第 9 章 策略梯度方法 (Policy Gradient Methods)

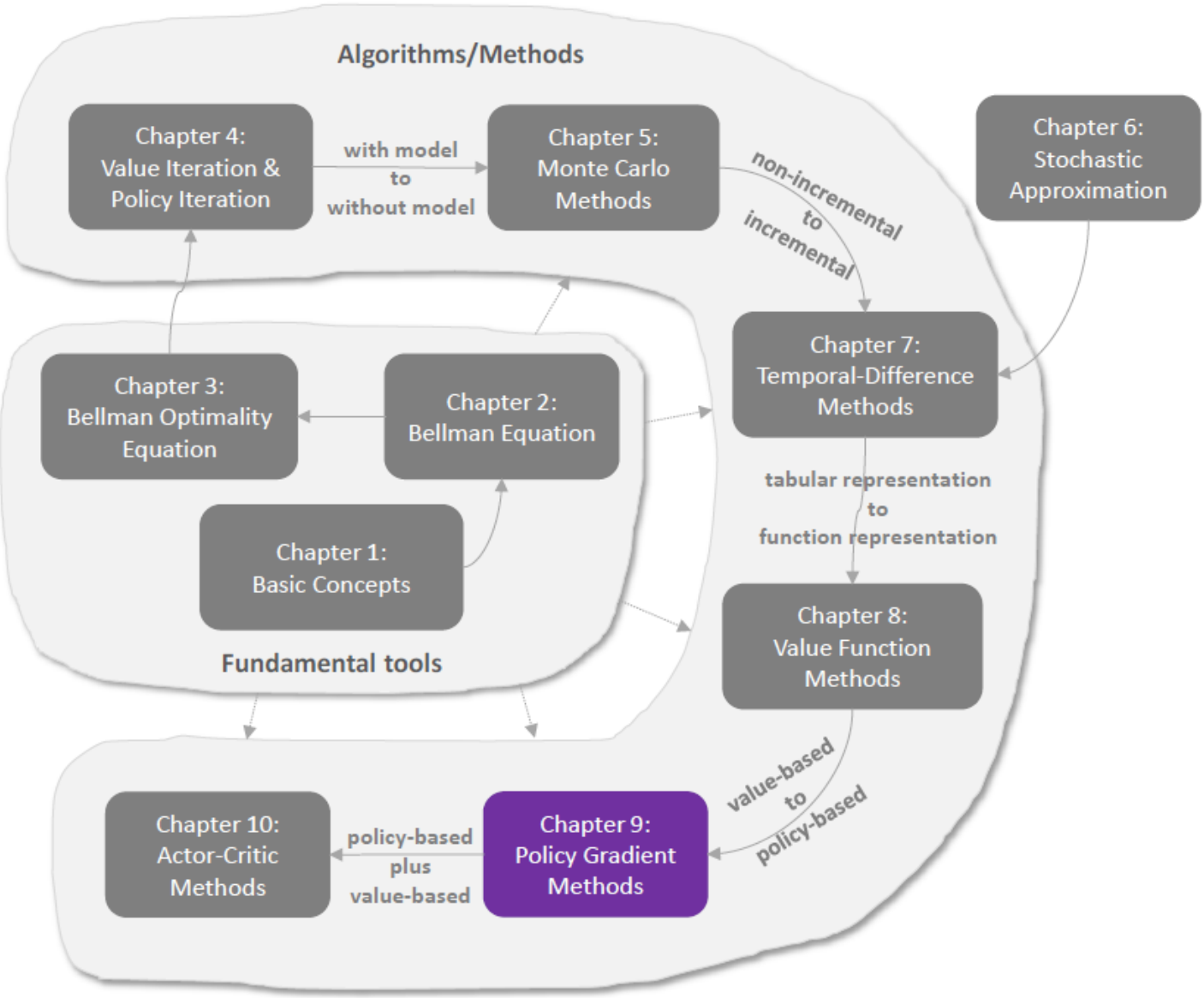

Figure 9.1: Where we are in this book.

图 9.1：我们在本书中的位置。

函数逼近的思想不仅可以应用于表示状态/动作价值（如第 8 章所介绍的），还可以应用于表示策略（如本章所介绍的）。到目前为止，在本书中，策略一直是由表格表示的：所有状态的动作概率都存储在一个表格中（例如，表 9.1）。在本章中，我们将展示策略可以由参数化函数表示，记为 $\pi(a|s,\theta)$，其中 $\theta \in \mathbb{R}^m$ 是一个参数向量。它也可以写成其他形式，例如 $\pi_\theta(a|s)$、$\pi_\theta(a,s)$ 或 $\pi(a,s,\theta)$。

当策略被表示为函数时，可以通过优化某些标量指标来获得最优策略。这种方法被称为 *策略梯度 (policy gradient)*。策略梯度方法是本书的一大进步，因为它是 **基于策略的 (policy-based)**。相比之下，本书之前的所有章节讨论的都是 *基于* **价值的 (value-based)** 方法。策略梯度方法有许多优点。例如，它在处理大的状态/动作空间时效率更高。它具有更强的泛化能力，因此在样本使用方面更高效。

表 9.1：策略的表格表示。共有九个状态，每个状态有五个动作。

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

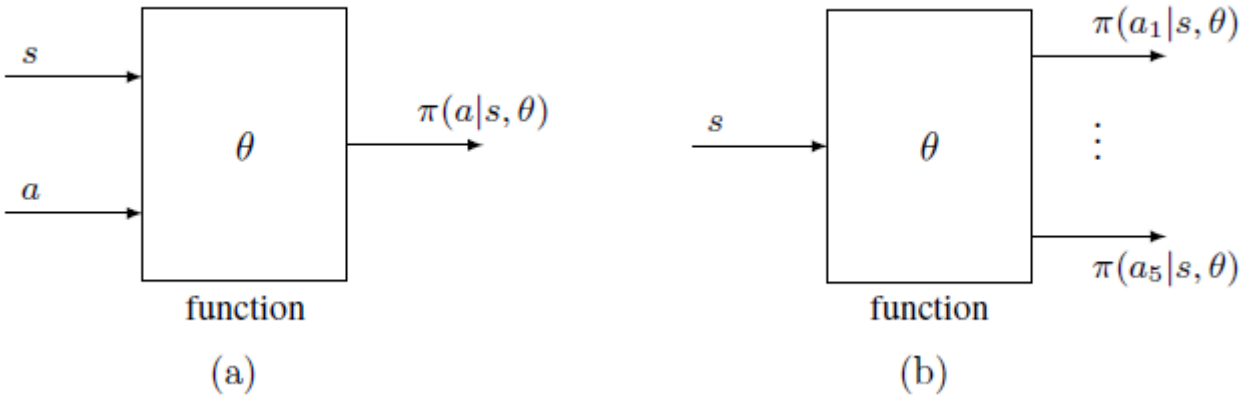

图 9.2：策略的函数表示。函数可能具有不同的结构。

9.1 策略表示：从表格到函数 (Policy representation: From table to function)

