LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

南京大学
NANJING UNIVERSITY

# An Algorithmic Perspective on Imitation Learning Part4-5

Presented by He Quan

2020.2.2

# Section 4: Inverse Reinforcement Learning

# 4.1 问题陈述



- Recovering the reward function can be beneficial when the reward function is the most parsimonious way to describe the desired behavior

1. 多任务学习，权衡不同因素之间的影响
2. 从专家的决策中量化回报函数

# 4.1 问题陈述



- **IRL过程：**
- **Given** 1) measurements of an agent's behavior over time, in a variety of circumstances, 2) measurements of the sensory inputs to that agent; 3) a model of the physical environment (including the agent's body).

  **Determine** the reward function that the agent is optimizing

专家的状态-动作轨迹+环境

# 4.1 问题陈述



- A common assumption in IRL is that the demonstrator utilizes a Markov decision process (MDP)

- MDP：$(\mathcal{X}, \mathcal{U}, \mathcal{P}, \gamma, D, R)$

- many IRL methods assume that there are vectors of features $\Phi: \mathcal{X} \mapsto [0,1]^k$

# 4.1 问题陈述



Goal :recover the unknown reward function from the expert's trajectories(Multiple solution)

原则：

1. margin between the optimal policy and others
2. maximize the entropy

,

# 4.1 问题陈述



**Algorithm 14** Abstract version of feature matching inverse reinforcement learning

**Input:** Expert trajectories $\mathcal{D} = \{\boldsymbol{\tau}_i\}_{i=1}^{N}$

Initialize the reward function and policy parameters $\boldsymbol{w}, \boldsymbol{\theta}$

**repeat**

Evaluate the state-action visitation frequency $\boldsymbol{\mu}$ of the current policy $\pi_\theta$

Evaluate the objective function $\mathcal{L}$ and its derivative $\nabla_{\boldsymbol{w}}\mathcal{L}$ by comparing $\boldsymbol{\mu}$ and the state-action distribution implied by $\mathcal{D}$

Update the reward function parameter $\boldsymbol{w}$

Update the policy parameter $\boldsymbol{\theta}$ with a reinforcement learning method

**until**

**return** optimized policy parameters $\boldsymbol{\theta}$ and reward function parameter $\boldsymbol{w}$

IRL过程，依据比较专家轨迹与习得策略轨迹更新奖励函数参数

RL过程，依据奖励函数更新策略参数

需要迭代式更新

# 4.2 基于模型-无模型的逆强化学习



Model-Based：需要学习系统动态转移概率（系统不能太过复杂），可以规划出学习者策略的生成轨迹分布

Model-Free：无需学习系统动态转移概率，需要从大量采样中估算学习者策略的生成轨迹分布

| | Model-free | Model-based |
|---|---|---|
| **Advantages** | Applicable to systems with nonlinear and unknown dynamics | Estimation of the trajectory distribution is data-efficient. |
| **Disadvantages** | It is necessary to sample many trajectories to estimate the trajectory distribution. | Model learning can be very difficult.<br>It is hard to apply to underactuated systems. |

# 4.3 逆强化学习方法设计原则



Q1：如何定义目标，使得奖励函数有唯一解？

| Objectives | Employed by |
|---|---|
| *Maximum margin* | [Ng and Russell, 2000, Abbeel and Ng, 2004, Ratliff et al., 2006b,a, 2009, Silver et al., 2010, Zucker et al., 2011] |
| *Maximum entropy* | [Ziebart et al., 2008, Ramachandran and Amir, 2007, Choi and Kim, 2011b, Ziebart, 2010, Boularias et al., 2011, Kitani et al., 2012, Shiarlis et al., 2016, Ho and Ermon, 2016, Finn et al., 2016b] |
| *Other* | [Doerr et al., 2015, Arenz et al., 2016] |

# 4.3 逆强化学习方法设计原则



Q2：使用线性/非线性奖励函数?

非线性奖励函数有着相对更好的表示性，但相对更难学习参数

| | Model-free | Model-based |
|---|---|---|
| **Linear reward** | [Boularias et al., 2011, Kalakrishnan et al., 2013] | [Abbeel and Ng, 2004, Ratliff et al., 2006b, Silver et al., 2010, Ramachandran and Amir, 2007, Choi and Kim, 2011b, Ziebart et al., 2008, Ziebart, 2010, Levine and Koltun, 2012, Hadfield-Menell et al., 2016] |
| **Nonlinear reward** | [Finn et al., 2016b, Ho and Ermon, 2016] | [Ratliff et al., 2006a, 2009, Silver et al., 2010, Grubb and Bagnell, 2010, Levine et al., 2011] |

# 4.4 基于模型的逆强化学习方法



## 4.4.1特征期望匹配

假设奖励是特征的线性函数：$r(x) = \mathrm{w}^{\mathrm{T}}\phi(x)$

则$\mathbb{E}[R|\pi] = \mathbb{E}[\sum_{t=0}^{T}\gamma^t r(x_t)|\pi] = w^T\mathbb{E}[\sum_{t=0}^{T}\gamma^t\phi(x_t)|\pi]$

令$[\sum_{t=0}^{T}\gamma^t\phi(x_t)|\pi] = \mu(\pi)$(特征期望)

则$\mathbb{E}[R|\pi] = w^T\ \mu(\pi)$

问题转化为求$\mu(\pi) = \mu(\pi_E)$

# 4.4 基于模型的逆强化学习方法



### 4.4.2最大间隔规划

寻找cost函数使演示轨迹的cost与其他轨迹的cost差距最大

$$C(\tau^{demo}) \leq \min\{C(\tau) - L(\tau)\}$$

若$C(\tau)=w^T\phi(\tau)$，$\phi(\tau)=F\mu$，$L(\tau)=l^T\mu$，其中$l$为损失向量，$\mu$为状态-动作对频率统计，F为特性矩阵，$\zeta$为松弛变量，$\lambda>0$为衡量偏离约束的惩罚和对权重向量的二范数正则的常数，则寻找cost函数参数$w$转化为优化如下问题：

$$\begin{aligned} &\min_{w,\zeta_i} \ \frac{\lambda}{2}\|w\|^2 + \frac{1}{N_i}\sum_{i=1}^{N}\zeta_i \\ &s.t. \ \ \forall i, w^T\phi(\tau_i) \leq \min\{w^T\phi(\tau) - l_i^T\mu\} + \zeta_i \end{aligned}$$

