

第 10 章演员-评论家方法 (Actor-Critic Methods)

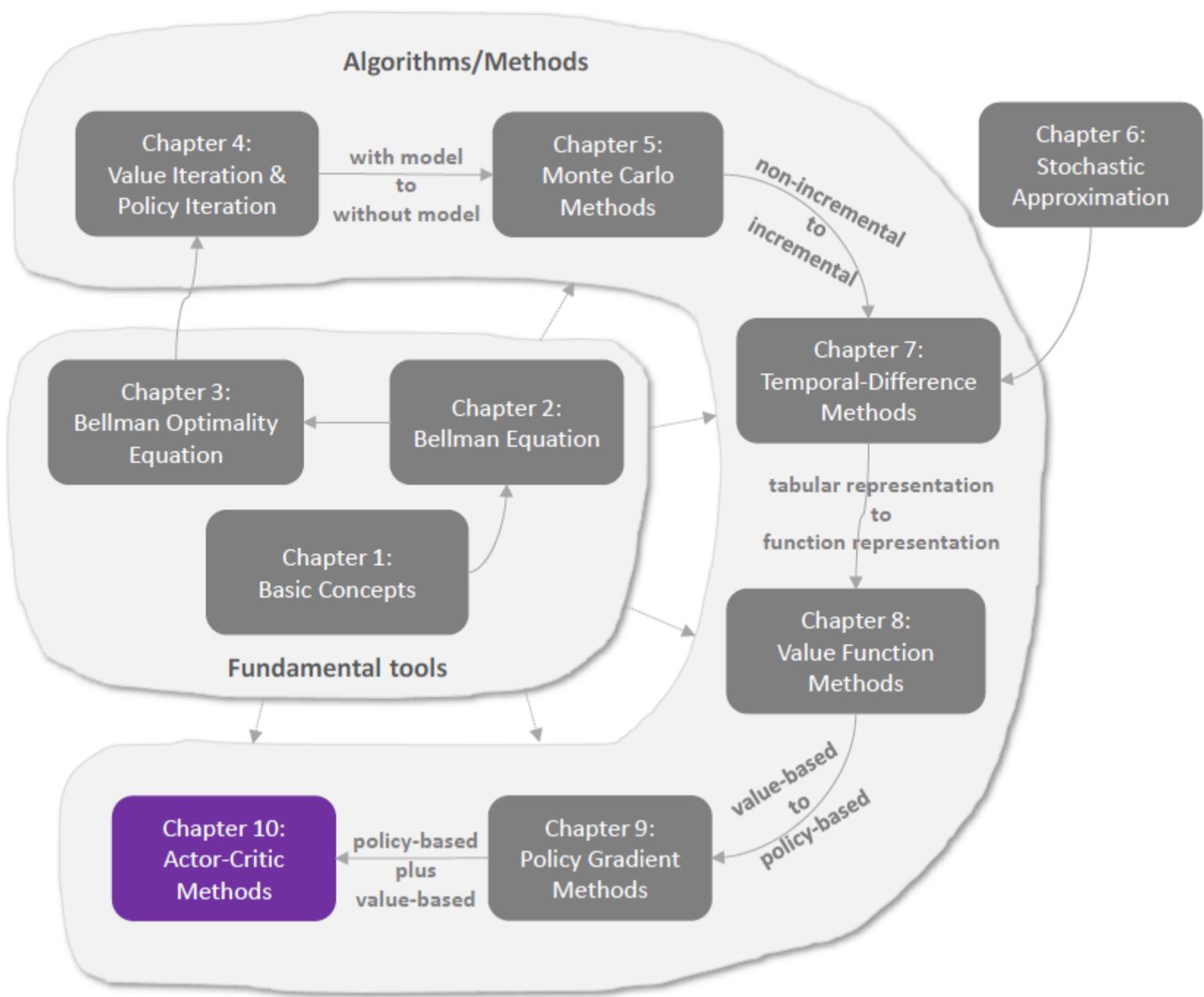


Figure 10.1: Where we are in this book.

图 10.1：我们在本书中的位置。

本章介绍演员-评论家方法 (actor-critic methods)。从一个角度来看，“演员-评论家”指的是一种结合了**基于策略** (policy-based) 和**基于价值** (value-based) 方法的结构。在这里，“演员” (actor) 指的是**策略更新步骤**。它被称为演员是因为动作是根据策略采取的。在这里，“评论家” (critic) 指的是**价值更新步骤**。它被称为评论家是因为它通过评估演员相应的价值来批评 (criticizes) 演员。从另一个角度来看，演员-评论家方法仍然是**策略梯度算法**。它们可以通过扩展第 9 章介绍的**策略梯度算法**而获得。在学习本章之前，读者充分理解第 8 章和第 9 章的内容是非常重要的。

10.1 最简单的演员-评论家算法 (QAC)

本节介绍最简单的演员-评论家算法。该算法可以通过扩展 (9.32) 中的**策略梯度算法**轻松获得。

回顾一下，**策略梯度方法**的核心思想是通过最大化标量指标 $J(\theta)$ 来搜索最优策略。用于最大化 $J(\theta)$ 的梯度上升算法为：

$$\begin{aligned}\theta_{t+1} &= \theta_t + \alpha \nabla_\theta J(\theta_t) \\ &= \theta_t + \alpha \mathbb{E}_{S \sim \eta, A \sim \pi} [\nabla_\theta \ln \pi(A|S, \theta_t) q_\pi(S, A)],\end{aligned}\quad (10.1)$$

其中 η 是状态的分布 (更多信息见定理 9.1)。由于真实梯度是未知的，我们可以使用随机梯度来近似它：

$$\theta_{t+1} = \theta_t + \alpha \nabla_\theta \ln \pi(a_t | s_t, \theta_t) q_t(s_t, a_t). \quad (10.2)$$

这就是(9.32)中给出的算法。

方程(10.2)非常重要，因为它清楚地展示了基于策略的方法和基于价值的方法是如何结合的。**一方面，它是一个基于策略的(policy-based)算法，因为它直接更新策略参数。另一方面，该方程需要知道** $q_t(s_t, a_t)$ **，这是动作价值** $q_\pi(s_t, a_t)$ **的估计值。**因此，需要另一个基于价值的(value-based)算法来生成 $q_t(s_t, a_t)$ 。到目前为止，本书已经研究了两种估计动作价值的方法。**第一种是基于蒙特卡洛学习，第二种是时序差分(TD)学习。**

- 如果 $q_t(s_t, a_t)$ 是通过蒙特卡洛学习估计的，对应的算法被称为**REINFORCE**或**蒙特卡洛策略梯度**(Monte Carlo policy gradient)，这在第9章中已经介绍过。
- 如果 $q_t(s_t, a_t)$ 是通过TD学习估计的，对应的算法通常被称为**演员-评论家**(actor-critic)。因此，**演员-评论家方法可以通过将基于TD的价值估计融入策略梯度方法中来获得。**

最简单的演员-评论家算法的流程总结在算法10.1中。**评论家**(critic)对应于通过(8.35)中提出的Sarsa算法进行的价值更新步骤。动作价值由参数化函数 $q(s, a, w)$ 表示。**演员**(actor)对应于(10.2)中的策略更新步骤。这个演员-评论家算法有时被称为**Q演员-评论家**(Q actor-critic，简称QAC)。尽管它很简单，但QAC揭示了演员-评论家方法的核心思想。它可以被扩展以生成许多高级算法，如本章后续部分所示。

算法10.1：最简单的演员-评论家算法(QAC)

初始化：策略函数 $\pi(a|s, \theta_0)$ ，其中 θ_0 是初始参数。价值函数 $q(s, a, w_0)$ ，其中 w_0 是初始参数。 $\alpha_w, \alpha_\theta > 0$ 。

目标：学习一个最大化 $J(\theta)$ 的最优策略。

在每个回合(episode)的时间步 t ，执行

根据 $\pi(a|s_t, \theta_t)$ 生成 a_t ，观测 r_{t+1}, s_{t+1} ，然后根据 $\pi(a|s_{t+1}, \theta_t)$ 生成 a_{t+1} 。

