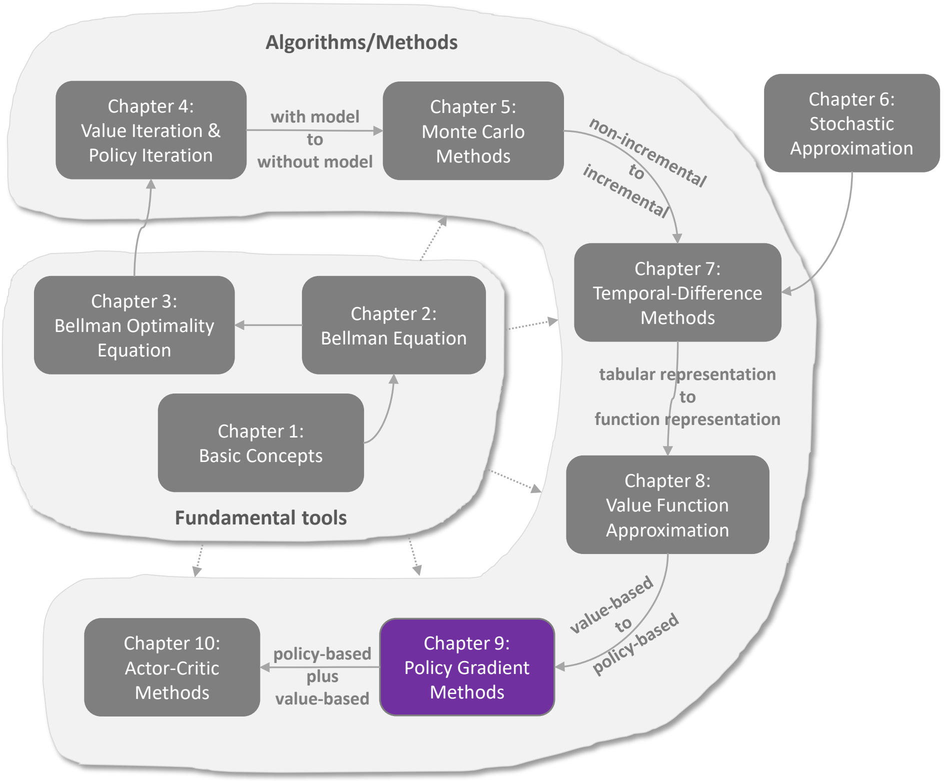
Chapter 9

Policy Gradient Methods



Algorithms/Methods
Chapter 4: Value Iteration & Policy Iteration
with model to without model
Chapter 5: Monte Carlo Methods
non-incremental to incremental
Chapter 6: Stochastic Approximation
Chapter 7: Temporal-Difference Methods
Chapter 3: Bellman Optimality Equation
Chapter 2: Bellman Equation
Chapter 1: Basic Concepts
Fundamental tools
tabular representation to function representation
Chapter 8: Value Function Approximation
value-based to policy-based
Chapter 10: Actor-Critic Methods
policy-based plus value-based
Chapter 9: Policy Gradient Methods

Figure 9.1: Where we are in this book.

The idea of function approximation can be applied not only to represent state/action values, as introduced in Chapter 8, but also to represent policies, as introduced in this chapter. So far in this book, policies have been represented by tables: the action probabilities of all states are stored in a table (e.g., Table 9.1). In this chapter, we show that policies can be represented by parameterized functions denoted as $\pi(a|s,\theta)$, where $\theta \in \mathbb{R}^m$ is a parameter vector. It can also be written in other forms such as $\pi_\theta(a|s)$, $\pi_\theta(a,s)$, or $\pi(a,s,\theta)$.

When policies are represented as functions, optimal policies can be obtained by optimizing certain scalar metrics. Such a method is called *policy gradient*. The policy

	a_1	a_2	a_3	a_4	a_5
s_1	$\pi(a_1\|s_1)$	$\pi(a_2\|s_1)$	$\pi(a_3\|s_1)$	$\pi(a_4\|s_1)$	$\pi(a_5\|s_1)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s_9	$\pi(a_1\|s_9)$	$\pi(a_2\|s_9)$	$\pi(a_3\|s_9)$	$\pi(a_4\|s_9)$	$\pi(a_5\|s_9)$

Table 9.1: A tabular representation of a policy. There are nine states and five actions for each state.

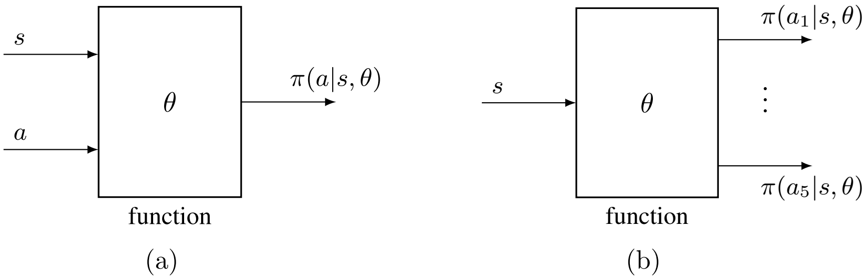


Figure 9.2: Function representations of policies. The functions may have different structures.

gradient method is a big step forward in this book because it is *policy-based*. By contrast, all the previous chapters in this book discuss *value-based* methods. The advantages of the policy gradient method are numerous. For example, it is more efficient for handling large state/action spaces. It has stronger generalization abilities and hence is more efficient in terms of sample usage.

9.1 Policy representation: From table to function

When the representation of a policy is switched from a table to a function, it is necessary to clarify the difference between the two representation methods.

- First, how to define optimal policies? When represented as a table, a policy is defined as optimal if it can maximize *every state value*. When represented by a function, a policy is defined as optimal if it can maximize certain *scalar metrics*.
- Second, how to update a policy? When represented by a table, a policy can be updated by directly changing the entries in the table. When represented by a parameterized function, a policy can no longer be updated in this way. Instead, it can only be updated by changing the parameter θ.
- Third, how to retrieve the probability of an action? In the tabular case, the probability of an action can be directly obtained by looking up the corresponding entry in the table. In the case of function representation, we need to input (s, a) into the function to calculate its probability (see Figure 9.2(a)). Depending on the structure of the function, we can also input a state and then output the probabilities of all actions (see Figure 9.2(b)).

The basic idea of the policy gradient method is summarized below. Suppose that $J(\theta)$ is a scalar metric. Optimal policies can be obtained by optimizing this metric via the gradient-based algorithm:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta J(\theta_t),$$

where $\nabla_\theta J$ is the gradient of J with respect to θ, t is the time step, and α is the optimization rate.

With this basic idea, we will answer the following three questions in the remainder of this chapter.

- What metrics should be used? (Section 9.2).
- How to calculate the gradients of the metrics? (Section 9.3)
- How to use experience samples to calculate the gradients? (Section 9.4)

9.2 Metrics for defining optimal policies

If a policy is represented by a function, there are two types of metrics for defining optimal policies. One is based on state values and the other is based on immediate rewards.

Metric 1: Average state value

The first metric is the *average state value* or simply called the *average value*. It is defined as

$$\bar{v}_\pi = \sum_{s \in \mathcal{S}} d(s) v_\pi(s),$$

where $d(s)$ is the weight of state s. It satisfies $d(s) \geq 0$ for any $s \in \mathcal{S}$ and $\sum_{s \in \mathcal{S}} d(s) = 1$. Therefore, we can interpret $d(s)$ as a probability distribution of s. Then, the metric can be written as

$$\bar{v}_\pi = \mathbb{E}_{S \sim d}[v_\pi(S)].$$

How to select the distribution d? This is an important question. There are two cases.

