Exercise 07 - Solution

DeepPoly and Abstract Interpretation

Reliable and Interpretable Artificial Intelligence
ETH Zurich

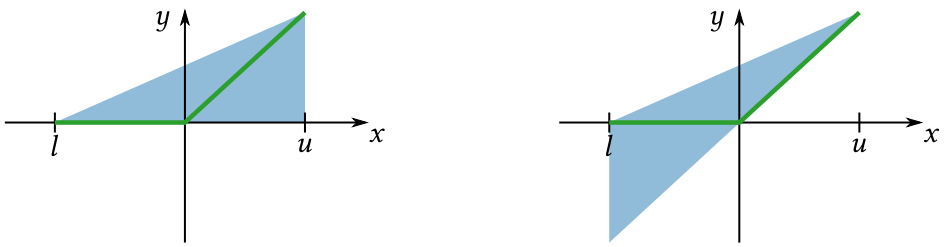
Problem 1 (Smaller Area)**.** Recall that DeepPoly decides between two options for relaxing the result of $y = \text{ReLU}(x)$ based on the area, shown in Fig. 1.

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.

Solution 1. The area for Option 1 is $A_1 = \frac{(-l+u)\cdot u}{2}$, while the area for Option 2 is $A_2 = \frac{-l\cdot -l}{2} + \frac{(-l+u)\cdot u}{2} - \frac{u\cdot u}{2}$.

This, we should pick Option 1 if

$$\begin{aligned} & A_1 \leq A_2 \\ \iff & -l \cdot u + u^2 \leq (-l)^2 - l \cdot u + u^2 - u^2 \\ \iff & u^2 \leq (-l)^2 \\ \iff & u \leq -l \end{aligned}$$



(a) Option 1 (b) Option 2

Figure 1: Options for triangle relaxations in DeepPoly.

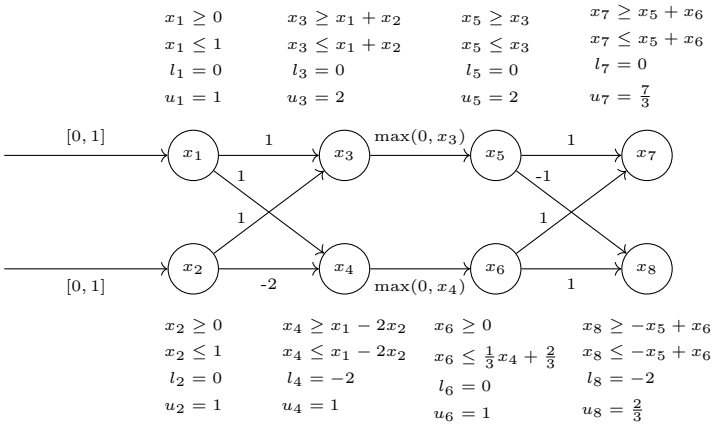


Figure 2: Neural network to be analyzed with DeepPoly.

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 $x_7 \geq x_8$.

Solution 2. Fig. 2 shows the result of our analysis.

For the ReLUs, we used that x_3 is strictly positive and that x_4 satisfies $-l_4 \geq u_4$ (hence we used Option 1 in Fig. 1b).

To compute the lower and upper bounds, we computed the following:

$$
\begin{aligned}
&x_1 \geq 0 =: l_1 \\
&x_1 \leq 1 =: u_1 \\
&x_2 \geq 0 =: l_2 \\
&x_2 \leq 1 =: u_1 \\
&x_3 \geq x_1 + x_2 \geq 0 + 0 = 0 =: l_3 \\
&x_3 \leq x_1 + x_2 \leq 1 + 1 = 2 =: u_3 \\
&x_4 \geq x_1 - 2x_2 \geq 0 - 2 \cdot 1 = -2 =: l_4 \\
&x_4 \leq x_1 - 2x_2 \leq 1 - 2 \cdot 0 = 1 =: u_4 \\
&x_5 \geq x_3 = x_1 + x_2 \geq 0 + 0 =: l_5 \\
&x_5 \leq x_3 = x_1 + x_2 \leq 1 + 1 =: u_5 \\
&x_6 \geq 0 =: l_6 \\
&x_6 \leq \tfrac{1}{3}x_4 + \tfrac{2}{3} \leq \ldots \text{(as above)} \leq \tfrac{1}{3} \cdot 1 + \tfrac{2}{3} = 1 =: u_6 \\
&x_7 \geq x_5 + x_6 \geq x_3 + 0 \geq \ldots \text{(as above)} \geq 0 := l_7 \\
&x_7 \leq x_5 + x_6 \leq x_3 + \tfrac{1}{3}x_4 + \tfrac{2}{3} \leq x_1 + x_2 + \tfrac{1}{3}(x_1 - 2x_2) + \tfrac{2}{3} = \tfrac{4}{3}x_1 + \tfrac{1}{3}x_2 + \tfrac{2}{3} \leq \tfrac{7}{3} := u_7 \\
&x_8 \geq -x_5 + x_6 \geq -x_3 + 0 \geq \ldots \text{(as above)} \geq -2 := l_8 \\
&x_8 \leq -x_5 + x_6 \leq -x_3 + \tfrac{1}{3}x_4 + \tfrac{2}{3} \leq -(x_1 + x_2) + \tfrac{1}{3}(x_1 - 2x_2) + \tfrac{2}{3} \\
&\quad = -\tfrac{2}{3}x_1 - \tfrac{5}{3}x_2 + \tfrac{2}{3} \leq \tfrac{2}{3} =: u_8
\end{aligned}
$$

