Inverse Q-Learning as a Tool to Investigate Behavior and its Neural Correlates

Joschka Boedecker, University of Freiburg

ABSTRACT

Inverse Reinforcement Learning offers the promise to recover the intention underlying an observed behavior, i.e. a reward signal that would explain the behavior of an agent, assuming it tries to maximize it in the long term. This has important applications for imitation learning in robotics, as well as behavior understanding in fields such as biology and neuroscience. Most approaches that allow to extract such a signal from observed data, however, need to solve a full reinforcement learning problem to convergence multiple times in an inner loop, making them computationally expensive and unsuitable for many application scenarios. In this talk, I will present our work on novel inverse RL algorithms which exploit the structure of Q-Learning and result in algorithms that speed up learning of the reward signal by several orders of magnitude, while also providing more accurate value estimates than prior work. I will also illustrate how this enables the analysis of behavior and the activity of brain regions of different animals as specific examples.