Learning from demonstrations: Is it worth estimating a reward function?

ECML 2013

Bilal Piot[1,2], Matthieu Geist[1] and Olivier Pietquin[1,2]
firstname.lastname@supelec.fr

1 Equipe IMS/MaLIS (Supélec), France
2 UMI 2958 (GeorgiaTech-CNRS)

1. Introduction

This poster provides a comparative study between Inverse Reinforcement Learning (IRL) and Apprenticeship Learning (AL). IRL and AL are two frameworks, using Markov Decision Processes (MDP), which are used for the imitation learning problem where an agent tries to learn from demonstrations of an expert. In the AL framework, the agent tries to learn the expert policy whereas in the IRL framework, the agent tries to learn a reward which can explain the behavior of the expert. This reward is then optimized to imitate the expert. One can wonder if it is worth estimating such a reward, or if estimating a policy is sufficient. This quite natural question has not really been addressed in the literature right now. We provide partial answers, both from a theoretical and empirical point of view.

2. Notations

- The set of probability distributions over $\mathcal{X}$ is noted $\Delta_{\mathcal{X}}$.
- A MDP is a tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma\}$
- A stationary and Markovian policy $\pi \in \Delta_{\mathcal{A}}^{\mathcal{S}}$ represents the behavior of an agent.
- $v_{\mathcal{R}}^{\pi}(s) = \mathbb{E}[\sum_{t\geq 0}\gamma^t \mathcal{R}(s_t, a_t)|s_0 = s, \pi]$.
- A policy $\pi_{\mathcal{R}}^*$ is said optimal if its value function $v_{\mathcal{R}}^*$ satisfies $v_{\mathcal{R}}^* \geq v_{\mathcal{R}}^{\pi}$ for any policy π and component wise.

3. AL and IRL

In the AL framework, the apprentice, given some observations of the expert policy π_E, tries to learn a policy π_A which is as good as the expert policy according to the unknown reward $\mathcal{R}$ that the expert is trying to optimize. The apprentice tries to find a policy π_A such that the quantity: $\mathbb{E}_{\nu}[v_{\mathcal{R}}^{\pi_E} - v_{\mathcal{R}}^{\pi_A}]$ is the lowest possible, where $\nu \in \Delta_S$. In the IRL framework, given some observations of the expert policy π_E, the apprentice is trying to learn $\hat{\mathcal{R}}$ such that $\pi_E \approx \pi_{\hat{\mathcal{R}}}^*$.The apprentice is trying to learn a reward $\hat{\mathcal{R}}$ such that the quantities $\mathbb{E}_{\nu}[v_{\hat{\mathcal{R}}}^{\pi_{\hat{\mathcal{R}}}^*} - v_{\hat{\mathcal{R}}}^{\pi_E}]$ or $\mathbb{E}_{\nu}[v_{\mathcal{R}}^{\pi_E} - v_{\mathcal{R}}^{\pi_{\hat{\mathcal{R}}}^*}]$ are the lowest possible.

4. Theoretical Study

[**Infinite Horizon bound**] We assume that some demonstrations examples $D_E = (s_i, a_i)_{\{1\leq i\leq N\}}$ where $a_i \sim \pi_E(.|s_i)$ are available. Let define the following concentration coefficient: $C_{\nu} = (1-\gamma)\sum_{t\geq 0}\gamma^t c_{\nu}(t)$ where $\forall t \in \mathbb{N}, c_{\nu}(t) = \max_{s\in\mathcal{S}} \frac{(\nu^T P_{\pi_E}^t)(s)}{\nu(s)}$. Let π_C be the classifier policy (trained on the data set D_E to imitate the expert policy π_E). Then $\forall \mathcal{R} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$:

$$\mathbf{E}_{\nu}[v_{\mathcal{R}}^{\pi_E} - v_{\mathcal{R}}^{\pi_C}] \leq \frac{2C_{\nu}\|\mathcal{R}\|_{\infty}}{(1-\gamma)^2}\epsilon_C. \tag{1}$$

[**Finite Horizon bound**] Let π_E be the expert non-stationary and Markovian expert policy, D_E a set of N trajectories with $s_1^i \sim \nu \in \Delta_{\mathcal{S}}$ and π_C the policy learnt by the H classifiers, then:

$$\mathbb{E}_{\nu}[v_{1,\mathcal{R}}^{\pi_E} - v_{1,\mathcal{R}}^{\pi_C}] \leq \min(2\sqrt{\epsilon_C}H^2, 4\epsilon_C H^3 + \delta_{\pi_E})\|\mathcal{R}\|_{\infty}, \tag{2}$$

where $\delta_{\pi_E} = \frac{\mathbb{E}_{\nu}[v_{1,\mathcal{R}}^* - v_{1,\mathcal{R}}^{\pi_E}]}{\|\mathcal{R}\|_{\infty}}$ represents the sub-optimality of the expert and H the horizon.

