Chapter 3

Optimal State Values and Bellman Optimality Equation

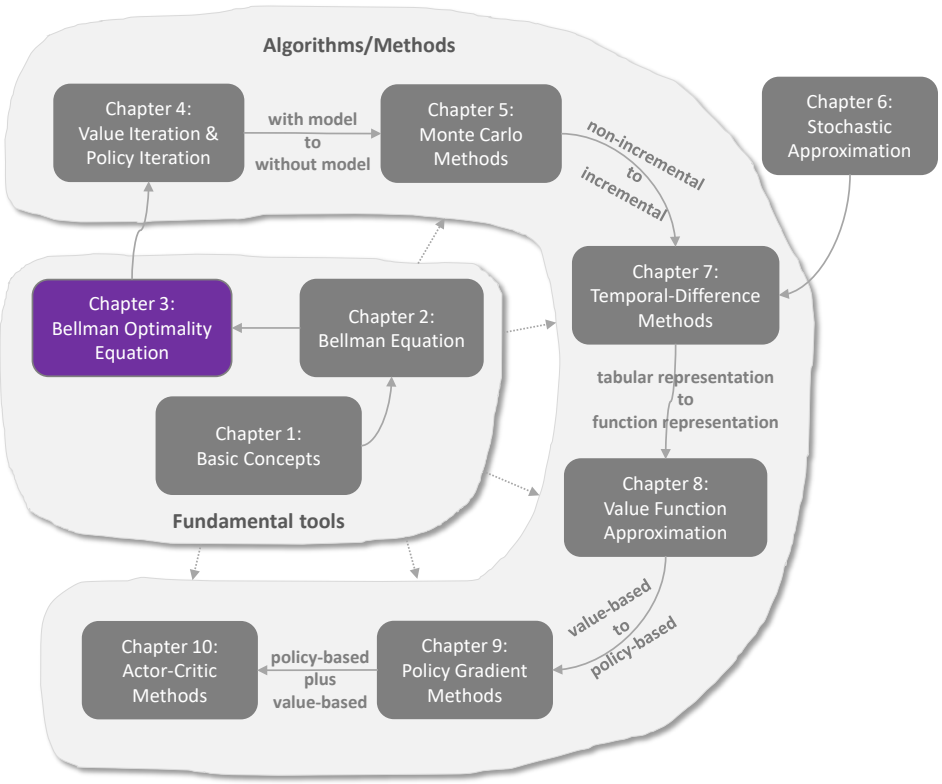


Figure 3.1: Where we are in this book.

The ultimate goal of reinforcement learning is to seek *optimal policies*. It is, therefore, necessary to define what optimal policies are. In this chapter, we introduce *a core concept* and *an important tool*. The core concept is the *optimal state value*, based on which we can define *optimal policies*. The important tool is the *Bellman optimality equation*, from which we can solve the optimal state values and policies.

The relationship between the previous, present, and subsequent chapters is as follows. The previous chapter (Chapter 2) introduced the Bellman equation of any given policy.

The present chapter introduces the Bellman optimality equation, which is a special Bellman equation whose corresponding policy is optimal. The next chapter (Chapter 4) will introduce an important algorithm called value iteration, which is exactly the algorithm for solving the Bellman optimality equation as introduced in the present chapter.

Be prepared that this chapter is slightly mathematically intensive. However, it is worth it because many fundamental questions can be clearly answered.

3.1 Motivating example: How to improve policies?

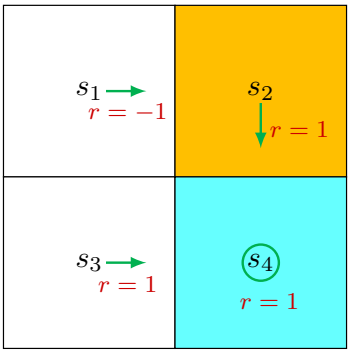


Figure 3.2: An example for demonstrating policy improvement.

Consider the policy shown in Figure 3.2. Here, the orange and blue cells represent the forbidden and target areas, respectively. The policy here is *not good* because it selects a_2 (rightward) at state s_1. How can we improve the given policy to obtain a better policy? The answer lies in state values and action values.

- ◇ *Intuition:* It is intuitively clear that the policy can improve if it selects a_3 (downward) instead of a_2 (rightward) at s_1. This is because moving downward enables the agent to avoid entering the forbidden area.
- ◇ *Mathematics:* The above intuition can be realized based on the calculation of state values and action values.

  First, we calculate the state values of the given policy. In particular, the Bellman equation of this policy is

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

Let $\gamma = 0.9$. It can be easily solved that

$$v_\pi(s_4) = v_\pi(s_3) = v_\pi(s_2) = 10,$$
$$v_\pi(s_1) = 8.$$

Second, we calculate the action values for state s_1:

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

It is notable that action a_3 has the greatest action value:

$$q_\pi(s_1, a_3) \geq q_\pi(s_1, a_i), \quad \text{for all } i \neq 3.$$

Therefore, we can update the policy to select a_3 at s_1.

This example illustrates that we can obtain a better policy if we update the policy to select the action with the *greatest action value*. This is the basic idea of many reinforcement learning algorithms.