# 4.4 基于模型的逆强化学习方法



令松弛变量$\xi_i = \boldsymbol{w}^\top F_i \boldsymbol{\mu}_i - \min_{\boldsymbol{\mu}} \{\boldsymbol{w}^\top F_i \boldsymbol{\mu} - \boldsymbol{l}_i^\top \boldsymbol{\mu}\}$,则最终优化目标：

$$\ell_{\text{MMP}}(w) = \frac{1}{N}\sum_{i=1}^{N} (w^\top F_i \mu_i - \min_{\mu} \{w^\top F_i \mu - l_i^\top \mu\}) + \frac{\lambda}{2}\|w\|^2$$

梯度法解该优化问题

**Algorithm 16** Maximum margin planning Ratliff et al. [2006b]

**input:** Training set $\mathcal{D} = \{F_i, \boldsymbol{\tau}_i, \boldsymbol{l}_i\}_{i=1}^N$, regularization parameter $\lambda > 0$, stepsize sequence $\{\alpha_t\}$, iteration $T$
**while** $t < T$ **do**
  **for** $i = 1, ..., N$ **do**
    Compute the loss-augmented cost map $\tilde{c}_i = \boldsymbol{w}^\top F_i - \boldsymbol{l}_i^\top$
    Compute the optimal trajectory $\boldsymbol{\tau}_i^* = \arg\min \tilde{c}_i \boldsymbol{\mu}$
    Compute the state-action frequency couts $\boldsymbol{\mu}_i^*$
  **end for**
  Compute the subgradient $\boldsymbol{g} \in \partial \mathcal{L}_{\text{MMP}}(\boldsymbol{w})$
  $\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha_t \boldsymbol{g}$
  (Optional) Project $\boldsymbol{w}$ on to any additional constraint
  $t \leftarrow t + 1$
**end while**
**return** $\boldsymbol{w}$

## 4.4.3最大熵原则

在特征期望匹配的条件下最大化策略生成轨迹的熵

$$\max_{p} H(p(\tau)) = \sum p(\tau) \ln \frac{1}{p(\tau)}$$

$$\mathbf{s.t.}\ \mathbb{E}_{\pi^L}[\phi(\tau)] = \mathbb{E}_{\pi^E}[\phi(\tau)]$$

$$\sum_{\tau} p(\tau) = 1, \forall \tau, p(\tau) > 0$$

可得$p(\tau)$服从指数族分布 $p(\tau) \propto \exp\left(w^T \phi(\tau)\right) = \exp\left(R(\tau)\right)$

W为拉格朗日乘子　　R为w参数的线性奖励

# 4.4 基于模型的逆强化学习方法



在状态转移确定下有

$$p(\tau \mid w) = \frac{1}{Z(w)} \exp\ (w^{\top} \phi(\tau))$$

随机环境下

$$p(\tau \mid w) = \frac{1}{Z(w)} \exp\ (w^{\top} \phi(\tau)) \prod_{x_{t+1}, u_t, x_t \in \tau} p(x_{t+1} \mid u_t, x_t)$$

由于配分函数未知，需要用次梯度的方式求解最优值，令$D_{x_i}$为访问$x_i$概率，则

$$\nabla l_{\mathrm{ME}}(w) = \mathbb{E}_{\pi^E}[\phi(\tau)] - \sum_{\tau} p(\tau \mid w) \phi(\tau) = \mathbb{E}_{\pi^E}[\phi(\tau)] - \sum_{x_i} D_{x_i} \phi(x_i)$$

# 4.4 基于模型的逆强化学习方法



**使用因果熵**：将优化目标从熵变为因果熵

$$\pi^*(u \mid x) = \underset{\pi^{\mathrm{L}}(u \mid x)}{\operatorname{argmax}} H(u_{1:T} \parallel x_{1:T})$$

其中因果熵为条件熵之和

$$H(u_{1:T} \parallel x_{1:T}) = \sum_{t=1}^{T} H(u_t \mid u_{1:t-1}, x_{1:t})$$

即任何时刻的动作分布只与之前时刻的动作-状态序列有关

# 4.4 基于模型的逆强化学习方法



从失败演示中学习

$$\max_{\pi^L(u \mid x), w, z} H(u_{1:T} \| x_{1:T}) + \sum_{k=1}^{K} w_k z_k - \frac{\lambda}{2} \| w \|^2$$

s.t.

$$\mathbb{E}_{\pi^L(u \mid x)}[\phi(\tau_S)] = \mathbb{E}_{\pi^E}[\phi(\tau_S^{\text{demo}})]$$
$$\mathbb{E}_{\pi^L(u \mid x)}[\phi(\tau_F)] - \mathbb{E}_{\pi^E}[\phi(\tau_F^{\text{demo}})] = z_k$$
$$\sum_u \pi^L(u \mid x) = 1, \pi^L(u \mid x) \geq 0$$

即加入了最大化与失败演示特征期望的距离

**最大熵与经济学的联系：**微观经济学下为理性人行为的奖励函数建模

## 4.4.4基于其他重要模型的IRL方法

**线性可解的MDP**：将线性可解系统分为被动系统 $P(x_{t+1}|x_t)$（与策略无关的基础转移）与主动系统$\pi(x_{t+1}|x_t)$（与策略有关）,直接求解原动态系统的值函数

损失函数变为

$$c(\boldsymbol{x}_t, \pi) = c(\boldsymbol{x}_t) + D_{\mathrm{KL}}(\pi \| p)$$

类似最大熵原理得

$$\pi(x_{t+1} \mid x_t) = \frac{p(x_{t+1} \mid x_t) \exp(V(x_{t+1}))}{Z}$$

通过该式用类似方法可直接求解值函数但值函数难以应用到类似奖励但具有不同动态的系统

# 4.4 基于模型的逆强化学习方法



## 贝叶斯框架下的逆强化学习

专家的行动被视为可用于更新奖励先验的“证据”：

$$p(R \mid \tau) = \frac{p(\tau \mid R)p(R)}{p(\tau)} = \frac{1}{Z}\exp(\alpha E(\tau, R))p(R)$$

