Transfer in Inverse Reinforcement Learning for Multiple Strategies

Ajay Kumar Tanwani, Aude Billard

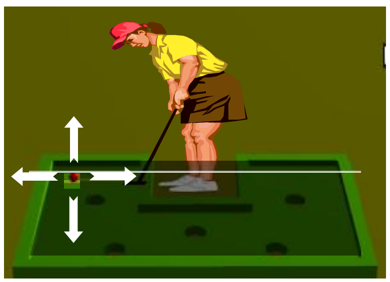
Learning Algorithms and Systems Laboratory (LASA)
Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland

{ajay.tanwani, aude.billard}@epfl.ch

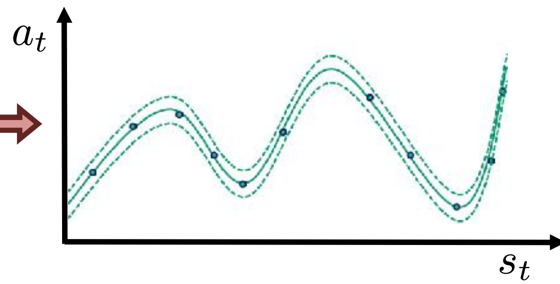
Inverse Reinforcement Learning

s_t - ball position
a_t - hitting speed and direction

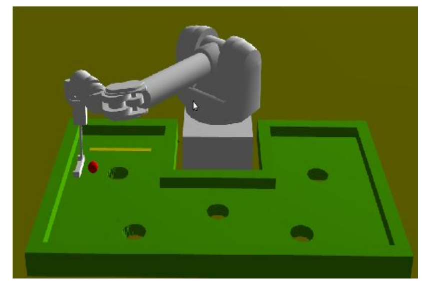
$(s_0, a_1), (s_1, a_2), \ldots$



Expert Demonstrations



Trajectory Encoding



Robot Control Policy

$a = \pi(s)$

Inverse Reinforcement Learning

s_t - ball position

a_t - hitting speed and direction

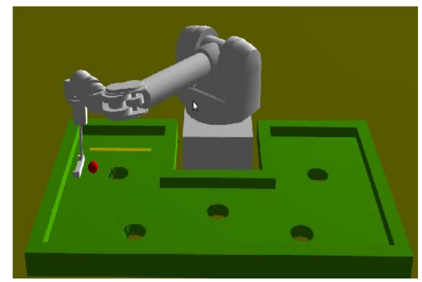
$\phi(s_t)$ - distance to each hole and wall segment

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

Expert Demonstrations

Reward Function

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



Robot Control Policy

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

$$= w^T \mu^\pi \qquad s_{t+1} \sim P^\pi(.|s_t)$$

feature-expectation vector

Inverse Reinforcement Learning

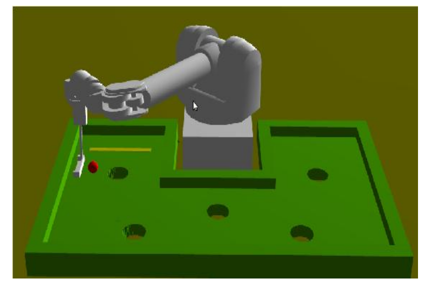
s_t - ball position
a_t - hitting speed and direction
$\phi(s_t)$ - distance to each hole and wall segment

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

Expert Demonstrations

Reward Function

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



Robot Control Policy

Metric-of-Imitation: $|V^{\pi_E} - V^{\pi_A}| \approx \|\mu^{\pi_E} - \mu^{\pi_A}\|_2$ $\quad \|w\|_1 \square 1$

known expert feature-expectation vector

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

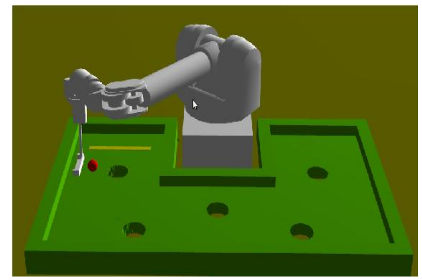
Learning Multiple Strategies

$\mu^{\pi_{E1}}$

$\vdots$

$\mu^{\pi_{En}}$

Reward Function(s)



