RSS Paper

Active Preference-Based Learning of Reward Functions

Nikita Jaipuria

Aerospace Controls Laboratory
Department of Mechanical Engineering
Massachusetts Institute of Technology

August 11, 2017

Overview

- Objective:
  - model a **human's preference** for how a dynamical system should act
  - learn
    $R_H(\xi) = R_H(x^0, \mathbf{u}_R, \mathbf{u}_H) = \sum_{t=0}^{N} r_H(x^t, u_R^t, u_H^t) = \sum_{t=0}^{N} \mathbf{w}^T \phi(x^t, u_R^t, u_H^t) = \mathbf{w}^T \Phi(\xi)$
- Problem Domain:
  - difficult to provide demonstrations of **desired** system trajectory (IRL)
  - assign numerical reward to an action/trajectory
- Main Idea: active preference-based learning
  - system decides on what preference queries to make (**active**)
  - build on label ranking; learn from preferences/comparisons (**preference-based**)
- Challenges/Contribution
  - complexity and continuous nature of **queries**
  - **active synthesis** of queries satisfying system dynamics: $x^{t+1} = f_{HR}(x^t, u_R^t, u_H^t)$
  - **maximize volume removed** from continuous hypothesis space by each query

Overview

- Two main sections of active preference-based learning:
  - **Active query synthesis**: generate query ξ_A vs ξ_B defined over same fixed scenario $\tau = (x^0, \mathbf{u}_R)$ to maximize volume removed from continous hypothesis space of rewards
  - model probability $p(I|\mathbf{w})$ as noisily capturing preference w.r.t. R_H

update function: $f_\varphi(\mathbf{w}) = p(I_t|\mathbf{w}) = \dfrac{1}{1+\exp(-I_t \mathbf{w}^T \varphi)}$,where $\varphi = \Phi(\xi_A) - \Phi(\xi_B)$

Algorithm

- Inputs: $\phi, N, f_{HR}, iter$
- Output: $p(\mathbf{w})$
- Step 1: Initialize $p(\mathbf{w}) \sim Uniform(B)$, for a unit ball B
- Step 2: **synthesize query** to remove as much volume as possible from the space of possible rewards (*constrained optimization*)

Slide Title 3

Slide Title 4

Questions?

Backup Slide 1

- Blah blah blah ...

References I