Maximum Entropy Semi Supervised Inverse Reinforcement Learning

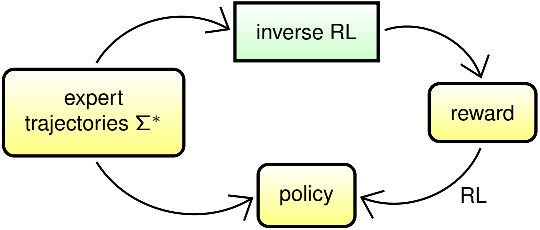
a.k.a. MESSI

J. Audiffren M. Valko A. Lazaric M. Ghavamzadeh

Apprenticeship learning (IRL)

Main idea

Learning from demonstrated behaviour (provided by an **expert**, or *teacher*).



Why ?

In many settings, the reward is very complex to define. Ex : Highway driving

$\rightarrow$ How can we force the agent to respect the highway code while traveling as fast as possible ?

The Maximum entropy principle [Ziebart & al, 2009.]

The Setting :
MDP with linear reward : $(\mathcal{S}, \mathcal{A}, T, f, \theta_*)$

→ $f : \mathcal{S} \to \mathbb{R}^k$ features of each state.

→ linear hypothesis : $\mathcal{R}(s) = \theta_*^T f_s$

The IRL problem reduces to find the θ_* encoding the reward the expert abide by.

Maximum entropy principle : Idea : Maximize the log-likelihood of the probability of the expert trajectories

$$\theta_* = \arg\max_\theta \sum \log \mathbf{P}(\xi_i^* | \theta)$$

→ Backward Pass : Given a reward, get expected feature frequency, knowing that a path with greater value is exponentially preferred.

→ Forward Pass : Update the reward with a gradient descent step.

Limitations of IRL

$\rightarrow$ IRL generally needs a lot of expert's data (and experts are very expensive).

$\rightarrow$ Is a feature never reached by an expert a bad feature ?

$\rightarrow$ What about near optimal expert ?

The Semi Supervised approach

Add to this problem

- → A set of unlabeled data $\Sigma = (\xi_i)_i$
- → A set of expert data $\Sigma^* = (\xi_i)_i$
- → A similarity function :

$$s : \Sigma \cup \Sigma^* \times \Sigma \cup \Sigma^* \mapsto [0, 1]$$

Smoothness assumption

Intrinsically similar trajectories have similar rewards.

The smoothness assumption becomes

$$s(\xi, \xi') \approx 1 \quad \Rightarrow \quad \mathcal{R}(\xi) \approx \mathcal{R}(\xi')$$

Semi Supervised IRL

Main idea

Add a regularization term to the maximum entropy objective to enforce the smoothness of the reward w.r.t. the similarity function.

$\rightarrow$ Regularization : pairwise penalty

$$PR(\theta|\Sigma) = \frac{1}{2|\Sigma \cup \Sigma^*|} \sum_{\xi,\xi'} s(\xi,\xi') \left(\theta^T (f_\xi - f_{\xi'})\right)^2$$

So the objective function becomes :

$$\arg\max_\theta \sum_{\xi \in \Sigma^*} \log\left(\mathbf{P}(\xi|\theta)\right) - \lambda PR(\theta|\Sigma)$$

The algorithm

Algorithm 1 pseudocode for MESSI

Input : Σ^* experts trajectories, Σ unlabelled trajectories, similarity function s, iteration number T, constraint $\theta_{max} > 0$, regularizer λ, random initial reward θ_0

for $t = 1$ to T **do**

Compute the expected feature count $\hat{f_{\theta_t}}$ with reward θ_t (Value iteration)

Update θ :

$$\theta_{t+1} = \theta_t + (f_* - \hat{f_{\theta_t}}) + \frac{\lambda}{\theta_{max}|\Sigma \cup \Sigma^*|} \sum_{\xi,\xi'} s(\xi,\xi')\theta^T(f_\xi - f_{\xi'})^2$$

If $\|\theta_{t+1}\| > \theta_{max}$, then

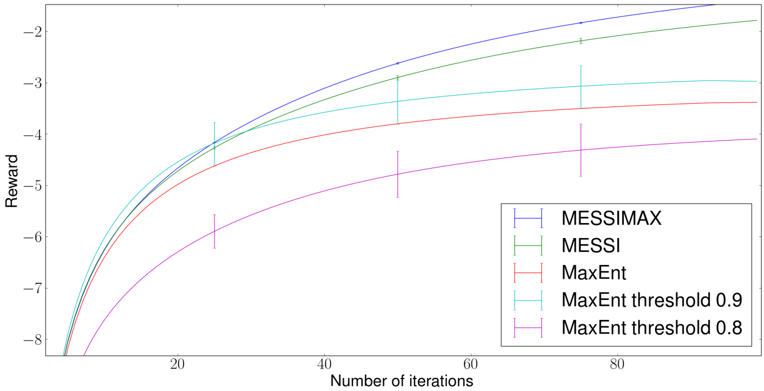
$$\theta_{t+1} = \frac{\theta_{t+1}\theta_{max}}{\|\theta_{t+1}\|}$$

end for

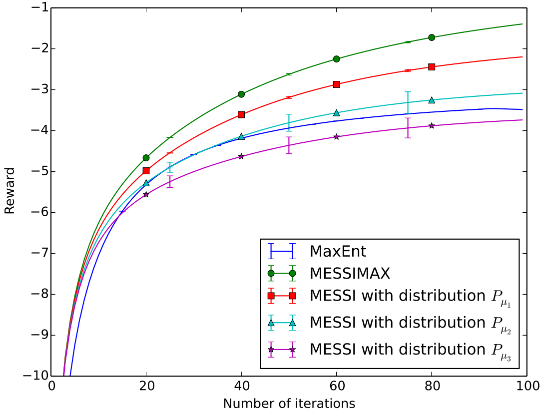
Does it really work ?

Experiments

Highway driving benchmark



$\longrightarrow$ Better than MaxEnt, even by aggregating near optimal trajectories



→ Works even for low quality unlabeled data and for generic similarity function

Conclusion

Strengths

$\rightarrow$ First implementable SSIRL approach.

$\rightarrow$ Works on small and average-sized problems

$\rightarrow$ Work with generic similarity function (ex : RBF)

$\rightarrow$ Work with average quality unlabeled data.

Weaknesses

$\rightarrow$ Do not scale to big MDP problems. (Future Work : solve the MDP with a model free approach)

$\rightarrow$ Many approximation to be computationally tractable...

$\rightarrow$ ...and thus no theoretical guarantees (for now).

Come to see our poster at panel 38 for more details !