Exercise 07

DeepPoly and Abstract Interpretation

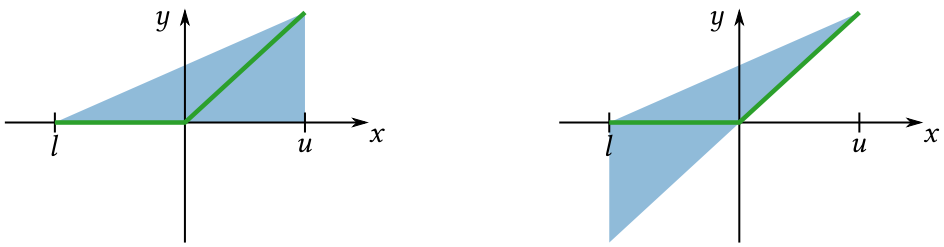
Reliable and Interpretable Artificial Intelligence
ETH Zurich

Problem 1 (Smaller Area)**.** Recall that DeepPoly decides between two options for relaxing the result of $y = \text{ReLU}(x)$ based on the area, shown in Fig. 1.

Derive a decision procedure depending on l and u which decides when Option 1 results in a smaller area. Break ties in favor of Option 1.

Problem 2 (DeepPoly Example)**.** Consider the fully connected neural network shown in Fig. 2. The neural network has two input neurons (x_1, x_2) and two output neurons (x_7, x_8).

Analyze this network using DeepPoly with respect to the input region spanned by $x_1 \in [0, 1]$ and $x_2 \in [0, 1]$. Then, use the result to show that $x_7 \geq x_8$.



(a) Option 1 (b) Option 2

Figure 1: Options for triangle relaxations in DeepPoly.

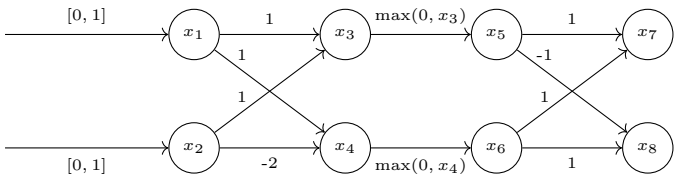


Figure 2: Neural network to be analyzed with DeepPoly.

Problem 3 (Abstract Interpretation)**.** In this problem, we consider a (toy) abstract domain A over $\mathbb{R}$ with abstract elements $\{+, -, 0, \top\}$ whose meaning is defined by the concretization γ: [1]

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

For instance, the abstract element + represents all positive real numbers.

1. Find sound abstract transformers for addition ($+^\sharp$), scalar multiplication with a constant ($\cdot^\sharp$), and ReLU ($\text{ReLU}^\sharp$) in the abstract domain A. The transformers should be as precise as possible.

2. Consider the single input neural network $N\colon \mathbb{R} \to \mathbb{R}$ defined as:

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

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

   $$\forall x \in \mathbb{R}.\quad x \geq 5 \quad \Longrightarrow \quad N(x) > 0$$

   Try to prove the claim 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 transformer $N^\#$ of N using the transformers from the previous step and apply $N^\#$ to the abstract input. Can you prove the claim?

3. Try to prove the claim using the box/interval domain. Can you prove the claim?

[1] For technical reasons, A should also include a dedicated element $\bot$ with concretization $\gamma(\bot) = \emptyset$. However, for this exercise, you do not need to consider this.