8 模仿学习

为了缓解深度强化学习中的低样本效率问题，模仿学习（Imitation Learning）或称学徒学习（Apprenticeship Learning）是一种可能的解决方式，在连续决策过程中利用专家示范来更快速地实现策略优化。为了让读者全面理解如何高效地从示范数据中提取信息，我们将介绍模仿学习中最重要的几类方法，包括行为克隆（Behavioral Cloning）、逆向强化学习（Inverse Reinforcement Learning）、从观察量（Observations）进行模仿学习、概率性方法和一些其他方法。在强化学习的范畴下，模仿学习可以用作对智能体训练的初始化或引导。在实践中，结合模仿学习和强化学习是一种可以有效学习并进行快速策略优化的方法。

8.1 简介

如我们所知，强化学习，尤其是无模型的强化学习，有着低样本效率的问题，如第 7 章所讨论的。通常用其解决一个不是很复杂的任务并达到人类级别的表现可能需要成百上千的样本。然而，人类可以用少得多的时间和样本来解决这些任务。为了改进强化学习的算法效率，除了通过更精细地设计强化学习算法本身，我们实际上可以让智能体利用一些额外的信息源，比如专家示范（Expert Demonstrations）。这些专家示范依据先验知识而对策略选择有一定偏向性，而这些有效的偏见可以通过一个适当的学习过程而被提取或转移到强化学习的智能体策略中。从专家示范中学习的任务被称为模仿学习（Imitation Learning，IL），也称为学徒学习（Apprenticeship Learning）。人类和动物天生就有模仿同类其他个体的能力，这启发了让智能体从其他个体的示范中进行模仿学习的方法。相比于强化学习，监督学习在数据使用方面是一种更加高效的方法，因为它可以利用有标签数据。因此，如果示范数据是以有标签的形式提供的，监督学习的方法可以被融合到智能体的学习过程中来改进它的学习效率。

本章中，我们将介绍不同的使用示范进行策略学习的方法。图 8.1 是对模仿学习中各个类别方法的概览。我们将在后续小节中详细介绍各种模仿学习方法，并将它们总结成几个主要的类别，包括（1）行为克隆（Behavioral Cloning，BC），（2）逆向强化学习（Inverse Reinforcement Learning，IRL），（3）从观察量进行模仿学习（Imitation Learning from Observations，IfO，（一些文献 (Sun et al., 2019b) 中称 ILFO），（4）概率推理，（5）其他方法。BC 是一种最简单和直接的通过监督学习方式利用示范数据的方法，由于它的简便性而被广泛使用并作为其他更高级方法的基石。IRL 对于某些应用情况是有用的，比如难以写出显式的奖励函数（Explicit Reward Function）来实现在不同的目标之间权衡的情况。举例来说，对于一个基于视觉观察量的自动驾驶车辆，多少注意力应当被分配到处理不同的反光镜上，这难以通过奖励函数工程的方式来定义。IRL 是一种可以从示范数据中恢复未知奖励函数的方法，从而促进强化学习过程。IfO 实际上解决了模仿学习的一个缺陷，即它通常要求每个状态输入都伴有动作标签，而这种方式在人类的模仿学习过程中也是经常发生的。从概率推理的角度出发的方法包括用高斯混合模型回归（Gaussian Mixture Model Regression）或高斯过程回归（Gaussian Process Regression）来表示示范数据并引导动作策略，在某些情况下这是比深度神经网络更高效的替代方法。也有一些其他方法，比如对离线策略（Off-Policy）强化学习直接将示范数据送入经验回放缓存（Replay Buffer）等。在介绍了不同模仿学习方法的基本类别后，我们将讨论模仿学习和强化学习的关系，比如将模拟学习用作强化学习的初始化，来提高强化学习的效率。最终，我们将介绍一些其他的伴随强化学习的具体模仿学习方法，它们可能是之前一些概念和方法的组合，或者我们之前总结过的方法类别之外的方法，如图 8.1 所总结的。

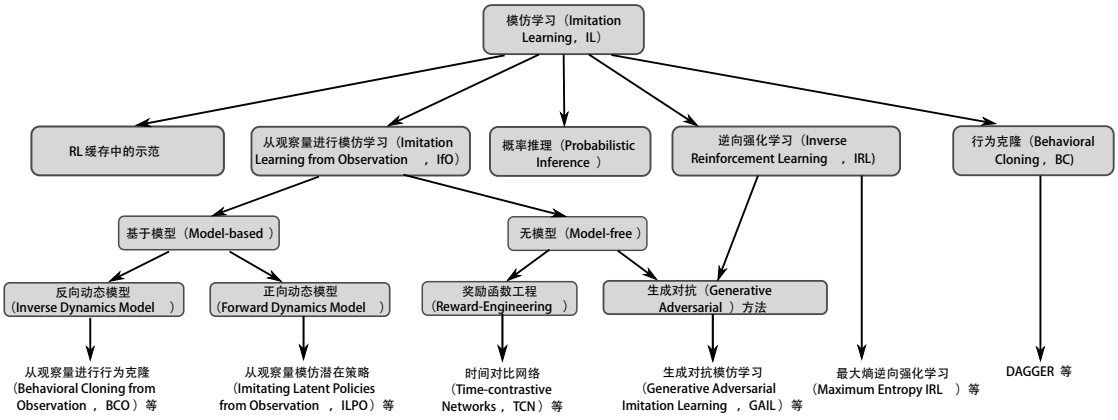

图 8.1 模仿学习算法概览

模仿学习的概念可以用学徒学习的形式 (Abbeel et al., 2004) 来定义：按照一个未知的奖励函数 $r(s,a)$，学习者找到一个策略 π 能够表现得和专家策略 π_{E} 相当。我们定义一个策略 $\pi \in \Pi$ 的占用率（Occupancy）的度量 $\rho_\pi \in \mathcal{D} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 为：$\rho_\pi(s,a) = \pi(a|s) \sum_{t=0}^{\infty} \gamma^t p(S_t = s|\pi)$ (Puterman,

2014)，这是一个用当前策略估计的状态和动作的联合分布。由于 Π 和 $\mathcal{D}$ 的一一对应关系，模仿学习的问题等价于 $\rho_\pi(s,a)$ 和 $\rho_{\pi_{\mathrm{E}}}(s,a)$ 之间的一个匹配问题。模仿学习的一个普遍目标是学习这样一个策略：

$$\hat{\pi} = \arg\min_{\pi\in\Pi} \psi^*(\rho_\pi - \rho_{\pi_{\mathrm{E}}}) - \lambda H(\pi) \tag{8.1}$$

其中 ψ^* 是一个 ρ_π 和 $\rho_{\pi_{\mathrm{E}}}$ 之间的距离度量，而 $\lambda H(\pi)$ 是一个有权衡因子 λ 的正则化项。举例来说，这个正则化项可以定义为策略 π 的 γ-折扣因果熵（Causal Entropy）：$H(\pi) \stackrel{\text{def}}{=} \mathbb{E}_\pi[-\log\pi(s,a)]$。模仿学习的整体目标就是增加从当前策略采样得到的 $\{(s,a)\}$ 分布和示范数据中分布的相似度，同时考虑到策略参数上的一些限制。

8.2　行为克隆方法

如果示范数据有相应标签的话（比如，对于给定状态的一个好的动作可以被看作一个标签），利用示范的模仿学习可以自然地被看作是一个监督学习任务。在强化学习的情况下，有标签的示范数据 $\mathcal{D}$ 通常包含配对的状态和动作：$\mathcal{D} = \{(s_i,a_i)|i=1,\cdots,N\}$，其中 N 是示范数据集的大小而指标 i 表示 s_i 和 a_i 是在同一个时间步的。在满足 MDP 假设的情况下（即最优动作只依赖于当前状态），状态-动作对的顺序在训练中可以被打乱。考虑强化学习设定下，有一个以 θ 参数化和 s 为输入状态的初始策略 π_θ，其输出的确定性动作为 $\pi_\theta(s)$，我们有专家生成的示范数据集 $\mathcal{D} = \{(s_i,a_i)|i=1,\cdots,N\}$，可以用来训练这个策略，其目标如下：

$$\min_\theta \sum_{(s_i,a_i)\sim\mathcal{D}} \|a_i - \pi_\theta(s_i)\|_2^2 \tag{8.2}$$

一些随机性策略 $\pi_\theta(\tilde{a}|s)$ 的具体形式，比如高斯策略等，可以用再参数化技巧来处理：

$$\min_\theta \sum_{\tilde{a_i}\sim\pi(\cdot|s_i),(s_i,a_i)\sim\mathcal{D}} \|a_i - \tilde{a_i}\|_2^2 \tag{8.3}$$

这个使用监督学习直接模仿专家示范的方法在文献中称为**行为克隆**（Behavioral Cloning，BC）。

8.2.1　行为克隆方法的挑战

- **协变量漂移**（Covariate Shift）：尽管模仿学习可以对与示范数据集（用于训练策略）相似的样本有较好的表现，对它在训练过程中未见过的样本可能会有较差的泛化表现，因为示范数据集中只能包含有限的样本。举例来说，如果数据分布是多模式的，测试中的新样本可能跟训练中的样本来自不同的群集（Cluster），比如，在实践中将一个不同猫的分类器用于

区分狗的种类。由于 BC 方法将决策问题归结为一个监督学习问题，机器学习中，众所周知的协变量漂移 (Ross et al., 2010) 的问题可能使通过监督学习方法学得的策略很脆弱，而这对 BC 方法是一个挑战。图 8.2 进一步阐释了 BC 中的协变量漂移。

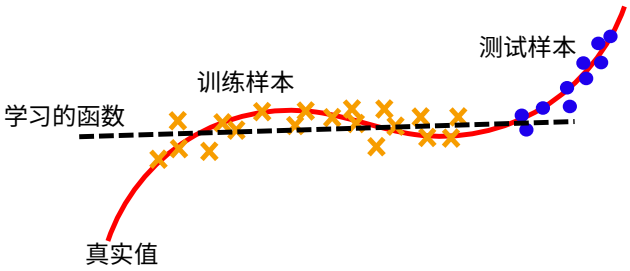

图 8.2 协变量漂移：所学的函数（虚线）对训练样本可以很好地拟合（交叉符号），但是对测试样本（点符号）有很大的预测偏差。线是真实值

- **复合误差**（Compounding Errors）：BC 方法在很大程度上受复合误差的影响，这是一种小误差可以随时间累积而最终导致显著不同的状态分布 (Ross et al., 2011) 的现象。强化学习任务的 MDP 性质是导致复合误差的主要因素，即连续误差的放大效应。而在 BC 方法中，实际上在每一个时间步上产生的误差主要可能是由上面所述的协变量漂移所造成的。图 8.3 展示了复合误差。

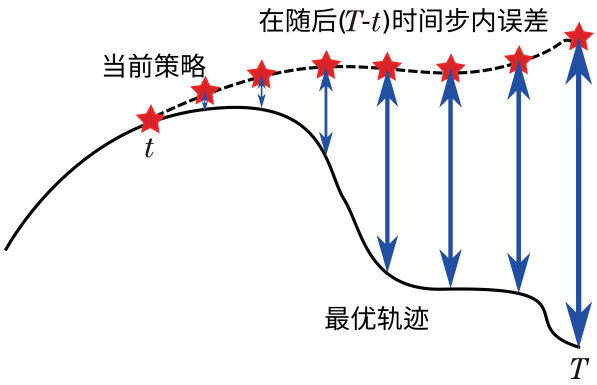

图 8.3 在一个连续决策任务中，复合误差沿着当前策略选择的轨迹逐渐增加

8.2.2 数据集聚合

数据集聚合（Dataset Aggregation，DAgger）(Ross et al., 2011) 是一种更先进的基于 BC 方法的从示范中模仿学习的算法，它是一种无悔的（No-Regret）迭代算法。根据先前的训练迭代过程，它主动选择策略，在随后过程中有更大几率遇到示范样本，这使得 DAgger 成为一种更有用且高