Using the analysis result, we can show that

$$x_7 - x_8 \geq x_5 + x_6 - (-x_5 + x_6) = 2x_5 \geq \ldots \text{(as above)} \geq 0. \tag{1}$$

Note that we perform symbolic simplifications during back-substitution whenever possible. For example, in Eq. (1), we simplified $x_6 - x_6$ to 0. Without these critical simplifications, we get a worse lower bound:

$$
\begin{aligned}
x_7 - x_8 &\geq x_5 + x_6 - (-x_5 + x_6) \\
&= x_5 + x_6 + x_5 - x_6 \\
&\geq x_3 + 0 + x_3 - \left(\tfrac{1}{3}x_4 + \tfrac{2}{3}\right) \\
&= x_3 + x_3 - \tfrac{1}{3}x_4 - \tfrac{2}{3} \\
&\geq x_1 + x_2 + x_1 + x_2 - \tfrac{1}{3}(x_1 - 2x_2) - \tfrac{2}{3} \\
&\geq x_1 + x_2 + x_1 + x_2 - \tfrac{1}{3}x_1 + \tfrac{2}{3}x_2 - \tfrac{2}{3} \\
&\geq 0 + 0 + 0 + 0 - \tfrac{1}{3} + 0 - \tfrac{2}{3} = -1
\end{aligned}
$$

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?

Solution 3.

1. If any operand of $+^\sharp$ is $\top$, the result is $\top$. The remaining transformers $+^\sharp$, $\cdot^\sharp$ and $\text{ReLU}^\sharp$ are shown in Fig. 3 (symmetric rules ommited).

2. We abstract the set $\mathcal{S}$ of real numbers greater or equal to 5 by +, since the latter is the element in A whose concretization is the least superset of $\mathcal{S}$ (most precise). Note that + is an overapproximation of $\mathcal{S}$. Applying $N^\#$ to + yields:

   $$N^\sharp(+) = \text{ReLU}^\sharp(\underbrace{(\underbrace{3 \cdot^\sharp +}_{+}) -^\sharp \underbrace{1}_{+}}_{\top}) +^\sharp 1 = \top +^\sharp \underbrace{1}_{+} = \top$$

   It is $\gamma(\top) = \mathbb{R}$, from which we cannot infer that $N(x) > 0$. This domain is too imprecise to 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.

$$
\begin{aligned}
+ +^\sharp + &= + \\
+ +^\sharp - &= \top \\
- +^\sharp - &= - \\
0 +^\sharp + &= + \\
0 +^\sharp - &= - \\
0 +^\sharp 0 &= 0
\end{aligned}
\qquad\qquad
+ \cdot^\sharp c = \begin{cases} + & \text{if } c > 0 \\ - & \text{if } c < 0 \\ 0 & \text{if } c = 0 \end{cases}
$$

$$
0 \cdot^\sharp c = 0 \quad \text{for any } c
$$

$$
\begin{aligned}
\mathrm{ReLU}^\sharp(+) &= + \\
\mathrm{ReLU}^\sharp(-) &= 0 \\
\mathrm{ReLU}^\sharp(0) &= 0 \\
\mathrm{ReLU}^\sharp(\top) &= \top
\end{aligned}
\qquad\qquad
- \cdot^\sharp c = \begin{cases} - & \text{if } c > 0 \\ + & \text{if } c < 0 \\ 0 & \text{if } c = 0 \end{cases}
$$

$$
\top \cdot^\sharp c = \begin{cases} 0 & \text{if } c = 0 \\ \top & \text{otherwise} \end{cases}
$$

Figure 3: Transformers $+^\sharp$, $\cdot^\sharp$ and $\mathrm{ReLU}^\sharp$.

3. We abstract the input set $\mathcal{S}$ by the interval $[5, \infty)$. Using the box transformers, we get:

$$
N([5,\infty)) = \mathrm{ReLU}(\underbrace{\underbrace{3 \cdot [5,\infty)}_{[15,\infty)} - 1}_{[14,\infty)}) + 1 \;=\; [14,\infty) + 1 \;=\; [15,\infty)
$$

   Hence, the output is guaranteed to be at least 15, which proves the property.