MESSI: Maximum Entropy Semi-Supervised Inverse Reinforcement Learning

Julien Audiffren
CMLA, ENS Cachan
julien.audiffren@cmla.ens-cachan.fr

Michal Valko
INRIA Lille – Nord Europe
michal.valko@inria.fr

Alessandro Lazaric
INRIA Lille – Nord Europe
alessandro.lazaric@inria.fr

Mohammad Ghavamzadeh
INRIA/Adobe Research
mohammad.ghavamzadeh@inria.fr

Introduction. The most common approach to solve a sequential decision-making problem is to formulate it as a Markov decision process (MDP). This process requires the definition of a reward function, but in many applications, such as driving or playing tennis, it is easier and more natural to learn how to perform such tasks by observing an expert's demonstration, rather than by definition of a reward function. The task of learning from an expert is called *apprenticeship learning*. A powerful and relatively novel approach to apprenticeship learning (AL) is to formulate it as an *inverse reinforcement learning* (IRL) problem [6]. The basic idea is to assume that the expert is trying to optimize an MDP and to derive an algorithm for learning the task demonstrated by the expert [6, 1]. This approach has been shown to be effective in learning non-trivial tasks such as inverted helicopter flight control [5], ball-in-a-cup [2], and driving on a highway [1, 4].

In the IRL approach to AL, we assume that several trajectories generated by an expert are available and the *unknown* reward function optimized by the expert can be specified as a linear combination of a number of state features. In many applications, in addition to the expert's trajectories, we may have access to a large number of trajectories that are not necessarily performed by an "expert". For example, in learning to drive, we may ask an expert driver to demonstrate a few trajectories and use them in an AL algorithm to mimic her behavior. At the same time, we may record trajectories from many other drivers for which we cannot assess their quality and that may or may not demonstrate an expert-level behavior. We will refer to them as *unsupervised trajectories* and to the task of learning with them as *semi-supervised apprenticeship learning* following Valko et al. [8] who combine the IRL approach of Abbeel and Ng [1] with semi-supervised SVMs. However, unlike in classification, we do not regard the unsupervised trajectories as being a mixture of expert and non-expert classes. This is because the unsupervised trajectories might have been generated by the expert herself, by another expert(s), by near-expert agents, by agents maximizing different reward functions, or simply they can be some noisy data. The objective of IRL is to find the reward function that expert trajectories maximize, and thus, semi-supervised apprenticeship learning cannot be considered as a special case of semi-supervised classification.

Maximum Entropy Semi-Supervised Inverse Reinforcement Learning. We propose the algorithm MESSI (MaxEnt Semi-Supervised IRL, see algorithm 1) to address the challenge above by combining the MaxEnt-IRL approach of Ziebart et al. [9] with SSL. MESSI integrates the unsupervised trajectories in a principled way such that it performs better than MaxEnt-IRL. For this purpose, we assume that the learner is provided with a set of expert trajectories $\Sigma^* = \{\zeta_i^*\}_{i=1}^l$ and a set of unsupervised trajectories $\widetilde{\Sigma} = \{\zeta_j\}_{j=1}^u$. We also assume that a function s is provided to measure the similarity $s(\zeta, \zeta')$ between any pair of trajectories (ζ, ζ'). We define the *pairwise penalty* R as

$$R(\boldsymbol{\theta}|\Sigma) = \frac{1}{|\Sigma|} \sum_{\zeta,\zeta' \in \Sigma} s\left(\zeta, \zeta'\right) \left(\boldsymbol{\theta}^\intercal (\mathbf{f}_\zeta - \mathbf{f}_{\zeta'})\right)^2, \tag{1}$$

where $\Sigma = \Sigma^* \cup \widetilde{\Sigma}$, and $\mathbf{f}_\zeta$ and $\mathbf{f}_{\zeta'}$ are the feature counts for trajectories $\zeta, \zeta' \in \Sigma$, and finally $(\boldsymbol{\theta}^\intercal (\mathbf{f}_\zeta - \mathbf{f}_{\zeta'}))^2 = (\bar{r}_{\boldsymbol{\theta}}(\zeta) - \bar{r}_{\boldsymbol{\theta}}(\zeta'))^2$ is the difference in rewards accumulated by the two trajectories