效的在线模仿学习方法，可以应用于像强化学习中的连续预测问题。示范数据集 $\mathcal{D}$ 会在每个时间步 i 连续地聚合新的数据集 $\mathcal{D}_i$，这些数据集包含当前策略在整个模仿学习过程中遇到的状态和相应的专家动作。因此，DAgger 同样有一个缺陷，即它需要不断地与专家交互，而这在现实应用中通常是一种苛求。DAgger 的伪代码如算法 8.29 所示，其中 π^* 是专家策略，而 β_i 是在迭代 i 时对策略软更新（Soft-Update）的参数。

算法 8.29 DAgger

初始化 $\mathcal{D} \leftarrow \emptyset$
初始化策略 $\hat{\pi}_1$ 为策略集 Π 中任意策略
for i $= 1, 2, \cdots, N$ **do**
　$\pi_i \leftarrow \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$
　用 π_i 采样几个 T 步的轨迹
　得到由 π_i 访问的策略和专家给出的动作组成的数据集 $\mathcal{D}_i = \{(s, \pi^*(s))\}$
　聚合数据集：$\mathcal{D} \leftarrow \mathcal{D} \cup \mathcal{D}_i$
　在 $\mathcal{D}$ 上训练策略 $\hat{\pi}_{i+1}$
end for
返回策略 $\hat{\pi}_{N+1}$

8.2.3 Variational Dropout

一种缓解模仿学习中泛化问题的方法是预训练并使用 **Variational Dropout** (Blau et al., 2018)，来替代 BC 方法中完全克隆专家示范的行为。在这个方法中，使用示范数据集预训练（模仿学习）得到的权重被参数化为高斯分布，并用一个确定的方差阈值来进行高斯 Dropout，然后用来初始化强化学习策略。对于模仿学习的 Variational Dropout 方法 (Molchanov et al., 2017) 可以被看作一种相比于在预训练的权重中加入噪声来说更高级的泛化方法，它可以减少对噪声大小选择的敏感性，因而是一种使用模仿学习来初始化强化学习的有用技巧。

8.2.4 行为克隆的其他方法

行为克隆方法也包含了其他一些概念。比如，一些方法提供了在一个任务中将示范数据泛化到更一般情形的方法，比如**动态运动基元**（Dynamic Movement Primitives，DMP）(Pastor et al., 2009) 法，它使用一系列微分方程（Differential Equations）来表示任何记录过的运动。DMP 中的微分方程通常包含可调整的权重，以及非线性函数来生成任意复杂运动。因此在行为克隆中，相比于“黑盒”深度学习方法，DMP 更像是一种解析形式的解决方法。此外，有一种单样本的（One-Shot）模仿学习方法 (Duan et al., 2017) 使用对示范数据的柔性注意力（Soft Attention）来将模型泛化到在训练数据中未见过的情景。它是一种元学习（Meta-Learning）的方法，在多个任务中将一个任务的一个示范映射到一个有效的策略上。相关的方法不限于此，在这里不做过多介绍。

8.3 逆向强化学习方法

8.3.1 简介

另一种主要的模仿学习方法基于**逆向强化学习**（Inverse Reinforcement Learning，IRL）(Ng et al., 2000; Russell, 1998)。IRL 可以归结为解决从观察到的最优行为中提取奖励函数（Reward Function）的问题，这些最优行为也可以表示为专家策略 π_{E}。基于 IRL 的方法反复地在两个过程中交替：一个是使用示范来推断一个隐藏的奖励或代价（Cost）函数，另一个是使用强化学习基于推断的奖励函数来学习一个模仿策略。IRL 选择奖励函数 R 来最优化策略，并且使得任何脱离于 π_{E} 的单步选择尽可能产生更大损失。对于所有满足 $|R(s)| \leqslant R_{\max}, \forall s$ 的奖励函数 R，IRL 用以下方式选择 R^*：

$$R^* = \arg\max_R \sum_{s\in\mathcal{S}} (Q^\pi(s, a_{\mathrm{E}}) - \max_{a\in A\setminus a_{\mathrm{E}}} Q^\pi(s, a)) \tag{8.4}$$

其中 $a_{\mathrm{E}} = \pi_{\mathrm{E}}(s)$ 或 $a_{\mathrm{E}} \sim \pi(\cdot|s)$ 是专家（最优的）动作。基于 IRL 的技术已经被用于许多任务，比如操控一个直升机 (Abbeel et al., 2004) 和物体控制 (Finn et al., 2016b)。IRL (Ng et al., 2000; Russell, 1998) 企图从观察到的最优行为，比如专家示范中提取一个奖励函数，但是这个奖励函数可能不是唯一的（在之后有所讨论）。IRL 中一个典型的方法是使用最大因果熵（Maximum Causal Entropy）正则化，即最大熵（Maximum Entropy, MaxEnt）IRL (Ziebart et al., 2010) 方法。MaxEnt IRL 可以表示为以下两个步骤：

$$\mathrm{IRL}(\pi_{\mathrm{E}}) = \arg\max_R \mathbb{E}_{\pi_{\mathrm{E}}}[R(s, a)] - \mathrm{RL}(R) \tag{8.5}$$

$$\mathrm{RL}(R) = \max_\pi H(\pi(\cdot|s)) + \mathbb{E}_\pi[R(s, a)] \tag{8.6}$$

这构成了 $\mathrm{RL} \circ \mathrm{IRL}(\pi_{\mathrm{E}})$ 策略学习架构。第一个式子 $\mathrm{IRL}(\pi_{\mathrm{E}})$ 学习一个奖励函数来最大化专家策略和强化学习策略间的奖励值差异，并且由于 Q 值是对奖励的估计，它可以被式 (8.4) 替代。第二个式子 $\mathrm{RL}(R)$ 是熵正则化（Entropy-Regularized）正向强化学习，而其奖励函数 R 是第一个式子学到的。这里的熵 $H(\pi(\cdot|s))$ 是给定状态下的策略分布的熵函数。

关于随机变量 X 的分布 $p(X)$ 的香农信息熵度量了这个概率分布的不确定性。

定义 8.1 *一个满足 p 分布的离散随机变量 X 的信息熵为*

$$H_p(X) = \mathbb{E}_{p(X)}[-\log p(X)] = -\sum_{X\in\mathcal{X}} p(X)\log p(X) \tag{8.7}$$

对于强化学习中随机策略的情况，表示动作分布的随机变量通常排列成一个与动作空间维数

相同的矢量。常用的分布有对角高斯分布和类别分布，导出它们的熵是很简单的。

代价函数 $c(s,a) = -R(s,a)$ 也很常见，它在强化学习的过程中被最小化：

$$\mathrm{RL}(c) = \arg\min_{\pi} -H(\pi) + \mathbb{E}_{\pi}[c(s,a)] \tag{8.8}$$

其中 $H(\pi) = \mathbb{E}_{\pi}[-\log \pi(a|s)]$ 是策略 π 的熵。代价函数 $c(s,a)$ 常用作当前策略 π 的分布和示范数据间相似度的度量。熵 $H(\pi)$ 可以被视作实现最优解的唯一性的正则化项。

把上式代入 IRL 公式 (8.5) 中，我们可以将 IRL 的目标表示成 max-min 的形式，它企图在最大化熵正则化奖励值的目标下学习一个状态 s 和动作 a 的代价函数 $c(s,a)$，以及进行策略 π 的学习。

$$\max_{c}(\min_{\pi} -\mathbb{E}_{\pi}[-\log \pi(a|s)] + \mathbb{E}_{\pi}[c(s,a)]) - \mathbb{E}_{\pi_{\mathrm{E}}}[c(s,a)] \tag{8.9}$$

其中 π_{E} 表示生成专家示范的专家策略，而 π 是强化学习过程训练的策略。所学的代价函数将给专家策略分配较高的熵而给其他策略较低的熵。

8.3.2　逆向强化学习方法的挑战

- **奖励函数的非唯一性**或**奖励歧义**（Reward Ambiguity）：IRL 的函数搜索是病态的（Ill-Posed），因为示范行为可以由多个奖励或代价函数导致。它始于奖励塑形（Reward Shaping）(Ng et al., 1999) 的概念，这个概念描述了一类能保持最优策略的奖励函数变换。主要的结果是，在以下奖励变换之下：

$$\hat{r}(s,a,s') = r(s,a,s') + \gamma\phi(s') - \phi(s), \tag{8.10}$$

  最优策略对任何函数 $\phi : \mathcal{S} \to \mathbb{R}$ 保持不变。只用示范数据通过 IRL 方法学到的奖励函数，是不能消除上面一类变换下奖励函数之间分歧的。

  因此，我们需要对奖励或者策略施加限制来保证示范行为最优解的唯一性。举例来说，奖励函数通常被定义为一个状态特征的线性组合 (Abbeel et al., 2004; Ng et al., 2000) 或凸的组合（Convex Combination）(Syed et al., 2008)。所学的策略也假设其满足最大熵 (Ziebart et al., 2008) 或者最大因果熵 (Ziebart et al., 2010) 规则。然而，这些显式的限制对所提出方法 (Ho et al., 2016) 的通用性有一定潜在限制。

- **较大的计算代价**：IRL 可以在一般强化学习过程中通过示范和交互学到一个更好的策略。然而，在推断出的奖励函数下，使用强化学习来优化策略要求智能体与它的环境交互，这从时间和安全性的角度考虑都可能是要付出较大代价的。此外，IRL 的步骤主要要求智能体在迭代优化奖励函数 (Abbeel et al., 2004; Ziebart et al., 2008) 的内循环中解决一个 MDP 问

题，而这从计算的角度也可能是有极大消耗的。然而，近来有一些方法被提出，以减轻这个要求 (Finn et al., 2016b; Ho et al., 2016)。其中一种方法称为生成对抗模仿学习（Generative Adversarial Imitation Learning，GAIL）(Ho et al., 2016)。

8.3.3 生成对抗模仿学习

生成对抗模仿学习（Generative Adversarial Imitation Learning，GAIL）(Ho et al., 2016) 采用了生产对抗网络（Generative Adversarial Networks，GANs）(Goodfellow et al., 2014) 中的生成对抗方法。相关算法可以被想成是企图引入一个对模仿者的状态-动作占用率（Occupancy）的度量，使之与示范者的相关特性类似。它使用一个 GAN 中的辨别器（Discriminator）来给出基于示范数据的动作-价值（Action Value）函数估计。对于一般基于动作价值函数的强化学习过程来说，动作-价值可以通过一种生成式方法来从示范中得到：

$$Q(s,a) = \mathbb{E}_{\mathcal{T}_i}[\log(D_{\omega_{i+1}}(s,a))], \tag{8.11}$$

其中 $\mathcal{T}_i$ 迭代次数为 i 时探索的样本集合，而 $D_{\omega_{i+1}}(s,a)$ 是来自辨别器的输出值 $D_{\omega_{i+1}}(s,a)$，辨别器的参数为 ω_{i+1}。ω_{i+1} 表示 Q 值是在更新了一步辨别器的参数过后再估计的，因此迭代次数是 $i+1$。辨别器的损失函数定义为一般形式：

$$\text{Loss} = \mathbb{E}_{\mathcal{T}_i}[\nabla_\omega \log(D_\omega(s,a))] + \mathbb{E}_{\mathcal{T}_\text{E}}[\nabla_\omega \log(1 - D_\omega(s,a))] \tag{8.12}$$

