6 APPENDIX

6.1 BOUND TIGHTENING

The quality of the bound in theorem 1 depends crucially on having bounds on all the intermediate pre and post activations z^l, x^l. In this section, we describe how these bounds may be derived. One simple way to compute bounds on neural activations is to use interval arithmetic given bounds on the input $\underline{x}^0 \leq x \leq \overline{x}^0$. Bounds at each layer can then be computed recursively for $l = 0, \ldots, L-1$ as follows:

$$\underline{z}^l = \left[W^l\right]_+ \underline{x}^l + \left[W^l\right]_- \overline{x}^l + b^l \tag{11a}$$

$$\overline{z}^l = \left[W^l\right]_+ \overline{x}^l + \left[W^l\right]_- \underline{z}^l + b^l \tag{11b}$$

$$\underline{x}^{l+1} = h^l\left(\underline{z}^l\right) \tag{11c}$$

$$\overline{x}^{l+1} = h^l\left(\overline{z}^l\right) \tag{11d}$$

However, these bounds could be quite loose and could be improved by solving the optimization problem

$$\max_{\substack{z^0,\ldots,z^{L-1}\\ x^1,\ldots,x^{L-1}}} x_k^l \tag{12a}$$

$$\text{s.t (5b), (5c), (5d)} \tag{12b}$$

We can relax this problem using dual relaxation approach from the previous section and the optimal value of the relaxation would provide a new (possibly tighter) upper bound on x_k^l than $\overline{x}_k^l$. Further, since our relaxation approach is anytime (ie for any choice of dual variables we obtain a valid bound), we can stop the computation at any time and use the resulting bounds. This is a significant advantage compared to previous approaches used in [Bunel et al., 2017]. Similarly, one can obtain tighter upper and lower bounds on x^l for each value of l, k. Given these bounds, one can infer tighter bounds on z^l using (11).

Plugging these tightened bounds back into (6) and rerunning the dual optimization, we can compute a tighter upper bound on the verification objective.

6.2 CONJUGATES OF TRANSFER FUNCTIONS

We are interested in computing $g^\star = \max_{y\in[\underline{y},\overline{y}]} g(y) = \mu y - \lambda h(y)$. This is a one dimensional optimization problem and can be computed via brute-force discretization of the input domain in general. However, for most commonly used transfer functions, this can be computed analytically. We derive the analytical solution for various commonly used transfer functions:

ReLUs: If h is a ReLU, $g(y)$ is piecewise linear and specifically is linear on $[\underline{y}, 0]$ and on $[0, \overline{y}]$ (assuming that $0 \in [\underline{y}, \overline{y}]$, else $g(y)$ is simply linear and can be optimized by setting y to one of its bounds). On each linear piece, the optimum is attained at one of the endpoints of the input domain. Thus, the overall maximum can be obtained by evaluating g at $\underline{y}, \overline{y}, 0$ (if $0 \in [\underline{y}, \overline{y}]$). Thus,

$$g^\star = \begin{cases} \max\left(g(\underline{y}), g(\overline{y}), g(0)\right) & \text{if } 0 \in [\underline{y}, \overline{y}] \\ \max\left(g(\underline{y}), g(\overline{y})\right) & \text{otherwise} \end{cases}.$$

Sigmoid: If h is a sigmoid, we can consider two cases: a) The optimum of g is obtained at one of its input bounds or b) The optimum of g is obtained at a point strictly inside the interval $[\underline{y}, \overline{y}]$. In case (b), we require that the derivative of g vanishes, i.e:$\mu - \lambda\sigma(y)(1-\sigma(y)) = 0$. Since $\sigma(y) \in [0,1]$, this equation only has a solution if $\lambda \neq 0$ and $\frac{\mu}{\lambda} \in \left[0, \frac{1}{4}\right]$ In this case, the solutions are $\sigma(y) = \frac{1\pm\sqrt{1-\frac{4\mu}{\lambda}}}{2}$ Solving for y, we obtain $y = \sigma^{-1}\left(\frac{1\pm\sqrt{1-\frac{4\mu}{\lambda}}}{2}\right)$ where σ^{-1} is the logit function $\sigma^{-1}(t) = \log\left(\frac{t}{1-t}\right)$. We only consider these solutions if they lie within the domain $[\underline{y}, \overline{y}]$. Define:

$$y_1(\mu,\lambda) = \max\left(\underline{y}, \min\left(\overline{y}, \frac{1-\sqrt{1-\frac{4\mu}{\lambda}}}{2}\right)\right)$$

$$y_2(\mu,\lambda) = \max\left(\underline{y}, \min\left(\overline{y}, \frac{1+\sqrt{1-\frac{4\mu}{\lambda}}}{2}\right)\right)$$

Thus, we obtain the following expression for $g^\star$:

$$\begin{cases} \max\left(g\left(\underline{y}\right), g\left(\overline{y}\right)\right) \text{ if } \lambda = 0 \text{ or } \frac{\mu}{\lambda} \notin [0, \frac{1}{4}] \\ \max\left(g\left(\underline{y}\right), g\left(\overline{y}\right), g\left(y_1(\mu,\lambda)\right), g\left(y_2(\mu,\lambda)\right)\right) \text{ otherwise} \end{cases}$$

Tanh: Define

$$y_1(\mu,\lambda) = \max\left(\underline{y}, \min\left(\overline{y}, \operatorname{arctanh}\left(\sqrt{1-\frac{\mu}{\lambda}}\right)\right)\right)$$

$$y_2(\mu,\lambda) = \max\left(\underline{y}, \min\left(\overline{y}, \operatorname{arctanh}\left(\sqrt{1-\frac{\mu}{\lambda}}\right)\right)\right)$$

and obtain $g^\star$ to be

$$\begin{cases} \max\left(g\left(\underline{y}\right), g\left(\overline{y}\right)\right) \text{ if } \lambda = 0 \text{ or } \frac{\mu}{\lambda} \notin [0, 1] \\ \max\left(g\left(\underline{y}\right), g\left(\overline{y}\right), g\left(y_1(\mu,\lambda)\right), g\left(y_2(\mu,\lambda)\right)\right) \\ \text{otherwise} \end{cases}$$

6.2.1 MaxPool

If h is a max-pool, we need to deal with it layer-wise and not component-wise. We have $h(y) = \max(y_1, \ldots, y_t)$ and are interested in solving for

$$g^\star = \max_{y\in[\underline{y},\overline{y}]} (\mu)^T y - \lambda h(y)$$

Note here that λ is a scalar while μ is a vector. This can be solved by considering the case of each component of y attaining the maximumum separately. We look at the case where y_i attains the maximum below:

$$\max_{y\in[\underline{y},\overline{y}],y_i\geq y_j\ \forall j\neq i} (\mu)^T y - \lambda y_i$$

Fixing y_i and optimizing the other coordinates, we obtain

$$\begin{aligned}\max_{y_i\in[\underline{y}_i,\overline{y}_i]} &\sum_{j\neq i,y_i\geq\overline{y}_j} \max\left(\mu_j\overline{y}_j,\mu_j\underline{y}_j\right)\\ &+\sum_{j\neq i,y_i\leq\overline{y}_j} \max\left(\mu_j y_i,\mu_j\underline{y}_j\right)\\ &-(\mu_i-\lambda)\,y_i\end{aligned}$$

This one-dimensional function can be optimized via binary search on y_i. After solving for each i, taking the maximum over i gives the value $g^\star$

6.2.2 Upper bounds for general nonlinearities

In general, we can compute an upper bound on g even if we cannot optimize it exactly. The idea is to decouple the y from the two terms in g: μy and $-\lambda h(y)$ and optimize each independently. However, this gives a very weak bound. This can be made much tighter by applying it separately to a decomposition of the input domain $[\underline{y},\overline{y}] = \cup_i[a_i,b_i]$:

$$\max_{y\in[a_i,b_i]} g(y) \leq \max(\mu a_i,\mu b_i)+\max\left(-\lambda h(a_i),-\lambda h(b_i)\right)$$

Finally we can bound $\max_{[\underline{y},\overline{y}]} g(y)$ using

$$\begin{aligned}&\max_{y\in[\underline{y},\overline{y}]} g(y) \leq\\ &\max_i \left(\max(\mu a_i,\mu b_i)+\max\left(-\lambda h(a_i),-\lambda h(b_i)\right)\right)\end{aligned}$$

As the decomposition gets finer, ie, $|a_i-b_i|\to 0$, we obtain an arbitrarily tight upper bound on g) this way.