This example is very simple in the sense that the given policy is only not good for state s_1. If the policy is also not good for the other states, will selecting the action with the greatest action value still generate a better policy? Moreover, whether there always exist optimal policies? What does an optimal policy look like? We will answer all of these questions in this chapter.

3.2 Optimal state values and optimal policies

While the ultimate goal of reinforcement learning is to obtain optimal policies, it is necessary to first define what an optimal policy is. The definition is based on state values. In particular, consider two given policies π_1 and π_2. If the state value of π_1 is greater than or equal to that of π_2 for any state:

$$v_{\pi_1}(s) \geq v_{\pi_2}(s), \quad \text{for all } s \in \mathcal{S},$$

then π_1 is said to be better than π_2. Furthermore, if a policy is better than all the other possible policies, then this policy is optimal. This is formally stated below.

Definition 3.1 (Optimal policy and optimal state value)**.** *A policy π^* is optimal if $v_{\pi^*}(s) \geq v_\pi(s)$ for all $s \in \mathcal{S}$ and for any other policy π. The state values of π^* are the optimal state values.*

The above definition indicates that an optimal policy has the greatest state value for every state compared to all the other policies. This definition also leads to many questions:

- ◇ Existence: Does the optimal policy exist?
- ◇ Uniqueness: Is the optimal policy unique?
- ◇ Stochasticity: Is the optimal policy stochastic or deterministic?
- ◇ Algorithm: How to obtain the optimal policy and the optimal state values?

These fundamental questions must be clearly answered to thoroughly understand optimal policies. For example, regarding the existence of optimal policies, if optimal policies do not exist, then we do not need to bother to design algorithms to find them.

We will answer all these questions in the remainder of this chapter.

3.3 Bellman optimality equation

The tool for analyzing optimal policies and optimal state values is the *Bellman optimality equation* (BOE). By solving this equation, we can obtain optimal policies and optimal state values. We next present the expression of the BOE and then analyze it in detail.

For every $s \in \mathcal{S}$, the elementwise expression of the BOE is

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

where $v(s), v(s')$ are unknown variables to be solved and

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

Here, $\pi(s)$ denotes a policy for state s, and $\Pi(s)$ is the set of all possible policies for s.

The BOE is an elegant and powerful tool for analyzing optimal policies. However, it may be nontrivial to understand this equation. For example, this equation has two unknown variables $v(s)$ and $\pi(a|s)$. It may be confusing to beginners how to solve two unknown variables from one equation. Moreover, the BOE is actually a special Bellman equation. However, it is nontrivial to see that since its expression is quite different from that of the Bellman equation. We also need to answer the following fundamental questions about the BOE.

- ◇ Existence: Does this equation have a solution?
- ◇ Uniqueness: Is the solution unique?
- ◇ Algorithm: How to solve this equation?
- ◇ Optimality: How is the solution related to optimal policies?

Once we can answer these questions, we will clearly understand optimal state values and optimal policies.

3.3.1 Maximization of the right-hand side of the BOE

We next clarify how to solve the maximization problem on the right-hand side of the BOE in (3.1). At first glance, it may be confusing to beginners how to solve *two* unknown variables $v(s)$ and $\pi(a|s)$ from *one* equation. In fact, these two unknown variables can be solved one by one. This idea is illustrated by the following example.

Example 3.1. *Consider two unknown variables $x, y \in \mathbb{R}$ that satisfy*

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

The first step is to solve y on the right-hand side of the equation. Regardless of the value of x, we always have $\max_y(2x - 1 - y^2) = 2x - 1$, where the maximum is achieved when $y = 0$. The second step is to solve x. When $y = 0$, the equation becomes $x = 2x - 1$, which leads to $x = 1$. Therefore, $y = 0$ and $x = 1$ are the solutions of the equation. □

We now turn to the maximization problem on the right-hand side of the BOE. The BOE in (3.1) can be written concisely as

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

Inspired by Example 3.1, we can first solve the optimal π on the right-hand side. How to do that? The following example demonstrates its basic idea.

Example 3.2. *Given $q_1, q_2, q_3 \in \mathbb{R}$, we would like to find the optimal values of c_1, c_2, c_3 to maximize*

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

where $c_1 + c_2 + c_3 = 1$ and $c_1, c_2, c_3 \geq 0$.

Without loss of generality, suppose that $q_3 \geq q_1, q_2$. Then, the optimal solution is $c_3^* = 1$ and $c_1^* = c_2^* = 0$. This is because

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

for any c_1, c_2, c_3. □

Inspired by the above example, since $\sum_a \pi(a|s) = 1$, we have

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

where equality is achieved when

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

Here, $a^* = \arg\max_a q(s,a)$. In summary, the optimal policy $\pi(s)$ is the one that selects the action that has the greatest value of $q(s,a)$.

3.3.2 Matrix-vector form of the BOE

The BOE refers to a set of equations defined for all states. If we combine these equations, we can obtain a concise matrix-vector form, which will be extensively used in this chapter.