其中 $\mathcal{T}_i$，$\mathcal{T}_\text{E}$ 分别是来自探索和专家示范的样本集合，而 ω 是辨别器的参数。图 8.4 展示了 GAIL 的结构。

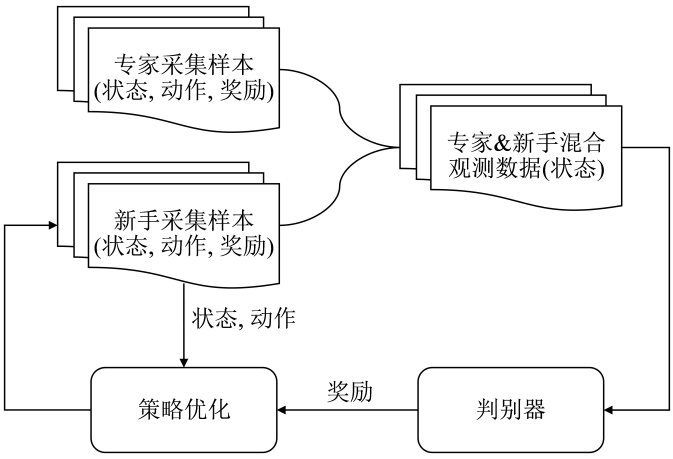

图 8.4 GAIL 的结构，改编自文献 (Ho et al., 2016)

通过 GAIL 方法，策略可以通过由示范数据泛化得到的样本进行学习，而且相比于使用 IRL 的方法有较低的计算消耗。它也不需要在训练中跟专家进行交互，而像 DAgger 等方法可能需要这种实际上有时难以得到的交互数据。

这种方法可以进一步推广到多模态的（Multi-Modal）策略来从多任务中学习。基于 GAN 的多模态模仿学习 (Hausman et al., 2017) 将一个更高级的目标函数（额外的潜在指标表示不同的任务）用于生成对抗过程中，从而自动划分来自不同任务的示范，并以模仿学习的方法学习一个多模态策略。

根据文献 (Goodfellow et al., 2014)，如果有无限的数据和无限的计算资源，在最优情况下，以 GAIL 的目标生成的状态-动作分布应当完全匹配示范数据的状态-动作对。然而，这种方法的缺点是，我们绕过了生成奖励的中间步骤，即我们不能从辨别器中提取奖励函数，因为 $D_\omega(s,a)$ 对于所有的 (s,a) 将收敛到 0.5。

8.3.4　生成对抗网络指导性代价学习

如上所述，GAIL 方法无法从示范数据中恢复奖励函数。一个类似的工作称为生成对抗网络指导性代价学习（Generative Adversarial Network Guided Cost Learning，GAN-GCL），它基于 GAN 的结构来优化一个指导性代价学习（Guided Cost Learning，GCL）方法，以此来从使用示范数据训练的最优辨别器中提取一个最优的奖励函数。我们将详细介绍该方法。

GAN-GCL 方法（具体来说 GCL 部分）是基于之前介绍的最大因果熵反向强化学习方法的，它考虑一个熵正则化马尔可夫决策过程（Markov Decision Process，MDP）。熵正则化 MDP 对于强化学习的目标是最大化熵正则化折扣奖励的期望（Expected Entropy-Regularized Discounted Reward）：

$$\pi^* = \arg\max_\pi \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\mathrm{T}} \gamma^t (r(S_t, A_t) + H(\pi(\cdot|S_t)))\right], \tag{8.13}$$

这是源自式 (8.5) 的用于实际学习策略的一个具体形式。可以看出最优策略 $\pi^*(a|s)$ 给出的轨迹分布满足 $\pi^*(a|s) \propto \exp(Q^*_{\text{soft}}(s,a))$ (Ziebart et al., 2010)，其中 $Q^*_{\text{soft}}(S_t, A_t) = r(S_t, A_t) + \mathbb{E}_{\tau\sim\pi}[\sum_{t'=t}^{\mathrm{T}} \gamma^{t'-t}(r(s_{t'}, a_{t'}) + H(\pi(\cdot|s_{t'})))]$ 表示柔性 Q 函数（Soft Q-Function），这在柔性 Actor-Critic 算法中也有用到。

IRL 问题可以被理解为解决如下一个极大似然估计（Maximum Likelihood Estimation，MLE）问题：

$$\max_\theta \mathbb{E}_{\tau\sim\pi_{\mathrm{E}}}[\log p_\theta(\tau)], \tag{8.14}$$

其中 π_{E} 是提供示范的专家策略，而 $p_\theta(\tau) \propto p(S_0)\prod_{t=0}^{\mathrm{T}} p(S_{t+1}|S_t, A_t)\mathrm{e}^{\gamma^t r_\theta(S_t, A_t)}$ 以奖励函数

$r_\theta(s,a)$ 的参数 θ 为参数，并且依赖 MDP 的初始状态分布和动态变化（或称状态转移）。$p_\theta(\tau)$ 是示范数据以轨迹为中心的（Trajectory-Centric）分布，这些数据是从以状态为中心的（State-Centric）π_{E} 得来的，即 $p_\theta(\tau) \sim \pi_{\mathrm{E}}$。根据确定性转移过程中 $p(S_{t+1}|S_t, A_t) = 1$，其简化为一个基于能量的模型 $p_\theta(\tau) \propto \mathrm{e}^{\sum_{t=0}^{\mathrm{T}} \gamma^t r_\theta(S_t, A_t)}$ (Ziebart et al., 2008) 。参数化的奖励函数可以按照上面的目标来优化参数 θ。与之前的过程类似，我们在这里可以引入代价函数作为累积折扣奖励（Cumulative Discounted Rewards）$c_\theta = -\sum_{t=0}^{\mathrm{T}} \gamma^t r_\theta(S_t, A_t)$ 的负值，它也由 θ 参数化。那么 MaxEnt IRL 可以看作是使用玻尔兹曼分布（Boltzmann Distribution）在以轨迹为中心的形式下对示范数据建模的结果，其中由代价函数 c_θ 给出的能量为

$$p_\theta(\tau) = \frac{1}{Z} \exp(-c_\theta(\tau)), \tag{8.15}$$

其中 τ 是状态-动作轨迹，而 $c_\theta(\tau) = \sum_t c_\theta(S_t, A_t)$ 总的代价函数，配分函数（Partition Function）Z 是 $\exp(-c_\theta(\tau))$ 对所有符合环境动态变化的轨迹的积分，用以归一化概率。对于大规模或连续空间的情况，准确估计配分函数 Z 会很困难，因为通过动态规划（Dynamic Programming）对 Z 的精确估计只适用于小规模离散情况。否则我们需要使用近似估计的方法，比如基于采样的（Sampling-Based）GCL 方法。

GCL 使用重要性采样（Importance Sampling）来以一个新的分布 $q(\tau)$（原来的示范数据分布为 $p(\tau)$）估计 Z，并采用 MaxEnt IRL 的形式：

$$\theta^* = \arg\min_\theta \mathbb{E}_{\tau\sim p}[-\log p_\theta(\tau)] \tag{8.16}$$

$$= \arg\min_\theta \mathbb{E}_{\tau\sim p}[c_\theta(\tau)] + \log Z \tag{8.17}$$

$$= \arg\min_\theta \mathbb{E}_{\tau\sim p}[c_\theta(\tau)] + \log\left(\mathbb{E}_{\tau'\sim q}\left[\frac{\exp(-c_\theta(\tau'))}{q(\tau')}\right]\right). \tag{8.18}$$

其中 τ' 是从分布 q 采样得到的，而 $q(\tau')$ 是其概率。因此 q 可以通过最小化 $q(\tau')$ 和 $\frac{1}{Z}\exp(-c_\theta(\tau'))$ 间的 KL 散度来优化，从而更新 θ 以学习 $q(\tau')$，其等价表示如下：

$$q^* = \min \mathbb{E}_{\tau\sim q}[c_\theta(\tau)] + \mathbb{E}_{\tau\sim q}[\log q(\tau)] \tag{8.19}$$

文献 (Finn et al., 2016a) 提出使用 GAN 的形式来解决上述优化问题，它使用 GAN 的结构优化 GCL，与 GAIL 方法类似但是有不同的具体形式。

注意，GAN 中的辨别器也可以实现用一个分布去拟合另一个的功能：

$$D^*(\tau) = \frac{p(\tau)}{p(\tau) + q(\tau)} \tag{8.20}$$

我们可以在这里将它用于 MaxEnt IRL 形式的 GCL。

$$D_\theta(\tau) = \frac{\frac{1}{Z}\exp(-c_\theta(\tau))}{\frac{1}{Z}\exp(-c_\theta(\tau)) + q(\tau)} \tag{8.21}$$

这产生了 GAN-GCL 方法。策略 π 被训练以最大化 $R_\theta(\tau) = \log(1 - D_\theta(\tau)) - \log D_\theta(\tau)$，从而奖励函数可以通过优化辨别器来学习。策略通过更新采样分布 $q(\tau)$ 来学习，这个采样分布是用来估计配分函数的。如果达到了最优情况，那么我们可以用所学的最优的代价函数 $c^*_\theta = -R^*_\theta(\tau) = -\sum_{t=0}^{\mathrm{T}} \gamma^t r^*_\theta(S_t, A_t)$ 来得到最优奖励函数，而最优策略可以通过 $\pi^* = q^*$ 得到。GAN-GCL 为解决 MaxEnt IRL 问题提供了一种除直接最大化似然（Maximum Likelihood）方法外的方法。

8.3.5　对抗性逆向强化学习

由于上面介绍的 GAN-GCL 是以轨迹为中心（Trajectory-Centric）的，这意味着完整的轨迹需要被估计，相比于估计单个状态动作对会有较大的估计方差。对抗性逆向强化学习（Adversarial Inverse Reinforcement Learning，AIRL）(Fu et al., 2017) 直接对单个状态和动作进行估计：

$$D_\theta(s, a) = \frac{\exp(f_\theta(s, a))}{\exp(f_\theta(s, a)) + \pi(a|s)} \tag{8.22}$$

其中 $\pi(a|s)$ 是待更新的采样分布而 $f_\theta(s, a)$ 是所学的函数。配分函数在上面式子中被忽略了，而概率值的归一性在实践中可以由 Softmax 函数或者 Sigmoid 输出激活函数来保证。经证明，在最优情况下，$f^*(s, a) = \log \pi^*(a|s) = A^*(s, a)$ 给出了最优策略的优势函数（Advantage Function）。然而，优势函数是一个高度纠缠的奖励函数减去一个基线值的结果。文献 (Fu et al., 2017) 论证说奖励函数从环境动态的变化中不能被鲁棒地恢复出来。因此，他们提出通过 AIRL 来从优势函数中解纠缠（Disentangle）以得到奖励函数：

$$D_{\theta,\phi}(s, a, s') = \frac{\exp(f_{\theta,\phi}(s, a, s'))}{\exp(f_{\theta,\phi}(s, a, s')) + \pi(a|s)} \tag{8.23}$$

其中，$f_{\theta,\phi}$ 被限制为一个奖励拟合器 g_θ 和一个塑形（Shaping）项 h_ϕ：

$$f_{\theta,\phi}(s, a, s') = g_\theta(s, a) + \gamma h_\phi(s') - h_\phi(s) \tag{8.24}$$