- The first and simplest case is that d is *independent* of the policy π. In this case, we specifically denote d as d_0 and $\bar{v}_\pi$ as $\bar{v}_\pi^0$ to indicate that the distribution is independent of the policy. One case is to treat all the states equally important and select $d_0(s) = 1/|\mathcal{S}|$. Another case is when we are only interested in a specific state s_0 (e.g., the agent always starts from s_0). In this case, we can design

$$d_0(s_0) = 1, \quad d_0(s \neq s_0) = 0.$$

- ◇ The second case is that d is *dependent* on the policy π. In this case, it is common to select d as d_π, which is the *stationary distribution* under π. One basic property of d_π is that it satisfies

$$d_\pi^T P_\pi = d_\pi^T,$$

where P_π is the state transition probability matrix. More information about the stationary distribution can be found in Box 8.1.

The interpretation of selecting d_π is as follows. The stationary distribution reflects the long-term behavior of a Markov decision process under a given policy. If one state is frequently visited in the long term, it is more important and deserves a higher weight; if a state is rarely visited, then its importance is low and deserves a lower weight.

As its name suggests, $\bar{v}_\pi$ is a weighted average of the state values. Different values of θ lead to different values of $\bar{v}_\pi$. Our ultimate goal is to find an optimal policy (or equivalently an optimal θ) to maximize $\bar{v}_\pi$.

We next introduce another two important equivalent expressions of $\bar{v}_\pi$.

- ◇ Suppose that an agent collects rewards $\{R_{t+1}\}_{t=0}^\infty$ by following a given policy $\pi(\theta)$. Readers may often see the following metric in the literature:

$$J(\theta) = \lim_{n\to\infty} \mathbb{E}\left[\sum_{t=0}^{n} \gamma^t R_{t+1}\right] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right]. \tag{9.1}$$

This metric may be nontrivial to interpret at first glance. In fact, it is equal to $\bar{v}_\pi$. To see that, we have

$$\begin{aligned}\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}\right] &= \sum_{s\in\mathcal{S}} d(s)\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{t+1}|S_0 = s\right]\\ &= \sum_{s\in\mathcal{S}} d(s) v_\pi(s)\\ &= \bar{v}_\pi.\end{aligned}$$

The first equality in the above equation is due to the law of total expectation. The second equality is by the definition of state values.

- ◇ The metric $\bar{v}_\pi$ can also be rewritten as the inner product of two vectors. In particular, let

$$\begin{aligned}v_\pi &= [\ldots, v_\pi(s), \ldots]^T \in \mathbb{R}^{|\mathcal{S}|},\\ d &= [\ldots, d(s), \ldots]^T \in \mathbb{R}^{|\mathcal{S}|}.\end{aligned}$$

Then, we have

$$\bar{v}_\pi = d^T v_\pi.$$

This expression will be useful when we analyze its gradient.

Metric 2: Average reward

The second metric is the *average one-step reward* or simply called the *average reward* [2,64,65]. In particular, it is defined as

$$\begin{aligned}\bar{r}_\pi &\doteq \sum_{s\in\mathcal{S}} d_\pi(s) r_\pi(s) \\ &= \mathbb{E}_{S\sim d_\pi}[r_\pi(S)], \end{aligned} \tag{9.2}$$

where d_π is the stationary distribution and

$$r_\pi(s) \doteq \sum_{a\in\mathcal{A}} \pi(a|s,\theta) r(s,a) = \mathbb{E}_{A\sim\pi(s,\theta)}[r(s,A)|s] \tag{9.3}$$

is the expectation of the immediate rewards. Here, $r(s,a) \doteq \mathbb{E}[R|s,a] = \sum_r rp(r|s,a)$.

We next present another two important equivalent expressions of $\bar{r}_\pi$.

- ◇ Suppose that the agent collects rewards $\{R_{t+1}\}_{t=0}^{\infty}$ by following a given policy $\pi(\theta)$. A common metric that readers may often see in the literature is

$$J(\theta) = \lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\sum_{t=0}^{n-1} R_{t+1}\right]. \tag{9.4}$$

It may seem nontrivial to interpret this metric at first glance. In fact, it is equal to $\bar{r}_\pi$:

$$\lim_{n\to\infty} \frac{1}{n} \mathbb{E}\left[\sum_{t=0}^{n-1} R_{t+1}\right] = \sum_{s\in\mathcal{S}} d_\pi(s) r_\pi(s) = \bar{r}_\pi. \tag{9.5}$$

The proof of (9.5) is given in Box 9.1.

- ◇ The average reward $\bar{r}_\pi$ in (9.2) can also be written as the inner product of two vectors. In particular, let

$$\begin{aligned} r_\pi &= [\dots, r_\pi(s), \dots]^T \in \mathbb{R}^{|\mathcal{S}|}, \\ d_\pi &= [\dots, d_\pi(s), \dots]^T \in \mathbb{R}^{|\mathcal{S}|}, \end{aligned}$$

where $r_\pi(s)$ is defined in (9.3). Then, it is clear that

$$\bar{r}_\pi = \sum_{s\in\mathcal{S}} d_\pi(s) r_\pi(s) = d_\pi^T r_\pi.$$

This expression will be useful when we derive its gradient.

Box 9.1: Proof of (9.5)

Step 1: We first prove that the following equation is valid for any starting state $s_0 \in \mathcal{S}$:

$$\bar{r}_\pi = \lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{t=0}^{n-1} R_{t+1} | S_0 = s_0\right]. \tag{9.6}$$

To do that, we notice

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\left[\sum_{t=0}^{n-1} R_{t+1} | S_0 = s_0\right] &= \lim_{n\to\infty} \frac{1}{n}\sum_{t=0}^{n-1}\mathbb{E}\left[R_{t+1} | S_0 = s_0\right] \\ &= \lim_{t\to\infty} \mathbb{E}\left[R_{t+1} | S_0 = s_0\right], \end{aligned} \tag{9.7}$$

where the last equality is due to the property of the Cesaro mean (also called the Cesaro summation). In particular, if $\{a_k\}_{k=1}^\infty$ is a convergent sequence such that $\lim_{k\to\infty} a_k$ exists, then $\{1/n\sum_{k=1}^n a_k\}_{n=1}^\infty$ is also a convergent sequence such that $\lim_{n\to\infty} 1/n\sum_{k=1}^n a_k = \lim_{k\to\infty} a_k$.

We next examine $\mathbb{E}\left[R_{t+1}|S_0 = s_0\right]$ in (9.7) more closely. By the law of total expectation, we have

$$\begin{aligned}\mathbb{E}\left[R_{t+1}|S_0 = s_0\right] &= \sum_{s\in\mathcal{S}} \mathbb{E}\left[R_{t+1}|S_t = s, S_0 = s_0\right] p^{(t)}(s|s_0) \\ &= \sum_{s\in\mathcal{S}} \mathbb{E}\left[R_{t+1}|S_t = s\right] p^{(t)}(s|s_0) \\ &= \sum_{s\in\mathcal{S}} r_\pi(s) p^{(t)}(s|s_0),\end{aligned}$$

where $p^{(t)}(s|s_0)$ denotes the probability of transitioning from s_0 to s using exactly t steps. The second equality in the above equation is due to the Markov memoryless property: the reward obtained at the next time step depends only on the current state rather than the previous ones.

Note that

$$\lim_{t\to\infty} p^{(t)}(s|s_0) = d_\pi(s)$$

by the definition of the stationary distribution. As a result, the starting state s_0 does not matter. Then, we have

$$\lim_{t\to\infty} \mathbb{E}\left[R_{t+1}|S_0=s_0\right] = \lim_{t\to\infty}\sum_{s\in\mathcal{S}} r_\pi(s)p^{(t)}(s|s_0) = \sum_{s\in\mathcal{S}} r_\pi(s)d_\pi(s) = \bar{r}_\pi.$$

Substituting the above equation into (9.7) gives (9.6).

