**Maximum Entropy Semi-Supervised Inverse Reinforcement Learning**

Julien Audiffren and Michal Valko and Alessandro Lazaric and Mohammad Ghavamzadeh

CMLA, ENS CACHAN and INRIA and ADOBE RESEARCH

---

**Contribution**

MESSI (Maximum Entropy Semi-Supervised Inverse Reinforcement Learning)

- is a novel algorithm exploiting unsupervised trajectories in apprenticeship learning.
- is a principled integration between MaxEnt-IRL and semi-supervised learning techniques.
- improves the performance of MaxEnt-IRL and other SSL baselines.
- is robust to different choices of similarity function and relatively poor quality unsupervised trajectories.

---

**SSL Apprenticeship Learning**

semi-supervised inverse RL

undirected RL

imitation learning

trajectory trajectories

- Problem: expert trajectories are expensive to get or not available

- Solution: learn also from unsupervised trajectories and use the structure in the feature counts.

---

**Background**

- Markov decision process (MDP) \( (S, A, r, p) \)
  - \( S \) state space
  - \( A \) action space
  - \( r: S \rightarrow \mathbb{R} \) state reward function
  - \( p: S \times A \rightarrow \Delta(S) \) the stochastic-dynamics

- Stochastic Policy \( \pi: S \rightarrow \Delta(A) \)

- Trajectory \( \zeta = (s_0, a_0, \ldots, s_{T-1}, r_T) \) is sequence of states encountered by an agent in a given interval of time.

- Features \( f: S \rightarrow \mathbb{R}^k \)

- Feature count of a trajectory \( \zeta \) is \( f(\zeta) = \sum_{i=1}^T f(s_i) \)

- Linear reward \( r^\pi(\zeta) \) such that \( r(s) = \theta(f(s)) \).

- Expert trajectories \( \Sigma^* = \{\zeta^*\} \) from expert.

- Use a similarity function \( s \) to measure the distance \( s(\zeta, \zeta') \) between any pair of trajectories \( \{\zeta, \zeta'\} \).

- Use the pairwise penalty forces similar trajectories to have similar rewards

\[
R(\theta(\Sigma^*)) = \frac{1}{2(1+\nu)} \sum_{\zeta \in \Sigma^*} s(\zeta, \zeta^*) (\theta(f(\zeta^*)) - \theta(f(\zeta)))^2
\]

- New optimization problem penalizes the likelihood of \( \theta \) by the similarity in unsupervised trajectories

\[
\theta^* = \arg \max \left( L(\theta(\Sigma^*)) - \lambda R(\theta(\Sigma^*)) \right)
\]

---

**MESSI**

- Integration of unsupervised trajectories in MaxEnt-IRL using a penalty function reflecting the geometry of the trajectories, similar to [Erikan and Altun, 2009], but on the dual problem to preserve a low computational complexity.

- Set of expert trajectories \( \Sigma^* = \{\zeta^*_i\} \) and unsupervised trajectories \( \Sigma = \{\zeta_i\}_{i=1}^T \).

- Use a similarity function \( s \) to measure the distance \( s(\zeta, \zeta') \) between any pair of trajectories \( \{\zeta, \zeta'\} \).

- Use the pairwise penalty forces similar trajectories to have similar rewards

\[
R(\theta(\Sigma^*)) = \frac{1}{2(1+\nu)} \sum_{\zeta \in \Sigma^*} s(\zeta, \zeta^*) (\theta(f(\zeta^*)) - \theta(f(\zeta)))^2
\]

- New optimization problem penalizes the likelihood of \( \theta \) by the similarity in unsupervised trajectories

\[
\theta^* = \arg \max \left( L(\theta(\Sigma^*)) - \lambda R(\theta(\Sigma^*)) \right)
\]

---

**Experimental settings**

- Two Benchmarks: Grid World [Abbeel and Ng, 2004] and Highway Driving [Szydlo et al., 2008].

- Unlabeled trajectories are drawn from three different distributions over policies
  - \( P_r = P(\theta^*) \) (expert)
  - \( P_l = P(\theta_l) \) (average quality)
  - \( P_d = P(\theta_d) \) (very different reward)

- MESSIMAX: MESSI with only near expert unlabeled trajectories (upper bound for MESSI performance).

- Parameters: MESSI is evaluated with respect to \( \theta_{max} \)

---

**Results**

- Number of iterations. MESSI improves at each iteration (unlike SSIRL). Advantage of MESSIMAX is clear starting from the beginning.

- Proportion of good unsupervised trajectories. Non-relevant distribution (as \( P_\mu \)) make MESSI performs worse than MaxEnt-IRL. However, improves quickly with even a few worthy trajectories.

---

**Comparison with EM baseline**

- SSIRL Cannot be compared to SSIRL [Valko et al., 2012] because it does not have a stopping criterion.

- EM Comparison to semi-supervised baseline inspired by EM [Zhu, 2005].

---

**Discussion**

- Not semi-supervised classification: Unsupervised trajectories come from the expert herself, another expert(s), near-expert, by agents maximizing different reward functions, or noisy data.

- Similarity functions are more efficient when hand-crafted to fit the problem, but still works for baseline like RBF.

- Improves MaxEnt IRL when the similarity function is meaningful and the distribution of unsupervised trajectories is informative.

---

**References**


