

第3章 最优状态价值与贝尔曼最优方程

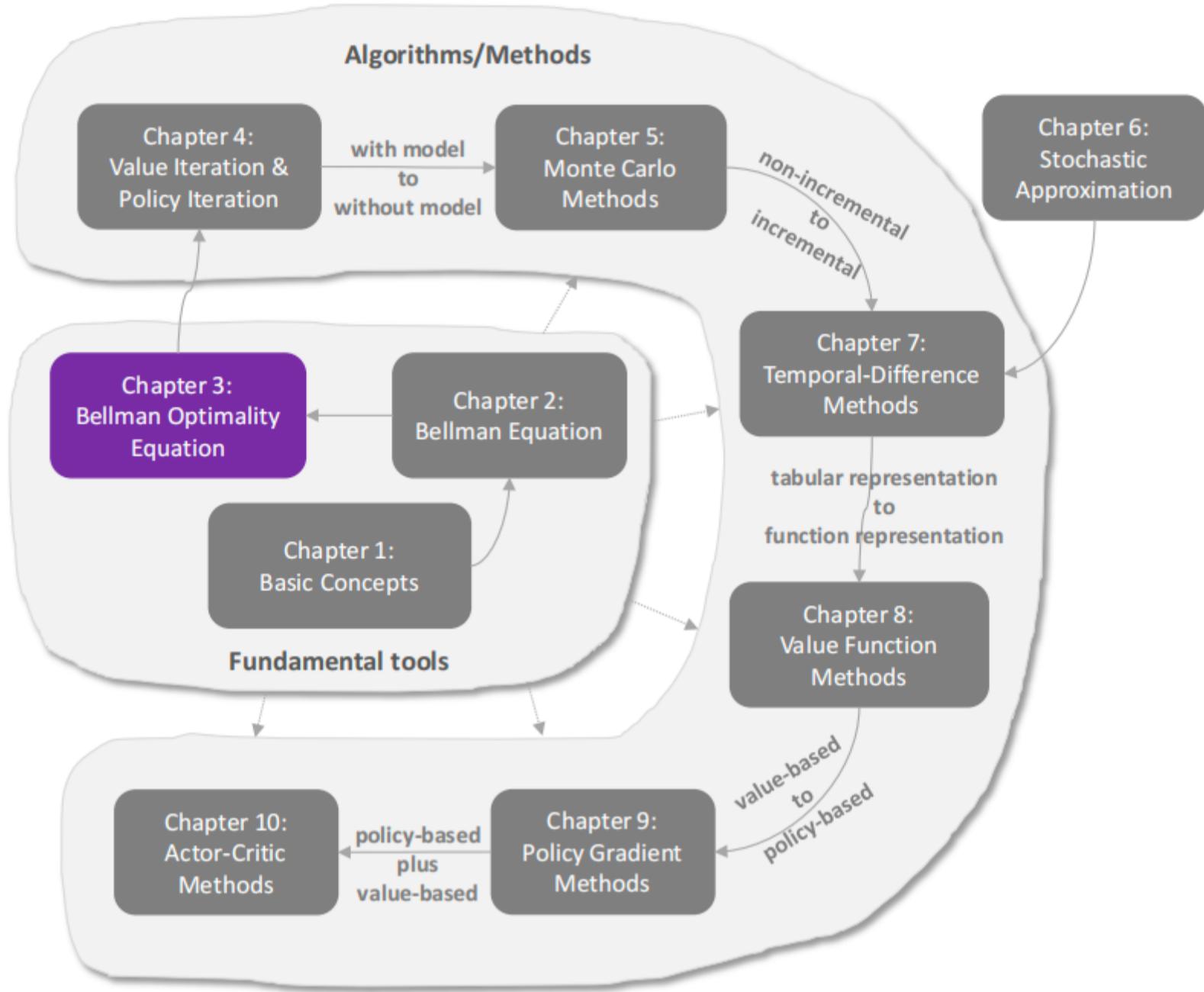


Figure 3.1: Where we are in this book.

图 3.1：我们在本书中的位置。

强化学习的最终目标是寻找**最优策略 (optimal policies)**。因此，有必要定义什么是**最优策略 (optimal policies)**。在本章中，我们介绍一个核心概念和一个重要工具。核心概念是

- **最优状态价值 (optimal state value)**，基于此我们可以定义**最优策略 (optimal policies)**。
- 重要工具是**贝尔曼最优方程 (Bellman optimality equation)**，通过它我们可以求解最优状态价值和策略。

前一章、本章和后一章之间的关系如下。前一章（第 2 章）介绍了任意给定策略的贝尔曼方程。

本章介绍了**贝尔曼最优方程 (Bellman optimality equation)**，这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）将介绍一种称为**价值迭代 (value iteration)** 的重要算法，正如本章所介绍的那样，它正是用于求解**贝尔曼最优方程**的算法。

请做好准备，本章的数学推导可能会稍微密集一些。然而，这是值得的，因为许多基本问题都可以得到清晰的解答。

3.1 激励性示例：如何改进策略？

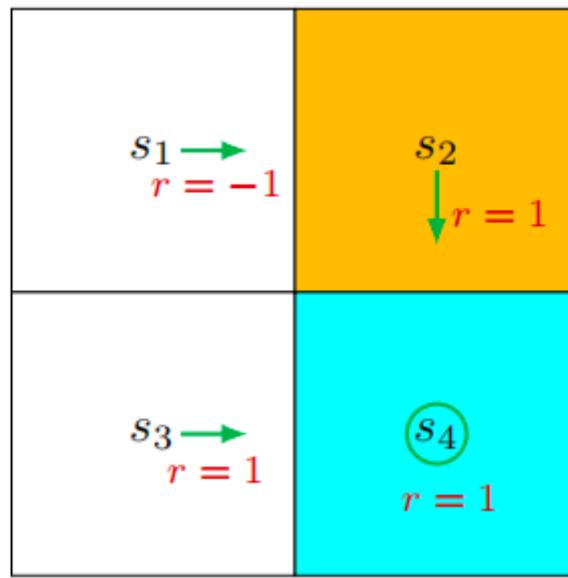


Figure 3.2: An example for demonstrating policy improvement.