6.3 OPTIMIZING OVER THE INPUT CONSTRAINTS

In this section, we discuss solving the optimization problem defining $f_0(\mu^0)$ in (7)

$$\max_{x\in\mathcal{S}_{in}} (W^T\mu)^T x + b^T\mu$$

where we dropped the superscript (0) for brevity. For commonly occuring constraint sets $\mathcal{S}_{in}$, this problem can be solved in closed form easily:
Norm constraints: Consider the case $\mathcal{S}_{in} = \|x-\hat{x}\| \leq \epsilon$. In this case, we by Holder's inequality, the objective is larger than or equal to

$$b^T\mu - \left\|W^T\mu\right\|_\star$$

where $\|\cdot\|_\star$ is the dual norm to the norm $\|\|$. Further this bound can be achieved for an appropriate choice of x. Hence, the optimal value is precisely $b^T\mu - \left\|W^T\mu\right\|_\star$.
Combinatorial objects: Linear objectives can be optimized efficiently over several combinatorial structures. For example, if x is indexed by the edges in a graph and $\mathcal{S}_{in}$ imposes constraints that x is binary valued and that the edges set to 1 should form a spanning tree of the graph, the optimal value can be computed using a maximum spanning tree approach.
Cardinality constraints: x may have cardinality constrained imposed on it: $\|x\|_0 \leq k$, saying that at most k elements of x can be non-zero. If further we have bounds on $x \in [\underline{x},\overline{x}]$, then the optimization problem can be solved as follows: Let $v(\mu) = \left[W^T\mu\right]_+ \odot \overline{x} + \left[W^T\mu\right]_- \odot \underline{x}$ and let $[v]_i$ denote the i-th largest component of v. Then the optimal value is $\sum_{i=1}^k [v(\mu)]_i + b^T\mu$.

6.4 PROOFS OF THEORETICAL RESULTS

6.4.1 NP-hardness of a verification of a single hidden layer network

Consider the case of sigmoid transfer function with an ∞ norm perturbation. Then, the verification problem reduces to :

$$\begin{aligned}&\max_{x^{in},z} \sum_i c_i h_i(z_i)\\ &\quad\text{s.t } z = Wx+b, \left\|x^{in}-x^{nom}\right\|_\infty \leq \epsilon\end{aligned}$$

This is an instance of a sigmoidal programming problem, which is proved to be NP-hard in Udell and Boyd [2013]

6.4.2 Proof of theorem 2

Proof. In this case, since h is a relu, $\tilde{f}$ can be written as (from section 6.2) as

$$\tilde{f}_{l,k}(\lambda,\mu) = \begin{cases}\max\left(\mu\underline{z}_k^l,(\mu-\lambda)\,\overline{z}_k^l,0\right) & \text{if } 0\in[\underline{z}_k^l,\overline{z}_k^l]\\ \max\left(\mu\overline{z}_k^l,\mu\underline{z}_k^l\right) & \text{if } 0 \text{ if } \overline{z}_k^l\leq 0\\ \max\left((\mu-\lambda)\,\underline{z}_k^l,(\mu-\lambda)\,\overline{z}_k^l\right) & \text{if } \underline{z}^l\geq 0\end{cases} \tag{13}$$

