A Relationship between Probabilistic Safety and other Measures

In Proposition 4 we show that probabilistic safety, as defined in Definition 1, gives a lower bound to other measures commonly used to guarantee the absence of adversarial examples.

Proposition 4. *For $S \subseteq \mathbb{R}^{n_c}$ it holds that*

$$P_{safe}(T,S) \leq \inf_{x \in T} Prob(f^{\mathbf{w}}(x) \in S).$$

Moreover, if for $i \in \{1,...,n_c\}$, we assume that $S = \{y \in \mathbb{R}^{n_c} \,|\, y^i > a\}$. Then, it holds that

$$P_{safe}(T,S) \leq \frac{\inf_{x \in T} \mathbb{E}_{w \sim \mathbf{w}}[f_i^w(x)]}{a}.$$

Proof of Proposition 4

$$\begin{aligned}
&P_{safe}(T,S) = \\
&1 - Prob_{w \sim \mathbf{w}}(\exists x \in T, f^w(x) \in S) = \\
&1 - \mathbb{E}_{w \sim \mathbf{w}}[\sup_{x \in T} \mathbf{1}_S[f^w(x)]] \leq \\
&1 - \sup_{x \in T} \mathbb{E}_{w \sim \mathbf{w}}[\mathbf{1}_S[f^w(x)]] = \\
&1 - \sup_{x \in T} Prob_{w \sim \mathbf{w}}(f^w(x) \in S) = \\
&\inf_{x \in T} Prob_{w \sim \mathbf{w}}(f^w(x) \in S) = \\
&\inf_{x \in T} Prob_{w \sim \mathbf{w}}(f_i^w(x) \geq a) \leq \\
&\frac{\inf_{x \in T} \mathbb{E}_{w \sim \mathbf{w}}[f_i^w(x)]}{a},
\end{aligned}$$

where the last inequality is due to Markov's inequality.

B Computational Complexity

Algorithm 1 has a complexity which is linear in the number of samples, N, taken from the posterior distribution of w. The computational complexity of the method is then determined by the computational complexity of the method used to propagate a given interval $\hat{H}$ (that is, line 5 in Algorithm 1). The cost of performing IBP is $\mathcal{O}(Knm)$ where K is the number of hidden layers and $n \times m$ is the size of the largest weight matrix $W^{(k)}$, for $k = 0, \ldots, K$. LBP is instead $\mathcal{O}(K^2nm)$.

C Proofs

In this section of the Supplementary Material we provide proofs for the main propositions stated in the paper.

C.1 Proposition 2

The bounding box can be computed iteratively in the number of hidden layers of the network, K. We show how to compute the lower bound of the bounding box; the computation for the maximum is analogous.

Consider the k-th network layer, for $k = 0, \ldots, K$, we want to find for $i = 1, \ldots n_{k+1}$:

$$\min_{\substack{W_{i:}^{(k)} \in [W_{i:}^{(k),L}, W_{i:}^{(k),U}] \\ z^{(k)} \in [z^{(k),L}, z^{(k),U}] \\ b_i^{(k)} \in [b_i^{(k),L}, b_i^{(k),U}]}} z_i^{(k+1)} = \sigma\left(\sum_{j=1}^{n_k} W_{ij}^{(k)} z_j^{(k)} + b_i^{(k)}\right).$$

As the activation function σ is monotonic, it suffice to find the minimum of: $\sum_{j=1}^{n_k} W_{ij}^{(k)} z_j^{(k)} + b_i^{(k)}$. Since $W_{ij}^{(k)} z_j^{(k)}$ is a bi-linear form defined on an hyper-rectangle, it follows that it obtains its minimum in one of the four corners of the rectangle $[W_{ij}^{(k),L}, W_{ij}^{(k),U}] \times [z_j^{(k),L}, z_j^{(k),U}]$.
Let $t_{ij}^{(k),L} = \min\{W_{ij}^{(k),L} z_j^{(k),L}, W_{ij}^{(k),U} z_j^{(k),L}, W_{ij}^{(k),L} z_j^{(k),U}, W_{ij}^{(k),U} z_j^{(k),U}\}$ we hence have:

$$\sum_{j=1}^{n_k} W_{ij}^{(k)} z_j^{(k)} + b_i^{(k)} \geq \sum_{j=1}^{n_k} t_{ij}^{(k),L} + b_i^{(k),L} =: \zeta_i^{(k+1),L}.$$

Thus for every $W_{i:}^{(k)} \in [W_{i:}^{(k),L}, W_{i:}^{(k),U}]$, $z^{(k)} \in [z^{(k),L}, z^{(k),U}]$ and $b_i^{(k)} \in [b_i^{(k),L}, b_i^{(k),U}]$ we have:

$$\sigma\left(\sum_{j=1}^{n_k} W_{ij}^{(k)} z_j^{(k)} + b_i^{(k)}\right) \geq \sigma\left(\zeta_i^{(k+1),L}\right)$$

that is $z_i^{(k+1),L} = \sigma\left(\zeta_i^{(k+1),L}\right)$ is a lower bound to the solution of the minimisation problem posed above.

