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.

Background

- **Markov decision process (MDP)** $\langle S, A, r, p\rangle$
  - S state space
  - A action space
  - $r : S \to \mathbb{R}$ state reward function
  - $p : S \times A \to \Delta(S)$ is the stochastic dynamics
- **Stochastic Policy** $\pi : S \to \Delta(A)$
- **Trajectory** $\zeta = (s_1, a_1, \ldots, a_{T-1}, s_T)$ is sequence of states encountered by an agent in a given interval of time.
- **Features** $f : S \to \mathbb{R}^d_+$
- **Feature count** of a trajectory ζ is $\mathbf{f}_\zeta = \sum_{t=1}^T f(s_t)$
- **Linear reward** $\exists \boldsymbol{\theta} \in \mathbb{R}^d$ such that $r(s) = \langle \boldsymbol{\theta}, f(s)\rangle$.
- **Expert trajectories** $\Sigma^* = \{\zeta^* \text{ from expert}\}$, i.e. realizations of the expert policy.
- **The objective of apprenticeship learning is to recover the reward followed by the expert.**
- **Ill-posed problem:** infinite possible solutions, some uninteresting or bad.
- **Solution:** Propose a reward, solve the RL problem, compare the trajectory obtained with the expert one, and adjust the reward. Iterate until convergence.

MaxEnt IRL [Ziebart et al., 2008]

Idea: Maximize the log-likelihood of $\boldsymbol{\theta}$ given Σ^*

$$\boldsymbol{\theta}^* = \arg\max_\theta \sum_{\boldsymbol{\xi}\in\Sigma^*} \log P(\boldsymbol{\xi}|\theta)$$

At each iteration, repeat

- **Compute** the probability of trajectories through maximum entropy principle
$$P(\zeta|\boldsymbol{\theta}) \approx \frac{\exp(\boldsymbol{\theta}^\mathsf{T} f_\zeta)}{Z(\boldsymbol{\theta})}\prod_{t=1}^T p(s_{t+1}|s_t, a_t),$$
- **Deduce** the expected feature count of the current candidate.
$$\mathbf{f}_t = \sum_\zeta P(\zeta|\boldsymbol{\theta}_t)\mathbf{f}_\zeta = \sum_{s\in S}\rho_t(s)\mathbf{f}(s)$$
- **Update** the value of $\boldsymbol{\theta}$ with a gradient descent step.

Trade-off : MESSI is based on the original MaxEnt IRL and do not use the Causal Entropy version to preserve a low computational complexity.

References

[Abbeel and Ng, 2004] Abbeel, P. and Ng, A. (2004).
Apprenticeship Learning via inverse reinforcement learning.
In *Proceedings of the 21st International Conference on Machine Learning.*

[Erkan and Altun, 2009] Erkan, A. and Altun, Y. (2009).
Semi-Supervised Learning via Generalized Maximum Entropy.
In *Proceedings of JMLR Workshop*, pages 209–216. New York University.

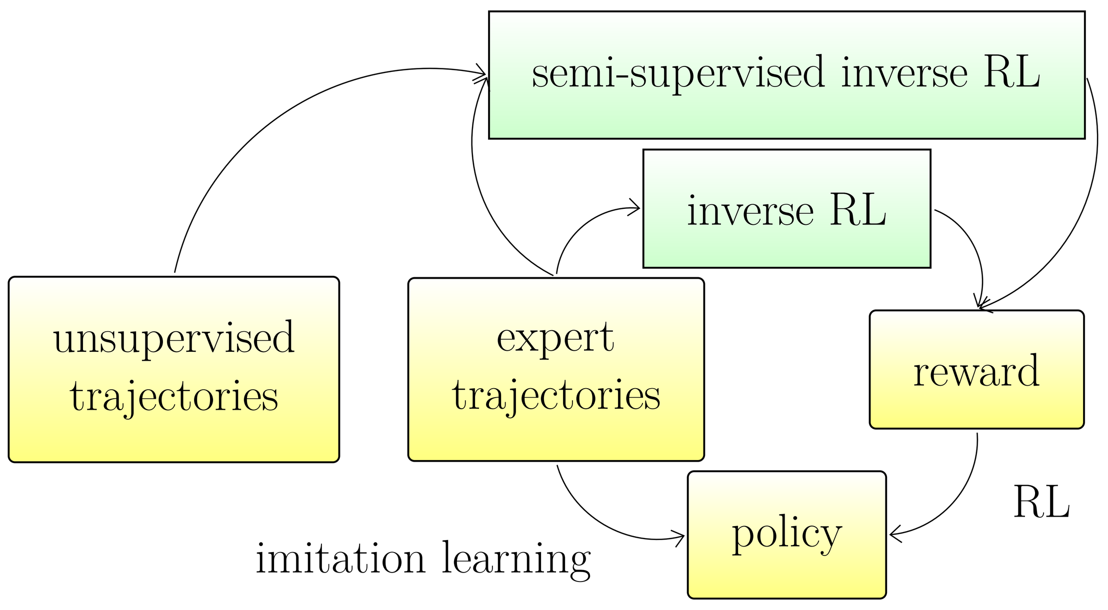
[Syed et al., 2008] Syed, U., Schapire, R., and Bowling, M. (2008).
Apprenticeship Learning Using Linear Programming.
In *Proceedings of the 25th International Conference on Machine Learning*, pages 1032–1039.

