Inverse Reinforcement Learning for Telepresence Robots

Shimon Whiteson
Dept. of Computer Science
University of Oxford

joint work with
Kyriacos Shiarlis
& João Messias
Acknowledgements

Kyriacos Shiarlis

João Messias
Outline

• **TERESA**

• Inverse Reinforcement Learning

• Inverse Reinforcement Learning from Failure

• Results

• Next Steps
What Is Telepresence?

“Skype on a stick”
Telepresence Benefits

• Greater physical presence: “Your alter ego on wheels”

• Mobility enables spontaneous interaction
Application: Education Accessibility
Application: Remote Health Care
Application: Elderly Accessibility

Les Arcades Prevention Centre in Troyes, France

“Cafe Philo" Discussion Group
Telepresence Limitations

- *Learning curve* for human controller

- Otherwise automatic behaviours (e.g., body language) require *manual execution*

- Human controller must simultaneously make *high-level* and *low-level* decisions

- Leads to *cognitive overload*: mistakes at the low level; less attention for the high level [Tsui et al. 2011]

- Result is *poor quality social interaction*
TERESA Solution

• A new telepresence system with partial autonomy: automate low-level decision making

• Free human controller to focus on high-level decisions

• Requires social intelligence:
  • Social navigation
  • Social conversation
Experiments with Real Subjects

Les Arcades Prevention Centre in Troyes, France
Video
Interface
Cognitive Architecture

- Human Controller
- Audiovisual Processing
- Environment
- Cost Estimator
- State Estimator
- Navigator
- Body Pose Controller

- Audio & video
- Controller reactions
- People behavior
- People reactions
- People locations
- Body poses
- Motion
- State estimates
- Cost estimates
- High-level control signals

- Cognitive Architecture

- Control signals
- Body poses
- Motion
- State estimates
- Cost estimates
- People behavior
- People locations
- People reactions
- Controller reactions
- Audio & video
- Audio & video
Outline

• TERESA

• Inverse Reinforcement Learning

• Inverse Reinforcement Learning from Failure

• Results

• Next Steps
Inverse Reinforcement Learning
Reward & Value

\[ V^\pi(s) = E\left\{ \sum_{t=1}^{h} R(s_t, a_t) \mid s_1 = s \right\} \]

\[ R(s, a) = \sum_{k=1}^{K} w_k \phi_k(s, a) \]

\[ V^\pi(s) = \sum_{k=1}^{K} w_k \left( E\left\{ \sum_{t=1}^{h} \phi_k(s_t, a_t) \mid s_1 = s \right\} \right) \]

\[ = \sum_{k=1}^{K} w_k \mu^{\pi,1:h}_{k} \mid_{s_1=s}, \]

unknown weight feature function

feature expectation under \( s_1 \)
Data & Feature Expectations

Dataset: \( \mathcal{D} = \{ \tau_1, \tau_2, \ldots \tau_N \} \)

Trajectory: \( \tau_i = \langle (s_{1i}, a_{1i}), (s_{2i}, a_{2i}), \ldots, (s_{hi}, a_{hi}) \rangle \)

Empirical feature expectation:
\[
\tilde{\mu}_k^\mathcal{D} = \frac{1}{N} \sum_{\tau \in \mathcal{D}} \sum_{t=1}^{h} \phi_k(s_{t}^{\tau}, a_{t}^{\tau})
\]

Feature expectation of \( \pi \):
\[
\mu_k^\pi|_\mathcal{D} = \sum_{s \in S} P_{\mathcal{D}}(s_1 = s) \mu_k^\pi|_{s_1=s}
\]
Maximum Causal Entropy IRL

find: \( \max_{\pi} H(\mathcal{A}^h||\mathcal{S}^h) \)

subject to: \( \tilde{\mu}_k^D = \mu_k^\pi|_D \) \( \forall k \)

and: \( \sum_{a \in \mathcal{A}} \pi(s, a) = 1 \) \( \forall s \in \mathcal{S} \)

and: \( \pi(s, a) \geq 0 \) \( \forall s \in \mathcal{S}, a \in \mathcal{A} \)

where \( H \) is the causal entropy:

\[
H(\mathcal{A}^h||\mathcal{S}^h) = - \sum_{t=1}^{h} \sum_{s_{1:t} \in \mathcal{S}^t, a_{1:t} \in \mathcal{A}^t} P(a_{1:t}, s_{1:t}) \log(P(a_t|s_t))
\]

depends only on previous states & actions

Lagrangian

\[ \mathcal{L}(\pi, w) = H(A^h || S^h) + \sum_{k=1}^{K} w_k(\mu_k^{\pi}|_D - \tilde{\mu}_k^D) \]

find: \( \min_w \left\{ \max_\pi (\mathcal{L}(\pi, w)) \right\} \)

subject to: \( \sum_{a \in A} \pi(s, a) = 1 \quad \forall s \in S \)

and: \( \pi(s, a) \geq 0 \quad \forall s \in S, a \in A \)
Solving the Lagrangian

Setting $\nabla_{\pi(s_t,a_t)} \mathcal{L}(\pi, w)$ to zero implies:

$$\pi(s_t, a_t) \propto \exp \left( H(A^{t:h} \| S^{t:h}) + \sum_{k=1}^{K} w_k \mu_k^{\pi, t:h | s_t, a_t} \right)$$

Finding $\pi$ corresponds to performing soft value iteration:

$$Q_w^{\pi}(s_t, a_t) = \sum_{k=1}^{K} w_k \phi_k(s, a) + \sum_{s'} T(s, a, s') V_w(s')$$

$$V_w(s)^{soft} = \log \sum_a \exp(Q_w(s, a))$$

$$\pi(s, a) = \exp(Q_w(s, a) - V_w(s))$$
Outer Loop

Finding $\pi$ for a fixed $w$ corresponds to the inner loop of:

$$\min_w \left\{ \max_\pi \left( \mathcal{L}(\pi, w) \right) \right\}$$

Gradient descent w.r.t. $w$ in the outer loop:

$$\nabla_w \mathcal{L}(\pi, w) = \mu^\pi|_\mathcal{D} - \tilde{\mu}\mathcal{D},$$
Outline

• TERESA

• Inverse Reinforcement Learning

  • Inverse Reinforcement Learning from Failure

• Results

• Next Steps
Failed Demonstrations

• Existing IRL methods learn only from success

• IRL motivation: describing reward is hard but demonstrating success is easy

• *Failed demonstrations* also often available whenever humans learn via trial and error
Inverse Reinforcement Learning from Failure

- We extend IRL to include learning from failure
- Finds cost function that encourages behaviour that:
  - Maximises similarity to successful demos
  - Minimises similarity to failed demos

Add inequality constraints:

\[
|\tilde{\mu}_k^F - \mu_k^\pi|_F > \alpha_k \quad \forall k
\]

Non-convex!

Convex relaxation:

\[
\max_{\pi, \alpha} H(A^h || S^h) + \sum_{k=1}^K \alpha_k
\]

Intractable!
IRLF

\[
\max_{\pi, \theta, z} \ H(A^h || S^h) + \sum_{k=1}^{K} \theta_k z_k - \frac{\lambda}{2} \|\theta\|^2
\]

subject to:

\[
\mu^\pi_k |_{\mathcal{D}} = \bar{\mu}^\mathcal{D}_k \quad \forall k
\]

and:

\[
\mu^\pi_k |_{\mathcal{F}} - \bar{\mu}^\mathcal{F}_k = z_k \quad \forall k
\]

and:

\[
\sum_{a \in \mathcal{A}} \pi(s, a) = 1 \quad \forall s \in \mathcal{S}
\]

and:

\[
\pi(s, a) \geq 0 \quad \forall s \in \mathcal{S}, a \in \mathcal{A},
\]
\[ \mathcal{L}(\pi, \theta, z, w^D, w^F) = H(A^h||S^h) - \frac{\lambda}{2} \|\theta\|^2 + \sum_{k=1}^{K} \theta_k z_k + \]
\[ \sum_{k=1}^{K} w_k^D (\mu_k^\pi|_D - \tilde{\mu}_k^D) + \sum_{k=1}^{K} w_k^F (\mu_k^\pi - \tilde{\mu}_k^F|_F - z_k) \]

find: \( \min_w \left\{ \max_{\pi, \theta, z} (\mathcal{L}(\pi, w)) \right\} \)

subject to: \( \sum_{a \in \mathcal{A}} \pi(s, a) = 1 \quad \forall s \in \mathcal{S} \)

and: \( \pi(s, a) \geq 0 \quad \forall s \in \mathcal{S}, a \in \mathcal{A} \)
Solving the Lagrangian