The matrix-vector form of the BOE is

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

where $v \in \mathbb{R}^{|\mathcal{S}|}$ and $\max_\pi$ is performed in an elementwise manner. The structures of r_π and P_π are the same as those in the matrix-vector form of the normal Bellman equation:

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

Since the optimal value of π is determined by v, the right-hand side of (3.2) is a function of v, denoted as

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

Then, the BOE can be expressed in a concise form as

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

In the remainder of this section, we show how to solve this nonlinear equation.

3.3.3 Contraction mapping theorem

Since the BOE can be expressed as a nonlinear equation $v = f(v)$, we next introduce the contraction mapping theorem [6] to analyze it. The contraction mapping theorem is a powerful tool for analyzing general nonlinear equations. It is also known as the fixed-point theorem. Readers who already know this theorem can skip this part. Otherwise, the reader is advised to be familiar with this theorem since it is the key to analyzing the

BOE.

Consider a function $f(x)$, where $x \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}^d$. A point x^* is called a *fixed point* if

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

The interpretation of the above equation is that the map of x^* is itself. This is the reason why x^* is called "fixed". The function f is a *contraction mapping* (or contractive function) if there exists $\gamma \in (0, 1)$ such that

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

for any $x_1, x_2 \in \mathbb{R}^d$. In this book, $\|\cdot\|$ denotes a vector or matrix norm.

Example 3.3. *We present three examples to demonstrate fixed points and contraction mappings.*

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

  It is easy to verify that $x = 0$ is a fixed point since $0 = 0.5 \cdot 0$. Moreover, $f(x) = 0.5x$ is a contraction mapping because $\|0.5x_1 - 0.5x_2\| = 0.5\|x_1 - x_2\| \leq \gamma\|x_1 - x_2\|$ for any $\gamma \in [0.5, 1)$.

- ◇ $x = f(x) = Ax$, *where* $x \in \mathbb{R}^n, A \in \mathbb{R}^{n\times n}$ *and* $\|A\| \leq \gamma < 1$.

  It is easy to verify that $x = 0$ is a fixed point since $0 = A0$. To see the contraction property, $\|Ax_1 - Ax_2\| = \|A(x_1 - x_2)\| \leq \|A\|\|x_1 - x_2\| \leq \gamma\|x_1 - x_2\|$. Therefore, $f(x) = Ax$ is a contraction mapping.

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

  It is easy to see that $x = 0$ is a fixed point since $0 = 0.5 \sin 0$. Moreover, it follows from the mean value theorem *[7, 8] that*

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

  As a result, $|0.5 \sin x_1 - 0.5 \sin x_2| \leq 0.5|x_1 - x_2|$ and hence $f(x) = 0.5 \sin x$ is a contraction mapping. □

The relationship between a fixed point and the contraction property is characterized by the following classic theorem.

Theorem 3.1 (Contraction mapping theorem)**.** *For any equation that has the form $x = f(x)$ where x and $f(x)$ are real vectors, if f is a contraction mapping, then the following properties hold.*

- ◇ *Existence: There exists a fixed point x^* satisfying $f(x^*) = x^*$.*

◇ *Uniqueness: The fixed point x^* is unique.*

◇ *Algorithm: Consider the iterative process:*

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

where $k = 0, 1, 2, \ldots$. Then, $x_k \to x^*$ as $k \to \infty$ for any initial guess x_0. Moreover, the convergence rate is exponentially fast.

The contraction mapping theorem not only can tell whether the solution of a nonlinear equation exists but also suggests a numerical algorithm for solving the equation. The proof of the theorem is given in Box 3.1.

The following example demonstrates how to calculate the fixed points of some equations using the iterative algorithm suggested by the contraction mapping theorem.

Example 3.4. *Let us revisit the abovementioned examples: $x = 0.5x$, $x = Ax$, and $x = 0.5 \sin x$. While it has been shown that the right-hand sides of these three equations are all contraction mappings, it follows from the contraction mapping theorem that they each have a* unique *fixed point, which can be easily verified to be $x^* = 0$. Moreover, the fixed points of the three equations can be iteratively solved by the following algorithms:*

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

given any initial guess x_0. □

Box 3.1: Proof of the contraction mapping theorem

Part 1: We prove that the consequence $\{x_k\}_{k=1}^{\infty}$ with $x_k = f(x_{k-1})$ is convergent.

The proof relies on *Cauchy sequences*. A sequence $x_1, x_2, \cdots \in \mathbb{R}$ is called *Cauchy* if for any small $\varepsilon > 0$, there exists N such that $\|x_m - x_n\| < \varepsilon$ for all $m, n > N$. The intuitive interpretation is that there exists a finite integer N such that all the elements after N are sufficiently close to each other. Cauchy sequences are important because it is guaranteed that a Cauchy sequence converges to a limit. Its convergence property will be used to prove the contraction mapping theorem. Note that we must have $\|x_m - x_n\| < \varepsilon$ for all $m, n > N$. If we simply have $x_{n+1} - x_n \to 0$, it is insufficient to claim that the sequence is a Cauchy sequence. For example, it holds that $x_{n+1} - x_n \to 0$ for $x_n = \sqrt{n}$, but apparently, $x_n = \sqrt{n}$ diverges.