其中还需要对 h_ϕ 进行额外拟合。

8.4 从观察量进行模仿学习

首先，从观察量进行模仿学习（Imitation Learning from Observation，IfO）是在没有完整可观察的动作的情况下进行的模仿学习。IfO 的一个例子是从视频中学习，其中物体的真实动作值是无法单纯地通过一些帧中的信息得到的，但人类仍旧能够从视频中学习，比如模仿动作，因此，在 IfO 相关文献中经常见到从视频中学习的例子。相比于其他前面介绍过的方法，IfO 从另一个角度来看待模仿学习。因而，这一小节所介绍的具体方法和之前介绍的方法有不可避免的重叠之处，但是，要注意这一小节的方法是在 IfO 的范畴之下的。当你阅读这一小节时，应当记得，这里的 IfO 方法与其他类别的方法大多是正交的关系，因为它是从另一个角度来处理模仿学习的，并且着重于解决不可观测动作的问题。

之前提到的算法，几乎都不能用于解决只包含部分可观测或不可观测动作的示范数据的情况。一个对于学习这种类型的示范数据的想法是先从状态中恢复动作，再采用标准的模仿学习算法从恢复出来的状态-动作对（State-Action Pairs）中进行策略学习。比如，文献 (Torabi et al., 2018a) 通过学习一个状态转移（State Transition）的动态模型来恢复动作，并使用 BC 算法来找到最优策略。然而，这种方法的性能极大地依赖于所学动态模型的好坏，对于状态转移中有噪声的情况则很可能失败。相反，文献 (Merel et al., 2017) 提出只通过状态（或状态的特征值）轨迹来学习。他们拓展了 GAIL 框架，并只通过采集运动示范数据的状态来学习控制策略，展示了只需要部分状态特征而不需要示范者的具体动作对对抗式模仿（Adversarial Imitation）也是足够的。相似地，文献 (Eysenbach et al., 2018) 指出策略应该可以控制智能体到达哪些状态，因而可通过最大化策略和状态轨迹间的互信息（Mutual Information）来仅仅通过状态训练策略。也有一些其他研究尝试只从观察量而不是真实状态中学习。比如，文献 (Stadie et al., 2017) 通过域自适应（Domain Adaption）方法从观察量中提取特征来保证专家（Experts）和新手（Novices）在同一个特征空间下。然而，只使用示范状态或状态特征在训练中可能需要大量的环境交互，因为任何来自动作的信息都被忽略了。

为了提供 IfO 方法的一个清楚的框架，我们把文献中的 IfO 方法总结为两大类：（1）基于模型（Model-Based）方法；（2）无模型（Model-Free）方法。这也与强化学习中的一种主要的分类方法吻合。随后，我们讨论每一类方法的特点，并提出相关文献中的算法作为例子。

8.4.1 基于模型方法

类似于基于模型的强化学习（如第 9 章），如果环境模型可以用较低的消耗来精确学习，这个模型可能对学习过程有利，因为通过它可以高效地做出规划。由于模仿学习在与环境交互的过程中模仿的是一系列的动作而非单个动作，所以它难以避免地涉及环境的动态变化，而这可以通过基于模型方法学习。根据不同的动态模型类型，基于模型的 IfO 方法可以被分类为：（1）逆向动态模型（Inverse Dynamics Models）和（2）正向动态模型（Forward Dynamics Models）。

逆向动态模型：一个逆向动态模型是从状态转移 $\{(S_t, S_{t+1})\}$ 到动作 $\{A_t\}$ 的映射 (Hanna et al., 2017)。在这一类中的一个工作如文献 (Nair et al., 2017) 提出的方法，它通过人类操作绳子从一个初始状态到目标状态的一系列图像，来学习预测绳结操作中的一系列动作，这需要学习如下的一个像素级（Pixel-Level）的逆向动态模型：

$$A_t = M_\theta(I_t, I_{t+1}) \tag{8.25}$$

以上面的任务为例，其中 A_t 是通过逆向动态模型 M 以输入的一对图片 I_t, I_{t+1} 所预测的动作，模型由 θ 参数化，卷积神经网络被用于学习逆向动态模型。机器人通过探索策略自动地收集绳结操作的样本，收集到的样本被用于学习逆向动态模型，随后机器人使用所学的模型和来自人类示范的期望状态进行规划。学到的逆向动态模型 M_θ^* 实际可以作为策略来根据期望帧 I^{e} 选择与示范相似的动作：

$$A_t = M_\theta^*(I_t, I_{t+1}^{\mathrm{e}}) \tag{8.26}$$

另一个工作叫作增强逆向动态建模（Reinforced Inverse Dynamics Modeling，RIDM）(Pavse et al., 2019)，它在使用预定义的探索策略所收集的样本进行训练的基础上，使用一个增强的后训练（Post-Training）过程来微调所学的逆向动态模型。如上所述，预训练的逆向动态模型被看作是强化学习设置下的一个智能体策略，这时可以用一个稀疏奖励函数 R 来基于强化学习对这个策略进行微调：

$$\theta^* = \arg\max_\theta \sum_t R(S_t, M_\theta^{\mathrm{pre}}(S_t, S_{t+1}^{\mathrm{e}})) \tag{8.27}$$

其中 M_θ^{pre} 是预训练模型，在这里通过强化学习的方式来进行微调，微调目标是最大化奖励函数 R。

协方差矩阵自适应进化策略（Covariance Matrix Adaptation Evolution Strategy，CMA-ES）或者贝叶斯优化（Bayesian Optimization，BO）方法可以用于在低维的情况下优化模型。然而，作者假设每个观察量转移（Observation Transition）都可以通过单个动作实现。为了消除这个不需要的假设，文献 (Pathak et al., 2018) 允许智能体执行多个动作直到它与下一个示范帧足够接近。

上面介绍的算法试图对每个示范状态使用逆向动态模型从而实现对策略的恢复。从观察量进行行为克隆（Behavioral Cloning from Observation，BCO）算法由文献 (Torabi et al., 2018a) 提出，这个算法则试图使用完整的观察量-动作对（Observation-Action Pair）和所学的逆向动态模型来恢复示范数据集，然后用常规模仿学习的形式使用这个增强后的示范数据集来学习策略，如图 8.5 所示。

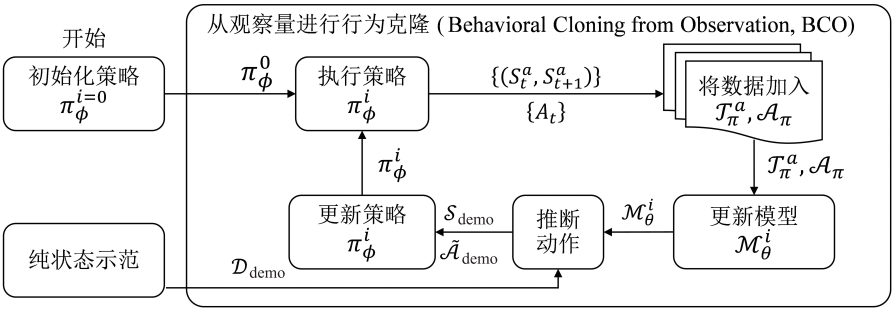

图 8.5 从观察量进行行为克隆（Behavioral Cloning from Observation，BCO）的学习框架，改编自文献 (Torabi et al., 2018a)

文献 (Guo et al., 2019) 提出使用一个基于张量的（Tensor-Based）模型来推理专家状态序列相应的未观测动作（即一个 IfO 问题），如图 8.6 所示。智能体的策略通过一个结合了强化学习和模仿学习的混合目标来优化：

$$\theta^* = \arg\min_\theta L_{\text{RL}}(\pi(a|s;\theta)) - \mathbb{E}_{(S_t^{\text{e}}, S_{t+1}^{\text{e}})\sim\mathcal{D}}[\log \pi_\theta(M(S_t^{\text{e}}, S_{t+1}^{\text{e}})|S_t^{\text{e}})] \tag{8.28}$$

其中 L_{RL} 是常规强化学习的损失项，其策略 π 由 θ 参数化。$\mathcal{D}$ 是示范数据集，而第二项是行为克隆损失函数，用于最大化基于专家状态 s^{e} 和逆向动态模型 M 预测专家动作的可能性（Likelihood）。文献 (Guo et al., 2019) 提出一种结合 RIDM 和 BCO 的方法。这里的逆向动态模型 M 是一个低秩的（Low-Rank）张量模型，而非像上面介绍的其他方法中的参数化（Parameterized）模型，它在某些情况下比深度神经网络有优势。类似于 RIDM，这个方法需要提供奖励信号（Reward Signals）来得到强化学习损失函数。

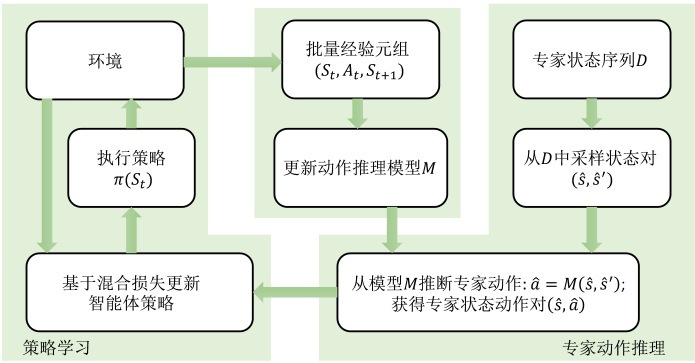

图 8.6 混合强化学习和专家状态序列的学习框架，改编自文献 (Guo et al., 2019)

正向动态模型：正向动态模型是从状态-动作对 $\{(S_t, A_t)\}$ 到下一个状态 $\{S_{t+1}\}$ 的映射。一

个典型的在 IfO 中使用正向动态模型的方法叫作从观察量模仿潜在策略（Imitating Latent Policies from Observation，ILPO）(Edwards et al., 2018)。ILPO 在其学习过程中使用两个网络：潜在策略（Latent Policy）网络和动作重映射（Action Remapping）网络。潜在策略网络包括一个动作推理（Action Inference）模块，它将状态 S_t 映射到一个潜在动作（Latent Action）z，而一个正向动态模块根据当前状态 S_t 和潜在动作 z 预测下一个状态 S_{t+1}。这两个模块的更新规则如下：

$$\omega^* = \arg\min \mathbb{E}_{(S_t^{\mathrm{e}}, S_{t+1}^{\mathrm{e}})\sim\mathcal{D}}[\|G_\omega(S_t^{\mathrm{e}}, z) - S_{t+1}^{\mathrm{e}}\|_2^2] \tag{8.29}$$

这是对于潜在动态模块 G_ω 的，而

$$\theta^* = \arg\max \mathbb{E}_{(S_t^{\mathrm{e}}, S_{t+1}^{\mathrm{e}})\sim\mathcal{D}} \left[\left\| \sum_z \pi_\theta(z|S_t^{\mathrm{e}}) G_\omega(S_t^{\mathrm{e}}, z) - S_{t+1}^{\mathrm{e}} \right\|_2^2 \right] \tag{8.30}$$

是对于潜在策略 $\pi_\theta(\cdot|z)$ 而言的，其中 $\mathcal{D}$ 是专家示范数据集。

然而，由于潜在策略网络产生的潜在动作可能并不是真正的环境动态中的真实动作，动作重映射网络被用来将潜在动作关联到真实动作。使用潜在动作不需要在学习潜在模型和潜在策略的过程中与环境进行交互，而动作网络重映射只需要跟环境交互有限的次数，这使得整个算法在学习过程中很高效（Efficient）。