First find critical points w.r.t. \( z \) and \( \theta \), then w.r.t. \( \pi \), yielding:

\[
\pi(s_t, a_t) \propto \exp \left( H(A_t^h \| S_t^h) + \sum_{k=1}^{K} (w_k^D + w_k^F) \mu^{\pi, t:h}_{k} | s_t, a_t \right)
\]

Yielding a new soft value iteration:

\[
Q_w(s, a)^{soft} = \sum_{k=1}^{K} (w_k^D + w_k^F) \phi_k(s, a) + \sum_{s'} T(s, a, s') V_w(s')
\]

\[
V_w(s)^{soft} = \log \sum_a \exp(Q_w(s, a))
\]

\[
\pi(s, a) = \exp(Q_w(s, a) - V_w(s))
\]
Outer Loop

Finding $\pi, \theta, z$ for a fixed $w$ corresponds to the inner loop of:

$$\min_w \left\{ \max_{\pi, \theta, z} (\mathcal{L}(\pi, w)) \right\}$$

Gradient descent w.r.t. $w^\mathcal{D}$ and $w^\mathcal{F}$ in the outer loop:

$$\nabla_{w_k^\mathcal{D}} L^*(w^\mathcal{D}, w^\mathcal{F}) = \mu_k^{\pi^*} |\mathcal{D} - \bar{\mu}_k^\mathcal{D}|,$$

$$\nabla_{w_k^\mathcal{F}} L^*(w^\mathcal{D}, w^\mathcal{F}) = \mu_k^{\pi^*} |\mathcal{F} - \bar{\mu}_k^\mathcal{F} - \lambda w_k^\mathcal{F}|.$$
Incremental IRLF

- Updates to each set of weights affect feature expectations and thus interact with each other
- Some failed demos may be better for fine tuning
- Solution: anneal $\lambda$
Outline

• TERESA
• Inverse Reinforcement Learning
• Inverse Reinforcement Learning from Failure
• Results
• Next Steps
Simulated Robot Domain
Simulated Robot Results

**Complementary:**
- E: reach target
- T: hit obstacle

**Contrasting:**
- E: reach target, avoid obstacles
- T: hit obstacle

**Overlapping:**
- E: reach target, avoid obstacles
- T: reach target, hit obstacle
Varying # of Successful Demos

![Graph showing varying number of successful demos across different training set sizes. The graph compares IRLF, IRLF (incremental), and IRL in a contrasting scenario.](image-url)
Learned Reward Functions

- Expert
- Taboo
- IRL
- IRLF (Contrasting Scenario)
Factory Domain Results

w.r.t. Expert

w.r.t. Taboo

Real Robot Data
Real-World IRL Evaluation

- Missing ground-truth reward functions
- Current practice: use qualitative and/or ad-hoc metrics

**New approach:**

- Apply IRL to D and F separately
- Resulting reward functions := ground truth
- Generate new data from corresponding policies
- Apply IRLF to new datasets
Real Robot Results

w.r.t. Expert

w.r.t. Taboo
Outline

• TERESA

• Inverse Reinforcement Learning

• Inverse Reinforcement Learning from Failure

• Results

• Next Steps
Rapidly Exploring Random Trees

Figure source: Wikipedia
Next Version of TERESA

concept D

GIRAFFE robot concepts
Extra Slides
Wizard of Oz Methodology

Human Obstacles

Interaction Target

Human Controller

Wizard of Oz
Deep IRLF

Figure source: Wulfmeier et al.
Markus Wulfmeier, Peter Ondruska, Ingmar Posner
Maximum Entropy Deep Inverse Reinforcement Learning, arXiv:1507.04888