Massively Multiagent Hierarchical Inverse Reinforcement Learning in Open-world Environments

William H. Guss[*]
wguss@cs.cmu.edu

Ruslan Salakhutdinov[*]
rsalakhu@cs.cmu.edu

January 17, 2018

1 Introduction

Over recent years inverse reinforcement learning (IRL) has revolutionized the state of the art in apprenticeship learning, cooperative and adversarial modeling, and the modeling of intent in human and animal behaviour**TODO: cite**. At its core, IRL solves the problem of learning expert policies indirectly: in direct juxtaposition to behavioral cloning IRL learns the reward function of an expert agent and then produces a policy which maximizes that reward. In this regime, the resultant policy is often far more interpratable, robust, and sample efficient than that of behavioural cloning.

Mechanically, inverse reinforcement learning is optimal not only for cloning expert policies in applications where thousands if not millions of demonstrations are not possible, but also as a substitute for traditional reinforcement learning when exploration is extremely expensive. For example, in robotic manipulation tasks, where typical ϵ-greedy exploration polices would result in potential damage to the robot, apprenticeship learning via IRL is a powerful alternative. Furthermore, by learning the reward function directly, IRL is an effective, interpritable forecasting mechanism in tasks such as epidimiological modeling, traffic prediciton, and first person activity forecasting.**TODO: cite**

Despite its numerous applications and avantages, IRL is an underconstrained optimzation problem; in particular, there are potentially inifinitely many reward functions which explain an expert policies behaviour. To see this formally, let (S, A, T, γ, D) be a rewardless Markov deciion process (MDP) with state space S, action space A, state-to-state transition function T, γ some marginal utility discount, and D an initial state distribution. Given some expert policy $\pi^*(a|s)$, where $a \in A, s \in S$, inverse reinforcement learning aims to find a reward function $R : S \times A \to \mathbb{R}$ such that

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t R(s_t, a_t)\right] \tag{1}$$

where $s_t \sim T(s_{t-1}, a_{t-1}), a_t \sim \pi(\cdot|s_t)$, and $s_0 \sim D$. In this setup, it is clear that degenerate reward functions such as $R = 0$ suffice in explainig π^*. Constraining the space of candidate reward functions has therefore become central to the application of inverse reinforcement learning.

Formally, let R

[*] Carnegie Mellon University, Machine Learning Department.

Due to the power of IRL, practitioners have been able to predict traffic patterns, I don't really know what I'm saying with this. So I think the best strategy is to write as much as possible and then cut. How about that. Bullshit whatever but spin this into a narative that you'd tell Core to the success of IRL is not only its ability to recapitulate expert policies from a relatively low number of samples but also

TODO: Provide a slightly more formal perspectrive to motivate why these methods could be interesting?

TODO: Introduce the problem statement, what is the problem of massively multi-agent hierarchical reinforcement learning in open-world environments and why would this be pertinent.

TODO: Introduce the implications of solving the MMHIRL problem in open-world environments, potentially reference the taxi problem.

2 Our Approach

TODO: To solve the problem of MMHIRL we will introduce methods which take advantage of the local connectedness of parameterized policy and reward space.

TODO:

2.1 Experiments

2.2 Metrics

References