Chapter 3

Optimal State Value and Bellman Optimality Equation

The ultimate goal of reinforcement learning (RL) is to seek *optimal policies*. It is, therefore, important to define "optimal". In this chapter, we will introduce a core concept and an important tool. The core concept is 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 as well as optimal policies. The contents of this chapter are fundamentally important for understanding the model-based RL algorithms that will be introduced in the next chapter. Be prepared that this chapter is a little mathematically intensive. However, it is worth because many fundamental and algorithmic questions will be clearly answered.

3.1 Motivating example: how to improve policies?

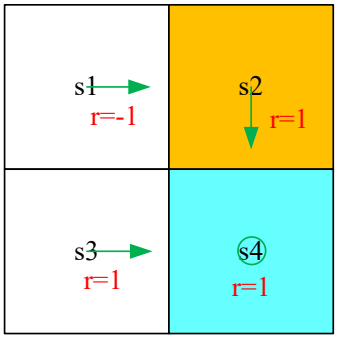


Figure 3.1: An example to demonstrate policy improvement.

Consider the example shown in Figure 3.1, where the target state is s_4. This policy is not good, because the policy at s_1 suggests moving rightwards to the forbidden area. How can we improve the given policy to make it better? The answer lies in 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 calculated that

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

Second, we calculate the action values for state s_1:

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

It is notable that action a_3 (moving downwards) has the greatest action value. That is

$$q_\pi(s_1, a_3) \geq q_\pi(s_1, a_i), \quad \text{for all } i \in \{1, 2, 3, 4, 5\}.$$

Therefore, if we update the policy so that it selects a_3 at s_1, then the updated policy becomes better since moving downwards at s_1 can avoid the forbidden area.

This example illustrates that, if we update the policy to select the action with the *greatest action value*, we could find a better policy. This is the basic idea of many RL algorithms. Of course, this example is very simple in the sense that the given policy is not good only 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 exist optimal or best policies? What does an optimal policy look like? We will answer all of these questions in this chapter.

3.2 Optimal state value and optimal policy

The goal of reinforcement learning is to find out optimal policies. It is, therefore, important to define what optimal policy is. While state values can be used to evaluate policies, they actually can also be used to define optimal policies. In particular, consider two given

policies π_1 and π_2. If

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

π_1 is said "better" than π_2. That is the state value of π_1 is greater than or equal to that of π_2 for any state. Furthermore, if a policy is better than all the other possible policies, then this policy is optimal. This idea 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.*

This definition indicates that an optimal policy has the greatest state value for every state compared to all the other policies. The 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?

Be patient. We will answer these questions one by one in the rest of the chapter.

3.3 Bellman optimality equation

In order to understand optimal state values and optimal policies, we need to explore a special Bellman equation called *Bellman optimality equation* (BOE). We first directly present this equation and then analyze why it describes optimal state values and optimal policies.

The BOE is

$$\begin{aligned} v(s) &= \max_\pi \sum_a \pi(a|s) \left(\sum_r p(r|s,a)r + \gamma \sum_{s'} p(s'|s,a)v(s') \right), \\ &= \max_\pi \sum_a \pi(a|s) q(s,a), \quad \text{for all } s \in \mathcal{S}. \end{aligned} \tag{3.1}$$

where $v(s), v(s')$ are unknowns and

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

The BOE is an elegant and powerful tool for analyzing optimal policies. However, it looks tricky at first glance. For example, there are two unknowns $v(s)$ and $\pi(a|s)$ in the equation. It may be confusing how to solve unknowns from one equation. Also, although this equation is a Bellman equation, it is nontrivial to see that since it involves

a maximization problem on the right-hand side. We also need to answer the following fundamental questions about the BOE.

– Existence: does this equation have solutions?

– Uniqueness: is the solution unique?

– Algorithm: how to solve this equation?

– Optimality: how is the solution related to optimal policy?

Once we can answer these questions, we will understand optimal state values and optimal policies clearly. At this moment, it is better to temporarily forget about the RL interpretations of the symbols in the BOE and study the equation from a pure algebraic perspective. The rest of the section is important but, in the meantime, a little mathematically intensive. The readers are suggested to read selectively.

