

Appendix¹

Contents

A.	Theorems and Algorithms	2
A.1	Vanilla Recurrent Neural Network	2
A.2	Algorithms for Recurrent Neural Networks	11
A.3	Long Short-term Memory Network	12
A.4	Bounding Planes for 2D Nonlinear Activation Functions	24
A.5	Algorithms for Long Short-term Memory Networks	26
A.6	Long Short-term Memory Network (1D bounding lines)	29
A.7	Gated Recurrent Unit Network	34
B.	Experiments	43
B.1	Significance of plane (2D)-bounding POPQORN certificates	43
B.2	CLEVER-RNN Score	44
B.3	Adapted C&W Attack	45
B.4	Experimental Details	46

¹Supplementary materials of *POPQORN: Quantifying Robustness of Recurrent Neural Networks*. Copyright 2019 by the authors.

A. Theorems and Algorithms

A.1 Vanilla Recurrent Neural Network

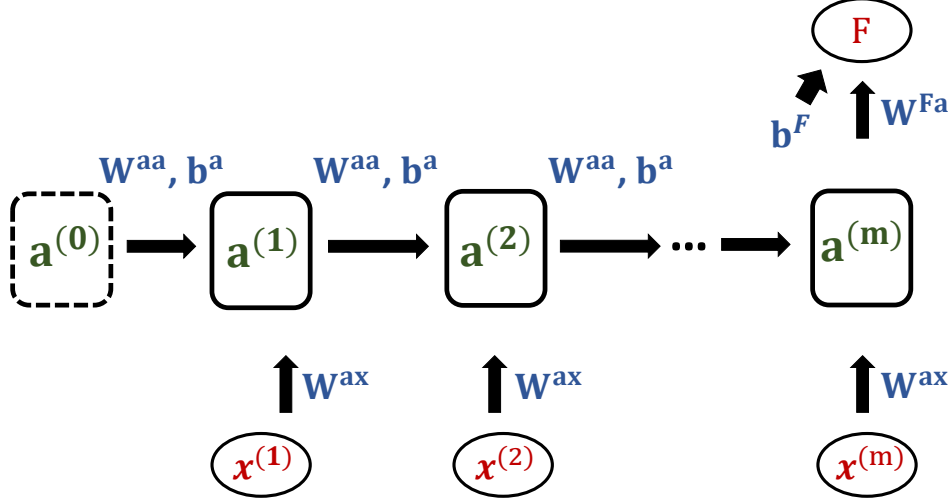


Figure 1: Graphical depiction of an m -th layer many-to-one recurrent neural network

Theorem A. 1 (Explicit output bound of recurrent neural network F) *Given an m -layer recurrent neural network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$, there exists two explicit functions $F_j^L : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ and $F_j^U : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ such that $\forall j \in [t]$ and $\forall \mathbf{X} \in \mathbb{R}^{n \times m}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, the inequity $F_j^L(\mathbf{X}) \leq F_j(\mathbf{X}) \leq F_j^U(\mathbf{X})$ holds true, where*

$$F_j^U(\mathbf{X}) = \Lambda_{j,:}^{(0)} \mathbf{a}^{(0)} + \sum_{k=1}^m \Lambda_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}^{(k)} + \sum_{k=1}^m \Lambda_{j,:}^{(k)} (\mathbf{b}^a + \Delta_{:,j}^{(k)}) + \mathbf{b}_j^F,$$

$$F_j^L(\mathbf{X}) = \Omega_{j,:}^{(0)} \mathbf{a}^{(0)} + \sum_{k=1}^m \Omega_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}^{(k)} + \sum_{k=1}^m \Omega_{j,:}^{(k)} (\mathbf{b}^a + \Theta_{:,j}^{(k)}) + \mathbf{b}_j^F,$$

$$\Lambda_{j,:}^{(k-1)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{(k-1)} & \text{if } k = m+1; \\ (\Lambda_{j,:}^{(k)} \mathbf{W}^{aa}) \odot \lambda_{j,:}^{(k-1)} & \text{if } k \in [m]; \end{cases} \quad \Omega_{j,:}^{(k-1)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{(k-1)} & \text{if } k = m+1; \\ (\Omega_{j,:}^{(k)} \mathbf{W}^{aa}) \odot \omega_{j,:}^{(k-1)} & \text{if } k \in [m]; \end{cases}$$

and $\forall r \in s$,

$$\lambda_{j,i}^{(m)} = \begin{cases} \alpha_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \alpha_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} \quad \Delta_{i,j}^{(m)} = \begin{cases} \beta_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \beta_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

$$\omega_{j,r}^{(m)} = \begin{cases} \alpha_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \alpha_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} \quad \Theta_{r,j}^{(m)} = \begin{cases} \beta_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \beta_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

and $\forall r \in s, k \in [m-1]$,

$$\lambda_{j,r}^{(k)} = \begin{cases} \alpha_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} \geq 0 \quad (\Leftrightarrow \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \alpha_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} < 0 \quad (\Leftrightarrow \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases}$$

$$\omega_{j,r}^{(k)} = \begin{cases} \alpha_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} \geq 0 \quad (\Leftrightarrow \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \alpha_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} < 0 \quad (\Leftrightarrow \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases}$$

$$\begin{aligned}\Delta_{r,j}^{(k)} &= \begin{cases} \beta_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} \geq 0 \quad (\Leftrightarrow \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \beta_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} < 0 \quad (\Leftrightarrow \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases} \\ \Theta_{r,j}^{(k)} &= \begin{cases} \beta_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} \geq 0 \quad (\Leftrightarrow \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \beta_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} < 0 \quad (\Leftrightarrow \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases}\end{aligned}$$

and $\forall r \in s, \lambda_{j,r}^{(0)} = \omega_{j,r}^{(0)} = 1$, where \odot is the Hadamard product.

Proof 1 Given an m -layer recurrent neural network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$ with pre-activation bounds $\mathbf{l}^{(k)}$ and $\mathbf{u}^{(k)}$ for $\mathbf{X} \in \mathbb{R}^{n \times m}$, where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, let the pre-activation inputs for the j -th neuron at the output layer be $F_j(\mathbf{X}) = \mathbf{W}_{j,:}^{Fa} \mathbf{a}^{(m)} + \mathbf{b}_j^F$. The j -th output of the recurrent neural network is the following:

$$\begin{aligned}F_j(\mathbf{X}) &= \mathbf{W}_{j,:}^{Fa} \mathbf{a}^{(m)} + \mathbf{b}_j^F, \tag{1} \\ &= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} \mathbf{a}_i^{(m)} + \mathbf{b}_j^F, \\ &= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} \sigma(\mathbf{y}_i^{(m)}) + \mathbf{b}_j^F, \\ &= \sum_{\mathbf{W}_{j,i}^{Fa} \geq 0} \mathbf{W}_{j,i}^{Fa} \sigma(\mathbf{y}_i^{(m)}) + \sum_{\mathbf{W}_{j,i}^{Fa} < 0} \mathbf{W}_{j,i}^{Fa} \sigma(\mathbf{y}_i^{(m)}) + \mathbf{b}_j^F. \tag{2}\end{aligned}$$

Assume the activation function $\sigma(\mathbf{y})$ is bounded by two linear functions $h_{L,i}^{(m)}(\mathbf{y})$ and $h_{U,i}^{(m)}(\mathbf{y})$, we have

$$h_{L,i}^{(m)}(\mathbf{y}_i^{(m)}) \leq \sigma(\mathbf{y}_i^{(m)}) \leq h_{U,i}^{(m)}(\mathbf{y}_i^{(m)}).$$

Thus, if the associated weight $\mathbf{W}_{j,i}^{Fa}$ to the i -th neuron is non-negative, we have

$$\mathbf{W}_{j,i}^{Fa} \cdot h_{L,i}^{(m)}(\mathbf{y}_i^{(m)}) \leq \mathbf{W}_{j,i}^{Fa} \sigma(\mathbf{y}_i^{(m)}) \leq \mathbf{W}_{j,i}^{Fa} \cdot h_{U,i}^{(m)}(\mathbf{y}_i^{(m)}); \tag{3}$$

otherwise, we have

$$\mathbf{W}_{j,i}^{Fa} \cdot h_{U,i}^{(m)}(\mathbf{y}_i^{(m)}) \leq \mathbf{W}_{j,i}^{Fa} \sigma(\mathbf{y}_i^{(m)}) \leq \mathbf{W}_{j,i}^{Fa} \cdot h_{L,i}^{(m)}(\mathbf{y}_i^{(m)}). \tag{4}$$

Upper bound. By using the right-hand side of Inequity (3) and Inequity (4) in Equation (2), we obtain an upper bound

$$\begin{aligned}F_j^{U,m}(\mathbf{X}) &= \sum_{\mathbf{W}_{j,i}^{Fa} \geq 0} \mathbf{W}_{j,i}^{Fa} \alpha_{U,i}^{(m)}(\mathbf{y}_i^{(m)}) + \beta_{U,i}^{(m)} \\ &\quad + \sum_{\mathbf{W}_{j,i}^{Fa} < 0} \mathbf{W}_{j,i}^{Fa} \alpha_{L,i}^{(m)}(\mathbf{y}_i^{(m)}) + \beta_{L,i}^{(m)} + \mathbf{b}_j^F.\end{aligned}$$

We use variables $\lambda_{j,i}^{(m)}$ and $\Delta_{i,j}^{(m)}$ to denote the slopes in front of $\mathbf{y}_i^{(m)}$ and intercepts in the parentheses:

$$\lambda_{j,i}^{(m)} = \begin{cases} \alpha_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \alpha_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

$$\Delta_{i,j}^{(m)} = \begin{cases} \beta_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \beta_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

and obtain

$$F_j^{U,m}(\mathbf{X}) = \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} \lambda_{j,i}^{(m)}(\mathbf{y}_i^{(m)}) + \Delta_{i,j}^{(m)} + \mathbf{b}_j^F.$$

We further let $\Lambda_{j,i}^{(m)} := \mathbf{W}_{j,i}^{Fa} \lambda_{j,i}^{(m)}$ and we have

$$\begin{aligned}
F_j^{U,m}(\mathbf{X}) &= \sum_{i=1}^s \Lambda_{j,i}^{(m)} (\mathbf{y}_i^{(m)} + \Delta_{i,j}^{(m)}) + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \Lambda_{j,i}^{(m)} (\mathbf{W}_{i,:}^{aa} \mathbf{a}^{(m-1)} + \mathbf{W}_{i,:}^{ax} \mathbf{x}^{(m)} + \mathbf{b}_i^a + \Delta_{i,j}^{(m)}) + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \Lambda_{j,i}^{(m)} \left(\sum_{r=1}^s \mathbf{W}_{i,r}^{aa} \mathbf{a}_r^{(m-1)} + \sum_{q_m=1}^n \mathbf{W}_{i,q_m}^{ax} \mathbf{x}_{q_m}^{(m)} + \mathbf{b}_i^a + \Delta_{i,j}^{(m)} \right) + \mathbf{b}_j^F, \\
&= \sum_{r=1}^s \left(\sum_{i=1}^s \Lambda_{j,i}^{(m)} \mathbf{W}_{i,r}^{aa} \right) \mathbf{a}_r^{(m-1)} + \sum_{q_m=1}^n \left(\sum_{i=1}^s \Lambda_{j,i}^{(m)} \mathbf{W}_{i,q_m}^{ax} \right) \mathbf{x}_{q_m}^{(m)} \\
&\quad + \sum_{i=1}^s \Lambda_{j,i}^{(m)} (\mathbf{b}_i^a + \Delta_{i,j}^{(m)}) + \mathbf{b}_j^F. \tag{5}
\end{aligned}$$