图 3.2：演示策略改进的示例。

考虑图 3.2 中所示的策略。这里，橙色和蓝色的单元格分别代表禁止区域和目标区域。这里的策略不好，因为它在状态 s_1 选择了 a_2 （向右）。我们如何改进给定的策略以获得更好的策略？答案在于状态价值和动作价值。

- **直觉：**直观上很清楚，如果策略在 s_1 选择 a_3 （向下）而不是 a_2 （向右），策略就可以得到改进。这是因为向下移动使智能体能够避免进入禁止区域。
- **数学：**上述直觉可以通过计算状态价值和动作价值来实现。

首先，我们计算给定策略的状态价值。特别地，该策略的贝尔曼方程为

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

令 $\gamma = 0.9$ 。可以很容易地解出

$$\begin{aligned} v_\pi(s_4) &= v_\pi(s_3) = v_\pi(s_2) = 10, \\ v_\pi(s_1) &= 8. \end{aligned}$$

其次，我们计算状态 s_1 的动作价值：

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

值得注意的是，动作 a_3 具有最大的动作价值：

$$q_\pi(s_1, a_3) \geq q_\pi(s_1, a_i), \quad \text{对于所有 } i \neq 3.$$

因此，我们可以更新策略以在 s_1 处选择 a_3 。

这个例子说明，如果我们更新策略以选择具有**最大动作价值**的动作，我们就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，因为给定的策略仅在状态 s_1 处表现不佳。如果策略在其他状态下也不好，选择具有最大动作价值的动作是否仍然会生成更好的策略？此外，是否总是存在最优策略？最优策略是什么样子的？我们将在本章回答所有这些问题。

3.2 最优状态价值与最优策略

虽然强化学习的最终目标是获得最优策略，但首先有必要定义什么是最优策略。该定义基于状态价值。特别地，考虑两个给定的策略 π_1 和 π_2 。如果对于任意状态， π_1 的状态价值都大于或等于 π_2 的状态价值：

$$v_{\pi_1}(s) \geq v_{\pi_2}(s), \quad \text{对于所有 } s \in \mathcal{S},$$

那么我们就说 π_1 比 π_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。这在下面进行了形式化的陈述。

定义 3.1 (最优策略与最优状态价值)。如果对于所有 $s \in \mathcal{S}$ 以及任何其他策略 π ，都有 $v_{\pi^*}(s) \geq v_{\pi}(s)$ ，则策略 π^* 是**最优的 (optimal)**。 π^* 的状态价值即为**最优状态价值** (optimal state values)。

上述定义表明，与所有其他策略相比，最优策略在每个状态下都具有最大的状态价值。这个定义也引出了许多问题：

- 存在性 (Existence)：最优策略是否存在？
- 唯一性 (Uniqueness)：最优策略是唯一的吗？
- 随机性 (Stochasticity)：最优策略是随机的还是确定性的？
- 算法 (Algorithm)：如何获得最优策略和最优状态价值？

必须清楚地回答这些基本问题，才能彻底理解最优策略。例如，关于最优策略的存在性，如果最优策略不存在，那么我们就不必费心设计算法去寻找它们。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**最优状态价值**的工具是**贝尔曼最优方程 (Bellman optimality equation, BOE)**。通过求解该方程，我们可以获得最优策略和最优状态价值。接下来我们给出 BOE 的表达式并对其进行详细分析。

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

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

其中 $v(s), v(s')$ 是待求解的未知变量，且

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

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

BOE 是分析最优策略的一种优雅且强大的工具。然而，理解这个方程可能并非易事。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，如何从一个方程中求解两个未知变量可能会令人困惑。此外，BOE 实际上是一种特殊的贝尔曼方程。然而，这一点并不直观，因为它的表达式与贝尔曼方程的表达式截然不同。我们还需要回答关于 BOE 的以下基本问题。

- 存在性 (Existence)：该方程有解吗？
- 唯一性 (Uniqueness)：解是唯一的吗？
- 算法 (Algorithm)：如何求解该方程？
- 最优性 (Optimality)：解与最优策略有什么关系？

一旦我们回答了这些问题，我们将清楚地理解最优状态价值和最优策略。

3.3.1 BOE 右侧的最大化

接下来我们阐明如何求解 (3.1) 中 BOE 右侧的最大化问题。对于初学者来说，如何从一个方程中求解两个未知变量 $v(s)$ 和 $\pi(a|s)$ 可能会令人困惑。事实上，这两个未知变量可以逐一求解。这个想法通过以下示例进行说明。

示例 3.1。 考虑满足以下条件的两个未知变量 $x, y \in \mathbb{R}$

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

第一步是求解方程右侧的 y 。无论 x 的值如何，我们总是有 $\max_y (2x - 1 - y^2) = 2x - 1$ ，其中最大值在 $y = 0$ 时取得。第二步是求解 x 。当 $y = 0$ 时，方程变为 $x = 2x - 1$ ，这导致 $x = 1$ 。因此， $y = 0$ 和 $x = 1$ 是方程的解。

我们现在转向 BOE 右侧的最大化问题。(3.1) 中的 BOE 可以简洁地写为

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

受示例 3.1 的启发，我们可以首先求解右侧的最优 π 。怎么做呢？下面的例子展示了它的基本思想。

示例 3.2。 给定 $q_1, q_2, q_3 \in \mathbb{R}$ ，我们想要找到 c_1, c_2, c_3 的最优点以最大化

$$\sum_{i=1}^3 c_i q_i = c_1 q_1 + c_2 q_2 + c_3 q_3,$$

其中 $c_1 + c_2 + c_3 = 1$ 且 $c_1, c_2, c_3 \geq 0$ 。

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

$$q_3 = (c_1 + c_2 + c_3)q_3 = c_1 q_3 + c_2 q_3 + c_3 q_3 \geq c_1 q_1 + c_2 q_2 + c_3 q_3$$