3.3.1 Maximization on the right-hand side of the BOE

First of all, we clarify how to solve the maximization problem on the right-hand side of the BOE. At first glance, it may be confusing to beginners how can we solve two unknowns $v(s)$ and $\pi(a|s)$ from one equation. In fact, the two unknowns can be solved one by one. This idea is illustrated by the following simple example.

Example 3.1. *Consider two variables $x, a \in \mathbb{R}$. Suppose they satisfy*

$$x = \max_a (2x - 1 - a^2).$$

This equation has two unknowns. To solve them, first consider the right hand side. Regardless the value of x, $\max_a(2x-1-a^2) = 2x-1$ where the maximization is achieved when $a = 0$. Second, when $a = 0$, the equation becomes $x = 2x - 1$, which leads to $x = 1$. Therefore, $a = 0$ and $x = 1$ are the solution of the equation.

From the above example, we know that we can first fix one variable and solve the maximization problem. We now come back to the maximization problem on the right-hand side of the BOE. The BOE in (3.1) can be written in short as

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

Inspired by Example 3.1, we can first fix $q(s,a)$ for all $a \in \mathcal{A}(s)$ and then find the optimal π. How to do that? The following example can demonstrate the basic idea.

Example 3.2. *Suppose $x_1, x_2, x_3 \in \mathbb{R}$ are given. Find c_1^*, c_2^*, c_3^* solving*

$$\max_{c_1,c_2,c_3} c_1 x_1 + c_2 x_2 + c_3 x_3.$$

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

Without loss of generality, suppose $x_3 \geq x_1, x_2$. *Then, the optimal solution is* $c_3^* = 1$ *and* $c_1^* = c_2^* = 0$. *That is because for any* c_1, c_2, c_3

$$x_3 = (c_1 + c_2 + c_3)x_3 = c_1 x_3 + c_2 x_3 + c_3 x_3 \geq c_1 x_1 + c_2 x_2 + c_3 x_3.$$

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

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

where the equality is achieved when

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

where $a^* = \arg\max_a q(s,a)$. Therefore, when $\{q(s,a)\}_{a \in \mathcal{A}(s)}$ are fixed, the optimal policy $\pi(s)$ is to choose the action that corresponds to the greatest value of $q(s,a)$.

3.3.2 Matrix-vector form of the BOE

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

The matrix-vector form of the BOE is

$$v = \max_\pi (r_\pi + \gamma P_\pi v),$$

where $v \in \mathbb{R}^{|\mathcal{S}|}$. The structure of r_π and P_π are the same as those in the matrix-vector form of a normal Bellman equation:

$$[r_\pi]_s \triangleq \sum_a \pi(a|s) \sum_r p(r|s,a) r, \qquad [P_\pi]_{s,s'} = p(s'|s) \triangleq \sum_a \pi(a|s) \sum_{s'} p(s'|s,a).$$

Here, $\max_\pi$ is performed elementwise. Furthermore, denote the right hand side as

$$f(v) \doteq \max_\pi (r_\pi + \gamma P_\pi v).$$

Then, the BOE becomes

$$v = f(v). \tag{3.2}$$

Therefore, the BOE is expressed as a simple nonlinear equation of v. In the rest, we show how to solve this nonlinear equation and what kind of properties it has.

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, which is a useful tool to analyze general nonlinear equations. This theorem is also known as the fixed-point theorem. Readers who already know this theorem can skip this part.

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 is that the map of x^* is itself. That is why it 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 $x_1, x_2 \in \mathbb{R}$. In this book, $\|\cdot\|$ denotes a vector or matrix norm.

Example 3.3. *Here are three examples to demonstrate the concept of fixed point and contraction mapping.*

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

  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 *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 the fixed point and the contraction property is given in the following classic theorem.

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

– *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 tells if the solution of a nonlinear equation exists but also suggests a numerical algorithm solving the equation. The proof of the theorem is given below. Readers who are not interested can skip the proof.

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

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

given any initial guess x_0.

Proof of the contraction mapping theorem

Part 1: the convergence of $\{x_k = f(x_{k-1})\}_{k=1}^{\infty}$.