8.4.2　无模型方法

除了使用所学动态模型进行基于模型的 IfO 方法，也有一些无模型 IfO 方法，这属于另一个主要的方法类别，即不使用模型进行学习。对于高度复杂的动态变化，模型可能很难学习，这与在常规强化学习设置中的情况一样。对于无模型 IfO 有两个主要的方法：（1）生成对抗（Generative Adversarial）方法和（2）奖励函数工程（Reward Engineering）方法。其中生成对抗方法类似于常规模仿学习中的，但是只有状态作为示范。

生成对抗方法：一种基本的生成对抗 IfO 的框架是由之前介绍的在常规模仿学习设置下 IRL 中的 GAIL 方法改进的。辨别器（Discriminator）只判别和比较当前策略探索到的样本的状态或专家示范数据中的状态，而非对状态-动作对进行判别，于是给出以下损失函数：

$$\text{Loss} = \mathbb{E}_{s\sim\mathcal{D}}[\nabla_\omega \log(D_\omega(s))] + \mathbb{E}_{s\sim\mathcal{D}^{\mathrm{e}}}[\nabla_\omega \log(1 - D_\omega(s))] \tag{8.31}$$

其中 $\mathcal{D}$ 是用当前策略探索到的样本集，而 $\mathcal{D}^{\mathrm{e}}$ 是示范数据集。不同的具体算法基于以上有不同的具体形式和修正方式。

举例来说，文献 (Merel et al., 2017) 发展了一个 GAIL 的变体，它只使用部分可观测的状态特征而不使用动作来给人类提供类似人的（Human-Like）运动轨迹，通过 GAN 的结构。它类似于

基于模型的 IfO 中的 RIDM 方法和混合（Hybrid）强化学习方法，也使用了一个强化学习模块和一个模仿学习模块，但是以一种层次化的结构使用的。强化学习模块是一个高阶的（High-Level）控制器，它基于一个低阶的（Low-Level）控制器，这个低阶控制器使用 BC 方法来采集人类的运动特征。状态和动作的轨迹在一个随机性策略 π 和环境的交互中被采集，这对应于 GAN 结构中的生成器（Generator）。状态-动作对随后被转化成特征 z，其中动作可能被除去。根据原文所述，示范数据和采集到的数据被假设在同一个特征空间（Feature Space）下。示范或生成数据由辨别器评估来得到这个数据属于示范数据的概率。辨别器的输出值随后被用作奖励来通过强化学习更新模仿策略，类似于 GAIL 中的式 (8.12)。如果学习多种行为的（Multi-Behavior）策略，那么可以添加一个额外的背景变量（Context Variable）。这个辨别器的损失函数可以写作：

$$\text{Loss} = \mathbb{E}_{z\sim s, s\sim\mathcal{D}}[\nabla_\omega \log(D_\omega(z, c))] + \mathbb{E}_{z^\text{e}\sim s^\text{e}, s^\text{e}\sim\mathcal{D}^\text{e}}[\nabla_\omega \log(1 - D_\omega(z^\text{e}, c^\text{e}))] \tag{8.32}$$

其中 z, z^e 是 s, s^e 的编码特征，而 s, s^e 分别来自强化学习探索得到的数据集 $\mathcal{D}$ 和专家示范数据集 $\mathcal{D}^\text{e}$，而 c, c^e 是表示不同行为的背景变量。

由文献 (Henderson et al., 2018) 提出的 OptionGAN 使用分层强化学习中的选项框架（Options Framework），从而基于只使用可观测状态的生成对抗式结构（Generative Adversarial Architecture）来恢复奖励-策略的联合选项（Joint Reward-Policy Options），如图 8.7 所示。经过策略分解（Decomposition），它不仅可以在简单的任务上学习得好，而且对于复杂的连续控制任务也能学得一个基于选项的一般策略（A General Policy over Options）。

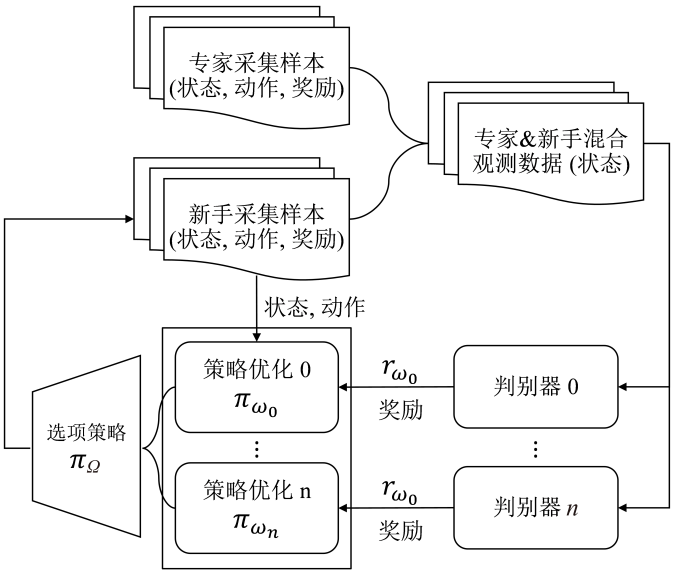

图 8.7 OptionGAN 的结构，改编自文献 (Henderson et al., 2018)

图 8.7 中 IfO 方法的一个潜在问题是，即使所学的最优策略能够生成一个与专家策略非常类似的状态分布，不意味着对于模仿策略和专家策略的所有状态，它相应的动作都是完全相同的。由文献 (Torabi et al., 2019d) 提出的一个简单例子是，在一个环状的（Ring-Like）环境中，两个智能体以相同的速度但是不同的方向移动（即一个为顺时针、另一个为逆时针），这将导致相同的状态分布，即使它们的行为与彼此相反（即在给定状态下有不同的动作分布）。

一种解决上述动作分布不匹配问题的方法是，给辨别器输入一系列状态而非单个状态，如文献 (Torabi et al., 2018b, 2019b) 所提出的一个相似算法，它只是将辨别器的输入改为状态转移 $\{(S_t, S_{t+1})\}$ 而非单个状态。这时辨别器的损失函数将变为

$$\mathbb{E}_{\mathcal{D}}[\nabla_\omega \log(D_\omega(S_t, S_{t+1}))] + \mathbb{E}_{\mathcal{D}^e}[\nabla_\omega \log(1 - D_\omega(S_t, S_{t+1}))] \tag{8.33}$$

其中状态序列在实践中也可以选择长度大于 2 的。

另一个由文献 (Torabi et al., 2019c) 提出的工作使用本体感觉（Proprioceptive）特征而非观察到的图像作为策略的状态输入，来在强化学习智能体中构建类似于人和动物的基于本体感觉控制（Proprioception-Based Control）的模型。由于本体感觉特征的低维性质，策略可以用一个简单的多层感知机（Multi-Layer Perceptron，MLP），而非一个卷积神经网络（Convolutional Neural Network，CNN）来表示，而辨别器仍旧以来自探索样本和专家示范的序列观测图像为输入，如图 8.8 所示。低维本体感觉特征也使得整个学习过程更高效。

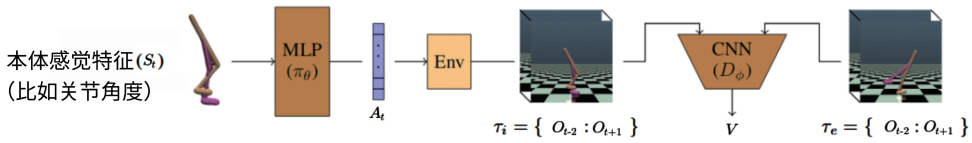

图 8.8　使用本体感觉状态，只从观察量进行模仿学习。图片改编自文献 (Torabi et al., 2019c)

如第 7 章中所提及的，较低的样本效率（Sample Efficiency）是当前强化学习算法的一个主要问题，这在模仿学习和 IfO 领域中也存在。由于生成对抗的方法属于 IRL 的范畴，上面介绍的这些方法可能有 8.3 节所提到的较大计算消耗的问题。这些对抗式模仿学习算法通常需要大量的示范样本和迭代学习来成功学会模仿示范者的行为。为了进一步提高上述方法的样本效率，文献 (Torabi et al., 2019a) 提出在策略学习中使用线性二次型调节器（Linear Quadratic Regulators，LQR）(Tassa et al., 2012) 作为一种基于轨迹的（Trajectory-Centric）强化学习方法，而这有可能使得真实机器人的模仿学习成为现实。

上述方法主要基于示范数据空间和模仿者学习的空间有一致性的基本假设。然而，当这两个空间不匹配时，比如在三维空间中由于提供观察量的摄像机位置不同而造成的视角变化，一般的模仿学习方法可能会有性能上的下降。示范和模仿的空间差异可能在动作空间，也可能在状态空

间。对于动作空间的差异，文献 (Zołna et al., 2018) 提出使用成对有任意时间间隔（Time Gaps）的状态替代连续不断的状态（Consecutive States）来作为辨别器的输入，这可以看作是用噪声进行数据集增强（Dataset Augmentation），从而有更鲁棒和通用的表现。在他们的实验中，这个方法确实展示出了在模仿者策略与示范数据有不同动作空间的情况下的性能提升。而对状态空间的差异，比如上面提及的视角变化，文献 (Stadie et al., 2017) 提出使用一个分类器（Classifier）来区分来自不同视角的样本，将辨别器最初的几个神经网络层的输出作为分类器的输入。这个方法使用了域混淆（Domain Confusion）的想法来学习域无关的（Domain Agnostic）特征，其中域在这种情况下指不同的视角。在辨别器的最初神经网络层（作为一个特征提取器）混淆被最大化，但对分类器混淆被最小化，因而这也利用了对抗式训练的框架。在训练之后，提取器（辨别器的最初几个神经网络层）所学特征对视角变化有了不变性。

这领域也有一些其他方法。Sun et al. (2019b) 提出 IfO 中第一个可证明高效的算法，叫作正向对抗式模仿学习（Forward Adversarial Imitation Learning，FAIL），它可以用跟所有相关参数有多项式（Polynomial）数量关系的样本量来学习一个近最优的策略，而不依赖于单一观察量（Unique Observations）的数量。FAIL 中的极小化极大（Minimax）方法学习一个策略，这个策略能够根据之前时间步的策略匹配下一个状态的概率分布。近来，一个称为动作指导性对抗式模仿学习（Action-Guided Adversarial Imitation Learning，AGAIL）由文献 (Sun et al., 2019a) 提出，它试图利用示范中的状态和不完整动作信息，因而是 IfO 跟传统 IL 的一个结合方法。辨别器被用来区分单个状态，类似于之前介绍的文献 (Merel et al., 2017) 的方法。此外，它还用一个指导性 Q 网络（Guided Q-Network）来以一种监督学习的方式学习 $p(a^{\mathrm{e}}|a \sim \pi(s^{\mathrm{e}}))$ 的真实后验（Posterior），其中 $(s^{\mathrm{e}}, a^{\mathrm{e}})$ 表示专家示范样本。

奖励函数工程方法：生成对抗方法自然地提供了可以让模仿策略以强化学习方式训练的奖励信号。除了生成对抗方法，也有像奖励函数工程（Reward Engineering）的方法来解决无模型 IfO。事实上，之前小节中提到的基于模型的 IfO 中的 RIDM 方法是一种奖励函数工程方法。这里的奖励函数工程指需要人为设计奖励函数来以强化学习的方式从专家示范中学习模仿策略的方法。奖励函数工程将模仿学习的监督学习方式转化为一个强化学习问题，通过给强化学习智能体构建一个奖励函数。需要注意的是，人为设计的奖励函数不需要是真实的产生专家策略的奖励函数，而更像是一个基于示范数据集或任务先验知识（Prior Knowledge）的估计。比如，文献 (Kimura et al., 2018) 提出使用预测的下一个状态和示范者的下一个真实状态间的欧氏距离（Euclidean Distance）作为奖励函数，随后根据这个奖励函数可以用一般强化学习的方式来学习一个模仿策略。

