Principles of Robot Autonomy II

Preference-Based Learning, Interactions, Intent Inference
Today’s itinerary

• Learning from other sources of data (preferences, physical feedback)

• Game-Theoretic Views on Multi-Agent Interactions

• Partner Modeling: Active Info Gathering over Human’s Intent

• Partner Modeling: Learning Latent Intents

• Influencing/Stabilizing Interactions
Maximum Margin Planning (MMP)

\[ w \cdot \mu(\pi) > w \cdot \mu(\pi^*) \rightarrow \frac{w \cdot \mu(\pi^*) - b}{||w||} > 0 \]

\[ w \cdot \mu(\pi) < w \cdot \mu(\pi^*) \rightarrow \frac{w \cdot \mu(\pi) - b}{||w||} < 0 \]

\[
\max_{w} \min_{\pi} \frac{\left(\frac{w \cdot \mu(\pi^*) - b}{||w||}\right) + \left(-\frac{w \cdot \mu(\pi) - b}{||w||}\right)}{2} \\
= \max_{w} \min_{\pi} \frac{w \cdot \mu(\pi^*) - w \cdot \mu(\pi)}{||w||} \\
= \min_{w} ||w|| \text{ s.t. } w \cdot \mu(\pi^*) - w \cdot \mu(\pi) \geq 1
\]

Maximally separate the policy induced by our learned reward functions from suboptimal policies.
Today’s itinerary

• Learning from other sources of data (preferences, physical feedback)

• Game-Theoretic Views on Multi-Agent Interactions

• Partner Modeling: Active Info Gathering over Human’s Intent

• Partner Modeling: Learning Latent Intents

• Influencing/Stabilizing Interactions
Learn from Different Sources of Data

Expert demonstrations
Learn from Different Sources of Data

Expert demonstrations
Suboptimal demonstrations, observations
Language instructions, narrations

Pairwise comparisons, rankings, ordinal data

Physical Corrections

X  ✓
Learn from Different Sources of Data

- Expert demonstrations
- Suboptimal demonstrations, observations
- Language instructions, narrations
- Pairwise comparisons, rankings, ordinal data
- Physical Corrections

Example: X ✓
Learning from Decomposition

Interaction

- Wash the coffee mug
- I’m sorry - I don’t understand!
- Go to the mug and pick it up
- Go to the sink and put it inside
- Turn on the faucet
- Turn it off
- Pick up the mug

Teaching

- Wash the coffee mug
- GOTO Mug; PICKUP Mug
- GOTO Sink; PUT Mug Sink
- TOGGLE Faucet
- TOGGLE Faucet
- PICKUP Mug

Decompose into simpler steps!

Model

Historical Interaction Data (Single-User)

Online Learning
Learn from Different Sources of Data

- Expert demonstrations
- Suboptimal demonstrations, observations
- Language instructions, narrations
- Pairwise comparisons, rankings, ordinal data
- Physical Corrections
Scoring Feasibility and Optimality

policy $\pi^*$

Imitation learning

from $p_w$

sample

feasibility $w_f$

$w$ ×

optimality $w_o$

demonstrations
Learn from Different Sources of Data

Expert demonstrations
Suboptimal demonstrations, observations
Language instructions, narrations

Pairwise comparisons, rankings, ordinal data

Physical Corrections
Humans often use *multiple correlated interactions* to correct the robot.
Learn from Different Sources of Data

- Expert demonstrations
- Suboptimal demonstrations, observations
- Language instructions, narrations

Pairwise comparisons, rankings, ordinal data

Physical Corrections
Learn from Different Sources of Data

- Expert demonstrations
- Suboptimal demonstrations, observations
- Language instructions, narrations

Pairwise comparisons, rankings, ordinal data

Physical Corrections

X ✓
Leverage *comparisons* as useful observations about the desired robot reward function.
Which *grasping* trajectory do you prefer?
\[ R = w \cdot \phi \]

- \( R_A \) or \( R_B \)?
$R_A \text{ or } R_B$?

