RSS Paper

Active Preference-Based Learning of Reward Functions [1]

Nikita Jaipuria

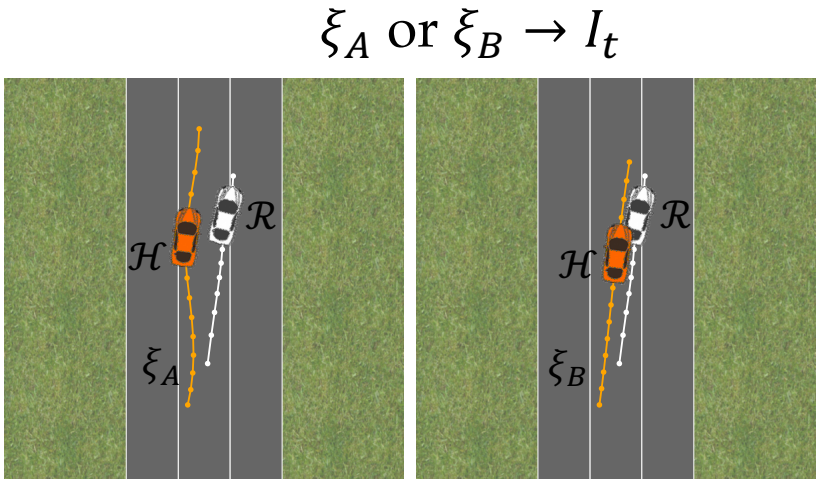
Aerospace Controls Laboratory
Department of Mechanical Engineering
Massachusetts Institute of Technology

August 11, 2017

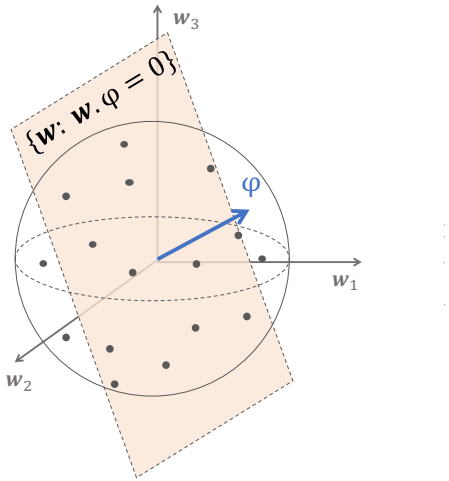
Overview

- 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: $x^{t+1} = f_{HR}(x^t, u_R^t, u_H^t)$ using continuous optimization
  - **maximize volume removed** from continuous hypothesis space of reward functions by each query

Algorithm



(a) Preference query.



(b) Query response effect.

Algorithm 1 Preference-Based Learning of Reward Functions

Input: Features ϕ, horizon N, dynamics f, *iter*
Output: Distribution of $\mathbf{w}$: $p(\mathbf{w})$
Initialize $p(\mathbf{w}) \sim \text{Uniform}(B)$, for a unit ball B
While $t <$ *iter*:
  $W \leftarrow M$ samples from AdaptiveMetropolis($p(\mathbf{w})$)
  $(x^0, \mathbf{u}_R, \mathbf{u}_H^A, \mathbf{u}_H^B) \leftarrow \text{SynthExps}(W, f)$
  $I_t \leftarrow \text{QueryHuman}(x^0, \mathbf{u}_R, \mathbf{u}_H^A, \mathbf{u}_H^B)$
  $\varphi = \Phi(x^0, \mathbf{u}_R, \mathbf{u}_H^A) - \Phi(x^0, \mathbf{u}_R, \mathbf{u}_H^B)$
  $f_\varphi(\mathbf{w}) = \min(1, I_t \exp(\mathbf{w}^\top \varphi))$
  $p(\mathbf{w}) \leftarrow p(\mathbf{w}) \cdot f_\varphi(\mathbf{w})$
  $t \leftarrow t + 1$
End for