当策略的表示形式从表格转换为函数时，有必要阐明这两种表示方法之间的区别。

- **第一，如何定义最优策略？** 当表示为表格时，如果一个策略能最大化 *每个状态的价值 (every state value)*，则该策略被定义为最优的。当表示为函数时，如果一个策略能最大化某些 *标量指标 (scalar metrics)*，则该策略被定义为最优的。
- **第二，如何更新策略？** 当表示为表格时，可以通过直接更改表格中的条目来更新策略。当表示为参数化函数时，策略不再能以这种方式更新。相反，它只能通过改变参数 θ 来更新。
- **第三，如何检索某个动作的概率？** 在表格情况下，可以通过查找表格中的相应条目直接获得动作的概率。在函数表示的情况下，我们需要将 (s, a) 输入到函数中以计算其概率（见图 9.2(a)）。根据函数的结构，我们也可以输入一个状态，然后输出所有动作的概率（见图 9.2(b)）。

策略梯度方法的基本思想总结如下。**假设 $J(\theta)$ 是一个标量指标。可以通过基于梯度的算法优化该指标来获得最优策略：**

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

其中 $\nabla_\theta J$ 是 J 关于 θ 的梯度，t 是时间步，α 是优化率。

基于这个基本思想，我们将在本章的其余部分回答以下三个问题。

- 应该使用什么指标？（9.2 节）
- 如何计算指标的梯度？（9.3 节）
- 如何利用经验样本计算梯度？（9.4 节）

9.2 定义最优策略的指标 (Metrics for defining optimal policies)

如果策略由函数表示，则有两种类型的指标用于**定义最优策略**。一种基于状态价值，另一种基于即时奖励。

指标 1：平均状态价值 (Metric 1: Average state value)

第一个指标是 *平均状态价值 (average state value)* 或简称为 *平均价值 (average value)*。其定义为

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

其中 $d(s)$ 是状态 s 的权重。它满足对于任意 $s \in \mathcal{S}$ 都有 $d(s) \geq 0$ 且 $\sum_{s \in \mathcal{S}} d(s) = 1$。

因此，我们可以将 $d(s)$ 解释为 s 的概率分布。那么，该指标可以写成

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

如何选择分布 d？这是一个重要的问题。有两种情况。

- 第一种也是最简单的情况是 d *独立于 (independent)* 策略 π。在这种情况下，我们特地将 d 记为 d_0，将 $\bar{v}_\pi$ 记为 $\bar{v}_\pi^0$，以表明分布独立于策略。**一种情况是将所有状态视为同等重要，并选择 $d_0(s) = 1/|\mathcal{S}|$。另一种情况是我们只对特定的状态 s_0 感兴趣（例如，智能体总是从 s_0 开始）**。在这种情况下，我们可以设计

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

- 第二种情况是 d *依赖于 (dependent)* 策略 π。在这种情况下，**通常选择 d 为 d_π，即 π 下的 *平稳分布 (stationary distribution)***。d_π 的一个基本性质是它满足

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

其中 P_π 是状态转移概率矩阵。关于平稳分布的更多信息可以在方框 8.1 中找到。

选择 d_π 的解释如下。**平稳分布反映了马尔可夫决策过程在给定策略下的长期行为**。如果在长期运行中某个状态被频繁访问，那么它就更重要，应该获得更高的权重；如果一个状态很少被访问，那么它的重要性就低，应该获得较低的权重。

顾名思义，$\bar{v}_\pi$ 是状态价值的加权平均。不同的 θ 值导致不同的 $\bar{v}_\pi$ 值。我们的最终目标是找到一个最优策略（或等价地找到一个最优的 θ ）来最大化 $\bar{v}_\pi$ 。

- 接下来我们将介绍 $\bar{v}_\pi$ 的另外两个重要的等价表达式。

  假设智能体通过遵循给定策略 $\pi(\theta)$ 收集奖励 $\{R_{t+1}\}_{t=0}^{\infty}$ 。读者经常会在文献中看到以下指标：

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

  乍一看，解释这个指标并非易事。事实上，它等于 $\bar{v}_\pi$ 。为了看出这一点，我们有

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

  上式中的第一个等式是由于全期望定律 (law of total expectation)。第二个等式是根据状态价值的定义（价值函数本来就是“从 s 出发的折扣回报期望”，和这里条件期望的表达完全一致。）。

  指标 $\bar{v}_\pi$ 也可以重写为两个向量的内积。特别地，令

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

那么，我们有

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

当我们分析其梯度时，这个表达式将非常有用。

指标 2：平均奖励 (Metric 2: Average reward)

第二个指标是 **平均单步奖励 (average one-step reward)** 或简称为 **平均奖励 (average reward)** [2, 64, 65]。具体定义为

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

其中 d_π 是平稳分布，且

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

是即时奖励的期望。这里，$r(s,a) \doteq \mathbb{E}[R|s,a] = \sum_r r p(r|s,a)$ 。

接下来我们给出 $\bar{r}_\pi$ 的另外两个重要的等价表达式。

- 假设智能体通过遵循给定策略 $\pi(\theta)$ 收集奖励 $\{R_{t+1}\}_{t=0}^{\infty}$ 。读者经常会在文献中看到的一个常见指标是

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

乍一看，解释这个指标似乎并非易事。事实上，它等于 $\bar{r}_\pi$：

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

(9.5) 的证明在方框 9.1 中给出。

- (9.2) 中的平均奖励 $\bar{r}_\pi$ 也可以写成两个向量的内积。特别地，令

$$r_\pi = [\ldots, r_\pi(s), \ldots]^T \in \mathbb{R}^{|\mathcal{S}|},$$
$$d_\pi = [\ldots, d_\pi(s), \ldots]^T \in \mathbb{R}^{|\mathcal{S}|},$$

其中 $r_\pi(s)$ 在 (9.3) 中定义。那么，显然有

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

当我们推导其梯度时，这个表达式将会非常有用。

方框 9.1：(9.5) 的证明 (Proof of (9.5))

第一步： 我们首先证明以下方程对于任意起始状态 $s_0 \in \mathcal{S}$ 都成立：

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

为此，我们注意到

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

其中最后一个等式是由于切萨罗平均 (Cesaro mean)（也称为切萨罗求和）的性质。特别地，如果 $\{a_k\}_{k=1}^\infty$ 是一个收敛序列，使得 $\lim_{k\to\infty} a_k$ 存在，那么 $\{1/n\sum_{k=1}^{n} a_k\}_{n=1}^\infty$ 也是一个收敛序列，且使得 $\lim_{n\to\infty} 1/n\sum_{k=1}^{n} a_k = \lim_{k\to\infty} a_k$。