We combine coefficients in Equation (5) into equivalent weights $\tilde{\mathbf{W}}_{j,r}^{aa(m)}$, $\tilde{\mathbf{W}}_{j,q_m}^{ax(m)}$ and bias $\tilde{\mathbf{b}}_j^{(m)}$ as defined by

$$\begin{aligned}
\tilde{\mathbf{W}}_{j,r}^{aa(m)} &= \sum_{i=1}^s \Lambda_{j,i}^{(m)} \mathbf{W}_{i,r}^{aa} = \Lambda_{j,:}^{(m)} \mathbf{W}_{:,r}^{aa}, \\
\tilde{\mathbf{W}}_{j,q_m}^{ax(m)} &= \sum_{i=1}^s \Lambda_{j,i}^{(m)} \mathbf{W}_{i,q_m}^{ax} = \Lambda_{j,:}^{(m)} \mathbf{W}_{:,q_m}^{ax}, \\
\tilde{\mathbf{b}}_j^{(m)} &= \sum_{i=1}^s \Lambda_{j,i}^{(m)} (\mathbf{b}_i^a + \Delta_{i,j}^{(m)}) + \mathbf{b}_j^F = \Lambda_{j,:}^{(m)} (\mathbf{b}^a + \Delta_{:,j}^{(m)}) + \mathbf{b}_j^F,
\end{aligned}$$

and obtain

$$\begin{aligned}
F_j^{U,m}(\mathbf{X}) &= \sum_{r=1}^s \tilde{\mathbf{W}}_{j,r}^{aa(m)} \mathbf{a}_r^{(m-1)} + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m)}, \tag{6} \\
&= \sum_{r=1}^s \tilde{\mathbf{W}}_{j,r}^{aa(m)} \sigma(\mathbf{y}_r^{(m-1)}) + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m)}, \\
&= \sum_{\tilde{\mathbf{W}}_{j,r}^{aa(m)} \geq 0} \tilde{\mathbf{W}}_{j,r}^{aa(m)} \sigma(\mathbf{y}_r^{(m-1)}) + \sum_{\tilde{\mathbf{W}}_{j,r}^{aa(m)} < 0} \tilde{\mathbf{W}}_{j,r}^{aa(m)} \sigma(\mathbf{y}_r^{(m-1)}) + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m)}. \tag{7}
\end{aligned}$$

Assume the activation unction $\sigma(\mathbf{y})$ is bounded by two linear functions $h_{L,r}^{(m-1)}(\mathbf{y})$ and $h_{U,r}^{(m-1)}(\mathbf{y})$, we have

$$h_{L,r}^{(m-1)}(\mathbf{y}_r^{(m-1)}) \leq \sigma(\mathbf{y}_r^{(m-1)}) \leq h_{U,r}^{(m-1)}(\mathbf{y}_r^{(m-1)}).$$

Thus, if the associated weight $\tilde{\mathbf{W}}_{j,r}^{aa(m)}$ to the r -th neuron is non-negative, we have

$$\tilde{\mathbf{W}}_{j,r}^{aa(m)} \cdot h_{L,r}^{(m-1)}(\mathbf{y}_r^{(m-1)}) \leq \tilde{\mathbf{W}}_{j,r}^{aa(m)} \sigma(\mathbf{y}_r^{(m-1)}) \leq \tilde{\mathbf{W}}_{j,r}^{aa(m)} \cdot h_{U,r}^{(m-1)}(\mathbf{y}_r^{(m-1)}); \tag{8}$$

otherwise, we have

$$\tilde{\mathbf{W}}_{j,r}^{aa(m)} \cdot h_{U,r}^{(m-1)}(\mathbf{y}_r^{(m-1)}) \leq \tilde{\mathbf{W}}_{j,r}^{aa(m)} \sigma(\mathbf{y}_r^{(m-1)}) \leq \tilde{\mathbf{W}}_{j,r}^{aa(m)} \cdot h_{L,r}^{(m-1)}(\mathbf{y}_r^{(m-1)}). \tag{9}$$

By using the right-hand side of Inequity (8) and Inequity (9) in Equation (7), an upper bound can be obtained similarly as

$$\begin{aligned}
F_j^{U,m-1}(\mathbf{X}) &= \sum_{\tilde{\mathbf{W}}_{j,r}^{aa(m)} \geq 0} \tilde{\mathbf{W}}_{j,r}^{aa(m)} \alpha_{U,r}^{(m-1)}(\mathbf{y}_r^{(m-1)} + \beta_{U,r}^{(m-1)}) \\
&\quad + \sum_{\tilde{\mathbf{W}}_{j,r}^{aa(m)} < 0} \tilde{\mathbf{W}}_{j,r}^{aa(m)} \alpha_{L,r}^{(m-1)}(\mathbf{y}_r^{(m-1)} + \beta_{L,r}^{(m-1)}) + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m)}.
\end{aligned}$$

We use variables $\lambda_{j,r}^{(m-1)}$ and $\Delta_{r,j}^{(m-1)}$ to denote the slopes in front of $\mathbf{y}_r^{(m-1)}$ and intercepts in the parentheses:

$$\lambda_{j,r}^{(m-1)} = \begin{cases} \alpha_{U,r}^{(m-1)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(m)} \geq 0 \quad (\Leftrightarrow \Lambda_{j,:}^{(m)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \alpha_{L,r}^{(m-1)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(m)} < 0 \quad (\Leftrightarrow \Lambda_{j,:}^{(m)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases}$$

$$\Delta_{r,j}^{(m-1)} = \begin{cases} \beta_{U,r}^{(m-1)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(m)} \geq 0 \quad (\Leftrightarrow \Lambda_{j,:}^{(m)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \beta_{L,r}^{(m-1)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(m)} < 0 \quad (\Leftrightarrow \Lambda_{j,:}^{(m)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases}$$

and obtain

$$F_j^{U,m-1}(\mathbf{X}) = \sum_{r=1}^s \tilde{\mathbf{W}}_{j,r}^{aa(m)} \lambda_{j,r}^{(m-1)} (\mathbf{y}_r^{(m-1)} + \Delta_{r,j}^{(m-1)}) + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m)}.$$

We further let $\Lambda_{j,r}^{(m-1)} := \tilde{\mathbf{W}}_{j,r}^{aa(m)} \lambda_{j,r}^{(m-1)}$ and we have

$$\begin{aligned} F_j^{U,m-1}(\mathbf{X}) &= \sum_{r=1}^s \Lambda_{j,r}^{(m-1)} (\mathbf{y}_r^{(m-1)} + \Delta_{r,j}^{(m-1)}) + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m)}, \\ &= \sum_{r=1}^s \Lambda_{j,r}^{(m-1)} (\mathbf{W}_{r,:}^{aa} \mathbf{a}^{(m-2)} + \mathbf{W}_{r,:}^{ax} \mathbf{x}^{(m-1)} + \mathbf{b}_r^a + \Delta_{r,j}^{(m-1)}) + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m)}, \\ &= \sum_{r=1}^s \Lambda_{j,r}^{(m-1)} \left(\sum_{i=1}^s \mathbf{W}_{r,i}^{aa} \mathbf{a}_i^{(m-2)} + \sum_{q_{m-1}=1}^n \mathbf{W}_{r,q_{m-1}}^{ax} \mathbf{x}_{q_{m-1}}^{(m-1)} + \mathbf{b}_r^a + \Delta_{r,j}^{(m-1)} \right) \\ &\quad + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m)}, \\ &= \sum_{i=1}^s \left(\sum_{r=1}^s \Lambda_{j,r}^{(m-1)} \mathbf{W}_{r,i}^{aa} \right) \mathbf{a}_i^{(m-2)} + \sum_{q_{m-1}=1}^n \left(\sum_{r=1}^s \Lambda_{j,r}^{(m-1)} \mathbf{W}_{r,q_{m-1}}^{ax} \right) \mathbf{x}_{q_{m-1}}^{(m-1)} \\ &\quad + \sum_{r=1}^s \Lambda_{j,r}^{(m-1)} (\mathbf{b}_r^a + \Delta_{r,j}^{(m-1)}) + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m)}. \end{aligned} \quad (10)$$

We combine coefficients in Equation (10) into equivalent weights $\tilde{\mathbf{W}}_{j,i}^{aa(m-1)}$, $\tilde{\mathbf{W}}_{j,q_{m-1}}^{ax(m-1)}$ and bias $\tilde{\mathbf{b}}_j^{(m-1)}$ as defined by

$$\begin{aligned} \tilde{\mathbf{W}}_{j,i}^{aa(m-1)} &= \sum_{r=1}^s \Lambda_{j,r}^{(m-1)} \mathbf{W}_{r,i}^{aa} = \Lambda_{j,:}^{(m-1)} \mathbf{W}_{:,i}^{aa}, \\ \tilde{\mathbf{W}}_{j,q_{m-1}}^{ax(m-1)} &= \sum_{r=1}^s \Lambda_{j,r}^{(m-1)} \mathbf{W}_{r,q_{m-1}}^{ax} = \Lambda_{j,:}^{(m-1)} \mathbf{W}_{:,q_{m-1}}^{ax}, \\ \tilde{\mathbf{b}}_j^{(m-1)} &= \sum_{r=1}^s \Lambda_{j,r}^{(m-1)} (\mathbf{b}_r^a + \Delta_{r,j}^{(m-1)}) + \tilde{\mathbf{b}}_j^{(m)} = \Lambda_{j,:}^{(m-1)} (\mathbf{b}^a + \Delta_{:,j}^{(m-1)}) + \tilde{\mathbf{b}}_j^{(m)}, \end{aligned}$$

and obtain

$$F_j^{U,m-1}(\mathbf{X}) = \sum_{i=1}^s \tilde{\mathbf{W}}_{j,i}^{aa(m-1)} \mathbf{a}_i^{(m-2)} + \sum_{q_{m-1}=1}^n \tilde{\mathbf{W}}_{j,q_{m-1}}^{ax(m-1)} \mathbf{x}_{q_{m-1}}^{(m-1)} + \sum_{q_m=1}^n \tilde{\mathbf{W}}_{j,q_m}^{ax(m)} \mathbf{x}_{q_m}^{(m)} + \tilde{\mathbf{b}}_j^{(m-1)}. \quad (11)$$

