



## Autonome Lernende Roboter (ALR)

Prof. Gerhard Neumann

### Project Type \_\_\_\_\_

- Master Thesis
- Bachelor Thesis
- Research Project

### Supervisors \_\_\_\_\_

-  Philipp Becker
-  philipp.becker@kit.edu

### Difficulty \_\_\_\_\_

Algorithmic



Math



Application



# Inverse Reinforcement Learning for Highly-Versatile Behavior

## Description

For many reinforcement learning scenarios exemplary data from (human) experts is available. One way to utilize this data is learning from demonstrations, i.e., learning a policy that mimics the expert behavior. The goal of Inverse Reinforcement Learning (IRL) is to not just copy the expert but to infer the underlying reward function, which gave rise to the observed behavior. This reward function can subsequently be used to refine the learned policy.

Current approaches [4, 3] are often limited to learning the reward function of a single behavior. Yet, human behavior data is often highly versatile, where the same task is solved using different motions. Recently, we proposed Expectation Information Maximization (EIM) [2], a novel approach for density estimation based on the I(nformation)-Projection. Similar to generative adversarial approaches [4], EIM employs a discriminator, yet it avoids an adversarial formulation. Using a previously introduced decomposition of the I-Projection objective [1] for mixture models, EIM can be applied to model highly versatile (multi-modal) data and is already applicable to behavioural cloning.

This project is concerned with developing a new technique for IRL in a contextual, trajectory-based setting and applying it to trajectory planning. In this setting only the trajectory needs to be planned. A controller able of following these trajectories can be assumed to be available, thus actions do not need to be considered explicitly. More specifically, the goal of this project is to extend EIM to IRL. To this end, different ways of realizing the discriminator, such that it learn a reasonable and precise reward function should be derived and implemented.

## Tasks

- Learning from Demonstrations. Use EIM to clone expert behaviour for trajectory planning tasks. To get policies that generalize to novel scenarios we work in a contextual setting.
- Designing Informative and Stable Discriminators. Different ways of realizing the discriminator, such that it learns reasonable, precise and stable reward functions are to be derived and evaluated.
- Inverse Reinforcement Learning. The new discriminators are used together with EIM to simultaneously clone the experts behaviour and extract its underlying reward. The learned reward functions are supposed to be used to further improve the policy.

## References

- [1] Oleg Arenz, Mingjun Zhong, and Gerhard Neumann. Trust-region variational inference with gaussian mixture models. *arXiv preprint arXiv:1907.04710*, 2019.
- [2] Philipp Becker, Oleg Arenz, and Gerhard Neumann. Expected information maximization: Using the i-projection for mixture density estimation. *arXiv preprint arXiv:2001.08682*, 2020.
- [3] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.
- [4] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.