The Price of Relaxation: Evaluating the Worst Case Error in Convex Relaxations of Neural Networks

Merkouris Papamichail[1,2], Konstantinos Varsos[1,2], Giorgos Flouris[1], João Marques-Silva[3,4]

[1] Institute of Computer Science, Foundation for Research and Technology – Hellas, [2] Computer Science Department, University of Crete, [3] Catalan Instutute for Research and Advance Studies, [4] University of Lleida

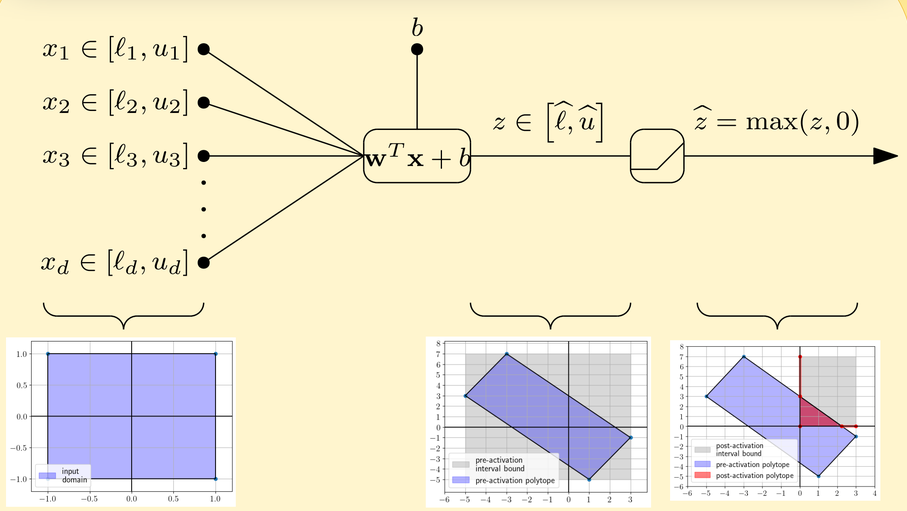
Abstract

NNs have become the de facto ML method. This led the researchers to develop techniques that formally assert the existence (or absence) of certain desirable properties. Such techniques rely on convex programming. However, the ReLU activation function lacks differentiability. Thus, in order to apply convex techniques, the network must first be (convexly) relaxed. In this work, we examine how the relaxed network differs from the original, in the worst case.

Contributions:

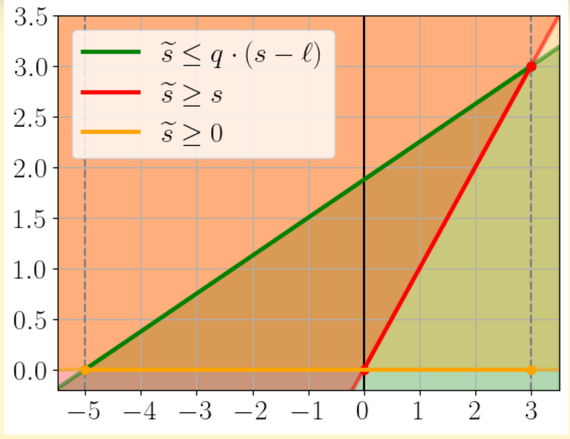
1. Tight (top) convex relaxation (TR).
2. Tight divergence metric (TD).
3. Theoretical lower/upper bounds, w.r.t. depth.
4. Predicting Exponential growth.
5. Testing the TD, w.r.t. NN's depth.
6. Testing the TR, w.r.t. mis-classifications.

Interval Bound Propagation



When given interval bounds for the input domain, we can *propagate* them through the perceptron to obtain bounds on the output. Here we consider a 2x2 neuron, with parameters: W = [1 2; 3 -4], b = [-1 1], and $x_1, x_2 \in [-1, 1]$.

