Exercise 08

Geometric robustness

Reliable and Interpretable Artificial Intelligence
ETH Zurich

Problem 1 (Verifying rotation)**.** In this problem we consider verifying robustness of network shown in Fig. 1 to rotation by angle $\theta \in [0.5, 1.5]$. We assume that there are two pixels, one at coordinate $(1, 0)$ whose value after the rotation is $\cos(\theta)$ and another at coordinate $(0, 1)$ whose value after the rotation is $\sin(\theta)$. Vaues of these two pixels are used as inputs to the network in Fig. 1, where their values are denoted as x_1 and x_2. Our goal here is to certify that $x_5 > -1.6$.

1. Try to prove the claim by using interval/box domain for both trignometric operations and operations in the neural network. Can you prove it?

2. Try to prove the claim by using interval/box domain for trigonometric operations and DeepPoly domain for operations in the network. Can you prove it?

3. Compute the tightest linear lower and upper bound for $\sin(\theta)$ and $\cos(\theta)$. By tightest lower bound, here we mean lower bound with smallest area between the lower bound line and the target function (and analogous for upper bound).

4. Use linear bounds computed in the previous step for trignometric functions together with DeepPoly domain for the network to prove the property.

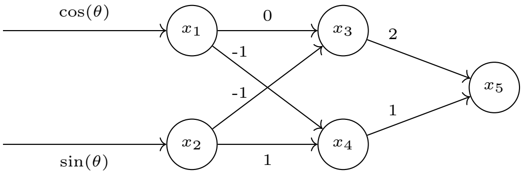

Figure 1: Neural network to be analyzed.

Problem 2. (Bounding functions) In this task we will prove statement from the lecture. The derived inequality will enable us to bound piecewise differentiable function given that we have a bound on its gradients. Let $f : [a_1, b_1] \times ... \times [a_k, b_k] \to \mathbb{R}$ be piecewise differentiable function defined as $f(x) = f_{i,j}(x)$ where $x \in D_{i,j}$. Here $D_1, ..., D_N$ are hyperrectangles which partition the function domain $[a_1, b_1] \times ...[a_k, b_k]$ into finite number of pieces.

1. Let $||\nabla f_{i,j}(z)||_\infty \leq L$ for all $z \in D_{i,j}$. Prove the following bound:

$$f_{i,j}(y) \leq f_{i,j}(x) + L||x - y||_1, \forall x, y \in D_{i,j}$$

2. Prove the following bound:

$$f(y) \leq f(x) + L||x - y||_1, \forall x, y \in [a_1, b_1] \times ...[a_k, b_k].$$

3. Prove that:

$$f(y) \leq f(c) + \frac{L}{2} \sum_{i=1}^{k} b_i - a_i, \forall y \in [a_1, b_1] \times ...[a_k, b_k].$$

Here c is center of the domain, meaning $c_i = \frac{1}{2}(a_i + b_i)$.