Now, the LP relaxation from [Ehlers, 2017] can be written as

$$\begin{aligned}
&\max c^T x^L \\
&\text{s.t } z^l = W^l x^l + b^l, l = 0, \ldots, L-1 \\
&x_k^{l+1} = z_k^l \quad \forall l, k \text{ s.t } \underline{z}_k^l \geq 0 \\
&x_k^{l+1} = 0 \text{ if } \quad \forall l, k \text{ s.t } \overline{z}_k^l \leq 0 \\
&x_k^{l+1} \geq 0, x_k^{l+1} \geq z_k^l, x_k^{l+1} \leq \left(z_k^l - \underline{z}_k^l\right)\left(\frac{\overline{z}_k^l}{\overline{z}_k^l - \underline{z}_k^l}\right) \\
&\qquad \forall l, k \text{ s.t } 0 \in [\underline{z}_k^l, \overline{z}_k^l] \\
&\underline{z}^l \leq z^l \leq \overline{z}^l, \underline{x}^l \leq x^l \leq \overline{x}^l \quad \forall l
\end{aligned}$$

We can rewrite this optimization problem as

$$\begin{aligned}
&\max c^T x^L \\
&\text{s.t } z^l = W^l x^l + b^l, l = 0, \ldots, L-1 \\
&x_k^{l+1} = z_k^l \quad \forall l, k \text{ s.t } \underline{z}_k^l \geq 0 \\
&x_k^{l+1} = 0 \text{ if } \quad \forall l, k \text{ s.t } \overline{z}_k^l \leq 0 \\
&x_k^{l+1} \geq \max\left(0, z_k^l\right), x_k^{l+1} \leq \left(z_k^l - \underline{z}_k^l\right)\left(\frac{\overline{z}_k^l}{\overline{z}_k^l - \underline{z}_k^l}\right) \\
&\qquad \forall l, k \text{ s.t } 0 \in [\underline{z}_k^l, \overline{z}_k^l] \\
&\underline{z}^l \leq z^l \leq \overline{z}^l, \underline{x}^l \leq x^l \leq \overline{x}^l \quad \forall l
\end{aligned}$$

which is still a convex optimization problem, since all the constraints are either linear of the form $\max(0, z) \leq x$ which is a convex constraint since the LHS is a convex function and the RHS is linear.

Taking the dual of this optimization problem, we obtain

$$\begin{aligned}
\max_{x,z}\ & c^T x^L + \sum_l \left(\mu^l\right)^T \left(z^l - W^l x^l - b^l\right) \\
&+ \sum_{l,k:\underline{z}_k^l \geq 0} \lambda_k^l \left(x_k^{l+1} - z_k^l\right) \\
&+ \sum_{l,k:\underline{z}_k^l \leq 0} \lambda_k^l \left(x_k^{l+1}\right) \\
&+ \sum_{l,k:0 \in [\underline{z}_k^l, \overline{z}_k^l]} \left(\lambda_{k;a}^l \left(x_k^{l+1} - \max\left(z_k^l, 0\right)\right)\right) \\
&+ \sum_{l,k:0 \in [\underline{z}_k^l, \overline{z}_k^l]} -\left(x_k^{l+1} - s_k^l \left(z_k^l - \underline{z}_k^l\right)\right) \lambda_{k;b}^l \\
\text{s.t}\quad & \underline{z}^l \leq z^l \leq \overline{z}^l, \underline{x}^l \leq x^l \leq \overline{x}^l \quad \forall l
\end{aligned}$$

where $s_k^l = \frac{\overline{z}_k^l}{\overline{z}_k^l - \underline{z}_k^l}$. Let $\lambda_k^l = \lambda_{k;a}^l - \lambda_{k;b}^l$ for l, k such that $0 \in [\underline{z}_k^l, \overline{z}_k^l]$. Further let $\mathcal{I}_{l,k} = [\underline{z}_k^l, \overline{z}_k^l]$ and let $\mathcal{I}_a$ denote the set of l, k such that $\underline{z}_k^l \geq 0$, $\mathcal{I}_b$ the set of l, k such that $\overline{z}_k^l \leq 0$ and $\mathcal{I}_c$ the set of l, k such that $0 \in \mathcal{I}_{l,k}$.