We next show that $\{x_k = f(x_{k-1})\}_{k=1}^{\infty}$ is a Cauchy sequence and hence converges.

First, since f is a contraction mapping, we have

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

Similarly, we have $\|x_k - x_{k-1}\| \leq \gamma \|x_{k-1} - x_{k-2}\|$, $\ldots$, $\|x_2 - x_1\| \leq \gamma \|x_1 - x_0\|$. Thus, we have

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

Since $\gamma < 1$, we know that $\|x_{k+1} - x_k\|$ converges to zero exponentially fast as $k \to \infty$ given any x_1, x_0. Notably, the convergence of $\{\|x_{k+1} - x_k\|\}$ is not sufficient for implying the convergence of $\{x_k\}$. Therefore, we need to further consider $\|x_m - x_n\|$ for any $m > n$. In particular,

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

As a result, for any ε, we can always find N such that $\|x_m - x_n\| < \varepsilon$ for all $m, n > N$. Therefore, this sequence is Cauchy and hence converges to a limit point denoted as $x^* = \lim_{k \to \infty} x_k$.

Part 2: We show that the limit $x^* = \lim_{k \to \infty} x_k$ is a fixed point. To do that, since

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

we know that $\|f(x_k) - x_k\|$ converges to zero exponentially fast. Hence, we have $f(x^*) = x^*$ at the limit.

Part 3: We show that the fixed point is unique. Suppose that there is another fixed point x' satisfying $f(x') = x'$. Then,

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

Since $\gamma < 1$, this inequality holds if and only if $\|x' - x^*\| = 0$. Therefore, $x' = x^*$.

Part 4: We show that x_k converges to x^* exponentially fast. Recall that $\|x_m - x_n\| \leq \frac{\gamma^n}{1-\gamma}\|x_1 - x_0\|$, as proven in (3.4). Since m can be arbitrarily large, we have

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

Since $\gamma < 1$, the error converges to zero exponentially fast as $n \to \infty$.

3.3.4 Contraction property of the right-hand side of the BOE

We next show that $f(v)$ in the BOE in (3.3) is a contraction mapping. Thus, the contraction mapping theorem introduced in the previous subsection can be applied.

Theorem 3.2 (Contraction property of $f(v)$)**.** *The function $f(v)$ on the right-hand side of the BOE in* (3.3) *is a contraction mapping. In particular, for any $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$, it holds that*

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

where $\gamma \in (0, 1)$ is the discount rate, and $\|\cdot\|_\infty$ is the maximum norm, which is the maximum absolute value of the elements of a vector.

The proof of the theorem is given in Box 3.2. This theorem is important because we can use the powerful contraction mapping theorem to analyze the BOE.

Box 3.2: Proof of Theorem 3.2

Consider any two vectors $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$, and suppose that $\pi_1^* \doteq \arg\max_\pi(r_\pi + \gamma P_\pi v_1)$ and $\pi_2^* \doteq \arg\max_\pi(r_\pi + \gamma P_\pi v_2)$. Then,

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

where $\geq$ is an elementwise comparison. As a result,

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

Similarly, it can be shown that $f(v_2) - f(v_1) \leq \gamma P_{\pi_2^*}(v_2 - v_1)$. Therefore,

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

Define

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

where $\max(\cdot)$, $|\cdot|$, and $\geq$ are all elementwise operators. By definition, $z \geq 0$. On the one hand, it is easy to see that

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

which implies

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

It then follows that

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

where $\|\cdot\|_\infty$ is the maximum norm.

On the other hand, suppose that z_i is the ith entry of z, and p_i^T and q_i^T are the ith row of $P_{\pi_1^*}$ and $P_{\pi_2^*}$, respectively. Then,

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

Since p_i is a vector with all nonnegative elements and the sum of the elements is equal to one, it follows that

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

Similarly, we have $|q_i^T(v_1 - v_2)| \leq \|v_1 - v_2\|_\infty$. Therefore, $z_i \leq \gamma\|v_1 - v_2\|_\infty$ and hence

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

Substituting this inequality to (3.5) gives

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

which concludes the proof of the contraction property of $f(v)$.

3.4 Solving an optimal policy from the BOE

With the preparation in the last section, we are ready to solve the BOE to obtain the optimal state value v^* and an optimal policy π^*.

◇ Solving v^*: If v^* is a solution of the BOE, then it satisfies

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

Clearly, v^* is a fixed point because $v^* = f(v^*)$. Then, the contraction mapping theorem suggests the following results.

Theorem 3.3 (Existence, uniqueness, and algorithm)**.** *For the BOE $v = f(v) = \max_{\pi \in \Pi}(r_\pi + \gamma P_\pi v)$, there always exists a unique solution v^*, which can be solved iteratively by*

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

The value of v_k converges to v^* exponentially fast as $k \to \infty$ given any initial guess v_0.

The proof of this theorem directly follows from the contraction mapping theorem since $f(v)$ is a contraction mapping. This theorem is important because it answers some fundamental questions.