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 exist N such that $\|x_m - x_n\| < \varepsilon$ for all $m, n > N$. The intuitive interpretation is that, in a Cauchy sequence, there exists a finite integer N such that all the elements after N are sufficiently close to each other. Cauchy sequences are important because a Cauchy sequence will converge to a limit. Its convergence property will be used to prove the contraction mapping theorem. It is worth mentioning 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 it 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 $\|x_k - x_{k-1}\| \leq \gamma\|x_{k-1} - x_{k-2}\|$, $\ldots$, $\|x_2 - x_1\| \leq \gamma\|x_1 - x_0\|$. As a result, 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. It is worth mentioning that the convergence of $\{\|x_{k+1} - x_k\|\}$ is not sufficient to imply the convergence of $\{x_k\}$. Therefore, we need to further consider $\|x_m - x_n\|$ for any $m > n$. In particular,

$$\begin{aligned}\|x_m - x_n\| &= \|x_m - x_{m-1} + x_{m-1} - \cdots - x_{n+1} + x_{n+1} - x_n\| \\ &\leq \|x_m - x_{m-1}\| + \cdots + \|x_{n+1} - x_n\| \\ &\leq \gamma^{m-1}\|x_1 - x_0\| + \cdots + \gamma^n\|x_1 - x_0\| \\ &= \gamma^n(\gamma^{m-1-n} + \cdots + 1)\|x_1 - x_0\| \\ &\leq \gamma^n(1 + \cdots + \gamma^{m-1-n} + \gamma^{m-n} + \gamma^{m-n+1} + \ldots)\|x_1 - x_0\| \\ &= \frac{\gamma^n}{1-\gamma}\|x_1 - x_0\|. \end{aligned} \tag{3.3}$$

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 it 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 see that, since

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

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

Part 3: we show that the fixed point is unique. To see that, suppose 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 (3.3). 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. The smaller γ is, the faster the convergence is.

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

We next show that $f(v)$ in the BOE (3.2) is a contraction mapping. Then, the contraction mapping theorem introduced in the last subsection can be applied to solve the BOE immediately.

Theorem 3.2 (Contraction property of $f(v)$)**.** *The function $f(v)$ in the BOE is a contraction mapping satisfying*

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

where $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$ are any two vectors and $\gamma \in (0, 1)$ is the discount rate.

Thanks to the contraction mapping property, the BOE can be analyzed by the contraction mapping theorem. The details of the analysis will be given in the next section. The proof of the contraction property of the BOE is given below.

Proof of Theorem 3.2

Consider any two vectors $v_1, v_2 \in \mathbb{R}^{|\mathcal{S}|}$, suppose $\pi_1^* \doteq \arg\max_\pi(r_\pi + \gamma P_\pi v_1)$ 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 elementwise. 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 $f(v_2) - f(v_1) \leq \gamma P_{\pi_2^*}(v_2 - v_1)$, which implies $f(v_1) - f(v_2) \geq \gamma P_{\pi_2^*}(v_1 - v_2)$. Therefore,

$$\gamma P_{\pi_2^*}(v_1 - v_2) \leq f(v_1) - f(v_2) \leq \gamma P_{\pi_1^*}(v_1 - v_2). \tag{3.4}$$

Define

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

By definition, $z \geq 0$. Here, $\max(\cdot)$, $|\cdot|$, and $\geq$ are all elementwise. 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

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

Since $|f(v_1) - f(v_2)| \leq z$ elementwise, it follows that

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

where $\|\cdot\|_\infty$ is the sup-norm, which is the maximum absolute value of the elements of a vector. Here, the inequality is still valid if the sup-norm is replaced by other norms such as $\|\cdot\|_2$ or $\|\cdot\|_1$.

On the other hand, suppose 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 the elements nonnegative 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 contraction property of $f(v)$.

The proof of the contraction property also suggests another useful fact. That is, $f(v)$ is monotonically non-decreasing.