对于任意 c_1, c_2, c_3 都成立。

(如果不理解，还是需要仔细看上述两个实例)

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

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

其中等号成立的条件是

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

这里， $a^* = \arg \max_a q(s, a)$ 。总之，最优策略 $\pi(s)$ 是选择具有最大 $q(s, a)$ 值的动作的策略。

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。如果我们把这些方程组合起来，我们可以得到一个简洁的矩阵-向量形式，这将在本章中被广泛使用。

BOE 的矩阵-向量形式为

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

其中 $v \in \mathbb{R}^{|\mathcal{S}|}$ ，且 \max_{π} 是以逐元素方式进行的（这个是对于理解整体性方程是有益的）。 r_π 和 P_π 的结构与普通贝尔曼方程的矩阵-向量形式中的结构相同：

$$[r_\pi]_s \doteq \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{r \in \mathcal{R}} p(r|s, a)r, \quad [P_\pi]_{s, s'} = p(s'|s) \doteq \sum_{a \in \mathcal{A}} \pi(a|s) p(s'|s, a).$$

由于 π 的最优点由 v 决定，(3.2) 的右侧是 v 的函数，记为

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

那么，BOE 可以表达为简洁形式

$$v = f(v). \tag{3.3}$$

在本节的剩余部分，我们将展示如何求解这个非线性方程。

3.3.3 压缩映射定理

由于 BOE 可以表示为非线性方程 $v = f(v)$ ，接下来我们介绍压缩映射定理 (contraction mapping theorem) [6] 来对其进行分析。压缩映射定理是分析一般非线性方程的有力工具。它也被称为不动点定理 (fixed-point theorem)。已经了解该定理的读者可以跳过这部分。否则，建议读者熟悉该定理，因为它是分析 BOE (贝尔曼最优方程) 的关键。

考虑一个函数 $f(x)$ ，其中 $x \in \mathbb{R}^d$ 且 $f : \mathbb{R}^d \rightarrow \mathbb{R}^d$ 。如果满足以下条件，则点 x^* 被称为不动点 (fixed point)：

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

上述方程的解释是， x^* 的映射是其自身。这就是为什么 x^* 被称为“固定”(fixed)的原因。如果存在 $\gamma \in (0, 1)$ 使得

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

对于任意 $x_1, x_2 \in \mathbb{R}^d$ 都成立，则函数 f 是一个压缩映射 (contraction mapping) (或收缩函数 (contractive function))。在本书中， $\|\cdot\|$ 表示向量范数或矩阵范数。

示例 3.3。 我们给出三个例子来演示不动点和压缩映射。

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

很容易验证 $x = 0$ 是一个不动点，因为 $0 = 0.5 \cdot 0$ 。此外， $f(x) = 0.5x$ 是一个压缩映射，因为对于任意 $\gamma \in [0.5, 1]$ ，都有 $\|0.5x_1 - 0.5x_2\| = 0.5\|x_1 - x_2\| \leq \gamma\|x_1 - x_2\|$ 。

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

很容易验证 $x = 0$ 是一个不动点，因为 $0 = A0$ 。为了查看压缩性质，我们有

$\|Ax_1 - Ax_2\| = \|A(x_1 - x_2)\| \leq \|A\|\|x_1 - x_2\| \leq \gamma\|x_1 - x_2\|$ 。因此， $f(x) = Ax$ 是一个压缩映射。

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

很容易看出 $x = 0$ 是一个不动点，因为 $0 = 0.5 \sin 0$ 。此外，根据中值定理 (mean value theorem) [7, 8] 可得

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

结果是， $|0.5 \sin x_1 - 0.5 \sin x_2| \leq 0.5|x_1 - x_2|$ ，因此 $f(x) = 0.5 \sin x$ 是一个压缩映射。

不动点与压缩性质之间的关系由以下经典定理刻画。

定理 3.1 (压缩映射定理)。 对于任何形式为 $x = f(x)$ 的方程，其中 x 和 $f(x)$ 是实向量，如果 f 是一个压缩映射，那么以下性质成立。

- 存在性 (Existence)：存在一个满足 $f(x^*) = x^*$ 的不动点 x^* 。
- 唯一性 (Uniqueness)：不动点 x^* 是唯一的。
- 算法 (Algorithm)：考虑迭代过程：

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

其中 $k = 0, 1, 2, \dots$ 。那么，对于任意初始猜测 x_0 ，当 $k \rightarrow \infty$ 时， $x_k \rightarrow x^*$ 。此外，收敛速度是指数级的。

压缩映射定理不仅能告诉我们非线性方程的解是否存在，还提供了一种求解该方程的数值算法。定理的证明在方框 3.1 中给出。

下面的例子演示了如何使用压缩映射定理建议的迭代算法来计算某些方程的不动点。

示例 3.4。 让我们回顾一下上述例子： $x = 0.5x$ ， $x = Ax$ ，以及 $x = 0.5 \sin x$ 。虽然已经表明这三个方程的右侧都是压缩映射，但根据压缩映射定理，它们每一个都有唯一的不动点，很容易验证该不动点为 $x^* = 0$ 。此外，这三个方程的不动点可以通过以下算法迭代求解：

$$\begin{aligned} x_{k+1} &= 0.5x_k, \\ x_{k+1} &= Ax_k, \\ x_{k+1} &= 0.5 \sin x_k, \end{aligned}$$