接下来我们更仔细地考察 (9.7) 中的 $\mathbb{E}[R_{t+1}|S_0 = s_0]$。根据全期望定律，我们有

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

其中 $p^{(t)}(s|s_0)$ 表示恰好使用 t 步从 s_0 转移到 s 的概率。上式中的第二个等式是由于马尔可夫无记忆性质 (Markov memoryless property)：在下一个时间步获得的奖励仅取决于当前状态，而不是之前的状态。

注意到

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

根据平稳分布的定义。因此，起始状态 s_0 无关紧要。那么，我们有

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

将上述方程代入 (9.7) 即可得到 (9.6)。

第 2 步： 考虑任意状态分布 d。根据全期望定律，我们有

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

由于 (9.6) 对任意起始状态都成立，将 (9.6) 代入上述方程可得

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

证明完毕。

一些注记 (Some remarks)

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

表 9.2：$\bar{v}_\pi$ 和 $\bar{r}_\pi$ 的不同但等价的表达式总结。

到目前为止，我们介绍了两种类型的指标：$\bar{v}_\pi$ 和 $\bar{r}_\pi$。每个指标都有几种不同但等价的表达式。它们总结在表 9.2 中。我们有时使用 $\bar{v}_\pi$ 特指状态分布为平稳分布 d_π 的情况，并使用 $\bar{v}_\pi^0$ 指代 d_0 独立于 π 的情况。关于这些指标的一些注记如下。

- 所有这些指标都是 π 的函数。由于 π 是由 θ 参数化的，因此这些指标是 θ 的函数。换句话说，不同的 θ 值可以生成不同的指标值。因此，**我们可以搜索 θ 的最优值来最大化这些指标。这是策略梯度方法的基本思想。**
- 这两个指标 $\bar{v}_\pi$ 和 $\bar{r}_\pi$ 在 $\gamma < 1$ 的折扣情况下是等价的。特别地，可以证明

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

上述方程表明这两个指标可以同时被最大化。该方程的证明将在稍后的引理 9.1 中给出。

9.3 指标的梯度 (Gradients of the metrics)

给定上一节介绍的指标，我们可以使用基于梯度的方法来最大化它们。为此，我们需要首先计算这些指标的梯度。本章最重要的理论结果是以下定理。

定理 9.1（策略梯度定理）。 $J(\theta)$ 的梯度为

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

其中 η 是一个状态分布，$\nabla_\theta \pi$ 是 π 关于 θ 的梯度。此外，(9.8) 有一个用期望表示的紧凑形式：

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

其中 $\ln$ 是自然对数。

关于定理 9.1 的一些重要注记如下。

- 值得注意的是，**定理 9.1 是对定理 9.2、定理 9.3 和定理 9.5 中结果的总结**。这三个定理解决了涉及不同指标以及折扣/非折扣情况的不同场景。所有这些场景中的梯度都具有类似的表达式，因此被归纳在定理 9.1 中。$J(\theta)$ 和 η 的具体表达式并未在定理 9.1 中给出，而是在定理 9.2、定理 9.3 和定理 9.5 中给出。特别地，$J(\theta)$ 可以是 $\bar{v}_\pi^0$、$\bar{v}_\pi$ 或 $\bar{r}_\pi$。(9.8) 中的等号可能会变成严格等号或近似等号。分布 η 在不同场景下也有所不同。

(9.8) 是如何给出的？（推导过程-9.3.1有类似的推导）

公式 (9.8) 的核心在于**如何处理目标函数 $J(\theta)$ 的梯度**。通常 $J(\theta)$ 被定义为初始状态 s_0 的价值 $V_\pi(s_0)$，或者是平均奖励。为了方便推导，假设 $J(\theta) = V_\pi(s_0)$。

1. 从价值函数的贝尔曼方程出发

回顾状态价值函数 $V_\pi(s)$ 的定义：

$$V_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s,\theta) q_\pi(s,a)$$

我们对参数 θ 求梯度。注意，这里有两部分依赖 θ：一是策略 π 本身，二是动作价值 q_π（因为 q 依赖于未来的策略）。利用乘法法则：

$$\nabla_\theta V_\pi(s) = \sum_a [\nabla_\theta \pi(a|s,\theta) q_\pi(s,a) + \pi(a|s,\theta) \nabla_\theta q_\pi(s,a)]$$

2. 展开 q_π 的梯度

动作价值函数 $q_\pi(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) V_\pi(s')$。

对其求导（奖励 R 和转移概率 P 与 θ 无关）：

$$\nabla_\theta q_\pi(s,a) = \gamma \sum_{s'} P(s'|s,a) \nabla_\theta V_\pi(s')$$

3. 代回并递归

将第2步的结果代回第1步的式子：

$$\nabla_\theta V_\pi(s) = \sum_a \nabla_\theta \pi(a|s,\theta) q_\pi(s,a) + \gamma \sum_a \pi(a|s,\theta) \sum_{s'} P(s'|s,a) \nabla_\theta V_\pi(s')$$

为了简化，令 $\phi(s) = \sum_a \nabla_\theta \pi(a|s,\theta) q_\pi(s,a)$，这是当前步的梯度贡献。

剩下的部分表示从状态 s 按照策略 π 转移到下一个状态 s' 的期望。式子变成了递归形式：

$$\nabla_\theta V_\pi(s) = \phi(s) + \gamma \sum_{s'} P(s \to s'|\pi) \nabla_\theta V_\pi(s')$$

4. 展开递归 (Unrolling)

如果我们不断展开这项递归（即 $\nabla V(s')$ 又是 $\phi(s') + \dots$）：

$$\nabla_\theta V_\pi(s_0) = \phi(s_0) + \gamma \sum_{s_1} P(s_0 \to s_1) [\phi(s_1) + \gamma \sum_{s_2} P(s_1 \to s_2) \dots]$$

$$= \sum_s \left(\sum_{t=0}^{\infty} \gamma^t P(s_0 \to s, t, \pi) \right) \phi(s)$$

其中 $P(s_0 \to s, t, \pi)$ 表示在策略 π 下，t 步后到达状态 s 的概率。

5. 引入状态分布 $\eta(s)$

括号里的部分 $\sum_{t=0}^{\infty} \gamma^t P(s_0 \to s, t, \pi)$ 就是所谓的 **Discounted State Distribution（折现状态分布）**，或者是平稳分布（取决于具体定义，这里统称为 $\eta(s)$）。

它表示了：**在当前策略下，智能体出现在状态 s 的频率（或概率权重）。**

于是我们得到了 (9.8)：

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

直观理解 (9.8)：总梯度 = (在这个状态出现的概率) × (在这个状态下动作概率变化带来的价值提升)，对所有状态求和。

梯度的推导是策略梯度方法中最复杂的部分。对于许多读者来说，只要熟悉定理 9.1 中的结果而无需了解证明就足够了。本节其余部分给出的推导细节在数学上是密集的。建议读者根据自己的兴趣选择性地学习。