Corollary 3.1 (Monotone property of $f(v)$)**.** *For any $v_1, v_2 \in \mathbb{R}^n$, $f(v_1) \geq f(v_2)$ if $v_1 \geq v_2$, and $f(v_1) \leq f(v_2)$ if $v_1 \leq v_2$, where $\geq$ and $\leq$ are elementwise.*

Proof of Corollary 3.1

If $v_1 \geq v_2$, then $P_{\pi_2^*}(v_1 - v_2) \geq 0$ because $P_{\pi_2^*}$ is a nonnegative matrix. Thus, (3.4) implies that

$$0 \leq \gamma P_{\pi_2^*}(v_1 - v_2) \leq f(v_1) - f(v_2) \leq \gamma P_{\pi_1^*}(v_1 - v_2)$$

and hence $f(v_1) \geq f(v_2)$. Similarly, if $v_1 \leq v_2$, (3.4) implies that

$$\gamma P_{\pi_2^*}(v_1 - v_2) \leq f(v_1) - f(v_2) \leq \gamma P_{\pi_1^*}(v_1 - v_2) \leq 0$$

and hence $f(v_1) \leq f(v_2)$.

The monotone property of $f(v)$ will be used frequently later in the analysis of policy optimality. It should be noted that the converse of Corollary 3.1 is not true. That is, $f(v_1) \geq f(v_2)$ may not imply $v_1 \geq v_2$, because it is possible that some elements of v_1 may be greater than their counterparts in v_2 while the others may be less.

3.4 Solutions of the BOE

With the preparation in the last section, we are ready now to solve the BOE. There are two unknowns in the BOE: v and π. We solve the two unknowns one by one as follows.

First, we solve the unknown v in the BOE.

Suppose v^* is a solution to $v = f(v) = \max_\pi(r_\pi + \gamma P_\pi v)$. Then, it holds that

$$v^* = \max_\pi(r_\pi + \gamma P_\pi v^*).$$

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 $v = f(v) = \max_\pi(r_\pi + \gamma P_\pi v)$, there always exists a unique solution v^*, which can be solved iteratively by*

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

The sequence $\{v_k\}$ *converges to* v^* *exponentially fast given any initial guess* v_0.

Proof. Since $f(v)$ is a contraction mapping as proven in Theorem 3.2, all the results here follow directly from the contraction mapping theorem. □

This theorem answers some fundamental questions.

– Existence: The solution v^* to the BOE always exists.

– Uniqueness: The solution v^* is always unique.

– Algorithm: v^* can be solved iteratively by the iterative algorithm in Theorem 3.3.

The iterative algorithm given in Theorem 3.3 is an important RL algorithm. This algorithm, which has a name called *value iteration*, will be studied in detail in the next chapter. In this chapter, we merely treat this algorithm as a numerical algorithm solving the BOE and focus on the fundamental properties of the BOE.

Second, we solve the unknown π in the BOE. Suppose v^* has been obtained and

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

Then, v^* and π^* satisfy

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

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

Up to now, we merely know that v^* is the solution to the BOE. Whether it is the optimal state value is still unclear. The following theorem shows that v^* is the optimal state value and π^* is an optimal policy.

Theorem 3.4 (Optimality)**.** *For any policy* π*, it holds that*

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

where v_π *is the state value of* π*.*

Proof of optimality

Since

$$\begin{aligned} v_\pi &= r_\pi + \gamma P_\pi v_\pi, \\ v^* &= \max_{\pi}(r_\pi + \gamma P_\pi v^*) = r_{\pi^*} + \gamma P_{\pi^*} v^* \geq r_\pi + \gamma P_\pi v^*, \end{aligned}$$

we have

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

Using the above inequality recursively 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 due to $\gamma < 1$ and P_π^n is a nonnegative matrix with all elements no greater than 1 (because $P_\pi^n \mathbf{1} = \mathbf{1}$).

Finally, we know that the two unknowns v and π in the BOE correspond to the optimal state value and optimal policy, respectively. This is the reason why it is important to study the BOE.

What does an optimal policy look like? The following theorem shows that a deterministic greedy policy 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 solving the BOE. Here,

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

where

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

Proof. The matrix-vector form of the optimal policy is $\pi^* = \arg\max_\pi (r_\pi + \gamma P_\pi v^*)$. Its elementwise form is

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

It is clear that $\sum_a \pi(a|s)q^*(s, a)$ is maximized if $\pi(s)$ selects the action with the greatest value of $q^*(s, a)$. □

The policy in (3.7) is called *greedy*, because it seeks the actions with the greatest $q^*(s, a)$.