- Existence of v^*: The solution of the BOE always exists.
- Uniqueness of v^*: The solution v^* is always unique.
- Algorithm for solving v^*: The value of v^* can be solved by the iterative algorithm suggested by Theorem 3.3. This iterative algorithm has a specific name called *value iteration*. Its implementation will be introduced in detail in Chapter 4. We mainly focus on the fundamental properties of the BOE in the present chapter.

◇ Solving π^*: Once the value of v^* has been obtained, we can easily obtain π^* by solving

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

The value of π^* will be given in Theorem 3.5. Substituting (3.6) into the BOE yields

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

Therefore, $v^* = v_{\pi^*}$ is the state value of π^*, and the BOE is a special Bellman equation whose corresponding policy is π^*.

At this point, although we can solve v^* and π^*, it is still unclear whether the solution is optimal. The following theorem reveals the optimality of the solution.

Theorem 3.4 (Optimality of v^* and π^*)**.** *The solution v^* is the optimal state value, and π^* is an optimal policy. That is, for any policy π, it holds that*

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

where v_π is the state value of π, and $\geq$ is an elementwise comparison.

Now, it is clear why we must study the BOE: its solution corresponds to optimal state values and optimal policies. The proof of the above theorem is given in the following box.

Box 3.3: Proof of Theorem 3.4

For any policy π, it holds that

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

Since

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

we have

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

Repeatedly applying the above inequality gives $v^* - v_\pi \geq \gamma P_\pi(v^* - v_\pi) \geq \gamma^2 P_\pi^2(v^* - v_\pi) \geq \cdots \geq \gamma^n P_\pi^n(v^* - v_\pi)$. It follows that

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

where the last equality is true because $\gamma < 1$ and P_π^n is a nonnegative matrix with all its elements less than or equal to 1 (because $P_\pi^n \mathbf{1} = \mathbf{1}$). Therefore, $v^* \geq v_\pi$ for any π.

We next examine π^* in (3.6) more closely. In particular, the following theorem shows that there always exists a deterministic greedy policy that is optimal.

Theorem 3.5 (Greedy optimal policy)**.** *For any $s \in \mathcal{S}$, the deterministic greedy policy*

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

is an optimal policy for solving the BOE. Here,

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

where

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

Box 3.4: Proof of Theorem 3.5

While the matrix-vector form of the optimal policy is $\pi^* = \arg\max_\pi(r_\pi + \gamma P_\pi v^*)$, its elementwise form is

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

It is clear that $\sum_{a\in\mathcal{A}} \pi(a|s)q^*(s, a)$ is maximized if $\pi(s)$ selects the action with the greatest $q^*(s, a)$.

The policy in (3.7) is called *greedy* because it seeks the actions with the greatest $q^*(s, a)$. Finally, we discuss two important properties of π^*.

- ◇ Uniqueness of optimal policies: Although the value of v^* is unique, the optimal policy that corresponds to v^* may not be unique. This can be easily verified by counterexamples. For example, the two policies shown in Figure 3.3 are both optimal.
- ◇ Stochasticity of optimal policies: An optimal policy can be either stochastic or deterministic, as demonstrated in Figure 3.3. However, it is certain that there always exists a deterministic optimal policy according to Theorem 3.5.

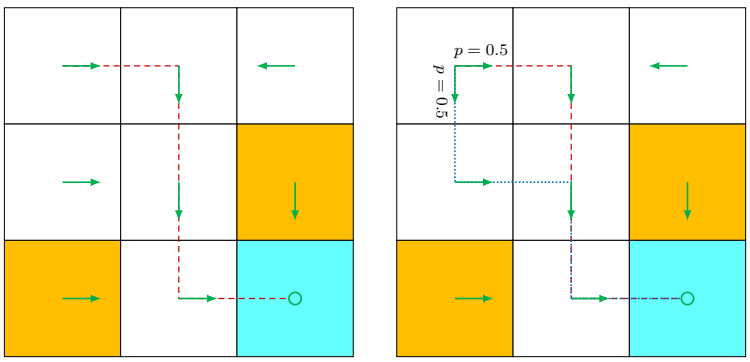


Figure 3.3: Examples for demonstrating that optimal policies may not be unique. The two policies are different but are both optimal.

3.5 Factors that influence optimal policies

The BOE is a powerful tool for analyzing optimal policies. We next apply the BOE to study what factors can influence optimal policies. This question can be easily answered by observing the elementwise expression of the BOE:

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

The optimal state value and optimal policy are determined by the following parameters: 1) the immediate reward r, 2) the discount rate γ, and 3) the system model $p(s'|s,a), p(r|s,a)$. While the system model is fixed, we next discuss how the optimal policy varies when we change the values of r and γ. All the optimal policies presented in this section can be obtained via the algorithm in Theorem 3.3. The implementation details of the algorithm will be given in Chapter 4. The present chapter mainly focuses on the fundamental properties of optimal policies.