Relaxing the ReLU



$$0 \leq \widetilde{z} \leq z, \ \widetilde{z} \leq \frac{\widehat{u}}{\widehat{u} - \widehat{\ell}} \cdot (z - \widehat{\ell})$$

Given the *pre-activation* bounds, we can *relax* the ReLU. This entails a set of inequalities *for each neuron*. Above, we show the feasible area of the inequalities.

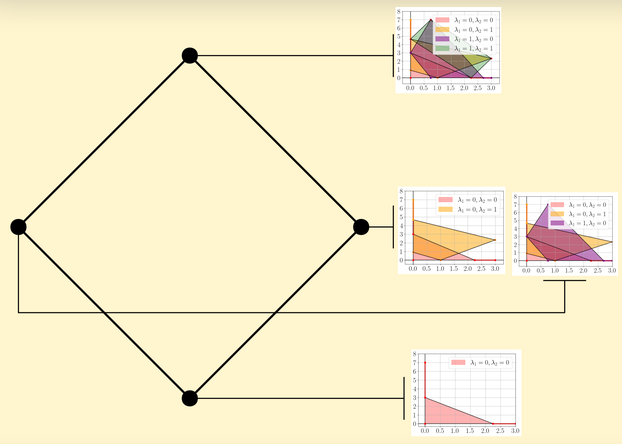
Convex Program

For each layer $i \in [L]$:

$$\left.\begin{aligned} z^{(i)} &= W^{(i)}\widehat{z}^{(i)} + b^{(i)} \\ 0 &\leq \widetilde{z}^{(i)} \leq z^{(i)}, \\ \widetilde{z}^{(i)} &\leq \frac{\widehat{u}^{(i)}}{\widehat{u}^{(i)} - \widehat{\ell}^{(i)}} \cdot \left(z^{(i)} - \widehat{\ell}^{(i)}\right) \end{aligned}\right\}$$

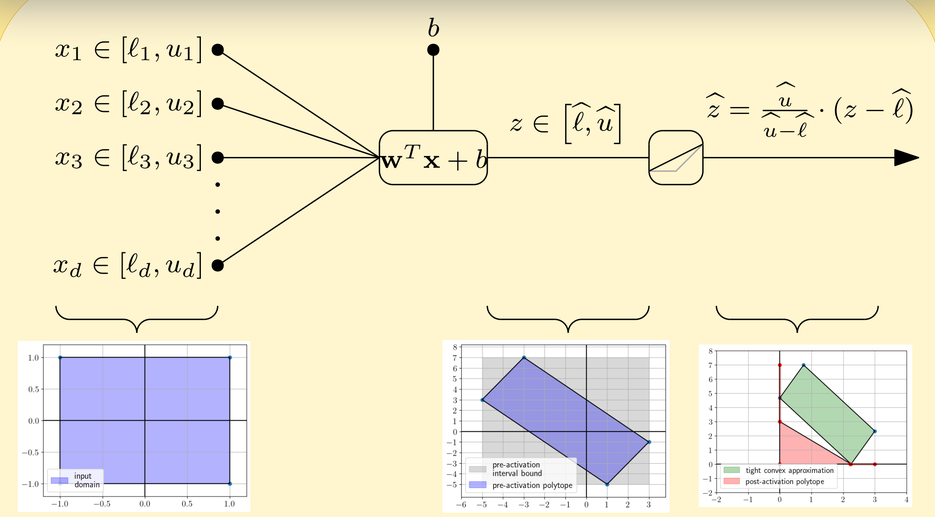
By grouping together the inequalities of each neuron, we end up with a *convex program*. This program has multiple solutions, each corresponding to a different relaxation. The original belongs to the solution set. Above we run Interval Bound Propagation *recursively* for each layer to obtain the pre-activation bounds.