Robot Control Policies

$\mu^{\pi_1}, \ldots, \mu^{\pi_m}$

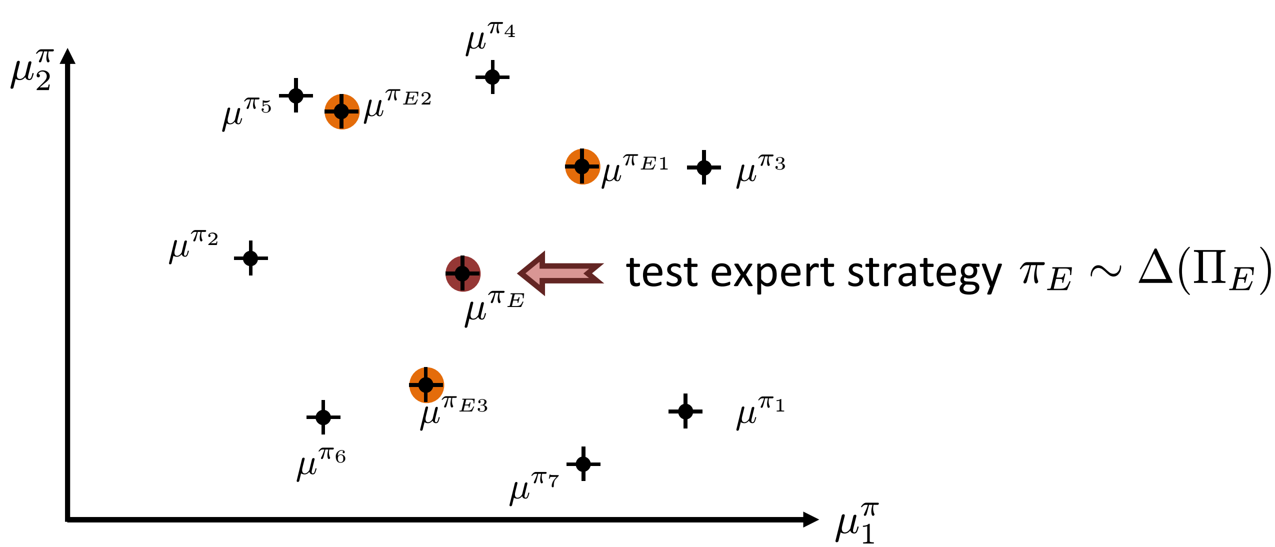
- Different humans have *different preferences*
- Humans can have *dynamic preferences*
- Humans *transfer knowledge* from the learned behavior

Problem Statement

- Expert strategies: $\{\mu^{\pi_{E1}}, \ldots, \mu^{\pi_{En}}\} \sim \Delta(\Pi_E)$ with Π_E unknown