演员(策略更新)：

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \ln \pi(a_t|s_t, \theta_t) q(s_t, a_t, w_t)$$

评论家(价值更新)：

$$w_{t+1} = w_t + \alpha_w [r_{t+1} + \gamma q(s_{t+1}, a_{t+1}, w_t) - q(s_t, a_t, w_t)] \nabla_w q(s_t, a_t, w_t)$$

10.2 优势演员-评论家(A2C)

我们现在介绍**优势演员-评论家**(advantage actor-critic)算法。该算法的核心思想是引入一个基线(baseline)以减少估计方差。

10.2.1 基线不变性(Baseline invariance)

策略梯度的一个有趣性质是它对于额外的基线(baseline)具有不变性。即

$$\mathbb{E}_{S \sim \eta, A \sim \pi} [\nabla_\theta \ln \pi(A|S, \theta_t) q_\pi(S, A)] = \mathbb{E}_{S \sim \eta, A \sim \pi} [\nabla_\theta \ln \pi(A|S, \theta_t) (q_\pi(S, A) - b(S))], \quad (10.3)$$

其中额外的基线 $b(S)$ 是一个关于 S 的标量函数。我们接下来回答关于基线的两个问题。

◇ 首先，为什么(10.3)是有效的？

方程(10.3)成立当且仅当

$$\mathbb{E}_{S \sim \eta, A \sim \pi} [\nabla_\theta \ln \pi(A|S, \theta_t) b(S)] = 0.$$

该等式之所以成立，是因为

$$\begin{aligned} \mathbb{E}_{S \sim \eta, A \sim \pi} [\nabla_\theta \ln \pi(A|S, \theta_t) b(S)] &= \sum_{s \in \mathcal{S}} \eta(s) \sum_{a \in \mathcal{A}} \pi(a|s, \theta_t) \nabla_\theta \ln \pi(a|s, \theta_t) b(s) \\ &= \sum_{s \in \mathcal{S}} \eta(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(a|s, \theta_t) b(s) \\ &= \sum_{s \in \mathcal{S}} \eta(s) b(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(a|s, \theta_t) \\ &= \sum_{s \in \mathcal{S}} \eta(s) b(s) \nabla_\theta \sum_{a \in \mathcal{A}} \pi(a|s, \theta_t) \\ &= \sum_{s \in \mathcal{S}} \eta(s) b(s) \nabla_\theta 1 = 0. \end{aligned}$$

◇ 其次，为什么基线是有用的？

基线之所以有用，是因为当我们使用样本来近似真实梯度时，它可以减少近似方差。具体来说，令

$$X(S, A) \doteq \nabla_{\theta} \ln \pi(A|S, \theta_t)[q_{\pi}(S, A) - b(S)]. \quad (10.4)$$

那么，真实梯度为 $\mathbb{E}[X(S, A)]$ 。由于我们需要使用随机样本 x 来近似 $\mathbb{E}[X]$ ，如果方差 $\text{var}(X)$ 很小，那将是有利的。例如，如果 $\text{var}(X)$ 接近于零，那么任何样本 x 都可以准确地近似 $\mathbb{E}[X]$ 。相反，如果 $\text{var}(X)$ 很大，样本的值可能会远离 $\mathbb{E}[X]$ 。

虽然 $\mathbb{E}[X]$ 对基线具有不变性，但方差 $\text{var}(X)$ 并不具有不变性。我们的目标是设计一个好的基线以最小化 $\text{var}(X)$ 。在 REINFORCE 和 QAC 算法中，我们设置 $b = 0$ ，这并不能保证是一个好的基线。

事实上，最小化 $\text{var}(X)$ 的最优基线是

$$b^*(s) = \frac{\mathbb{E}_{A \sim \pi}[\|\nabla_{\theta} \ln \pi(A|s, \theta_t)\|^2 q_{\pi}(s, A)]}{\mathbb{E}_{A \sim \pi}[\|\nabla_{\theta} \ln \pi(A|s, \theta_t)\|^2]}, \quad s \in \mathcal{S}. \quad (10.5)$$

证明在方框 10.1 中给出。

虽然 (10.5) 中的基线是最优的，但它过于复杂，在实践中难以使用。如果从 (10.5) 中移除权重 $\|\nabla_{\theta} \ln \pi(A|s, \theta_t)\|^2$ ，我们可以得到一个具有简洁表达式的次优基线：

$$b^{\dagger}(s) = \mathbb{E}_{A \sim \pi}[q_{\pi}(s, A)] = v_{\pi}(s), \quad s \in \mathcal{S}.$$

有趣的是，这个次优基线就是状态价值。

■ 框 10.1：证明 (10.5) 中的 $b^*(s)$ 是最优基线

令 $\bar{x} \doteq \mathbb{E}[X]$ ，对于任何 $b(s)$ 而言，它都是不变的。如果 X 是一个向量，其方差是一个矩阵。通常选择 $\text{var}(X)$ 的迹 (trace) 作为优化的标量目标函数：

$$\begin{aligned} \text{tr}[\text{var}(X)] &= \text{tr}\mathbb{E}[(X - \bar{x})(X - \bar{x})^T] \\ &= \text{tr}\mathbb{E}[XX^T - \bar{x}X^T - X\bar{x}^T + \bar{x}\bar{x}^T] \\ &= \mathbb{E}[X^T X - X^T \bar{x} - \bar{x}^T X + \bar{x}^T \bar{x}] \\ &= \mathbb{E}[X^T X] - \bar{x}^T \bar{x}. \quad (10.6) \end{aligned}$$

- 矩阵没有天然的“大小”标量，所以需要选一个标量指标，迹就是各主方向方差的总和，衡量整体散布。
- 迹对旋转/正交变换不敏感：不依赖坐标系

在推导上述方程时，我们利用了迹的性质：对于任何维度适当的方阵 A, B ，都有 $\text{tr}(AB) = \text{tr}(BA)$ 。由于 \bar{x} 是不变的，方程 (10.6) 表明我们只需要最小化 $\mathbb{E}[X^T X]$ 。根据 (10.4) 中 X 的定义，我们有

$$\begin{aligned} \mathbb{E}[X^T X] &= \mathbb{E}[(\nabla_{\theta} \ln \pi)^T (\nabla_{\theta} \ln \pi)(q_{\pi}(S, A) - b(S))^2] \\ &= \mathbb{E}[\|\nabla_{\theta} \ln \pi\|^2 (q_{\pi}(S, A) - b(S))^2], \end{aligned}$$

其中 $\pi(A|S, \theta)$ 简写为 π 。由于 $S \sim \eta$ 且 $A \sim \pi$ ，上述方程可以重写为

$$\mathbb{E}[X^T X] = \sum_{s \in \mathcal{S}} \eta(s) \mathbb{E}_{A \sim \pi}[\|\nabla_{\theta} \ln \pi\|^2 (q_{\pi}(s, A) - b(s))^2].$$

为了确保 $\nabla_b \mathbb{E}[X^T X] = 0$ ，对于任意 $s \in \mathcal{S}$ ， $b(s)$ 应满足

$$\mathbb{E}_{A \sim \pi}[\|\nabla_{\theta} \ln \pi\|^2 (b(s) - q_{\pi}(s, A))] = 0, \quad s \in \mathcal{S}.$$

上述方程可以很容易地求解，从而得到最优基线 (optimal baseline)：

$$b^*(s) = \frac{\mathbb{E}_{A \sim \pi}[\|\nabla_{\theta} \ln \pi\|^2 q_{\pi}(s, A)]}{\mathbb{E}_{A \sim \pi}[\|\nabla_{\theta} \ln \pi\|^2]}, \quad s \in \mathcal{S}.$$