(9.9) 中的表达式比 (9.8) 更优，因为它被表示为期望的形式。我们将在 9.4 节中展示，这个真实梯度（true gradient）可以通过随机梯度（stochastic gradient）来近似。

为什么 (9.8) 可以表示为 (9.9)？证明如下。根据期望的定义，(9.8) 可以重写为：

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

此外，$\ln \pi(a|s,\theta)$ 的梯度是(根据链式法则)：

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

由此可得：

$$\nabla_\theta \pi(a|s,\theta) = \pi(a|s,\theta) \nabla_\theta \ln \pi(a|s,\theta). \quad (9.11)$$

将 (9.11) 代入 (9.10) 可得：

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

值得注意的是，$\pi(a|s,\theta)$ 对于所有的 (s,a) 必须是**正值**（positive），以确保 $\ln \pi(a|s,\theta)$ 有效。这可以通过使用 **softmax 函数**（softmax functions）来实现：

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

其中 $h(s,a,\theta)$ 是一个函数，表示在状态 s 下选择动作 a 的偏好程度。(9.12) 中的策略满足 $\pi(a|s,\theta) \in (0,1)$ 且对于任意 $s \in \mathcal{S}$ 都有 $\sum_{a\in\mathcal{A}} \pi(a|s,\theta) = 1$。这种策略可以通过神经网络来实现。

网络的输入是 s 。输出层是一个 softmax 层，使得网络输出所有 a 对应的 $\pi(a|s,\theta)$ ，并且输出之和等于 1。参见图 9.2(b) 的图示。

由于对于所有的 a 都有 $\pi(a|s,\theta)>0$ ，因此该策略是**随机的**（stochastic），也因此具有**探索性**（exploratory）。该策略并不直接告知应采取哪个动作。相反，动作应根据策略的概率分布生成。

9.3.1 折扣情形下梯度的推导 (Derivation of the gradients in the discounted case)

接下来我们推导折扣情形下（其中 $\gamma\in(0,1)$ ）指标的梯度。折扣情形下的状态价值和动作价值定义为

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

我们有 $v_\pi(s)=\sum_{a\in\mathcal{A}}\pi(a|s,\theta)q_\pi(s,a)$ ，且状态价值满足贝尔曼方程。

首先，我们证明 $\bar{v}_\pi(\theta)$ 和 $\bar{r}_\pi(\theta)$ 是等价的指标。

引理 9.1 ($\bar{v}_\pi(\theta)$ 和 $\bar{r}_\pi(\theta)$ 之间的等价性)。 在 $\gamma\in(0,1)$ 的折扣情形下，成立

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

证明。 注意 $\bar{v}_\pi(\theta)=d_\pi^Tv_\pi$ 且 $\bar{r}_\pi(\theta)=d_\pi^Tr_\pi$ ，其中 v_π 和 r_π 满足贝尔曼方程 $v_\pi=r_\pi+\gamma P_\pi v_\pi$ 。在贝尔曼方程两边左乘 d_π^T 可得

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

这蕴含了 (9.13)。 □

其次，下面的引理给出了对于任意 s 的 $v_\pi(s)$ 的梯度。

引理 9.2 ($v_\pi(s)$ 的梯度)。 在折扣情形下，对于任意 $s\in\mathcal{S}$ 都有

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

其中

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

是在策略 π 下从 s 转移到 s' 的折扣总概率。这里，$[\cdot]_{ss'}$ 表示第 s 行和第 s' 列的元素，$[P_\pi^k]_{ss'}$ 是在策略 π 下恰好使用 k 步从 s 转移到 s' 的概率。

方框 9.2：引理 9.2 的证明

首先，对于任意 $s\in\mathcal{S}$ ，有：

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

其中 $q_\pi(s,a)$ 是动作价值，由下式给出：

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

由于 $r(s,a) = \sum_r r p(r|s,a)$ 与 θ 无关，我们有：

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

将此结果代入 (9.15) 可得：

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

值得注意的是，$\nabla_\theta v_\pi$ 出现在上述等式的两边。计算它的一种方法是使用**展开技巧**（unrolling technique）[64]。这里，我们使用另一种基于**矩阵-向量形式**（matrix-vector form）的方法，我们认为这更直观易懂。具体来说，令

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

$$\nabla_\theta v_\pi(s) = u + \gamma \sum_{a\in\mathcal{A}} \pi(a|s,\theta) \sum_{s'\in\mathcal{S}} p(s'|s,a)\nabla_\theta v_\pi(s'). \quad (9.16)$$

由于

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

等式 (9.16) 可以写成矩阵-向量形式如下：

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

可以简洁地写为

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

这里，$n = |\mathcal{S}|$，m 是参数向量 θ 的维度。方程中出现克罗内克积 (Kronecker product) $\otimes$ 的原因是因为 $\nabla_\theta v_\pi(s)$ 是一个向量。上述方程是关于 $\nabla_\theta v_\pi$ 的线性方程，其解为

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

对于任意状态 s ，由 (9.17) 可得（看下一个框以明确 $u(s')$ ）

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

数量 $[(I_n - \gamma P_\pi)^{-1}]_{ss'}$ 具有清晰的概率解释。特别地，由于 $(I_n - \gamma P_\pi)^{-1} = I + \gamma P_\pi + \gamma^2 P_\pi^2 + \dots$，我们有

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

注意 $[P_\pi^k]_{ss'}$ 是恰好使用 k 步从 s 转移到 s' 的概率（见方框 8.1）。因此，$[(I_n - \gamma P_\pi)^{-1}]_{ss'}$ 是使用任意步数从 s 转移到 s' 的折扣总概率。通过记 $[(I_n - \gamma P_\pi)^{-1}]_{ss'} \doteq \Pr_\pi(s'|s)$ ，方程 (9.18) 变为 (9.14)。

矩阵-向量形式的进一步明晰

让我们通过代数推导把这一点彻底看清楚。为了不被复杂的求和符号淹没，我们使用简化一点的记号：

记 $P(s'|s) = \sum_a \pi(a|s) p(s'|s,a)$ 为状态转移概率。

回顾递归公式 (9.16) 的核心结构：

$$\nabla v(s) = u(s) + \gamma \sum_{s'} P(s'|s) \nabla v(s')$$

这里的 $\nabla v(s')$ 实际上和左边的 $\nabla v(s)$ 拥有完全相同的定义结构，只不过是针对状态 s' 的。

1. 第一层展开（代入 s' 的定义）

根据定义，对于状态 s' ，它的梯度是：

$$\nabla v(s') = u(s') + \gamma \sum_{s''} P(s''|s') \nabla v(s'')$$

一定要注意：这里的 $u(s')$ 正是你所问的“对应的 u 项”。

现在，我们将这个 $\nabla v(s')$ 代回最上面的公式中：

$$\nabla v(s) = u(s) + \gamma \sum_{s'} P(s'|s) \left[u(s') + \gamma \sum_{s''} P(s''|s') \nabla v(s'') \right]$$

$$= \underbrace{u(s)}_{\text{第0步的贡献}} + \underbrace{\gamma \sum_{s'} P(s'|s) u(s')}_{\text{第1步的贡献}} + \gamma^2 \sum_{s'} \sum_{s''} P(s'|s) P(s''|s') \nabla v(s'')$$

