Exercise 06

Abstract Interpretation

Reliable and Interpretable Artificial Intelligence
ETH Zurich

Problem 1 (Interval analysis on a Neural Network)**.** The Rectified Linear Unit (ReLU) is widely used as an activation function in Neural Networks. It is defined as:

$$\mathrm{ReLU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Consider the following two input Neural Network N with one hidden layer:

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

In this problem, we are applying abstract interpretation on N using the interval domain.

1. In the lecture, we have discussed transformers for addition and inequality in the interval domain. Write a python program that performs abstract interpretation of the network N using the interval domain. This is, implement a program that evaluates $N^{\#}(I_1 \times I_2)$ for input intervals I_1 and I_2.

2. Apply interval analysis by hand to find $N^{\#}([0, 2] \times [0, 1])$. Compare this to the output of your program.

3. The above analysis is not very precise: for example, the value 5 is contained in the output interval even though it is infeasible in the concrete domain. By cleverly splitting the input box, one can refine the analysis. Use such an approach to prove that $5 \notin N([0, 2] \times [0, 1])$.

Problem 2 (Alternative abstract domain)**.** In this problem, we consider a simple abstract domain $A = \{+, -, 0, \top, \bot\}$ to be used for sign analysis of integers. The lattice $(A, \sqsubseteq)$ is visualized in Fig. 1. Its meaning is defined by its concretization γ as follows:

$$\gamma(+) = \{x \mid x \in \mathbb{Z},\ x > 0\} \qquad \gamma(\top) = \mathbb{Z}$$
$$\gamma(-) = \{x \mid x \in \mathbb{Z},\ x < 0\} \qquad \gamma(\bot) = \emptyset$$
$$\gamma(0) = \{0\}$$

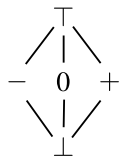

Figure 1: The lattice $(A, \sqsubseteq)$.

1. Find a transformer for addition ($\oplus$), subtraction ($\ominus$), multiplication ($\odot$) and ReLU in the abstract domain A.

2. Consider the single input Neural Network N' defined as:

$$N'(x) = \text{ReLU}(2 - x) + \text{ReLU}(3x + 1)$$

   Assume we want to prove that the output of N' is positive for all odd inputs greater or equal to 5, this is:

$$\forall x \in \mathbb{Z}. \quad x \bmod 2 = 1 \ \wedge \ x \geq 5 \quad \Longrightarrow \quad N'(x) > 0$$

   a) Try to prove the fact using the domain A. First, find a suitable abstraction of the set of inputs satisfying the left hand side of the implication. Then, construct the abstract function $N'^{\#}$ using the transformers from the previous step and apply $N'^{\#}$ to the abstract input. Can you prove the claim?

   b) Perform the same task using the interval abstract domain. Can you prove the claim now?

   c) What are the advantages of using the interval domain over A and vice versa?