关于策略梯度方法中最优基线的更多讨论可以在文献 [69, 70] 中找到。

10.2.2 算法描述

当 $b(s) = v_{\pi}(s)$ 时，(10.1) 中的梯度上升算法变为：

$$\begin{aligned} \theta_{t+1} &= \theta_t + \alpha \mathbb{E}[\nabla_{\theta} \ln \pi(A|S, \theta_t)[q_{\pi}(S, A) - v_{\pi}(S)]] \\ &\doteq \theta_t + \alpha \mathbb{E}[\nabla_{\theta} \ln \pi(A|S, \theta_t) \delta_{\pi}(S, A)]. \quad (10.7) \end{aligned}$$

这里，

$$\delta_\pi(S, A) \doteq q_\pi(S, A) - v_\pi(S)$$

被称为**优势函数 (advantage function)**，它反映了某一动作相对于其他动作的优势。具体来说，注意到

$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s, a)$ 是动作价值的均值。如果 $\delta_\pi(s, a) > 0$ ，这意味着对应的动作具有比均值更大的价值。

(10.7) 的随机梯度版本为：

$$\begin{aligned} \theta_{t+1} &= \theta_t + \alpha \nabla_\theta \ln \pi(a_t | s_t, \theta_t) [q_t(s_t, a_t) - v_t(s_t)] \\ &= \theta_t + \alpha \nabla_\theta \ln \pi(a_t | s_t, \theta_t) \delta_t(s_t, a_t), \end{aligned} \quad (10.8)$$

其中 s_t, a_t 是时刻 t 时 S, A 的样本。这里， $q_t(s_t, a_t)$ 和 $v_t(s_t)$ 分别是 $q_{\pi(\theta_t)}(s_t, a_t)$ 和 $v_{\pi(\theta_t)}(s_t)$ 的近似值。**(10.8) 中的算法基于 q_t 相对于 v_t 的相对值而非 q_t 的绝对值来更新策略。** 这在直觉上是合理的，因为当我们试图在某个状态下选择一个动作时，我们只关心哪个动作相对于其他动作具有最大的价值。

- 如果 $q_t(s_t, a_t)$ 和 $v_t(s_t)$ 是通过蒙特卡洛 (Monte Carlo) 学习估计的，(10.8) 中的算法被称为**带基线的 REINFORCE (REINFORCE with a baseline)**。
- 如果 $q_t(s_t, a_t)$ 和 $v_t(s_t)$ 是通过 TD 学习估计的，该算法通常被称为**优势演员-评论家 (advantage actor-critic, A2C)**。A2C 的实现总结在算法 10.2 中。**需要注意的是，在此实现中，优势函数通过 TD 误差 (TD error) 来近似：**

$$q_t(s_t, a_t) - v_t(s_t) \approx r_{t+1} + \gamma v_t(s_{t+1}) - v_t(s_t).$$

这种近似是合理的，因为

$$q_\pi(s_t, a_t) - v_\pi(s_t) = \mathbb{E}[R_{t+1} + \gamma v_\pi(s_{t+1}) - v_\pi(s_t) | S_t = s_t, A_t = a_t],$$

这根据 $q_\pi(s_t, a_t)$ 的定义是成立的。**使用 TD 误差的一个优点是我们只需要使用单个神经网络来表示 $v_\pi(s)$ 。** 否则，如果 $\delta_t = q_t(s_t, a_t) - v_t(s_t)$ ，我们需要维护两个网络来分别表示 $v_\pi(s)$ 和 $q_\pi(s, a)$ 。当使用 TD 误差时，该算法也可以被称为**TD 演员-评论家 (TD actor-critic)**。此外，值得注意的是，策略 $\pi(\theta_t)$ 是随机的，因此具有探索性。所以，它可以直接用于生成经验样本，而不依赖于诸如 ϵ -贪心 (ϵ -greedy) 之类的技术。A2C 有一些变体，例如异步优势演员-评论家 (*asynchronous advantage actor-critic, A3C*)。感兴趣的读者可以查阅 [71, 72]。



算法 10.2：优势演员-评论家 (A2C) 或 TD 演员-评论家

初始化：策略函数 $\pi(a|s, \theta_0)$ ，其中 θ_0 为初始参数。价值函数 $v(s, w_0)$ ，其中 w_0 为初始参数。 $\alpha_w, \alpha_\theta > 0$ 。

目标：学习一个最优策略以最大化 $J(\theta)$ 。

在每一回合 (episode) 的时间步 t ，执行：

根据 $\pi(a|s_t, \theta_t)$ 生成 a_t ，然后观测 r_{t+1}, s_{t+1} 。

优势 (TD 误差)：

$$\delta_t = r_{t+1} + \gamma v(s_{t+1}, w_t) - v(s_t, w_t)$$

演员 (策略更新)：

$$\theta_{t+1} = \theta_t + \alpha_\theta \delta_t \nabla_\theta \ln \pi(a_t | s_t, \theta_t)$$

评论家 (价值更新)：

$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w v(s_t, w_t)$$

10.3 Off-policy演员-评论家 (Off-policy actor-critic)

我们目前学习到的策略梯度方法，包括 REINFORCE、QAC 和 A2C，都是on-policy的。其原因可以从真实梯度的表达式中看出：

$$\nabla_\theta J(\theta) = \mathbb{E}_{S \sim \eta, A \sim \pi} [\nabla_\theta \ln \pi(A|S, \theta_t) (q_\pi(S, A) - v_\pi(S))].$$

为了使用样本来近似这个真实梯度，我们必须通过遵循 $\pi(\theta)$ 来生成动作样本。因此， $\pi(\theta)$ 是行为策略 (behavior policy)。由于 $\pi(\theta)$ 也是我们旨在改进的目标策略 (target policy)，所以这些策略梯度方法是On-policy的。

在我们就已经拥有一些由给定的行为策略生成的样本的情况下，策略梯度方法仍然可以被应用以利用这些样本。为此，我们可以采用一种称为**重要性采样 (importance sampling)** 的技术。值得一提的是，重要性采样技术并不局限于强化学习领域。它是一种使

用从另一个分布中抽取的样本来估计定义在一个概率分布上的期望值的通用技术。

10.3.1 重要性采样 (Importance sampling)

接下来我们介绍重要性采样技术。考虑一个随机变量 $X \in \mathcal{X}$ 。假设 $p_0(X)$ 是一个概率分布。我们的目标是估计 $\mathbb{E}_{X \sim p_0}[X]$ 。假设我们有一些独立同分布 (i.i.d.) 的样本 $\{x_i\}_{i=1}^n$ 。

- 首先，如果样本 $\{x_i\}_{i=1}^n$ 是通过遵循 p_0 生成的，那么平均值 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 可以用来近似 $\mathbb{E}_{X \sim p_0}[X]$ ，因为 \bar{x} 是 $\mathbb{E}_{X \sim p_0}[X]$ 的无偏估计，且随着 $n \rightarrow \infty$ ，估计方差收敛于零（更多信息请参见框 5.1 中的大数定律）。
- 其次，考虑一个新的场景，**其中样本 $\{x_i\}_{i=1}^n$ 不是由 p_0 生成的**。相反，它们是由另一个分布 p_1 生成的。我们还能使用这些样本来近似 $\mathbb{E}_{X \sim p_0}[X]$ 吗？答案是肯定的。然而，我们不能再使用 $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ 来近似 $\mathbb{E}_{X \sim p_0}[X]$ ，因为 $\bar{x} \approx \mathbb{E}_{X \sim p_1}[X]$ 而不是 $\mathbb{E}_{X \sim p_0}[X]$ 。

在第二种场景中， $\mathbb{E}_{X \sim p_0}[X]$ 可以基于**重要性采样 (importance sampling)** 技术进行近似。特别是， $\mathbb{E}_{X \sim p_0}[X]$ 满足

$$\mathbb{E}_{X \sim p_0}[X] = \sum_{x \in \mathcal{X}} p_0(x)x = \sum_{x \in \mathcal{X}} p_1(x) \underbrace{\frac{p_0(x)}{p_1(x)}}_{f(x)} x = \mathbb{E}_{X \sim p_1}[f(X)]. \quad (10.9)$$

