Inverse Reinforcement Learning from Failure

Kyriacos Shiarlis, João Messias, Maarten van Someren, and Shimon Whiteson
Informatics Institute, University of Amsterdam
{k.c.shiarlis, jmessias, m.w.vansomeren, s.a.whiteson}@uva.nl

Abstract—**In this paper, we approach the problem of Inverse Reinforcement Learning (IRL) from a rather different perspective. Instead of trying to only mimic an expert as in traditional IRL, we present a method that can also utilise failed or bad demonstrations of a task. In particular, we propose a new IRL algorithm that extends the state-of-the-art method of Maximum Causal Entropy Inverse Reinforcement Learning to exploit such failed demonstrations. Furthermore, we present experimental results showing that our method can learn faster and better than its original counterpart.**

I. Introduction

In Inverse Reinforcement Learning (IRL) [4], an *apprentice* aims to learn a policy for acting in an environment modelled by a Markov Decision Process (MDP) for which the reward function is not available, but samples from the policy of an *expert* performing the task are given instead. An IRL algorithm tries to find a reward function that leads the apprentice to exhibit behaviour that is similar to the expert's, and that generalises well to situations for which expert data is not available. IRL methods have been applied to simulated car driving [1] and socially appropriate navigation [2, 8]. Existing IRL methods leverage concepts from the wider Machine Learning community, such as maximum entropy [10] and Bayesian formulations [5], structured classification [7], boosting [6] and Gaussian processes [3].

Existing IRL algorithms learn only from successful trials, i.e., from data gathered by an expert performing the task well. This is consistent with the main motivation of IRL since it allows learning in tasks where the reward cannot be trivially hard-coded. For example, the reward function that allows an agent to perform complicated manoeuvres while flying a helicopter cannot be trivially determined, but example demonstrations can be easily obtained from an expert.

However, in many realistic scenarios, failed trials are also readily available. Consider for example tasks such as driving a car. Since humans also learn this task by trial and error, demonstrations of both successful and failed behaviour are available. Moreover, although hand-coding a reward function for this task would be infeasible, labelling each trial as successful or failed is straightforward.

In this paper, we present the first IRL algorithm that can learn from both successful and failed demonstrations. In doing so, we address a key difficulty in IRL: the problem is typically under-constrained since many reward functions are consistent with the expert's behaviour. By using failed trials, our method reduces this ambiguity, resulting in faster and better learning.

To derive an IRL algorithm for learning from failure, we start from the state-of-the-art method of Maximum Causal Entropy Inverse Reinforcement Learning [10]. This approach starts by formulating a constrained optimisation problem that seeks a reward function that yields behaviour consist with that of the expert. We formulate a new optimisation problem with additional constraints that require the resulting behaviour to also be maximally different from the failed demonstrations. Then, by applying the method of Lagrangian multipliers, we produce our new method, which learns from both successful and failed demonstrations. Our empirical results show that utilising failed trials enables learning a reward function that is closer to that of the expert, in fewer iterations.

II. Method

Ziebart proposed an IRL method based on the principle of Maximum Causal Entropy [9]. The main idea is to find a stochastic policy $\pi(a, s)$ that maximises the causal entropy $H(\mathbf{A}^T||\mathbf{S}^T)$ of all actions $\mathbf{A}^T$ taken in a time horizon T, given the visited states $\mathbf{S}^T$, while constrained to match certain statistics from the expert dataset $\mathcal{D}$. These statistics are called the *empirical feature expectations* $\widetilde{\Phi}_{\mathcal{D}}$. This task is formalised as a constrained convex optimisation problem which is then addressed using the method of Lagrangian multipliers.

Each assignment of these multipliers defines a specific reward function, which is optimised as follows. A policy is found under the current reward function using a Bellman equation with a softmax operator instead of a maximum. The feature expectations of the model are then calculated from $\mathcal{D}$. Finally, the multipliers are updated based on the comparison between the empirical and model feature expectations.

Our method extends this approach in two key ways. First, we relax the constraint that the model must exactly match $\widetilde{\Phi}_{\mathcal{D}}$ by introducing slack variables (ζ^+ and ζ^-). Second, we introduce new constraints that enforce dissimilarity between the model and $\widetilde{\Phi}_{\mathcal{D}_b}$, the empirical feature expectations of the failed demonstrations. This is done through a second set of slacks (α^+ and α^-). The resulting constrained optimisation problem is:

$$\underset{\pi(a,s),\alpha_k^+,\alpha_k^-,\zeta_k^+,\zeta_k^-}{\text{argmax}} \quad H(\mathbf{A}^T||\mathbf{S}^T)-\sum_k \left(C(\zeta_k^+ + \zeta_k^-) - D(\alpha_k^+ \alpha_k^-)\right)$$