Notice that Equation (6) and Equation (11) are in similar forms. Thus, we can repeat the above procedure iteratively until we obtain the final upper bound $F_j^{U,1}(\mathbf{X})$, where $F_j(\mathbf{X}) \leq F_j^{U,F}(\mathbf{X}) \leq F_j^{U,m}(\mathbf{X}) \leq \dots \leq F_j^{U,1}(\mathbf{X})$. We let $F_j^U(\mathbf{X})$ denote the final upper bound $F_j^{U,1}(\mathbf{X})$, and we have

$$F_j^U(\mathbf{X}) = \Lambda_{j,:}^{(0)} \mathbf{a}^{(0)} + \sum_{k=1}^m \Lambda_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}^{(k)} + \sum_{k=1}^m \Lambda_{j,:}^{(k)} (\mathbf{b}^a + \Delta_{:,j}^{(k)}) + \mathbf{b}_j^F,$$

and (\odot is the Hadamard product)

$$\mathbf{\Lambda}_{j,:}^{(k-1)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{(k-1)} & \text{if } k = m + 1; \\ (\mathbf{\Lambda}_{j,:}^{(k)} \mathbf{W}^{aa}) \odot \lambda_{j,:}^{(k-1)} & \text{if } k \in [m]; \end{cases}$$

and $\forall r \in s$,

$$\lambda_{j,i}^{(m)} = \begin{cases} \alpha_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \alpha_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

$$\Delta_{i,j}^{(m)} = \begin{cases} \beta_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \beta_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

and $\forall r \in s, k \in [m - 1]$,

$$\lambda_{j,r}^{(k)} = \begin{cases} \alpha_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} \geq 0 \quad (\Leftrightarrow \mathbf{\Lambda}_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \alpha_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} < 0 \quad (\Leftrightarrow \mathbf{\Lambda}_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases}$$

$$\Delta_{r,j}^{(k)} = \begin{cases} \beta_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} \geq 0 \quad (\Leftrightarrow \mathbf{\Lambda}_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \beta_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} < 0 \quad (\Leftrightarrow \mathbf{\Lambda}_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases}$$

and $\forall r \in s, \lambda_{j,r}^{(0)} = 1$.

Lower bound. The above derivations of upper bound can be applied similarly to derive lower bounds of $F_j(\mathbf{X})$, and the only difference is now we need to use the left-hand side of Inequity (3), (4), (8), (9) (rather than right-hand side when deriving upper bound) to bound the terms in (2) and (7). Thus, following the same procedure in deriving the upper bounds, we can iteratively unwrap the activation functions and obtain a final lower bound $F_j^{L,1}(\mathbf{X})$, where $F_j(\mathbf{X}) \geq F_j^{L,F}(\mathbf{X}) \geq F_j^{L,m}(\mathbf{X}) \geq \dots \geq F_j^{L,1}(\mathbf{X})$. We let $F_j^L(\mathbf{X})$ denote the final lower bound $F_j^{L,1}(\mathbf{X})$, and we have

$$F_j^L(\mathbf{X}) = \mathbf{\Omega}_{j,:}^{(0)} \mathbf{a}^{(0)} + \sum_{k=1}^m \mathbf{\Omega}_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}^{(k)} + \sum_{k=1}^m \mathbf{\Omega}_{j,:}^{(k)} (\mathbf{b}^a + \mathbf{\Theta}_{:,j}^{(k)}) + \mathbf{b}_j^F,$$

and (\odot is the Hadamard product)

$$\mathbf{\Omega}_{j,:}^{(k-1)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{(k-1)} & \text{if } k = m + 1; \\ (\mathbf{\Omega}_{j,:}^{(k)} \mathbf{W}^{aa}) \odot \omega_{j,:}^{(k-1)} & \text{if } k \in [m]; \end{cases}$$

and $\forall r \in s$,

$$\omega_{j,r}^{(m)} = \begin{cases} \alpha_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \alpha_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

$$\Theta_{r,j}^{(m)} = \begin{cases} \beta_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \beta_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

and $\forall r \in s, k \in [m - 1]$,

$$\omega_{j,r}^{(k)} = \begin{cases} \alpha_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} \geq 0 \quad (\Leftrightarrow \mathbf{\Omega}_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \alpha_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} < 0 \quad (\Leftrightarrow \mathbf{\Omega}_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases}$$

$$\Theta_{r,j}^{(k)} = \begin{cases} \beta_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} \geq 0 \quad (\Leftrightarrow \mathbf{\Omega}_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0); \\ \beta_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{j,r}^{aa(k+1)} < 0 \quad (\Leftrightarrow \mathbf{\Omega}_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0); \end{cases}$$

and $\forall r \in s, \omega_{j,r}^{(0)} = 1$.

Corollary A. 1 (Closed-form global bounds) Given data $\mathbf{X} \in \mathbb{R}^{n \times m}$, l_p ball parameters $p \geq 1$ and $\epsilon \geq 0$. For an m -layer recurrent neural network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$, there exists two fixed values γ_j^L and γ_j^U such that $\forall \mathbf{X} \in \mathbb{R}^{n \times m}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, and $\forall j \in [t]$, $1/q = 1 - 1/p$, the inequity $\gamma_j^L \leq F_j(\mathbf{X}) \leq \gamma_j^U$ holds true, where

$$\begin{aligned}\gamma_j^U &= \Lambda_{j,:}^{(0)} \mathbf{a}^{(0)} + \sum_{k=1}^m \epsilon \|\Lambda_{j,:}^{(k)} \mathbf{W}^{ax}\|_q + \sum_{k=1}^m \Lambda_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}_0^{(k)} + \sum_{k=1}^m \Lambda_{j,:}^{(k)} (\mathbf{b}^a + \Delta_{:,j}^{(k)}) + \mathbf{b}_j^F, \\ \gamma_j^L &= \Omega_{j,:}^{(0)} \mathbf{a}^{(0)} - \sum_{k=1}^m \epsilon \|\Omega_{j,:}^{(k)} \mathbf{W}^{ax}\|_q + \sum_{k=1}^m \Omega_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}_0^{(k)} + \sum_{k=1}^m \Omega_{j,:}^{(k)} (\mathbf{b}^a + \Theta_{:,j}^{(k)}) + \mathbf{b}_j^F.\end{aligned}$$

Theorem A. 2 (Closed-form pre-activation bounds of recurrent neural network F) For the pre-activation input $\mathbf{y}_j^{(v)}$ of the v -th layer in a given m -layer ($v \leq m$) recurrent neural network, there exists two explicit functions $g_j^{L,v} : \mathbb{R}^{n \times v} \rightarrow \mathbb{R}^s$ and $g_j^{U,v} : \mathbb{R}^{n \times v} \rightarrow \mathbb{R}^s$ and two fixed values $\mathbf{l}_j^{(v)}$ and $\mathbf{u}_j^{(v)}$ such that $\forall j \in [s]$ and $\forall \mathbf{X}^{(1:v)} \in \mathbb{R}^{n \times v}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$ and $1/q = 1 - 1/p$, the inequity $\mathbf{l}_j^{(v)} \leq g_j^{L,v}(\mathbf{X}^{(1:v)}) \leq \mathbf{y}_j^{(v)} \leq g_j^{U,v}(\mathbf{X}^{(1:v)}) \leq \mathbf{u}_j^{(v)}$ holds true, where

$$\begin{aligned}\mathbf{u}_j^{(v)} &= \Lambda_{j,:}^{(0)} \mathbf{a}^{(0)} + \sum_{k=1}^v \epsilon \|\Lambda_{j,:}^{(k)} \mathbf{W}^{ax}\|_q + \sum_{k=1}^v \Lambda_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}_0^{(k)} + \sum_{k=1}^v \Lambda_{j,:}^{(k)} (\mathbf{b}^a + \Delta_{:,j}^{(k)}), \\ \mathbf{l}_j^{(v)} &= \Omega_{j,:}^{(0)} \mathbf{a}^{(0)} - \sum_{k=1}^v \epsilon \|\Omega_{j,:}^{(k)} \mathbf{W}^{ax}\|_q + \sum_{k=1}^v \Omega_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}_0^{(k)} + \sum_{k=1}^v \Omega_{j,:}^{(k)} (\mathbf{b}^a + \Theta_{:,j}^{(k)}),\end{aligned}\quad (12)$$

$$\Lambda_{j,:}^{(k-1)} = \begin{cases} \mathbf{e}_j^\top & \text{if } k = v + 1; \\ \mathbf{W}_{j,:}^{aa} \odot \lambda_{j,:}^{(k-1)} & \text{if } k = v; \\ (\Lambda_{j,:}^{(k)} \mathbf{W}^{aa}) \odot \lambda_{j,:}^{(k-1)} & \text{if } k \in [v - 1]; \end{cases} \quad \Omega_{j,:}^{(k-1)} = \begin{cases} \mathbf{e}_j^\top & \text{if } k = v + 1; \\ \mathbf{W}_{j,:}^{aa} \odot \omega_{j,:}^{(k-1)} & \text{if } k = v; \\ (\Omega_{j,:}^{(k)} \mathbf{W}^{aa}) \odot \omega_{j,:}^{(k-1)} & \text{if } k \in [v - 1]; \end{cases}\quad (13)$$

and $\forall r \in s, k \in [v - 1]$,

$$\begin{aligned}\lambda_{j,r}^{(k)} &= \begin{cases} \alpha_{U,r}^{(k)} & \text{if } \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0; \\ \alpha_{L,r}^{(k)} & \text{if } \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0; \end{cases} \\ \omega_{j,r}^{(k)} &= \begin{cases} \alpha_{L,r}^{(k)} & \text{if } \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0; \\ \alpha_{U,r}^{(k)} & \text{if } \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0; \end{cases} \\ \Delta_{r,j}^{(k)} &= \begin{cases} \beta_{U,r}^{(k)} & \text{if } \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0; \\ \beta_{L,r}^{(k)} & \text{if } \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0. \end{cases} \\ \Theta_{r,j}^{(k)} &= \begin{cases} \beta_{L,r}^{(k)} & \text{if } \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0; \\ \beta_{U,r}^{(k)} & \text{if } \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0; \end{cases}\end{aligned}\quad (14)$$

and $\forall r \in s, \lambda_{j,r}^{(0)} = \omega_{j,r}^{(0)} = 1, \Delta_{r,j}^{(v)} = \Theta_{r,j}^{(v)} = 0$, where \odot is the Hadamard product.

