CAREER: Towards Reliable Machine Learning in Feedback Systems

Sarah Dean

Overview

Machine learning (ML) advancements on tasks like image recognition and click prediction have enabled applications like self-driving cars and personalized social media feeds. However, once ML *predictions* are deployed to make *decisions*, downstream impacts can lead to catastrophic failures, like pedestrian fatalities and radicalization. Problems arise when seemingly inconsequential errors in prediction result in bad decisions; then, bad decisions have negative outcomes and affect future predictions. In such *feedback loops*, naive decision-making causes compounding errors and leads to failures in safety, equity, and performance. Ensuring desirable outcomes is one of the most important challenges for the modern practice and theory of machine learning—a challenge that this proposal aims to address.

How should decision algorithms make use of possibly unreliable ML predictions while ensuring good outcomes? How can we reliably predict the long term impacts of decisions with models learned from temporally correlated data? How do we adaptively make decisions while learning about initially unknown impacts? This proposal introduces a research program to address these three layered questions through the development of theory and algorithms.

Intellectual Merit

The proposed work will leverage ideas and techniques from online optimization, control theory, system identification, and reinforcement learning. The algorithmic and theoretical frameworks will be developed in tandem with applications in weather prediction, autonomous aerial navigation, recommendation systems, and human-robot interaction, along the following thrusts:

1. *Decision-making with ML predictions:* Reliably leveraging unreliable predictions requires accounting for potential errors to guard against bad outcomes. We propose to 1) develop algorithms that robustly guarantee performance and safety while benefiting from predictions when they are accurate and 2) use decision performance as a metric to evaluate prediction quality.
2. *Learning transition and measurement models:* Understanding the long term impacts of decisions on individuals is crucial, and yet human activities are non-stationary, correlated, and partially observed. We will develop reliable learning algorithms for data arising from such processes, with a particular focus on finite sample uncertainty quantification and bounds.
3. *Sample-efficient reinforcement learning:* Adaptive decision-making requires simultaneously learning from data while making decisions. We propose to develop model-based algorithms which can operate even in partially observed settings, such as the non-stationary user behaviors important for applications like recommendation and human-robot collaboration.

Broader Impacts

The algorithms and theory developed in this proposal have direct impact on applications like weather prediction, recommendation systems, and human-robot collaboration, in which ML is deployed for decision-making. The proposed work will also promote cross-pollination among interdisciplinary research communities in ML, EconCS, and Systems & Control, where the PI serves as a regular organizer of workshops and conferences.

Education The proposal includes an integrated education plan at the undergraduate and graduate levels which promotes *feedback* as a foundational intellectual concept for computer scientists. The proposal also includes an initiative for high school outreach through partnership with residential summer programs at Cornell Engineering. The project will develop a hands-on project and interactive education modules about weather forecasting & balloon control. The initiative will seek to broaden both participation in and understanding of computing. It will provide a particular focus on predictive technologies, feedback, dynamics, and bias.