Exercise 05

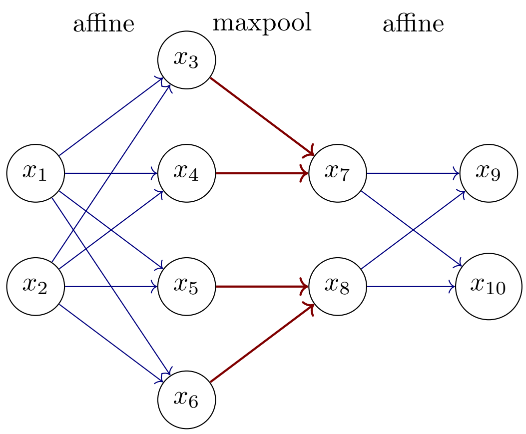
Certification with Box and MILP

Reliable and Interpretable Artificial Intelligence
ETH Zurich

Problem 1 (Box Verification for Maxpool)**.** Consider the maxpool operation defined as $y := \max(x_1, x_2)$, which computes the maximum of two input neurons $x_1, x_2 \in \mathbb{R}$. This operation is typically used in neural networks to reduce dimensionality. In this task, you are going to extend box verification to the maxpool operation.

1. Derive a sound abstract transformer $\max^\sharp$ for the maxpool operation in the box domain. That is, derive expressions for y_1, y_2 such that $[y_1, y_2] = \max^\sharp([a_1, b_1], [a_2, b_2])$ for $a_1, b_1, a_2, b_2 \in \mathbb{R}$. Your transformer should be as precise as possible.

2. Consider the neural network defined below. The network takes inputs x_1, x_2 and produces outputs x_9, x_{10}. It consists of both affine and maxpool layers.



$$x_3 := x_1 + x_2 \qquad x_7 := \max(x_3, x_4)$$
$$x_4 := x_1 - 2 \qquad x_8 := \max(x_5, x_6)$$
$$x_5 := x_1 - x_2 \qquad x_9 := x_7$$
$$x_6 := x_2 \qquad x_{10} := -x_7 + x_8 - 0.5$$

Assume we want to prove that for all values of $x_1, x_2 \in [0,1]$, the output satisfies $x_9 > x_{10}$. Using your abstract transformer from above, try to prove the property by performing verification in the box domain. Does the proof succeed?

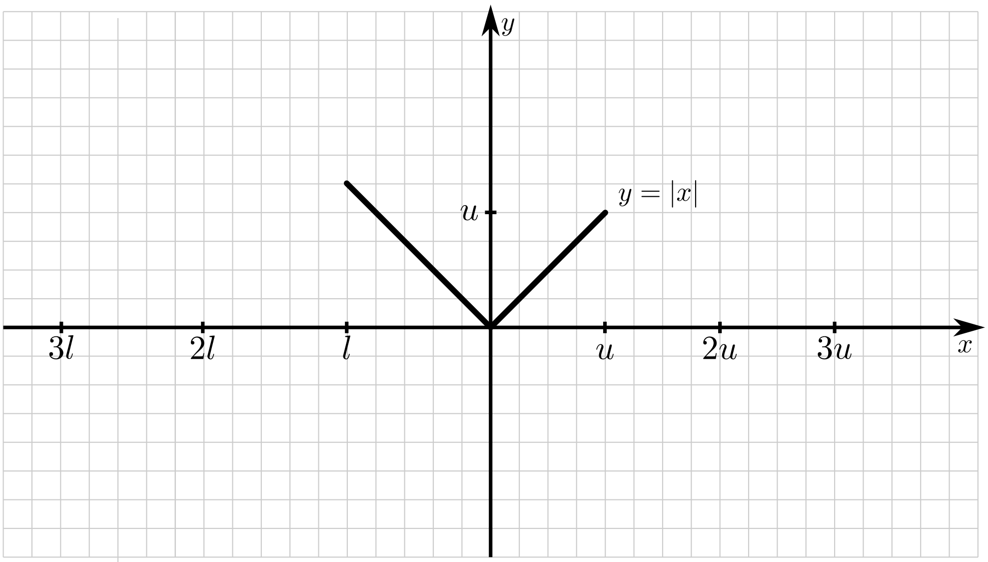
Problem 2 (MILP for Absolute Function—*from a previous exam*)**.** Consider the absolute function $y = |x|$, which computes the absolute value of a neuron $x \in \mathbb{R}$. Assume we know that x takes values in the range $l \leq x \leq u$ (e.g., computed using box verification).

1. In the coordinate system below (where $l \leq 0 \leq u$), draw the two lines indicated by

$$\frac{y}{2} = -\frac{x}{2} + u \cdot a \qquad \text{for } a \in \{0, 1\}.$$

   Indicate which points satisfy the following Mixed Integer Linear Program (MILP) constraints (here, ignore that $l \leq x \leq u$):

$$\frac{y}{2} \leq -\frac{x}{2} + u \cdot a, \qquad a \in \{0, 1\}.$$



2. Starting from the constraints above, find an exact MILP encoding of the absolute function. That is, provide a set of MILP constraints with solution $y = |x|$.