另一种奖励函数工程方法称为时间对比网络（Time-Contrastive Networks，TCN），由文献 (Sermanet et al., 2018) 提出，如图 8.9 所示。为了解决前面提及的多视角问题，而这个问题对于学习人的行为很重要，TCN 方法通过学习一个视角不变的表示来获取物体之间的关系，它通过 TCN 网络处理从不同视角获得的几个（原文中是两个）同步的相机视野。对抗式训练因此可以用在嵌入式表示空间（Embedded Representation Space），而非原来的状态空间（如其他方法中所用的）。这个表示是通过一个三重（Triplet）损失函数和 TCN 嵌入网络（Embedding Network）来学到的。这

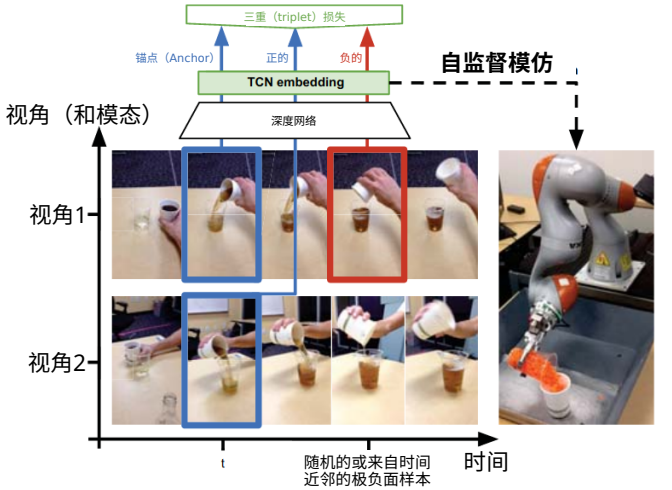

图 8.9　使用三重损失函数的时间对比网络（TCN）的学习框架，它以一种自监督式的学习，用于只从观察量进行的模仿学习（IfO）中的观察量嵌入（Observation Embedding）。图片来自文献 (Sermanet et al., 2018)（见彩插）

个三重损失被设定为在视频示范数据中驱散（Disperse）连续帧的短时近邻（Temporal Neighbors），而这些近邻满足有相似的视觉特征但是不同的实际动态状态，同时吸引（Attract）那些不同视角下同时发生的帧，这些帧在嵌入空间中有相同的动态状态。因此，模仿策略能够用无标签的人类示范视频以自监督（Self-Supervised Learning）的方式进行学习。类似文献 (Kimura et al., 2018) 中描述的工作，奖励函数定义为同一时间步下示范状态和智能体实际状态的欧氏距离，但它是在嵌入空间而不是状态空间。TCN 被设计成用于单帧状态嵌入（Single Frame State Embedding）。Dwibedi et al. (2018) 扩展了 TCN 的工作，使其可以对多个帧进行嵌入，从而更好地表示轨迹的模式（Patterns in Trajectory）。文献 (Aytar et al., 2018) 也采用了一个相似的方法，从 YouTube 视频帧中基于示范数据来学习嵌入函数，从而解决难以探索的任务，比如 Montezuma's Revenge 和 Pitfall，这些任务在第 7 章的探索挑战中有所提及。它可以解决较小的变化，比如视频的失真和颜色变化。模仿者嵌入状态和示范者嵌入状态的距离测度（Measurement）也被用作奖励函数。

如之前所介绍的，可以用一个分类器来区分来自不同视角的观察量。文献 (Goo et al., 2019) 提出，分类器也可以用于预测示范数据中帧的顺序，通过一种打乱学习（Shuffle-and-Learn）的训练方式 (Misra et al., 2016)。奖励函数可以根据所学的分类器来定义，并用于训练模拟者策略。同时，在之前生成对抗方法的描述中，状态空间的不匹配，比如由视角不同造成，可以通过不变的特征表示（Invariant Feature Representation）来解决。然而，它也可以用一个定义为示范状态和模仿者状态在表征空间下的欧氏距离作为奖励函数，来训练模仿策略，而非使用辨别器并以示范状态和模仿者状态作为输入时的输出值为奖励，这在文献 (Gupta et al., 2017; Liu et al., 2018) 中都有提到。

8.4.3 从观察量模仿学习的挑战

根据以上所提及的 IfO 中的方法，智能体能够只从观察到的状态来学习策略，但是仍旧存在文献 (Torabi et al., 2019d) 所提到的问题。

- **具象不匹配**（Embodiment Mismatch）：具象不匹配通常用来描述外观（对于基于视觉的控制）、动态过程和其他特征在模仿者域和示范者域间的差异。一个典型的例子是让机械臂模仿人的手臂执行动作。由于控制动力学和观察智能体的视角会有显著的差别，所以这样的模仿学习过程可能很难实现。即使是确认机器人和人的手臂是否在同一个状态都会有困难。一个解决这个问题的方法是学习隐藏对应关系（Correspondences）或潜在表示（Latent Representations），这个关系或表示能够对两个域的差异产生不变性，然后基于这个关系或者在所学的表征空间内进行模仿学习。一个用来解决这个问题的 IfO 方法 (Gupta et al., 2017) 用自动编码器（Autoencoder）来学习不同的具象之间的对应关系以一种监督学习的方式。自动编码器被训练使得编码后的表示对具象特征有不变性。另一个方法 (Sermanet et al., 2018) 使用少量人类监督和无监督的学习方式来学习对应关系。
- **视角差异**：在上面提到的几个方法中，比如 TCN 和一些其他基于模型的 IfO 方法，对于基于视觉的控制，由于示范数据由相机采集的图像或视频给出，视角的差异可能导致模仿策略表现显著下降。通常来讲，需要有一个在对视角不变的（Viewpoint Invariant）空间中表征状态的编码模型（Encoding Model），如文献 (Sieb et al., 2019) 中提到的，或者一个能够根据某一帧预测具体视角的分类器，如文献 (Stadie et al., 2017) 所提到。另一种试图解决这个问题的 IfO 方法是去学习一个背景转化（Context Translation）模型，从而根据一个观察量预测它在目标背景中的表示 (Liu et al., 2018)。这个转化是通过包含源背景和目标背景下的图像数据来学习的，而任务是将源背景转化到目标背景。这需要收集源背景和目标背景下相似的样本来实现。

8.5 概率性方法

除了使用神经网络的参数化方法，许多概率推理方法也可以被用于模仿学习，尤其是在机器人运动领域，这些方法包括高斯混合回归（Gaussian Mixture Regression，GMR）(Calinon, 2016)、动态运动基元（Dynamic Movement Primitives，DMP）(Pastor et al., 2009)、概率性运动基元（Probabilistic Movement Primitives，ProMP）(Paraschos et al., 2013)、核运动基元（Kernelized Movement Primitives，KMP）(Huang et al., 2019)、高斯过程回归（Gaussian Process Regression，GPR）(Schneider et al., 2010)、基于 GMR 的高斯过程 (Jaquier et al., 2019) 等。由于本书主要是介绍使用深度神经网络参数化的深度强化学习，所以我们将仅简单介绍这些概率性方法，而将概率性方法和深度强化学习结合起来本身就不是平庸的（Non-Trivial），不像在本章中介绍的其他方法那样直接。

然而，即使将概率性方法用于深度强化学习任务可能是不容易实现的，概率性方法由于其一

些优点还是很值得研究的，具体表现讨论如下。

不同于深度神经网络给出确定性的预测结果，由 GRM、ProMP 和 KMP 计算得到预测分布的协方差矩阵（Covariance Matrices）编码了预测轨迹的变化性。而这在使用所学模型来预测或做决策且其决策的置信度同样重要时会很有用，比如在机器人操作或车辆驾驶的情形中为了保证安全，每个指令的可行性和风险都需要以概率模型的方式来分析。除此之外，概率性方法根据概率论的支持通常有解析解，这与基于深度神经网络的“黑盒”优化过程不同。而这也使得概率性方法能够在数据量较小时用较短时间求解。此外，像基于 GMR 的高斯过程类的概率性方法对未见过的输入数据点有快速的适应能力，这在下面小节中将会讨论。对于模仿学习中的概率性方法，数据集被默认为是以有标签数据类型来提供的，即输入和输出的配对，对于一般强化学习，它通常是状态-动作对 $\{(s_i, a_i)|i = 0, \cdots, N\}$，而对按时间排列的示范数据，它可以是时间-状态对 $\{(t, S_t)|t = 0, \cdots, N\}$ (Jaquier et al., 2019)。

基于高斯混合回归（GMR）的高斯回归（GPR）是一种结合了高斯混合回归和高斯过程回归的方法。GMR 利用了高斯条件定理（Gaussian Conditioning Theorem）来估计给定输入数据的输出分布。高斯混合模型（Gaussian Mixture Model，GMM）通过期望最大化算法（Expectation Maximization，EM）来拟合输入输出数据点的联合分布（Joint Distribution）。给定观察输入，基于条件的（Conditional）均值和方差可以有封闭解，其输出结果因而可以通过基于条件的期望的线性组合来得到，使用测试数据点作为输入。GP 如同深度神经网络一样，是针对学习确定性（Deterministic）输入-输出关系问题的方法，它基于可能的目标函数的高斯先验（Prior）来计算。基于 GMR 的 GP（GMR-Based GP）是种结合的方法，它的 GP 先验均值等于 GMR 模型基于条件的均值，而 GP 的核（Kernel）是相应 GMM 各个组分单独的核的叠加。这种结合使得基于 GMR 的 GP 方法有 GP 通过均值和核来编码多种先验置信（Prior Beliefs）的能力，并且允许 GMR 估计的多样化信息被封装到 GP 的不确定性（Uncertainty）估计中。当给出新的未见过的输入观察数据点时，基于 GMR 的 GP 能够快速适应它们并给出合理预测输出，如图 8.10 所示。对于一个

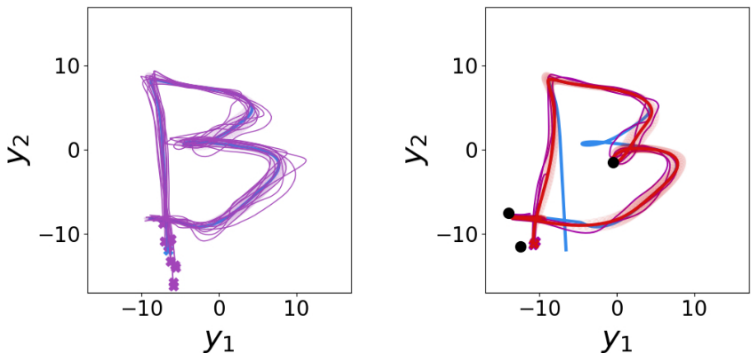

图 8.10　模仿学习中基于 GMR 的 GP 方法。左边图中，先验均值为蓝色，采样轨迹为紫色。右边图中，先验均值（与左图相同）为蓝色，采样轨迹为粉色，预测轨迹为红色，有三个黑色的点为新观察量。图片来自文献 (Jaquier et al., 2019)（见彩插）

二维轨迹的估计过程，图 8.10 中的左边用紫色线展示了所给的样本，而蓝色线展示了先验均值。右边的图是基于 GMR 的 GP 方法，其中有 3 个新的观察数据点被标为黑色，粉色线展示了采样轨迹，而红色线是预测轨迹。这个方法经证实对使用示范数据进行学习并快速适应到新的数据点的情况有很好的表现，而这可以用于操作机器人基于示范规避障碍物。