Solutions' Lattice



The convex program has multiple solutions, organized in a *lattice*. Only the vertices of the lattice are able to be chosen, when applying a linear objective. The *top* solution corresponds to the *fully-relaxed* network, where the inequalities of every neuron are tight. The bottom relaxation corresponds to the original network.

Tight (Top) Convex Perceptron



By replacing the activation function, in each neuron, with the right hand side of the neuron's convex inequality, we end up with the fully relaxed network, corresponding to the lattice's top solution.

Tight (Top) Convex Network

$$\widetilde{z}(\mathbf{x}) = \underbrace{\left(\prod_{i=1}^{L} W^{(i)} \frac{\widehat{u}^{(i)}}{\widehat{u}^{(i)} - \widehat{\ell}^{(i)}}\right)}_{\text{Aggregated Weights}} \mathbf{x} + \underbrace{\sum_{i=1}^{L} \left(\prod_{i=1}^{L} W^{(i)} \frac{\widehat{u}^{(i)}}{\widehat{u}^{(i)} - \widehat{\ell}^{(i)}}\right) \left(b^{(i)} - \ell^{(i)}\right)}_{\text{Aggregated Bias}}$$

The fully relaxed network can be re-written as a single perceptron. We expect this to influence its predictive capabilities. We consider this to be the *worst-case* relaxation, w.r.t. a given input.

Tight Divergence (TD)

$$\text{dvg}(\mathbf{x}) = \|\widehat{z}(\mathbf{x}) - \widetilde{z}(\mathbf{x})\|_\infty$$

original network; tight relaxation

We define the *tight divergence* as a measure of error, i.e. the infinity distance between the output of the tight relaxation and the original netowrk.

Wort Case Tight Divergence (WCTD)

WCTD is computed as the *maximum* TD w.r.t. an input domain D:

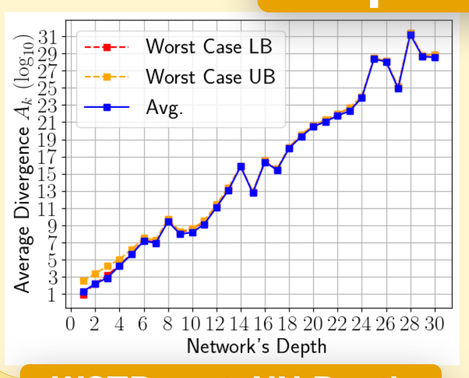
$$\|\widehat{u}^{(L)}\|_\infty \geq \mathcal{E} = \sup_{\mathbf{x} \in D} \text{dvg}(\mathbf{x}) \geq \text{dvg}(\mathbf{0}) = \|\widetilde{z}(\mathbf{0})\|_\infty$$

For the value of the tight relaxation at **0**, we have:

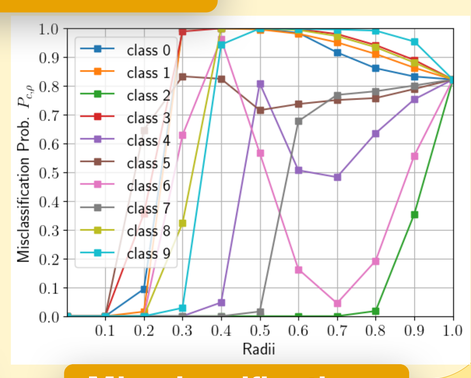
$$\widetilde{z}(\mathbf{0}) = \sum_{i=1}^{L} \left(\prod_{i=1}^{L} W^{(i)} \frac{\widehat{u}^{(i)}}{\widehat{u}^{(i)} - \widehat{\ell}^{(i)}}\right) \left(b^{(i)} - \ell^{(i)}\right)$$

Predicting an **exponential WCTD, w.r.t. the NN's depth!**

Experiments



WCTD w.r.t. NN Depth



Mis-classifications

Conclusions:

Convex relaxations perform well only for shallow networks and for small input domains.

View our code and experiments here!

FORTH FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS

ICREA

Universitat de Lleida