Step 2: Consider an arbitrary state distribution d. By the law of total expectation, we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{t=0}^{n-1}R_{t+1}\right] &= \lim_{n\to\infty}\frac{1}{n}\sum_{s\in\mathcal{S}} d(s)\mathbb{E}\left[\sum_{t=0}^{n-1}R_{t+1}|S_0=s\right]\\ &= \sum_{s\in\mathcal{S}} d(s)\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{t=0}^{n-1}R_{t+1}|S_0=s\right].\end{aligned}$$

Since (9.6) is valid for any starting state, substituting (9.6) into the above equation yields

$$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\left[\sum_{t=0}^{n-1}R_{t+1}\right] = \sum_{s\in\mathcal{S}} d(s)\bar{r}_\pi = \bar{r}_\pi.$$

The proof is complete.

Some remarks

Metric	Expression 1	Expression 2	Expression 3
$\bar{v}_\pi$	$\sum_{s\in\mathcal{S}} d(s)v_\pi(s)$	$\mathbb{E}_{S\sim d}[v_\pi(S)]$	$\lim_{n\to\infty}\mathbb{E}\big[\sum_{t=0}^{n}\gamma^t R_{t+1}\big]$
$\bar{r}_\pi$	$\sum_{s\in\mathcal{S}} d_\pi(s)r_\pi(s)$	$\mathbb{E}_{S\sim d_\pi}[r_\pi(S)]$	$\lim_{n\to\infty}\frac{1}{n}\mathbb{E}\big[\sum_{t=0}^{n-1}R_{t+1}\big]$

Table 9.2: Summary of the different but equivalent expressions of $\bar{v}_\pi$ and $\bar{r}_\pi$.

Up to now, we have introduced two types of metrics: $\bar{v}_\pi$ and $\bar{r}_\pi$. Each metric has several different but equivalent expressions. They are summarized in Table 9.2. We sometimes use $\bar{v}_\pi$ to specifically refer to the case where the state distribution is the stationary distribution d_π and use $\bar{v}_\pi^0$ to refer to the case where d_0 is independent of π. Some remarks about the metrics are given below.

◇ All these metrics are functions of π. Since π is parameterized by θ, these metrics are functions of θ. In other words, different values of θ can generate different metric values. Therefore, we can search for the optimal values of θ to maximize these metrics. This is the basic idea of policy gradient methods.

- ◇ The two metrics $\bar{v}_\pi$ and $\bar{r}_\pi$ are equivalent in the discounted case where $\gamma < 1$. In particular, it can be shown that

$$\bar{r}_\pi = (1-\gamma)\bar{v}_\pi.$$

The above equation indicates that these two metrics can be simultaneously maximized. The proof of this equation is given later in Lemma 9.1.

9.3 Gradients of the metrics

Given the metrics introduced in the last section, we can use gradient-based methods to maximize them. To do that, we need to first calculate the gradients of these metrics. The most important theoretical result in this chapter is the following theorem.

Theorem 9.1 (Policy gradient theorem)**.** *The gradient of $J(\theta)$ is*

$$\nabla_\theta J(\theta) = \sum_{s\in\mathcal{S}} \eta(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a), \tag{9.8}$$

where η is a state distribution and $\nabla_\theta \pi$ is the gradient of π with respect to θ. Moreover, (9.8) *has a compact form expressed in terms of expectation:*

$$\nabla_\theta J(\theta) = \mathbb{E}_{S\sim\eta, A\sim\pi(S,\theta)}\Big[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\Big], \tag{9.9}$$

where ln *is the natural logarithm.*

Some important remarks about Theorem 9.1 are given below.

- ◇ It should be noted that Theorem 9.1 is a summary of the results in Theorem 9.2, Theorem 9.3, and Theorem 9.5. These three theorems address different scenarios involving different metrics and discounted/undiscounted cases. The gradients in these scenarios all have similar expressions and hence are summarized in Theorem 9.1. The specific expressions of $J(\theta)$ and η are not given in Theorem 9.1 and can be found in Theorem 9.2, Theorem 9.3, and Theorem 9.5. In particular, $J(\theta)$ could be $\bar{v}_\pi^0$, $\bar{v}_\pi$, or $\bar{r}_\pi$. The equality in (9.8) may become a strict equality or an approximation. The distribution η also varies in different scenarios.

  The derivation of the gradients is the most complicated part of the policy gradient method. For many readers, it is sufficient to be familiar with the result in Theorem 9.1 without knowing the proof. The derivation details presented in the rest of this section are mathematically intensive. Readers are suggested to study selectively based on their interests.

- ◇ The expression in (9.9) is more favorable than (9.8) because it is expressed as an expectation. We will show in Section 9.4 that this true gradient can be approximated by a stochastic gradient.

  Why can (9.8) be expressed as (9.9)? The proof is given below. By the definition of expectation, (9.8) can be rewritten as

$$\begin{aligned}\nabla_\theta J(\theta) &= \sum_{s\in\mathcal{S}} \eta(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) \\ &= \mathbb{E}_{S\sim\eta}\left[\sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|S,\theta) q_\pi(S,a)\right]. \end{aligned} \tag{9.10}$$

  Furthermore, the gradient of $\ln \pi(a|s,\theta)$ is

$$\nabla_\theta \ln \pi(a|s,\theta) = \frac{\nabla_\theta \pi(a|s,\theta)}{\pi(a|s,\theta)}.$$

  It follows that

$$\nabla_\theta \pi(a|s,\theta) = \pi(a|s,\theta) \nabla_\theta \ln \pi(a|s,\theta). \tag{9.11}$$

  Substituting (9.11) into (9.10) gives

$$\begin{aligned}\nabla_\theta J(\theta) &= \mathbb{E}\left[\sum_{a\in\mathcal{A}} \pi(a|S,\theta) \nabla_\theta \ln \pi(a|S,\theta) q_\pi(S,a)\right] \\ &= \mathbb{E}_{S\sim\eta, A\sim\pi(S,\theta)}\Big[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\Big].\end{aligned}$$

- ◇ It is notable that $\pi(a|s,\theta)$ must be *positive* for all (s,a) to ensure that $\ln \pi(a|s,\theta)$ is valid. This can be achieved by using *softmax functions*:

$$\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\sum_{a'\in\mathcal{A}} e^{h(s,a',\theta)}}, \quad a \in \mathcal{A}, \tag{9.12}$$

  where $h(s,a,\theta)$ is a function indicating the preference for selecting a at s. The policy in (9.12) satisfies $\pi(a|s,\theta) \in [0,1]$ and $\sum_{a\in\mathcal{A}} \pi(a|s,\theta) = 1$ for any $s \in \mathcal{S}$. This policy can be realized by a neural network. The input of the network is s. The output layer is a softmax layer so that the network outputs $\pi(a|s,\theta)$ for all a and the sum of the outputs is equal to 1. See Figure 9.2(b) for an illustration.

  Since $\pi(a|s,\theta) > 0$ for all a, the policy is *stochastic* and hence *exploratory*. The policy does not directly tell which action to take. Instead, the action should be generated according to the probability distribution of the policy.

9.3.1 Derivation of the gradients in the discounted case

We next derive the gradients of the metrics in the discounted case where $\gamma \in (0,1)$. The state value and action value in the discounted case are defined as

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s], \\ q_\pi(s,a) &= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s, A_t = a]. \end{aligned}$$

It holds that $v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s,\theta) q_\pi(s,a)$ and the state value satisfies the Bellman equation.

First, we show that $\bar{v}_\pi(\theta)$ and $\bar{r}_\pi(\theta)$ are equivalent metrics.

