Safety Validation of Learning-Based Autonomous Systems: A Multi-Fidelity Approach

Ali Baheri*

Rochester Institute of Technology
akbeme@rit.edu

Abstract

In recent years, learning-based autonomous systems have emerged as a promising tool for automating many crucial tasks. The key question is how we can build trust in such systems for safety-critical applications. My research aims to focus on the creation and validation of safety frameworks that leverage multiple sources of information. The ultimate goal is to establish a solid foundation for a long-term research program aimed at understanding the role of fidelity in simulators for safety validation and robot learning.

Introduction

The deployment of learning-enabled decision-making systems into the real world can be risky and error-prone. Reasoning about the safety behavior of a complex autonomous system in a dynamic environment is a challenging task. Our goal is to develop safety verification and validation algorithms for complex autonomous systems operating in a highly evolving and stochastic environment. Specifically, our focus is on the creation of principled general frameworks that fuse insights from machine learning, formal methods, control theory, and optimization communities for safety verification and validation of learning-enabled autonomous systems.

One potential approach to the process of ensuring the safe behavior of autonomous systems is through the use of white box approaches, also referred to as formal methods, that can provide formal guarantees and proofs for the safe behavior of autonomous systems. However, increasing the level of complexity in autonomous systems is a barrier to the use of formal methods for safety assessment due to the problem of scalability. Black-box safety validation has recently gained interest to assess the safety behavior of an autonomous system where the only thing needed to examine the system's safety is a transition function that generates the next state of the system. The black-box safety validation approaches, while dispersed across many application domains, face a key drawback: the current methods do not take into account data from multiple sources, including varying levels of fidelity in simulated environments.

*Ali Baheri is with the Department of Aeronautics & Astronautics at Stanford University.

Research Vision

Fidelity indicates the degree to which the simulator takes simplifications and assumptions into account when modeling the system. Low-fidelity simulators make strong assumptions and simplifications about the underlying system, resulting in relatively fast execution. On the other hand, high-fidelity simulators consider fewer assumptions about the underlying system and exhibit much more realistic behaviors and dynamics. However, they are slower to execute than low-fidelity simulators. With this insight in mind, our focus is to develop frameworks that aim to optimally tradeoff between low-fidelity, computationally cheap and high-fidelity, computationally expensive simulators to maximize the discovered number of most likely failure events while minimizing the associated simulation time and cost.

We argue that a robot, or a cyber-physical system (CPS), can query data from multiple sources, including different levels of simulators, real-world data, and/or human expert inputs. Our hypothesis is that incorporating data from different sources could accelerate the certification task of a CPS and, broadly speaking, the learning process of an autonomous system. With multiple testing environments available at different fidelity levels, monetary costs, and test times, our goal is to answer the following research questions: (i) How can we optimally coordinate and use these testing environments to achieve the safety validation and certification objectives, such as identifying realistic failure modes? (ii) Can we get the benefit of running a large number of low-fidelity simulations with a cheaper cost and shorter time as well as adaptively performing as few high-fidelity tests as possible to identify the failure modes? In that line, our vision is to develop a new class of validation and verification algorithms that optimally take into account multiple sources of information with varying degrees of fidelity in simulated environments.

In summary, the goal of this research is to build trust in learning-enabled decision-making systems for safety-critical applications. We believe that systematic approaches that capture the information from multiple simulated environments could significantly speed up the certification process and reduce the overall computational time and cost. Hence, the outcome of this research will establish new results and contribute to filling a gap in the state-of-the-art for the safety validation of autonomous systems.