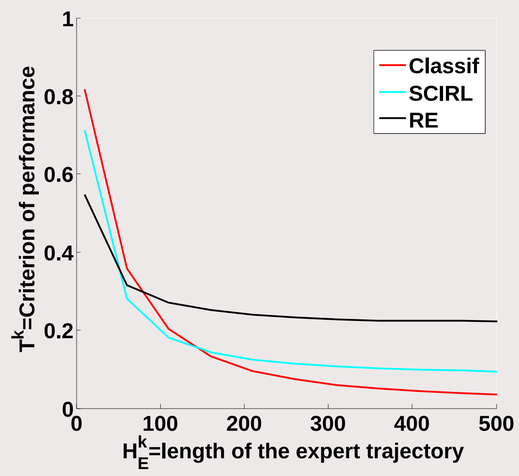
[**SCIRL bound**] The reward outputted by the SCIRL algorithm is $\mathcal{R}_C$. Then, the performance bound for this algorithm is:

$$0 \leq \mathbb{E}_{\rho_E}[v_{\mathcal{R}_C}^* - v_{\mathcal{R}_C}^{\pi_E}] \leq \frac{C_f}{(1-\gamma)}\left(\frac{2\|\mathcal{R}_C\|_{\infty}\epsilon_C}{1-\gamma} + \bar{\epsilon}_Q\right), \tag{3}$$

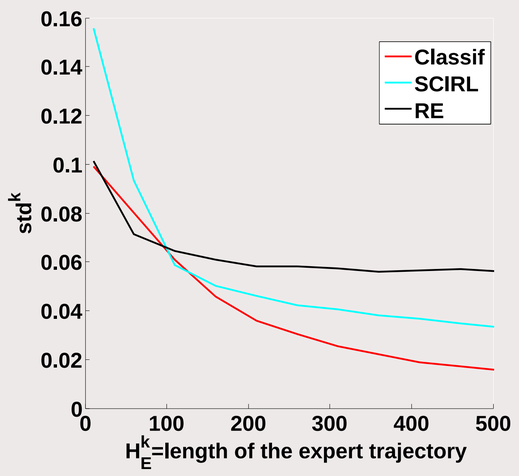
With $C_f = (1-\gamma)\sum_{t\geq 0}\gamma^t c_f(t)$ where $\forall t \in \mathbb{N}, c_f(t) = \max_{s\in\mathcal{S}} \frac{(\rho_E^T P_{\pi_{\mathcal{R}_C}^*}^t)(s)}{\rho_E(s)}$. $\bar{\epsilon}_Q$ is a measure of the error estimation of the feature expectation.

5. Empirical Study

The Garnet problems are a class of randomly constructed finite MDPs meant to be totally abstract while remaining representative of the kind of finite MDPs that might be encountered in practice. The experiment consists in resolving the learning from demonstrations problem on hundreds of Garnets with IRL and classification algorithms for different type of rewards.

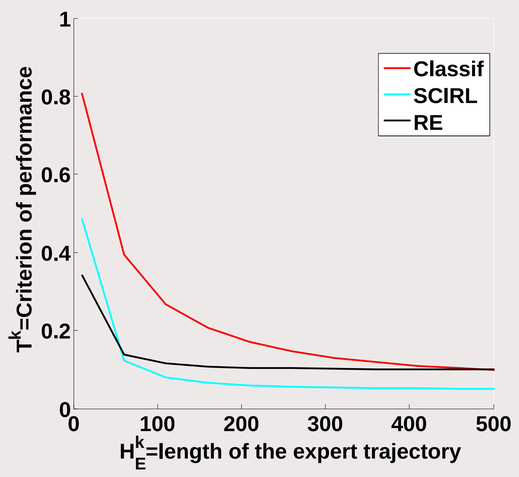


(a) Performance

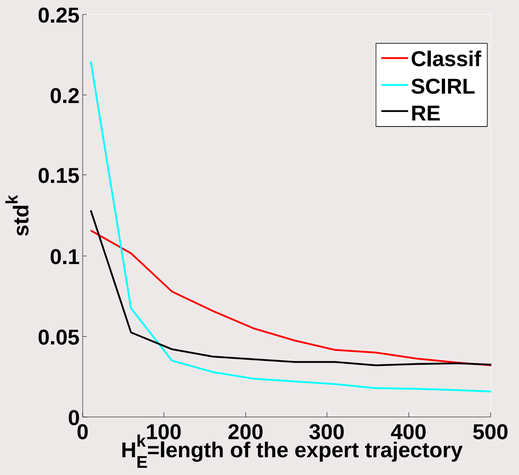


(b) Standard deviation

Figure: Garnets experiment: normally distributed reward.



