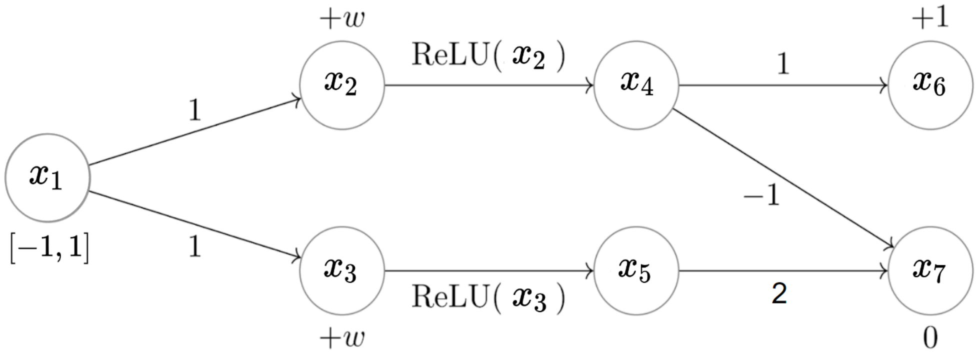
Exercise 05 - Solution

Continuity of Certified Training

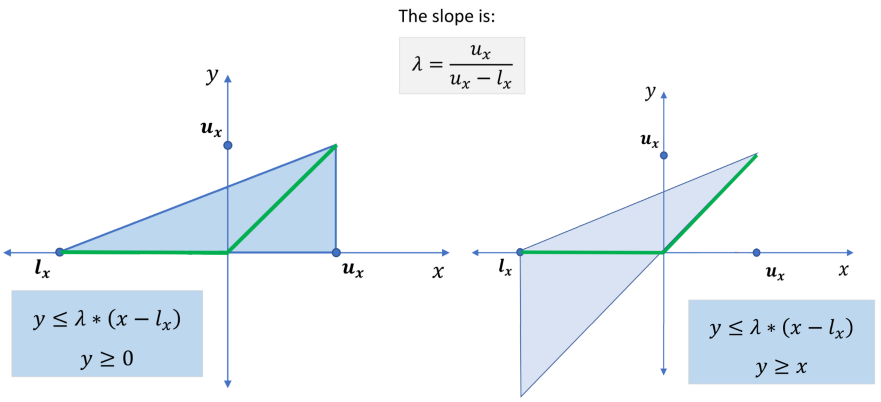
Reliable and Trustworthy Artificial Intelligence
ETH Zurich

Problem 1 (Continuity of DeepPoly in the Weight Space)**.** In this exercise, we demonstrate one common problem that arises during certified training with many precise convex relations, namely, that they are not continuous functions of the network weights. This phenomena in turn results in a much harder optimization problem that needs to be solved by the certified training algorithm. Consider the following neural network:



Here, the network has single input neuron $x_1 \in [-1, 1]$ and 1 parameter w such that $x_2 = x_1+w$, $x_3 = x_1+w$, $x_4 = \text{ReLU}(x_2)$, $x_5 = \text{ReLU}(x_3)$, $x_6 = x_4+1$ and $x_7 = 2x_5-x_4$.

(a) Recall that in DeepPoly in order to compute the linear bounds of a ReLU node whose sign cannot be determined, we need to choose which of the two different convex relaxations for the ReLU, shown in (a) and (b) in Fig. 1, is applied. A common efficient heuristic to choose between the two relaxations is based on selecting the triangle with the smaller area in Fig. 1 which is equivalent to comparing u_x and $-l_x$ (see Exercise 03). Using this heuristic, compute the DeepPoly upper bound for x_7 as a function of the network parameter w. Is the function continuous? Why or why not?


The slope is:
$\lambda = \frac{u_x}{u_x - l_x}$
y
u_x
l_x
u_x
x
$y \leq \lambda * (x - l_x)$
$y \geq 0$
y
u_x
l_x
u_x
x
$y \leq \lambda * (x - l_x)$
$y \geq x$