[Valko et al., 2012] Valko, M., Ghavamzadeh, M., and Lazaric, A. (2012).
Semi-Supervised Apprenticeship Learning.
In *Proceedings of the 10th European Workshop on Reinforcement Learning*, volume 24, pages 131–241.

[Zhu, 2005] Zhu, X. (2005).
Semi-supervised learning literature survey.
Technical Report 1530, Computer Sciences, University of Wisconsin-Madison.

[Ziebart et al., 2008] Ziebart, B., Maas, A., Bagnell, A., and Dey, A. (2008).
Maximum Entropy Inverse Reinforcement Learning.
In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence.*

SSL Apprenticeship Learning



- **Problem:** expert trajectories are expensive to get or not available
- **Solution:** learn also from unsupervised trajectories and use the structure in the feature counts.

MESSI

- **Integration** of unsupervised trajectories in MaxEnt-IRL using a penalty function reflecting the geometry of the trajectories, similar to [Erkan and Altun, 2009], but on the **dual problem** to preserve a low computational complexity.
- Set of expert trajectories $\Sigma^* = \{\zeta_i\}_{i=1}^l$ and unsupervised trajectories $\widetilde{\Sigma} = \{\zeta_j\}_{j=1}^u$.
- **Use a similarity function** s to measure the distance $s(\zeta, \zeta')$ between any pair of trajectories (ζ, ζ').
- **The pairwise penalty** forces similar trajectories to have similar rewards
$$R(\boldsymbol{\theta}|\Sigma) = \frac{1}{2(l+u)}\sum_{\zeta,\zeta'\in\Sigma} s(\zeta,\zeta')\left(\boldsymbol{\theta}^\mathsf{T}(\mathbf{f}_\zeta - \mathbf{f}_{\zeta'})\right)^2$$
- **New optimization problem** penalizes the likelihood of θ by the similarity in unsupervised trajectories
$$\boldsymbol{\theta}^* = \operatorname*{argmax}_{\boldsymbol{\theta}}\ \left(L(\boldsymbol{\theta}|\Sigma^*) - \lambda R(\boldsymbol{\theta}|\Sigma)\right)$$

The MESSI Algorithm

1: **Input:** l expert trajectories $\Sigma^* = \{\zeta_i^*\}_{i=1}^l$, u unsupervised trajectories $\widetilde{\Sigma} = \{\zeta_j\}_{j=1}^u$, similarity function s, number of iterations T, constraint $\theta_{\max}$, regularizer λ_0
2: **Initialization:**
3: Compute $\{\mathbf{f}_{\zeta_i^*}\}_{i=1}^l$, $\{\mathbf{f}_{\zeta_j}\}_{j=1}^u$ and $\mathbf{f}^* = 1/l\sum_{i=1}^l \mathbf{f}_{\zeta_i^*}$
4: Generate a random reward vector $\boldsymbol{\theta}_0$
5: **for** $t = 1$ **to** T **do**
6: Compute policy π_{t-1} from $\boldsymbol{\theta}_{t-1}$ (backward pass)
$$\pi(a|s;\boldsymbol{\theta}) = \sum_{\zeta\in\Sigma_{s,a}} P(\zeta|\boldsymbol{\theta})$$
7: Compute feature counts $\mathbf{f}_{t-1}$ of π_{t-1} (forward pass)
$$\mathbf{f}_t = \sum_\zeta P(\zeta|\boldsymbol{\theta}_t)\mathbf{f}_\zeta = \sum_{s\in S}\rho_t(s)\mathbf{f}(s)$$
8: Update the reward vector as follows
$$\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} + (\mathbf{f}^* - \mathbf{f}_{t-1}) + \frac{\lambda_0}{\theta_{\max}(l+u)}\sum_{\zeta,\zeta'\in\Sigma} s(\zeta,\zeta')\left(\boldsymbol{\theta}_{t-1}^\mathsf{T}(\mathbf{f}_\zeta - \mathbf{f}_{\zeta'})\right)^2.$$
9: If $\|\boldsymbol{\theta}_t\|_\infty > \theta_{\max}$, project back by $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t\frac{\theta_{\max}}{\|\boldsymbol{\theta}_t\|_\infty}$
10: **end for**

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 is 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.