给定任意初始猜测 x_0 。

方框 3.1：压缩映射定理的证明

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

证明依赖于柯西序列 (Cauchy sequences)。如果对于任意小的 $\varepsilon > 0$ ，存在 N 使得对于所有 $m, n > N$ 都有 $\|x_m - x_n\| < \varepsilon$ ，则序列 x_1, x_2, \dots 被称为柯西序列。

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，因为柯西序列保证收敛到一个极限。其收敛性质将用于证明压缩映射定理。注意，我们必须对所有 $m, n > N$ 都有 $\|x_m - x_n\| < \varepsilon$ 。如果我们仅有 $x_{n+1} - x_n \rightarrow 0$ ，不足以宣称该序列是柯西序列。

例如，对于 $x_n = \sqrt{n}$ ，虽然 $x_{n+1} - x_n \rightarrow 0$ 成立，但显然 $x_n = \sqrt{n}$ 发散。

接下来我们证明 $\{x_k = f(x_{k-1})\}_{k=1}^{\infty}$ 是一个柯西序列，因此是收敛的。

首先，由于 f 是一个压缩映射，我们有

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

同理，我们有 $\|x_k - x_{k-1}\| \leq \gamma \|x_{k-1} - x_{k-2}\|, \dots, \|x_2 - x_1\| \leq \gamma \|x_1 - x_0\|$ 。因此，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 $\gamma < 1$ ，我们知道随着 $k \rightarrow \infty$, $\|x_{k+1} - x_k\|$ 以指数速度收敛到零。值得注意的是， $\{\|x_{k+1} - x_k\|\}$ 的收敛并不足以推导出 $\{x_k\}$ 的收敛。因此，我们需要进一步考虑任意 $m > n$ 时的 $\|x_m - x_n\|$ 。特别地，

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

结果是，对于任意 ε ，我们总是可以找到 N 使得对于所有 $m, n > N$ 都有 $\|x_m - x_n\| < \varepsilon$ 。因此，该序列是柯西序列，并由此收敛到一个极限点，记为 $x^* = \lim_{k \rightarrow \infty} x_k$ 。

第2部分：我们证明极限 $x^* = \lim_{k \rightarrow \infty} x_k$ 是一个不动点。

为此，由于

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

我们知道 $\|f(x_k) - x_k\|$ 以指数速度收敛到零。因此，我们在极限处有 $\$f(x^*) = x^*\$$ 。

第3部分：我们证明不动点是唯一的。

假设存在另一个满足 $f(x') = x'$ 的不动点 x' 。那么，

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

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 以指数速度收敛到 x^* 。 回顾在 (3.4) 中证明的 $\|x_m - x_n\| \leq \frac{\gamma^n}{1 - \gamma} \|x_1 - x_0\|$ 。由于 m 可以任意大，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 $\gamma < 1$ ，误差随着 $n \rightarrow \infty$ 以指数速度收敛到零。

3.3.4 BOE 右侧的压缩性质

接下来我们证明 (3.3) 中 BOE 的 $f(v)$ 是一个压缩映射。因此，上一小节介绍的压缩映射定理可以被应用。

定理 3.2 ($f(v)$ 的压缩性质)。(3.3) 中 BOE 右侧的函数 $f(v)$ 是一个压缩映射。特别地，对于任意 $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$ ，成立

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

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

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

方框 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 $\pi_1^* \doteq \arg \max_{\pi} (r_{\pi} + \gamma P_{\pi} v_1)$ 以及 $\pi_2^* \doteq \arg \max_{\pi} (r_{\pi} + \gamma P_{\pi} v_2)$ 。那么，

$$\begin{aligned} f(v_1) &= \max_{\pi} (r_{\pi} + \gamma P_{\pi} v_1) = r_{\pi_1^*} + \gamma P_{\pi_1^*} v_1 \geq r_{\pi_2^*} + \gamma P_{\pi_2^*} v_1, \\ f(v_2) &= \max_{\pi} (r_{\pi} + \gamma P_{\pi} v_2) = r_{\pi_2^*} + \gamma P_{\pi_2^*} v_2 \geq r_{\pi_1^*} + \gamma P_{\pi_1^*} v_2, \end{aligned}$$

其中 \geq 是逐元素比较。结果是，很重要

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

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

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

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 都是逐元素算子。 根据定义， $z \geq 0$ 。一方面，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

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

由此可得

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

其中 $\|\cdot\|_{\infty}$ 是最大范数。

另一方面，假设 z_i 是 z 的第 i 个分量，且 p_i^T 和 q_i^T 分别是 $P_{\pi_1^*}$ 和 $P_{\pi_2^*}$ 的第 i 行。那么，

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

由于 p_i 是一个所有元素均非负且元素之和等于 1 的向量，因此可得

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

同理，我们有 $|q_i^T (v_1 - v_2)| \leq \|v_1 - v_2\|_{\infty}$ 。因此， $z_i \leq \gamma \|v_1 - v_2\|_{\infty}$ ，故

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

将此不等式代入 (3.5) 可得

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

这就完成了对 $f(v)$ 压缩性质的证明。

3.4 从 BOE 求解最优策略

经过上一节的准备，我们现在准备求解 BOE 以获得**最优状态价值** v^* 和**最优策略** π^* 。

- 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^* 是一个不动点，因为 $v^* = f(v^*)$ 。因此，压缩映射定理表明以下结果。

定理 3.3 (存在性、唯一性和算法)。对于 BOE $v = f(v) = \max_{\pi \in \Pi} (r_{\pi} + \gamma P_{\pi} v)$ ，总是存在唯一的解 v^* ，可以通过以下方式迭代求解