Proof 2 For a given m -layer ($v \leq m$) recurrent neural network, let the pre-activation input for the j -th neuron at the k -th layer be $\mathbf{y}_j^{(v)}$, where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon), \forall k \in [v]$. Then we have the following:

$$\begin{aligned}\mathbf{y}_j^{(v)} &= \mathbf{W}_{j,:}^{aa} \mathbf{a}^{(v-1)} + \mathbf{W}_{j,:}^{ax} \mathbf{x}^{(v)} + \mathbf{b}_j^a, \\ &= \sum_{r=1}^s \mathbf{W}_{j,r}^{aa} \mathbf{a}_r^{(v-1)} + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \mathbf{b}_j^a,\end{aligned}\quad (15)$$

$$\begin{aligned}&= \sum_{r=1}^s \mathbf{W}_{j,r}^{aa} \sigma(\mathbf{y}_r^{(v-1)}) + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \mathbf{b}_j^a, \\ &= \sum_{\mathbf{W}_{j,r}^{aa} \geq 0} \mathbf{W}_{j,r}^{aa} \sigma(\mathbf{y}_r^{(v-1)}) + \sum_{\mathbf{W}_{j,r}^{aa} < 0} \mathbf{W}_{j,r}^{aa} \sigma(\mathbf{y}_r^{(v-1)}) + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \mathbf{b}_j^a.\end{aligned}\quad (16)$$

Assume the activation unction $\sigma(\mathbf{y}_r^{(v-1)})$ is bounded by two linear functions $h_{L,r}^{(v-1)}(\mathbf{y})$ and $h_{U,r}^{(v-1)}(\mathbf{y})$, we have

$$h_{L,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}) \leq \sigma(\mathbf{y}_r^{(v-1)}) \leq h_{U,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}).$$

Thus, if the associated weight $\mathbf{W}_{j,r}^{aa}$ to the r -th neuron is non-negative, we have

$$\mathbf{W}_{j,r}^{aa} \cdot h_{L,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}) \leq \mathbf{W}_{j,r}^{aa} \sigma(\mathbf{y}_r^{(v-1)}) \leq \mathbf{W}_{j,r}^{aa} \cdot h_{U,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}); \quad (17)$$

otherwise, we have

$$\mathbf{W}_{j,r}^{aa} \cdot h_{U,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}) \leq \mathbf{W}_{j,r}^{aa} \sigma(\mathbf{y}_r^{(v-1)}) \leq \mathbf{W}_{j,r}^{aa} \cdot h_{L,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}). \quad (18)$$

By using the right-hand side of Inequity (17) and Inequity (18) in Equation (16), an upper bound can be obtained similarly as

$$\begin{aligned} g_j^{U,v,v}(\mathbf{X}^{(1:v)}) &= \sum_{\mathbf{W}_{j,r}^{aa} \geq 0} \mathbf{W}_{j,r}^{aa} \alpha_{U,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}) + \beta_{U,r}^{(v-1)} \\ &\quad + \sum_{\mathbf{W}_{j,r}^{aa} < 0} \mathbf{W}_{j,r}^{aa} \alpha_{L,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}) + \beta_{L,r}^{(v-1)} + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \mathbf{b}_j^a. \end{aligned}$$

We use variables $\lambda_{j,r}^{(v-1)}$ and $\Delta_{r,j}^{(v-1)}$ to denote the slops in front of $\mathbf{y}_r^{(v-1)}$ and intercepts in the parentheses:

$$\lambda_{j,r}^{(v-1)} = \begin{cases} \alpha_{U,r}^{(v-1)} & \text{if } \mathbf{W}_{j,r}^{aa} \geq 0; \\ \alpha_{L,r}^{(v-1)} & \text{if } \mathbf{W}_{j,r}^{aa} < 0; \end{cases}$$

$$\Delta_{r,j}^{(v-1)} = \begin{cases} \beta_{U,r}^{(v-1)} & \text{if } \mathbf{W}_{j,r}^{aa} \geq 0; \\ \beta_{L,r}^{(v-1)} & \text{if } \mathbf{W}_{j,r}^{aa} < 0; \end{cases}$$

and obtain

$$g_j^{U,v,v}(\mathbf{X}^{(1:v)}) = \sum_{r=1}^s \mathbf{W}_{j,r}^{aa} \lambda_{j,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}) + \Delta_{r,j}^{(v-1)} + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \mathbf{b}_j^a.$$

We further let $\Lambda_{j,r}^{(v-1)} := \mathbf{W}_{j,r}^{aa} \lambda_{j,r}^{(v-1)}$ and we have

$$\begin{aligned} g_j^{U,v,v}(\mathbf{X}^{(1:v)}) &= \sum_{r=1}^s \Lambda_{j,r}^{(v-1)}(\mathbf{y}_r^{(v-1)}) + \Delta_{r,j}^{(v-1)} + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \mathbf{b}_j^a, \\ &= \sum_{r=1}^s \Lambda_{j,r}^{(v-1)}(\mathbf{W}_{r,:}^{aa} \mathbf{a}^{(v-2)} + \mathbf{W}_{r,:}^{ax} \mathbf{x}^{(v-1)} + \mathbf{b}_r^a + \Delta_{r,j}^{(v-1)}) + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \mathbf{b}_j^a, \\ &= \sum_{r=1}^s \Lambda_{j,r}^{(v-1)} \left(\sum_{i=1}^s \mathbf{W}_{r,i}^{aa} \mathbf{a}_i^{(v-2)} + \sum_{q_{v-1}=1}^n \mathbf{W}_{r,q_{v-1}}^{ax} \mathbf{x}_{q_{v-1}}^{(v-1)} + \mathbf{b}_r^a + \Delta_{r,j}^{(v-1)} \right) \\ &\quad + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \mathbf{b}_j^a, \\ &= \sum_{i=1}^s \left(\sum_{r=1}^s \Lambda_{j,r}^{(v-1)} \mathbf{W}_{r,i}^{aa} \right) \mathbf{a}_i^{(v-2)} + \sum_{q_{v-1}=1}^n \left(\sum_{r=1}^s \Lambda_{j,r}^{(v-1)} \mathbf{W}_{r,q_{v-1}}^{ax} \right) \mathbf{x}_{q_{v-1}}^{(v-1)} \\ &\quad + \sum_{r=1}^s \Lambda_{j,r}^{(v-1)} (\mathbf{b}_r^a + \Delta_{r,j}^{(v-1)}) + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \mathbf{b}_j^a. \end{aligned} \quad (19)$$

We combine coefficients in Equation (19) into equivalent weights $\tilde{\mathbf{W}}_{j,i}^{aa(v-1)}$, $\tilde{\mathbf{W}}_{j,q_2}^{ax(v-1)}$ and bias $\tilde{\mathbf{b}}_j^{(v-1)}$ as defined by

$$\begin{aligned}\tilde{\mathbf{W}}_{j,i}^{aa(v-1)} &= \sum_{r=1}^s \Lambda_{j,r}^{(v-1)} \mathbf{W}_{r,i}^{aa} = \Lambda_{j,:}^{(v-1)} \mathbf{W}_{:,i}^{aa}, \\ \tilde{\mathbf{W}}_{j,q_{v-1}}^{ax(v-1)} &= \sum_{r=1}^s \Lambda_{j,r}^{(v-1)} \mathbf{W}_{r,q_{v-1}}^{ax} = \Lambda_{j,:}^{(v-1)} \mathbf{W}_{:,q_{v-1}}^{ax}, \\ \tilde{\mathbf{b}}_j^{(v-1)} &= \sum_{r=1}^s \Lambda_{j,r}^{(v-1)} (\mathbf{b}_r^a + \Delta_{r,j}^{(v-1)}) + \mathbf{b}_j^a = \Lambda_{j,:}^{(v-1)} (\mathbf{b}^a + \Delta_{:,j}^{(v-1)}) + \mathbf{b}_j^a,\end{aligned}$$

and obtain

$$g_j^{U,v,v}(\mathbf{X}^{(1:v)}) = \sum_{i=1}^s \tilde{\mathbf{W}}_{j,i}^{aa(v-1)} \mathbf{a}_i^{(v-2)} + \sum_{q_{v-1}=1}^n \tilde{\mathbf{W}}_{j,q_{v-1}}^{ax(v-1)} \mathbf{x}_{q_{v-1}}^{(v-1)} + \sum_{q_v=1}^n \mathbf{W}_{j,q_v}^{ax} \mathbf{x}_{q_v}^{(v)} + \tilde{\mathbf{b}}_j^{(v-1)}. \quad (20)$$

Notice that Equation (15) and Equation (20) are in similar forms. Thus, we can repeat the above procedure iteratively until we obtain the final upper bound $g_j^{U,v,1}(\mathbf{X}^{(1:v)})$, where $\mathbf{y}_j^{(v)} \leq g_j^{U,v,v}(\mathbf{X}^{(1:v)}) \leq \dots \leq g_j^{U,v,1}(\mathbf{X}^{(1:v)})$. We let $\mathbf{u}_j^{(v)} = g_j^{U,v}(\mathbf{X}^{(1:v)})$ denote the final upper bound $g_j^{U,v,1}(\mathbf{X}^{(1:v)})$, and we have

$$\mathbf{u}_j^{(v)} = g_j^{U,v}(\mathbf{X}^{(1:v)}) = \Lambda_{j,:}^{(0)} \mathbf{a}^{(0)} + \sum_{k=1}^v \Lambda_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}^{(k)} + \sum_{k=1}^v \Lambda_{j,:}^{(k)} (\mathbf{b}^a + \Delta_{:,j}^{(k)})$$

and (\odot is the Hadamard product)

$$\Lambda_{j,:}^{(k-1)} = \begin{cases} \mathbf{e}_j^\top & \text{if } k = v + 1; \\ \mathbf{W}_{j,:}^{aa} \odot \lambda_{j,:}^{(k-1)} & \text{if } k = v; \\ (\Lambda_{j,:}^{(k)} \mathbf{W}^{aa}) \odot \lambda_{j,:}^{(k-1)} & \text{if } k \in [v-1]; \end{cases}$$

and $\forall r \in s, k \in [v-1]$,

$$\lambda_{j,r}^{(k)} = \begin{cases} \alpha_{U,r}^{(k)} & \text{if } \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0; \\ \alpha_{L,r}^{(k)} & \text{if } \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0; \end{cases}$$

$$\Delta_{r,j}^{(k)} = \begin{cases} \beta_{U,r}^{(k)} & \text{if } \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0; \\ \beta_{L,r}^{(k)} & \text{if } \Lambda_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0; \end{cases}$$

and $\forall r \in s, \lambda_{j,r}^{(0)} = 1, \Delta_{r,j}^{(v)} = 0$.

Lower bound. The above derivations of upper bound can be applied similarly to derive lower bounds of $f_j(\mathbf{X})$, and the only difference is now we need to use the left-hand side of Inequity (17), (18) (rather than right-hand side when deriving upper bound) to bound the terms in (16). Thus, following the same procedure in deriving the upper bounds, we can iteratively unwrap the activation functions and obtain a final lower bound $g_j^{L,v,1}(\mathbf{X}^{(1:v)})$, where $\mathbf{y}_j^{(v)} \geq g_j^{L,v,v}(\mathbf{X}^{(1:v)}) \geq \dots \geq g_j^{L,v,1}(\mathbf{X}^{(1:v)})$. We let $\mathbf{l}_j^{(v)} = g_j^{L,v}(\mathbf{X}^{(1:v)})$ denote the final upper bound $g_j^{L,v,1}(\mathbf{X}^{(1:v)})$, and we have

$$\mathbf{l}_j^{(v)} = g_j^{L,v}(\mathbf{X}^{(1:v)}) = \Omega_{j,:}^{(0)} \mathbf{a}^{(0)} + \sum_{k=1}^v \Omega_{j,:}^{(k)} \mathbf{W}^{ax} \mathbf{x}^{(k)} + \sum_{k=1}^v \Omega_{j,:}^{(k)} (\mathbf{b}^a + \Theta_{:,j}^{(k)})$$

and (\odot is the Hadamard product)

$$\Omega_{j,:}^{(k-1)} = \begin{cases} \mathbf{e}_j^\top & \text{if } k = v + 1; \\ \mathbf{W}_{j,:}^{aa} \odot \omega_{j,:}^{(k-1)} & \text{if } k = v; \\ (\Omega_{j,:}^{(k)} \mathbf{W}^{aa}) \odot \omega_{j,:}^{(k-1)} & \text{if } k \in [v-1]; \end{cases}$$

and $\forall r \in s, k \in [v - 1]$,

$$\omega_{j,r}^{(k)} = \begin{cases} \alpha_{L,r}^{(k)} & \text{if } \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0; \\ \alpha_{U,r}^{(k)} & \text{if } \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0; \end{cases}$$

$$\Theta_{r,j}^{(k)} = \begin{cases} \beta_{L,r}^{(k)} & \text{if } \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} \geq 0; \\ \beta_{U,r}^{(k)} & \text{if } \Omega_{j,:}^{(k+1)} \mathbf{W}_{:,r}^{aa} < 0. \end{cases}$$

and $\forall r \in s, \omega_{j,r}^{(0)} = 1, \Theta_{r,j}^{(v)} = 0$.