Lemma 9.1 (Equivalence between $\bar{v}_\pi(\theta)$ and $\bar{r}_\pi(\theta)$). *In the discounted case where* $\gamma \in (0,1)$*, it holds that*

$$\bar{r}_\pi = (1-\gamma)\bar{v}_\pi. \tag{9.13}$$

Proof. Note that $\bar{v}_\pi(\theta) = d_\pi^T v_\pi$ and $\bar{r}_\pi(\theta) = d_\pi^T r_\pi$, where v_π and r_π satisfy the Bellman equation $v_\pi = r_\pi + \gamma P_\pi v_\pi$. Multiplying d_π^T on both sides of the Bellman equation yields

$$\bar{v}_\pi = \bar{r}_\pi + \gamma d_\pi^T P_\pi v_\pi = \bar{r}_\pi + \gamma d_\pi^T v_\pi = \bar{r}_\pi + \gamma \bar{v}_\pi,$$

which implies (9.13). □

Second, the following lemma gives the gradient of $v_\pi(s)$ for any s.

Lemma 9.2 (Gradient of $v_\pi(s)$). *In the discounted case, it holds for any* $s \in \mathcal{S}$ *that*

$$\nabla_\theta v_\pi(s) = \sum_{s'\in\mathcal{S}} \Pr_\pi(s'|s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s',\theta) q_\pi(s',a), \tag{9.14}$$

where

$$\Pr_\pi(s'|s) \doteq \sum_{k=0}^{\infty} \gamma^k [P_\pi^k]_{ss'} = \left[(I_n - \gamma P_\pi)^{-1}\right]_{ss'}$$

is the discounted total probability of transitioning from s to s' under policy π. Here, $[\cdot]_{ss'}$ denotes the entry in the sth row and s'th column, and $[P_\pi^k]_{ss'}$ is the probability of transitioning from s to s' using exactly k steps under π.

Box 9.2: Proof of Lemma 9.2

First, for any $s \in \mathcal{S}$, it holds that

$$\begin{aligned}\nabla_\theta v_\pi(s) &= \nabla_\theta \left[\sum_{a\in\mathcal{A}} \pi(a|s,\theta) q_\pi(s,a)\right] \\ &= \sum_{a\in\mathcal{A}} \left[\nabla_\theta \pi(a|s,\theta) q_\pi(s,a) + \pi(a|s,\theta)\nabla_\theta q_\pi(s,a)\right], \qquad (9.15)\end{aligned}$$

where $q_\pi(s,a)$ is the action value given by

$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a) v_\pi(s').$$

Since $r(s,a) = \sum_r rp(r|s,a)$ is independent of θ, we have

$$\nabla_\theta q_\pi(s,a) = 0 + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\nabla_\theta v_\pi(s').$$

Substituting this result into (9.15) yields

$$\begin{aligned}\nabla_\theta v_\pi(s) &= \sum_{a\in\mathcal{A}} \left[\nabla_\theta \pi(a|s,\theta) q_\pi(s,a) + \pi(a|s,\theta)\gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)\nabla_\theta v_\pi(s')\right] \\ &= \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) + \gamma \sum_{a\in\mathcal{A}} \pi(a|s,\theta) \sum_{s'\in\mathcal{S}} p(s'|s,a)\nabla_\theta v_\pi(s'). \qquad (9.16)\end{aligned}$$

It is notable that $\nabla_\theta v_\pi$ appears on both sides of the above equation. One way to calculate it is to use the *unrolling technique* [64]. Here, we use another way based on the *matrix-vector form*, which we believe is more straightforward to understand. In particular, let

$$u(s) \doteq \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a).$$

Since

$$\sum_{a\in\mathcal{A}} \pi(a|s,\theta) \sum_{s'\in\mathcal{S}} p(s'|s,a)\nabla_\theta v_\pi(s') = \sum_{s'\in\mathcal{S}} p(s'|s)\nabla_\theta v_\pi(s') = \sum_{s'\in\mathcal{S}} [P_\pi]_{ss'}\nabla_\theta v_\pi(s'),$$

equation (9.16) can be written in matrix-vector form as

$$\underbrace{\begin{bmatrix} \vdots \\ \nabla_\theta v_\pi(s) \\ \vdots \end{bmatrix}}_{\nabla_\theta v_\pi \in \mathbb{R}^{mn}} = \underbrace{\begin{bmatrix} \vdots \\ u(s) \\ \vdots \end{bmatrix}}_{u \in \mathbb{R}^{mn}} + \gamma (P_\pi \otimes I_m) \underbrace{\begin{bmatrix} \vdots \\ \nabla_\theta v_\pi(s') \\ \vdots \end{bmatrix}}_{\nabla_\theta v_\pi \in \mathbb{R}^{mn}},$$

which can be written concisely as

$$\nabla_\theta v_\pi = u + \gamma(P_\pi \otimes I_m)\nabla_\theta v_\pi.$$

Here, $n = |\mathcal{S}|$, and m is the dimension of the parameter vector θ. The reason that the Kronecker product $\otimes$ emerges in the equation is that $\nabla_\theta v_\pi(s)$ is a vector. The above equation is a linear equation of $\nabla_\theta v_\pi$, which can be solved as

$$\begin{aligned}\nabla_\theta v_\pi &= (I_{nm} - \gamma P_\pi \otimes I_m)^{-1} u \\ &= (I_n \otimes I_m - \gamma P_\pi \otimes I_m)^{-1} u \\ &= \left[(I_n - \gamma P_\pi)^{-1} \otimes I_m\right] u. \end{aligned} \tag{9.17}$$

For any state s, it follows from (9.17) that

$$\begin{aligned}\nabla_\theta v_\pi(s) &= \sum_{s'\in\mathcal{S}} \left[(I_n - \gamma P_\pi)^{-1}\right]_{ss'} u(s') \\ &= \sum_{s'\in\mathcal{S}} \left[(I_n - \gamma P_\pi)^{-1}\right]_{ss'} \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s',\theta) q_\pi(s',a). \end{aligned} \tag{9.18}$$

The quantity $[(I_n - \gamma P_\pi)^{-1}]_{ss'}$ has a clear probabilistic interpretation. In particular, since $(I_n - \gamma P_\pi)^{-1} = I + \gamma P_\pi + \gamma^2 P_\pi^2 + \cdots$, we have

$$\left[(I_n - \gamma P_\pi)^{-1}\right]_{ss'} = [I]_{ss'} + \gamma [P_\pi]_{ss'} + \gamma^2 [P_\pi^2]_{ss'} + \cdots = \sum_{k=0}^{\infty} \gamma^k [P_\pi^k]_{ss'}.$$

Note that $[P_\pi^k]_{ss'}$ is the probability of transitioning from s to s' using exactly k steps (see Box 8.1). Therefore, $[(I_n - \gamma P_\pi)^{-1}]_{ss'}$ is the discounted total probability of transitioning from s to s' using any number of steps. By denoting $[(I_n - \gamma P_\pi)^{-1}]_{ss'} \doteq \Pr_\pi(s'|s)$, equation (9.18) becomes (9.14).

With the results in Lemma 9.2, we are ready to derive the gradient of $\bar{v}_\pi^0$.