因此，估计 $\mathbb{E}_{X \sim p_0}[X]$ 就变成了估计 $\mathbb{E}_{X \sim p_1}[f(X)]$ 的问题。令

$$\bar{f} \doteq \frac{1}{n} \sum_{i=1}^n f(x_i).$$

由于 \bar{f} 可以有效地近似 $\mathbb{E}_{X \sim p_1}[f(X)]$ ，由 (10.9) 可得

$$\mathbb{E}_{X \sim p_0}[X] = \mathbb{E}_{X \sim p_1}[f(X)] \approx \bar{f} = \frac{1}{n} \sum_{i=1}^n f(x_i) = \frac{1}{n} \sum_{i=1}^n \underbrace{\frac{p_0(x_i)}{p_1(x_i)}}_{\text{importance weight}} x_i. \quad (10.10)$$

方程 (10.10) 表明 $\mathbb{E}_{X \sim p_0}[X]$ 可以通过 x_i 的加权平均来近似。在这里， $\frac{p_0(x_i)}{p_1(x_i)}$ 被称为**重要性权重 (importance weight)**。当 $p_1 = p_0$ 时，重要性权重为 1， \bar{f} 变为 \bar{x} 。当 $p_0(x_i) \geq p_1(x_i)$ 时， x_i 被 p_0 采样的频率较高，而被 p_1 采样的频率较低。在这种情况下，重要性权重大于 1，强调了这个样本的重要性。

一些读者可能会问这样一个问题：既然 (10.10) 中需要 $p_0(x)$ ，为什么我们不利用定义 $\mathbb{E}_{X \sim p_0}[X] = \sum_{x \in \mathcal{X}} p_0(x)x$ 直接计算 $\mathbb{E}_{X \sim p_0}[X]$ 呢？答案如下。要使用定义，我们需要知道 p_0 的解析表达式，或者知道每一个 $x \in \mathcal{X}$ 的 $p_0(x)$ 值。然而，当分布由例如神经网络表示时，**很难获得 p_0 的解析表达式**。当 \mathcal{X} 很大时，也很难获得每一个 $x \in \mathcal{X}$ 的 $p_0(x)$ 值。相比之下，(10.10) 仅需要一些样本的 $p_0(x_i)$ 值，在实践中更容易实现。

 比如在LLM强化学习时，使用的 x_i 也不是从 p_0 和 p_1 采样的，而是根据事先提供的数据集进行query收集，输出到新旧两个模型，获取响应，用样本生成的概率（或概率密度）进行计算

一个说明性示例

接下来我们展示一个例子来演示重要性采样技术。考虑 $X \in \mathcal{X} \doteq \{+1, -1\}$ 。假设 p_0 是一个满足以下条件的概率分布：

$$p_0(X = +1) = 0.5, \quad p_0(X = -1) = 0.5.$$

X 在 p_0 下的期望为

$$\mathbb{E}_{X \sim p_0}[X] = (+1) \cdot 0.5 + (-1) \cdot 0.5 = 0.$$

假设 p_1 是另一个满足以下条件的分布：

$$p_1(X = +1) = 0.8, \quad p_1(X = -1) = 0.2.$$

X 在 p_1 下的期望为

$$\mathbb{E}_{X \sim p_1}[X] = (+1) \cdot 0.8 + (-1) \cdot 0.2 = 0.6.$$

假设我们有一些从 p_1 中抽取的样本 $\{x_i\}$ 。我们的目标是利用这些样本估计 $\mathbb{E}_{X \sim p_0}[X]$ 。如图 10.2 所示， $+1$ 的样本比 -1 多。这是因为 $p_1(X = +1) = 0.8 > p_1(X = -1) = 0.2$ 。如果我们直接计算样本的平均值 $\sum_{i=1}^n x_i/n$ ，该值会收敛到 $\mathbb{E}_{X \sim p_1}[X] = 0.6$ （见图 10.2 中的虚线）。相比之下，如果我们按照 (10.10) 计算加权平均值，该值可以成功收敛到 $\mathbb{E}_{X \sim p_0}[X] = 0$ （见图 10.2 中的实线）。

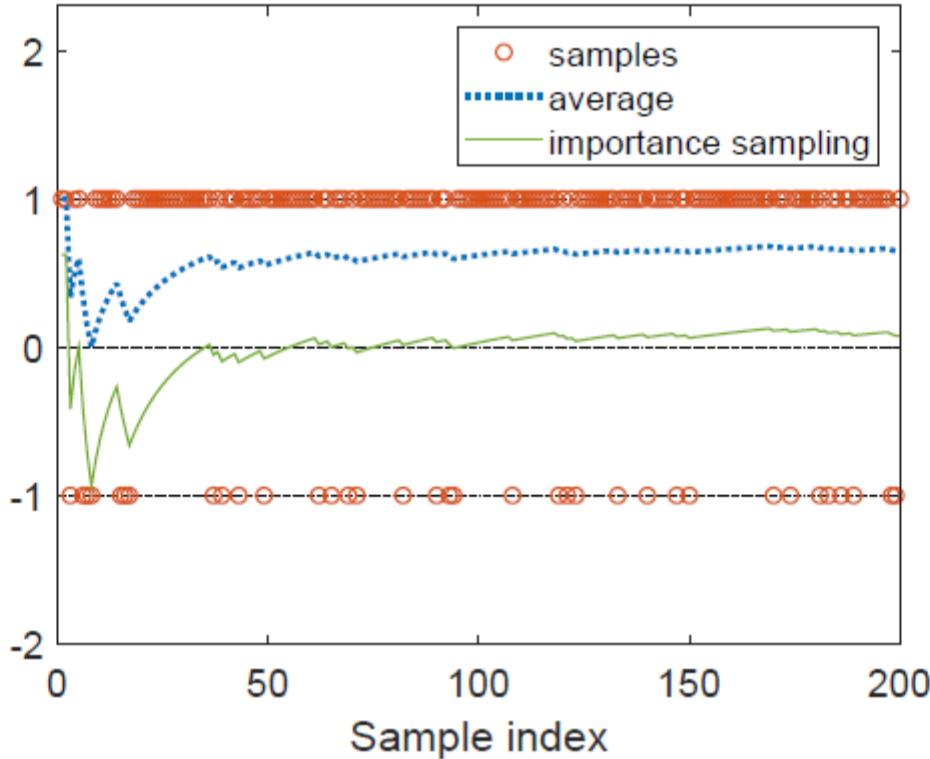


Figure 10.2: An example for demonstrating the importance sampling technique. Here, $X \in \{+1, -1\}$ and $p_0(X = +1) = p_0(X = -1) = 0.5$. The samples are generated according to p_1 where $p_1(X = +1) = 0.8$ and $p_1(X = -1) = 0.2$. The average of the samples converges to $\mathbb{E}_{X \sim p_1}[X] = 0.6$, but the weighted average calculated by the importance sampling technique in (10.10) converges to $\mathbb{E}_{X \sim p_0}[X] = 0$.