A.2 Algorithms for Recurrent Neural Networks

Algorithm 1 Compute POPQORN robustness bound of recurrent neural network (true label: c , target label: i)

Input: weight matrices \mathbf{W}^{aa} , \mathbf{W}^{Fa} , \mathbf{W}^{ax} and bias vectors \mathbf{b}^a , \mathbf{b}^f of an m -layer recurrent neural network f , input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}]$, p -norm, ϵ_0 .

Output: a robustness bound $\tilde{\epsilon}$.

- 1: Replace the last layer weights $\bar{\mathbf{w}} \leftarrow \mathbf{W}_{c,:}^{Fa} - \mathbf{W}_{i,:}^{Fa}$
- 2: Initialize $\epsilon \leftarrow \epsilon_0$
- 3: **while** ϵ has not achieved a desired accuracy and iteration limit has not reached **do**
- 4: **for** $v=1, \dots, m$ **do**
- 5: **for** $j=1, \dots, s$ **do**
- 6: $\mathbf{l}_j^{(v)}, \mathbf{u}_j^{(v)} \leftarrow$ compute the pre-activation bound for the j -th neuron of the v -th layer according to Equation (12), (13) and (14).
- 7: $\alpha_{L,j}^{(v)}, \alpha_{U,j}^{(v)}, \beta_{L,j}^{(v)}, \beta_{U,j}^{(v)} \leftarrow$ compute the two bounding linear function $h_{L,j}^{(v)}, h_{U,j}^{(v)}$
- 8: **end for**
- 9: **end for**
- 10: $\gamma^U, \gamma^L \leftarrow$ Corollary A.1
- 11: **if** $\gamma^L \geq 0$ **then**
- 12: ϵ is a lower bound; increase ϵ using a binary search procedure
- 13: **else**
- 14: ϵ is not a lower bound; decrease ϵ using a binary search procedure
- 15: **end if**
- 16: $\tilde{\epsilon} \leftarrow \epsilon$
- 17: **end while**

A.3 Long Short-term Memory Network

Notations In the derivation of long short-term memory network (LSTM), we define the equations as follows:

$$\begin{aligned}
\text{Input gate: } \mathbf{i}^{(k)} &= \sigma(\mathbf{y}^{i(k)}) = \sigma(\mathbf{W}^{ix}\mathbf{x}^{(k)} + \mathbf{W}^{ia}\mathbf{a}^{(k-1)} + \mathbf{b}^i); \\
\text{Forget gate: } \mathbf{f}^{(k)} &= \sigma(\mathbf{y}^f(k)) = \sigma(\mathbf{W}^{fx}\mathbf{x}^{(k)} + \mathbf{W}^{fa}\mathbf{a}^{(k-1)} + \mathbf{b}^f); \\
\text{Cell gate: } \mathbf{g}^{(k)} &= \tanh(\mathbf{y}^g(k)) = \tanh(\mathbf{W}^{gx}\mathbf{x}^{(k)} + \mathbf{W}^{ga}\mathbf{a}^{(k-1)} + \mathbf{b}^g); \\
\text{Output gate: } \mathbf{o}^{(k)} &= \sigma(\mathbf{y}^o(k)) = \sigma(\mathbf{W}^{ox}\mathbf{x}^{(k)} + \mathbf{W}^{oa}\mathbf{a}^{(k-1)} + \mathbf{b}^o); \\
\text{Cell state: } \mathbf{c}^{(k)} &= \mathbf{f}^{(k)} \odot \mathbf{c}^{(k-1)} + \mathbf{i}^{(k)} \odot \mathbf{g}^{(k)}; \\
\text{Hidden state: } \mathbf{a}^{(k)} &= \mathbf{o}^{(k)} \odot \tanh(\mathbf{c}^{(k)}).
\end{aligned}$$

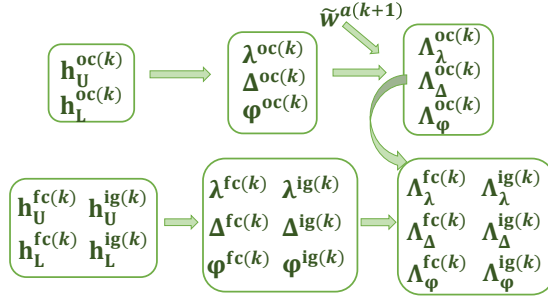


Figure 2: Steps in computing bounds for a long short-term memory network.

Definition 1 (Linear bounds on 2D activation function) For the r -th neuron in k -th layer with pre-activation bounds of $\mathbf{y}_r^{o(k)}, \mathbf{c}_r^{(k)}: \mathbf{l}_r^{o(k)}, \mathbf{u}_r^{o(k)}, \mathbf{l}_r^{c(k)}, \mathbf{u}_r^{c(k)}$ and the activation functions, define two linear functions $h_{L,r}^{oc(k)}, h_{U,r}^{oc(k)}: \mathbb{R}^2 \rightarrow \mathbb{R}$,

$$\begin{aligned}
h_{L,r}^{oc(k)}(\mathbf{y}_r^{o(k)}, \mathbf{c}_r^{(k)}) &= \alpha_{L,r}^{oc(k)} \mathbf{y}_r^{o(k)} + \beta_{L,r}^{oc(k)} \mathbf{c}_r^{(k)} + \gamma_{L,r}^{oc(k)}, \\
h_{U,r}^{oc(k)}(\mathbf{y}_r^{o(k)}, \mathbf{c}_r^{(k)}) &= \alpha_{U,r}^{oc(k)} \mathbf{y}_r^{o(k)} + \beta_{U,r}^{oc(k)} \mathbf{c}_r^{(k)} + \gamma_{U,r}^{oc(k)},
\end{aligned}$$

such that for $\mathbf{y}_r^{o(k)} \in [\mathbf{l}_r^{o(k)}, \mathbf{u}_r^{o(k)}], \mathbf{c}_r^{(k)} \in [\mathbf{l}_r^{c(k)}, \mathbf{u}_r^{c(k)}]$,

$$h_{L,r}^{oc(k)}(\mathbf{y}_r^{o(k)}, \mathbf{c}_r^{(k)}) \leq \sigma(\mathbf{y}_r^{o(k)}) \tanh(\mathbf{c}_r^{(k)}) \leq h_{U,r}^{oc(k)}(\mathbf{y}_r^{o(k)}, \mathbf{c}_r^{(k)}).$$

Similarly, we can define linear functions $h_{L,r}^{fc(k)}, h_{U,r}^{fc(k)}, h_{L,r}^{ig(k)}$, and $h_{U,r}^{ig(k)}$ such that for $\mathbf{y}_r^{f(k)} \in [\mathbf{l}_r^{f(k)}, \mathbf{u}_r^{f(k)}], \mathbf{c}_r^{(k-1)} \in [\mathbf{l}_r^{c(k-1)}, \mathbf{u}_r^{c(k-1)}]$,

$$h_{L,r}^{fc(k)}(\mathbf{y}_r^{f(k)}, \mathbf{c}_r^{(k-1)}) \leq \sigma(\mathbf{y}_r^{f(k)}) \mathbf{c}_r^{(k-1)} \leq h_{U,r}^{fc(k)}(\mathbf{y}_r^{f(k)}, \mathbf{c}_r^{(k-1)});$$

and for $\mathbf{y}_r^{i(k)} \in [\mathbf{l}_r^{i(k)}, \mathbf{u}_r^{i(k)}], \mathbf{y}_r^{g(k)} \in [\mathbf{l}_r^{g(k)}, \mathbf{u}_r^{g(k)}]$,

$$h_{L,r}^{ig(k)}(\mathbf{y}_r^{i(k)}, \mathbf{y}_r^{g(k)}) \leq \sigma(\mathbf{y}_r^{i(k)}) \tanh(\mathbf{y}_r^{g(k)}) \leq h_{U,r}^{ig(k)}(\mathbf{y}_r^{i(k)}, \mathbf{y}_r^{g(k)}).$$

Theorem A. 3 (Explicit output bound of long short-term memory network F) Given an m -layer long short-term memory network $F: \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$, we can derive two explicit functions $F_j^L: \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ and $F_j^U: \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ such that $\forall j \in [t]$ and $\forall \mathbf{X} \in \mathbb{R}^{n \times m}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, the inequity $F_j^L(\mathbf{X}) \leq F_j(\mathbf{X}) \leq F_j^U(\mathbf{X})$ holds true. We can obtain them through following steps: Starting from $k = m, \forall r \in s$ we

1. define unified slopes and intercepts of cross terms of the cell state and the output gate;

$$\begin{aligned}\lambda_{j,r}^{oc(k)} &= \begin{cases} \alpha_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \alpha_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \omega_{j,r}^{oc(k)} &= \begin{cases} \alpha_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \alpha_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\ \Delta_{j,r}^{oc(k)} &= \begin{cases} \beta_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \beta_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \Theta_{j,r}^{oc(k)} &= \begin{cases} \beta_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \beta_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\ \varphi_{j,r}^{oc(k)} &= \begin{cases} \gamma_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \gamma_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \psi_{j,r}^{oc(k)} &= \begin{cases} \gamma_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \gamma_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases}\end{aligned}$$

where $\tilde{\mathbf{W}}_{U,j,r}^{a(m+1)} = \tilde{\mathbf{W}}_{L,j,r}^{a(m+1)} = \mathbf{W}_{j,r}^{Fa}$;

2. collect coefficients in front of the cell state and the output gate *pre-activation*, and constants;

$$\begin{aligned}\Lambda_{\lambda,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \lambda_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\omega,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \omega_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases} \\ \Lambda_{\Delta,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Delta_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \Delta_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\Theta,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Theta_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \Theta_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases} \\ \Lambda_{\varphi,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \varphi_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \varphi_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\psi,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \psi_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \psi_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases}\end{aligned}$$