Theorem 9.2 (Gradient of $\bar{v}_\pi^0$ in the discounted case)**.** *In the discounted case where $\gamma \in (0,1)$, the gradient of $\bar{v}_\pi^0 = d_0^T v_\pi$ is*

$$\nabla_\theta \bar{v}_\pi^0 = \mathbb{E}\big[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\big],$$

where $S \sim \rho_\pi$ and $A \sim \pi(S,\theta)$. Here, the state distribution ρ_π is

$$\rho_\pi(s) = \sum_{s'\in\mathcal{S}} d_0(s') \Pr_\pi(s|s'), \qquad s \in \mathcal{S}, \tag{9.19}$$

where $\Pr_\pi(s|s') = \sum_{k=0}^{\infty} \gamma^k [P_\pi^k]_{s's} = [(I - \gamma P_\pi)^{-1}]_{s's}$ is the discounted total probability of *transitioning from s' to s under policy π.*

Box 9.3: Proof of Theorem 9.2

Since $d_0(s)$ is independent of π, we have

$$\nabla_\theta \bar{v}_\pi^0 = \nabla_\theta \sum_{s\in\mathcal{S}} d_0(s) v_\pi(s) = \sum_{s\in\mathcal{S}} d_0(s) \nabla_\theta v_\pi(s).$$

Substituting the expression of $\nabla_\theta v_\pi(s)$ given in Lemma 9.2 into the above equation yields

$$\begin{aligned}\nabla_\theta \bar{v}_\pi^0 = \sum_{s\in\mathcal{S}} d_0(s)\nabla_\theta v_\pi(s) &= \sum_{s\in\mathcal{S}} d_0(s) \sum_{s'\in\mathcal{S}} \Pr_\pi(s'|s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s',\theta) q_\pi(s',a) \\ &= \sum_{s'\in\mathcal{S}} \left(\sum_{s\in\mathcal{S}} d_0(s) \Pr_\pi(s'|s) \right) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s',\theta) q_\pi(s',a) \\ &\doteq \sum_{s'\in\mathcal{S}} \rho_\pi(s') \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s',\theta) q_\pi(s',a) \\ &= \sum_{s\in\mathcal{S}} \rho_\pi(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) \qquad \text{(change } s' \text{ to } s) \\ &= \sum_{s\in\mathcal{S}} \rho_\pi(s) \sum_{a\in\mathcal{A}} \pi(a|s,\theta) \nabla_\theta \ln \pi(a|s,\theta) q_\pi(s,a) \\ &= \mathbb{E}\left[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A) \right],\end{aligned}$$

where $S \sim \rho_\pi$ and $A \sim \pi(S,\theta)$. The proof is complete.

With Lemma 9.1 and Lemma 9.2, we can derive the gradients of $\bar{r}_\pi$ and $\bar{v}_\pi$.

Theorem 9.3 (Gradients of $\bar{r}_\pi$ and $\bar{v}_\pi$ in the discounted case)**.** *In the discounted case where $\gamma \in (0,1)$, the gradients of $\bar{r}_\pi$ and $\bar{v}_\pi$ are*

$$\begin{aligned}\nabla_\theta \bar{r}_\pi = (1-\gamma)\nabla_\theta \bar{v}_\pi &\approx \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) \\ &= \mathbb{E}\big[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\big],\end{aligned}$$

where $S \sim d_\pi$ and $A \sim \pi(S,\theta)$. Here, the approximation is more accurate when γ is closer to 1.

Box 9.4: Proof of Theorem 9.3

It follows from the definition of $\bar{v}_\pi$ that

$$\begin{aligned}\nabla_\theta \bar{v}_\pi &= \nabla_\theta \sum_{s\in\mathcal{S}} d_\pi(s) v_\pi(s) \\ &= \sum_{s\in\mathcal{S}} \nabla_\theta d_\pi(s) v_\pi(s) + \sum_{s\in\mathcal{S}} d_\pi(s) \nabla_\theta v_\pi(s). \end{aligned} \tag{9.20}$$

This equation contains two terms. On the one hand, substituting the expression of $\nabla_\theta v_\pi$ given in (9.17) into the second term gives

$$\begin{aligned}\sum_{s\in\mathcal{S}} d_\pi(s)\nabla_\theta v_\pi(s) &= (d_\pi^T \otimes I_m)\nabla_\theta v_\pi \\ &= (d_\pi^T \otimes I_m)\left[(I_n - \gamma P_\pi)^{-1} \otimes I_m\right] u \\ &= \left[d_\pi^T (I_n - \gamma P_\pi)^{-1}\right] \otimes I_m u. \end{aligned} \tag{9.21}$$

It is noted that

$$d_\pi^T (I_n - \gamma P_\pi)^{-1} = \frac{1}{1-\gamma} d_\pi^T,$$

which can be easily verified by multiplying $(I_n - \gamma P_\pi)$ on both sides of the equation. Therefore, (9.21) becomes

$$\begin{aligned}\sum_{s\in\mathcal{S}} d_\pi(s)\nabla_\theta v_\pi(s) &= \frac{1}{1-\gamma} d_\pi^T \otimes I_m u \\ &= \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a). \end{aligned}$$

On the other hand, the first term of (9.20) involves $\nabla_\theta d_\pi$. However, since the second term contains $\frac{1}{1-\gamma}$, the second term becomes dominant, and the first term becomes negligible when $\gamma \to 1$. Therefore,

$$\nabla_\theta \bar{v}_\pi \approx \frac{1}{1-\gamma} \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a).$$

Furthermore, it follows from $\bar{r}_\pi = (1-\gamma)\bar{v}_\pi$ that

$$\begin{aligned}\nabla_\theta \bar{r}_\pi = (1-\gamma)\nabla_\theta \bar{v}_\pi &\approx \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) \\ &= \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \pi(a|s,\theta) \nabla_\theta \ln \pi(a|s,\theta) q_\pi(s,a) \\ &= \mathbb{E}\left[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\right]. \end{aligned}$$

The approximation in the above equation requires that the first term does not go to infinity when $\gamma \to 1$. More information can be found in [66, Section 4].

9.3.2 Derivation of the gradients in the undiscounted case

We next show how to calculate the gradients of the metrics in the undiscounted case where $\gamma = 1$. Readers may wonder why we suddenly start considering the undiscounted case while we have only considered the discounted case so far in this book. The reasons are as follows. First, for continuing tasks, it may be inappropriate to introduce the discount rate and we need to consider the undiscounted case. Second, the definition of the average reward $\bar{r}_\pi$ is valid for both discounted and undiscounted cases. While the gradient of $\bar{r}_\pi$ in the discounted case is an approximation, we will see that its gradient in the undiscounted case is more elegant.

State values and the Poisson equation

In the undiscounted case, it is necessary to redefine state and action values. Since the undiscounted sum of the rewards, $\mathbb{E}[R_{t+1} + R_{t+2} + R_{t+3} + \dots | S_t = s]$, may diverge, the state and action values are defined in a special way [64]:

$$\begin{aligned} v_\pi(s) &\doteq \mathbb{E}[(R_{t+1} - \bar{r}_\pi) + (R_{t+2} - \bar{r}_\pi) + (R_{t+3} - \bar{r}_\pi) + \dots | S_t = s], \\ q_\pi(s,a) &\doteq \mathbb{E}[(R_{t+1} - \bar{r}_\pi) + (R_{t+2} - \bar{r}_\pi) + (R_{t+3} - \bar{r}_\pi) + \dots | S_t = s, A_t = a], \end{aligned}$$

where $\bar{r}_\pi$ is the average reward, which is determined when π is given. There are different names for $v_\pi(s)$ in the literature such as the differential reward [65] or bias [2, Section 8.2.1]. It can be verified that the state value defined above satisfies the following Bellman-like equation:

$$v_\pi(s) = \sum_a \pi(a|s,\theta) \left[\sum_r p(r|s,a)(r - \bar{r}_\pi) + \sum_{s'} p(s'|s,a) v_\pi(s') \right]. \tag{9.22}$$

Since $v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s,\theta) q_\pi(s,a)$, it holds that $q_\pi(s,a) = \sum_r p(r|s,a)(r - \bar{r}_\pi) + \sum_{s'} p(s'|s,a) v_\pi(s')$. The matrix-vector form of (9.22) is

$$v_\pi = r_\pi - \bar{r}_\pi \mathbf{1}_n + P_\pi v_\pi, \tag{9.23}$$

where $\mathbf{1}_n = [1, \dots, 1]^T \in \mathbb{R}^n$. Equation (9.23) is similar to the Bellman equation and it has a specific name called the *Poisson equation* [65, 67].

How to solve v_π from the Poisson equation? The answer is given in the following theorem.