图 10.2：演示重要性采样技术的示例。这里， $X \in \{+1, -1\}$ 且 $p_0(X = +1) = p_0(X = -1) = 0.5$ 。样本是根据 p_1 生成的，其中 $p_1(X = +1) = 0.8$ 且 $p_1(X = -1) = 0.2$ 。样本的平均值收敛到 $\mathbb{E}_{X \sim p_1}[X] = 0.6$ ，但是通过 (10.10) 中的重要性采样技术计算的加权平均值收敛到 $\mathbb{E}_{X \sim p_0}[X] = 0$ 。

最后，用于生成样本的分布 p_1 必须满足：当 $p_0(x) \neq 0$ 时， $p_1(x) \neq 0$ 。如果 $p_1(x) = 0$ 而 $p_0(x) \neq 0$ ，估计结果可能会有问题。例如，如果

$$p_1(X = +1) = 1, \quad p_1(X = -1) = 0,$$

那么由 p_1 生成的样本全为正数： $\{x_i\} = \{+1, +1, \dots, +1\}$ 。这些样本无法正确估计 $\mathbb{E}_{X \sim p_0}[X] = 0$ ，因为

$$\frac{1}{n} \sum_{i=1}^n \frac{p_0(x_i)}{p_1(x_i)} x_i = \frac{1}{n} \sum_{i=1}^n \frac{p_0(+1)}{p_1(+1)} 1 = \frac{1}{n} \sum_{i=1}^n \frac{0.5}{1} 1 \equiv 0.5,$$

无论 n 有多大。

10.3.2 off-policy 策略梯度定理 (The off-policy policy gradient theorem)

有了重要性采样技术，我们就可以介绍 off-policy 策略梯度定理了。假设 β 是一个行为策略 (behavior policy)。我们的目标是利用由 β 生成的样本来学习一个目标策略 (target policy) π ，使其能够最大化以下指标：

$$J(\theta) = \sum_{s \in \mathcal{S}} d_\beta(s) v_\pi(s) = \mathbb{E}_{S \sim d_\beta}[v_\pi(S)],$$

其中 d_β 是策略 β 下的平稳分布， v_π 是策略 π 下的状态价值。该指标的梯度由以下定理给出。

定理 10.1 (off-policy 策略梯度定理)。在折扣因子 $\gamma \in (0, 1)$ 的折扣情况下， $J(\theta)$ 的梯度为

$$\nabla_\theta J(\theta) = \mathbb{E}_{S \sim \rho, A \sim \beta} \left[\underbrace{\frac{\pi(A|S, \theta)}{\beta(A|S)}}_{\text{importance weight}} \nabla_\theta \ln \pi(A|S, \theta) q_\pi(S, A) \right], \quad (10.11)$$

其中状态分布 ρ 定义为

$$\rho(s) \doteq \sum_{s' \in \mathcal{S}} d_\beta(s') \Pr_\pi(s|s'), \quad s \in \mathcal{S},$$

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

(10.11) 中的梯度与定理 9.1 中 on-policy 情况下的梯度相似，但有两点不同。第一点不同是重要性权重。第二点不同是 $A \sim \beta$ 而不是 $A \sim \pi$ 。因此，我们可以通过遵循 β 生成的样本来近似真实梯度。该定理的证明在框 10.2 中给出。

■ 框 10.2：定理 10.1 的证明

由于 d_β 独立于 θ ， $J(\theta)$ 的梯度满足

$$\nabla_\theta J(\theta) = \nabla_\theta \sum_{s \in \mathcal{S}} d_\beta(s) v_\pi(s) = \sum_{s \in \mathcal{S}} d_\beta(s) \nabla_\theta v_\pi(s). \quad (10.12)$$

根据引理 9.2， $\nabla_\theta v_\pi(s)$ 的表达式为

$$\nabla_\theta v_\pi(s) = \sum_{s' \in \mathcal{S}} \Pr_\pi(s'|s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(a|s', \theta) q_\pi(s', a), \quad (10.13)$$

其中 $\Pr_\pi(s'|s) \doteq \sum_{k=0}^{\infty} \gamma^k [P_\pi^k]_{ss'} = [(I - \gamma P_\pi)^{-1}]_{ss'}$ 。将 (10.13) 代入 (10.12) 得到

$$\begin{aligned} \nabla_\theta J(\theta) &= \sum_{s \in \mathcal{S}} d_\beta(s) \nabla_\theta v_\pi(s) = \sum_{s \in \mathcal{S}} d_\beta(s) \sum_{s' \in \mathcal{S}} \Pr_\pi(s'|s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(a|s', \theta) q_\pi(s', a) \\ &= \sum_{s' \in \mathcal{S}} \left(\sum_{s \in \mathcal{S}} d_\beta(s) \Pr_\pi(s'|s) \right) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(a|s', \theta) q_\pi(s', a) \\ &\doteq \sum_{s' \in \mathcal{S}} \rho(s') \sum_{a \in \mathcal{A}} \nabla_\theta \pi(a|s', \theta) q_\pi(s', a) \\ &= \sum_{s \in \mathcal{S}} \rho(s) \sum_{a \in \mathcal{A}} \nabla_\theta \pi(a|s, \theta) q_\pi(s, a) \quad (\text{将 } s' \text{ 换为 } s) \\ &= \mathbb{E}_{S \sim \rho} \left[\sum_{a \in \mathcal{A}} \nabla_\theta \pi(a|S, \theta) q_\pi(S, a) \right]. \end{aligned}$$

利用重要性采样技术，上述方程可进一步重写为

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

证明完成。上述证明与定理 9.1 的证明类似。

10.3.3 算法描述

基于 off-policy 策略梯度定理，我们准备介绍异策略演员-评论家 (off-policy actor-critic) 算法。由于 off-policy 情况与 on-policy 情况非常相似，我们仅展示一些关键步骤。

首先，off-policy 策略梯度对于任何额外的基线 $b(s)$ 都是不变的。特别地，我们有

$$\nabla_\theta J(\theta) = \mathbb{E}_{S \sim \rho, A \sim \beta} \left[\frac{\pi(A|S, \theta)}{\beta(A|S)} \nabla_\theta \ln \pi(A|S, \theta) (q_\pi(S, A) - b(S)) \right],$$

因为 $\mathbb{E} \left[\frac{\pi(A|S, \theta)}{\beta(A|S)} \nabla_\theta \ln \pi(A|S, \theta) b(S) \right] = 0$ 。为了减小估计方差，我们可以选择基线为 $b(S) = v_\pi(S)$ 并得到

$$\nabla_{\theta} J(\theta) = \mathbb{E} \left[\frac{\pi(A|S, \theta)}{\beta(A|S)} \nabla_{\theta} \ln \pi(A|S, \theta) (q_{\pi}(S, A) - v_{\pi}(S)) \right].$$

相应的随机梯度上升算法为

$$\theta_{t+1} = \theta_t + \alpha_{\theta} \frac{\pi(a_t|s_t, \theta_t)}{\beta(a_t|s_t)} \nabla_{\theta} \ln \pi(a_t|s_t, \theta_t) (q_t(s_t, a_t) - v_t(s_t)),$$

其中 $\alpha_{\theta} > 0$ 。与on-policy情况类似，优势函数 $q_t(s, a) - v_t(s)$ 可以被 TD 误差替换。即

$$q_t(s_t, a_t) - v_t(s_t) \approx r_{t+1} + \gamma v_t(s_{t+1}) - v_t(s_t) \doteq \delta_t(s_t, a_t).$$

那么，算法变为

$$\theta_{t+1} = \theta_t + \alpha_{\theta} \frac{\pi(a_t|s_t, \theta_t)}{\beta(a_t|s_t)} \nabla_{\theta} \ln \pi(a_t|s_t, \theta_t) \delta_t(s_t, a_t).$$

off-policy演员-评论家算法的实现总结在算法 10.3 中。可以看出，该算法与优势演员-评论家（advantage actor-critic）算法相同，除了在评论家（critic）和演员（actor）中都包含了一个额外的重要性权重。必须注意的是，除了演员之外，评论家也通过重要性采样技术从on-policy转换为off-policy。事实上，重要性采样是一种通用技术，既可以应用于基于策略的算法，也可以应用于基于价值的算法。最后，算法 10.3 可以通过多种方式扩展以结合更多技术，例如资格迹（eligibility traces）[73]。

■ 算法 10.3：基于重要性采样的off-policy演员-评论家算法

初始化：给定的行为策略 $\beta(a|s)$ 。目标策略 $\pi(a|s, \theta_0)$ ，其中 θ_0 是初始参数。价值函数 $v(s, w_0)$ ，其中 w_0 是初始参数。 $\alpha_w, \alpha_{\theta} > 0$ 。