(a) $u_x \leq -l_x$ (b) $u_x > -l_x$

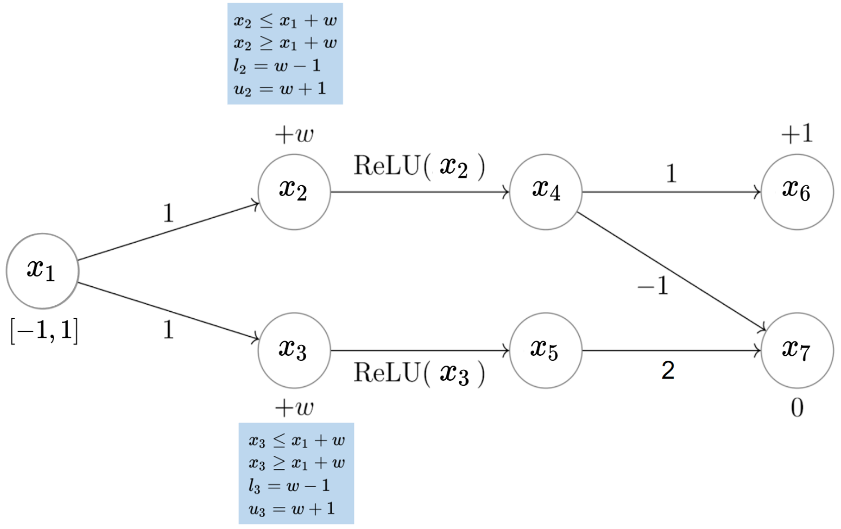
Figure 1: DeepPoly ReLU approximations

(b) Consider another heuristic for DeepPoly that always selects the relaxation from (a) in Fig. 1 regardless of the area. Compute the upper bound for x_7 as a function of the network parameter w using this simpler heuristic. Is the function continuous? Why or why not?

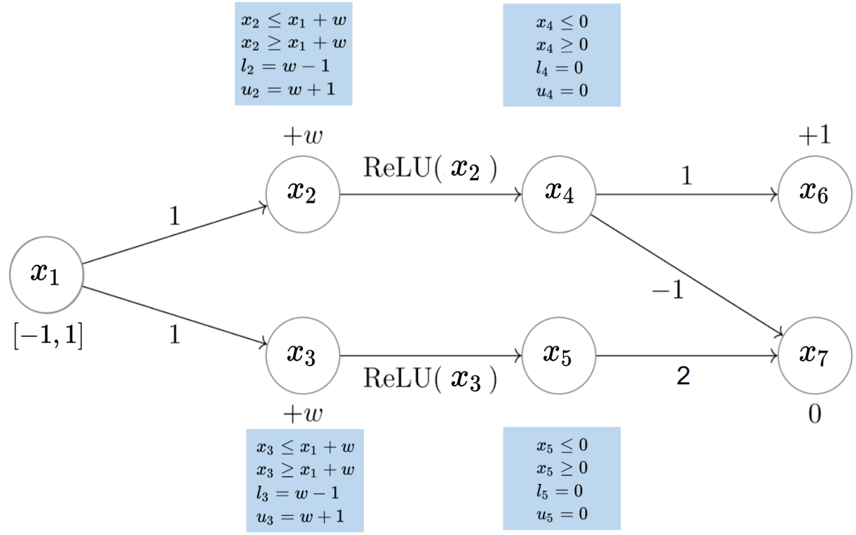
(c) Compute the upper bound for x_7 as a function of the network parameter w using the Box domain. Is the function continuous? Why or why not?

(d) Plot the functions for the different methods above (i.e the two DeepPoly heuristics and Box) on the same graph. What do you observe?

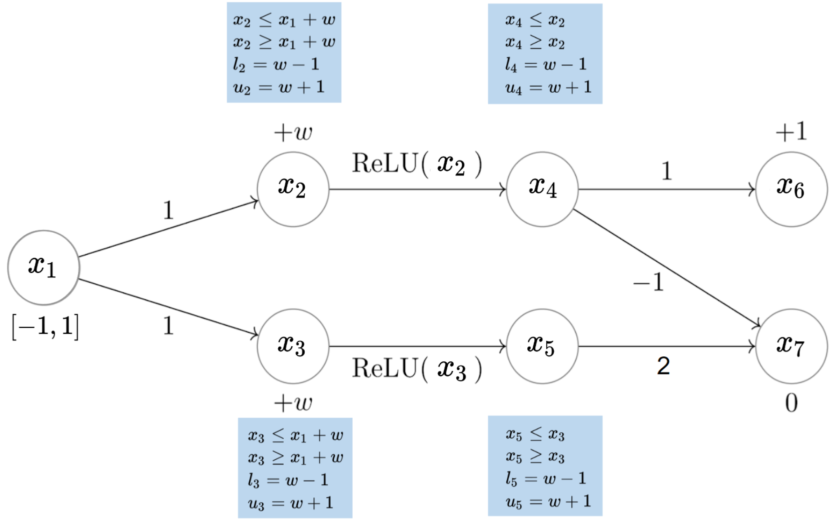
Solution 1. (a) We obtain the bounds of x_2 and x_3 with regular DeepPoly:



We then look at cases depending on the sign of the bounds of x_2 and x_3. First if $w < -1$, we have:



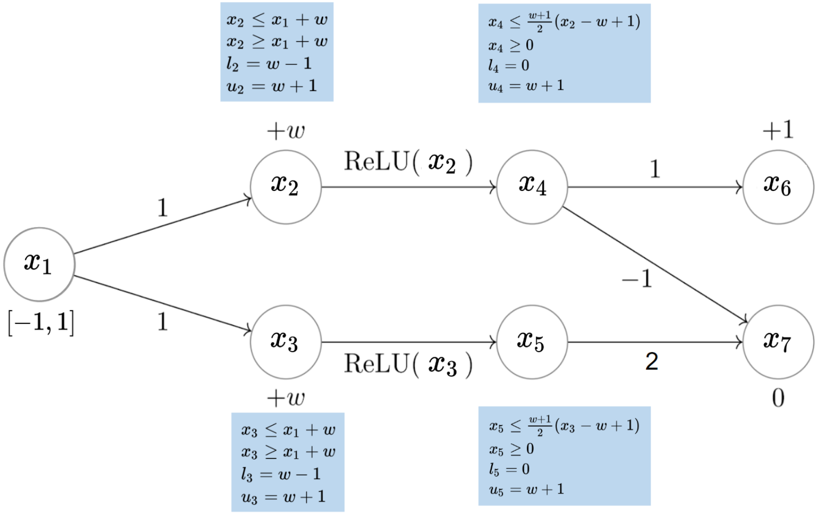
We calculate $u_7 = 0$. Similarly, when $w > 1$, we have:



We again use backsubstitution to calculate $x_7 \leq w + x_1$, so $u_7 = w + 1$ in this case.

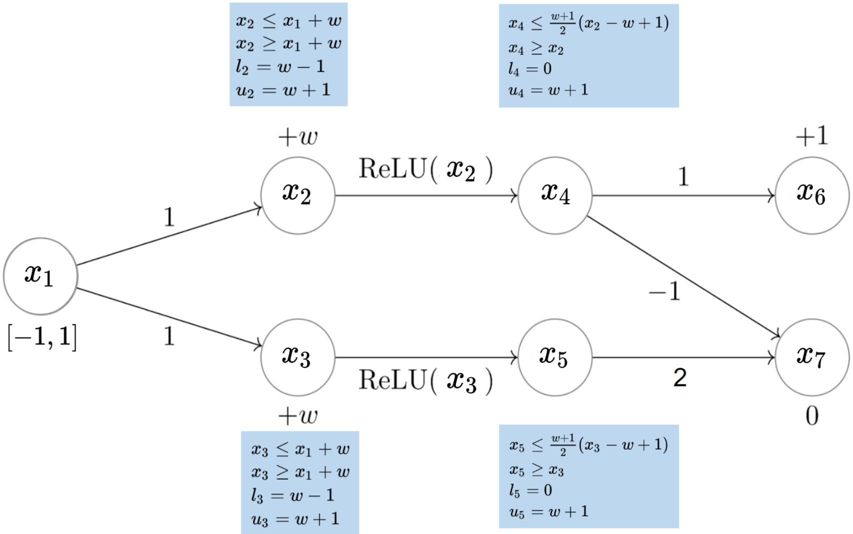
Finally, we look at the case when $-1 \leq w \leq 1$. To determine according to the heuristic which convex relaxation we will use in this case, we need to compare $u = w + 1$ and $-l = 1 - w$. It is easy to see that when $w \leq 0$, we are in case (a) in Fig. 1 and otherwise we are in case (b).

Let's first look at the subcase $-1 \leq w \leq 0$. Then, we have:



The backsubstitution in this case results in $x_7 \leq (w+1)x_1 + w + 1$. As $w + 1 \geq 0$, we get $u_7 = 2w + 2$.

Finally, when $0 < w \leq 1$ we have:



and $x_7 \leq wx_1 + 1$. As $w < 1$, we get $u_7 = w + 1$.

Putting everything together we have:

$$u_7(w) = \begin{cases} 0 & \text{if } w \in (-\infty, -1) \\ 2w+2 & \text{if } w \in [-1, 0] \\ w+1 & \text{if } w \in (0, \infty) \end{cases} . \tag{1}$$

The function is displayed in Fig. 2. Note that the DeepPoly area heuristic introduced a discontinuity at $w = 0$.

(b) The only difference with the previous analysis is that we apply the same convex relaxation for all $-1 \leq w \leq 1$, where as before we obtain $u_7 = 2w + 2$. Note this is due to the coefficient before x_1 in the upper bound being non-negative for all $-1 \leq w \leq 1$. The overall function then looks like:

$$u_7(w) = \begin{cases} 0 & \text{if } w \in (-\infty, -1) \\ 2w+2 & \text{if } w \in [-1, 1] \\ w+1 & \text{if } w \in (1, \infty) \end{cases} . \tag{2}$$

The function is displayed in Fig. 2. Note that this simple heuristic also introduces a discontinuity. This time at $w = 1$.