Finally, we discuss the last two questions about the BOE.

– Uniqueness of optimal policies: is the optimal policy π^* unique? Although the value of v^* is unique, the optimal policy corresponding to v^* may not be unique. It can be easily verified by counterexamples. For example, the two policies shown in Figure 3.2 are both optimal.

– Stochasticity of optimal policies: is the optimal policy stochastic or deterministic? The optimal policy can be either stochastic or deterministic as demonstrated in Figure 3.2. However, it is certain that there always exists a deterministic optimal policy as shown in Theorem 3.5.

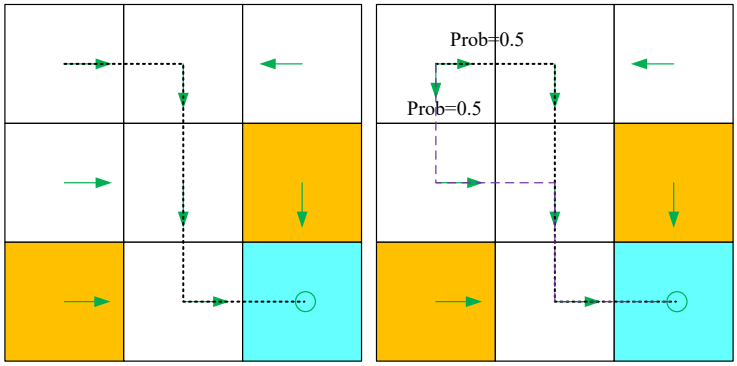


Figure 3.2: Examples to demonstrate that optimal policies may not unique.

3.5 Factors that influence optimal policies

The BOE is a powerful tool to analyze optimal policies. We next further use the BOE to study what kind of factors can influence optimal policies. This question can be answered by observing the elementwise form of the BOE:

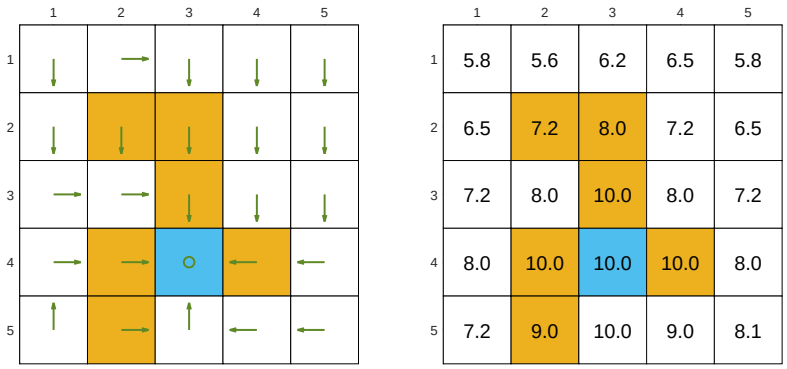
$$v(s) = \max_{\pi} \sum_{a} \pi(a|s) \left(\sum_{r} p(r|s,a) r + \gamma \sum_{s'} p(s'|s,a) v(s') \right), \quad s \in \mathcal{S}.$$

The parameters in this equation include the immediate reward r, the discount rate γ, and the system model $p(s'|s,a), p(r|s,a)$. While the system model is fixed in general, we next discuss how the optimal policy changes if we change the values of r and γ. All the optimal policies presented in this section can be obtained by the algorithm in Theorem 3.3. The details of the algorithm are given in the next chapter. This chapter mainly focuses on fundamental problems rather than algorithms.