C.2 Proposition 3

We first state the following Lemma that follows directly from the definition of linear functions:

Lemma 2. *Let $f^L(t) = \sum_j a_j^L t_j + b^L$ and $f^U(t) = \sum_j a_j^U t_j + b^U$ be lower and upper LBFs to a function $g(t)$ $\forall t \in \mathcal{T}$, i.e. $f^L(t) \leq g(t) \leq f^U(t)$ $\forall t \in \mathcal{T}$. Consider two real coefficients $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$. Define*

$$\bar{a}_j^L = \begin{cases} \alpha a_j^L \text{ if } \alpha \geq 0 \\ \alpha a_j^U \text{ if } \alpha < 0 \end{cases} \quad \bar{b}^L = \begin{cases} \alpha b^L + \beta \text{ if } \alpha \geq 0 \\ \alpha b^U + \beta \text{ if } \alpha < 0 \end{cases} \tag{9}$$

$$\bar{a}_j^U = \begin{cases} \alpha a_j^U \text{ if } \alpha \geq 0 \\ \alpha a_j^L \text{ if } \alpha < 0 \end{cases} \quad \bar{b}^U = \begin{cases} \alpha b^U + \beta \text{ if } \alpha \geq 0 \\ \alpha b^L + \beta \text{ if } \alpha < 0 \end{cases} \tag{10}$$

Then:

$$\bar{f}^L(t) := \sum_j \bar{a}_j^L t_j + \bar{b}^L \leq \alpha g(t) + \beta \leq \sum_j \bar{a}_j^U t_j + \bar{b}^U =: \bar{f}^U(t)$$

That is, LBFs can be propagated through linear transformation by redefining the coefficients through Equations (9)–(10).

We now proof Proposition 3 iteratively on $k = 1, \ldots, K$ that is that for $i = 1, \ldots, n_k$ there exist $f_i^{(k),L}(x, W)$ and $f_i^{(k),U}(x, W)$ lower and upper LBFs such that:

$$\zeta_i^{(k)} \geq f_i^{(k),L}(x, W) := \mu_i^{(k),L} \cdot x + \sum_{l=0}^{k-2} \langle \nu_i^{(l,k),L}, W^{(l)} \rangle + \nu_i^{(k-1,k),L} \cdot W_{i:}^{(k-1)} + \lambda_i^{(k),L} \tag{11}$$

$$\zeta_i^{(k)} \leq f_i^{(k),U}(x, W) := \mu_i^{(k),U} \cdot x + \sum_{l=0}^{k-2} \langle \nu_i^{(l,k),U}, W^{(l)} \rangle + \nu_i^{(k-1,k),U} \cdot W_{i:}^{(k-1)} + \lambda_i^{(k),U} \tag{12}$$

and iteratively find valid values for the LBFs coefficients, i.e., $\mu_i^{(k),L}$, $\nu_i^{(l,k),L}$, $\lambda_i^{(k),L}$, $\mu_i^{(k),U}$, $\nu_i^{(l,k),U}$ and $\lambda_i^{(k),U}$.

For the first hidden-layer we have that $\zeta_i^{(1)} = \sum_j W_{ij}^{(0)} x_j + b_i^{(0)}$. By inequality (7) and using the lower bound for $b_i^{(0)}$ we have:

$$\zeta_i^{(1)} \geq \sum_j \left(W_{ij}^{(0),L} x_j + W_{ij}^{(0)} x_j^L - W_{ij}^{(0),L} x_j^L \right) + b_i^{(0),L}$$
$$= W_{i:}^{(0),L} \cdot x + W_{i:}^{(0)} \cdot x^L - W_{i:}^{(0),L} \cdot x^L + b_i^{(0),L}$$

which is a lower LBF on $\zeta^{(1)}$. Similarly using Equation (8) we obtain:

$$\zeta_i^{(1)} \leq W_{i:}^{(0),U} \cdot x + W_{i:}^{(0)} \cdot x^L - W_{i:}^{(0),U} \cdot x^L + b_i^{(0),U}$$

which is an upper LBF on $\zeta^{(1)}$. By setting:

$$\begin{aligned} \mu_i^{(1),L} &= W_{i:}^{(0),L} \quad , \quad \mu_i^{(1),U} = W_{i:}^{(0),U} \\ \nu_i^{(0,1),L} &= z^{(0),L} \quad , \quad \nu_i^{(0,1),U} = x^L \\ \lambda_i^{(1),L} &= -W_{i:}^{(0),L} \cdot x^L + b_i^{(0),L} \\ \lambda_i^{(1),U} &= -W_{i:}^{(0),U} \cdot x^L + b_i^{(0),U} \end{aligned}$$

we obtains LBFs $f_i^{(1),L}(x, W)$ and $f_i^{(1),U}(x, W)$ of the form (11)–(12).