其中E为玻尔兹曼式分布，前沿分布可为均匀或高斯：使用Markov chain Monte Carlo法求解

或者通过maximum-a-posterior（后验最大）方法：

$$R_{\mathrm{MAP}} = \arg\max_R p(R \mid \mathcal{D}) = \arg\max_R [\ln p(\mathcal{D} \mid R) + \ln p(R)]$$

其中D为专家样本状态-动作对数据

## 4.4.5 学习非线性奖励函数

**集成方法：**使用多个监督学习算法合成高度非线性函数

**神经网络方法：**使用反向传播进行学习

**高斯过程逆强化学习：**假设输出为基于输入特征的高斯分布

## 4.4.6 指导成本学习

训练一个新的采样分布并使用重要性采样，来对归一化系数Z进行估算：

$$\mathcal{L}_{\mathrm{GCL}} = \frac{1}{N} \sum_{\tau_j \in \mathcal{D}_{\mathrm{demo}}} c_w(\tau_i) + \ln Z$$

$$\approx \frac{1}{N} \sum_{\tau_i \in \mathcal{D}_{\mathrm{demo}}} c_w(\tau_i) + \ln \frac{1}{M} \sum_{\tau_j \in \mathcal{D}_{\mathrm{samp}}} \frac{\exp\left(-c_w(\tau_j)\right)}{q(\tau_j)}$$

迭代式更新$c_w$与$q(\tau_j)$

# 4.5 无模型的逆强化学习方法



## 4.5.1 相对熵逆强化学习

最小化基准策略的先验轨迹分布$q_0(\tau)$与学习者策略的轨迹分布$p(\tau)$的相对熵：

$$\min \quad \sum p(\tau) \ln \frac{p(\tau)}{q_0(\tau)}$$

$$\text{s.t.} \quad \forall i \in \{1, \ldots k\}, \left| \mathbb{E}_{\pi^L}[\phi_i(\tau)] - \mathbb{E}_{\pi^E}[\phi_i(\tau)] \right| \leq \epsilon_i$$
$$\sum_{\tau \in \mathbb{T}} p(\tau) = 1$$
$$\forall \tau \in \mathbb{T}, p(\tau) \geq 0$$

其中$\epsilon_i$为给定置信度下专家轨迹特征期望与学习者轨迹特征期望差距的上界：通过重要性采样的方式来估算优化目标的梯度

## 4.5.2 生成对抗式模仿学习

最小化与专家策略的occupancy measure的JS散度

**Algorithm 1** Generative adversarial imitation learning

**Input:** Expert trajectories $\tau_E \sim \pi_E$, initial policy and discriminator parameters $\theta_0, w_0$
**for** $i = 0, 1, 2, \ldots$ **do**
  Sample trajectories $\tau_i \sim \pi_{\theta_i}$
  Update the discriminator parameters from $w_i$ to $w_{i+1}$ with the gradient

$$\hat{\mathbb{E}}_{\tau_i}[\nabla_w \log(D_w(s,a))] + \hat{\mathbb{E}}_{\tau_E}[\nabla_w \log(1 - D_w(s,a))] \tag{17}$$

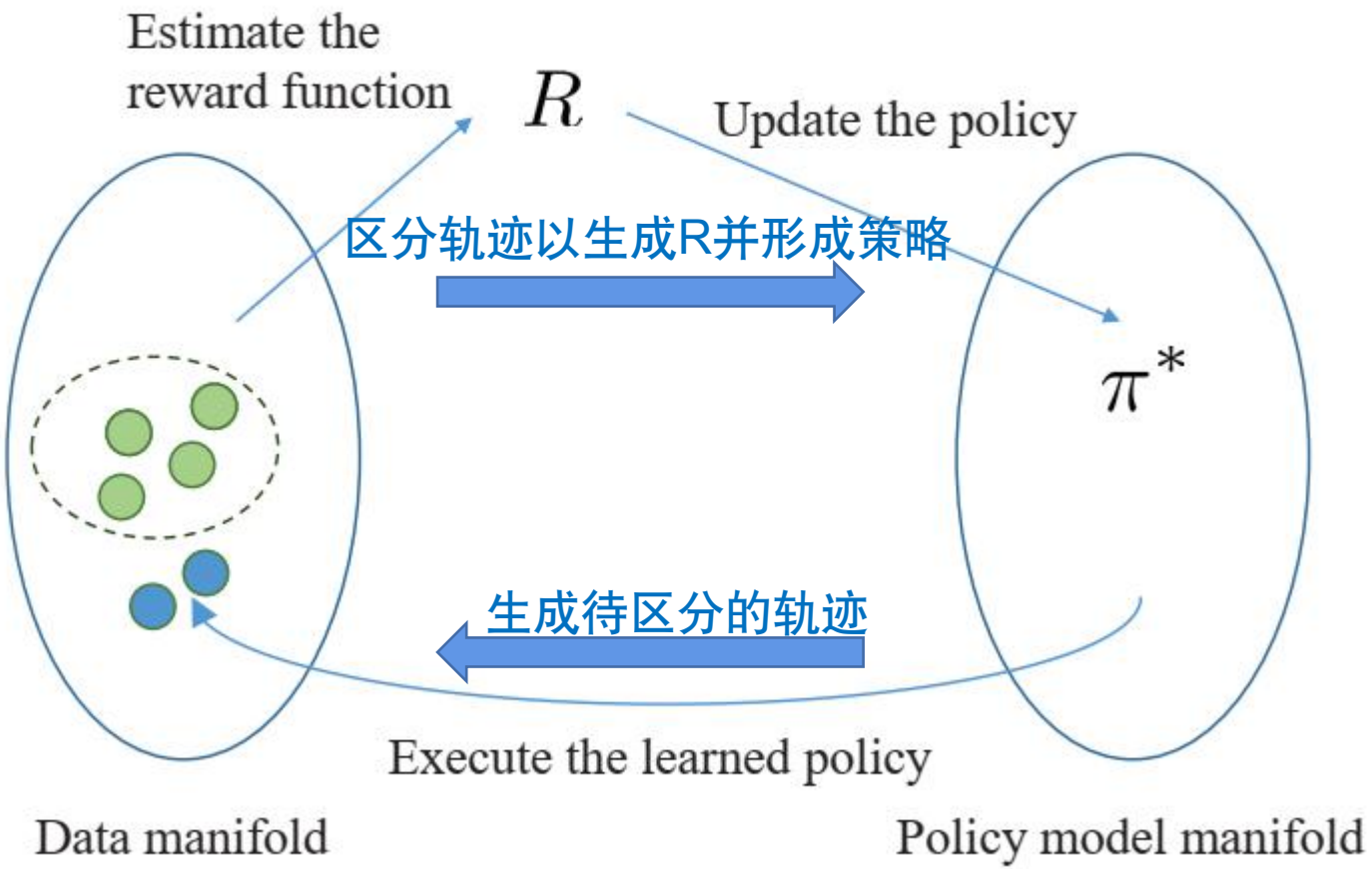
  Take a policy step from $\theta_i$ to $\theta_{i+1}$, using the TRPO rule with cost function $\log(D_{w_{i+1}}(s,a))$. Specifically, take a KL-constrained natural gradient step with