A baseline example

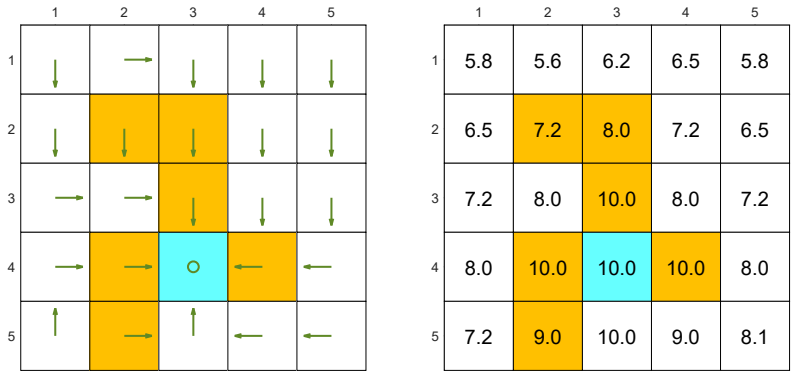
Consider the example in Figure 3.4. The reward settings are $r_{\text{boundary}} = r_{\text{forbidden}} = -1$ and $r_{\text{target}} = 1$. In addition, the agent receives a reward of $r_{\text{other}} = 0$ for every movement step. The discount rate is selected as $\gamma = 0.9$.

With the above parameters, the optimal policy and optimal state values are given in Figure 3.4(a). It is interesting that the agent is not afraid of passing through forbidden areas to reach the target area. More specifically, starting from the state at (row=4, column=1), the agent has two options for reaching the target area. The first option is to avoid all the forbidden areas and travel a long distance to the target area. The second option is to pass through forbidden areas. Although the agent obtains negative rewards when entering forbidden areas, the cumulative reward of the second trajectory is greater than that of the first trajectory. Therefore, the optimal policy is *far-sighted* due to the relatively large value of γ.

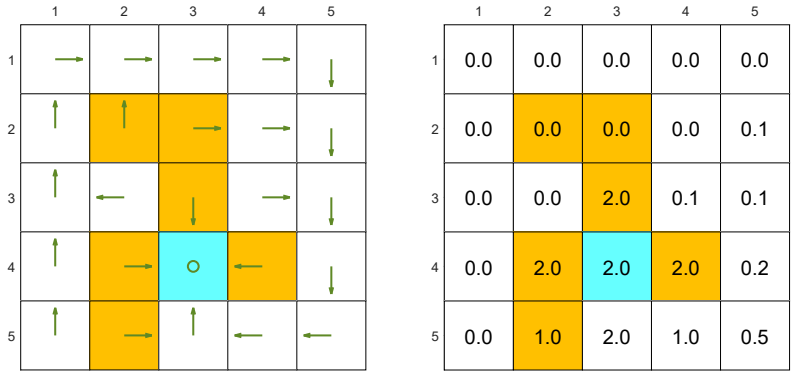
Impact of the discount rate

If we change the discount rate from $\gamma = 0.9$ to $\gamma = 0.5$ and keep other parameters unchanged, the optimal policy becomes the one shown in Figure 3.4(b). It is interesting that the agent does not dare to take risks anymore. Instead, it would travel a long distance to reach the target while avoiding all the forbidden areas. This is because the optimal policy becomes *short-sighted* due to the relatively small value of γ.

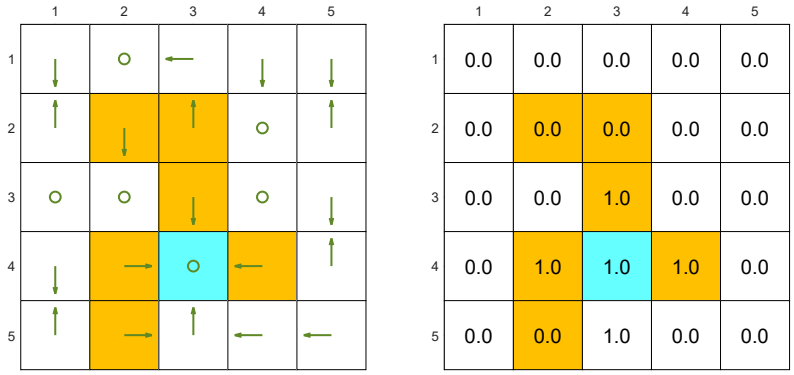
In the extreme case where $\gamma = 0$, the corresponding optimal policy is shown in Figure 3.4(c). In this case, the agent is not able to reach the target area. This is



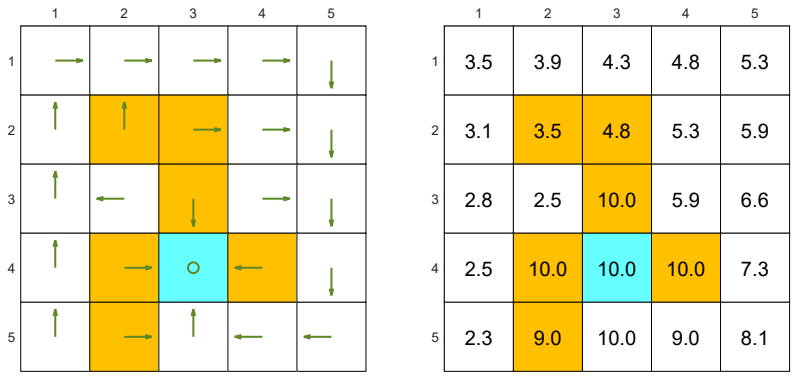
(a) Baseline example: $r_{\text{boundary}} = r_{\text{forbidden}} = -1$, $r_{\text{target}} = 1$, $\gamma = 0.9$.