Algorithm 1 MESSI - MaxEnt SSIRL

Input: Set of l expert trajectories $\Sigma^* = \{\zeta_i^*\}_{i=1}^l$, set of u unsupervised trajectories $\widetilde{\Sigma} = \{\zeta_j\}_{j=1}^u$, similarity function s, number of iterations T, constraint $\theta_{\max}$, regularizer λ_0
Initialization:
Compute $\{\mathbf{f}_{\zeta_i^*}\}_{i=1}^l$, $\{\mathbf{f}_{\zeta_j}\}_{j=1}^u$ and $\mathbf{f}^* = 1/l \sum_{i=1}^l \mathbf{f}_{\zeta_i^*}$ and generate a random reward vector $\boldsymbol{\theta}_0$
for $t = 1$ **to** T **do**
1. Compute policy π_{t-1} from $\boldsymbol{\theta}_{t-1}$ (Solving the MDP)
2. Compute feature counts $\mathbf{f}_{t-1}$ of π_{t-1} (forward pass of MaxEnt)
3. Update the reward vector by doing a gradient descent step on (2)
4. If $\|\boldsymbol{\theta}_t\|_\infty > \theta_{\max}$, project back by $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_t \frac{\theta_{\max}}{\|\boldsymbol{\theta}_t\|_\infty}$

end for

w.r.t. the reward vector $\boldsymbol{\theta}$. The purpose of the pairwise penalty is to penalize reward vectors $\boldsymbol{\theta}$ that assign very different rewards to similar trajectories (as measured by $s(\zeta, \zeta')$). We then follow the framework of Erkan and Altun [3] to integrate the regularization R into the learning objective. This leads to the following optimization problem :

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \ \left(L(\boldsymbol{\theta}|\Sigma^*) - \lambda R(\boldsymbol{\theta}|\Sigma)\right), \tag{2}$$

where L is the the log-likelihood of $\boldsymbol{\theta}$ w.r.t. the expert's trajectories and λ is a parameter trading off between L and the coherence with the similarity between the provided trajectories both in $\widetilde{\Sigma}$ and Σ^*. Although hand-crafted similarity functions usually perform better, our experiments show that even the simple RBF $s\left(\zeta, \zeta'\right) = \exp(-\|\mathbf{f}_\zeta - \mathbf{f}_{\zeta'}\|^2/2\sigma)$, (where σ is the bandwidth) is an effective similarity for the feature counts.

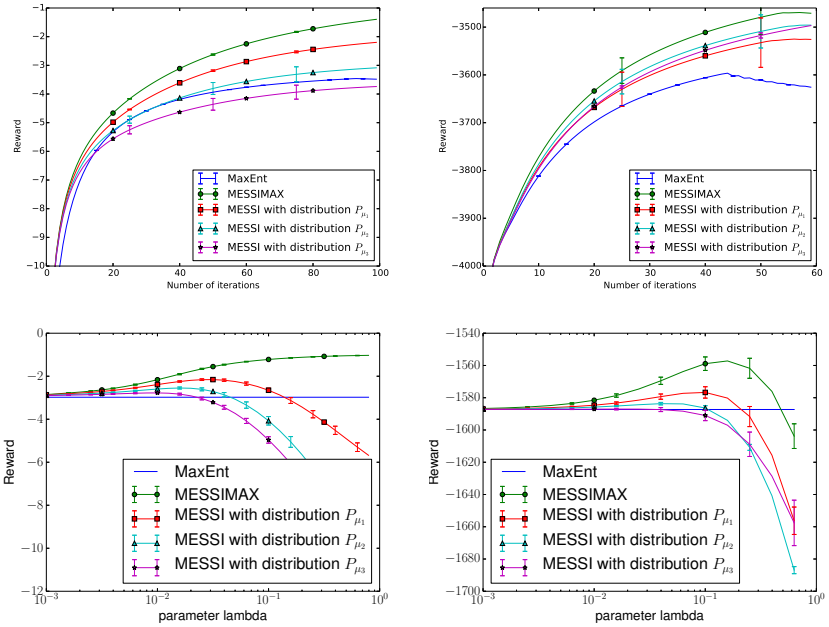

Figure 1: Results as a function of number of iterations (up) and the parameter lambda (down) of the MaxEnt, MESSIMAX (MESSI with all unsupervised trajectories drawn from expert) and MESSI (with different distributions of unsupervised trajectories) algorithms on the Highway driving (left) and the gridworld (right) dataset.

Experimental Results.

Our experiments shows that MESSI takes advantage of unsupervised trajectories and can perform better and more efficiently than MaxEnt-IRL in the highway driving problem of Syed et al. [7] and the grid-world domain in Abbeel and Ng [1] (see fig 1).

References

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

[2] A. Boularias, J. Kober, and J. Peters. Relative Entropy Inverse Reinforcement Learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, volume 15, pages 182–189, 2011.

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

[4] S. Levine, Z. Popovic, and V. Koltun. Nonlinear Inverse Reinforcement Learning with Gaussian Processes. In *Advances in Neural Information Processing Systems 24*, pages 1–9, 2011.

[5] A. Ng, A. Coates, M. Diel, V. Ganapathi, J. Schulte, B. Tse, E. Berger, and E. Liang. Inverted Autonomous Helicopter Flight via Reinforcement Learning. In *International Symposium on Experimental Robotics*, 2004.

[6] A. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning*, pages 663–670, 2000.

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

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

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