Theorem 9.4 (Solution of the Poisson equation)**.** *Let*

$$v_\pi^* = (I_n - P_\pi + \mathbf{1}_n d_\pi^T)^{-1} r_\pi. \tag{9.24}$$

Then, v_π^* is a solution of the Poisson equation in (9.23)*. Moreover, any solution of the Poisson equation has the following form:*

$$v_\pi = v_\pi^* + c\mathbf{1}_n,$$

where $c \in \mathbb{R}$.

This theorem indicates that the solution of the Poisson equation may not be unique.

Box 9.5: Proof of Theorem 9.4

We prove using three steps.

◇ Step 1: Show that v_π^* in (9.24) is a solution of (9.25).

For the sake of simplicity, let

$$A \doteq I_n - P_\pi + \mathbf{1}_n d_\pi^T.$$

Then, $v_\pi^* = A^{-1} r_\pi$. The fact that A is invertible will be proven in Step 3. Substituting $v_\pi^* = A^{-1} r_\pi$ into (9.25) gives

$$A^{-1} r_\pi = r_\pi - \mathbf{1}_n d_\pi^T r_\pi + P_\pi A^{-1} r_\pi.$$

This equation is valid as proven below. Recognizing this equation gives $(-A^{-1} + I_n - \mathbf{1}_n d_\pi^T + P_\pi A^{-1}) r_\pi = 0$, and consequently,

$$(-I_n + A - \mathbf{1}_n d_\pi^T A + P_\pi) A^{-1} r_\pi = 0.$$

The term in the brackets in the above equation is zero because $-I_n + A - \mathbf{1}_n d_\pi^T A + P_\pi = -I_n + (I_n - P_\pi + \mathbf{1}_n d_\pi^T) - \mathbf{1}_n d_\pi^T (I_n - P_\pi + \mathbf{1}_n d_\pi^T) + P_\pi = 0$. Therefore, v_π^* in (9.24) is a solution.

◇ Step 2: General expression of the solutions.

Substituting $\bar{r}_\pi = d_\pi^T r_\pi$ into (9.23) gives

$$v_\pi = r_\pi - \mathbf{1}_n d_\pi^T r_\pi + P_\pi v_\pi \tag{9.25}$$

and consequently

$$(I_n - P_\pi) v_\pi = (I_n - \mathbf{1}_n d_\pi^T) r_\pi. \tag{9.26}$$

It is noted that $I_n - P_\pi$ is singular because $(I_n - P_\pi)\mathbf{1}_n = 0$ for any π. Therefore, the solution of (9.26) is not unique: if v_π^* is a solution, then $v_\pi^* + x$ is also a solution for any $x \in \text{Null}(I_n - P_\pi)$. When P_π is irreducible, $\text{Null}(I_n - P_\pi) = \text{span}\{\mathbf{1}_n\}$. Then, any solution of the Poisson equation has the expression $v_\pi^* + c\mathbf{1}_n$ where $c \in \mathbb{R}$.

◇ Step 3: Show that $A = I_n - P_\pi + \mathbf{1}_n d_\pi^T$ is invertible.

Since v_π^* involves A^{-1}, it is necessary to show that A is invertible. The analysis is summarized in the following lemma.

Lemma 9.3. *The matrix $I_n - P_\pi + \mathbf{1}_n d_\pi^T$ is invertible and its inverse is*

$$\left[I_n - (P_\pi - \mathbf{1}_n d_\pi^T)\right]^{-1} = \sum_{k=1}^{\infty}(P_\pi^k - \mathbf{1}_n d_\pi^T) + I_n.$$

Proof. First of all, we state some preliminary facts without proof. Let $\rho(M)$ be the spectral radius of a matrix M. Then, $I - M$ is invertible if $\rho(M) < 1$. Moreover, $\rho(M) < 1$ if and only if $\lim_{k\to\infty} M^k = 0$.

Based on the above facts, we next show that $\lim_{k\to\infty}(P_\pi - \mathbf{1}_n d_\pi^T)^k \to 0$, and then the invertibility of $I_n - (P_\pi - \mathbf{1}_n d_\pi^T)$ immediately follows. To do that, we notice that

$$(P_\pi - \mathbf{1}_n d_\pi^T)^k = P_\pi^k - \mathbf{1}_n d_\pi^T, \quad k \geq 1, \tag{9.27}$$

which can be proven by induction. For instance, when $k = 1$, the equation is valid. When $k = 2$, we have

$$\begin{aligned}(P_\pi - \mathbf{1}_n d_\pi^T)^2 &= (P_\pi - \mathbf{1}_n d_\pi^T)(P_\pi - \mathbf{1}_n d_\pi^T)\\ &= P_\pi^2 - P_\pi \mathbf{1}_n d_\pi^T - \mathbf{1}_n d_\pi^T P_\pi + \mathbf{1}_n d_\pi^T \mathbf{1}_n d_\pi^T\\ &= P_\pi^2 - \mathbf{1}_n d_\pi^T,\end{aligned}$$

where the last equality is due to $P_\pi \mathbf{1}_n = \mathbf{1}_n$, $d_\pi^T P_\pi = d_\pi^T$, and $d_\pi^T \mathbf{1}_n = 1$. The case of $k \geq 3$ can be proven similarly.

Since d_π is the stationary distribution of the state, it holds that $\lim_{k\to\infty} P_\pi^k = d_\pi^T \mathbf{1}_n$ (see Box 8.1). Therefore, (9.27) implies that

$$\lim_{k\to\infty}(P_\pi - \mathbf{1}_n d_\pi^T)^k = \lim_{k\to\infty} P_\pi^k - d_\pi^T \mathbf{1}_n = 0.$$

As a result, $\rho(P_\pi - \mathbf{1}_n d_\pi^T) < 1$ and hence $I_n - (P_\pi - \mathbf{1}_n d_\pi^T)$ is invertible. Furthermore, the inverse of this matrix is given by

$$\begin{aligned}(I_n - (P_\pi - \mathbf{1}_n d_\pi^T))^{-1} &= \sum_{k=0}^{\infty}(P_\pi - \mathbf{1}_n d_\pi^T)^k \\ &= I_n + \sum_{k=1}^{\infty}(P_\pi - \mathbf{1}_n d_\pi^T)^k \\ &= I_n + \sum_{k=1}^{\infty}(P_\pi^k - \mathbf{1}_n d_\pi^T) \\ &= \sum_{k=0}^{\infty}(P_\pi^k - \mathbf{1}_n d_\pi^T) + \mathbf{1}_n d_\pi^T.\end{aligned}$$

The proof is complete. □

The proof of Lemma 9.3 is inspired by [66]. However, the result $(I_n - P_\pi + \mathbf{1}_n d_\pi^T)^{-1} = \sum_{k=0}^{\infty}(P_\pi^k - \mathbf{1}_n d_\pi^T)$ given in [66] (the statement above equation (16) in [66]) is inaccurate because $\sum_{k=0}^{\infty}(P_\pi^k - \mathbf{1}_n d_\pi^T)$ is singular since $\sum_{k=0}^{\infty}(P_\pi^k - \mathbf{1}_n d_\pi^T)\mathbf{1}_n = 0$. Lemma 9.3 corrects this inaccuracy.

Derivation of gradients

Although the value of v_π is not unique in the undiscounted case, as shown in Theorem 9.4, the value of $\bar{r}_\pi$ is unique. In particular, it follows from the Poisson equation that

$$\begin{aligned}\bar{r}_\pi \mathbf{1}_n &= r_\pi + (P_\pi - I_n)v_\pi \\ &= r_\pi + (P_\pi - I_n)(v_\pi^* + c\mathbf{1}_n) \\ &= r_\pi + (P_\pi - I_n)v_\pi^*.\end{aligned}$$

Notably, the undetermined value c is canceled and hence $\bar{r}_\pi$ is unique. Therefore, we can calculate the gradient of $\bar{r}_\pi$ in the undiscounted case. In addition, since v_π is not unique, $\bar{v}_\pi$ is not unique either. We do not study the gradient of $\bar{v}_\pi$ in the undiscounted case. For interested readers, it is worth mentioning that we can add more constraints to uniquely solve v_π from the Poisson equation. For example, by assuming that a recurrent state exists, the state value of this recurrent state is zero [65, Section II], and hence c can be determined. There are also other ways to uniquely determine v_π. See, for example, equations (8.6.5)-(8.6.7) in [2].