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



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



(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.

because the optimal policy for each state is *extremely short-sighted* and merely selects the action with the greatest *immediate* reward instead of the greatest *total* reward.

In addition, the spatial distribution of the state values exhibits an interesting pattern: the states close to the target have greater state values, whereas those far away have lower values. This pattern can be observed from all the examples shown in Figure 3.4. It can be explained by using the discount rate: if a state must travel along a longer trajectory to reach the target, its state value is smaller due to the discount rate.

Impact of the reward values

If we want to strictly prohibit the agent from entering any forbidden area, we can increase the punishment received for doing so. For instance, if $r_{\text{forbidden}}$ is changed from -1 to -10, the resulting optimal policy can avoid all the forbidden areas (see Figure 3.4(d)).

However, changing the rewards does not always lead to different optimal policies. One important fact is that optimal policies are *invariant* to affine transformations of the rewards. In other words, if we scale all the rewards or add the same value to all the rewards, the optimal policy remains the same.

Theorem 3.6 (Optimal policy invariance)**.** *Consider a Markov decision process with $v^* \in \mathbb{R}^{|\mathcal{S}|}$ as the optimal state value satisfying $v^* = \max_{\pi\in\Pi}(r_\pi + \gamma P_\pi v^*)$. If every reward $r \in \mathcal{R}$ is changed by an affine transformation to $\alpha r + \beta$, where $\alpha, \beta \in \mathbb{R}$ and $\alpha > 0$, then the corresponding optimal state value v' is also an affine transformation of v^*:*

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

where $\gamma \in (0, 1)$ is the discount rate and $\mathbf{1} = [1, \dots, 1]^T$. Consequently, the optimal policy derived from v' is invariant to the affine transformation of the reward values.

Box 3.5: Proof of Theorem 3.6

For any policy π, define $r_\pi = [\dots, r_\pi(s), \dots]^T$ where

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

If $r \to \alpha r + \beta$, then $r_\pi(s) \to \alpha r_\pi(s) + \beta$ and hence $r_\pi \to \alpha r_\pi + \beta\mathbf{1}$, where $\mathbf{1} = [1, \dots, 1]^T$. In this case, the BOE becomes

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

We next solve the new BOE in (3.9) by showing that $v' = \alpha v^* + c\mathbf{1}$ with $c = \beta/(1-\gamma)$ is a solution of (3.9). In particular, substituting $v' = \alpha v^* + c\mathbf{1}$ into (3.9) gives

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

where the last equality is due to the fact that $P_\pi\mathbf{1} = \mathbf{1}$. The above equation can be reorganized as

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

which is equivalent to

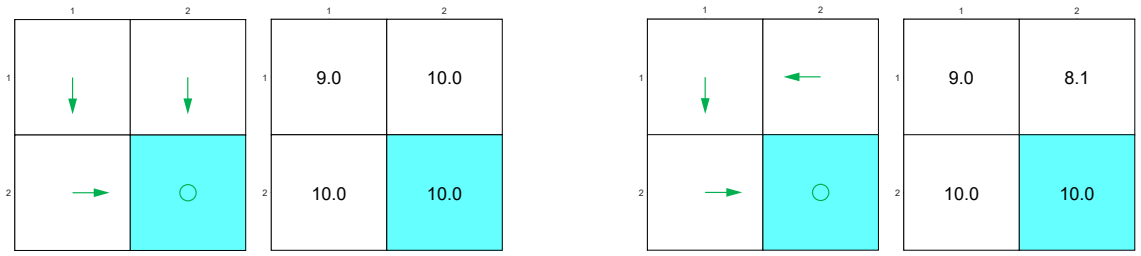
$$\beta\mathbf{1} + c\gamma\mathbf{1} - c\mathbf{1} = 0.$$

Since $c = \beta/(1-\gamma)$, the above equation is valid and hence $v' = \alpha v^* + c\mathbf{1}$ is the solution of (3.9). Since (3.9) is the BOE, v' is also the unique solution. Finally, since v' is an affine transformation of v^*, the relative relationships between the action values remain the same. Hence, the greedy optimal policy derived from v' is the same as that from v^*: $\arg\max_{\pi \in \Pi}(r_\pi + \gamma P_\pi v')$ is the same as $\arg\max_\pi(r_\pi + \gamma P_\pi v^*)$.

Readers may refer to [9] for a further discussion on the conditions under which modifications to the reward values preserve the optimal policy.

Avoiding meaningless detours

In the reward setting, the agent receives a reward of $r_{\text{other}} = 0$ for every movement step (unless it enters a forbidden area or the target area or attempts to go beyond the boundary). Since a zero reward is not a punishment, would the optimal policy take meaningless detours before reaching the target? Should we set r_{other} to be negative to encourage the agent to reach the target as quickly as possible?



(a) Optimal policy (b) Non-optimal policy

Figure 3.5: Examples illustrating that optimal policies do not take meaningless detours due to the discount rate.