3. define unified slopes and intercepts of cross terms of the cell state and the forget gate, the input gate and the cell gate;

$$\begin{aligned}\lambda_{j,r}^{fc(k)} &= \begin{cases} \alpha_{U,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \alpha_{L,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \omega_{j,r}^{fc(k)} &= \begin{cases} \alpha_{L,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \alpha_{U,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \Delta_{j,r}^{fc(k)} &= \begin{cases} \beta_{U,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \beta_{L,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \Theta_{j,r}^{fc(k)} &= \begin{cases} \beta_{L,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \beta_{U,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \varphi_{j,r}^{fc(k)} &= \begin{cases} \gamma_{U,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \gamma_{L,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \psi_{j,r}^{fc(k)} &= \begin{cases} \gamma_{L,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \gamma_{U,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \lambda_{j,r}^{ig(k)} &= \begin{cases} \alpha_{U,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \alpha_{L,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \omega_{j,r}^{ig(k)} &= \begin{cases} \alpha_{L,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \alpha_{U,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \Delta_{j,r}^{ig(k)} &= \begin{cases} \beta_{U,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \beta_{L,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \Theta_{j,r}^{ig(k)} &= \begin{cases} \beta_{L,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \beta_{U,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \varphi_{j,r}^{ig(k)} &= \begin{cases} \gamma_{U,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \gamma_{L,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \psi_{j,r}^{ig(k)} &= \begin{cases} \gamma_{L,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \gamma_{U,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases}\end{aligned}$$

4. collect coefficients in front of the cell state, the forget gate *pre-activation*, the input gate *pre-activation*, the cell gate *pre-activation*, and constants of the two couplings;

$$\begin{aligned}\Lambda_{\lambda,j,:}^{fc(k)} &= \Lambda_{\Delta,j,:}^{oc(k)} \odot \lambda_{j,:}^{fc(k)}; & \Omega_{\omega,j,:}^{fc(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \omega_{j,:}^{fc(k)}; \\ \Lambda_{\Delta,j,:}^{fc(k)} &= \Lambda_{\Delta,j,:}^{oc(k)} \odot \Delta_{j,:}^{fc(k)}; & \Omega_{\Theta,j,:}^{fc(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \Theta_{j,:}^{fc(k)}; \\ \Lambda_{\varphi,j,:}^{fc(k)} &= \Lambda_{\Delta,j,:}^{oc(k)} \odot \varphi_{j,:}^{fc(k)}; & \Omega_{\psi,j,:}^{fc(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \psi_{j,:}^{fc(k)}; \\ \Lambda_{\lambda,j,:}^{ig(k)} &= \Lambda_{\Delta,j,:}^{oc(k)} \odot \lambda_{j,:}^{ig(k)}; & \Omega_{\omega,j,:}^{ig(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \omega_{j,:}^{ig(k)};\end{aligned}$$

$$\begin{aligned}
\Lambda_{\Delta,j,:}^{ig(k)} &= \Lambda_{\Delta,j,:}^{oc(k)} \odot \Delta_{j,:}^{ig(k)}; & \Omega_{\Theta,j,:}^{ig(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \Theta_{j,:}^{ig(k)}; \\
\Lambda_{\varphi,j,:}^{ig(k)} &= \Lambda_{\Delta,j,:}^{oc(k)} \odot \varphi_{j,:}^{ig(k)}; & \Omega_{\psi,j,:}^{ig(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \psi_{j,:}^{ig(k)};
\end{aligned}$$

5. define equivalent weight matrices and biases;

$$\begin{aligned}
\tilde{\mathbf{W}}_{U,j,q_k}^{x(k)} &= \Lambda_{\lambda,j,:}^{oc(k)} \mathbf{W}_{:,q_k}^{ox} + \Lambda_{\lambda,j,:}^{fc(k)} \mathbf{W}_{:,q_k}^{fx} + \Lambda_{\lambda,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{ix} + \Lambda_{\Delta,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{gx}; \\
\tilde{\mathbf{W}}_{U,j,r}^{a(k)} &= \Lambda_{\lambda,j,:}^{oc(k)} \mathbf{W}_{:,r}^{oa} + \Lambda_{\lambda,j,:}^{fc(k)} \mathbf{W}_{:,r}^{fa} + \Lambda_{\lambda,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ia} + \Lambda_{\Delta,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ga}; \\
\tilde{\mathbf{b}}_{U,j}^{(k)} &= \Lambda_{\lambda,j,:}^{oc(k)} \mathbf{b}^o + \Lambda_{\lambda,j,:}^{fc(k)} \mathbf{b}^f + \Lambda_{\lambda,j,:}^{ig(k)} \mathbf{b}^i + \Lambda_{\Delta,j,:}^{ig(k)} \mathbf{b}^g + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{oc(k)} + \Lambda_{\varphi,j,i}^{fc(k)} + \Lambda_{\varphi,j,i}^{ig(k)}); \\
\tilde{\mathbf{W}}_{L,j,q_k}^{x(k)} &= \Omega_{\omega,j,:}^{oc(k)} \mathbf{W}_{:,q_k}^{ox} + \Omega_{\omega,j,:}^{fc(k)} \mathbf{W}_{:,q_k}^{fx} + \Omega_{\omega,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{ix} + \Omega_{\Theta,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{gx}; \\
\tilde{\mathbf{W}}_{L,j,r}^{a(k)} &= \Omega_{\omega,j,:}^{oc(k)} \mathbf{W}_{:,r}^{oa} + \Omega_{\omega,j,:}^{fc(k)} \mathbf{W}_{:,r}^{fa} + \Omega_{\omega,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ia} + \Omega_{\Theta,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ga}; \\
\tilde{\mathbf{b}}_{L,j}^{(k)} &= \Omega_{\omega,j,:}^{oc(k)} \mathbf{b}^o + \Omega_{\omega,j,:}^{fc(k)} \mathbf{b}^f + \Omega_{\omega,j,:}^{ig(k)} \mathbf{b}^i + \Omega_{\Theta,j,:}^{ig(k)} \mathbf{b}^g + \sum_{i=1}^s (\Omega_{\psi,j,i}^{oc(k)} + \Omega_{\psi,j,i}^{fc(k)} + \Omega_{\psi,j,i}^{ig(k)});
\end{aligned}$$

6. After looping steps 1 to 5 from $k = m$ to $k = 1$, the bounds are given by

$$\begin{aligned}
F_j^U(\mathbf{X}) &= \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \Lambda_{\Delta,j,:}^{fc(1)} \mathbf{c}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{U,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{U,j}^{(k)} + \mathbf{b}_j^F, \\
F_j^L(\mathbf{X}) &= \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} + \Omega_{\Theta,j,:}^{fc(1)} \mathbf{c}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{L,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{L,j}^{(k)} + \mathbf{b}_j^F.
\end{aligned}$$

Though we exemplify by an m -layer many-to-one long short-term memory network, the above steps can also be used to derive bounds for the k -th output $\mathbf{W}^{Fa} \mathbf{a}^{(k)} + \mathbf{b}^F$ in a many-to-many long short-term memory network.

Proof 3 Given an m -layer long short-term memory network $F: \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$ with pre-activation bounds $\mathbf{l}^{o(k)}, \mathbf{u}^{o(k)}, \mathbf{l}^{c(k)}, \mathbf{u}^{c(k)}, \mathbf{l}^{f(k)}, \mathbf{u}^{f(k)}, \mathbf{l}^{i(k)}, \mathbf{u}^{i(k)}, \mathbf{l}^{g(k)}$ and $\mathbf{u}^{g(k)}$ for $\mathbf{X} \in \mathbb{R}^{n \times m}$, where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, let the pre-activation inputs for the j -th neuron at the output layer be $F_j(\mathbf{X}) = \mathbf{W}_{j,:}^{Fa} \mathbf{a}^{(m)} + \mathbf{b}_j^F$. The j -th output of the recurrent neural network is the following:

$$\begin{aligned}
F_j(\mathbf{X}) &= \mathbf{W}_{j,:}^{Fa} \mathbf{a}^{(m)} + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} \mathbf{a}_i^{(m)} + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} [\mathbf{o}_i^{(m)} \tanh(\mathbf{c}_i^{(m)})] + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} [\sigma(\mathbf{y}_i^{o(m)}) \tanh(\mathbf{c}_i^{(m)})] + \mathbf{b}_j^F, \\
&= \sum_{\mathbf{W}_{j,i}^{Fa} \geq 0} \mathbf{W}_{j,i}^{Fa} [\sigma(\mathbf{y}_i^{o(m)}) \tanh(\mathbf{c}_i^{(m)})] + \sum_{\mathbf{W}_{j,i}^{Fa} < 0} \mathbf{W}_{j,i}^{Fa} [\sigma(\mathbf{y}_i^{o(m)}) \tanh(\mathbf{c}_i^{(m)})] + \mathbf{b}_j^F.
\end{aligned}$$

Assume the activation function $\sigma(\mathbf{y}) \tanh(\mathbf{x})$ is bounded by two linear functions $h_{L,i}^{oc(m)}(\mathbf{y}, \mathbf{x}), h_{U,i}^{oc(m)}(\mathbf{y}, \mathbf{x})$, we have

$$h_{L,i}^{oc(m)}(\mathbf{y}_i^{o(m)}, \mathbf{c}_i^{(m)}) \leq \sigma(\mathbf{y}_i^{o(m)}) \tanh(\mathbf{c}_i^{(m)}) \leq h_{U,i}^{oc(m)}(\mathbf{y}_i^{o(m)}, \mathbf{c}_i^{(m)}).$$

Upper bound. We can then obtain

$$F_j(\mathbf{X}) \leq \sum_{\mathbf{W}_{j,i}^{Fa} \geq 0} \mathbf{W}_{j,i}^{Fa} (\alpha_{U,i}^{oc(m)} \mathbf{y}_i^{o(m)} + \beta_{U,i}^{oc(m)} \mathbf{c}_i^{(m)} + \gamma_{U,i}^{oc(m)}) \\ + \sum_{\mathbf{W}_{j,i}^{Fa} < 0} \mathbf{W}_{j,i}^{Fa} (\alpha_{L,i}^{oc(m)} \mathbf{y}_i^{o(m)} + \beta_{L,i}^{oc(m)} \mathbf{c}_i^{(m)} + \gamma_{L,i}^{oc(m)}) + \mathbf{b}_j^F.$$

We then define $\lambda_{j,i}^{oc(m)}$, $\Delta_{j,i}^{oc(m)}$ and $\varphi_{j,i}^{oc(m)}$ in the parentheses:

$$\lambda_{j,i}^{oc(m)} = \begin{cases} \alpha_{U,i}^{oc(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \alpha_{L,i}^{oc(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} \\ \Delta_{j,i}^{oc(m)} = \begin{cases} \beta_{U,i}^{oc(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \beta_{L,i}^{oc(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} \\ \varphi_{j,i}^{oc(m)} = \begin{cases} \gamma_{U,i}^{oc(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \gamma_{L,i}^{oc(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

and obtain

$$F_j(\mathbf{X}) \leq \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} (\lambda_{j,i}^{oc(m)} \mathbf{y}_i^{o(m)} + \Delta_{j,i}^{oc(m)} \mathbf{c}_i^{(m)} + \varphi_{j,i}^{oc(m)}) + \mathbf{b}_j^F.$$