Subject to:

$$\widetilde{\Phi}_{\mathcal{D},k} - \Phi_{\pi,s_0,k} = \zeta_k^+ - \zeta_k^- \quad \forall k,$$

$$\widetilde{\Phi}_{\mathcal{D}_b,k} - \Phi_{\pi,s_0,k} = \alpha_k^+ - \alpha_k^- \quad \forall k,$$

$$-\zeta_k^+ \leq 0 \quad \text{and} \quad -\zeta_k^- \leq 0 \quad \forall k,$$

$$-\alpha_k^+ \leq 0 \quad \text{and} \quad -\alpha_k^- \leq 0 \quad \forall k,$$

$$\alpha_k^- \alpha_k^+ = 0 \quad \forall k,$$

$$\sum_a P(a|s) = 1 \quad \forall s, \quad \text{and} \quad P(a|s) > 0 \quad \forall s, a,$$

where s and a are states and actions in the MDP; D and C are constants that specify how much we prioritise minimising the ζ slacks versus maximising the α slacks. Optimisation is performed using Lagrangian multipliers, similarly to Ziebart's method. The Lagrangian is first maximised with respect to the variables in the objective function to obtain the dual. We then update the Lagrange multipliers by taking a step in the direction that minimises this dual.

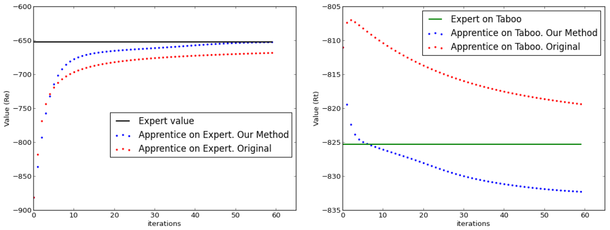
III. Results & Discussion

We compared the performance of our method to Ziebart's original IRL method on a simulated robot control task that requires avoiding moving obstacles. In particular, the robot aims to reach a target state (T) of high reward, while avoiding some states of negative reward (N). The (N) states change throughout time.

To set up the experiments, we defined two reward functions, R_e and R_t, both of which are unknown to the learning algorithms. R_e, the expert's reward function, is the true reward function while R_t is the reward function used by a *taboo agent* that generates the failed demonstrations. Specifically, the taboo agent attempts to approach the (N) states rather than avoid them, and ignores the target. Using these reward functions, we generate the empirical feature expectations $\widetilde{\Phi}_{\mathcal{D},k}$ and $\widetilde{\Phi}_{\mathcal{D}_b,k}$ by allowing agents using soft-optimal policies w.r.t. R_e and R_t to start from an initial state distribution $b_{0,train}$ and proceed for T timesteps. Based on this data, we train two apprentices, one using Ziebart's method and one using ours. Finally, we evaluate performance on a test distribution of initial states $b_{0,test}$.

Figure 1a shows the average value, with respect to R_e, of the policies found by the two methods, at each iteration. Our approach learns more quickly that Ziebart's method and ultimately matches the value of the expert's policy, while Ziebart's method plateaus lower. Figure 1b shows the average value of the two methods with respect to R_t. In this case, lower values are better as they correspond to performing badly on a bad reward function. The results indicate that our method does a better job of minimising value with respect to R_t than Ziebart's method. This figure also shows the performance of the expert's policy with respect to R_t (green). Unlike Ziebart's method, our method is able to accumulate less value with respect to R_t than the expert's policy.

These results demonstrate that taking into account failed demonstrations of a task can allow an apprentice to generalise better to new initial states and learn in fewer iterations than previously possible. In the future, we aim to apply this method to more realistic scenarios such as those involving real robots. We also hope to examine how the similarity between the expert and taboo datasets affect the apprentice's ability to learn. Finally, we would like to extend our modifications to other popular IRL algorithms.



(a) Expert Reward Function (b) Taboo Reward Function

Fig. 1: *Results from the moving obstacle avoidance task. Learning from failure (blue) yields behaviour that is more similar to that of the expert and more dissimilar to that of the taboo agent than Ziebart's IRL method (red).*

Acknowledgments

This research is part of the TERESA project, which has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 611153.

References

[1] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

[2] Peter Henry, Christian Vollmer, Brian Ferris, and Dieter Fox. Learning to navigate through crowded environments. In *ICRA, 2010*, pages 981–986. IEEE, 2010.

[3] Sergey Levine, Zoran Popovic, and Vladlen Koltun. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Information Processing Systems*, pages 19–27, 2011.

[4] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, pages 663–670, 2000.

[5] Deepak Ramachandran and Eyal Amir. Bayesian inverse reinforcement learning. *Urbana*, 51:61801, 2007.

[6] Nathan Ratliff, David Bradley, J Andrew Bagnell, and Joel Chestnutt. Boosting structured prediction for imitation learning. *Robotics Institute*, page 54, 2007.

[7] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning*, pages 729–736. ACM, 2006.

[8] Dizan Vasquez, Billy Okal, and Kai O Arras. Inverse reinforcement learning algorithms and features for robot navigation in crowds: An experimental comparison. In *Intelligent Robots and Systems (IROS 2014)*, pages 1341–1346. IEEE, 2014.

[9] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, Carnegie Mellon University, 2010.

[10] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, pages 1433–1438, 2008.