Transfer in Inverse Reinforcement Learning for Multiple Strategies

Ajay Kumar Tanwani, Aude Billard

Learning Algorithms and Systems Laboratory (LASA)
Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

{ajay.tanwani, aude.billard}@epfl.ch
Inverse Reinforcement Learning

$s_t$ - ball position
$a_t$ - hitting speed and direction

$$(s_0, a_1), (s_1, a_2), \ldots$$

Expert Demonstrations  Trajectory Encoding  Robot Control Policy

$$a = \pi(s)$$
Inverse Reinforcement Learning

$s_t$ - ball position
$a_t$ - hitting speed and direction
$\phi(s_t)$ - distance to each hole and wall segment

$(s_0, a_1), (s_1, a_2), \ldots$

**Reward Function**

$$R(s) = w^T \phi(s)$$

**Expert Demonstrations**

$$V^\pi = w^T E(\sum_{t=0}^{T} \gamma^t \phi(s_t)|s_0 \sim \alpha, a = \pi(s_t))$$

$$= w^T \mu^\pi$$

**Robot Control Policy**

$$\pi = \arg \max_{\pi \in \Pi} V^\pi$$

$${s_{t+1} \sim P^\pi(.|s_t)}$$
Inverse Reinforcement Learning

$s_t$ - ball position  
$a_t$ - hitting speed and direction  
$\phi(s_t)$ - distance to each hole and wall segment  
$(s_0, a_1), (s_1, a_2), \ldots$

**Reward Function**

$$R(s) = w^T \phi(s)$$

**Expert Demonstrations**

**Robot Control Policy**

**Metric-of-Imitation:**

$$|V^{\pi E} - V^{\pi A}| \approx \left\| \mu^{\pi E} - \mu^{\pi A} \right\|_2$$

known expert feature-expectation vector

[Abbeel and Ng, 2004] [Syed and Schapire, 2008] [Ziebart et. al., 2008]
Learning Multiple Strategies

- Different humans have *different preferences*
- Humans can have *dynamic preferences*
- Humans *transfer knowledge* from the learned behavior
Problem Statement

- Expert strategies: \( \{ \mu^{\pi E_1}, \ldots, \mu^{\pi E_n} \} \sim \Delta(\Pi_E) \) with \( \Pi_E \) unknown

- Learn robot policies: \( \{ \mu^{\pi_1}, \ldots, \mu^{\pi_m} \} \in \Pi_A \)

\[
|V^{\pi_E} - V^{\pi_A}| \approx \|\mu^{\pi_E} - \mu^{\pi_A}\|_2 \quad \pi_A \sim \Delta(\Pi_A)
\]

![Diagram showing distribution of expert and robot policies](image.png)

Test expert strategy \( \pi_E \sim \Delta(\Pi_E) \)
Learning Multiple Strategies

- Enclose all the expert strategies with a set of optimal policies
- Extend projection algorithm [Abbeel and Ng, 2004] for multiple expert strategies
- Approximate any new expert strategy by convex combination of policies

\[ \mu_{\pi} = \sum_{i=1}^{\left| \Pi_A \right|} \lambda_i \mu_{\pi_i} \]

- Computational complexity
  - Reuse learned policies
- Number of policies
  - Store only distinct optimal policies
Optimal Policy Transfer

Optimal policy $\pi$ with transition dynamics $P^\pi$ is $\epsilon$-better policy

$$\alpha^T \left( (I - \gamma P^\pi)^{-1} - (I - \gamma P^\pi\text{init})^{-1} \right) R \geq \epsilon$$
Experimental Study

- Sink the ball in each hole same number of times as the expert does in his strategy

Learning multiple expert strategies helps to infer intention of *unseen* experts
Experimental Study

Optimal policy transfer significantly improves learning time and stored policies.
RunTime: 9.1 (-1.0)
FPS: 30
FPS: 419

Timers:
Rendering: 392us (0.01s)
Processing: 2510us (1.05s)

Modules:
IRLWorldModule: 0 - 0 us
PolicyEvalMAT: 46 - 0 us

Expert Strategy 1
Conclusions

• Incremental learning of multiple expert strategies with optimal policy transfer

Learning multiple expert strategies helps to infer intention of unseen experts

Optimal policy transfer significantly improves learning time and stored policies