Exercise 8

Abstract Interpretation of Neural Networks

Program Analysis for System Security and Reliability 2018
ETH Zürich

May 5, 2018

Recall that the activation function $\text{ReLU}(x) = \left\{ \begin{array}{ll} x & x \geq 0 \\ 0 & \text{otherwise.} \end{array} \right\}$

Let N_θ be a two input and one hidden layer neural network defined as

$$N(x_1, x_2) = \text{ReLU}(\text{ReLU}(-x_1 + x_2 + 2) + \text{ReLU}(x_1 - 2x_2))$$

Problem 1. Evaluate the neural network on the input $(x_1, x_2) = (-1, 1)$

Problem 2. Write a short program to compute the interval analysis of this network.

Problem 3. Apply standard interval analysis to find $N^{\#}([0, 2] \times [0, 1])$.

Problem 4. Use interval analysis creatively to demonstrate that $5 \notin N([0, 2] \times [0, 1])$.

Problem 5. Represent every Interval domain perfectly as a Zonotope.

Problem 6. Suppose we were to define the point-wise ReLU transformer for Zonotope as follows: $\text{ReLU}^{\#}_n(a)^n_i = \text{ReLU}(a^n_i)$ for error terms i and variables n. Prove that this is unsound.

Problem 7. Design a very simple and sound point-wise transformer for ReLU for Zonotopes, and show that it is at least as precise as Interval.