8.6 模仿学习作为强化学习的初始化

使用模仿学习的基本设定是在不使用任何强化信号而只有示范数据的情况下学习一个策略，这意味着通过模仿学习所学策略是来自示范数据的最终策略。然而，在实际中，来自模仿学习的策略通常没有足够的泛化能力，尤其是对于未见过的情况。因此，我们可以在强化学习的过程中使用模仿学习，以此来提高强化学习的效率。举例来说，使用示范数据的预训练策略可以用来初始化强化学习的策略。关于这些方法的细节将在随后讨论。因此，我们并不需要模仿学习给出的策略是最优的，而是通过一个相对简单的学习过程得到一个足够好的策略，比如使用监督学习的模仿学习方法。所以，我们在下面只选择一些简单直接的方法来作为后续强化学习过程的初始化方法。模仿学习中更精致的方法毫无疑问会成为更好的初始化策略，但是也会相应带来如较长的预训练时间等缺点。

总体来说，通过监督学习方式模仿示范数据而学到的策略，可以使用包括 BC、DAgger、Variational Dropout 等方法，它们被看作是对强化学习策略较好的初始化，具体地，通过下面小节中描述的策略替换（Policy Replacement）或者残差策略学习（Residual Policy Learning）方法。

除了用策略替换来初始化强化学习（模拟学习策略在强化学习初始时替换其策略），残差策略学习 (Johannink et al., 2019; Silver et al., 2018) 是另一种实现初始化的方法。比如对于机器人控制任务，它通常基于一个较好但是不完美的控制器，并以这个初始控制器为基础学习一个残差策略。对于现实世界的机器人控制，初始控制器可以是一个模拟器中预训练的策略；对于模拟的机器人控制，初始控制器可以用监督学习的方式基于专家轨迹预训练得到，如 8.2 节中的方法。

残差策略学习中的动作遵循结合式策略，即由初始策略（Initial Policy）π_{ini} 和残差策略 π_{res} 求和得到：

$$a = \pi_{\text{ini}}(s) + \pi_{\text{res}}(s). \tag{8.34}$$

通过这种方式，残差策略学习能够尽可能地保持初始策略的表现。

例子：使用 DDPG 的残差策略学习

这里我们使用深度确定性策略梯度（DDPG）算法来实现基于示范的残差策略学习。根据残差策略学习方法，DDPG 中的行动者（Actor）策略将包含两部分：一个是预训练得到的初始策略，在初始化后将被固定；另一个是后面学习过程中将训练的残差策略。初始策略通过模仿学习根据

示范数据训练得到。这个预训练的初始策略只用于 DDPG 的行动者部分。基于 DDPG 算法使用残差策略学习的过程如下：

（1）以残差学习的方式初始化 DDPG 中的所有网络，包括对批判者（Critic）、目标批判者（Target Critic）的一般初始化，以及对残差策略（Residual Policy）和目标残差策略（Target Residual Policy）的最后网络层（Final Layers）进行零值初始化，还有将通过模仿学习得到的策略作为初始策略和目标初始策略（Target Initial Policy），一共是六个网络。这时固定住初始策略和目标策略，开始训练过程。

（2）让智能体与环境交互，动作值是初始化策略和残差策略的和值：$a = a_{\text{ini}} + a_{\text{res}}$；将样本以 $(s, a_{\text{res}}, s', r, \text{done})$ 的形式存储。

（3）从经验回放缓存中采样 $(s, a_{\text{res}}, s', r, \text{done})$，有

$$Q_{\text{target}}(s, a_{\text{res}}) = r + \gamma Q^{\text{T}}(s, \pi_{\text{res}}^{\text{T}}(s)) \tag{8.35}$$

其中 $Q^{\text{T}}, \pi_{\text{res}}^{\text{T}}$ 分别表示目标批判者和目标残差策略。批判损失函数是 $\text{MSE}(Q_{\text{target}}(s, a_{\text{res}}), Q(s, a_{\text{res}}))$。行动者的目标是最大化状态 s 和动作 a_{res} 的动作价值函数，如下：

$$\max_{\theta} Q(s, a_{\text{res}}) = \max_{\theta} Q(s, \pi_{\text{res}}(s|\theta)) \tag{8.36}$$

这可以通过确定性策略梯度（Deterministic Policy Gradient）来优化。

（4）重复上面的第（2）（3）步，直到策略收敛到接近最优。

对比一般的 DDPG 算法，使用残差策略学习的不同只是对残差策略的动作 a_{res}，而非智能体的整个动作 a 来学习动作价值函数和策略。

8.7　强化学习中利用示范数据的其他方法

8.7.1　将示范数据导入经验回放缓存

基于示范的深度 Q-Learning（Deep Q-Learning from Demonstrations，DQfD）(Hester et al., 2018) 通过直接将专家轨迹导入离线（Off-Policy）强化学习的记忆缓存（Memory Buffer）中来利用示范数据，而非预训练一个策略来初始化强化学习策略。它使用 DQN 来解决只有离散动作空间的应用。DQfD 使用一个由所有专家示范初始化的经验回放缓存（Experience Replay Buffer），并不断向其添加采集到的新样本。DQfD 使用优先经验回放（Prioritized Experience Replay）(Schaul et al., 2015) 来从回放缓存中采样训练批，且它使用一个监督式折页损失函数（Hinge Loss）来模仿示范数据和一个一般的 TD 损失函数的结合来训练策略。

基于示范的深度确定性策略梯度（Deep Deterministic Policy Gradient from Demonstrations，DDPGfD）(Večerík et al., 2017) 是一种与上面 DQfD 类似的方法，但是使用 DDPG 来处理连续动作

空间的应用。DDPGfD 通过直接将专家策略输入离线强化学习（即 DDPG）的缓存来利用示范数据，从而通过示范和探索数据一同训练策略。优先经验回放被用来平衡两种训练数据。DDPGfD 可以用于强化学习中的简单、易解决的任务，而对从稀疏奖励学习等较难任务需要在训练中进行更积极的探索。

文献 (Nair et al., 2018) 提出一个基于 DQfD 和 DDPGfD 的方法，对较难的任务有更好的学习效率，这些任务需要基于示范数据进一步探索去解决。它的策略损失函数是策略梯度损失（Policy Gradient Loss）和行为克隆损失（Behavioral Cloning Loss）的结合，其梯度如下：

$$\lambda_1 \nabla_\theta J - \lambda_2 \nabla_\theta L_{\text{BC}} \tag{8.37}$$

其中 J 是一般的强化学习目标（最大化的），而 L_{BC}（最小化的）是本章开始时定义的行为克隆损失。

此外，这个方法也使用了 Q-Filter 技术，它要求行为克隆损失函数只用于部分状态，在这些状态下所学的批判者 $Q(s,a)$ 判定示范者动作比行动者动作更好：

$$L_{\text{BC}} = \sum_{i=1}^{N_D} \|\pi(s_i|\theta_\pi) - a_i\|^2 \mathbb{1}_{Q(s_i,a_i)>Q(s_i,\pi(s_i))} \tag{8.38}$$

其中 N_D 是示范数据集中样本的数量，而 (s_i, a_i) 是从示范数据集中采样得到的。这保证了策略能够探索到更好的动作，而不是被示范数据所限制。

以同样的方式，QT-Opt (Kalashnikov et al., 2018) 和分位数 QT-Opt（Quantile QT-Opt）(Bodnar et al., 2019) 算法也使用在线缓存和离线示范缓存混合的方式来实现离线学习，通过一种无行动者（Actor-Free）的交叉熵方法和 DQN，可以在现实世界中基于图像的机器人学习任务上达到当时最先进的（State-of-the-Art）表现。

8.7.2 标准化 Actor-Critic

标准化 Actor-Critic（Normalized Actor-Critic，NAC）(Gao et al., 2018) 是另一个利用示范数据来进行高效强化学习的方法，它先预训练一个策略作为改进强化学习过程的初始化。NAC 与其他方法的差异是它在使用示范数据预训练初始化策略和改进强化学习的过程中使用完全相同的目标函数，这使得 NAC 对包含次优样本的示范数据也表现得很鲁棒。

另一方面，NAC 方法类似于 DDPGfD 和 DQfD 方法，但是它依次使用示范数据和交互样本进行训练，而不是同时使用这两类样本数据。

8.7.3　用示范数据进行奖励塑形

用示范数据进行奖励塑形（Reward Shaping with Demonstrations）(Brys et al., 2015) 是一个专注于初始化强化学习中价值函数而非动作策略的方法。它给智能体提供了一个中间的奖励来丰富稀疏奖励信号：

$$R_F(s,a,s') = R(s,a,s') + F^D(s,a,s') \tag{8.39}$$

其中基于示范数据 D 的塑形奖励 F^D 通过势函数 ϕ 来定义并保证其收敛性，其形式如下：

$$F^D(s,a,s',a') = \gamma\phi^D(s',a') - \phi^D(s,a) \tag{8.40}$$

而 ϕ^D 定义为

$$\phi^D(s,a) = \max_{(s^d,a)} \mathrm{e}^{-\frac{1}{2}(s-s^d)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(s-s^d)} \tag{8.41}$$

它被用来最大化最接近示范状态 s^d 的状态 s 的势值。优化后的势函数被用来初始化强化学习中的动作价值函数 Q：

$$Q_0(s,a) = \phi^D(s,a) \tag{8.42}$$

奖励塑形的直观理解是使探索到的样本倾向于那些等于或接近示范数据的状态-动作对，从而加速强化学习的训练过程。奖励塑形提供了一种在强化学习过程中初始化价值估计函数的较好方式。

其他方法像无监督感知奖励（Unsupervised Perceptual Rewards）(Sermanet et al., 2016) 也用于通过示范数据学习一个密集且平滑的奖励函数，使用的是一个预训练的深度学习模型得出的特征。

8.8　总结

由于第 7 章中提到的强化学习低学习效率的挑战，我们介绍模仿学习来作为一种可能的解决方案，它需要使用专家示范。本章整体可以总结为几个主要类别。8.2 节中介绍的行为克隆方法是以监督学习方式进行模仿学习的最直接方法，它可以进一步与强化学习结合，比如 8.6 节中介绍的将其作为强化学习的初始化。一个更先进的结合模仿学习和强化学习的方式是通过 IRL 来显式或隐式地从示范中恢复奖励函数，如 8.3 节所介绍的。像 MaxEnt IRL 方法可以显式地学习奖励函数，但是可能有较大计算消耗。其他的生成对抗式方法，如 GAIL、GAN-GCL、AIRL 则能更高效地学习奖励函数和策略。另一个问题是如果示范数据集中的动作是缺失的，比如只从视频中学习，那么怎样合理地进行模仿学习？这实际是 IfO 的研究范畴，如 8.4 节所介绍。由于 IfO

问题是从另一个角度来看模仿学习的，之前介绍的方法像 BC、IRL 同样可以经过适当修改用于 IfO。IfO 中的方法基本可以总结为基于模型和无模型两类。基于模型的方法从样本中学习动态模型，而且它可以通过模型中状态-动作关系从只有观察量的示范数据中恢复动作，以显式或者隐式的方法。随后，如果动作被显式地恢复了，就可以使用常规的模仿学习方法。像 RIDM、BCO、ILPO 等方法属于这个基于模型的 IfO 范畴。对于 IfO 中的无模型方法，奖励函数工程或者生成对抗式方法可以用来提供奖励函数从而进行强化学习。像 OptionGAN、FAIL、AGAIL 等方法属于生成对抗式 IfO，而 TCN 和一些其他方法属于 IfO 的奖励函数工程一类。这里对 IfO 的两个类别实际对一般的模仿学习也适用，比如 GAIL 是一种生成对抗式方法，而最近提出的对比正向动态（Contrastive Forward Dynamics，CFD）(Jeong et al., 2019) 是模仿学习的一种从观察量和动作示范中学习的奖励函数工程方法。概率性方法包括 GMR、GPR 和基于 GMR 的 GP 方法作为一般的模仿学习方法而在本章中有所介绍，它们对于相对低维的情况有较高的学习效率，如 8.5 节所讨论的。最终，一些其他方法像 DDPGfD 和 DQfD 将示范数据直接导入离线强化学习的回放缓存中，等等，都在 8.7 节中介绍。模仿学习作为一种解决学习问题的高效方式，可以与强化学习有机结合，相关研究领域依然十分活跃，

参考文献

ABBEEL P, NG A Y, 2004. Apprenticeship learning via inverse reinforcement learning[C]//Proceedings of the twenty-first international conference on Machine learning. ACM: 1.