$$\begin{aligned} &\hat{\mathbb{E}}_{\tau_i}\left[\nabla_\theta \log \pi_\theta(a|s) Q(s,a)\right] - \lambda \nabla_\theta H(\pi_\theta), \\ &\text{where } Q(\bar{s},\bar{a}) = \hat{\mathbb{E}}_{\tau_i}[\log(D_{w_{i+1}}(s,a)) \,|\, s_0 = \bar{s}, a_0 = \bar{a}] \end{aligned} \tag{18}$$

**end for**

# 4.6 最大熵原则的信息论解释





# 4.6 最大熵原则的信息论解释



最大熵原则：

$$H(p(\tau)) = \sum p(\tau) \ln \frac{1}{p(\tau)}$$

最小化学习者分布$p(\boldsymbol{\tau})$与先验分布$p_0(\boldsymbol{\tau})$的散度：

$$D_{\mathrm{KL}}(p(\boldsymbol{\tau}) \parallel p_0(\boldsymbol{\tau})) \simeq \sum p(\boldsymbol{\tau}) \ln \frac{p(\boldsymbol{\tau})}{p_0(\boldsymbol{\tau})}$$

当先验分布为均匀分布时，最大熵原则等价于最小化学习者分布与先验分布的散度

# 4.7 部分可观测下的逆强化学习



某些时候，由于人和物体引起的传感器噪声和阻塞等问题，环境状态不是完全已知的：同时，逆强化过程可以被视为智能体对奖励函数有着不完全的观测

**4.7.1专家轨迹部分可观测：**通常为专家轨迹带有噪音

使用隐变量马尔科夫过程：类似最大熵有$p(\tau o, \theta) \approx \frac{\exp(w^T \phi_\tau')}{Z(w)}$,但现在的状态包括了观测概率

使用Markov random fields 基于距离对状态进行纠正：通常需要有易定义的距离（例如导航系统），缺点是需要较大计算量

不考虑被遮挡的状态或动作，仅针对可观察状态计算特征期望值

# 4.7 部分可观测下的逆强化学习



**4.7.2专家的观察部分可观测：**专家本身处于部分可观测环境

将专家环境建模为POMDP环境

IRL问题在POMDP上的拓展分为两类：1.已知专家策略 2.已知专家轨迹

J. Choi and K. Kim. Inverse reinforcement learning in partially observable environments. *Journal of Machine Learning Research*, 12(Mar):691–730, 2011a.

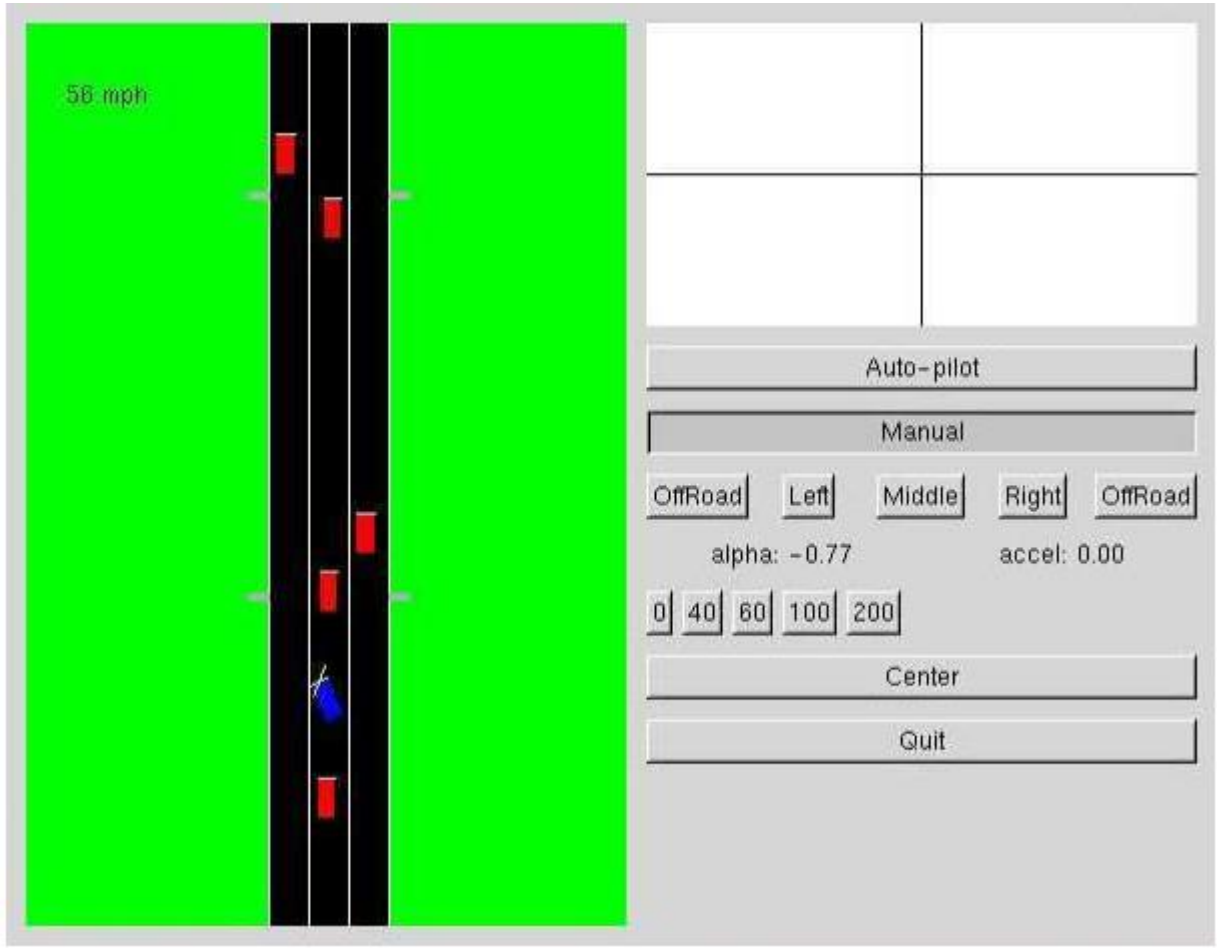
**4.7.3作为POMDP的主动逆强化学习：**主动逆强化学习可以建模为获取奖励函数信息的POMDP过程