看到了吗？当你展开 $\nabla v(s')$ 时，$u(s')$ **立刻就被“释放”出来了**，作为第二项出现在了等式中。

2. 继续展开（第二层、第三层...）

剩余的尾巴项包含 $\nabla v(s'')$。同样地，$\nabla v(s'')$ 内部也包含它自己的 $u(s'')$。如果我们继续展开：

$$\nabla v(s'') = u(s'') + \dots$$

代入后，整个式子变成了：

$$\nabla v(s) = u(s) + \gamma \sum_{s'} P(s'|s) u(s') + \gamma^2 \sum_{s''} P(s''|s \text{ via } s') u(s'') + \dots$$

这个无限级数的每一项都有明确的物理意义：

1. **第一项** $u(s)$：我在当前状态 s 产生的直接梯度贡献。
2. **第二项** $\gamma \sum P(s'|s) u(s')$：我在下一步可能到达的所有状态 s' 中，它们产生的 $u(s')$ 贡献（乘上到达概率和折扣）。
3. **第三项** $\dots u(s'')$：我在下下步可能到达的状态 s'' 中，它们产生的 $u(s'')$ 贡献。

3. 总结与对应关系

所以，你的理解是完全正确的：

> **递归公式右边的 $\nabla v(s')$ 就是一个“压缩包”。**

- 当你只看 (9.16) 时，它是一个未解压缩的状态：当前梯度 = 当前 u + 未来的梯度总和。
- 当你解这个方程（无论是用矩阵求逆，还是用级数展开）得到 (9.18) 时，你实际上就是把这个“压缩包”一层层打开了。
- 打开后，你发现里面**包含了所有未来可能到达的状态 x 对应的 $u(x)$**。

这也是为什么最终公式 (9.18) 变成了对所有状态的加权求和：

$$\nabla v(s) = \sum_{x \in \mathcal{S}} \Pr(s \to x) \cdot u(x)$$

- **(9.16) 是“局部视角”**：当前的梯度 = 当前的 $u(s)$ + 下一步的梯度。
- **(9.18) 是“全局视角”**：当前的梯度 = 所有未来可能到达的状态 s' 的 $u(s')$ 的累加（加权和）。

定义变量：

- $\nabla_\theta v_\pi$ 是一个长向量（大小为 $nm \times 1$），它是由所有状态的梯度向量堆叠而成的：

$$\nabla_\theta v_\pi = \begin{bmatrix} \nabla_\theta v_\pi(s_1) \\ \nabla_\theta v_\pi(s_2) \\ \vdots \\ \nabla_\theta v_\pi(s_n) \end{bmatrix}$$

- 同理，u 也是一个长向量：

$$u = \begin{bmatrix} u(s_1) \\ u(s_2) \\ \vdots \\ u(s_n) \end{bmatrix}$$

- 令矩阵 $M = (I_n - \gamma P_\pi)^{-1}$。这是一个 $n \times n$ 的标量矩阵。$M_{ss'}$ 表示第 s 行第 s' 列的元素，也就是引理中提到的折扣总概率。

方程 (9.17) 是：

$$\nabla_\theta v_\pi = [M \otimes I_m]u$$

这里的 $\otimes$ 是克罗内克积（Kronecker Product）。$M \otimes I_m$ 会生成一个巨大的分块矩阵。对于 M 中的每一个元素 $M_{ss'}$，它都会扩展成一个 $m \times m$ 的对角阵 $M_{ss'}I_m$。

这个大矩阵的样子如下（假设状态是 $s_1, s_2, \ldots$）：

$$M \otimes I_m = \begin{bmatrix} M_{s_1 s_1} I_m & M_{s_1 s_2} I_m & \cdots \\ M_{s_2 s_1} I_m & M_{s_2 s_2} I_m & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}$$

进行矩阵-向量乘法：

现在我们计算 $\nabla_\theta v_\pi = (M \otimes I_m)u$。

我们要看结果向量中对应状态 s 的那一部分块（即 $\nabla_\theta v_\pi(s)$）。根据矩阵乘法规则（行乘以列），结果的第 s 个块等于大矩阵的第 s 块行乘以整个向量 u。

$$\nabla_\theta v_\pi(s) = \sum_{s' \in \mathcal{S}} (\text{对应 } s, s' \text{ 的矩阵块}) \times (\text{对应 } s' \text{ 的向量块})$$

代入具体的值：

- 对应 s, s' 的矩阵块是 $M_{ss'}I_m$。
- 对应 s' 的向量块是 $u(s')$。

于是：

$$\nabla_\theta v_\pi(s) = \sum_{s' \in \mathcal{S}} (M_{ss'}I_m)u(s')$$

由于 I_m 是单位阵，$I_m u(s') = u(s')$，且 $M_{ss'}$ 是一个标量（概率值），我们可以把它写在前面：

$$\nabla_\theta v_\pi(s) = \sum_{s' \in \mathcal{S}} M_{ss'} u(s')$$

把 $M_{ss'}$ 换回原来的符号 $[(I_n - \gamma P_\pi)^{-1}]_{ss'}$，这就是 **(9.18)**。

矩阵 M (2×2)： 这是一个标量矩阵，里面全是概率相关的数值。

$$M = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix}$$

矩阵 I_m (2×2)： 这是单位阵。

$$I_m = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

运算 $M \otimes I_m$： 它的结果是一个巨大的矩阵。维数变成了 $(n \times m) \times (n \times m) = 4 \times 4$。规则是把 M 的每一个元素 m_{ij} 变成 $m_{ij} \cdot I_m$：

$$M \otimes I_m = \begin{bmatrix} m_{11} I_m & m_{12} I_m \\ m_{21} I_m & m_{22} I_m \end{bmatrix}$$

展开看，就是：

$$\left[\begin{array}{cc|cc} m_{11} \cdot 1 & m_{11} \cdot 0 & m_{12} \cdot 1 & m_{12} \cdot 0 \\ m_{11} \cdot 0 & m_{11} \cdot 1 & m_{12} \cdot 0 & m_{12} \cdot 1 \\ \hline m_{21} \cdot 1 & m_{21} \cdot 0 & m_{22} \cdot 1 & m_{22} \cdot 0 \\ m_{21} \cdot 0 & m_{21} \cdot 1 & m_{22} \cdot 0 & m_{22} \cdot 1 \end{array}\right]$$