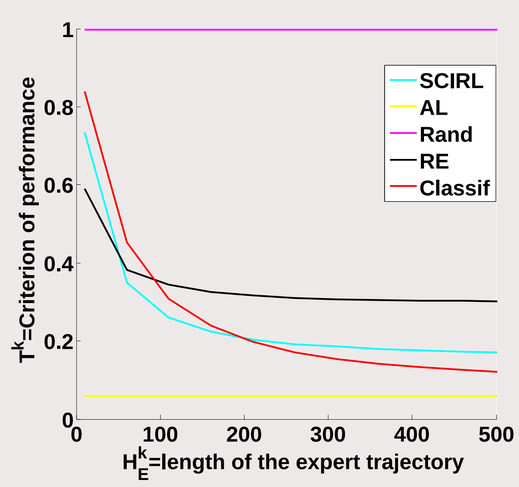
(a) Performance



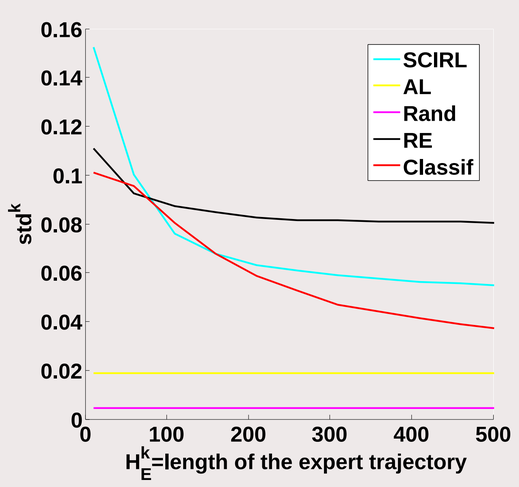
(b) Standard deviation

Figure: Garnets experiment: state-only-dependent reward.

The second experiment consists in perturbing the dynamics of the underlying MDP in order to see how the algorithms can adapt to this perturbation. Here, we see that the well behavior of the algorithms depends strongly on the reward shape.

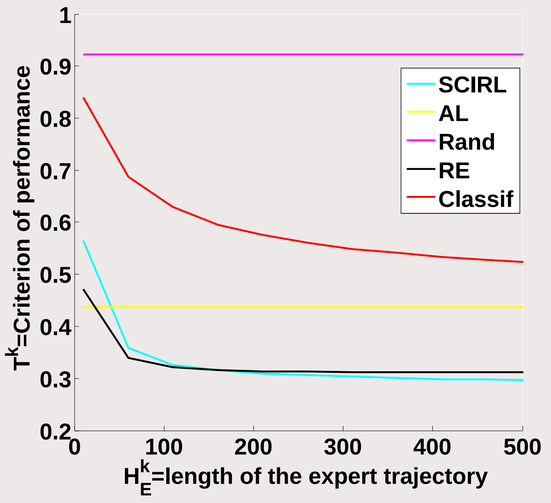


(a) Performance

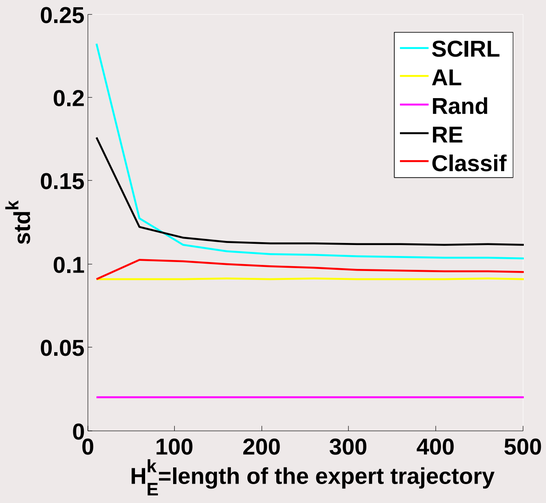


(b) Standard deviation

Figure: Perturbed dynamics: normally distributed reward.



(a) Performance



(b) Standard deviation

Figure: Perturbed dynamics: state-only-dependent reward.