目标：学习一个最优策略以最大化 $J(\theta)$ 。

在每个回合（episode）的时间步 t ，执行：

遵循 $\beta(s_t)$ 生成 a_t ，并观察 r_{t+1}, s_{t+1} 。

优势（TD 误差）：

$$\delta_t = r_{t+1} + \gamma v(s_{t+1}, w_t) - v(s_t, w_t)$$

演员（策略更新）：

$$\theta_{t+1} = \theta_t + \alpha_{\theta} \frac{\pi(a_t|s_t, \theta_t)}{\beta(a_t|s_t)} \delta_t \nabla_{\theta} \ln \pi(a_t|s_t, \theta_t)$$

评论家（价值更新）：

$$w_{t+1} = w_t + \alpha_w \frac{\pi(a_t|s_t, \theta_t)}{\beta(a_t|s_t)} \delta_t \nabla_w v(s_t, w_t)$$

10.4 确定性演员-评论家 (Deterministic actor-critic)

到目前为止，策略梯度方法中使用的策略都是随机的（stochastic），因为要求对于每一个 (s, a) 都有 $\pi(a|s, \theta) > 0$ 。本节表明确定性（deterministic）策略也可以用于策略梯度方法。这里，“确定性”意味着，对于任何状态，给出一个概率为 1 的单一动作，而所有其他动作的概率为 0。

 研究确定性情况很重要，因为它自然是异策略的（off-policy），并且可以有效地处理连续动作空间。

输入是 S 输出是 a （对应于动作本身）

我们一直使用 $\pi(a|s, \theta)$ 来表示一般策略，它可以是随机的也可以是确定性的。在本节中，我们使用

$$a = \mu(s, \theta)$$

来专门表示确定性策略。与给出动作概率的 π 不同， μ 直接给出动作，因为它是从 \mathcal{S} 到 \mathcal{A} 的映射。这个确定性策略可以由（例如）一个神经网络表示，其中 s 作为输入， a 作为输出， θ 作为其参数。为了简单起见，我们通常将 $\mu(s, \theta)$ 简写为 $\mu(s)$ 。

10.4.1 确定性策略梯度定理 (The deterministic policy gradient theorem)

上一章介绍的策略梯度定理仅对随机策略有效。当我们要求策略是确定性的时候，必须推导出一个新的策略梯度定理。

定理 10.2 (确定性策略梯度定理)。 $J(\theta)$ 的梯度为

$$\begin{aligned}\nabla_{\theta} J(\theta) &= \sum_{s \in \mathcal{S}} \eta(s) \nabla_{\theta} \mu(s) (\nabla_a q_{\mu}(s, a))|_{a=\mu(s)} \\ &= \mathbb{E}_{S \sim \eta} [\nabla_{\theta} \mu(S) (\nabla_a q_{\mu}(S, a))|_{a=\mu(S)}],\end{aligned}\quad (10.14)$$

其中 η 是状态的分布。

定理 10.2 是定理 10.3 和定理 10.4 中提出的结果的总结，因为这两个定理中的梯度具有相似的表达式。 $J(\theta)$ 和 η 的具体表达式可以在定理 10.3 和 10.4 中找到。

与随机情况不同，(10.14) 中显示的确定性情况下的梯度不涉及动作随机变量 A 。结果是，**当我们使用样本来近似真实梯度时，不需要对动作进行采样**。因此，**确定性策略梯度方法是off-policy的**。此外，一些读者可能会想知道为什么 $(\nabla_a q_{\mu}(s, a))|_{a=\mu(s)}$ 不能写成 $\nabla_a q_{\mu}(S, \mu(S))$ ，这看起来更简洁。这仅仅是因为，如果我们这样做，就不清楚 $q_{\mu}(S, \mu(S))$ 是如何作为 a 的函数的。一个简洁且较少引起混淆的表达可能是 $\nabla_a q_{\mu}(S, a = \mu(S))$ 。

在本小节的其余部分，我们将介绍定理 10.2 的推导细节。

特别是，我们推导了两个常见指标的梯度：**第一个是平均价值 (average value)**，**第二个是平均奖励 (average reward)**。由于这两个指标在 9.2 节中已经详细讨论过，我们有时会直接使用它们的性质而不加证明。对于大多数读者来说，熟悉定理 10.2 而不需要了解其推导细节就足够了。感兴趣的读者可以选择性地阅读本节剩余部分的细节。

指标 1：平均价值 (Average value)

我们首先推导平均价值的梯度：

$$J(\theta) = \mathbb{E}[v_{\mu}(s)] = \sum_{s \in \mathcal{S}} d_0(s) v_{\mu}(s), \quad (10.15)$$

其中 d_0 是状态的概率分布。在这里，为了简单起见，选择 d_0 为独立于 μ 。选择 d_0 有两种特殊但重要的情况。第一种情况是 $d_0(s_0) = 1$ 且 $d_0(s \neq s_0) = 0$ ，其中 s_0 是感兴趣的特定状态。在这种情况下，策略旨在最大化从 s_0 开始时可以获得的折扣回报。第二种情况是 d_0 是给定的行为策略（与目标策略不同）的分布。

为了计算 $J(\theta)$ 的梯度，我们需要首先计算任意 $s \in \mathcal{S}$ 的 $v_{\mu}(s)$ 的梯度。考虑 $\gamma \in (0, 1)$ 的折扣情况。

引理 10.1 ($v_{\mu}(s)$ 的梯度)。在折扣情况下，对于任意 $s \in \mathcal{S}$ ，以下等式成立：

$$\nabla_{\theta} v_{\mu}(s) = \sum_{s' \in \mathcal{S}} \Pr_{\mu}(s'|s) \nabla_{\theta} \mu(s') (\nabla_a q_{\mu}(s', a))|_{a=\mu(s')}, \quad (10.16)$$

其中

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

是在策略 μ 下从 s 转移到 s' 的折扣总概率。这里， $[\cdot]_{ss'}$ 表示矩阵第 s 行和第 s' 列的元素。

框 10.3：引理 10.1 的证明

由于策略是确定性的，我们有(不需要再取动作概率)

$$v_{\mu}(s) = q_{\mu}(s, \mu(s)).$$

由于 q_{μ} 和 μ 都是 θ 的函数，我们有

$$\nabla_{\theta} v_{\mu}(s) = \nabla_{\theta} q_{\mu}(s, \mu(s)) = (\nabla_{\theta} q_{\mu}(s, a))|_{a=\mu(s)} + \nabla_{\theta} \mu(s) (\nabla_a q_{\mu}(s, a))|_{a=\mu(s)}. \quad (10.17)$$

根据动作价值的定义，对于任意给定的 (s, a) ，我们有

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

其中 $r(s, a) = \sum_r r p(r|s, a)$ 。由于 $r(s, a)$ 独立于 μ ，我们有

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

将上述方程代入 (10.17) 可得

$$\nabla_\theta v_\mu(s) = \gamma \sum_{s' \in \mathcal{S}} p(s'|s, \mu(s)) \nabla_\theta v_\mu(s') + \underbrace{\nabla_\theta \mu(s)(\nabla_a q_\mu(s, a))|_{a=\mu(s)}}_{u(s)}, \quad s \in \mathcal{S}.$$

由于上述方程对所有 $s \in \mathcal{S}$ 都成立，我们可以结合这些方程得到矩阵-向量形式：

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

其中 $n = |\mathcal{S}|$, m 是 θ 的维度, P_μ 是状态转移矩阵, 且 $[P_\mu]_{ss'} = p(s'|s, \mu(s))$, \otimes 是克罗内克积 (Kronecker product)。上述矩阵-向量形式可以简洁地写为

$$\nabla_\theta v_\mu = u + \gamma(P_\mu \otimes I_m) \nabla_\theta v_\mu,$$

这是关于 $\nabla_\theta v_\mu$ 的一个线性方程。然后, $\nabla_\theta v_\mu$ 可以求解为

$$\begin{aligned} \nabla_\theta v_\mu &= (I_{mn} - \gamma P_\mu \otimes I_m)^{-1} u \\ &= (I_n \otimes I_m - \gamma P_\mu \otimes I_m)^{-1} u \\ &= [(I_n - \gamma P_\mu)^{-1} \otimes I_m] u. \end{aligned} \quad (10.18)$$