**4.7.4协同逆向强化学习：**人与机器人互动，人观察到奖励功能而机器人没有观察到：人可能选择给予机器人更多信息的对自身而言的次优解：寻找统合的最优解过程类似于POMDP过程

# 4.8 逆强化学习的应用



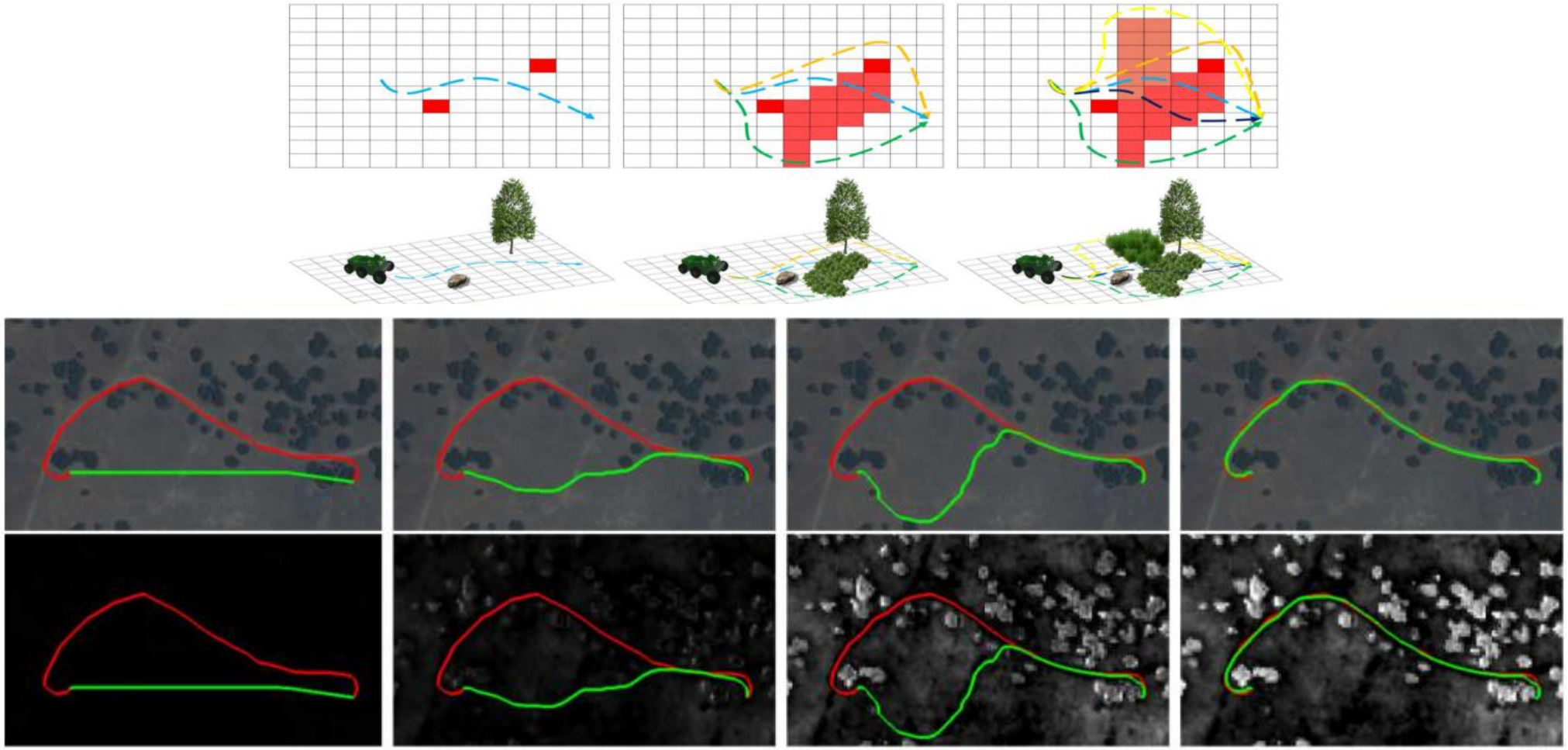
Model-based：学习不同驾驶风格的自动驾驶



# 4.8 逆强化学习的应用



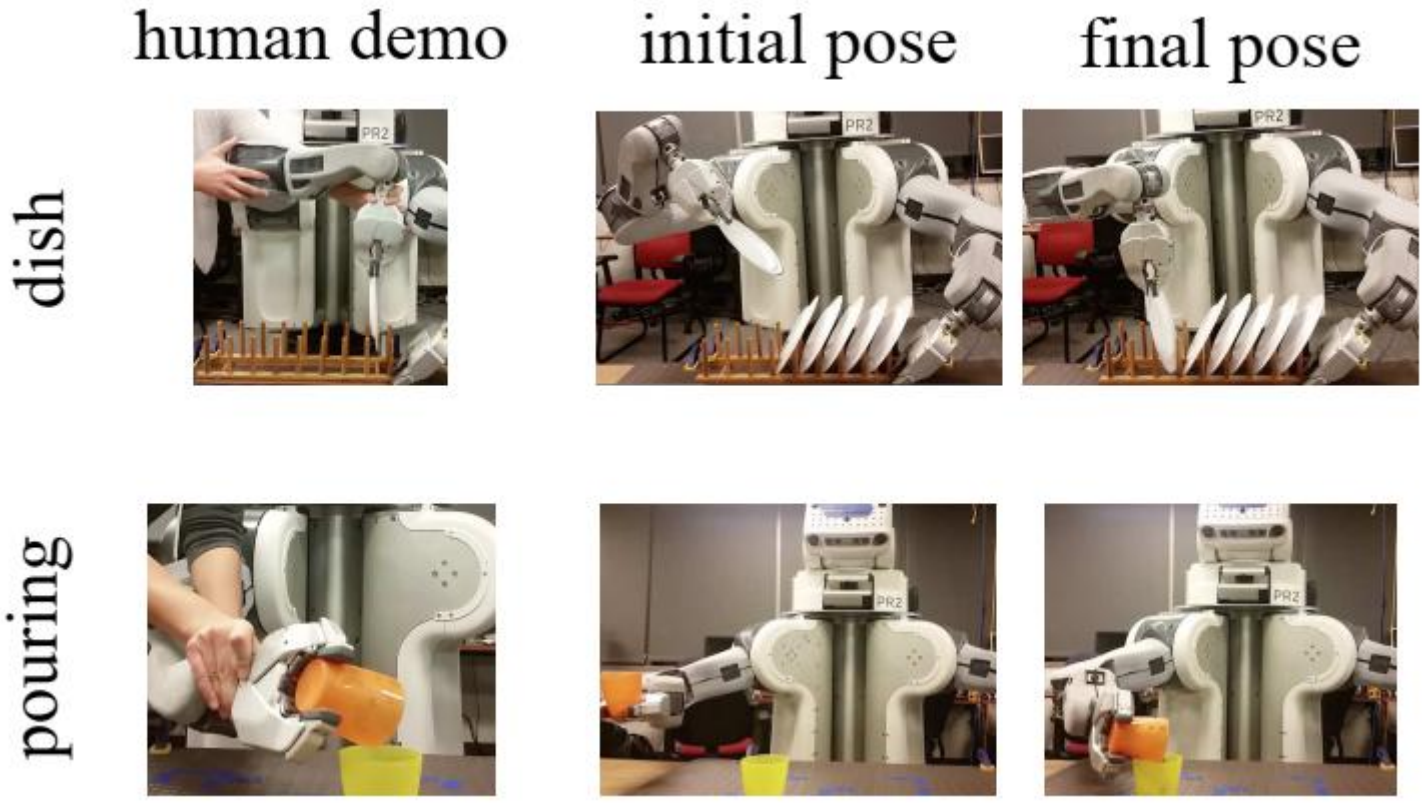
Model-based：学习避开地势的路线规划



# 4.8 逆强化学习的应用



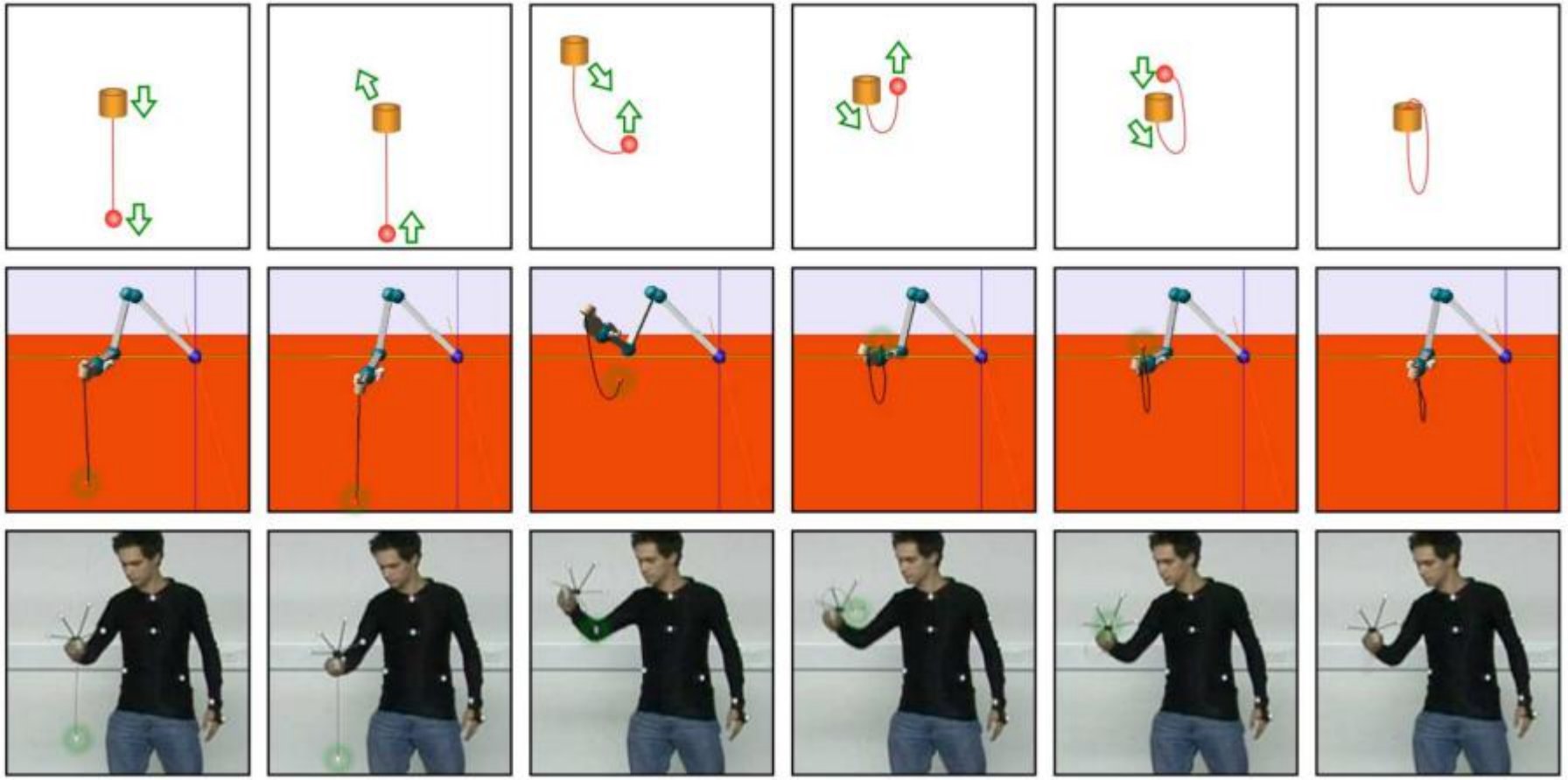
## 指导成本学习：学习非线性奖励的家务



# 4.8 逆强化学习的应用



## Model-free相对熵学习：学习引球入杯



# Section 5: Challenges in Imitation Learning for Robotics

# 5.1行为克隆 vs 逆强化学习



1.无法使用行为克隆的场景

