COLLEGE OF ENGINEERING
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

Dissertation Defense

Shun Zhang

Efficiently Finding Approximately-Optimal Queries for Improving Policies and Guaranteeing Safety

Wednesday, April 22, 2020
3:30 pm – 5:30 pm
bluejeans.com/438179098 (Digital only)

ABSTRACT: When a computational agent (called the "robot") takes actions on behalf of a human user, it may be uncertain about the human's preferences. The human may initially specify her preferences incompletely or inaccurately. In this case, the robot's performance may be unsatisfactory or even cause negative side effects to the environment. There are approaches in the literature that may solve this problem. For example, the human can provide some demonstrations or give real-time feedback to the robot's behavior. However, these methods typically require much of the human's attention.

In this thesis, I consider a querying approach. Before taking any actions, the robot has a chance to query the human about her preferences. For example, the robot may query the human about which trajectory in a set of trajectories she likes the most, or whether the human cares about some side effects to the domain. After the human responds to the query, the robot expects to improve its performance and/or guarantee that its behavior is considered safe by the human.

Finding a provably optimal query can be challenging since it is usually a combinatorial optimization problem. In this thesis, I contribute to providing efficient query selection algorithms under uncertainty. I first formulate the robot's uncertainty as reward uncertainty and safety-constraint uncertainty. Under only reward uncertainty, I provide a query selection algorithm that finds approximately-optimal k-response queries. Under only safety-constraint uncertainty, I provide a query selection algorithm that finds an optimal k-element query to improve a known safe policy, or use a set-cover-based query selection algorithm to find a safe policy. Under both types of uncertainty, I provide a batch-query-based querying method that empirically outperforms other baseline querying methods.

Chairs: Profs. Edmund Durfee and Satinder Singh Baveja