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

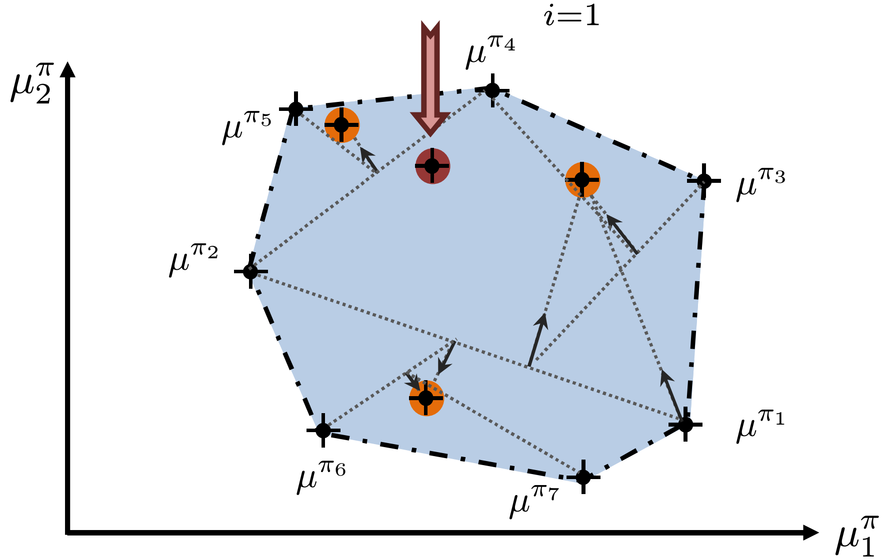
$$|V^{\pi_E} - V^{\pi_A}| \approx \|\mu^{\pi_E} - \mu^{\pi_A}\|_2 \qquad \pi_A \sim \Delta(\Pi_A)$$



Learning Multiple Strategies

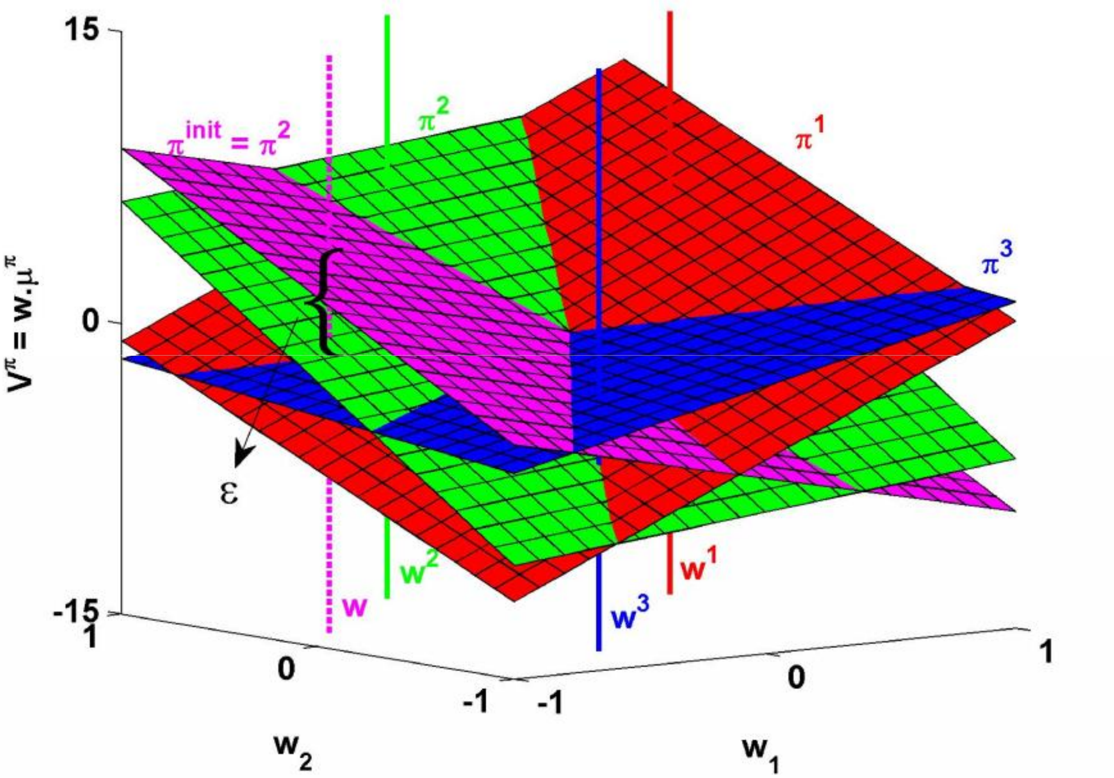
- Enclose all the expert strategies with a set of optimal policies
- Extend projection algorithm[Abbeel and Ng, 2004] for multiple expert strategies
- Approximate any new expert strategy by convex combination of policies

$$\mu^{\pi_E} = \sum_{i=1}^{|\Pi_A|} \lambda_i \mu^{\pi_i}$$



- computational complexity ⇑
  - reuse learned policies
- number of policies ⇑
  - store only distinct optimal policies