Results

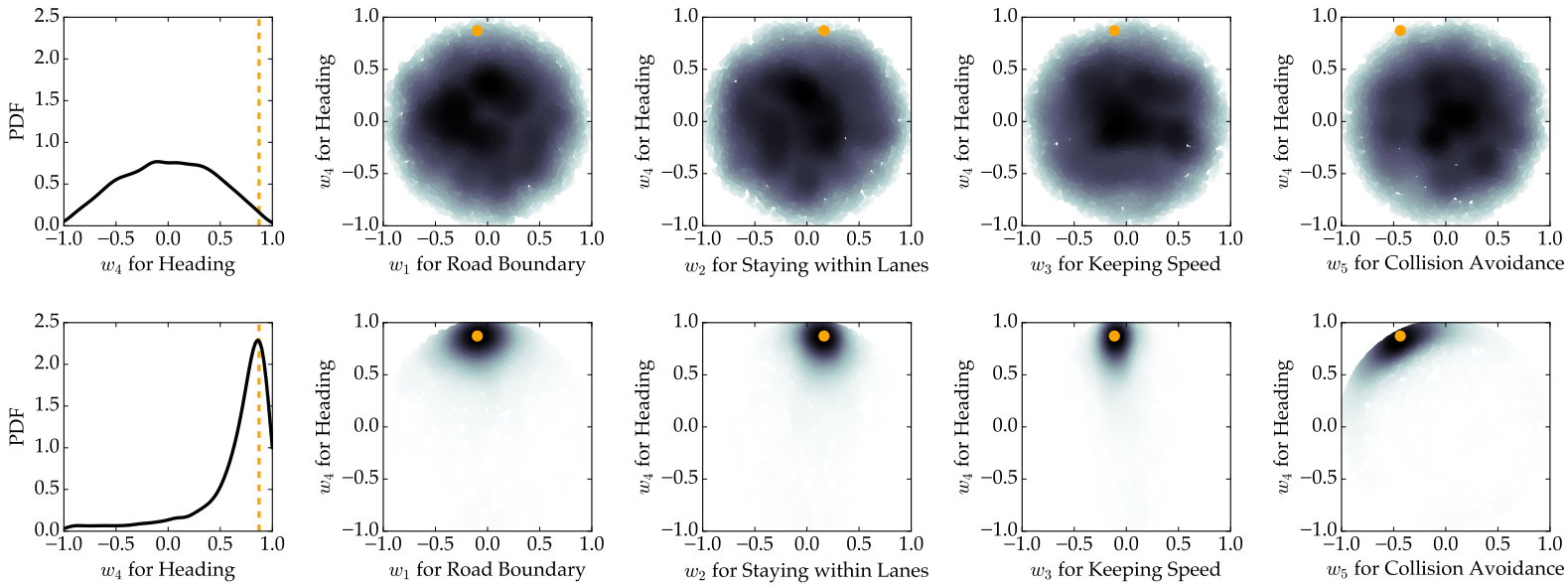
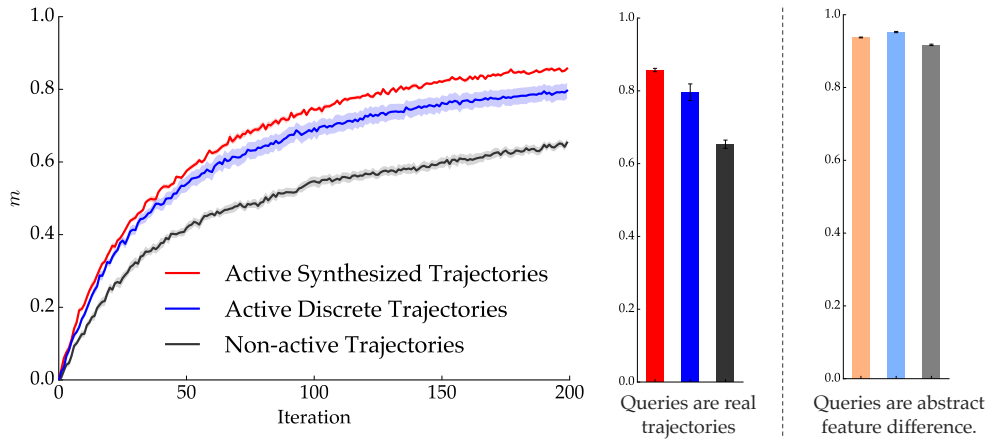


Fig. 2: Distribution of w_4, the weight for the heading feature, relative to the other features. The top plots shows the starting distribution, and the bottom plot shows the distribution at convergence. The orange dot and dotted line show the ground truth for the weights.



Discussion

- What is novel/interesting?
  - **preference-based** approach allows users to compare two trajectories in various scenarios without having to demonstrate a full trajectory (of how they would like to drive, not how they actually drive)
  - **active** enables choosing informative test cases otherwise difficult to encounter in driving scenarios; addresses the limitation of training data sparsity/informativeness
  - In-depth investigation of the effect of **synthesis** of real trajectories for dynamical systems compared to relying on a discrete set
- Limitations
  - Markov assumption is inherently built into system dynamics
  - Feature selection is not active/based on expert knowledge
  - Reward function is constrained to be linear in the feature space
  - The entire formulation is for a human H with only one other robot R
  - Formulation of $\mathbf{w}_{true}$ was unclear
  - Experiments:
    - potential functions' based reward function
    - f_{HR} assumed as a simple point mass-dynamics model

Questions?

References I

[1] Anca D Dragan Dorsa Sadigh, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions.