虽然每一步的即时奖励并不能鼓励智能体尽快接近目标，但折扣因子确实鼓励它这样做。

初学者可能存在一个误解是，在每个动作获得的奖励之上添加负奖励（例如 -1）对于鼓励智能体尽快达到目标是必要的。这是一个误解，因为在所有奖励之上添加相同的奖励是一种仿射变换，它保留了最优策略。此外，最优策略不会因为折扣因子而走无意义的弯路，尽管弯路可能不会立即获得任何负面奖励。

1. H. K. Khalil, Nonlinear systems (3rd Edition). Patience Hall, 2002. ↪

2. A. Y. Ng, D. Harada, and S. Russell, “Policy invariance under reward transformations: Theory and application to reward shaping,” in International Conference on Machine Learning, vol. 99, pp. 278–287, 1999. ↪ ↪