We can then rewrite the dual as

$$\begin{aligned}
&\max_x c^T x^L + \sum_{l=0}^{L-1} \left(x^l\right)^T \left(\lambda^l - \left(W^l\right)^T \mu^l\right) - \sum_l \left(\mu^l\right)^T b^l \\
&+ \sum_{l,k \in \mathcal{I}_a} \max_{z_k^l \in [\underline{z}_k^l, \overline{z}_k^l]} \left(\mu_k^l - \lambda_k^l\right) z_k^l \\
&+ \sum_{l,k \in \mathcal{I}_b} \max_{z_k^l \in [\underline{z}_k^l, \overline{z}_k^l]} \left(\mu_k^l\right) z_k^l \\
&+ \sum_{l,k \in \mathcal{I}_c} \max_{z_k^l \in \mathcal{I}_{l,k}} \left(\mu_k^l + \lambda_{k;b}^l s_k^l\right) z_k^l - \left(\lambda_{k;a}^l\right) \max\left(z_k^l, 0\right) \\
&+ \sum_{l,k \in \mathcal{I}_c} -\lambda_{k,b}^l s_k^l \underline{z}_k^l \\
&\text{s.t } \underline{x}^l \leq x^l \leq \overline{x}^l \quad \forall l
\end{aligned}$$

We now solve for the maximum over z_k^l considering three cases:
(a) $l, k \in \mathcal{I}_a$: The maximization is over a linear function and the maximum is attained at one of the bounds, hence the maximum evaluates to

$$\max\left(\left(\mu_k^l - \lambda_k^l\right) \underline{z}_k^l, \left(\mu_k^l - \lambda_k^l\right) \overline{z}_k^l\right)$$

This evaluates to $\tilde{f}_{l,k}\left(\lambda_k^l, \mu_k^l\right)$ for $l, k \in \mathcal{I}_a$. (b) $l, k \in \mathcal{I}_b$: The maximization is over a linear function and the maximum is attained at one of the bounds, hence the maximum evaluates to

$$\max\left(\left(\mu_k^l\right) \underline{z}_k^l, \left(\mu_k^l\right) \overline{z}_k^l\right)$$

This evaluates to $\tilde{f}_{l,k}\left(\lambda_k^l, \mu_k^l\right)$ for $l, k \in \mathcal{I}_b$. (c) $l, k \in \mathcal{I}_b$: The maximization is over a piecewise linear function and the maximum is attained at one of the bounds or at the breakpoint 0, hence the maximum evaluates to

$$\max\left(0, \left(\mu_k^l + \lambda_{k,b}^l s_k^l\right) \underline{z}_k^l, \left(\mu_k^l + \lambda_{k,b}^l \left(s_k^l - 1\right) - \lambda_k^l\right) \overline{z}_k^l\right)$$

Adding the constant $-\lambda_{k,b}^l s_k^l \underline{z}_k^l$ we obtain

$$\max\left(-\lambda_{k,b}^l s_k^l \underline{z}_k^l, \mu_k^l \underline{z}_k^l, \left(\mu_k^l - \lambda_k^l\right) \overline{z}_k^l\right)$$

Minimizing the above expression with respect to $\lambda_{k;b}^l$ subject to $\lambda_{k;b}^l \geq \max\left(0, -\lambda_k^l\right)$, we obtain

$$\max\left(0, \lambda_k^l \frac{\overline{z}_k^l \underline{z}_k^l}{\overline{z}_k^l - \underline{z}_k^l}, \mu_k^l \underline{z}_k^l, \left(\mu_k^l - \lambda_k^l\right) \overline{z}_k^l\right)$$

(since the expression is monotonically increasing in $\lambda_{k;b}^l$, minimizing it subject to these constraints we just set $\lambda_{k;b}^l$ to the larger of its lower bounds). Now, for the second term to attain the maximum, we require that $\lambda_k^l \leq 0, \mu_k^l = s_k^l \lambda_k^l$, in which case all the last three terms attain the maximum. Thus, the maximum is equal to the max of three terms:

$$\max\left(0, \mu_k^l \underline{z}_k^l, \left(\mu_k^l - \lambda_k^l\right) \overline{z}_k^l\right)$$

showing that the expression evaluts to $\tilde{f}_{l,k}\left(\lambda_k^l, \mu_k^l\right)$ for $l, k \in \mathcal{I}_c$.