AYTAR Y, PFAFF T, BUDDEN D, et al., 2018. Playing hard exploration games by watching youtube[C]// Advances in Neural Information Processing Systems. 2930-2941.

BLAU T, OTT L, RAMOS F, 2018. Improving reinforcement learning pre-training with variational dropout[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE: 4115-4122.

BODNAR C, LI A, HAUSMAN K, et al., 2019. Quantile QT-Opt for risk-aware vision-based robotic grasping[J]. arXiv preprint arXiv:1910.02787.

BRYS T, HARUTYUNYAN A, SUAY H B, et al., 2015. Reinforcement learning from demonstration through shaping[C]//Twenty-Fourth International Joint Conference on Artificial Intelligence.

CALINON S, 2016. A tutorial on task-parameterized movement learning and retrieval[J]. Intelligent Service Robotics, 9(1): 1-29.

DUAN Y, ANDRYCHOWICZ M, STADIE B, et al., 2017. One-shot imitation learning[C]//Advances in Neural Information Processing Systems. 1087-1098.

DWIBEDI D, TOMPSON J, LYNCH C, et al., 2018. Learning actionable representations from visual observations[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE: 1577-1584.

EDWARDS A D, SAHNI H, SCHROECKER Y, et al., 2018. Imitating latent policies from observation[J]. arXiv preprint arXiv:1805.07914.

EYSENBACH B, GUPTA A, IBARZ J, et al., 2018. Diversity is all you need: Learning skills without a reward function[J]. arXiv preprint arXiv:1802.06070.

FINN C, CHRISTIANO P, ABBEEL P, et al., 2016a. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models[J]. arXiv preprint arXiv:1611.03852.

FINN C, LEVINE S, ABBEEL P, 2016b. Guided cost learning: Deep inverse optimal control via policy optimization[C]//International Conference on Machine Learning. 49-58.

FU J, LUO K, LEVINE S, 2017. Learning robust rewards with adversarial inverse reinforcement learning[J]. arXiv preprint arXiv:1710.11248.

GAO Y, LIN J, YU F, et al., 2018. Reinforcement learning from imperfect demonstrations[J]. arXiv preprint arXiv:1802.05313.

GOO W, NIEKUM S, 2019. One-shot learning of multi-step tasks from observation via activity localization in auxiliary video[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE: 7755-7761.

GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al., 2014. Generative Adversarial Nets[C]// Proceedings of the Neural Information Processing Systems (Advances in Neural Information Processing Systems) Conference.

GUO X, CHANG S, YU M, et al., 2019. Hybrid reinforcement learning with expert state sequences[J]. arXiv preprint arXiv:1903.04110.

GUPTA A, DEVIN C, LIU Y, et al., 2017. Learning invariant feature spaces to transfer skills with reinforcement learning[J]. arXiv preprint arXiv:1703.02949.

HANNA J P, STONE P, 2017. Grounded action transformation for robot learning in simulation[C]// Thirty-First AAAI Conference on Artificial Intelligence.

HAUSMAN K, CHEBOTAR Y, SCHAAL S, et al., 2017. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets[C]//Advances in Neural Information Processing Systems. 1235-1245.

HENDERSON P, CHANG W D, BACON P L, et al., 2018. OptionGAN: Learning joint reward-policy options using generative adversarial inverse reinforcement learning[C]//Thirty-Second AAAI Conference on Artificial Intelligence.

HESTER T, VECERIK M, PIETQUIN O, et al., 2018. Deep Q-learning from demonstrations[C]// Thirty-Second AAAI Conference on Artificial Intelligence.

HO J, ERMON S, 2016. Generative adversarial imitation learning[C]//Advances in Neural Information Processing Systems. 4565-4573.

HUANG Y, ROZO L, SILVÉRIO J, et al., 2019. Kernelized movement primitives[J]. The International Journal of Robotics Research, 38(7): 833-852.

JAQUIER N, GINSBOURGER D, CALINON S, 2019. Learning from demonstration with model-based gaussian process[J]. arXiv preprint arXiv:1910.05005.

JEONG R, AYTAR Y, KHOSID D, et al., 2019. Self-supervised sim-to-real adaptation for visual robotic manipulation[J]. arXiv preprint arXiv:1910.09470.

JOHANNINK T, BAHL S, NAIR A, et al., 2019. Residual reinforcement learning for robot control[C]// 2019 International Conference on Robotics and Automation (ICRA). IEEE: 6023-6029.

KALASHNIKOV D, IRPAN A, PASTOR P, et al., 2018. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation[J]. arXiv preprint arXiv:1806.10293.

KIMURA D, CHAUDHURY S, TACHIBANA R, et al., 2018. Internal model from observations for reward shaping[J]. arXiv preprint arXiv:1806.01267.

LIU Y, GUPTA A, ABBEEL P, et al., 2018. Imitation from observation: Learning to imitate behaviors from raw video via context translation[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE: 1118-1125.

MEREL J, TASSA Y, SRINIVASAN S, et al., 2017. Learning human behaviors from motion capture by adversarial imitation[J]. arXiv preprint arXiv:1707.02201.

MISRA I, ZITNICK C L, HEBERT M, 2016. Shuffle and learn: unsupervised learning using temporal order verification[C]//European Conference on Computer Vision. Springer: 527-544.

MOLCHANOV D, ASHUKHA A, VETROV D, 2017. Variational dropout sparsifies deep neural networks[C]//Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org: 2498-2507.

NAIR A, CHEN D, AGRAWAL P, et al., 2017. Combining self-supervised learning and imitation for vision-based rope manipulation[C]//2017 IEEE International Conference on Robotics and Automation (ICRA). IEEE: 2146-2153.

NAIR A, MCGREW B, ANDRYCHOWICZ M, et al., 2018. Overcoming exploration in reinforcement learning with demonstrations[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE: 6292-6299.

NG A Y, HARADA D, RUSSELL S, 1999. Policy invariance under reward transformations: Theory and application to reward shaping[C]//Proceedings of the International Conference on Machine Learning (ICML): volume 99. 278-287.

NG A Y, RUSSELL S J, et al., 2000. Algorithms for inverse reinforcement learning.[C]//Proceedings of the International Conference on Machine Learning (ICML): volume 1. 2.

PARASCHOS A, DANIEL C, PETERS J R, et al., 2013. Probabilistic movement primitives[C]//Advances in Neural Information Processing Systems. 2616-2624.

PASTOR P, HOFFMANN H, ASFOUR T, et al., 2009. Learning and generalization of motor skills by learning from demonstration[C]//2009 IEEE International Conference on Robotics and Automation. IEEE: 763-768.

PATHAK D, MAHMOUDIEH P, LUO G, et al., 2018. Zero-shot visual imitation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2050-2053.

PAVSE B S, TORABI F, HANNA J P, et al., 2019. Ridm: Reinforced inverse dynamics modeling for learning from a single observed demonstration[J]. arXiv preprint arXiv:1906.07372.

PUTERMAN M L, 2014. Markov decision processes: Discrete stochastic dynamic programming[M]. John Wiley & Sons.

ROSS S, BAGNELL D, 2010. Efficient reductions for imitation learning[C]//Proceedings of the thirteenth international conference on artificial intelligence and statistics. 661-668.

ROSS S, GORDON G, BAGNELL D, 2011. A reduction of imitation learning and structured prediction to no-regret online learning[C]//Proceedings of the fourteenth international conference on artificial intelligence and statistics. 627-635.

RUSSELL S J, 1998. Learning agents for uncertain environments[C]//COLT: volume 98. 101-103.

SCHAUL T, QUAN J, ANTONOGLOU I, et al., 2015. Prioritized experience replay[C]//arXiv preprint arXiv:1511.05952.

SCHNEIDER M, ERTEL W, 2010. Robot learning by demonstration with local gaussian process regression[C]//2010 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE: 255-260.

SERMANET P, XU K, LEVINE S, 2016. Unsupervised perceptual rewards for imitation learning[J]. arXiv preprint arXiv:1612.06699.

SERMANET P, LYNCH C, CHEBOTAR Y, et al., 2018. Time-contrastive networks: Self-supervised learning from video[C]//2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE: 1134-1141.

SIEB M, XIAN Z, HUANG A, et al., 2019. Graph-structured visual imitation[J]. arXiv preprint arXiv:1907.05518.

SILVER T, ALLEN K, TENENBAUM J, et al., 2018. Residual policy learning[J]. arXiv preprint arXiv:1812.06298.

STADIE B C, ABBEEL P, SUTSKEVER I, 2017. Third-person imitation learning[J]. arXiv preprint arXiv:1703.01703.

SUN M, MA X, 2019a. Adversarial imitation learning from incomplete demonstrations[J]. arXiv preprint arXiv:1905.12310.

SUN W, VEMULA A, BOOTS B, et al., 2019b. Provably efficient imitation learning from observation alone[J]. arXiv preprint arXiv:1905.10948.

SYED U, BOWLING M, SCHAPIRE R E, 2008. Apprenticeship learning using linear programming[C]// Proceedings of the 25th international conference on Machine learning. ACM: 1032-1039.

TASSA Y, EREZ T, TODOROV E, 2012. Synthesis and stabilization of complex behaviors through online trajectory optimization[C]//2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE: 4906-4913.

TORABI F, WARNELL G, STONE P, 2018a. Behavioral cloning from observation[J]. arXiv preprint arXiv:1805.01954.

TORABI F, WARNELL G, STONE P, 2018b. Generative adversarial imitation from observation[J]. arXiv preprint arXiv:1807.06158.

TORABI F, GEIGER S, WARNELL G, et al., 2019a. Sample-efficient adversarial imitation learning from observation[J]. arXiv preprint arXiv:1906.07374.

TORABI F, WARNELL G, STONE P, 2019b. Adversarial imitation learning from state-only demonstrations[C]//Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems. International Foundation for Autonomous Agents and Multiagent Systems: 2229-2231.

TORABI F, WARNELL G, STONE P, 2019c. Imitation learning from video by leveraging proprioception[J]. arXiv preprint arXiv:1905.09335.

TORABI F, WARNELL G, STONE P, 2019d. Recent advances in imitation learning from observation[J]. arXiv preprint arXiv:1905.13566.

VEČERÍK M, HESTER T, SCHOLZ J, et al., 2017. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards[J]. arXiv preprint arXiv:1707.08817.

ZIEBART B D, MAAS A L, BAGNELL J A, et al., 2008. Maximum entropy inverse reinforcement learning.[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 8. Chicago, IL, USA: 1433-1438.

ZIEBART B D, BAGNELL J A, DEY A K, 2010. Modeling interaction via the principle of maximum causal entropy[J].

ZOŁNA K, ROSTAMZADEH N, BENGIO Y, et al., 2018. Reinforced imitation learning from observations[J].