The gradient of $\bar{r}_\pi$ in the undiscounted case is given below.

Theorem 9.5 (Gradient of $\bar{r}_\pi$ in the undiscounted case)**.** *In the undiscounted case, the* *gradient of the average reward $\bar{r}_\pi$ is*

$$\begin{aligned}\nabla_\theta \bar{r}_\pi &= \sum_{s\in\mathcal{S}} d_\pi(s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) \\ &= \mathbb{E}\big[\nabla_\theta \ln \pi(A|S,\theta) q_\pi(S,A)\big], \end{aligned} \tag{9.28}$$

where $S \sim d_\pi$ and $A \sim \pi(S,\theta)$.

Compared to the discounted case shown in Theorem 9.3, the gradient of $\bar{r}_\pi$ in the undiscounted case is more elegant in the sense that (9.28) is strictly valid and S obeys the stationary distribution.

Box 9.6: Proof of Theorem 9.5

First of all, it follows from $v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s,\theta) q_\pi(s,a)$ that

$$\begin{aligned}\nabla_\theta v_\pi(s) &= \nabla_\theta \left[\sum_{a\in\mathcal{A}} \pi(a|s,\theta) q_\pi(s,a)\right] \\ &= \sum_{a\in\mathcal{A}} \left[\nabla_\theta \pi(a|s,\theta) q_\pi(s,a) + \pi(a|s,\theta)\nabla_\theta q_\pi(s,a)\right], \end{aligned} \tag{9.29}$$

where $q_\pi(s,a)$ is the action value satisfying

$$\begin{aligned} q_\pi(s,a) &= \sum_r p(r|s,a)(r - \bar{r}_\pi) + \sum_{s'} p(s'|s,a) v_\pi(s') \\ &= r(s,a) - \bar{r}_\pi + \sum_{s'} p(s'|s,a) v_\pi(s'). \end{aligned}$$

Since $r(s,a) = \sum_r rp(r|s,a)$ is independent of θ, we have

$$\nabla_\theta q_\pi(s,a) = 0 - \nabla_\theta \bar{r}_\pi + \sum_{s'\in\mathcal{S}} p(s'|s,a) \nabla_\theta v_\pi(s').$$

Substituting this result into (9.29) yields

$$\begin{aligned}\nabla_\theta v_\pi(s) &= \sum_{a\in\mathcal{A}} \left[\nabla_\theta \pi(a|s,\theta) q_\pi(s,a) + \pi(a|s,\theta)\left(-\nabla_\theta \bar{r}_\pi + \sum_{s'\in\mathcal{S}} p(s'|s,a)\nabla_\theta v_\pi(s')\right)\right] \\ &= \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) - \nabla_\theta \bar{r}_\pi + \sum_{a\in\mathcal{A}} \pi(a|s,\theta) \sum_{s'\in\mathcal{S}} p(s'|s,a)\nabla_\theta v_\pi(s'). \end{aligned} \tag{9.30}$$

Let

$$u(s) \doteq \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s,\theta) q_\pi(s,a).$$

Since $\sum_{a\in\mathcal{A}}\pi(a|s,\theta)\sum_{s'\in\mathcal{S}}p(s'|s,a)\nabla_\theta v_\pi(s') = \sum_{s'\in\mathcal{S}}p(s'|s)\nabla_\theta v_\pi(s')$, equation (9.30) can be written in matrix-vector form as

$$\underbrace{\begin{bmatrix}\vdots\\ \nabla_\theta v_\pi(s)\\ \vdots\end{bmatrix}}_{\nabla_\theta v_\pi\in\mathbb{R}^{mn}} = \underbrace{\begin{bmatrix}\vdots\\ u(s)\\ \vdots\end{bmatrix}}_{u\in\mathbb{R}^{mn}} - \mathbf{1}_n\otimes\nabla_\theta\bar{r}_\pi + (P_\pi\otimes I_m)\underbrace{\begin{bmatrix}\vdots\\ \nabla_\theta v_\pi(s')\\ \vdots\end{bmatrix}}_{\nabla_\theta v_\pi\in\mathbb{R}^{mn}},$$

where $n=|\mathcal{S}|$, m is the dimension of θ, and $\otimes$ is the Kronecker product. The above equation can be written concisely as

$$\nabla_\theta v_\pi = u - \mathbf{1}_n\otimes\nabla_\theta\bar{r}_\pi + (P_\pi\otimes I_m)\nabla_\theta v_\pi,$$

and hence

$$\mathbf{1}_n\otimes\nabla_\theta\bar{r}_\pi = u + (P_\pi\otimes I_m)\nabla_\theta v_\pi - \nabla_\theta v_\pi.$$

Multiplying $d_\pi^T\otimes I_m$ on both sides of the above equation gives

$$\begin{aligned}(d_\pi^T\mathbf{1}_n)\otimes\nabla_\theta\bar{r}_\pi &= d_\pi^T\otimes I_m u + (d_\pi^T P_\pi)\otimes I_m\nabla_\theta v_\pi - d_\pi^T\otimes I_m\nabla_\theta v_\pi\\ &= d_\pi^T\otimes I_m u,\end{aligned}$$

which implies

$$\begin{aligned}\nabla_\theta\bar{r}_\pi &= d_\pi^T\otimes I_m u\\ &= \sum_{s\in\mathcal{S}}d_\pi(s)u(s)\\ &= \sum_{s\in\mathcal{S}}d_\pi(s)\sum_{a\in\mathcal{A}}\nabla_\theta\pi(a|s,\theta)q_\pi(s,a).\end{aligned}$$

9.4 Monte Carlo policy gradient (REINFORCE)

With the gradient presented in Theorem 9.1, we next show how to use the gradient-based method to optimize the metrics to obtain optimal policies.

The gradient-ascent algorithm for maximizing $J(\theta)$ is

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha\nabla_\theta J(\theta_t)\\ &= \theta_t + \alpha\mathbb{E}\Big[\nabla_\theta\ln\pi(A|S,\theta_t)q_\pi(S,A)\Big],\end{aligned}\tag{9.31}$$

where $\alpha>0$ is a constant learning rate. Since the true gradient in (9.31) is unknown, we can replace the true gradient with a stochastic gradient to obtain the following algorithm:

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t|s_t, \theta_t) q_t(s_t, a_t), \tag{9.32}$$

where $q_t(s_t, a_t)$ is an approximation of $q_\pi(s_t, a_t)$. If $q_t(s_t, a_t)$ is obtained by Monte Carlo estimation, the algorithm is called *REINFORCE* [68] or *Monte Carlo policy gradient*, which is one of earliest and simplest policy gradient algorithms.

The algorithm in (9.32) is important since many other policy gradient algorithms can be obtained by extending it. We next examine the interpretation of (9.32) more closely. Since $\nabla_\theta \ln \pi(a_t|s_t, \theta_t) = \frac{\nabla_\theta \pi(a_t|s_t,\theta_t)}{\pi(a_t|s_t,\theta_t)}$, we can rewrite (9.32) as

$$\theta_{t+1} = \theta_t + \alpha \underbrace{\left(\frac{q_t(s_t, a_t)}{\pi(a_t|s_t, \theta_t)} \right)}_{\beta_t} \nabla_\theta \pi(a_t|s_t, \theta_t),$$

which can be further written concisely as

$$\theta_{t+1} = \theta_t + \alpha \beta_t \nabla_\theta \pi(a_t|s_t, \theta_t). \tag{9.33}$$

Two important interpretations can be seen from this equation.