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

给定任意初始猜测 v_0 ，当 $k \rightarrow \infty$ 时， v_k 的值以指数速度收敛到 v^* 。

该定理的证明直接遵循压缩映射定理，因为 $f(v)$ 是一个压缩映射。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 的解总是存在的。
- v^* 的唯一性：解 v^* 总是唯一的。
- 求解 v^* 的算法： v^* 的值可以通过定理 3.3 建议的迭代算法求解。这个迭代算法有一个特定的名称，叫做价值迭代（value iteration）。其实现将在第 4 章详细介绍。我们在本章主要关注 BOE 的基本性质。
- **求解 π^* ：一旦获得 v^* 的值，我们可以通过求解下式轻松获得 π^***

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

π^* 的值将在定理 3.5 中给出。将 (3.6) 代入 BOE 可得

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

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

至此，虽然我们可以求解 v^* 和 π^* ，但尚不清楚该解是否是最优的。下面的定理揭示了该解的最优化。

定理 3.4 (v^* 和 π^* 的最优化)。 解 v^* 是最优状态价值， π^* 是最优策略。也就是说，对于任意策略 π ，都有

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

其中 v_π 是 π 的状态价值， \geq 是逐元素比较。

现在，我们清楚了为什么要研究 BOE：它的解对应于最优状态价值和最优策略。上述定理的证明在下方的方框中给出。

方框 3.3：定理 3.4 的证明

对于任意策略 π ，都有

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

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

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

反复应用上述不等式可得 $v^* - v_\pi \geq \gamma P_\pi (v^* - v_\pi) \geq \gamma^2 P_\pi^2 (v^* - v_\pi) \geq \dots \geq \gamma^n P_\pi^n (v^* - v_\pi)$ 。由此可得

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

其中最后一个等式成立是因为 $\gamma < 1$ 且 P_π^n 是一个非负矩阵，其所有元素均小于或等于 1（因为 $P_\pi^n \mathbf{1} = \mathbf{1}$ ）。因此，对于任意 π ，都有 $v^* \geq v_\pi$ 。

接下来我们更仔细地考察 (3.6) 中的 π^* 。特别地，下面的定理表明总是存在一个确定性的贪婪策略是最优的。

定理 3.5 (贪婪最优策略)。 对于任意 $s \in \mathcal{S}$ ，确定性贪婪策略

(注：这个策略最大化了最优的动作价值)

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

是求解 BOE 的最优策略。这里，

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

其中

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

方框 3.4：定理 3.5 的证明

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

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

很明显，如果 $\pi(s)$ 选择具有最大 $q^*(s, a)$ 的动作，则 $\sum_{a \in \mathcal{A}} \pi(a|s)q^*(s, a)$ 达到最大值。（这句话很重要，这暗示，使得最大动作价值的动作具备最大概率，因此是一个贪婪的策略）

(3.7) 中的策略被称为贪婪 (greedy) 策略，因为它寻找具有最大 $q^*(s, a)$ 的动作。最后，我们讨论 π^* 的两个重要性质。

- 最优策略的唯一性 (Uniqueness of optimal policies)：虽然 v^* 的值是唯一的，但对应于 v^* 的最优策略可能不是唯一的。这可以通过反例轻松验证。例如，图 3.3 中所示的两个策略都是最优的。
- 最优策略的随机性 (Stochasticity of optimal policies)：最优策略可以是随机的或确定性的，如图 3.3 所示。然而，根据定理 3.5，可以确定总是存在一个确定性的最优策略。

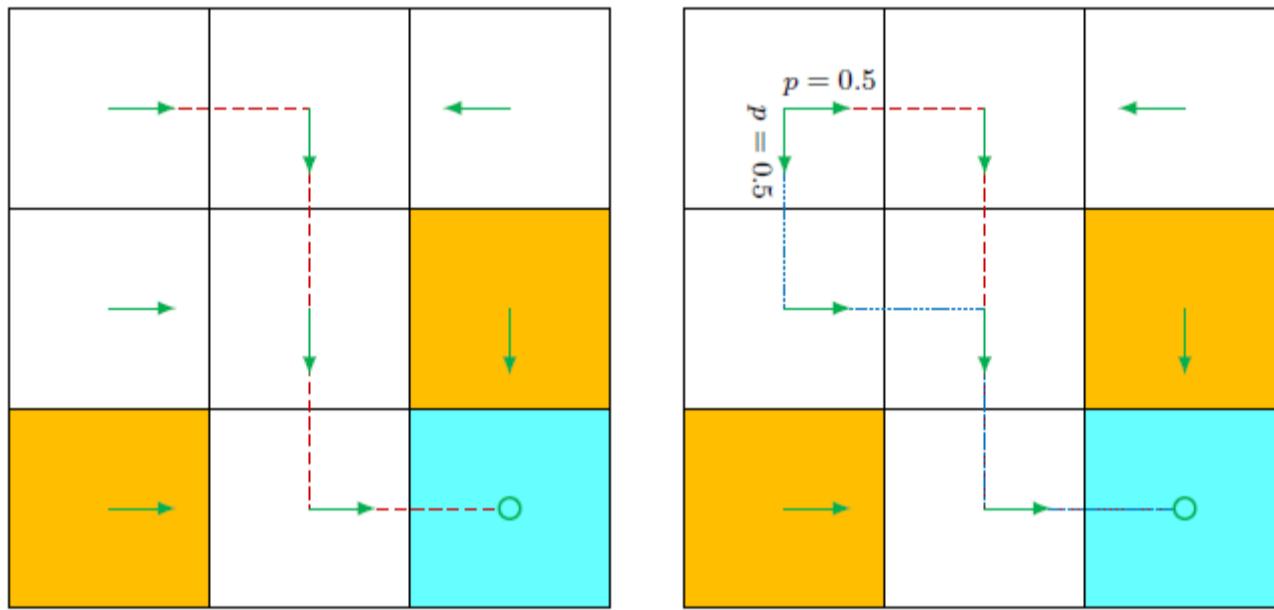


图 3.3：演示最优策略可能不唯一的示例。这两个策略不同，但都是最优的。

3.5 影响最优策略的因素

BOE 是分析最优策略的有力工具。接下来我们应用 BOE 来研究哪些因素会影响最优策略。通过观察 BOE 的元素形式表达式，这个问题很容易回答：

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

最优状态价值和最优策略由以下参数决定：

- 1) 即时奖励 r
- 2) 折扣率 γ ，
- 3) 系统模型 $p(s'|s, a), p(r|s, a)$ 。

在系统模型固定的情况下，我们接下来讨论当我们改变 r 和 γ 的值时，最优策略如何变化。本节中展示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性质。