- $w \cdot \phi_A > w \cdot \phi_B$
- $w \cdot \phi_A < w \cdot \phi_B$
\[ \varphi = (\phi_A - \phi_B) \]

or \[ R_A \text{ or } R_B \]

\[ w \cdot \varphi > 0 \]

or \[ w \cdot \varphi < 0 \]
$w_1$  

$w_2$  

$w_3$
queries correspond to a separating hyperplane
queries correspond to a separating hyperplane
can be noisy!

\[ f_\varphi(w) = p(I_t|w) = \frac{1}{1 + \exp(-I_t w^\top \varphi)} \]

\[ f_\psi^1(w) = p(I_t | w) = \frac{1}{1 + \exp(-I_t w^T \psi)} \]

\[ f_\psi^2(w) = \min(1, \exp(I_t w^T \psi)) \]
$\phi$ determines how informative the query is (in expectation)
\( \varphi \) determines how informative the query is (in expectation).
queries correspond to a separating hyperplane

\( \varphi \) determines how informative the query is (in expectation)
Queries should be *actively synthesized*
Actively synthesizing queries

update function \( f_\phi(w) = \min(1, \exp(I_t w^T \phi)) \)

minimum expected volume removed

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

Subject to \( \varphi \in \mathcal{F} \)
\( \mathcal{F} = \{ \varphi : \varphi = \Phi(\xi_A) - \Phi(\xi_B), \xi_A, \xi_B \in \Xi \} \)

Active Preference-Based Learning of Reward Functions. Sadigh, et al. RSS 2017
No prior preference

Learns *heading* preferences

Learns *collision avoidance* preferences
No prior preference

Pretting \textit{green} basket over the \textit{red} one.

\textbf{Features:} \textit{max altitude, final distance to the closest basket, max horizontal range, total angular displacement}
\[ R(\xi) = \theta \cdot \phi(\xi) \]
$R(\xi_A) > R(\xi_B)$
\[ R(\xi_A) = \theta \cdot \phi(\xi_A) \]

\[ e^{-c_4d_4} \text{ where } d_4 \text{ is the final horizontal distance between the object and the center of the closest basket, and } c_4 = 3. \]

Designing features is hard.

Feature generation? Deep learning?

\[ R(\xi_A) > R(\xi_B) \]
Trajectory Features: Shot Speed, Shot Angle

\[ R(\xi_A) = \theta(\phi(\xi_{AA})) \]

\[ R(\xi_A) > R(\xi_B) \]

[Biyik, Huynh, Kochenderfer, Sadigh. RSS20]
Training: An Optimal Query with GP Reward

*GP Reward* enables the exploration of different trajectories (not just the boundaries).
Online: Final policy based on learned reward

Linear Reward

GP Reward
Actively synthesizing queries

• Going *beyond linear reward functions* and using Gaussian Processes?
• Queries that are *easy to answer* for the human? (CoRL’19)
$$\max \ I(\text{response}; \omega)$$

$$= \max \ H(\text{response}) - H(\text{response}|\omega)$$

Robot's uncertainty

Human's uncertainty
Preferences:
Easier and more accurate to use – but *gives one bit of information.*
Preferences:
Easier and more accurate to use – but *gives one bit of information*.

Demonstrations:
Rich and informative – but *noisy* and *inaccurate*.
Preferences:
Easier and more accurate to use – but *gives one bit of information*.

Demonstrations:
Rich and informative – but *noisy* and *inaccurate*.
Preferences:
Easier and more accurate to use – but gives one bit of information.

Demonstrations:
Rich and informative – but noisy and inaccurate.
Preferences:
Easier and more accurate to use – but *gives one bit of information*.

Demonstrations:
Rich and informative – but *noisy* and *inaccurate*.
Preferences:
Easier and more accurate to use – but gives one bit of information.

Demonstrations:
Rich and informative – but noisy and inaccurate.
Preferences: Easier and more accurate to use – but gives one bit of information.