图x M 与 I_m 外积的例子

有了引理 9.2 的结果，我们准备推导 $\bar{v}_\pi^0$ 的梯度。

定理 9.2（折扣情形下 $\bar{v}_\pi^0$ 的梯度）。 在 $\gamma \in (0,1)$ 的折扣情形下，$\bar{v}_\pi^0 = d_0^T v_\pi$ 的梯度为

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

其中 $S \sim \rho_\pi$ 且 $A \sim \pi(S,\theta)$。这里，状态分布 ρ_π 为

$$\rho_\pi(s) = \sum_{s' \in \mathcal{S}} d_0(s') \mathrm{Pr}_\pi(s|s'), \quad s \in \mathcal{S}, \quad (9.19)$$

其中 $\Pr_\pi(s|s') = \sum_{k=0}^{\infty} \gamma^k [P_\pi^k]_{s's} = [(I-\gamma P_\pi)^{-1}]_{s's}$ **是在策略** π **下从** s' **转移到** s **的折扣总概率**

参考公式

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

方框 9.3：定理 9.2 的证明

由于 $d_0(s)$ 与 π 无关，我们有：

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

将引理 9.2（$\nabla_\theta v_\pi(s) = \sum_{s'\in\mathcal{S}} \Pr_\pi(s'|s) \sum_{a\in\mathcal{A}} \nabla_\theta \pi(a|s',\theta) q_\pi(s',a), \quad (9.14)$）中给出的 $\nabla_\theta v_\pi(s)$ 表达式代入上式可得：

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

其中 $S\sim\rho_\pi$ 且 $A\sim\pi(S,\theta)$。证明完毕。

利用引理 9.1（$\bar{r}_\pi = (1-\gamma)\bar{v}_\pi. \quad (9.13)$）和引理 9.2，我们可以推导出 $\bar{r}_\pi$ 和 $\bar{v}_\pi$ 的梯度。

定理 9.3（折扣情形下 $\bar{r}_\pi$ 和 $\bar{v}_\pi$ 的梯度）。 在 $\gamma\in(0,1)$ 的折扣情形下，$\bar{r}_\pi$ 和 $\bar{v}_\pi$ 的梯度为：

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

其中 $S\sim d_\pi$ 且 $A\sim\pi(S,\theta)$。这里，当 γ 越接近 1 时，近似越准确。

方框 9.4：定理 9.3 的证明

由 $\bar{v}_\pi$ 的定义可得：

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

该方程包含两项。一方面，将 (9.17) 中给出的 $\nabla_\theta v_\pi$ 表达式代入第二项，可得：

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

注意到：

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

这一点可以通过在方程两边同时乘以 $(I_n - \gamma P_\pi)$ 来轻松验证。因此，(9.21) 变为：

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

另一方面，(9.20) 的第一项涉及 $\nabla_\theta d_\pi$。然而，由于第二项包含 $\frac{1}{1-\gamma}$，当 $\gamma \to 1$ 时，第二项变为**主导项**（dominant），而第一项变得**可忽略**（negligible）。因此：

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

此外，由 $\bar{r}_\pi = (1-\gamma)\bar{v}_\pi$ 可得：

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

上述方程中的近似要求当 $\gamma \to 1$ 时第一项不会趋于无穷大。更多信息可以在 [66, Section 4] 中找到。

9.3.2 无折扣情形下的梯度推导

接下来我们展示如何在 $\gamma = 1$ 的无折扣情形（undiscounted case）下计算指标的梯度。读者可能会想，为什么我们在本书中到目前为止只考虑了折扣情形，却突然开始考虑无折扣情形。事实上，平均奖励 $\bar{r}_\pi$ 的定义对于折扣和无折扣情形都是有效的。虽然折扣情形下 $\bar{r}_\pi$ 的梯度是一个近似值，但我们将看到它在无折扣情形下的梯度更加优雅。

状态价值与泊松方程

在无折扣情形下，有必要重新定义状态和动作价值。**由于无折扣的奖励之和 $\mathbb{E}[R_{t+1} + R_{t+2} + R_{t+3} + \ldots | S_t = s]$ 可能会发散，因此状态和动作价值以一种特殊的方式定义 [64]：**

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

其中 $\bar{r}_\pi$ 是平均奖励，当 π 给定时它也就确定了。文献中对于 $v_\pi(s)$ 有不同的称呼，例如**差分奖励**（differential reward）[65] 或 **偏差**（bias）[2, Section 8.2.1]。可以验证，上述定义的状态价值满足以下类似贝尔曼方程的等式：

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

由于 $v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s,\theta) q_\pi(s,a)$，因此成立 $q_\pi(s,a) = \sum_r p(r|s,a)(r - \bar{r}_\pi) + \sum_{s'} p(s'|s,a) v_\pi(s')$。(9.22) 的矩阵-向量形式为

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

其中 $\mathbf{1}_n = [1, \ldots, 1]^T \in \mathbb{R}^n$。方程 (9.23) 与贝尔曼方程相似，它有一个特定的名称叫做**泊松方程**（Poisson equation）[65, 67]。

如何从泊松方程中解出 v_π？答案在下面的定理中给出。

定理 9.4（泊松方程的解）。 令

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

那么，v_π^* 是泊松方程 (9.23) 的一个解。此外，泊松方程的任意解都具有以下形式：

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

其中 $c \in \mathbb{R}$。该定理表明泊松方程的解可能是不唯一的。

方框 9.5：定理 9.4 的证明

我们分三个步骤进行证明。

◇ 步骤 1：证明 (9.24) 中的 v_π^* 是泊松方程的一个解。

为了简便起见，令

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

那么，$v_\pi^* = A^{-1} r_\pi$。A 是可逆的事实将在步骤 3 中证明。将 $v_\pi^* = A^{-1} r_\pi$ 和 $\bar{r}_\pi = d_\pi^T r_\pi$ 代入 (9.23) 可得

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

如下文证明，该等式成立。对该方程进行整理可得 $(-A^{-1} + I_n - \mathbf{1}_n d_\pi^T + P_\pi A^{-1}) r_\pi = 0$，从而有，

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

上述方程括号中的项为零，因为:

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

。

因此，(9.24) 中的 v_π^* 是一个解。

◇ 步骤 2：解的通项表达式。

将 $\bar{r}_\pi = d_\pi^T r_\pi$ 代入 (9.23) 可得

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

进而得到

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

值得注意的是，$I_n - P_\pi$ 是**奇异的**（singular），因为对于任意 π，都有 $(I_n - P_\pi)\mathbf{1}_n = 0$。因此，(9.26) 的解不是唯一的：如果 v_π^* 是一个解，那么对于任意 $x \in \mathrm{Null}(I_n - P_\pi)$，$v_\pi^* + x$ 也是一个解。当 P_π 是**不可约的**（irreducible）时，$\mathrm{Null}(I_n - P_\pi) = \mathrm{span}\{\mathbf{1}_n\}$。此时，泊松方程的任意解都具有 $v_\pi^* + c\mathbf{1}_n$ 的形式，其中 $c \in \mathbb{R}$。（看注释）