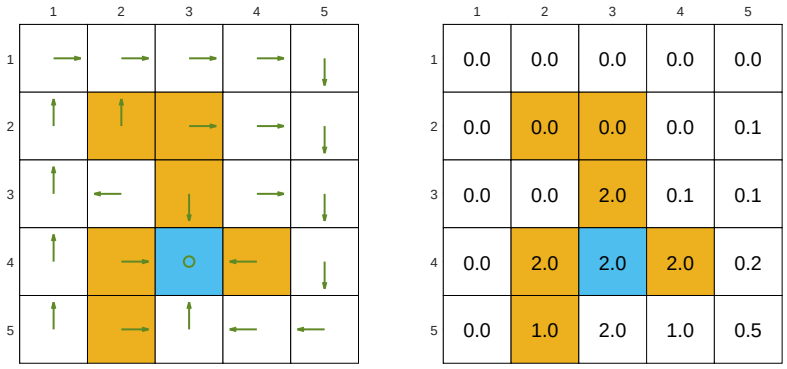
A baseline example

Consider the example in Figure 3.3. The reward setting is $r_{\text{boundary}} = r_{\text{forbidden}} = -1$ and $r_{\text{target}} = 1$. Besides, the agent receives a reward of $r = 0$ for every step of movements. The discount rate is selected as $\gamma = 0.9$.

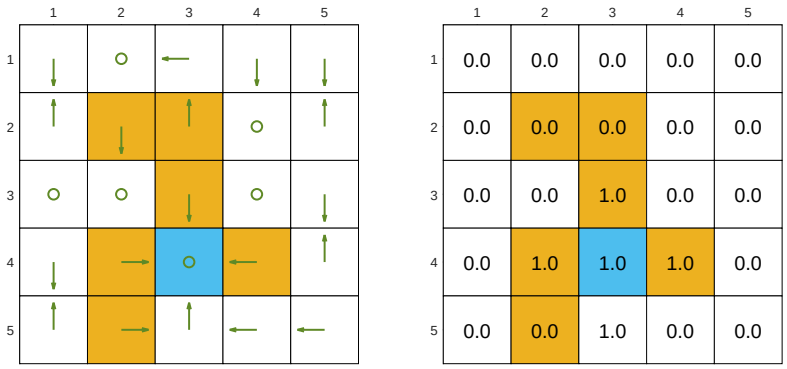
With the above parameters, the optimal policy and optimal state values are given in Figure 3.3(a). It is interesting to note that the agent is not afraid of passing through forbidden areas to reach the target area. For example, starting from the state at (row=4, column=1), the agent has two options to reach the target area. The first is to avoid all



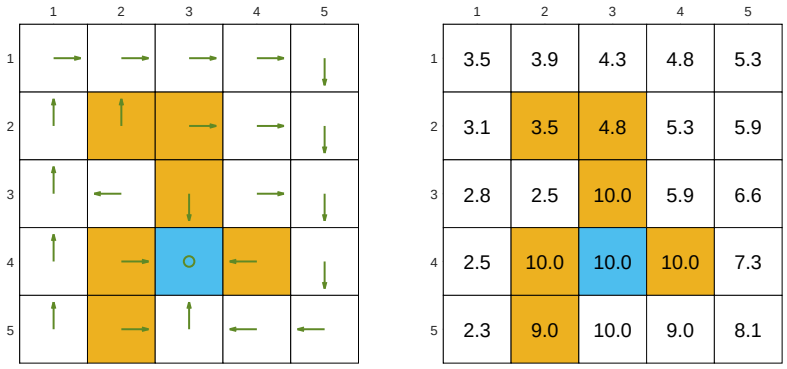
(a) Baseline example: $r_{\text{boundary}} = r_{\text{forbidden}} = -1$, $r_{\text{target}} = 1$, $\gamma = 0.9$



(b) The discount rate is $\gamma = 0.5$. Other parameters are the same as (a).



(c) The discount rate is $\gamma = 0$. Other parameters are the same as (a).



(d) $r_{\text{forbidden}} = -10$. Other parameters are the same as (a).

Figure 3.3: Optimal policies and optimal state values given different values of parameters.

the forbidden areas and travel a long distance to the target area. The second is to pass through the forbidden areas. Although the agent gets negative rewards when entering forbidden areas, the overall cumulative reward of the second trajectory is greater than that of the first trajectory.

The impact of discount rate

If we change the discount rate from $\gamma = 0.9$ to $\gamma = 0.5$ and keep other parameters unchanged, the optimal policy is shown in Figure 3.3(b). It is interesting to see that the agent would not take risks in this case. It instead avoids all the forbidden areas to reach the target. That is simply because the agent becomes short-sighted. It would not take any risk to get punished even though it can get greater rewards afterward.

