Introduction

Inverse Reinforcement Learning (IRL)
- Learn reward function from human demonstrations
- Applications include inferring human goals/preferences, learning policies that generalize to unobserved situations, transferring biological motion to robots

Letter Handwriting Task
- Teach robot the task of letter writing in a compliant way
- Complex task involving grasping force between the hand and the pen, pen-tip interaction force, orientation of hand

Goals and Objectives
- Humans adapt stiffness during contact with environment
- Humans and robots have different dynamics
- Use reward function in human demonstrations to teach optimal compliance to robot

Methodology: 3-Step Model-Based IRL Framework

1 - Learn Interaction Dynamics
- Collect 3 human demonstrations \( \{x_1(t), u_1(t)\} \) for writing each letter to learn the dynamics model:
\[
x = f_h(x, u)
\]
- \( u \in \mathbb{R}^3 \Rightarrow \) Interaction force \( \{F_x, F_y, F_z\} \)
- \( x \in \mathbb{R}^3 \Rightarrow \) Pen-tip position
- \( \dot{x} \in \mathbb{R}^3 \Rightarrow \) Pen-tip velocity
- \( f_h \) is learned from demonstrations using Support Vector Regression

2 - Learn Reward Function
- Reward function scores \( \{x_1(t), u_1(t)\} \) on its similarity to known reference letter trajectory \( \{x_d(t), u_d(t)\} \)
\[
r(x_1, u_1) = -[x_1 - x_d(t)]^T Q[x_1 - x_d(t)] - \cdots - [u_1 - u_d(t)]^T R[u_1 - u_d(t)]
Q = diag(w_1 w_2 w_3)
R = diag(w_4 w_5 w_6)
\]
- Find \( Q, R \) to maximize the probability of \( \{x_h(t), u_h(t)\} \) given \( \hat{x} = f_h(x, u) \) under the maximum entropy distribution [1]:
\[
P(\{x_h(t), u_h(t)\}|Q, R) = \frac{\exp \left( \sum_t r(x_h(t), u_h(t)) \right)}{\int \exp \left( \sum_t r(x_h(t), u_h(t)) \right) dx}
\]
\[
\pi = \begin{bmatrix} u_1, \ldots, u_T \end{bmatrix}
\]

3 - Learn Optimal Policy
- Optimal policy \( u = \pi(t) \) for the robot dynamics \( \dot{x} = f(x, u) \) and the learned reward function using iterative Linear Quadratic Regulator [2] gives the time-varying stiffness profiles
\[
\pi(t) = u_t + k_t + K_t(x - x_d)
\]
- \( k_t \Rightarrow \) open-loop input correction
- \( K_t \Rightarrow \) variable stiffness profiles

Results

- Learned Reward Function:
\[
Q = diag(0.53 0.26 0.0004) \quad R = diag(0.27 0.73 0.22)
\]

- Optimal Policy Response: Preliminary simulation results suggest learning variable stiffness profiles helps to counter perturbations in a compliant manner

---