◇ 步骤 3：证明 $A = I_n - P_\pi + \mathbf{1}_n d_\pi^T$ 是可逆的。

由于 v_π^* 涉及 A^{-1}，因此有必要证明 A 是可逆的。分析过程总结在下面的引理中。

引理 9.3。 矩阵 $I_n - P_\pi + \mathbf{1}_n d_\pi^T$ 是可逆的，其逆矩阵为

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

证明。 首先，我们陈述一些预备事实而不加证明。令 $\rho(M)$ 为矩阵 M 的**谱半径**（spectral radius）。如果 $\rho(M) < 1$，则 $I - M$ 是可逆的。此外，$\rho(M) < 1$ 当且仅当 $\lim_{k\to\infty} M^k = 0$。

基于上述事实，我们接下来证明 $\lim_{k\to\infty} (P_\pi - \mathbf{1}_n d_\pi^T)^k \to 0$，进而 $I_n - (P_\pi - \mathbf{1}_n d_\pi^T)$ 的可逆性便直接成立。为此，我们注意到

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

这可以通过归纳法证明。例如，当 $k=1$ 时，等式成立。当 $k=2$ 时，我们有

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

其中最后一个等式是因为 $P_\pi \mathbf{1}_n = \mathbf{1}_n$ ， $d_\pi^T P_\pi = d_\pi^T$ ，以及 $d_\pi^T \mathbf{1}_n = 1$ 。 $k \geq 3$ 的情况也可以类似证明。

由于 d_π 是状态的平稳分布，因此成立 $\lim_{k\to\infty} P_\pi^k = \mathbf{1}_n d_\pi^T$ （见方框 8.1）。因此，(9.27) 意味着

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

结果是， $\rho(P_\pi - \mathbf{1}_n d_\pi^T) < 1$ ，因此 $I_n - (P_\pi - \mathbf{1}_n d_\pi^T)$ 是可逆的。此外，

该矩阵的逆由下式给出：

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

证明完毕。

引理 9.3 的证明受到 [66] 的启发。然而，[66] 中给出的结果 $(I_n - P_\pi + \mathbf{1}_n d_\pi^T)^{-1} = \sum_{k=0}^{\infty} (P_\pi^k - \mathbf{1}_n d_\pi^T)$ （[66] 中方程 (16) 上方的陈述）是不准确的，因为 $\sum_{k=0}^{\infty} (P_\pi^k - \mathbf{1}_n d_\pi^T)$ 是**奇异的**（singular），毕竟 $\sum_{k=0}^{\infty} (P_\pi^k - \mathbf{1}_n d_\pi^T)\mathbf{1}_n = 0$ 。引理 9.3 修正了这一不准确之处。

梯度的推导 (Derivation of gradients)

尽管如定理 9.4 所示，在无折扣情形下 v_π 的值不是唯一的，但 $\bar{r}_\pi$ 的值是唯一的。特别地，由泊松方程可得：

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

值得注意的是，未定值 c 被消去了，因此 $\bar{r}_\pi$ 是唯一的。因此，我们可以计算无折扣情形下 $\bar{r}_\pi$ 的梯度。此外，由于 v_π 不是唯一的，$\bar{v}_\pi$ 也不是唯一的。我们在无折扣情形下不研究 $\bar{v}_\pi$ 的梯度。对于感兴趣的读者，值得一提的是，我们可以添加更多约束来从泊松方程中唯一地解出 v_π。例如，通过假设存在一个**常返状态**（recurrent state），该常返状态的状态价值可以被确定 [65, Section II]，从而 c 也可以被确定。还有其他方法可以唯一确定 v_π。参见例如 [2] 中的方程 (8.6.5)-(8.6.7)。

无折扣情形下 $\bar{r}_\pi$ 的梯度如下给出。

定理 9.5（无折扣情形下 $\bar{r}_\pi$ 的梯度）。 在无折扣情形下，平均奖励 $\bar{r}_\pi$ 的梯度为：

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

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

其中 $S \sim d_\pi$ 且 $A \sim \pi(S,\theta)$。

与定理 9.3 中展示的折扣情形相比，无折扣情形下 $\bar{r}_\pi$ 的梯度更加优雅，**因为 (9.28) 是严格成立的**，并且 S 服从平稳分布。

方框 9.6：定理 9.5 的证明

首先，由 $v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s,\theta) q_\pi(s,a)$ 可得：

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

其中 $q_\pi(s,a)$ 是满足下式的动作价值：

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

由于 $r(s,a) = \sum_r r p(r|s,a)$ 与 θ 无关，我们有：

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

将此结果代入 (9.29) 可得：

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

令

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

由于 $\sum_{a\in\mathcal{A}}\pi(a|s,\theta)\sum_{s'\in\mathcal{S}} p(s'|s,a)\nabla_\theta v_\pi(s') = \sum_{s'\in\mathcal{S}} p(s'|s)\nabla_\theta v_\pi(s')$，方程 (9.30) 可以写成如下矩阵-向量形式：

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

其中 $n = |\mathcal{S}|$，m 是 θ 的维数，$\otimes$ 是克罗内克积（Kronecker product）。上述方程可以简写为

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

因此

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

在上述方程两边同时乘以 $d_\pi^T\otimes I_m$ 可得

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

这意味着

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

9.4 蒙特卡洛策略梯度（REINFORCE）

利用定理 9.1 中给出的梯度，我们接下来展示如何使用基于梯度的方法来优化指标以获得最优策略。

用于最大化 $J(\theta)$ 的梯度上升算法为

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

其中 $\alpha > 0$ 是一个常数学习率（learning rate）。由于 (9.31) 中的真实梯度是未知的，我们

我们可以用随机梯度代替真实梯度，从而得到以下算法：

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

其中 $q_t(s_t,a_t)$ 是 $q_\pi(s_t,a_t)$ 的近似值。如果 $q_t(s_t,a_t)$ 是通过蒙特卡洛估计得到的，该算法被称为 **REINFORCE** [68] 或 **蒙特卡洛策略梯度**（Monte Carlo policy gradient），这是最早也是最简单的策略梯度算法之一。

(9.32) 中的算法非常重要，因为许多其他策略梯度算法都可以通过扩展它来得到。接下来我们更仔细地审视 (9.32) 的解释。

由于 $\nabla_\theta \ln \pi(a_t|s_t,\theta_t) = \dfrac{\nabla_\theta \pi(a_t|s_t,\theta_t)}{\pi(a_t|s_t,\theta_t)}$，我们可以将 (9.32) 重写为

