Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Inverse Reinforcement Learning by Matching Feature Distributions

Project Manager
Dr. Oleg Arenz

Principal Investigator
Dr. Oleg Arenz

Project Term
2019 - 2020

Clusters
Lichtenberg Cluster Darmstadt

Institute
Intelligent Autonomous Systems

University
Technische Universität Darmstadt

Introduction

Reinforcement Learning is a powerful approach to achieve optimal behaviour. However, it typically requires a manual specification of a reward function which often contains several objectives, such as reaching goal positions at different time steps or energy efficiency. Manually trading-off these objectives is often difficult and requires a high engineering effort and, hence, inverse reinforcement learning aims to infer the reward function from expert demonstrations.

State-of-the-art inverse reinforcement learning methods require solving adversarial games resulting in unstable optimization and, furthermore, they often require to directly observe the state and actions of the expert. Instead, we develop a non-adversarial approach that is based on matching arbitrary observations of the expert trajectories in terms of the Kullback-Leibler divergence between sample-based estimates of the learned trajectory distribution and the expert's distribution. The cluster is needed to evaluate design choices during the development of the algorithm and for obtaining statistically significant results for comparing the method with competing approaches.

Methods

Many modern methods in imitation learning and inverse reinforcement learning are based on an adversarial formulation. These methods frame the problem of distribution-matching as a minimax game between a policy and a discriminator, and rely on small policy updates for showing convergence to a Nash equilibrium. In contrast to these methods, we formulate

distribution-matching as an iterative lower-bound optimization by alternating between maximizing and tightening a bound on the reverse Kullback-Leibler divergence. This non-adversarial formulation enables us to drop the requirement of “sufficiently small” policy updates for proving convergence. Algorithmically, our non-adversarial formulation is very similar to previous adversarial formulations and differs only due to an additional reward term that penalizes deviations from the previous policy.

Results

We presented a non-adversarial formulation for imitation learning and used it to derive a new actor-critic based method for offline imitation learning, O-NAIL. We compared O-NAIL with a similar – but adversarial – method, ValueDice on the Lichtenberg high performance computer and found that the non-adversarial formulation may indeed be beneficial.

Discussion

As the resulting algorithms are very similar to their adversarial counterparts, it can be difficult to show significant differences in practice. Hence, in this work, we focused on the insights gained from the non- adversarial formulation. For example, we showed that adversarial inverse reinforcement learning, which was previously not well understood, can be straightforwardly derived from our non-adversarial formulation. However, eventually we would like to derive stronger practical advantages from our formulation.

We demonstrated that the non-adversarial formulation can be used to derive novel algorithms by presenting O-NAIL, an actor-critic based offline imitation learning method and our comparisons with ValueDice suggest that the non-adversarial formulation may indeed be beneficial. However, we hope to further distinguish O-NAIL from prior work by building on the close connection between non-adversarial imitation learning and inverse reinforcement learning in order to learn generalizable reward functions offline.

Adversarial methods have been suggested for a variety of different divergences, including but not limited to the family of f-divergences. The non-adversarial formulation is currently limited to the reverse KL divergence and penalizes deviations from the previous policy based on the reverse KL divergence. It is an open question, whether our lower bound can be generalized to other divergences, for example, when penalizing deviations based on different divergences.

Publications

Arenz, O.; Neumann, G. (2020). Non-Adversarial Imitation Learning and its Connections to Adversarial Methods, arXiv. Submitted to the Journal of Machine Learning. https://arxiv.org/abs/2008.03525

Last Update: 2022-01-04 10:46