In the extreme case where we set $\gamma = 0$, the optimal policy is given in Figure 3.3(c). In this case, the agent will not be able to reach the target area. That is because the optimal policy for each state is extremely short-sighted and merely selects the action that has the greatest immediate reward (instead of the greatest return). As a result, only those states that are adjacent to the target have nonzero state values and can select correct actions. Those states that are not adjacent to the target randomly select an action that has the greatest rewards.

The impact of reward signals

If we want to strictly prohibit the agent from entering any forbidden area, we can increase the punishment. For instance, if the reward of entering a forbidden area is changed from -1 to -10, the optimal policy in this case would avoid all the forbidden areas (see Figure 3.3(d)).

Changing rewards does not always lead to different optimal policies. One important fact is that optimal policies are *invariant* to affine transformations of rewards. In other words, if we scale up or down all the rewards or add the same number to all the rewards, the optimal policy would remain 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 $v^* = \max_\pi(r_\pi + \gamma P_\pi v^*)$. If every reward r is changed by an affine transformation to $ar + b$, where $a, b \in \mathbb{R}$ and $a > 0$, then the corresponding optimal state value v' is also an affine transformation of v^*:*

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

where $\gamma \in (0, 1)$ is the discount rate and $\mathbf{1} = [1, \ldots, 1]^T$. Consequently, the optimal policies are invariant to the affine transformation of the reward signals.

Proof of the optimal policy invariance theorem

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

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

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

$$v' = \max_\pi (ar_\pi + b\mathbf{1} + \gamma P_\pi v'). \tag{3.9}$$

We next solve the new BOE in (3.9). To do that, we verify that $v' = av^* + k\mathbf{1}$ with $k = b/(1-\gamma)$ is the solution of (3.9). In particular, substituting $v' = av^* + k\mathbf{1}$ into (3.9) gives

$$av^* + k\mathbf{1} = \max_\pi [ar_\pi + b\mathbf{1} + \gamma P_\pi (av^* + k\mathbf{1})] = \max_\pi (ar_\pi + b\mathbf{1} + a\gamma P_\pi v^* + k\gamma\mathbf{1}),$$

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

$$av^* = \max_\pi (ar_\pi + a\gamma P_\pi v^*) + b\mathbf{1} + k\gamma\mathbf{1} - k\mathbf{1},$$

which is equivalent to

$$b\mathbf{1} + k\gamma\mathbf{1} - k\mathbf{1} = 0.$$

Since $k = b/(1-\gamma)$, the above equation is valid and hence $v' = av^* + k\mathbf{1}$ is the solution to (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 relationship among the action values remain the same. Hence, v' would lead to the same optimal policies as v^*.

Designing appropriate rewards in RL is of practical challenge. Interested readers may see more discussion in [5].

Avoiding meaningless detour

In the reward setting, the agent would receive a reward of $r = 0$ for every step unless it enters a forbidden or the target area or attempts to get out of the boundary. Since a zero reward is not punishment, would the optimal policy take meaningless detours before reaching the target?

To be specific, consider the examples in Figure 3.4, where the target cell is the right bottom one. The two policies here are the same except for state s_2. In particular, the agent would move downwards at s_2 following the first policy in Figure 3.4(a) and

move downwards following the second policy in Figure 3.4(b). As a result, the resulting trajectory is $s_2 \to s_4$ by the first policy and $s_2 \to s_1 \to s_3 \to s_4$ by the second policy. It is noted that the second policy takes a detour before reaching the target area.