(10.18) 的逐元素形式为

$$\begin{aligned} \nabla_\theta v_\mu(s) &= \sum_{s' \in \mathcal{S}} [(I - \gamma P_\mu)^{-1}]_{ss'} u(s') \\ &= \sum_{s' \in \mathcal{S}} [(I - \gamma P_\mu)^{-1}]_{ss'} [\nabla_\theta \mu(s')(\nabla_a q_\mu(s', a))|_{a=\mu(s')}] . \end{aligned} \quad (10.19)$$

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

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

注意 $[P_\mu^k]_{ss'}$ 是恰好使用 k 步从 s 转移到 s' 的概率 (更多信息见框 8.1)。因此, $[(I - \gamma P_\mu)^{-1}]_{ss'}$ 是使用任意步数从 s 转移到 s' 的折扣总概率。通过记 $[(I - \gamma P_\mu)^{-1}]_{ss'} \doteq \Pr_\mu(s'|s)$, 方程 (10.19) 导出 (10.16)。

做好了引理 10.1 的准备工作后, 我们要准备推导 $J(\theta)$ 的梯度了。

定理 10.3 (折扣情况下的确定性策略梯度定理)。在折扣因子 $\gamma \in (0, 1)$ 的折扣情况下, (10.15) 中 $J(\theta)$ 的梯度为

$$\begin{aligned} \nabla_\theta J(\theta) &= \sum_{s \in \mathcal{S}} \rho_\mu(s) \nabla_\theta \mu(s)(\nabla_a q_\mu(s, a))|_{a=\mu(s)} \\ &= \mathbb{E}_{S \sim \rho_\mu} [\nabla_\theta \mu(S)(\nabla_a q_\mu(S, a))|_{a=\mu(S)}], \end{aligned}$$

其中状态分布 ρ_μ 为

$$\rho_\mu(s) = \sum_{s' \in \mathcal{S}} d_0(s') \Pr_\mu(s|s'), \quad s \in \mathcal{S}.$$

这里, $\Pr_\mu(s|s') = \sum_{k=0}^{\infty} \gamma^k [P_\mu^k]_{s's} = [(I - \gamma P_\mu)^{-1}]_{s's}$ 是在策略 μ 下从 s' 转移到 s 的折扣总概率。

■ 框 10.4：定理 10.3 的证明

由于 d_0 独立于 μ , 我们有

$$\nabla_\theta J(\theta) = \sum_{s \in \mathcal{S}} d_0(s) \nabla_\theta v_\mu(s).$$

将引理 10.1 给出的 $\nabla_\theta v_\mu(s)$ 的表达式代入上述方程, 得

$$\begin{aligned}
\nabla_{\theta} J(\theta) &= \sum_{s \in \mathcal{S}} d_0(s) \nabla_{\theta} v_{\mu}(s) \\
&= \sum_{s \in \mathcal{S}} d_0(s) \sum_{s' \in \mathcal{S}} \Pr_{\mu}(s'|s) \nabla_{\theta} \mu(s') (\nabla_a q_{\mu}(s', a))|_{a=\mu(s')} \\
&= \sum_{s' \in \mathcal{S}} \left(\sum_{s \in \mathcal{S}} d_0(s) \Pr_{\mu}(s'|s) \right) \nabla_{\theta} \mu(s') (\nabla_a q_{\mu}(s', a))|_{a=\mu(s')} \\
&\doteq \sum_{s' \in \mathcal{S}} \rho_{\mu}(s') \nabla_{\theta} \mu(s') (\nabla_a q_{\mu}(s', a))|_{a=\mu(s')} \\
&= \sum_{s \in \mathcal{S}} \rho_{\mu}(s) \nabla_{\theta} \mu(s) (\nabla_a q_{\mu}(s, a))|_{a=\mu(s)} \quad (\text{将 } s' \text{ 换为 } s) \\
&= \mathbb{E}_{S \sim \rho_{\mu}} [\nabla_{\theta} \mu(S) (\nabla_a q_{\mu}(S, a))|_{a=\mu(S)}].
\end{aligned}$$

证明完成。上述证明与文献 [74] 中定理 1 的证明一致。这里，我们考虑的是状态和动作有限的情况。当它们是连续的时，证明是类似的，但求和应替换为积分 [74]。

指标 2：平均奖励 (Average reward)

接下来我们推导平均奖励的梯度：

$$\begin{aligned}
J(\theta) = \bar{r}_{\mu} &= \sum_{s \in \mathcal{S}} d_{\mu}(s) r_{\mu}(s) \\
&= \mathbb{E}_{S \sim d_{\mu}} [r_{\mu}(S)],
\end{aligned} \tag{10.20}$$

其中

$$r_{\mu}(s) = \mathbb{E}[R|s, a = \mu(s)] = \sum_r r p(r|s, a = \mu(s))$$

是即时奖励的期望。关于该指标的更多信息可以在 9.2 节中找到。

$J(\theta)$ 的梯度在以下定理中给出。

定理 10.4 (非折扣情况下的确定性策略梯度定理)。在非折扣情况下，(10.20) 中 $J(\theta)$ 的梯度为

$$\begin{aligned}
\nabla_{\theta} J(\theta) &= \sum_{s \in \mathcal{S}} d_{\mu}(s) \nabla_{\theta} \mu(s) (\nabla_a q_{\mu}(s, a))|_{a=\mu(s)} \\
&= \mathbb{E}_{S \sim d_{\mu}} [\nabla_{\theta} \mu(S) (\nabla_a q_{\mu}(S, a))|_{a=\mu(S)}],
\end{aligned}$$

其中 d_{μ} 是策略 μ 下状态的平稳分布。

■ 框 10.5：定理 10.4 的证明

由于策略是确定性的，我们有

$$v_{\mu}(s) = q_{\mu}(s, \mu(s)).$$

由于 q_{μ} 和 μ 都是 θ 的函数，我们有

$$\nabla_{\theta} v_{\mu}(s) = \nabla_{\theta} q_{\mu}(s, \mu(s)) = (\nabla_{\theta} q_{\mu}(s, a))|_{a=\mu(s)} + \nabla_{\theta} \mu(s) (\nabla_a q_{\mu}(s, a))|_{a=\mu(s)}. \tag{10.21}$$

在非折扣情况下，根据动作价值的定义 (9.3.2 节) 可得

$$\begin{aligned}
q_{\mu}(s, a) &= \mathbb{E}[R_{t+1} - \bar{r}_{\mu} + v_{\mu}(S_{t+1})|s, a] \\
&= \sum_r p(r|s, a)(r - \bar{r}_{\mu}) + \sum_{s'} p(s'|s, a)v_{\mu}(s') \\
&= r(s, a) - \bar{r}_{\mu} + \sum_{s'} p(s'|s, a)v_{\mu}(s').
\end{aligned}$$

由于 $r(s, a) = \sum_r r p(r|s, a)$ 独立于 θ ，我们有

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

将上述方程代入 (10.21) 得

$$\nabla_{\theta} v_{\mu}(s) = -\nabla_{\theta} \bar{r}_{\mu} + \sum_{s'} p(s'|s, \mu(s)) \nabla_{\theta} v_{\mu}(s') + \underbrace{\nabla_{\theta} \mu(s) (\nabla_a q_{\mu}(s, a))|_{a=\mu(s)}}_{u(s)}, \quad s \in \mathcal{S}.$$

虽然上述方程对所有 $s \in \mathcal{S}$ 都成立，我们可以结合这些方程得到矩阵-向量形式：

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

其中 $n = |\mathcal{S}|$, m 是 θ 的维度, P_{μ} 是状态转移矩阵且 $[P_{\mu}]_{ss'} = p(s'|s, \mu(s))$, \otimes 是克罗内克积 (Kronecker product)。上述矩阵-向量形式可以简洁地写为