(c) Using standard box propagation, we have:

$$\begin{aligned} x_2 &= [w-1, w+1] \\ x_3 &= [w-1, w+1] \end{aligned}$$

Now we split the analysis into different cases.

First we look, at $w < -1$. Then we have:

$$\begin{aligned} x_4 &= [0, 0] \\ x_5 &= [0, 0] \\ x_7 &= [0, 0] \end{aligned}$$

Therefore, in this case $u_7 = 0$.

Next we look, at $-1 \leq w \leq 1$. Then we have:

$$\begin{aligned} x_4 &= [0, w+1] \\ x_5 &= [0, w+1] \\ x_7 &= [-w-1, 2w+2] \end{aligned}$$

Therefore, in this case $u_7 = 2w + 2$.

Finally, when $w > 1$ we have:

$$\begin{aligned} x_4 &= [w-1, w+1] \\ x_5 &= [w-1, w+1] \\ x_7 &= [w-3, w+3] \end{aligned}$$

Therefore, in this case $u_7 = w + 3$. Putting it together we have:

$$u_7(w) = \begin{cases} 0 & \text{if } w \in (-\infty, -1) \\ 2w+2 & \text{if } w \in [-1, 1] \\ w+3 & \text{if } w \in (1, \infty) \end{cases} . \tag{3}$$

Note that the function is continuous. This is always the case for bound functions produced using Box analysis, as the value of the function at the splits (where a lower or an upper bound of the analysis changes sign) must be 0 for both branches. For example, when the upper bound for x_4 switches from 0 to $w+1$, $w+1$ is 0 by construction at the point of switching ($w = -1$). Check [1] for detailed proof. The function is displayed in Fig. 2.

(d) We observe that both versions of DeepPoly are not continuous unlike Box which is always continuous. The simple heuristic that always picks the convex relaxation from (a) in Fig. 1 also has much worse upper bound (in the interval $w \in (0, 1]$), that is closer to the Box approximation, while still suffering from the discontinuity. In practice, the discontinuity of DeepPoly, perhaps counter-intuitively results in worse-performing certified trained networks compared to training with Box. There exists strategies to mitigate these discontinuities. Check [1] for more details.

[1] Jovanović, N., Balunović, M., Baader, M. and Vechev, M., On the Paradox of Certified Training.

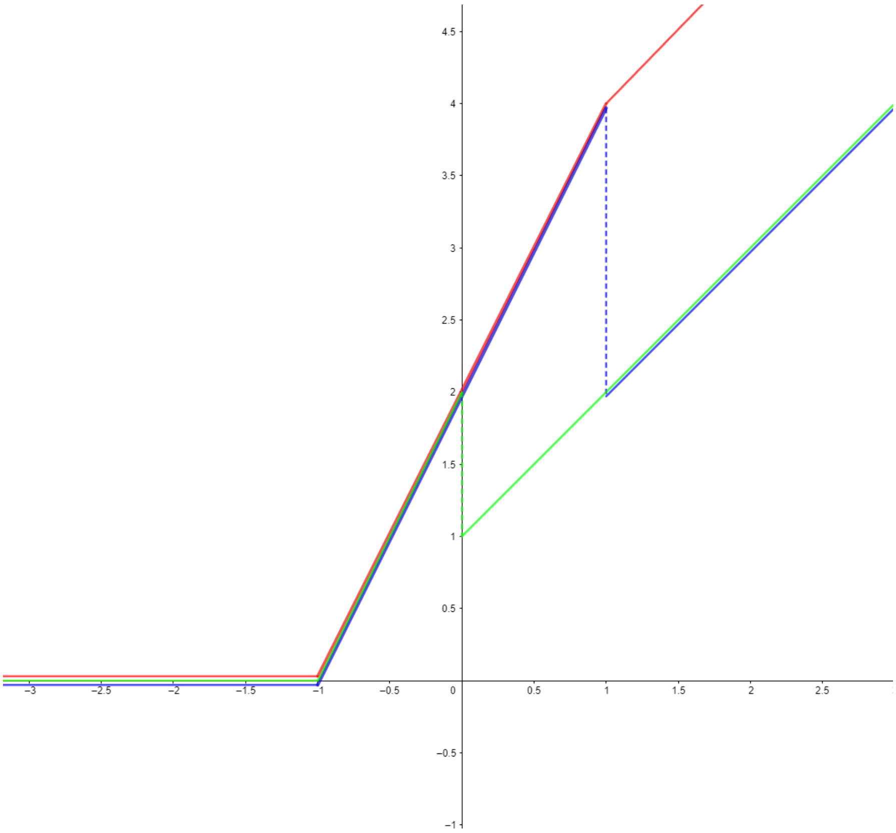


Figure 2: u_7 as a function of w. Red is box, green is DeepPoly with area approximation, blue is DeepPoly with simple heuristic