Learning Interpretable Models for Black-Box Agents
Pulkit Verma and Siddharth Srivastava, Arizona State University, Tempe, AZ, USA

Introduction
Can a non-expert determine if an AI agent is reliable/safe for a task?
- Objective: Learn an interpretable model of a non-stationary black-box agent by interrogating it.
- Key technical challenge:
  - Which sequence of questions to ask?

Abstraction in Space of Models

Algorithm
- Start with the most abstracted node in lattice.
- Pick abstraction candidates in some order.
- For each candidate, generate three models and for each pair of models:
  - Generate a distinguishing query $Q$ and pose it to the agent.
  - Get the response $R$ from the agent.
  - Prune out the incorrect variants of candidate models.
  - Repeat steps 3-6 till the model is fully estimated.
- Return the final set of model(s).

Example of Agent Interrogation

Plan Outcome Query: Asks the outcome of a plan.
Query: Initial state, plan.
Response: Length of successful execution, final state.

Key Algorithmic Principle

Key feature of the algorithm
Each time we prune an abstracted model, we prune a very large number of models at the most concrete node.

Results
- Randomly generate an agent and environment from the IPC benchmark suite.
- Algorithm learns this agent’s model.
- Theorem: The algorithm will always terminate and return a set of models, each of which are functionally equivalent to agent’s model.

Salient Features
- Needs no prior knowledge of the agent model.
- Requires an agent to have only rudimentary query answering capabilities.
- Queries can be answered by the agent using a simulator.
- Works for non-stationary environments.

This work was supported in part by the NSF under grants IIS 1844325 and IIS 1909370