Experimental settings

- **Two Benchmarks** : Grid World [Abbeel and Ng, 2004] and Highway Driving [Syed et al., 2008].
- **Unlabeled trajectories** are drawn from three different distributions over policies
  - $P_{u^*} = P(\cdot|\boldsymbol{\theta}^*)$ *(expert)*
  - $P_1 = P(\cdot|\boldsymbol{\theta}_1)$ *(average quality)*
  - $P_2 = P(\cdot|\boldsymbol{\theta}_2)$ *(very different reward)*
- **MESSIMAX:** MESSI with only near expert unlabeled trajectories (upper bound for MESSI performance)
- **Parameters:** MESSI is evaluated with respect to θ_{max}, λ, the number of iteration, the distribution over unlabeled trajectories

Results

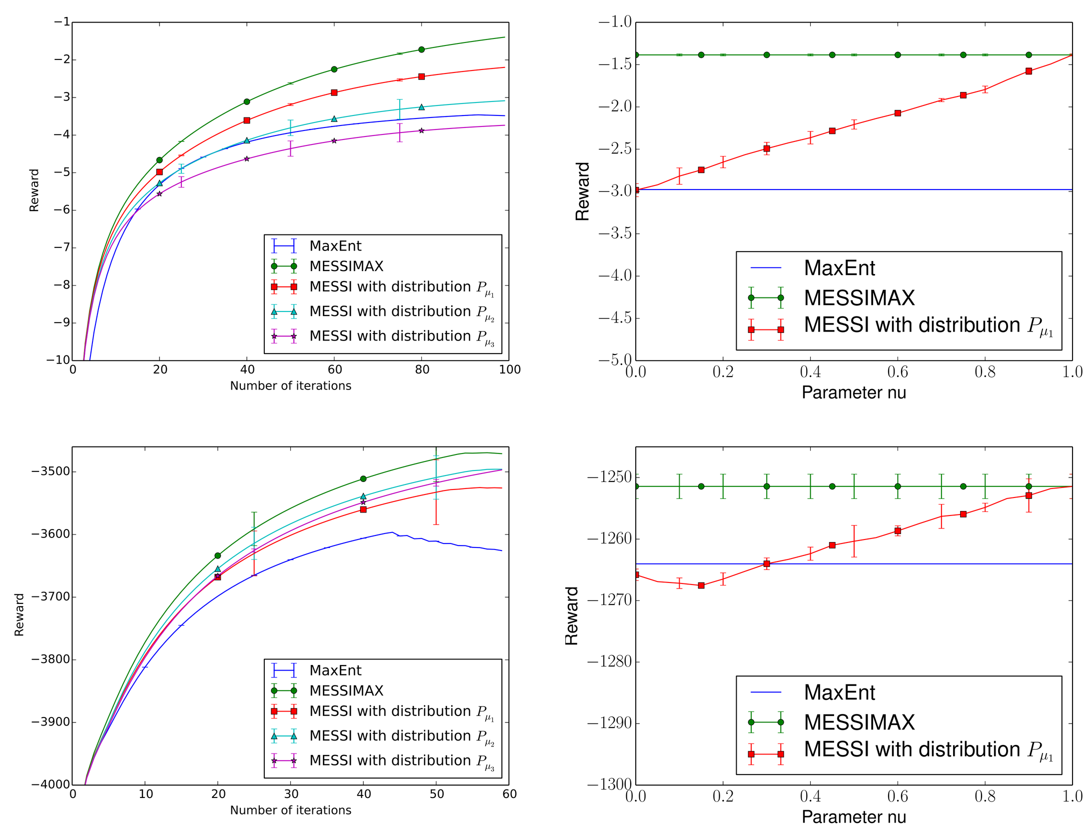


Figure 1: Results as a function of number of iterations (left), the distribution μ of the unsupervised data (right), a of the MaxEnt, MESSIMAX and MESSI on the Highway driving dataset (up) and the gridworld (down) dataset.

- **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_{μ_3}) 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] :
  - *Maximization step* : using belief on nature of trajectories, solve one iteration of MaxEnt.
  - *Expectation step*: Given the current reward, update the belief on the nature of the trajectories.

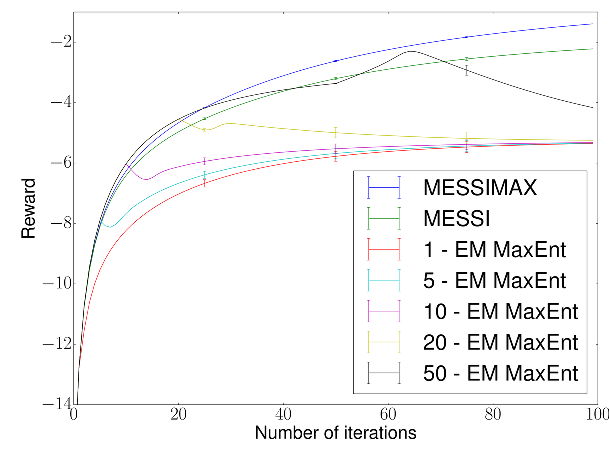


Figure 2: Comparison between MESSI and EM

Results: For all the respective frequencies of Maximization and Expectation steps, EM performs worse than MESSI (Fig. 2).

IJCAI INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE BUENOS AIRES, 2015