Demonstrations: Rich and informative – but noisy and inaccurate.
Actively synthesizing queries

*Inverse Reinforcement Learning Policy*  
*DemPref Policy*
Learn from Different Sources of Data

- Expert demonstrations
- Suboptimal demonstrations, observations
- Language instructions, narrations

Pairwise comparisons, rankings, ordinal data

Physical Corrections
Learn from Different Sources of Data

Expert demonstrations
Suboptimal demonstrations, observations
Language instructions, narrations

Pairwise comparisons, rankings, ordinal data

Physical Corrections
Learning from Interactions

• During interaction, users *change* the robot’s behavior

• Compliance alone does not provide an *intelligent* response to these physical changes

• How can robots *learn from* and intelligently respond to physical interactions?
carry cup closer to the table
match the human’s preference
carry cup closer to the table
return to original behavior...
Robots can learn by recognizing that *interactions* are often intentional *corrections*.
Formalizing Physical Corrections

- State: $s$
- Reward: $r(s, \theta^*)$
- Correction: $a_h$
- Human model: $P(a_h | s, \theta^*)$
Formalizing Physical Corrections

- state: $s$
- reward: $r(s, \theta^*)$
- correction: $a_h$
- human model: $P(a_h \mid s, \theta^*)$
Formalizing Physical Corrections
Formalizing Physical Corrections

- State: $s$
- Reward: $r(s, \theta^*)$
- Correction: $a_h$
- Human model: $P(a_h | s, \theta^*)$

$$P(a_h | s, \theta^*)$$
arg max_{\xi \in \Xi} R(\xi, \theta)
\[ R(\xi_h, \theta^*) \geq R(\xi_r, \theta^*) \]
Value Alignment

\[ \theta^{t+1} = \arg \max_{\theta \in \Theta} P(\xi_r^0, \dots, \xi_r^t | \xi_h^0, \dots, \xi_h^t; \theta) P(\theta) \]
Value Alignment

\[ \theta^{t+1} = \arg \max_{\theta \in \Theta} \sum_{\tau=0}^{t} \ln P(x_{h}^{\tau} | x_{r}^{\tau}; \theta) + \ln P(\theta) \]
Value Alignment

\[ \theta^{t+1} = \arg \max_{\theta \in \Theta} \sum_{\tau=0}^{t} \ln P(\xi_h^\tau | \xi_r^\tau; \theta) + \ln P(\theta) \]

Assume a prior

\[ P(\theta) = \exp\left(-\frac{1}{2\beta} \|\theta - \theta^0\|^2\right) \]
Value Alignment

\[ \theta^{t+1} = \arg \max_{\theta \in \Theta} \theta \cdot \sum_{\tau=0}^{t} (\Phi(\xi_h^\tau) - \Phi(\xi_r^\tau)) - \frac{1}{2\beta} \| \theta - \theta^0 \|^2 \]
Value Alignment

Gradient Descent

\[ 0 = \nabla_\theta \left( \theta \cdot \sum_{\tau=0}^{t} (\Phi(\xi_h^\tau) - \Phi(\xi_r^\tau)) - \frac{1}{2\beta} \| \theta - \theta^0 \|^2 \right) \]
Value Alignment

Learning Rule

$$\theta^{t+1} = \theta^t + \beta(\Phi(\xi^t_h) - \Phi(\xi^t_r))$$
\[ \theta + \beta (\Phi(\xi_h) - \Phi(\xi_r)) \]
\[
\arg \max_{\xi \in \Xi} R(\xi, \theta')
\]
Task 2: stay close to the table
Task 1: keep the cup upright
Task 3: don't move over the laptop
Humans often use *multiple correlated interactions* to correct the robot.
I should go to the blue region.

Robot’s goal
I want to go to the green region, and not squeeze/stretch the bag.

Human’s goal
$P(\theta \mid A_H, \xi^0_R) \propto P(\theta) P(A_H \mid \xi^0_R, \theta)$