Does this detour matter? If we merely consider the immediate rewards, taking this detour would 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

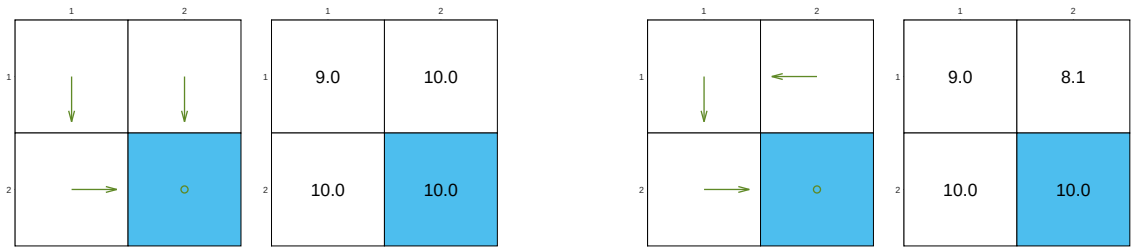
$$\text{return} = 1 + \gamma 1 + \gamma^2 1 + \cdots = 1/(1-\gamma) = 10.$$

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.

A misunderstanding that beginners may have is that adding a negative reward (say -1) on top of the rewards for every move is necessary to encourage the agent to reach the target as quickly as possible. This is a mistake because adding the same amount of reward on top of all rewards is an affine transformation. We have shown that affine transformation of the rewards would not change the optimal policies. As we analyzed above, due to the discount rate, optimal policies would not take meaningless detours before reaching the target even though a detour may not get any negative immediate rewards.



(a) Optimal policy (b) Not optimal

Figure 3.4: Examples illustrating that optimal policies do not take meaningless detours in the presence of discount rate.

An interesting pattern of the spatial distribution of the state values is that the states close to the target have greater state values, whereas those far away have less values. This pattern can be observed from all the examples shown in Figure 3.3. This interesting pattern can also be explained by using the discount rate. That is, if a state has to travel a longer trajectory to the target, its state value would be less due to the discount rate.

3.6 Summary

The core concept in this chapter is optimal policy, which is defined based on optimal state values. In particular, a policy is optimal if its state values are greater than or equal to those of all the other policies. In order to analyze optimal policies, we have to study the BOE. This equation is a nonlinear equation with a nice contraction property. Based on this property, we can apply the contraction mapping theorem to solve this equation. We proved that the solution to the BOE always exists and is unique. Its solution is the optimal state value, which is the greatest state value that can be achieved by any policy. The corresponding optimal policies may not be unique. While optimal policies may be either stochastic or deterministic, a nice property is that there always exist deterministic greedy optimal policies.

The contents in this chapter are important for thoroughly understanding many fundamental ideas of RL. For example, Theorem 3.3 suggests an iterative algorithm for solving the BOE. This algorithm is exactly the value iteration algorithm introduced in the next chapter.

3.7 Q&A

– Q: What is the definition of optimal policy?

  A: A policy is optimal if the corresponding state values are greater than or equal to all the other policies. It should be noted that this specific definition of optimality is valid only for value-based RL algorithms. When policies are approximated by functions, different metrics will be used to define optimal policies. Details will be given when we introduce policy gradient algorithms.

– Q: Do optimal policies exist?

  A: Yes. There always exist optimal policies according to the BOE.

– Q: Are optimal policies unique?

  A: No. There may exist multiple or infinite optimal policies which 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 is one way. The value iteration algorithm as introduced in the next chapter is an algorithm solving the BOE. Of course, many other RL algorithms can also obtain optimal policies as shown later in this book.

– Q: Why is the Bellman optimality equation important?

  A: It is important to study the BOE because it characterizes both optimal policies and optimal state values.

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

  A: Yes. The Bellman optimality equation is a special Bellman equation. The corresponding policy is optimal. The corresponding state value is the optimal state value.

– Q: Is the solution to the Bellman optimality equation unique?

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

– Q: What is the key property of the Bellman optimality equation for us 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 or called the fixed-point theorem to analyze its solution.

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

  A: The optimal policy will become short-sighted when we reduce the discount rate. That is, the agent does not dare to take high risks even though they may get greater cumulative rewards afterward.

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

  A: The agent would become extremely short-sighted. The optimal policy would be purely based on immediate rewards. That is the agent would take the action that has 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 will 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 would reach the target as quickly as possible?

  A: No, it is useless to introduce an additional negative reward to every step. Such an operation is an affine transformation of the rewards, which would not change the optimal policies. In fact, the discount rate can do the job to encourage the agent to reach the target as soon as possible. That is because meaningless detours would increase the length of the trajectory and hence reduce the discounted return.