We further let

$$\Lambda_{\lambda,j,i}^{oc(m)} := \mathbf{W}_{j,i}^{Fa} \lambda_{j,i}^{oc(m)}, \Lambda_{\Delta,j,i}^{oc(m)} := \mathbf{W}_{j,i}^{Fa} \Delta_{j,i}^{oc(m)}, \Lambda_{\varphi,j,i}^{oc(m)} := \mathbf{W}_{j,i}^{Fa} \varphi_{j,i}^{oc(m)},$$

and we have

$$F_j(\mathbf{X}) \leq \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{oc(m)} \mathbf{y}_i^{o(m)} + \Lambda_{\Delta,j,i}^{oc(m)} \mathbf{c}_i^{(m)} + \Lambda_{\varphi,j,i}^{oc(m)}) + \mathbf{b}_j^F, \quad (21) \\ = \sum_{i=1}^s [\Lambda_{\lambda,j,i}^{oc(m)} (\mathbf{W}_{i,:}^{ox} \mathbf{x}^{(m)} + \mathbf{W}_{i,:}^{oa} \mathbf{a}^{(m-1)} + \mathbf{b}_i^o) + \Lambda_{\Delta,j,i}^{oc(m)} (\mathbf{f}_i^{(m)} \mathbf{c}_i^{(m-1)} + \mathbf{i}_i^{(m)} \mathbf{g}_i^{(m)}) + \Lambda_{\varphi,j,i}^{oc(m)}] + \mathbf{b}_j^F, \\ = \sum_{q_m=1}^n (\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,q_m}^{ox}) \mathbf{x}_{q_m}^{(m)} + \sum_{r=1}^s (\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,r}^{oa}) \mathbf{a}_r^{(m-1)} + \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{oc(m)} \mathbf{b}_i^o + \Lambda_{\varphi,j,i}^{oc(m)}) \\ + \sum_{\Lambda_{\Delta,j,i}^{oc(m)} \geq 0} \Lambda_{\Delta,j,i}^{oc(m)} (\sigma(\mathbf{y}_i^{f(m)}) \mathbf{c}_i^{(m-1)} + \sigma(\mathbf{y}_i^{i(m)}) \tanh(\mathbf{y}_i^{g(m)})) \\ + \sum_{\Lambda_{\Delta,j,i}^{oc(m)} < 0} \Lambda_{\Delta,j,i}^{oc(m)} (\sigma(\mathbf{y}_i^{f(m)}) \mathbf{c}_i^{(m-1)} + \sigma(\mathbf{y}_i^{i(m)}) \tanh(\mathbf{y}_i^{g(m)})) + \mathbf{b}_j^F.$$

Assume the activation function $\sigma(\mathbf{y})\mathbf{x}$ is bounded by two linear functions $h_{L,i}^{fc(m)}(\mathbf{y}, \mathbf{x})$, $h_{U,i}^{fc(m)}(\mathbf{y}, \mathbf{x})$ and $\sigma(\mathbf{y}) \tanh(\mathbf{x})$ is bounded by two linear functions $h_{L,i}^{ig(m)}(\mathbf{y}, \mathbf{x})$, $h_{U,i}^{ig(m)}(\mathbf{y}, \mathbf{x})$, we have

$$h_{L,i}^{fc(m)}(\mathbf{y}_i^{f(m)}, \mathbf{c}_i^{(m-1)}) \leq \sigma(\mathbf{y}_i^{f(m)}) \mathbf{c}_i^{(m-1)} \leq h_{U,i}^{fc(m)}(\mathbf{y}_i^{f(m)}, \mathbf{c}_i^{(m-1)}), \\ h_{L,i}^{ig(m)}(\mathbf{y}_i^{i(m)}, \mathbf{y}_i^{g(m)}) \leq \sigma(\mathbf{y}_i^{i(m)}) \tanh(\mathbf{y}_i^{g(m)}) \leq h_{U,i}^{ig(m)}(\mathbf{y}_i^{i(m)}, \mathbf{y}_i^{g(m)})$$

We then have

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{q_m=1}^n \left(\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,q_m}^{ox} \right) \mathbf{x}_{q_m}^{(m)} + \sum_{r=1}^s \left(\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,r}^{oa} \right) \mathbf{a}_r^{(m-1)} + \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{oc(m)} \mathbf{b}_i^o + \Lambda_{\varphi,j,i}^{oc(m)}) + \mathbf{b}_j^F \\
&+ \sum_{\Lambda_{\Delta,j,i}^{oc(m)} \geq 0} \Lambda_{\Delta,j,i}^{oc(m)} (\alpha_{U,i}^{fc(m)} \mathbf{y}_i^{f(m)} + \beta_{U,i}^{fc(m)} \mathbf{c}_i^{(m-1)} + \gamma_{U,i}^{fc(m)} + \alpha_{U,i}^{ig(m)} \mathbf{y}_i^{i(m)} + \beta_{U,i}^{ig(m)} \mathbf{y}_i^{g(m)} + \gamma_{U,i}^{ig(m)}) \\
&+ \sum_{\Lambda_{\Delta,j,i}^{oc(m)} < 0} \Lambda_{\Delta,j,i}^{oc(m)} (\alpha_{L,i}^{fc(m)} \mathbf{y}_i^{f(m)} + \beta_{L,i}^{fc(m)} \mathbf{c}_i^{(m-1)} + \gamma_{L,i}^{fc(m)} + \alpha_{L,i}^{ig(m)} \mathbf{y}_i^{i(m)} + \beta_{L,i}^{ig(m)} \mathbf{y}_i^{g(m)} + \gamma_{L,i}^{ig(m)}).
\end{aligned}$$

We then define $\lambda_{j,i}^{fc(m)}$, $\Delta_{j,i}^{fc(m)}$, $\varphi_{j,i}^{fc(m)}$, $\Lambda_{\lambda,j,i}^{ig(m)}$, $\Lambda_{\Delta,j,i}^{ig(m)}$ and $\Lambda_{\varphi,j,i}^{ig(m)}$ in the parentheses:

$$\begin{aligned}
\lambda_{j,i}^{fc(m)} &= \begin{cases} \alpha_{U,i}^{fc(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} \geq 0; \\ \alpha_{L,i}^{fc(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} < 0; \end{cases} & \lambda_{j,i}^{ig(m)} &= \begin{cases} \alpha_{U,i}^{ig(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} \geq 0; \\ \alpha_{L,i}^{ig(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} < 0; \end{cases} \\
\Delta_{j,i}^{fc(m)} &= \begin{cases} \beta_{U,i}^{fc(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} \geq 0; \\ \beta_{L,i}^{fc(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} < 0; \end{cases} & \Delta_{j,i}^{ig(m)} &= \begin{cases} \beta_{U,i}^{ig(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} \geq 0; \\ \beta_{L,i}^{ig(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} < 0; \end{cases} \\
\varphi_{j,i}^{fc(m)} &= \begin{cases} \gamma_{U,i}^{fc(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} \geq 0; \\ \gamma_{L,i}^{fc(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} < 0; \end{cases} & \varphi_{j,i}^{ig(m)} &= \begin{cases} \gamma_{U,i}^{ig(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} \geq 0; \\ \gamma_{L,i}^{ig(m)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m)} < 0; \end{cases}
\end{aligned}$$

and obtain

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{q_m=1}^n \left(\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,q_m}^{ox} \right) \mathbf{x}_{q_m}^{(m)} + \sum_{r=1}^s \left(\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,r}^{oa} \right) \mathbf{a}_r^{(m-1)} + \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{oc(m)} \mathbf{b}_i^o + \Lambda_{\varphi,j,i}^{oc(m)}) + \mathbf{b}_j^F \\
&+ \sum_{i=1}^s \Lambda_{\Delta,j,i}^{oc(m)} (\lambda_{j,i}^{fc(m)} \mathbf{y}_i^{f(m)} + \Delta_{j,i}^{fc(m)} \mathbf{c}_i^{(m-1)} + \varphi_{j,i}^{fc(m)} + \lambda_{j,i}^{ig(m)} \mathbf{y}_i^{i(m)} + \Delta_{j,i}^{ig(m)} \mathbf{y}_i^{g(m)} + \varphi_{j,i}^{ig(m)}).
\end{aligned}$$

We further define

$$\begin{aligned}
\Lambda_{\lambda,j,i}^{fc(m)} &:= \Lambda_{\Delta,j,i}^{oc(m)} \lambda_{j,i}^{fc(m)}; \Lambda_{\Delta,j,i}^{fc(m)} := \Lambda_{\Delta,j,i}^{oc(m)} \Delta_{j,i}^{fc(m)}; \Lambda_{\varphi,j,i}^{fc(m)} := \Lambda_{\Delta,j,i}^{oc(m)} \varphi_{j,i}^{fc(m)}; \\
\Lambda_{\lambda,j,i}^{ig(m)} &:= \Lambda_{\Delta,j,i}^{oc(m)} \lambda_{j,i}^{ig(m)}; \Lambda_{\Delta,j,i}^{ig(m)} := \Lambda_{\Delta,j,i}^{oc(m)} \Delta_{j,i}^{ig(m)}; \Lambda_{\varphi,j,i}^{ig(m)} := \Lambda_{\Delta,j,i}^{oc(m)} \varphi_{j,i}^{ig(m)};
\end{aligned}$$

and obtain

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{q_m=1}^n \left(\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,q_m}^{ox} \right) \mathbf{x}_{q_m}^{(m)} + \sum_{r=1}^s \left(\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,r}^{oa} \right) \mathbf{a}_r^{(m-1)} + \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{oc(m)} \mathbf{b}_i^o + \Lambda_{\varphi,j,i}^{oc(m)}) + \mathbf{b}_j^F \\
&+ \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{fc(m)} \mathbf{y}_i^{f(m)} + \Lambda_{\Delta,j,i}^{fc(m)} \mathbf{c}_i^{(m-1)} + \Lambda_{\varphi,j,i}^{fc(m)} + \Lambda_{\lambda,j,i}^{ig(m)} \mathbf{y}_i^{i(m)} + \Lambda_{\Delta,j,i}^{ig(m)} \mathbf{y}_i^{g(m)} + \Lambda_{\varphi,j,i}^{ig(m)}) \\
&= \sum_{q_m=1}^n \left(\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,q_m}^{ox} \right) \mathbf{x}_{q_m}^{(m)} + \sum_{r=1}^s \left(\sum_{i=1}^s \Lambda_{\lambda,j,i}^{oc(m)} \mathbf{W}_{i,r}^{oa} \right) \mathbf{a}_r^{(m-1)} + \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{oc(m)} \mathbf{b}_i^o + \Lambda_{\varphi,j,i}^{oc(m)}) + \mathbf{b}_j^F \\
&+ \sum_{i=1}^s [\Lambda_{\lambda,j,i}^{fc(m)} (\mathbf{W}_{i,:}^{fx} \mathbf{x}^{(m)} + \mathbf{W}_{i,:}^{fa} \mathbf{a}^{(m-1)} + \mathbf{b}_i^f) + \Lambda_{\Delta,j,i}^{fc(m)} \mathbf{c}_i^{(m-1)} + \Lambda_{\varphi,j,i}^{fc(m)} \\
&+ \Lambda_{\lambda,j,i}^{ig(m)} (\mathbf{W}_{i,:}^{ix} \mathbf{x}^{(m)} + \mathbf{W}_{i,:}^{ia} \mathbf{a}^{(m-1)} + \mathbf{b}_i^i) + \Lambda_{\Delta,j,i}^{ig(m)} (\mathbf{W}_{i,:}^{gx} \mathbf{x}^{(m)} + \mathbf{W}_{i,:}^{ga} \mathbf{a}^{(m-1)} + \mathbf{b}_i^g) + \Lambda_{\varphi,j,i}^{ig(m)}] \\
&= \sum_{q_m=1}^n (\Lambda_{\lambda,j,:}^{oc(m)} \mathbf{W}_{:,q_m}^{ox} + \Lambda_{\lambda,j,:}^{fc(m)} \mathbf{W}_{:,q_m}^{fx} + \Lambda_{\lambda,j,:}^{ig(m)} \mathbf{W}_{:,q_m}^{ix} + \Lambda_{\Delta,j,:}^{ig(m)} \mathbf{W}_{:,q_m}^{gx}) \mathbf{x}_{q_m}^{(m)} \\
&+ \sum_{r=1}^s (\Lambda_{\lambda,j,:}^{oc(m)} \mathbf{W}_{:,r}^{oa} + \Lambda_{\lambda,j,:}^{fc(m)} \mathbf{W}_{:,r}^{fa} + \Lambda_{\lambda,j,:}^{ig(m)} \mathbf{W}_{:,r}^{ia} + \Lambda_{\Delta,j,:}^{ig(m)} \mathbf{W}_{:,r}^{ga}) \mathbf{a}_r^{(m-1)}
\end{aligned}$$