$A_H$: Sequence of corrections
\[ P(\theta | A_H, \xi_R^0) \propto P(\theta) P(A_H | \xi_R^0, \theta) \]
\[ \approx P(\theta) P(\xi_1^H, ..., \xi^K_H | \xi_R^0, \theta) \]

\[ A_H = \{\xi_1^H, ..., \xi^K_H\}: \text{Sequence of corrections} \]
\[
P(\theta \mid A_H, \xi_R^0) \propto P(\theta) P(A_H \mid \xi_R^0, \theta)
\]
\[
\approx P(\theta) P(\xi_H^1, \ldots, \xi_H^K \mid \xi_R^0, \theta)
\]

\[
P(\xi_H^1, \ldots, \xi_H^K \mid \xi_R^0, \theta) \propto \exp(D(\xi_H^1, \ldots, \xi_H^K, \theta))
\]

\[A_H = \{\xi_H^1, \ldots, \xi_H^K\}: \text{Sequence of corrections}\]
\[ P(\theta \mid A_H, \xi^0_R) \propto P(\theta) P(A_H \mid \xi^0_R, \theta) \]

\[ \approx P(\theta) P(\xi^1_H, \ldots, \xi^K_H \mid \xi^0_R, \theta) \]

\[
P(\xi^1_H, \ldots, \xi^K_H \mid \xi^0_R, \theta) \propto \exp(D(\xi^1_H, \ldots, \xi^K_H, \theta))
\]

\[
D(\xi^1_H, \ldots, \xi^K_H, \theta) = \sum_{t=1}^{K} \alpha^{K-t} R(\xi^t_H, \theta) - \gamma \sum_{t=1}^{K} ||a_H||_2
\]

\[ A_H = \{\xi^1_H, \ldots, \xi^K_H\}: \text{ Sequence of corrections} \]
Independent baseline

sequence model (ours)

[Li, Canberk, Losey, Sadigh. Submitted to ICRA21]
Summary so far...

- We need to learn *policies* or *reward* functions to model human preferences.

- One can leverage different sources of data (*expert demonstrations*, *pairwise comparison queries*, *rankings*, *physical feedback*) to better learn the underlying human preferences.
Today’s itinerary

• Learning from other sources of data (preferences, physical feedback)

• Game-Theoretic Views on Multi-Agent Interactions

• Partner Modeling: Active Info Gathering over Human’s Intent

• Partner Modeling: Learning Latent Intents

• Influencing/Stabilizing Interactions
Today’s itinerary

• Learning from other sources of data (preferences, physical feedback)

• Game-Theoretic Views on Multi-Agent Interactions

• Partner Modeling: Active Info Gathering over Human’s Intent

• Partner Modeling: Learning Latent Intents

• Influencing/Stabilizing Interactions
Learning from Humans

Existing research explores how robots *adapt* to humans

- Imitation learning
- Learning from demonstrations
Influencing Humans

Far less studies how robots *influence* humans
Nth order Theory of Mind
Nth order Theory of Mind
Nth order Theory of Mind

[Sadigh, Sastry, Seshia, Dragan, RSS 2016, IROS 2016, AURO 2018]
An autonomous car’s actions will affect the actions of other drivers.
Interaction as a Dynamical System

- Direct control over $\mathcal{U}_R$
- Indirect control over $\mathcal{U}_H$
Find optimal actions for the robot while accounting for the human response $u^*_H$.

Model $u^*_H$ as optimizing the human reward function $R_H$.

$$u^*_H (x, u_R) \approx \arg\max_{u_H} R_H (x, u_R, u_H)$$

$$u^*_R = \arg\max_{u_R} R_R (x, u_R, u^*_H (x, u_R))$$
Learning Driver Models

Learn Human’s reward function based on Inverse Reinforcement Learning:

\[
P(u_H | x, w) = \frac{\exp(R_H(x, u_R, u_H))}{\int \exp(R_H(x, u_R, \bar{u}_H)) d \bar{u}_H}
\]

\[
R_H(x, u_R, u_H) = w^\top \phi(x, u_R, u_H)
\]

- Features for the boundaries of the road.
- Features for staying inside the lanes.
- Features for avoiding other vehicles.

[Ziebart’ 09] [Levine’10]
Model $u^*_H$ as optimizing the human reward function $R_H$.

$$u^*_H(x, u_R) \approx \arg\max_{u_H} R_H(x, u_R, u_H)$$

Find optimal actions for the robot while accounting for the human response $u^*_R$. 

$$u^*_R = \arg\max_{u_R} R_R(x, u_R, u^*_H(x, u_R))$$