EFFICIENT INVERSE REINFORCEMENT LEARNING WITHOUT COMPOUNDING ERRORS

Nicolas Espinosa Dice, Sanjiban Choudhury, Wen Sun
Department of Computer Science
Cornell University
Ithaca, NY 14850, USA
{ne229, sanjibanc, ws455}@cornell.edu

Gokul Swamy
Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
gswamy@cmu.edu

ABSTRACT

Inverse reinforcement learning (IRL) is an on-policy approach to imitation learning (IL) that allows the learner to observe the consequences of their actions at train-time. Accordingly, there are two seemingly contradictory desiderata for IRL algorithms: *(a)* preventing the compounding errors that stymie offline approaches like behavioral cloning and *(b)* avoiding the worst-case exploration complexity of reinforcement learning (RL). Prior work has been able to achieve either *(a)* or *(b)* but not both simultaneously. In our work, we first present a negative result showing that, without further assumptions, there are no efficient IRL algorithms that avoid compounding errors in the worst case. We then provide a positive result: under a novel structural condition we term *reward-agnostic policy completeness*, we prove that efficient IRL algorithms *do* avoid compounding errors, giving us the best of both worlds. We then address a practical constraint—the case of limited expert data—and propose a principled method for using sub-optimal data to further improve the sample-efficiency of IRL algorithms. Finally, we corroborate our theory with experiments on a suite of continuous control tasks.

1 INTRODUCTION

Inverse reinforcement learning (IRL) is an on-policy approach to imitation learning that involves simultaneously learning a reward function from expert demonstrations and a policy that optimizes the learned reward (Ziebart et al., 2008a). IRL has been applied to a diverse set of applications, including robotics (Ratliff et al., 2007; Abbeel & Ng, 2008; Ratliff et al., 2009; Silver et al., 2010; Zucker et al., 2011), autonomous driving (Bronstein et al., 2022; Igl et al., 2022; Vinitsky et al., 2022), and route finding (Ziebart et al., 2008a;b; Barnes et al., 2023).

Compared to offline imitation learning methods such as behavior cloning, IRL offers the following advantages. First, IRL is more sample efficient, with respect to expert samples, than behavior cloning (Swamy et al., 2022). Second, IRL offers better error scaling, with respect to the horizon, than behavior cloning (Ross & Bagnell, 2010; Swamy et al., 2021; 2022). In other words, for a fixed number of expert samples, IRL achieves a tighter performance gap with the expert policy compared to behavior cloning.

However, the expert sample efficiency of traditional IRL comes at the cost of environment interactions. Because the reward function and policy are learned simultaneously, IRL requires policy optimization to be performed repeatedly, making it susceptible to the worst-case exploration complexity of reinforcement learning (RL) (Swamy et al., 2023). Traditional IRL methods can require an exponential number of environment interactions in the worst case (Kakade, 2003; Swamy et al., 2023). In order to focus the exploration on useful states, prior work has leveraged the expert's state