Efficient Interaction-Aware Interval Analysis of Neural Network Feedback Loops

Saber Jafarpour*, *Member, IEEE*, Akash Harapanahalli*, *Graduate Student Member, IEEE*, and Samuel Coogan, *Senior Member, IEEE*

***Abstract*— In this paper, we propose a computationally efficient framework for interval reachability of systems with neural network controllers. Our approach leverages inclusion functions for the open-loop system and the neural network controller to embed the closed-loop system into a larger-dimensional embedding system, where a single trajectory over-approximates the original system's behavior under uncertainty. We propose two methods for constructing closed-loop embedding systems, which account for the interactions between the system and the controller in different ways. The interconnection-based approach considers the worst-case evolution of each coordinate separately by substituting the neural network inclusion function into the open-loop inclusion function. The interaction-based approach uses novel Jacobian-based inclusion functions to capture the first-order interactions between the open-loop system and the controller by leveraging state-of-the-art neural network verifiers. Finally, we implement our approach in a Python framework called `ReachMM` to demonstrate its efficiency and scalability on benchmarks and examples ranging to 200 state dimensions.**

***Index Terms*— Reachability analysis, Inclusion functions, Neural networks, Interconnected systems.**

reliability of learning algorithms is an essential challenge in their integration into safety-critical systems.

Typical learning architectures involve high-dimensional nonlinear function approximators, such as neural networks, necessitating special tools for their input-output analysis. The machine learning and control community have made significant progress in analyzing the safety of neural networks in isolation, including efficient and sound input-output bounds, worst-case adversarial guarantees, and sampling-based stochastic guarantees (cf. [4]). However, most of these existing frameworks for standalone neural networks do not address the unique challenges for closed-loop analysis—namely information propagation, non-stationarity of the bounds, and complex interactions between the system and the controller. In these cases, it is essential to understand and capture the nature of the interactions between the system and the neural network. Recently, several frameworks have emerged for safety verification of learning algorithms in the closed-loop. These frameworks usually incorporate neural network verification algorithms into the closed-loop safety analysis by studying