Consider the examples in Figure 3.5, where the bottom-right cell is the target area

to reach. The two policies here are the same except for state s_2. By the policy in Figure 3.5(a), the agent moves downward at s_2 and the resulting trajectory is $s_2 \to s_4$. By the policy in Figure 3.5(b), the agent moves leftward and the resulting trajectory is $s_2 \to s_1 \to s_3 \to s_4$.

It is notable that the second policy takes a detour before reaching the target area. If we merely consider the immediate rewards, taking this detour does not matter because no negative immediate rewards will be obtained. However, if we consider the discounted return, then this detour matters. In particular, for the first policy, the discounted return is

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

As a comparison, the discounted return for the second policy is

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

It is clear that the shorter the trajectory is, the greater the return is. Therefore, although the immediate reward of every step does not encourage the agent to approach the target as quickly as possible, the discount rate does encourage it to do so.

A misunderstanding that beginners may have is that adding a negative reward (e.g., -1) on top of the rewards obtained for every movement is necessary to encourage the agent to reach the target as quickly as possible. This is a misunderstanding because adding the same reward on top of all rewards is an affine transformation, which preserves the optimal policy. Moreover, optimal policies do not take meaningless detours due to the discount rate, even though a detour may not receive any immediate negative rewards.

3.6 Summary

The core concepts in this chapter include optimal policies and optimal state values. In particular, a policy is optimal if its state values are greater than or equal to those of any other policy. The state values of an optimal policy are the optimal state values. The BOE is the core tool for analyzing optimal policies and optimal state values. This equation is a nonlinear equation with a nice contraction property. We can apply the contraction mapping theorem to analyze this equation. It was shown that the solutions of the BOE correspond to the optimal state value and optimal policy. This is the reason why we need to study the BOE.

The contents of this chapter are important for thoroughly understanding many fundamental ideas of reinforcement learning. For example, Theorem 3.3 suggests an iterative algorithm for solving the BOE. This algorithm is exactly the value iteration algorithm that will be introduced in Chapter 4. A further discussion about the BOE can be found in [2].

3.7 Q&A

◇ Q: What is the definition of optimal policies?

A: A policy is optimal if its corresponding state values are greater than or equal to any other policy.

It should be noted that this specific definition of optimality is valid only for tabular reinforcement learning algorithms. When the values or policies are approximated by functions, different metrics must be used to define optimal policies. This will become clearer in Chapters 8 and 9.

◇ Q: Why is the Bellman optimality equation important?

A: It is important because it characterizes both optimal policies and optimal state values. Solving this equation yields an optimal policy and the corresponding optimal state value.

◇ Q: Is the Bellman optimality equation a Bellman equation?

A: Yes. The Bellman optimality equation is a special Bellman equation whose corresponding policy is optimal.

◇ Q: Is the solution of the Bellman optimality equation unique?

A: The Bellman optimality equation has two unknown variables. The first unknown variable is a value, and the second is a policy. The value solution, which is the optimal state value, is unique. The policy solution, which is an optimal policy, may not be unique.

◇ Q: What is the key property of the Bellman optimality equation for analyzing its solution?

A: The key property is that the right-hand side of the Bellman optimality equation is a contraction mapping. As a result, we can apply the contraction mapping theorem to analyze its solution.

◇ Q: Do optimal policies exist?

A: Yes. Optimal policies always exist according to the analysis of the BOE.

◇ Q: Are optimal policies unique?

A: No. There may exist multiple or infinite optimal policies that have the same optimal state values.

◇ Q: Are optimal policies stochastic or deterministic?

A: An optimal policy can be either deterministic or stochastic. A nice fact is that there always exist deterministic greedy optimal policies.

◇ Q: How to obtain an optimal policy?

A: Solving the BOE using the iterative algorithm suggested by Theorem 3.3 yields an optimal policy. The detailed implementation of this iterative algorithm will be given in Chapter 4. Notably, all the reinforcement learning algorithms introduced in this book aim to obtain optimal policies under different settings.

◇ Q: What is the general impact on the optimal policies if we reduce the value of the discount rate?

A: The optimal policy becomes more short-sighted when we reduce the discount rate. That is, the agent does not dare to take risks even though it may obtain greater cumulative rewards afterward.

◇ Q: What happens if we set the discount rate to zero?

A: The resulting optimal policy would become extremely short-sighted. The agent would take the action with the greatest immediate reward, even though that action is not good in the long run.

◇ Q: If we increase all the rewards by the same amount, will the optimal state value change? Will the optimal policy change?

A: Increasing all the rewards by the same amount is an affine transformation of the rewards, which would not affect the optimal policies. However, the optimal state value would increase, as shown in (3.8).

◇ Q: If we hope that the optimal policy can avoid meaningless detours before reaching the target, should we add a negative reward to every step so that the agent reaches the target as quickly as possible?

A: First, introducing an additional negative reward to every step is an affine transformation of the rewards, which does not change the optimal policy. Second, the discount rate can automatically encourage the agent to reach the target as quickly as possible. This is because meaningless detours would increase the trajectory length and reduce the discounted return.