Inverse Reinforcement Learning

逆强化学习（Inverse Reinforcement Learning, IRL）作为一种典型的模仿学习方法，顾名思义，逆强化学习的学习过程与正常的强化学习利用奖励函数学习策略相反，不利用现有的奖励函数，而是试图学出一个奖励函数，并以之指导基于奖励函数的强化学习过程。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的方法交替地在两个过程中交替：一个阶段是使用示范数据来推断一个隐藏的奖励（Reward）或代价（ Cost）函数，另一个阶段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的基本准则是：IRL选择奖励函数 R 来优化策略，并且使得任何不同于 π_E （ $a_E \sim \pi_E$ ）的动作决策（ $a \in A \backslash a_{\mathrm{E}}$ ）,其中尽可能产生更大损失。

形式化上，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^* :

$$R^*=\arg \max _R \sum_{s \in \mathcal{S}}\left(Q^\pi\left(s, a_{\mathrm{E}}\right)-\max _{a \in A \backslash a_{\mathrm{E}}} Q^\pi(s, a)\right)$$

其中 $a_E=\pi_E(s)$ 或 $a_E \sim \pi_E(\cdot|s)$ 是专家最优动作。基于此类形式化优化的学徒学习作为最早期的逆强化学习方法之一，在2004年由Andrew Y.Ng与Pieter Abbeel提出。该算法的核心思想是学习一个能够使得专家策略下的轨迹的期望回报远高于非专家策略的奖励函数，从而达到无监督学习奖励函数的目的。在这样的优化目标下，习得的奖励函数会使得专家和非专家的差距不断增大，因此这种方法也叫做基于最大边际规划的逆强化学习方法 (Maximum Margin Planning Inverse Reinforcement Learning, MMPIRL)。但存在以下挑战性问题：

1. 无论是对奖励函数做线性组合或凸组合假设，还是对策略的最大熵约束，这类显式的规则给逆强化学习的通用性带有一定的限制；
2. 在估计出来的奖励函数下，使用强化学习交互环境来优化策略，这从时间和安全性的角度带来较大代价，同时要求迭代优化奖励函数的内循环中解决一个MDP的问题，将带来极大的计算消耗。

这也是2016年提出生成式对抗模仿学习提出的动机。

对抗式生成模仿学习简介

生成式对抗模仿学习(Generative Adversarial Imitation Learning, GAIL)[1] 采用了生成对抗网络(Generative Adversarial Networks, GANs)[2]中的生成对抗方法。通过生成对抗网络训练过程，建立生成式对抗模仿学习与GANs中的概率分布散度优化的联系，从而极大提升了模仿策略的学习效率。

对于生成对抗式神经网络，其主体部分由判别器（Discriminator）网络 $D(x)$ 和生成器（Generator）网络 $G(x)$ 。训练过程中，生成器将采样自服从高斯先验分布的隐变量 z 映射到跟真实

样本同维度的高维变量（比如图片），而判别器则区分样本来自生成器的输出还是来自真实样本。随着生成器与判别器的不断交替迭代训练，生成器逐渐产生使对抗器无法辨别的数据，进而实现了对复杂后验分布的建模。其优化目标可以数学形式化为：

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1-D(G(\boldsymbol{z})))],$$

其中 $\boldsymbol{x} \sim p_{\text{data}}$ 代表 $\boldsymbol{x}$ 来自真实样本的分布，p_z 代表生成器的输入隐变量 z 的先验分布，一般为标准高斯分布。

我们需要将示范样本中的状态--动作的成对数据视作生成对抗式网络中的图像。对应到生成式对抗模仿学习方法中，强化学习的优化过程类似于 GANs 的生成器优化过程，判别器区分数据来自专家示范样本还是来自强化学习探索产生的新样本。

具体对于优化策略生成器 π_θ 而言：

$$\max_{\pi_\theta} \mathbb{E}_{(s,a)\sim\rho(s,a)}[\log \pi_\theta(a|s) Q(s,a)],$$

其中 $Q(s,a) = \mathbb{E}_{(s,a)\sim\rho(s,a)}[\log(D(s,a))]$， $\rho(s,a)$ 为智能体以 π_θ 策略与环境交互生成的新样本，策略优化过程是以判别器的输出值 $\log(D(s,a))$ 为奖惩函数的强化学习优化过程。