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

因此

$$1_n \otimes \nabla_{\theta} \bar{r}_{\mu} = u + (P_{\mu} \otimes I_m) \nabla_{\theta} v_{\mu} - \nabla_{\theta} v_{\mu}. \quad (10.22)$$

由于 d_{μ} 是平稳分布，我们有 $d_{\mu}^T P_{\mu} = d_{\mu}^T$ 。在 (10.22) 两边同时左乘 $d_{\mu}^T \otimes I_m$ 得

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

由于 $d_{\mu}^T 1_n = 1$ ，上述方程变为

$$\begin{aligned} \nabla_{\theta} \bar{r}_{\mu} &= d_{\mu}^T \otimes I_m u \\ &= \sum_{s \in \mathcal{S}} d_{\mu}(s) u(s) \\ &= \sum_{s \in \mathcal{S}} d_{\mu}(s) \nabla_{\theta} \mu(s) (\nabla_a q_{\mu}(s, a))|_{a=\mu(s)} \\ &= \mathbb{E}_{S \sim d_{\mu}} [\nabla_{\theta} \mu(S) (\nabla_a q_{\mu}(S, a))|_{a=\mu(S)}]. \end{aligned}$$

证明完成。

10.4.2 算法描述

基于定理 10.2 给出的梯度，我们可以应用梯度上升算法来最大化 $J(\theta)$ ：

$$\theta_{t+1} = \theta_t + \alpha_{\theta} \mathbb{E}_{S \sim \eta} [\nabla_{\theta} \mu(S) (\nabla_a q_{\mu}(S, a))|_{a=\mu(S)}].$$

相应的随机梯度上升算法为

$$\theta_{t+1} = \theta_t + \alpha_{\theta} \nabla_{\theta} \mu(s_t) (\nabla_a q_{\mu}(s_t, a))|_{a=\mu(s_t)}.$$

算法 10.4 总结了该算法的实现。需要注意的是，该算法是 off-policy 的，因为行为策略 β 可能与 μ 不同。首先，演员 (actor) 是 off-policy 的。我们在介绍定理 10.2 时已经解释了原因。

其次，评论家 (critic) 也是 off-policy 的。必须特别注意为什么评论家是 off-policy 的但不需要重要性采样技术。具体来说，评论家所需的经验样本是 $(s_t, a_t, r_{t+1}, s_{t+1}, \tilde{a}_{t+1})$ 。其中 $\tilde{a}_{t+1} = \mu(s_{t+1})$ 这个经验样本的生成涉及两个策略。第一个是在 s_t 处生成 a_t 的策略，第二个是在 s_{t+1} 处生成 \tilde{a}_{t+1} 的策略。生成 a_t 的第一个策略是行为策略，因为 a_t 用于与环境进行交互。第二个策略必须是 μ ，因为它是评论家旨在评估的策略。值得注意的是， \tilde{a}_{t+1} 并不用于在下一个时间步与环境进行交互。因此， μ 不是行为策略。所以，评论家是异策略的。

如何选择函数 $q(s, a, w)$? 提出确定性策略梯度方法的原始研究工作 [74] 采用了线性函数： $q(s, a, w) = \phi^T(s, a)w$ ，其中 $\phi(s, a)$ 是特征向量。正如深度确定性策略梯度 (DDPG) 方法 [75] 所建议的那样，目前流行使用神经网络来表示 $q(s, a, w)$ 。

■ 算法 10.4：确定性策略梯度或确定性演员-评论家

初始化：给定的行为策略 $\beta(a|s)$ 。确定性目标策略 $\mu(s, \theta_0)$ ，其中 θ_0 是初始参数。价值函数 $q(s, a, w_0)$ ，其中 w_0 是初始参数。 $\alpha_w, \alpha_{\theta} > 0$ 。

目标：学习一个最优策略以最大化 $J(\theta)$ 。

在每个回合 (episode) 的时间步 t ，执行：

遵循 β 生成 a_t ，并观察 r_{t+1}, s_{t+1} 。

TD 误差：

$$\delta_t = r_{t+1} + \gamma q(s_{t+1}, \mu(s_{t+1}, \theta_t), w_t) - q(s_t, a_t, w_t)$$

演员（策略更新）：

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu(s_t, \theta_t) (\nabla_a q(s_t, a, w_t))|_{a=\mu(s_t)}$$

评论家（价值更新）：

$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w q(s_t, a_t, w_t)$$

如何选择行为策略 β ？它可以是任何探索性策略。它也可以是通过向 μ 添加噪声而获得的随机策略 [75]。在这种情况下， μ 也是行为策略，因此这种方式是on-policy实现。

10.5 总结

在本章中，我们介绍了演员-评论家 (actor-critic) 方法。内容总结如下。

- 10.1 节介绍了最简单的演员-评论家算法，称为 QAC。该算法类似于上一章介绍的策略梯度算法 REINFORCE。唯一的区别在于，QAC 中的 Q 值估计依赖于 TD 学习，而 REINFORCE 依赖于蒙特卡洛估计。
- 10.2 节将 QAC 扩展为优势演员-评论家 (advantage actor-critic)。结果表明，策略梯度对于任何额外的基线都是不变的。随后表明，最优基线有助于减少估计方差。
- 10.3 节进一步将优势演员-评论家算法扩展到异策略 (off-policy) 情况。为此，我们引入了一种称为重要性采样 (importance sampling) 的重要技术。
- 最后，虽然之前介绍的所有策略梯度算法都依赖于随机策略，但我们在 10.4 节中表明，策略可以被强制为确定性的。推导了相应的梯度，并介绍了确定性策略梯度算法。

策略梯度和演员-评论家方法在现代强化学习中被广泛使用。文献中存在大量高级算法，如 SAC [76, 77]、TRPO [78]、PPO [79] 和 TD3 [80]。此外，单智能体情况也可以扩展到多智能体强化学习的情况 [81–85]。经验样本也可用于拟合系统模型，以实现基于模型的强化学习 [15, 86, 87]。分布强化学习 (distributional reinforcement learning) 提供了与传统观点截然不同的视角 [88, 89]。强化学习与控制理论之间的关系已在 [90–95] 中进行了讨论。本书无法涵盖所有这些主题。希望本书奠定的基础能帮助读者在未来更好地学习它们。

10.6 问答

- 问：演员-评论家方法 (actor-critic methods) 和策略梯度方法 (policy gradient methods) 之间有什么关系？

答：演员-评论家方法实际上就是策略梯度方法。有时，我们会交替使用这两个术语。在任何策略梯度算法中，都需要估计动作价值。当使用带有价值函数近似的时序差分学习 (temporal-difference learning) 来估计动作价值时，这种策略梯度算法被称为演员-评论家算法。“演员-评论家”这个名字突出了其结合策略更新和价值更新组件的算法结构。这种结构也是所有强化学习算法中使用的基本结构。

- 问：为什么在演员-评论家方法中引入额外的基线很重要？

答：由于策略梯度对于任何额外的基线都是不变的，我们可以利用基线来减小估计方差。由此产生的算法被称为优势演员-评论家 (advantage actor-critic) 算法。

- 问：重要性采样 (importance sampling) 可以用于策略基方法以外的基于价值的算法吗？

答：答案是肯定的。这是因为重要性采样是一种通用技术，用于利用从另一个分布中抽取的一些样本来估计随机变量在一个分布上的期望。这种技术在强化学习中之所以有用，是因为强化学习中的许多问题都是估计期望。例如，在基于价值的方法中，动作或状态价值被定义为期望。在策略梯度方法中，真实的梯度也是一个期望。因此，重要性采样既可以应用于基于价值的算法，也可以应用于基于策略的算法。事实上，它已经被应用于算法 10.3 的基于价值的组件中。

- 问：为什么确定性策略梯度方法是异策略的 (off-policy)？

答：确定性情况下的真实梯度不涉及动作随机变量。结果是，当我们使用样本来近似真实梯度时，不需要对动作进行采样，因此可以使用任何策略。所以，确定性策略梯度方法是异策略的。