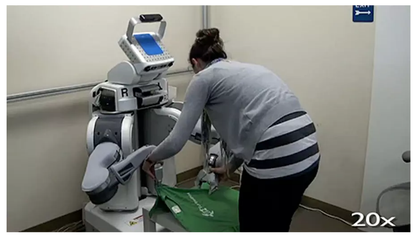
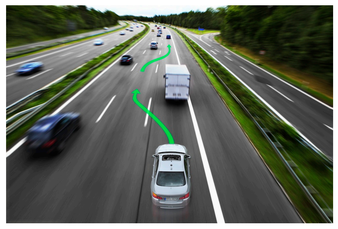
Active Preference-Based Learning of Reward Functions

Dorsa Sadigh
Anca Dragan
Shankar Sastry
Sanjit Seshia



Kinesthetic teaching difficult for high-DOF robots

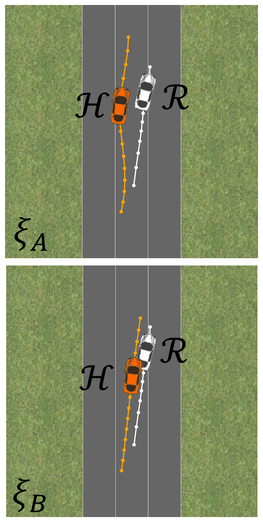


Want cars to drive differently than we drive

Hard to collect data in extreme scenarios

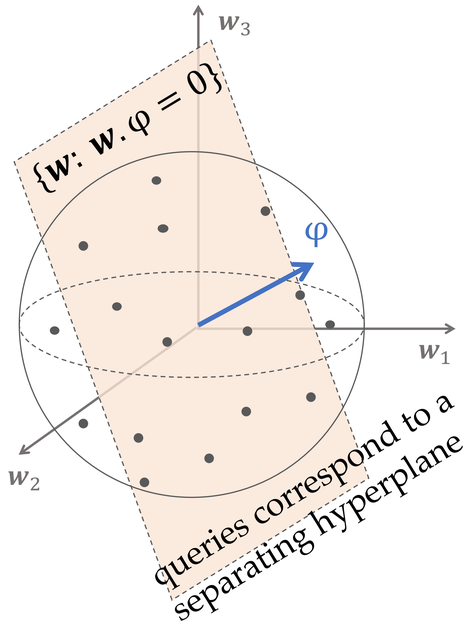
People can't always provide demonstrations of what they actually want a robot to do. We instead leverage ***comparisons*** as useful observations about the desired robot reward function.

I. Query the human



ξ_A or ξ_B ? $\rightarrow I_t$

II. Update belief over **w**



III. Actively synthesize queries

$$\max_{\varphi} \quad \min\{\mathbb{E}[1 - f_{\varphi}(w)], \mathbb{E}[1 - f_{-\varphi}(w)]\}$$

expected volume removed

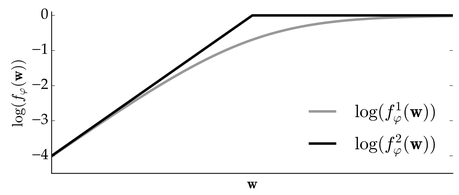
subject to $\varphi \in \mathbb{F}$

$\mathbb{F} = \{\varphi : \varphi = \Phi(\xi_A) - \Phi(\xi_B), \xi_A, \xi_B \in \Xi\}$

$$f_{\varphi}^{1}(\boldsymbol{w}) = p\,(I_t|\boldsymbol{w}) = \frac{1}{1 + \exp(-I_t \boldsymbol{w}^{\top}\varphi)}$$

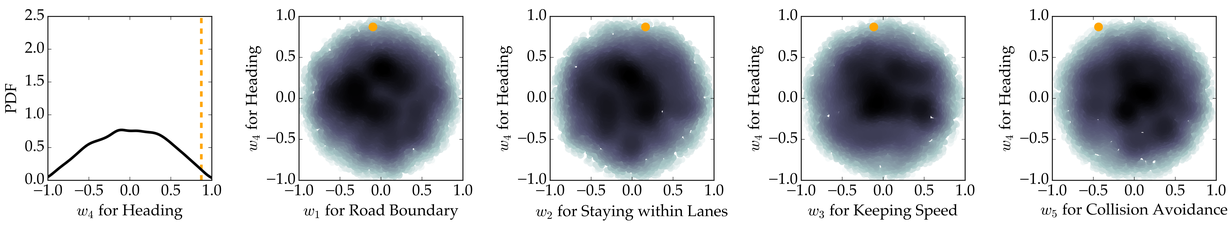
$$f_{\varphi}^{2}(\boldsymbol{w}) = \min(1, \exp(I_t \boldsymbol{w}^{\top}\varphi))$$

Log-concave update function for **w** enables bound on the number of iterations to converge

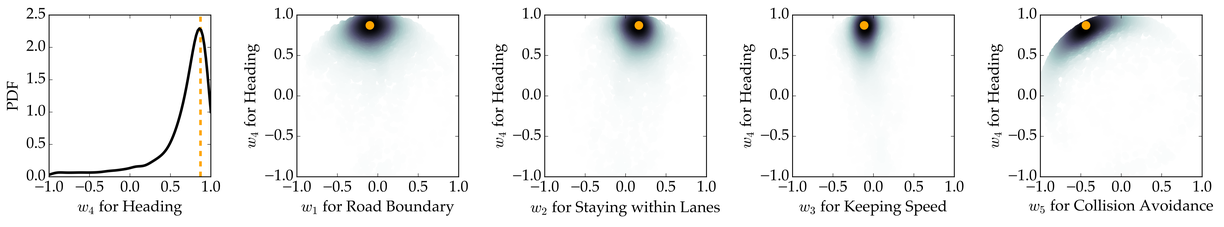


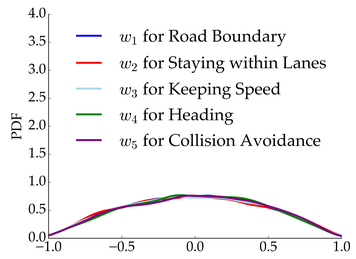
Converging to ground-truth reward

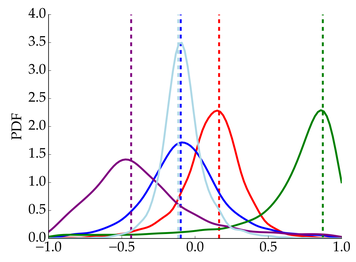
before

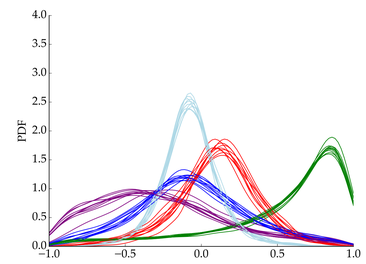


after



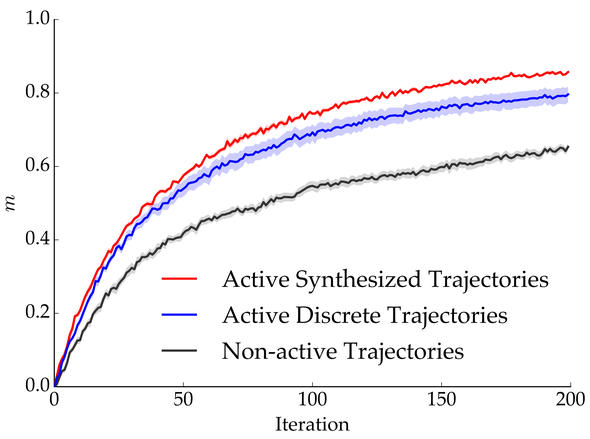




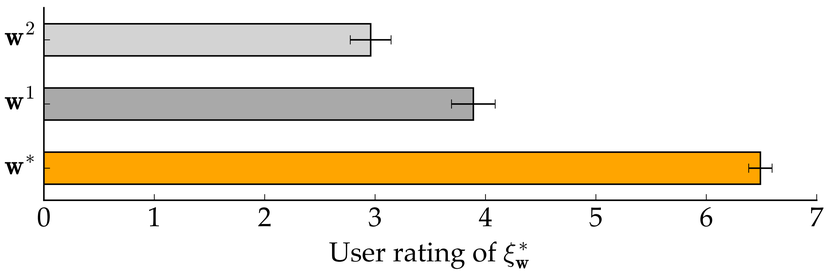


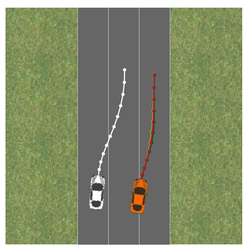
Active synthesis helps

$$m = \mathbb{E}\left[\frac{\boldsymbol{w} \cdot \boldsymbol{w}_{\text{true}}}{|\boldsymbol{w}|\,|\boldsymbol{w}_{\text{true}}|}\right]$$



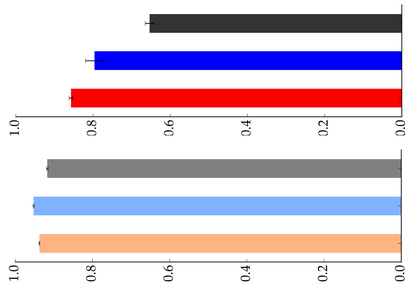
Usability study





Learned $\boldsymbol{w}^*$

… when queries = real trajectories



Queries are real trajectories

Queries are abstract feature difference