对于优化判别器 $D(s,a)$ 的目标函数可以定义为以下形式：

$$\min_D \mathbb{E}_{(s,a)\sim\rho(s,a)}[\log(D(s,a))] + \mathbb{E}_{(s,a)\sim\rho_E(s,a)}[\log(1-D(s,a))],$$

其中 $\rho_E(s,a)$ 是来自专家示范的样本。

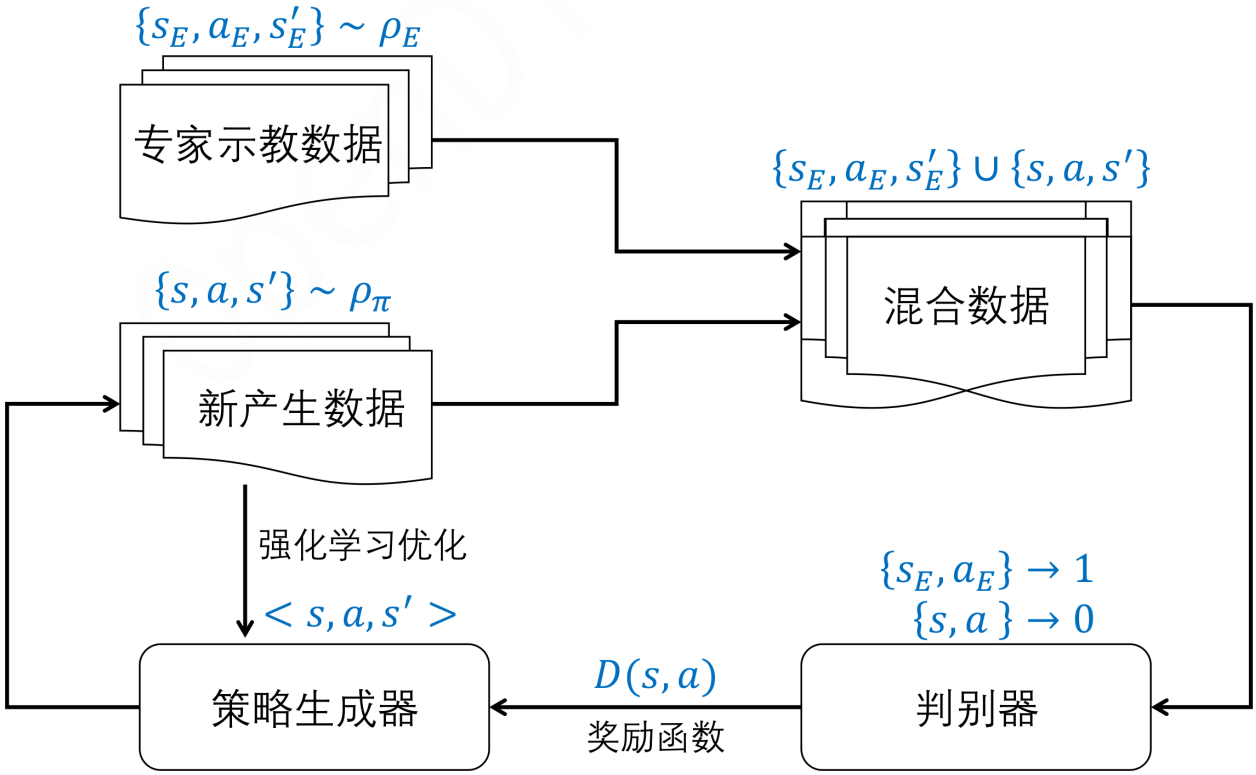

图1：生成式对抗模仿学习流程示意图

生成式对抗模仿学习的整体优化流程如图1所示。通过 GAIL 方法，策略生成器通过生成类似专家示教样本的探索样本，泛化示教样本的概率分布， 逼近专家示范行为数据，进而实现模仿专家技能的目的。该过程直接优化采样样本的概率分布，计算代价较小且算法通用性更强，实际模仿效果也更好。

多模态对抗模仿学习

这类方法还可以进一步推广到多行为模态（Multi-Modal）的模仿学习[3-5]中，此处的模态代表模仿专家样本的不同偏好或者倾向。InfoGAIL[3]与多模态行为模仿[45]方法中引入额外的潜在变量，以刻画示教数据中的模态信息，不同的潜在变量代表不同模态的示范数据， 对应不同示教者的偏好示范特性。基于生成模型向多模态生成模型拓展的思路，在多模态模仿学习中，通过潜在变量与新样本数据的状态-动作对（(s_t, a_t)）数据的互信息建立二者的关联关系，最后通过最大化互信息建立模态隐变量 z 与交互数据 (s_t, a_t) 关系，进而实现对多模态示教数据的模仿与逼近。除了这类单阶段的多模态模仿学习方法外，两阶段多模态模仿学习方法[4]，通过信息重构和变分推断预测模态变量，达到实现策略行为多样化的目的。

对抗式模仿学习GAIL，虽然在理论情况下可以完成模仿者产生的状态-动作分布 $\rho(s, a)$ 完全匹配示范数据的状态-动作分布 $\rho_E(s, a)$，但是根据 min-max 优化的均衡理论，判别器 $D(s, a)$ 最终对所有的状态-动作对数据(s,a)将收敛至0.5，显然优化出的判别器不能代表实际 MDP 中的奖励函数，导致判别器缺乏可解释性。在生成对抗网络指导下代价学习(Generative Adversarial Networks Guided Cost Learning, GAN-GCL)[6], 同样是基于 GAN 方法优化代价学习，最终学习最优奖励函数形式为：$R_\theta(\tau) = \log(1 - D_\theta(\tau)) - \log(D_\theta(\tau))$。此外，为了缓解 GCL 方法中以轨迹为中心的估计方法带来的估计方差较大的影响，对抗性逆强化学习 (Adversarial Inverse Reinforcement Learning, AIRL)[7]基于单一的状态-动作对应的特性函数优化判别器。

基于散度逼近的模仿学习

从数学层面上看，无论是行为克隆还是逆强化学习方法，模仿学习其本质都是在某种度量空间下，刻画模仿者与示教者的统计特征相似度问题。二者统计特征越接近，说明模仿者的技能越接近示教者，模仿学习越接近最优解。 在统计信息论中，这种统计特征相似度可以采用概率分布的距离度量方法，在不同的散度距离上建立流行空间上一个概率分布到另外一个概率分布的距离。

不同散度定义下的模仿学习方法列举在下表中。

行为克隆 BC	$E_{\rho^{exp}(s,a)} D_{KL}[\pi^{exp}(a\|s)\|\|\pi(a\|s)] = -E_{\rho^{exp}(s,a)}\ \pi(a\|s) + C$
DAgger	在第 $n+1$ 次迭代时: $E_{\rho^{agg_{1:n}}(s)} KL[\pi^{exp}(a\|s)\|\|\pi(a\|s)]$
AIRL	$KL(\rho^{\pi}(s,a)\|\|\rho^{exp}(s,a)) = -E_{\rho^{\pi}(s,a)}[log\ \rho^{exp}(s,a)] - \mathcal{H}(\rho^{\pi}(s,a))$
GAIL	$D_{JS}(\rho^{\pi}(s,a)\|\|\rho^{\exp}(s,a)) - \lambda\mathcal{H}^{causal}(\pi)$
FAIRL	$KL(\rho^{exp}(s,a)\|\|\rho^{\pi}(s,a)) = -E_{\rho^{\pi}(s,a)}[log\ \rho^{\pi}(s,a)] - \mathcal{H}(\rho^{exp}(s,a))$
对称 f - 散度	$D_f(\rho^{\pi}(s,a)\|\|\rho^{exp}(s,a)) - \lambda\mathcal{H}^{\text{causal}}(\pi)$
f - MAX	$D_f(\rho^{\pi}(s,a)\|\|\rho^{exp}(s,a))$
PWIL	$D_{\mathcal{W}_1}\ (\hat{\rho}^{\pi}(s,a), \hat{\rho}^{\exp}(s,a))$
SIL	$D_{\mathcal{W}_s^{\beta}}\ (\rho^{\pi}(s,a), \rho^{\exp}(s,a))_{c_w}$
GWIL	$D_{\mathcal{GW}}\ (\pi,\pi') = \mathcal{GW}\ ((S_E \times A_E, d_E, \rho_{\pi_E}), (S_A \times A_A, d_A, \rho_{\pi_A}))$

1. **行为克隆**：优化专家策略 $\pi^{\exp}(a|s)$ 与模仿策略 $\pi(a|s)$ 之间的 KL 散度距离，利用KL散度定义展开优化目标可知：

   $$E_{\rho^{exp}(s,a)} D_{KL}[\pi^{exp}(a|s)||\pi(a|s)] = -E_{\rho^{exp}(s,a)}\ \pi(a|s) + C + \mathcal{H}^{\exp}(s,a)$$

   其中在给定专家数据集情况下， (s,a) 联合概率分布固定不变，因此其熵信息确定不动，即 $\mathcal{H}^{\exp}(s,a)$ 为固定常数，最小化专家策略与模仿策略的KL散度距离的优化过程，等效于专家样本下对模仿策略的极大似然估计，

2. **DAgger**：迭代优化二者策略的 KL 散度距离，其中 $\mathbb{E}_{\rho^{\text{agg}1:n}(s)} = \frac{1}{n}\sum_{i=1}^{n} \rho^{\pi^{(i)}}(s)$, π^i 为交互过程中第 i 迭代中专家策略。 $\mathcal{H}^{\text{causal}}(\cdot)$ 代表信息熵， $\rho(s,a)$ 等效于状态-动作联合分布。

3. **AIRL** 与 **FAIRL**：分别代表专家与模仿者策略采样状态-动作联合分布 $\rho_E(s,a)$，$\rho(s,a)$ 的正向KL散度与反向KL散度距离。

4. **GAIL**：本质优化专家与模仿者策略采样状态--动作联合分布的JS散度距离或对称散度距离

   上一节介绍通过 GAN 网络优化专家数据的状态-动作对与生成样本之间的概率分布距离，不过为了高效探索与全局最优求解，在最小化 JS 散度基础上最大化策略信息熵。值得注意的是，对于 GAIL 中的对抗式网络优化中，不少方法选择 Wasserstein 散度距离优化，避免模式崩塌等典型训练问题。同时现有方法在应用GAIL方法时, 也通常额外引入梯度惩罚(Gradient Penalty)作为优化目标的正则项约束，保证模仿学习过程的稳定性。

5. **PWIL** 与 **SIL**：分别在其方法中采用近似 Wasserstein 优化方法与 Sinkhorn-Knopp 迭代求解散度逼近问题的方式，进一步优化专家与模仿策略之间的覆盖性测度距离。

6. **GWIL**：采用 Gromov Wasserstein 统计散度表示示教者与模仿者之间的距离主要针对的是示教者与模仿者动力学系统不相同的跨域模仿学习。

以散度逼近理论的视角将几乎所有的模仿学习统一在相同的框架下，为后续的模仿学习算法发展提供了很好的理论基础。

[1] Ho J, Ermon S. Generative adversarial imitation learning[C]//Advances in Neural Information Processing Systems (NeurIPS). 2016.

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in neural information processing systems, 2014, 27.

[3] Li Y, Song J, Ermon S. Infogail: Interpretable imitation learning from visual demonstrations [J]. Advances in Neural Information Processing Systems, 2017, 30.

[4] Wang Z, Merel J S, Reed S E, et al. Robust imitation of diverse behaviors[J]. Advances in Neural Information Processing Systems, 2017, 30.

[5] Hausman K, Chebotar Y, Schaal S, et al. Multi-modal imitation learning from unstructured demonstrationsusinggenerativeadversarialnets[J]. Advancesinneuralinformationprocessing systems, 2017, 30.

[6] Finn C, Levine S, Abbeel P. Guided cost learning: Deep inverse optimal control via policy optimization[C]//International conference on machine learning. PMLR, 2016: 49-58.

[7] Fu J, Luo K, Levine S. Learning robust rewards with adversarial inverse reinforcement learning[J]. arXiv preprint arXiv:1710.11248, 2017.