从专家样本中恢复奖励函数可以解释为推断专家的意图，因为奖励函数隐含的任务目标。例如，当从一系列图像中进行学习时，若没有专家的运动学信息，则难以进行行为克隆。 在这种情况下，我们需要推断出专家所需的条件，然后规划出达到推断目标的策略。因此，**逆强化学习**是解决此类问题的合理选择。

2.行为克隆与逆强化学习均可使用的场景

考虑奖励与策略哪一种是“对期望的行为的最简约的描述”；

例1：操纵任务（如书写字符），没有显式定义的奖励函数，但有明显的演示轨迹：此时恢复奖励函数通常是不必要的

例2：四足机器人轨迹规划任务，奖励函数与经过的地形是否平坦有关；此时使用行为克隆难以描述这一偏好

# 5.2模仿学习的其他开放式问题



## 5.2.1与演示数据相关的问题

**如何从多个专家中学习**?

从多个专家的演示数据中进行学习的效果往往差于从单个专家中学习

"Third, in all domains, the best clones were obtained when examples from a single human only were used."?

**如何应对专家演示数据中的缺陷**?

专家演示数据中可能含有一些不良行为，导致模仿学习获得行为存在瓶颈

一种解决方法是在模型已知的情形下与强化学习结合

# 5.2模仿学习的其他开放式问题



## 5.2.1与演示数据相关的问题

**如何从原始传感器输入中学习**?

例如仅从视觉数据中学习时，我们无法直接获得专家的运动学信息。

一种解决方法是使用深度网络提取图像特征，再做逆强化学习推断奖励函数

**如何应对不同视点的问题**?

专家演示数据通常为第一人称，但学习者视角可能为第三人称，需要从第三人称中推断任务是如何进行的

# 5.2模仿学习的其他开放式问题



## 5.2.1与演示数据相关的问题

**如何从原始传感器输入中学习**?

例如仅从视觉数据中学习时，我们无法直接获得专家的运动学信息。

一种解决方法是使用深度网络提取图像特征，再做逆强化学习推断奖励函数

**如何利用其他相关任务的过去演示，快速了解当前任务**?

人类可以从很少的演示数据中进行学习，因为他们通常具有很多先验知识；如何从不同任务的演示中获得可以迁移的知识?