一个基准示例

考虑图 3.4 中的示例。奖励设置如下： $r_{\text{boundary}} = r_{\text{forbidden}} = -1$ 且 $r_{\text{target}} = 1$ 。此外，智能体每移动一步都会收到 $r_{\text{other}} = 0$ 的奖励。折扣率选择为 $\gamma = 0.9$ 。

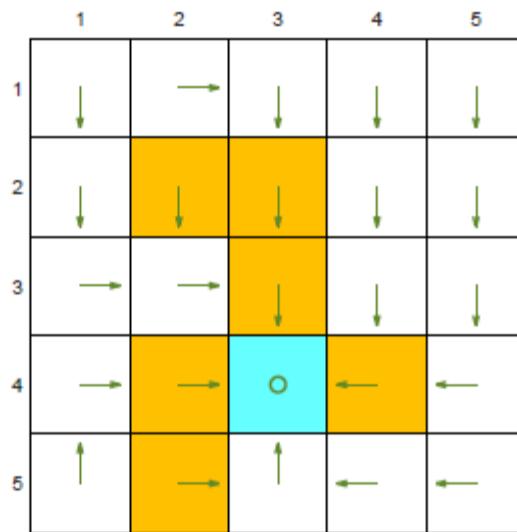
在上述参数下，最优策略和最优状态价值如图 3.4(a) 所示。有趣的是，智能体并不害怕穿过禁止区域以到达目标区域。具体来说，从 (第4行, 第1列) 的状态出发，智能体有两个到达目标区域的选项。第一个选项是避开所有禁止区域，长途跋涉到达目标区域。第二个选项是穿过禁止区域。虽然智能体在进入禁止区域时会获得负奖励，但第二条轨迹的累积奖励大于第一条轨迹的累积奖励。因此，由于 γ 的值相对较大，最优策略是远视的 (far-sighted)。

折扣率的影响

如果我们保持其他参数不变，将折扣率从 $\gamma = 0.9$ 更改为 $\gamma = 0.5$ ，最优策略将变为图 3.4(b) 所示的策略。有趣的是，智能体不再敢于冒险。相反，它会走很长的路去到达目标，同时避开所有的禁止区域。这是因为由于 γ 的值相对较小，最优策略变得近视 (short-sighted)。

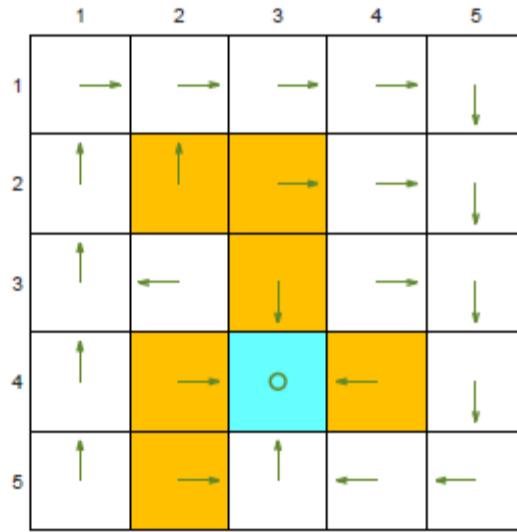
在 $\gamma = 0$ 的极端情况下，对应的最优策略如图 3.4(c) 所示。在这种情况下，智能体无法到达目标区域。**这是因为每个状态的最优策略都是极度短视的 (extremely short-sighted)**，仅仅选择具有最大即时奖励 (immediate reward) 的动作，而不是最大总回报 (total reward)。

此外，状态价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价值，而远离目标的状态具有较低的价值。这种模式可以从图 3.4 所示的所有示例中观察到。这可以通过折扣率来解释：如果一个状态必须沿着较长的轨迹到达目标，由于折扣率的存在，其状态价值会较小。



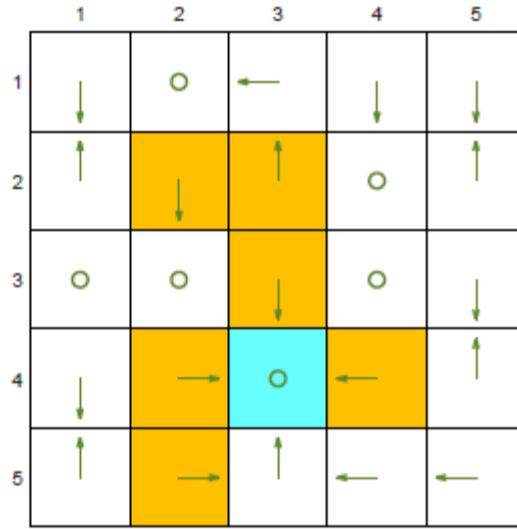
	1	2	3	4	5
1	5.8	5.6	6.2	6.5	5.8
2	6.5	7.2	8.0	7.2	6.5
3	7.2	8.0	10.0	8.0	7.2
4	8.0	10.0	10.0	10.0	8.0
5	7.2	9.0	10.0	9.0	8.1

(a) Baseline example: $r_{\text{boundary}} = r_{\text{forbidden}} = -1$, $r_{\text{target}} = 1$, $\gamma = 0.9$.