Optimal Policy Transfer



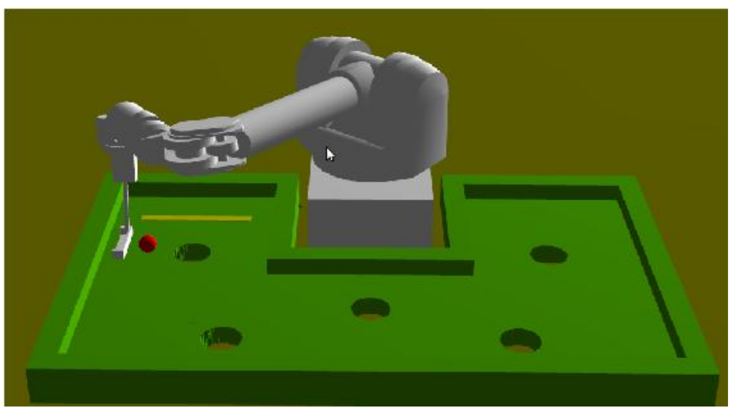
Optimal policy π with transition dynamics P^π is ϵ- better policy

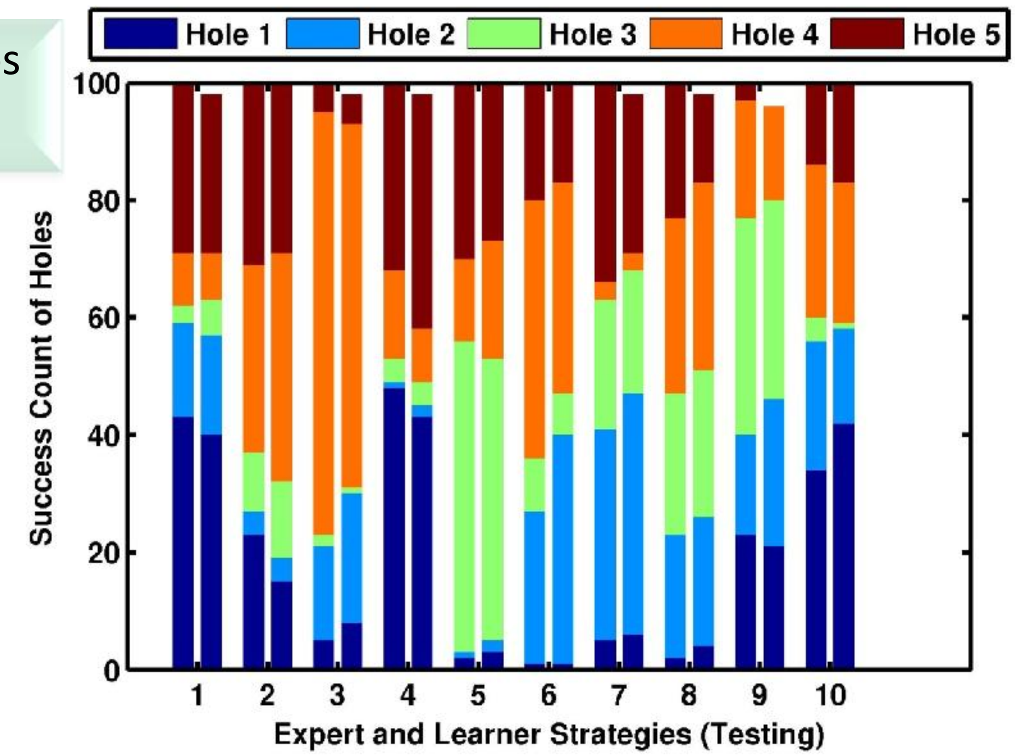
$$\alpha^T \left((I - \gamma P^\pi)^{-1} - (I - \gamma P^{\pi_{init}})^{-1} \right) R \geq \epsilon$$