Given the validity of Equations (11)–(12) up to a certain k we now show how to compute the LBF for layer $k+1$, that is given $f_i^{(k),L}(x, W)$ and $f_i^{(k),U}(x, W)$ we explicitly compute $f_i^{(k+1),L}(x, W)$ and $f_i^{(k+1),U}(x, W)$. Let $\zeta_i^{(k),L} = \min f_i^{(k),L}(x, W)$ and $\zeta_i^{(k),U} = \max f_i^{(k),U}(x, W)$ the minimum and maximum of the two LBFs (that can be computed analytically as the functions are linear). For Lemma 1 there exists a set of coefficients such that $z_i^{(k)} = \sigma(\zeta_i^{(k)}) \geq \alpha_i^{(k),L} \zeta_i^{(k)} + \beta_i^{(k),L}$. By Lemma 2 we know that there exists $\bar{f}_i^{(k),L}(x, W)$ with coefficients $\bar{\mu}_i^{(k),L}$, $\bar{\nu}_i^{(l,k),L}$, $\bar{\lambda}_i^{(k),L}$ obtained through Equations 9–10 such that:

$$z_i^{(k)} \geq \alpha_i^{(k),L} f_i^{(k),L}(x, W) + \beta_i^{(k),L} \geq \bar{f}_i^{(k),L}(x, W)$$

that is $\bar{f}_i^{(k),L}(x, W)$ is a lower LBF on $z_i^{(k)}$ with coefficients $\bar{\mu}_i^{(k),L}$, $\bar{\nu}_i^{(l,k),L}$, $\bar{\lambda}_i^{(k),L}$. Analogously let $\bar{f}_i^{(k),U}(x, W)$ be the upper LBF on $z_i^{(k)}$ computed in a similar way.

Consider now the bi-linear layer $\zeta_i^{(k+1)} = \sum_j W_{ij}^{(k)} z_j^{(k)} + b_i^{(k)}$. From Equation (7) we know that: $W_{ij}^{(k)} z_j^{(k)} \geq W_{ij}^{(k),L} z_j^{(k)} + W_{ij}^{(k)} z_j^{(k),L} - W_{ij}^{(k),L} z_j^{(k),L}$. By applying Lemma 2 with $\alpha = W_{ij}^{(k),L}$ and $\beta = 0$ we know that there exists a lower LBF $\hat{f}_{ij}^{(k),L}(x, W)$ with a set of coefficients $a_{ij}^{(k),L}$, $b_{ij}^{(l,k),L}$ and $c_{ij}^{(k),L}$ computed applying Equations (9)–(10) to $\bar{\mu}_i^{(k),L}$, $\bar{\nu}_i^{(l,k),L}$, $\bar{\lambda}_i^{(k),L}$ such that: $W_{ij}^{(k),L} z_j^{(k)} \geq \hat{f}_{ij}^{(k),L}(x, W)$. Hence we have:

$$\begin{aligned} \zeta_i^{(k+1)} &= \sum_j W_{ij}^{(k)} z_j^{(k)} + b_i^{(k)} \geq \sum_j \big(W_{ij}^{(k),L} z_j^{(k)} + \\ & W_{ij}^{(k)} z_j^{(k),L} - W_{ij}^{(k),L} z_j^{(k),L} \big) + b_i^{(k),L} \geq \\ & \sum_j \hat{f}_{ij}^{(k),L}(x, W) + \sum_j W_{ij}^{(k)} z_j^{(k),L} - \\ & \sum_j W_{ij}^{(k),L} z_j^{(k),L} + b_i^{(k),L} = \\ & \sum_j \big(a_{ij}^{(k),L} \cdot x + \sum_{l=0}^{k-2} \langle b_{ij}^{(l,k),L}, W^{(l)} \rangle \\ & + b_{ij}^{kl-1,k),L} \cdot W_{j:}^{(k-1)} + c_{ij}^{(k),L} \big) + \\ & W_{i:}^{(k)} \cdot z^{(k),L} - W_{i:}^{(k),L} z^{(k),L}. \end{aligned}$$