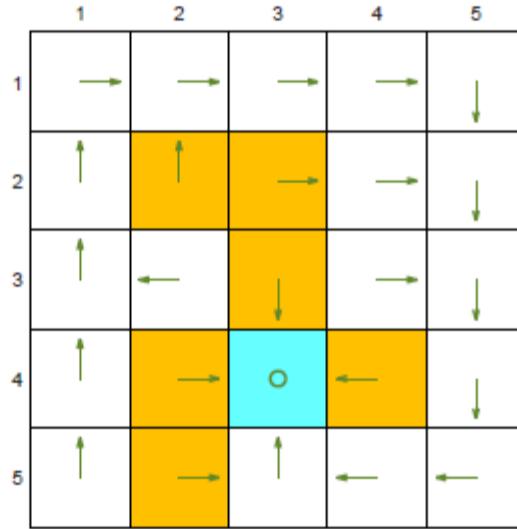
	1	2	3	4	5
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.1
3	0.0	0.0	2.0	0.1	0.1
4	0.0	2.0	2.0	2.0	0.2
5	0.0	1.0	2.0	1.0	0.5

(b) The discount rate is changed to $\gamma = 0.5$. The other parameters are the same as those in (a).



	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	0.0
3	0.0	0.0	1.0	0.0	0.0
4	0.0	1.0	1.0	1.0	0.0
5	0.0	0.0	1.0	0.0	0.0

(c) The discount rate is changed to $\gamma = 0$. The other parameters are the same as those in (a).



	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

(d) $r_{\text{forbidden}}$ is changed from -1 to -10 . The other parameters are the same as those in (a).

Figure 3.4: The optimal policies and optimal state values given different parameter values.

奖励值的影响

如果我们要严格禁止智能体进入任何禁止区域，我们可以增加这样做所受到的惩罚。例如，如果 $r_{\text{forbidden}}$ 从 -1 变为 -10 ，由此产生的最优策略可以避开所有禁止区域（见图 3.4(d)）。

然而，改变奖励并不总是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（affine transformations）是不变的（invariant）。换句话说，如果我们缩放所有奖励或对所有奖励加上相同的值，最优策略保持不变。

定理 3.6 (最优策略的不变性)。 考虑一个马尔可夫决策过程，其最优状态价值 $v^* \in \mathbb{R}^{|\mathcal{S}|}$ 满足 $v^* = \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v^*)$ 。如果每个奖励 $r \in \mathcal{R}$ 通过仿射变换变为 $\alpha r + \beta$ ，其中 $\alpha, \beta \in \mathbb{R}$ 且 $\alpha > 0$ ，那么对应的最优状态价值 v' 也是 v^* 的仿射变换：

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

其中 $\gamma \in (0, 1)$ 是折扣率， $\mathbf{1} = [1, \dots, 1]^T$ 。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。

方框 3.5：定理 3.6 的证明

对于任意策略 π ，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 变为

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

接下来，我们通过证明 $v' = \alpha v^* + c \mathbf{1}$ （其中 $c = \beta/(1 - \gamma)$ ）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) 可得

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

其中最后一个等式是由于 $P_\pi \mathbf{1} = \mathbf{1}$ 这一事实。上述方程可以重组为

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

这等价于

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

由于 $c = \beta/(1 - \gamma)$ ，上述方程成立，因此 $v' = \alpha v^* + c \mathbf{1}$ 是 (3.9) 的解。由于 (3.9) 是 BOE，因此 v' 也是唯一解。最后，由于 v' 是 v^* 的仿射变换，动作价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^* 导出的相同： $\arg \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v')$ 与 $\arg \max_{\pi \in \Pi} (r_\pi + \gamma P_\pi v^*)$ 相同。

读者可以参考 [9] 以进一步讨论修改奖励值在何种条件下能保持最优策略不变。

避免无意义的绕路

在奖励设置中，智能体每移动一步都会获得 $r_{\text{other}} = 0$ 的奖励（除非它进入禁止区域或目标区域，或者试图越过边界）。由于零奖励不是惩罚，最优策略会在到达目标之前进行无意义的绕路吗？我们是否应该将 r_{other} 设置为负值，以鼓励智能体尽快到达目标？

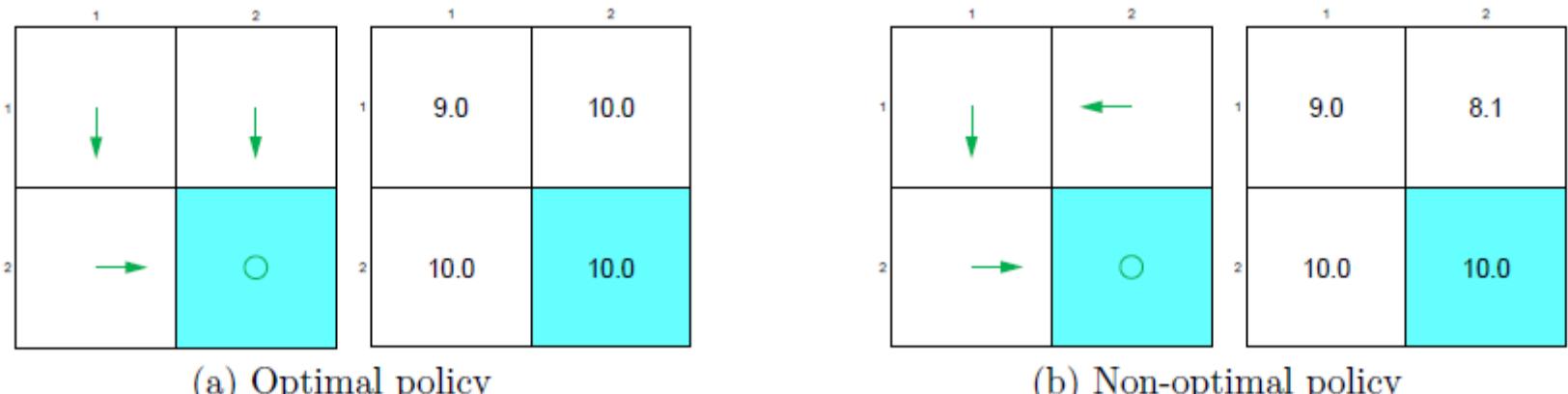


图 3.5：说明由于折扣率的存在，最优策略不会进行无意义绕路的示例。(a) 最优策略，(b) 非最优策略。

考虑图 3.5 中的示例，其中右下角的单元格是目标区域。这里的两个策略除了在状态 s_2 之外都是相同的。在图 3.5(a) 的策略中，智能体在 s_2 处向下移动，产生的轨迹是 $s_2 \rightarrow s_4$ 。在图 3.5(b) 的策略中，智能体向左移动，产生的轨迹是 $s_2 \rightarrow s_1 \rightarrow s_3 \rightarrow s_4$ 。