Experimental Study

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

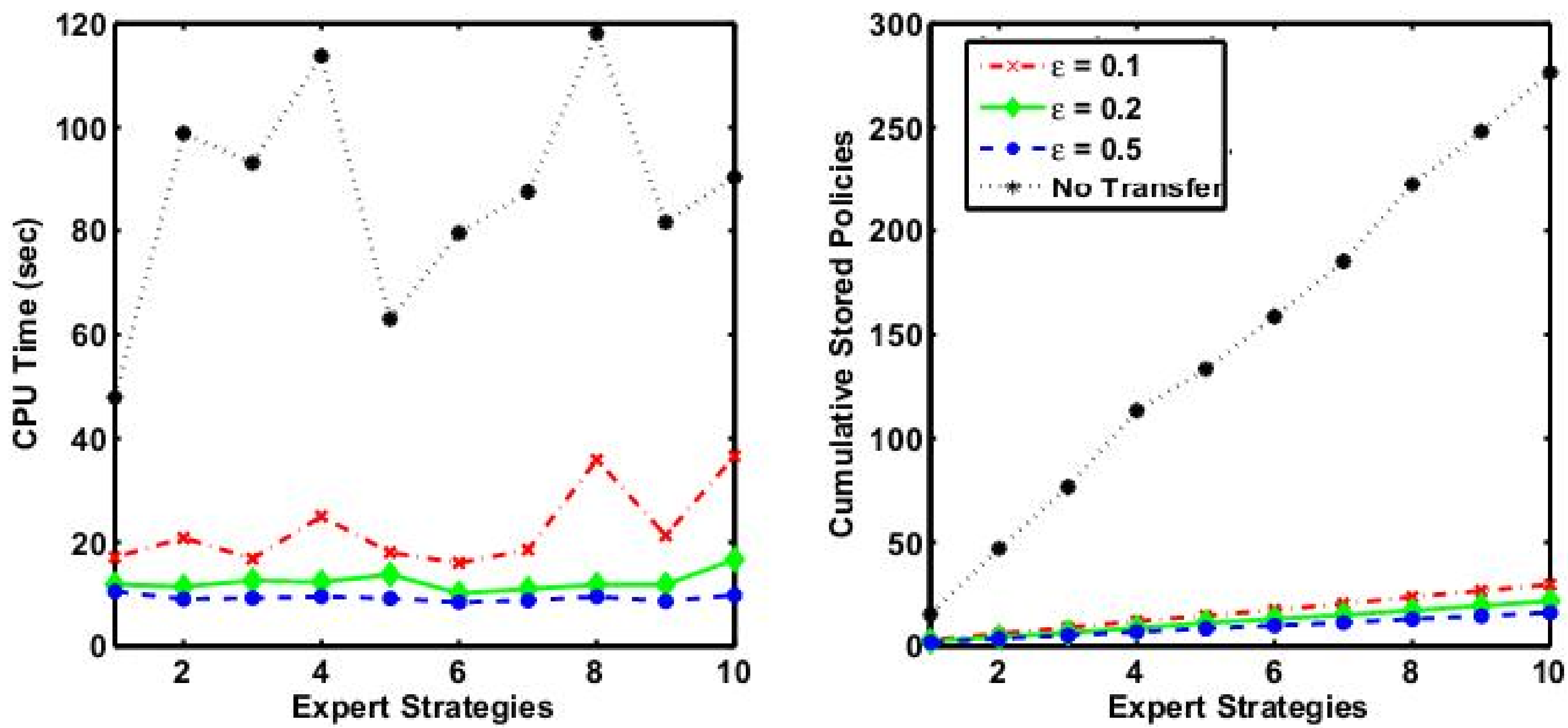
Learning multiple expert strategies helps to infer intention of *unseen* experts

