Principles of Robot Autonomy II

Preference-Based Learning, Interactions, Intent Inference
Logistics

• **March 9, Monday:** In class exam (Reinforcement Learning, Formal Methods (LTL, Synthesis), Grasping, Imitation Learning, POMDPs)

• **March 11, Wednesday:** Project Demo Day (9am – noon) – No class on March 11

• **March 15, Sunday:** Last homework is due.

• **March 20, Friday:** Paper review is due.
Today’s itinerary

- Learning from other sources of data (preferences, physical feedback)
- Planning for robots based on human models
- Modeling intent inference as a POMDP problem
- Shared Autonomy
Today’s itinerary

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

• Planning for robots based on human models

• Modeling intent inference as a POMDP problem

• Shared Autonomy
Task:

1. Reach the goal

2. Avoid the obstacle

3. Keep the arm low

\[ a^*_H = \max_{a_H} R_H(s, a_H) \]
Collect Expert Demonstrations
Max Ent Inverse Reinforcement Learning

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

\[ P(\xi) \propto \exp(R_H(\xi)) \]

\[ R_H(s, a_H) = w^\top \phi(s, a_H) \]

\[ \pi^* = \max_{\pi} R_H(s, a_H) \]

[Ziebart’ 08] [Levine’10]
Learned Policy from IRL
Leverage *comparisons* as useful observations about the desired robot reward function.
$R_A$ or $R_B$?

\[ R = w \cdot \phi \]
$R_A$ or $R_B$?

$w \cdot \phi_A > w \cdot \phi_B$

or

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

\( R_A \) or \( R_B \)?

\( \mathbf{w} \cdot \varphi > 0 \)

or \( \mathbf{w} \cdot \varphi < 0 \)
$w_1 \quad w_2 \quad w_3$
queries correspond to a separating hyperplane
queries correspond to a separating hyperplane
\[ f_\varphi(w) = p(I_t|w) = \frac{1}{1 + \exp(-I_t w^\top \varphi)} \]

can be noisy!

\( \varphi \text{ determines how informative the query is (in expectation)} \)
\( \boldsymbol{w} \cdot \varphi = 0 \) queries correspond to a separating hyperplane. \( \varphi \) determines how informative the query is (in expectation).
φ determines how informative the query is (in expectation)
Queries should be actively synthesized
Actively synthesizing queries

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

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
before

after
Actively synthesizing queries

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

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

• Generating a batch of queries?
Actively synthesizing queries

• Generating a *batch* of queries? (Biyik et al. CoRL 2018)
Generating Batches

\[
\max_{\varphi_{ib+1}, \ldots, \varphi_{(i+1)b}} "\text{minimum expected volume removal}\"
\]

2^b possible outcomes

Subject to \( \varphi_{ib+1}, \ldots, \varphi_{(i+1)b} \in F \)

\[
F = \{ \varphi: \varphi = \phi(\xi_A) - \phi(\xi_B), \xi_A, \xi_B \in \Xi_S \}
\]
Generating Batches

\[
\max_{\varphi_{ib+1}, \ldots, \varphi_{(i+1)b}} "\text{minimum expected volume removal}"
\]

Subject to \(\varphi_{ib+1}, \ldots, \varphi_{(i+1)b} \in F\)

\[F = \{\varphi: \varphi = \phi(\xi_A) - \phi(\xi_B), \xi_A, \xi_B \in \Xi_S\}\]

- Similar to joint entropy maximization.
- It is intractable.

*Discriminative Batch Mode Active Learning*

Y. Guo, D. Schuurmans. NIPS'08.
Greedy Selection
Successive Elimination

Choose the closest two pairs
Remove the one with the lowest entropy.
Theoretical Guarantees

Successive elimination always keeps the most informative query in the batch.

Ignoring the errors from

- Discretization of trajectory space,
- Sampling of weight vectors, and
- Human noise model,

successive elimination always converges.
No prior preference

Learns *heading* preferences

Learns *collision avoidance* preferences
No prior preference

Pretending *green* basket over the *red* one.

**Features:** max altitude, final distance to the closest basket, max horizontal range, total angular displacement

[Biyik, Sadigh. CoRL 2018]
Actively synthesizing queries

- Generating a *batch* of queries? (CoRL'18)
- Going *beyond linear reward functions* and using Gaussian Processes?
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

• Generating a batch of queries? (CoRL'18)
• Going beyond linear reward functions and using Gaussian Processes?
• Queries that are easy to answer for the human? (CoRL'19)
\[
\max I(\text{response}; \omega) = \max H(\text{response}) - H(\text{response} \mid \omega)
\]
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

- Generating a *batch* of queries? (CoRL’18)
- Going *beyond linear reward functions* and using Gaussian Processes?
- Queries that are *easy to answer* for the human? (CoRL’19)
- Integrating *demonstrations + preferences*? (RSS’19)
Today’s itinerary

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

• Planning for robots based on human models

• Modeling intent inference as a POMDP problem

• Shared Autonomy
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 $P(a_h \mid s, \theta^*)$
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: $P(a_h | s, \theta^*)$
Formalizing Physical Corrections

$S$  $\theta^*$

state $s$

reward $r(s, \theta^*)$

correction $a_h$

human model $P(a_h | s, \theta^*)$
Value Alignment

Human Compatible

ARTIFICIAL INTELLIGENCE
AND THE
PROBLEM OF CONTROL

Stuart Russell

“The most important book I have read in quite some time.”
Daniel Kahneman, author of THINKING, FAST AND SLOW
arg max_{\xi \in \Xi} R(\xi, \theta)
$R(\xi_h, \theta^*) \geq R(\xi_r, \theta^*)$
\[ \theta^{t+1} = \arg \max_{\theta \in \Theta} P(\xi^0_h, \ldots, \xi^t_h \mid \xi^0_r, \ldots, \xi^t_r; \theta) P(\theta) \]
Value Alignment

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

Conditionally Independent
Value Alignment

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

\[ P(\xi_h \mid \xi_r; \theta) = \frac{\exp(\theta^T \Phi(\xi_h) - \lambda \|\xi_h - \xi_r\|^2)}{\int \exp(\theta^T \Phi(\xi) - \lambda \|\xi - \xi_r\|^2) d\xi} \approx \exp(\theta^T (\Phi(\xi_h) - \Phi(\xi_r)) - \lambda \|\xi_h - \xi_r\|^2) \]

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

\[ \theta^{t+1} = \arg \max_{\theta \in \Theta} \theta \cdot \sum_{\tau=0}^{t} (\Phi(\xi^T_h) - \Phi(\xi^T_r)) - \frac{1}{2\beta} \| \theta - \theta^0 \|^2 \]
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

\[ \theta^{t+1} = \theta^t + \beta (\Phi(\xi_h^t) - \Phi(\xi_r^t)) \]
\[ \theta + \beta (\Phi(\xi_h) - \Phi(\xi_r)) \]
\[ \arg \max_{\xi \in \Xi} R(\xi, \theta') \]
Task 2: stay close to the table
Task 1: keep the cup upright
Task 3: don't move over the laptop
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)

• Planning for robots based on human models

• Modeling intent inference as a POMDP problem

• Shared Autonomy
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
An autonomous car’s actions will affect the actions of other drivers.
Interaction as a Dynamical System

- *direct* control over $u_R$
- *indirect* control over $u_H$
Interaction as a Dynamical System

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

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

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

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

• Intent inference using POMDPs
• Shared Autonomy