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

这可以进一步简洁地写为

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

从这个方程中可以看出两个重要的解释。

◇ 首先，由于 (9.33) 是一个简单的梯度上升算法，可以得出以下观察结果。

- 如果 $\beta_t \geq 0$，则选择 (s_t,a_t) 的概率会增加。即

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

  β_t 越大，这种增强效果越强。

- 如果 $\beta_t < 0$，则选择 (s_t,a_t) 的概率会降低。即

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

上述观察结果证明如下。当 $\theta_{t+1} - \theta_t$ 足够小时，根据泰勒展开可得

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

很明显，当 $\beta_t \geq 0$ 时，$\pi(a_t|s_t,\theta_{t+1}) \geq \pi(a_t|s_t,\theta_t)$；

当 $\beta_t < 0$ 时，$\pi(a_t|s_t,\theta_{t+1}) < \pi(a_t|s_t,\theta_t)$。(很接近重要性采样)

算法 9.1：基于蒙特卡洛的策略梯度 (REINFORCE)

初始化：初始参数 θ ； $\gamma \in (0,1)$ ； $\alpha > 0$ 。

目标：学习一个最大化 $J(\theta)$ 的最优策略。

对每个回合（episode），执行

根据 $\pi(\theta)$ 生成一个回合 $\{s_0, a_0, r_1, \ldots, s_{T-1}, a_{T-1}, r_T\}$ 。

对 $t = 0, 1, \ldots, T-1$:

价值更新： $q_t(s_t, a_t) = \sum_{k=t+1}^{T} \gamma^{k-t-1} r_k$

策略更新： $\theta \leftarrow \theta + \alpha \nabla_\theta \ln \pi(a_t|s_t, \theta) q_t(s_t, a_t)$

◇ 其次，该算法可以在**探索**（exploration）和**利用**（exploitation）之间取得平衡，以某种程度上归因于下式：

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

一方面， β_t 与 $q_t(s_t, a_t)$ 成正比（proportional）。结果是，如果 (s_t, a_t) 的动作价值很大，那么 $\pi(a_t|s_t, \theta_t)$ 会被增强，从而使得选择 a_t 的概率增加。因此，算法试图利用（exploit）具有更大价值的动作。

另一方面，当 $q_t(s_t, a_t) > 0$ 时， β_t 与 $\pi(a_t|s_t, \theta_t)$ 成反比（inversely proportional）。结果是，如果选择 a_t 的概率很小，那么 $\pi(a_t|s_t, \theta_t)$ 会被增强，从而使得选择 a_t 的概率增加。**因此，算法试图探索（explore）低概率的动作。**

此外，由于 (9.32) 使用样本来近似 (9.31) 中的真实梯度，理解应该如何获取样本是很重要的。

- 如何采样 S？真实梯度 $\mathbb{E}[\nabla_\theta \ln \pi(A|S, \theta_t) q_\pi(S, A)]$ 中的 S 应该服从分布 η ，即平稳分布 d_π 或折扣总概率分布 ρ_π （见 (9.19)）。 d_π 或 ρ_π 都代表了在 π 下表现出的长期行为。
- 如何采样 A？ $\mathbb{E}[\nabla_\theta \ln \pi(A|S, \theta_t) q_\pi(S, A)]$ 中的 A 应该服从分布 $\pi(A|S, \theta)$ 。**采样 A 的理想方式是根据 $\pi(a|s_t, \theta_t)$ 选择 a_t 。因此，策略梯度算法是on-policy。**

不幸的是，**采样 S 和 A 的理想方式在实践中并没有被严格遵循，因为它们的样本利用效率较低。**(9.32) 的一个样本效率更高的实现给在算法 9.1 中。**在这个实现中，首先根据 $\pi(\theta)$ 生成一个episode。然后，利用episode中的每一个经验样本多次更新 θ 。(PPO)**

9.5 本章小结

本章介绍了策略梯度方法（policy gradient method），它是许多现代强化学习算法的基础。策略梯度方法是**基于策略的**（policy-based）。这是本书的一大进步，因为前几章中的所有方法都是**基于价值的**（value-based）。策略梯度方法的基本思想很简单。**即选择一个合适的标量指标（scalar metric），然后通过梯度上升算法（gradient-ascent algorithm）对其进行优化。**

策略梯度方法最复杂的部分是指标梯度的推导。这是因为我们必须区分具有不同指标以及折扣/无折扣情况的各种场景。幸运的是，不同场景下的梯度表达式是相似的。因此，我们在定理 9.1 中总结了这些

表达式，这是本章最重要的理论结果。对于许多读者来说，了解这个定理就足够了。其证明是**非平凡的**（nontrivial），并不要求所有读者都去学习。

必须正确理解 (9.32) 中的策略梯度算法，因为它是许多高级策略梯度算法的基础。在下一章中，该算法将被扩展到另一种重要的策略梯度方法，称为 **演员-评论家**（actor-critic）。

9.6 问答（Q&A）

◇ **问：策略梯度方法的基本思想是什么？**

答：基本思想很简单。**即定义一个合适的标量指标，推导其梯度，然后使用梯度上升方法来优化该指标**。关于该方法最重要的理论结果是定理 9.1 中给出的策略梯度。

◇ **问：策略梯度方法最复杂的部分是什么？**

答：策略梯度方法的基本思想很简单。然而，梯度的推导过程相当复杂。这是因为我们必须区分许多不同的场景。每个场景中的数学推导过程都是非平凡的。对于许多读者来说，熟悉定理 9.1 中的结果而无需了解其证明就足够了。

◇ **问：策略梯度方法应该使用什么指标？**

答：我们在本章中介绍了三个常用的指标：$\bar{v}_\pi$，$\bar{v}_\pi^0$ 和 $\bar{r}_\pi$。由于它们都会导出相似的策略梯度，因此它们都可以被用于策略梯度方法中。更重要的是，(9.1) 和 (9.4) 中的表达式经常在文献中遇到。

◇ **问：为什么策略梯度中包含自然对数函数？**

答：引入自然对数函数是为了将梯度表示为期望值的形式。通过这种方式，我们可以用随机梯度来近似真实梯度。

◇ **问：为什么在推导策略梯度时我们需要研究无折扣情形？**

答：平均奖励 $\bar{r}_\pi$ 的定义对于折扣和无折扣情形都是有效的。虽然 $\bar{r}_\pi$ 在折扣情形下的梯度是一个近似值，但它在无折扣情形下的梯度更加优雅。

◇ **问：(9.32) 中的策略梯度算法在数学上做了什么？**

答：为了更好地理解该算法，建议读者考察其在 (9.33) 中的简洁表达，这清楚地表明它是用于更新 $\pi(a_t|s_t,\theta_t)$ 值的梯度上升算法。也就是说，当获得样本 (s_t,a_t) 时，可以更新策略，使得 $\pi(a_t|s_t,\theta_{t+1}) \geq \pi(a_t|s_t,\theta_t)$ 或 $\pi(a_t|s_t,\theta_{t+1}) < \pi(a_t|s_t,\theta_t)$，具体取决于系数。