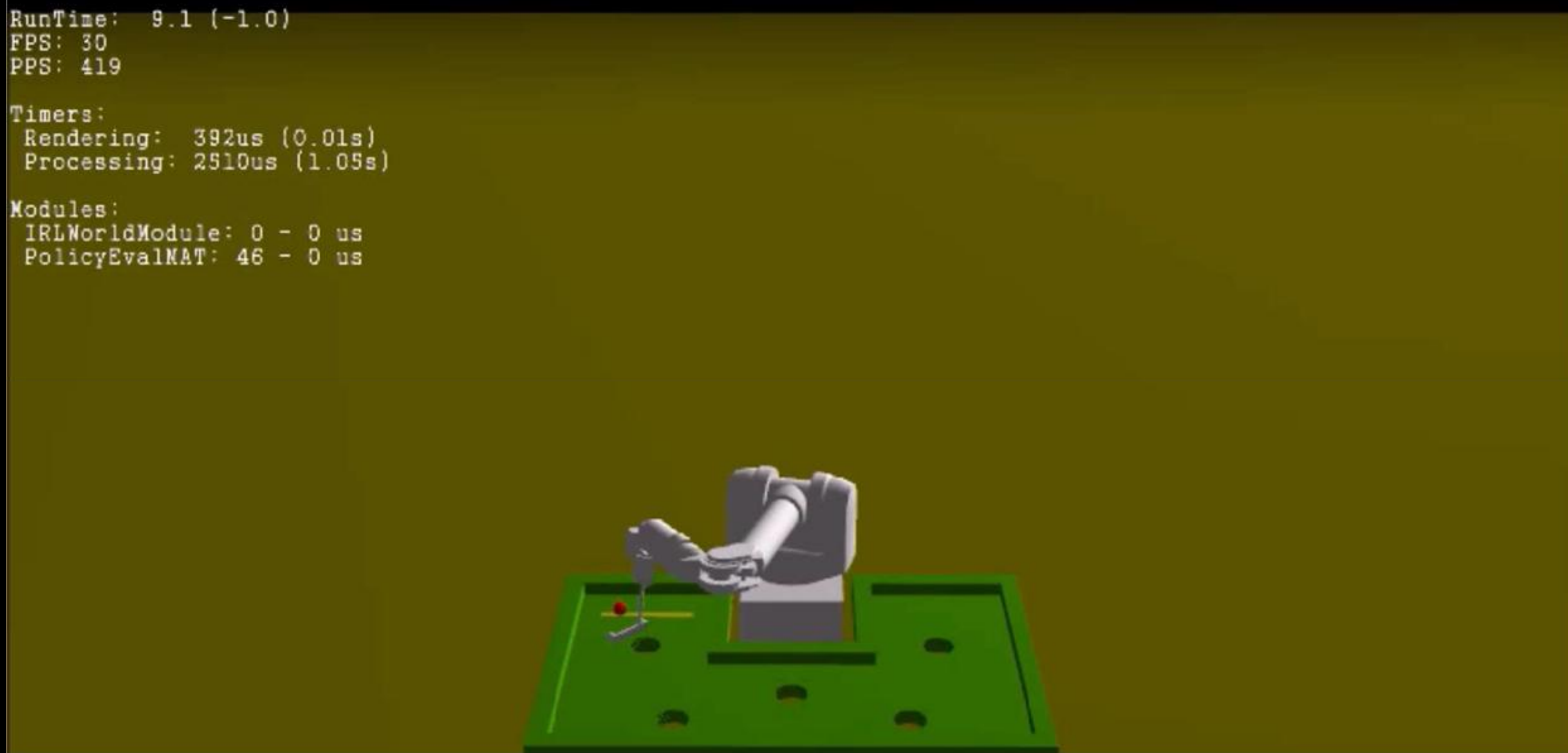




Experimental Study

Optimal policy transfer significantly improves *learning time* and *stored policies*





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*