◇ First, since (9.33) is a simple gradient-ascent algorithm, the following observations can be obtained.

- If $\beta_t \geq 0$, the probability of choosing (s_t, a_t) is enhanced. That is

$$\pi(a_t|s_t, \theta_{t+1}) \geq \pi(a_t|s_t, \theta_t).$$

  The greater β_t is, the stronger the enhancement is.

- If $\beta_t < 0$, the probability of choosing (s_t, a_t) decreases. That is

$$\pi(a_t|s_t, \theta_{t+1}) < \pi(a_t|s_t, \theta_t).$$

The above observations can be proven as follows. When $\theta_{t+1} - \theta_t$ is sufficiently small, it follows from the Taylor expansion that

$$\begin{aligned} \pi(a_t|s_t, \theta_{t+1}) &\approx \pi(a_t|s_t, \theta_t) + (\nabla_\theta \pi(a_t|s_t, \theta_t))^T (\theta_{t+1} - \theta_t) \\ &= \pi(a_t|s_t, \theta_t) + \alpha \beta_t (\nabla_\theta \pi(a_t|s_t, \theta_t))^T (\nabla_\theta \pi(a_t|s_t, \theta_t)) \quad \text{(substituting (9.33))} \\ &= \pi(a_t|s_t, \theta_t) + \alpha \beta_t \|\nabla_\theta \pi(a_t|s_t, \theta_t)\|_2^2. \end{aligned}$$

It is clear that $\pi(a_t|s_t, \theta_{t+1}) \geq \pi(a_t|s_t, \theta_t)$ when $\beta_t \geq 0$ and $\pi(a_t|s_t, \theta_{t+1}) < \pi(a_t|s_t, \theta_t)$ when $\beta_t < 0$.

◇ Second, the algorithm can strike a balance between *exploration* and *exploitation* to a

Algorithm 9.1: Policy Gradient by Monte Carlo (REINFORCE)

```
Initialization: Initial parameter $\theta$; $\gamma \in (0,1)$; $\alpha > 0$.
Goal: Learn an optimal policy for maximizing $J(\theta)$.

For each episode, do
    Generate an episode $\{s_0, a_0, r_1, \dots, s_{T-1}, a_{T-1}, r_T\}$ following $\pi(\theta)$.
    For $t = 0, 1, \dots, T-1$:
        Value update: $q_t(s_t, a_t) = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k$
        Policy update: $\theta \leftarrow \theta + \alpha \nabla_\theta \ln \pi(a_t|s_t, \theta) q_t(s_t, a_t)$
```

certain extent due to the expression of

$$\beta_t = \frac{q_t(s_t, a_t)}{\pi(a_t|s_t, \theta_t)}.$$

On the one hand, β_t is *proportional* to $q_t(s_t, a_t)$. As a result, if the action value of (s_t, a_t) is large, then $\pi(a_t|s_t, \theta_t)$ is enhanced so that the probability of selecting a_t increases. Therefore, the algorithm attempts to *exploit* actions with greater values. One the other hand, β_t is *inversely proportional* to $\pi(a_t|s_t, \theta_t)$ when $q_t(s_t, a_t) > 0$. As a result, if the probability of selecting a_t is small, then $\pi(a_t|s_t, \theta_t)$ is enhanced so that the probability of selecting a_t increases. Therefore, the algorithm attempts to *explore* actions with low probabilities.

Moreover, since (9.32) uses samples to approximate the true gradient in (9.31), it is important to understand how the samples should be obtained.

- ◇ How to sample S? S in the true gradient $\mathbb{E}[\nabla_\theta \ln \pi(A|S, \theta_t) q_\pi(S, A)]$ should obey the distribution η which is either the stationary distribution d_π or the discounted total probability distribution ρ_π in (9.19). Either d_π or ρ_π represents the long-term behavior exhibited under π.
- ◇ How to sample A? A in $\mathbb{E}[\nabla_\theta \ln \pi(A|S, \theta_t) q_\pi(S, A)]$ should obey the distribution of $\pi(A|S, \theta)$. The ideal way to sample A is to select a_t following $\pi(a|s_t, \theta_t)$. Therefore, the policy gradient algorithm is on-policy.

Unfortunately, the ideal ways for sampling S and A are not strictly followed in practice due to their low efficiency of sample usage. A more sample-efficient implementation of (9.32) is given in Algorithm 9.1. In this implementation, an episode is first generated by following $\pi(\theta)$. Then, θ is updated multiple times using every experience sample in the episode.

9.5 Summary

This chapter introduced the policy gradient method, which is the foundation of many modern reinforcement learning algorithms. Policy gradient methods are *policy-based.* It is a big step forward in this book because all the methods in the previous chapters are *value-based.* The basic idea of the policy gradient method is simple. That is to select an appropriate scalar metric and then optimize it via a gradient-ascent algorithm.

The most complicated part of the policy gradient method is the derivation of the gradients of the metrics. That is because we have to distinguish various scenarios with different metrics and discounted/undiscounted cases. Fortunately, the expressions of the gradients in different scenarios are similar. Hence, we summarized the expressions in Theorem 9.1, which is the most important theoretical result in this chapter. For many readers, it is sufficient to be aware of this theorem. Its proof is nontrivial, and it is not required for all readers to study.

The policy gradient algorithm in (9.32) must be properly understood since it is the foundation of many advanced policy gradient algorithms. In the next chapter, this algorithm will be extended to another important policy gradient method called actor-critic.

9.6 Q&A

◇ Q: What is the basic idea of the policy gradient method?

A: The basic idea is simple. That is to define an appropriate scalar metric, derive its gradient, and then use gradient-ascent methods to optimize the metric. The most important theoretical result regarding this method is the policy gradient given in Theorem 9.1.

◇ Q: What is the most complicated part of the policy gradient method?

A: The basic idea of the policy gradient method is simple. However, the derivation procedure of the gradients is quite complicated. That is because we have to distinguish numerous different scenarios. The mathematical derivation procedure in each scenario is nontrivial. It is sufficient for many readers to be familiar with the result in Theorem 9.1 without knowing the proof.

◇ Q: What metrics should be used in the policy gradient method?

A: We introduced three common metrics in this chapter: $\bar{v}_\pi$, $\bar{v}_\pi^0$, and $\bar{r}_\pi$. Since they all lead to similar policy gradients, they all can be adopted in the policy gradient method. More importantly, the expressions in (9.1) and (9.4) are often encountered in the literature.

◇ Q: Why is a natural logarithm function contained in the policy gradient?

A: A natural logarithm function is introduced to express the gradient as an expected value. In this way, we can approximate the true gradient with a stochastic one.

◇ Q: Why do we need to study undiscounted cases when deriving the policy gradient?

A: First, for continuing tasks, it may be inappropriate to introduce the discount rate and we need to consider the undiscounted case. Second, the definition of the average reward $\bar{r}_\pi$ is valid for both discounted and undiscounted cases. While the gradient of $\bar{r}_\pi$ in the discounted case is an approximation, we will see that its gradient in the undiscounted case is more elegant.

◇ Q: What does the policy gradient algorithm in (9.32) do mathematically?

A: To better understand this algorithm, readers are recommended to examine its concise expression in (9.33), which clearly shows that it is a gradient-ascent algorithm for updating the value of $\pi(a_t|s_t, \theta_t)$. That is, when a sample (s_t, a_t) is available, the policy can be updated so that $\pi(a_t|s_t, \theta_{t+1}) \geq \pi(a_t|s_t, \theta_t)$ or $\pi(a_t|s_t, \theta_{t+1}) < \pi(a_t|s_t, \theta_t)$ depending on the coefficients.