# 5.2模仿学习的其他开放式问题



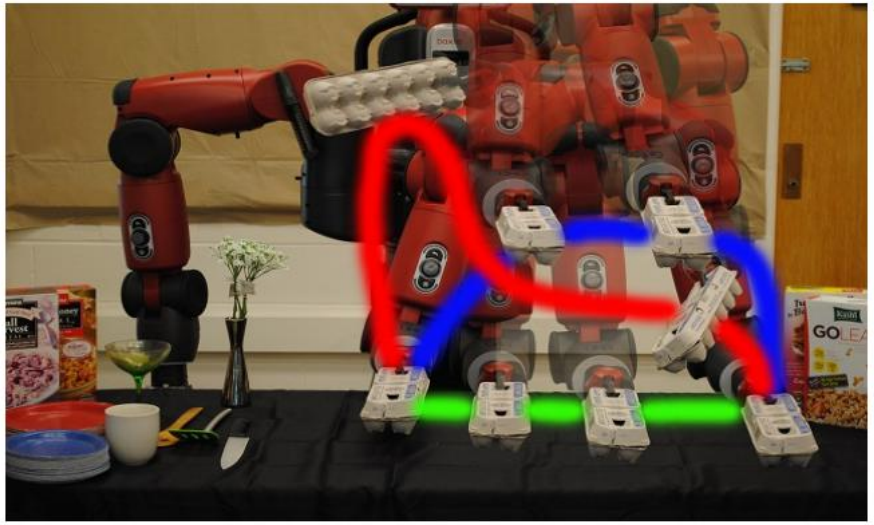
## 5.2.2与设计相关的问题

**如何选择策略的相似性度量**?

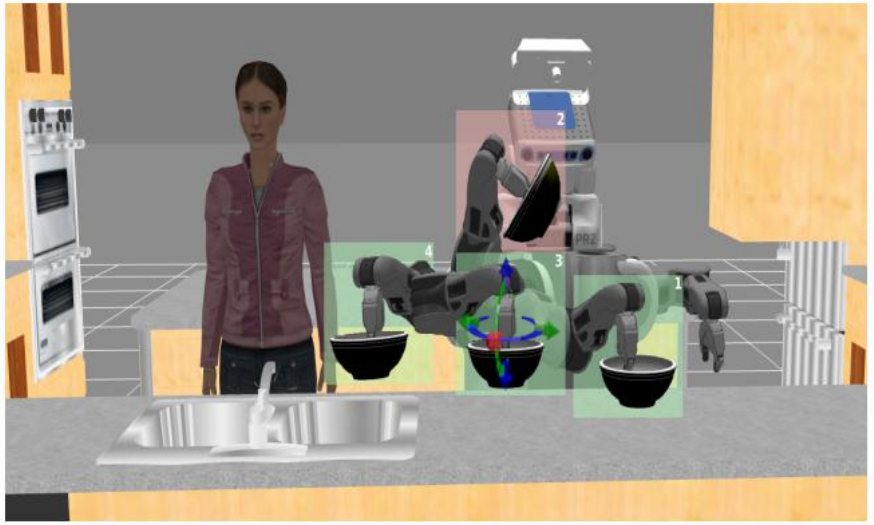
本文提到的相似度度量主要有KL散度与欧氏距离；仍有其他可以利用的距离如JS距离，Wasserstein距离等

**如何从多种指令方式中学习**?

评价式指令（选择最优轨迹）



交互式指令（更改路径点）



# 5.2模仿学习的其他开放式问题



## 5.2.2与设计相关的问题

**如何整合先验知识**?

尽管对系统或环境有先验知识，例如运动学或机械手的质量，很多模仿学习方法仅利用了演示轨迹。 另一方面，许多方法使用隐含的先验知识，例如假设了先验或者噪音的高斯分布。明确纳入先验知识的方法可以减少所需的演示数据量，并开发新的机器人应用。

**如何从多种传感器中学习**?

通常模仿学习的专家样本来源于同一种传感器，但可能同时存在多种传感器信息（例如触觉信息，视觉信息，音频信息可同时存在），如何利用额外的传感器信息?

（多模态融合问题：如双线性模型）

# 5.2模仿学习的其他开放式问题



## 5.2.2与设计相关的问题

**如何学习到人类无法操作的任务**?

人类可能无法给出一些任务的演示，特别是当机器人有着不同的机械能力时：例如机器人可能有两只以上的手臂

需要有能自行迭代优化策略的方法

**如何选择轨迹表示**?

| | Time dependence | Stabile attraction to a target position | Stochasticity of trajectories | Encoding spatial co-ordination patterns |
|---|---|---|---|---|
| Way points / Keyframe [Abbeel et al., 2010, Nakaoka et al., 2007] | ✓ | - | - | - |
| HMMs [Inamura et al., 2004, Takano and Nakamura, 2015] | ( ✓) | - | ✓ | ✓ |
| DMP [Schaal et al., 2004, Ijspeert et al., 2013] | ✓ | ✓ | - | - |
| ProMP [Paraschos et al., 2013, Maeda et al., 2016] | ✓ | - | ✓ | ✓ |
| SEDS [Khansari-Zadeh and Billard, 2011, 2014] | - | ✓ | - | ✓ |

# 5.2模仿学习的其他开放式问题



## 5.2.3与算法相关的问题

**如何在复杂条件下获得通用化的技术**?

DMP/ProMP等方法可以获得不同起点/终点的轨迹，但不能获得更高层次的技能（例如书写相似的字符):

使用深度网络从视觉中学习可能获得更好的通用性

**如何获得学习的理论保证**?

DMP/Dagger等算法具有理论性能的保证，但许多算法都没有相关保证，使得难以应用到需求稳定性的机器人领域：可能需要基于模仿学习的通用的理论保证基础

# 5.2模仿学习的其他开放式问题



## 5.2.3与算法相关的问题

**如何在复杂条件下获得通用化的技术**?

DMP/ProMP等方法可以获得不同起点/终点的轨迹，但不能获得更高层次的技能（例如书写相似的字符):

使用深度网络从视觉中学习可能获得更好的通用性

**如何获得学习的理论保证**?

DMP/Dagger等算法具有理论性能的保证，但许多算法都没有相关保证，使得难以应用到需求稳定性的机器人领域：可能需要基于模仿学习的通用的理论保证基础

# 5.2模仿学习的其他开放式问题



## 5.2.3与算法相关的问题

**如何基于维数扩大规模?**

例如类人机器人通常具有50多个关节，但现有模仿学习方法通常不适合学习如此高的维数的专家轨迹；神经网络可以处理高维的图像输入问题，但局限于2D图像。另一种可能的处理思路是进行降维

**如何在高维空间中找到全局最优解?**

机器人领域最优控制的可行空间可能为连续高维空间，而模仿学习通常会找到接近于专家行为的局部最优解：如何获得全局最优解?

# 5.2模仿学习的其他开放式问题



## 5.2.3与算法相关的问题

**如何在多智能体环境下进行模仿学习**？

多智能体环境下单个智能体的行为需要考虑其他智能体的影响；需要从多智能体形成的均衡行为中推断奖励函数

**如何在在模仿学习中应用主动/增量学习？**

当初始数据集中模仿到的策略性能不达到要求时，可使用主动/增量学习来针对特定情形扩充数据集；但相关增量式IRL方法尚未有充分的研究

# 5.2模仿学习的其他开放式问题



## 5.2.4与性能评估相关的问题

**如何建立模仿学习的基准评价问题**?

在模仿学习方面，缺乏类似计算机视觉/数据挖掘/自然语言处理的可以通用的Benchmark：不同方法可能适用的任务不同

**如何建立模仿学习的性能评价标尺?**

存在着不同种类的模仿学习的量化评价标准；如轨迹评价中有基于路径/基于速度等量化标准。

也存在着与最优轨迹间的相似度衡量基准选择等问题

# Thanks!