By setting

$$\begin{aligned} \mu_i^{(k+1),L} &= \sum_j a_{ij}^{(k),L} \\ \nu_i^{(l,k+1),L} &= \sum_j b_{ij}^{(l,k),L} \quad k = 0, \ldots, l-2 \\ \nu_i^{(k-1,k+1),L} &= b_i^{(k-1,k),L} \\ \nu_i^{(k,k+1),L} &= z^{(k),L} \\ \lambda_i^{(k+1),L} &= \sum_j c_{ij}^{(k),L} - W_{i:}^{(k),L} \cdot z^{(k),L} + b_i^{(k),L} \end{aligned}$$

and re-arranging the elements in the above inequality, we finally obtain:

$$\zeta_i^{(k+1)} \geq \mu_i^{(k+1),L} \cdot x + \sum_{l=0}^{k-1} \langle \nu_i^{(l,k+1),L}, W^{(l)} \rangle + \nu_i^{(k,k+1),L} \cdot W_{i:}^{(k)} + \lambda_i^{k+1),L} =: f_i^{(k+1),L}(x, W)$$

which is of the form of Equation (11) for the lower LBF for the $k+1$-th layer. Similarly an upper LBF of the form of Equation (12) can be obtained by using Equation (8) in the chain of inequalities above.

D Linear Specifications

In this section of the Supplementary Material we discuss how the output of IBP and LBP can be used to check against specification of the form of Assumption 1, that is of the form:

$$C_S f^w(x) + d_S \geq 0 \quad \forall x \in T \, \forall w \in \bar{H}$$

with $C_S \in \mathbb{R}^{n_S \times n_c}$ and $d_S \in \mathbb{R}^{n_S}$, $\bar{H} = [w^L, w^U]$ and $T = [x^L, x^U]$. Let $i = 1, \ldots, n_S$ then we need to check for every i whether: $C_{S,i} \cdot f^w(x) + d_{S,i} \geq 0$. Which is equivalent to compute

$$\min_{x \in T w \in \bar{H}} C_{S,i} \cdot f^w(x) + d_{S,i} = \sum_{j=1}^{n_c} C_{S,ij} f_j^w(x) + d_{S,i} \tag{13}$$

and checking whether that is greater or equal to zero or not. As presented in the main paper, Propositions 2 and 3 return a bounding box for the final output. Though this bounding box can be directly used to compute the minimum in Equation (13) (as this entails simply the minimisation of a linear function on a rectangular space), tighter bounds can be obtained both for IBP and for LBP. This is described in the following two subsections.

D.1 IBP

For IBP we can do something similar to what is done in the case of IBP for deterministic NN [13]. Instead of propagating the bounding box up until the very last layer to $f_i^w(x) = \zeta_i^{K+1}$, we stop at the last hidden activation function values $z_i^{(K)}$. We thus have:

$$\sum_{j=1}^{n_c} C_{S,ij} f_j^w(x) + d_{S,i} = \sum_{j=1}^{n_c} C_{S,ij} \zeta_j^{(K+1)} + d_{S,i} =$$
$$\sum_{j=1}^{n_c} C_{S,ij} \sum_{l=1}^{n_K} \left(W_{jl}^{(K)} z_l^{(K)} + b_j^{(K)} \right) + d_{S,i} =$$
$$\sum_{l=1}^{n_K} \left(\sum_{j=1}^{n_c} C_{S,ij} W_{jl}^{(K)} \right) z_l^{(K)} + \sum_{j=1}^{n_c} C_{S,ij} b_j^{(K)} + d_{S,i}.$$

Notice that $\sum_{j=1}^{n_c} C_{S,ij} W_{jl}^{(K)}$ and $\sum_{j=1}^{n_c} C_{S,ij} b_j^{(K)}$ are linear transformation, of the weights and biases. The lower and upper bounds on $W^{(K)}$ and $b^{(K)}$ can hence be propagated through this two functions to obtain lower and upper bounds that account for the specification $W_S^{(K),L}$, $W_S^{(K),U}$, $b_S^{(K),L}$ and $b_S^{(K),U}$. Propagating that interval through the layer:

$$\min_{\substack{z^{(K)} \in [z^{(K),L}, z^{(K),U}] \\ W_S^{(K)} \in [W_S^{(K),L}, W_S^{(K),U}] \\ b_S^{(K)} \in [b_S^{(K),L}, b_S^{(K),U}]}} W_S^{(K)} \cdot z^{(K)} + b_S^{(K)}$$

gives a solution to Equation (13).

D.2 LBP

For LBP one can simply proceed by propagating the linear bound obtained. In fact, Proposition 3 yields an upper and lower LBFs, $f^{(K+1),L}(x, w)$ and $f^{(K+1),U}(x, w)$ on $f^w(x)$ for every $x \in T$ and $w \in \hat{H}$. By Lemma 2 those two LBFs can simply be propagated through the linear specification of Equation (13) hence obtaining lower and upper LBFs on the full specification, which can then be minimised to checked for the validity of the interval $\hat{H}$.

E Computational Resources

All our experiments were conducted on a server equipped with two 24 core Intel Xenon 6252 processors and 256GB of RAM. For VCAS experiments no parallelization was necessary, whereas MNIST was parallelized over 25 concurrent threads.