Thus, the dual objective is equal to

$$\begin{aligned}
&\max_{x:\underline{x}^l \leq x^l \leq \overline{x}^l} c^T x^L + \sum_{l=0}^{L-1} \left(x^l\right)^T \left(\lambda^l - \left(W^l\right)^T \mu^l\right) \\
&- \sum_l \left(\mu^l\right)^T b^l \\
&+ \sum_{l,k} \tilde{f}_{l,k}\left(\lambda_k^l, \mu_k^l\right)
\end{aligned}$$

Given this, the rest of the dual exactly matches the calculations from section 3.3.

□

6.4.3 Proof of theorem 3

Proof. We leverage results from [Polyak, 2003] which argues that a smooth nonlinear function can be efficicently optimized over a "small enough" ball. Specifically, we use theorem 7 from [Polyak, 2003]. In order to apply the theorem, we need to bound the Lipschitz costant of the derivative of the function $f(x) = \sum_i c_i h_i(W_i x + b_i)$. Writing down the derivative, we obtain

$$f'(x) = W^T \operatorname{diag}(c)\, h'(Wx + b)$$

Thus,

$$\begin{aligned}
&f'(x) - f'(y) \\
&= W^T \operatorname{diag}(c)\left(h'(Wx + b) - h'(Wy + b)\right)
\end{aligned}$$

We have

$$h_i'(W_i x + b_i) - h_i'(W_i y + b_i) = h_i''(t)\left(W_i(x - y)\right)$$

where $t \in [W_i x + b_i, W_i y + b_i]$. Thus, we have

$$|h_i'(W_i x + b_i) - h_i'(W_i y + b_i)| \leq \gamma_i |W_i(x - y)|^2$$

So that

$$\begin{aligned}
&\|f'(x) - f'(y)\|_2 \\
&\leq \sigma_{\max}\left(W^T \operatorname{diag}(c)\right) \|h'(Wx + b) - h'(Wy + b)\|_2 \\
&= \sigma_{\max}\left(W^T \operatorname{diag}(c)\right) \times \\
&\qquad \sqrt{\sum_i \left(h_i'(W_i x + b_i) - h_i'(W_i y + b_i)\right)^2} \\
&\leq \sigma_{\max}\left(W^T \operatorname{diag}(c)\right) \gamma \sqrt{\sum_i |\gamma_i W_i(x - y)|^2} \\
&= \sigma_{\max}\left(W^T \operatorname{diag}(c)\right) \|\operatorname{diag}(\gamma)\, W(x - y)\|_2\, \gamma \\
&\leq \sigma_{\max}\left(W^T \operatorname{diag}(c)\right) \sigma_{\max}\left(\operatorname{diag}(\gamma)\, W\right) \|x - y\|_2
\end{aligned}$$

Thus f' is Lipschitz with Lipschitz constant $\sigma_{\max}\left(W^T \operatorname{diag}(c)\right) \sigma_{\max}(W)\, \gamma$. Further

$$\left\|f'\left(x^0\right)\right\| = \left\|W^T \operatorname{diag}(c)\, h'\left(Wx^0 + b\right)\right\|$$

Hence by theorem 7 from [Polyak, 2003], the theorem follows. □

6.4.4 PROOF OF THEOREM 4

Proof. We have $\|z\|_2 \leq \epsilon \implies |W_i z| \leq \|W_i\|_2 \epsilon$ (by Cauchy-Schwartz). Thus, for eacn i, we have

$$\begin{aligned}
&h_i(z_i^{nom} + W_i z) = \\
&= h_i(z_i^{nom}) + h_i'(z_i^{nom}) W_i z + \frac{h_i''(z_i^{nom})}{2} \left(W_i z\right)^2 \\
&+ \frac{1}{6} h_i'''(\tilde{x}_i^0 + t_i\left(W_i z\right)) \left(W_i z\right)^3
\end{aligned}$$

The last term can be bounded above by

$$\frac{\eta_i}{6} |W_i z|^3 \leq \frac{\eta_i}{6} \zeta_i^3 \epsilon^3$$

Adding the error terms over all terms in the objective function, we obtain the result. □