值得注意的是，第二个策略在到达目标区域之前绕了路。如果我们仅仅考虑即时奖励，绕路并不重要，因为不会获得负的即时奖励。然而，如果我们考虑折扣回报，那么绕路就很重要了。特别是，对于第一个策略，折扣回报为

$$\text{return} = 1 + \gamma 1 + \gamma^2 1 + \cdots = 1/(1 - \gamma) = 10.$$

作为比较，第二个策略的折扣回报为

$$\text{return} = 0 + \gamma 0 + \gamma^2 1 + \gamma^3 1 + \cdots = \gamma^2/(1 - \gamma) = 8.1.$$

很明显，轨迹越短，回报越大。因此，虽然每一步的即时奖励并不鼓励智能体尽快接近目标，但折扣率确实鼓励它这样做。

初学者可能会有一个误解，认为有必要在每次移动获得的奖励之上增加一个负奖励（例如 -1 ），以鼓励智能体尽快到达目标。这是一个误解，因为在所有奖励之上增加相同的奖励是一种仿射变换，它保持最优策略不变。此外，由于折扣率的存在，最优策略不会进行无意义的绕路，即使绕路可能不会收到任何即时的负奖励。

3.6 总结

本章的核心概念包括最优策略和最优状态价值。特别是，如果一个策略的状态价值大于或等于任何其他策略的状态价值，则该策略是最优的。最优策略的状态价值即为最优状态价值。BOE 是分析最优策略和最优状态价值的核心工具。该方程是一个具有良好压缩性质的非线性方程。我们可以应用压缩映射定理来分析该方程。结果表明，BOE 的解对应于最优状态价值和最优策略。这就是我们需要研究 BOE 的原因。

本章的内容对于彻底理解强化学习的许多基本思想非常重要。例如，定理 3.3 提出了一种求解 BOE 的迭代算法。该算法正是将在第 4 章中介绍的价值迭代算法。关于 BOE 的进一步讨论可以在文献 [2] 中找到。

3.7 问与答

- 问：最优策略的定义是什么？

答：如果一个策略对应的状态价值大于或等于任何其他策略的状态价值，则该策略是最优的。

需要注意的是，这种特定的最优化定义仅适用于表格型（tabular）强化学习算法。当价值或策略由函数近似时，必须使用不同的指标来定义最优策略。这一点在第 8 章和第 9 章中会变得更加清晰。

- 问：为什么贝尔曼最优方程很重要？

答：它很重要，因为它刻画了最优策略和最优状态价值。求解该方程可以得到最优策略及对应的最优状态价值。

- 问：贝尔曼最优方程是贝尔曼方程吗？

答：是的。贝尔曼最优方程是一种特殊的贝尔曼方程，其对应的策略是最优的。

- 问：贝尔曼最优方程的解是唯一的吗？

答：贝尔曼最优方程有两个未知变量。第一个未知变量是价值（value），第二个是策略（policy）。价值解（即最优状态价值）是唯一的。策略解（即最优策略）可能不是唯一的。

- 问：分析贝尔曼最优方程解的关键性质是什么？

答：关键性质是贝尔曼最优方程的右侧是一个压缩映射（contraction mapping）。因此，我们可以应用压缩映射定理来分析其解。

- 问：最优策略存在吗？

答：是的。根据对 BOE（贝尔曼最优方程）的分析，最优策略总是存在的。

- 问：最优策略是唯一的吗？

答：不一定。可能存在多个或无限个具有相同最优状态价值的最优策略。

- 问：最优策略是随机的还是确定性的？

答：最优策略既可以是确定性的，也可以是随机的。一个很好的事实是，总是存在确定性的贪婪最优策略。

- 问：如何获得最优策略？

答：使用定理 3.3 建议的迭代算法求解 BOE 可以产生最优策略。该迭代算法的详细实现将在第 4 章中给出。值得注意的是，本书介绍的所有强化学习算法都旨在不同设置下获得最优策略。

- 问：如果我们降低折扣率的值，对最优策略有什么普遍影响？

答：当我们降低折扣率时，最优策略会变得更加“近视”（short-sighted）。也就是说，即使以后可能获得更大的累积奖励，智能体也不敢冒险。

- 问：如果我们把折扣率设为零会发生什么？

答：由此产生的最优策略将变得极其短视。智能体将采取具有最大即时奖励的动作，即使该动作从长远来看并不好。

- 问：如果我们把所有奖励增加相同的数量，最优状态价值会改变吗？最优策略会改变吗？

答：将所有奖励增加相同的数量是奖励的仿射变换（affine transformation），这不会影响最优策略。然而，如 (3.8) 所示，最优状态价值会增加。

- 问：如果我们希望最优策略能够避免在到达目标前进行无意义的绕路，我们是否应该在每一步都增加一个负奖励，以便智能体尽快到达目标？

答：首先，在每一步引入额外的负奖励是奖励的仿射变换，这不会改变最优策略。其次，折扣率可以自动鼓励智能体尽快到达目标。这是因为无意义的绕路会增加轨迹长度并降低折扣回报。