$$\begin{aligned}
& + \Lambda_{\lambda,j,:}^{oc(m)} \mathbf{b}^o + \Lambda_{\lambda,j,:}^{fc(m)} \mathbf{b}^f + \Lambda_{\lambda,j,:}^{ig(m)} \mathbf{b}^i + \Lambda_{\Delta,j,:}^{ig(m)} \mathbf{b}^g + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{oc(m)} + \Lambda_{\varphi,j,i}^{fc(m)} + \Lambda_{\varphi,j,i}^{ig(m)}) + \mathbf{b}_j^F \\
& + \Lambda_{\Delta,j,:}^{fc(m)} \mathbf{c}^{(m-1)}. \tag{22}
\end{aligned}$$

We combine coefficients in Equation (22) into equivalent weights $\tilde{\mathbf{W}}_{U,j,q_m}^{x(m)}$, $\tilde{\mathbf{W}}_{U,j,r}^{a(m)}$ and bias $\tilde{\mathbf{b}}_{U,j}^{(m)}$ as defined by

$$\begin{aligned}
\tilde{\mathbf{W}}_{U,j,q_m}^{x(m)} &= \Lambda_{\lambda,j,:}^{oc(m)} \mathbf{W}_{:,q_m}^{ox} + \Lambda_{\lambda,j,:}^{fc(m)} \mathbf{W}_{:,q_m}^{fx} + \Lambda_{\lambda,j,:}^{ig(m)} \mathbf{W}_{:,q_m}^{ix} + \Lambda_{\Delta,j,:}^{ig(m)} \mathbf{W}_{:,q_m}^{gx}, \\
\tilde{\mathbf{W}}_{U,j,r}^{a(m)} &= \Lambda_{\lambda,j,:}^{oc(m)} \mathbf{W}_{:,r}^{oa} + \Lambda_{\lambda,j,:}^{fc(m)} \mathbf{W}_{:,r}^{fa} + \Lambda_{\lambda,j,:}^{ig(m)} \mathbf{W}_{:,r}^{ia} + \Lambda_{\Delta,j,:}^{ig(m)} \mathbf{W}_{:,r}^{ga}, \\
\tilde{\mathbf{b}}_{U,j}^{(m)} &= \Lambda_{\lambda,j,:}^{oc(m)} \mathbf{b}^o + \Lambda_{\lambda,j,:}^{fc(m)} \mathbf{b}^f + \Lambda_{\lambda,j,:}^{ig(m)} \mathbf{b}^i + \Lambda_{\Delta,j,:}^{ig(m)} \mathbf{b}^g + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{oc(m)} + \Lambda_{\varphi,j,i}^{fc(m)} + \Lambda_{\varphi,j,i}^{ig(m)}),
\end{aligned}$$

and obtain

$$\begin{aligned}
F_j(\mathbf{X}) &\leq F_j^{U,m}(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{W}}_{U,j,:}^{a(m)} \mathbf{a}^{(m-1)} + \Lambda_{\Delta,j,:}^{fc(m)} \mathbf{c}^{(m-1)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F, \tag{23} \\
&= \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F + \Lambda_{\Delta,j,:}^{fc(m)} \mathbf{c}^{(m-1)} + \tilde{\mathbf{W}}_{U,j,:}^{a(m)} [\mathbf{o}^{(m-1)} \tanh(\mathbf{c}^{(m-1)})] \\
&= \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F + \Lambda_{\Delta,j,:}^{fc(m)} \mathbf{c}^{(m-1)} \\
&+ \sum_{\tilde{\mathbf{W}}_{U,j,i}^{a(m)} \geq 0} \tilde{\mathbf{W}}_{U,j,i}^{a(m)} [\sigma(\mathbf{y}_i^{o(m-1)}) \tanh(\mathbf{c}_i^{(m-1)})] + \sum_{\tilde{\mathbf{W}}_{U,j,i}^{a(m)} < 0} \tilde{\mathbf{W}}_{U,j,i}^{a(m)} [\sigma(\mathbf{y}_i^{o(m-1)}) \tanh(\mathbf{c}_i^{(m-1)})].
\end{aligned}$$

Assume the activation function $\sigma(\mathbf{y}) \tanh(\mathbf{x})$ is bounded by two linear functions $h_{L,i}^{oc(m)}(\mathbf{y}, \mathbf{x})$, $h_{U,i}^{oc(m)}(\mathbf{y}, \mathbf{x})$, we have

$$h_{L,i}^{oc(m-1)}(\mathbf{y}_i^{o(m-1)}, \mathbf{c}_i^{(m-1)}) \leq \sigma(\mathbf{y}_i^{o(m-1)}) \tanh(\mathbf{c}_i^{(m-1)}) \leq h_{U,i}^{oc(m-1)}(\mathbf{y}_i^{o(m-1)}, \mathbf{c}_i^{(m-1)}).$$

We can then obtain

$$\begin{aligned}
F_j^{U,m}(\mathbf{X}) &\leq \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F + \Lambda_{\Delta,j,:}^{fc(m)} \mathbf{c}^{(m-1)} \\
&+ \sum_{\tilde{\mathbf{W}}_{U,j,i}^{a(m)} \geq 0} \tilde{\mathbf{W}}_{U,j,i}^{a(m)} (\alpha_{U,i}^{oc(m-1)} \mathbf{y}_i^{o(m-1)} + \beta_{U,i}^{oc(m-1)} \mathbf{c}_i^{(m-1)} + \gamma_{U,i}^{oc(m-1)}) \\
&+ \sum_{\tilde{\mathbf{W}}_{U,j,i}^{a(m)} < 0} \tilde{\mathbf{W}}_{U,j,i}^{a(m)} (\alpha_{L,i}^{oc(m-1)} \mathbf{y}_i^{o(m-1)} + \beta_{L,i}^{oc(m-1)} \mathbf{c}_i^{(m-1)} + \gamma_{L,i}^{oc(m-1)}).
\end{aligned}$$

We then define $\lambda_{j,i}^{oc(m-1)}$, $\Delta_{j,i}^{oc(m-1)}$ and $\varphi_{j,i}^{oc(m-1)}$ in the parentheses:

$$\begin{aligned}
\lambda_{j,i}^{oc(m-1)} &= \begin{cases} \alpha_{U,i}^{oc(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \geq 0; \\ \alpha_{L,i}^{oc(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} < 0; \end{cases} \\
\Delta_{j,i}^{oc(m-1)} &= \begin{cases} \beta_{U,i}^{oc(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \geq 0; \\ \beta_{L,i}^{oc(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} < 0; \end{cases} \\
\varphi_{j,i}^{oc(m-1)} &= \begin{cases} \gamma_{U,i}^{oc(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \geq 0; \\ \gamma_{L,i}^{oc(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} < 0; \end{cases}
\end{aligned}$$

and obtain

$$\begin{aligned}
F_j^{U,m}(\mathbf{X}) &\leq \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F + \Lambda_{\Delta,j,:}^{fc(m)} \mathbf{c}^{(m-1)} \\
&+ \sum_{i=1}^s \tilde{\mathbf{W}}_{U,j,i}^{a(m)} (\lambda_{j,i}^{oc(m-1)} \mathbf{y}_i^{o(m-1)} + \Delta_{j,i}^{oc(m-1)} \mathbf{c}_i^{(m-1)} + \varphi_{j,i}^{oc(m-1)}).
\end{aligned}$$

We further let

$$\Lambda_{\lambda,j,i}^{oc(m-1)} := \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \lambda_{j,i}^{oc(m-1)}, \Lambda_{\Delta,j,i}^{oc(m-1)} := \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \Delta_{j,i}^{oc(m-1)} + \Lambda_{\Delta,j,i}^{fc(m)}, \Lambda_{\varphi,j,i}^{oc(m-1)} := \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \varphi_{j,i}^{oc(m-1)},$$

and we have

$$F_j^{U,m}(\mathbf{X}) \leq \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{oc(m-1)} \mathbf{y}_i^{o(m-1)} + \Lambda_{\Delta,j,i}^{oc(m-1)} \mathbf{c}_i^{(m-1)} + \Lambda_{\varphi,j,i}^{oc(m-1)}) + \mathbf{b}_j^F. \quad (24)$$

Notice that Equation (21) and Equation (24) are in similar forms. Thus, $\forall r \in s$ we can define variables as follows:

$$\begin{aligned} \lambda_{j,i}^{fc(m-1)} &= \begin{cases} \alpha_{U,i}^{fc(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} \geq 0; \\ \alpha_{L,i}^{fc(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} < 0; \end{cases} & \lambda_{j,i}^{ig(m-1)} &= \begin{cases} \alpha_{U,i}^{ig(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} \geq 0; \\ \alpha_{L,i}^{ig(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} < 0; \end{cases} \\ \Delta_{j,i}^{fc(m-1)} &= \begin{cases} \beta_{U,i}^{fc(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} \geq 0; \\ \beta_{L,i}^{fc(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} < 0; \end{cases} & \Delta_{j,i}^{ig(m-1)} &= \begin{cases} \beta_{U,i}^{ig(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} \geq 0; \\ \beta_{L,i}^{ig(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} < 0; \end{cases} \\ \varphi_{j,i}^{fc(m-1)} &= \begin{cases} \gamma_{U,i}^{fc(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} \geq 0; \\ \gamma_{L,i}^{fc(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} < 0; \end{cases} & \varphi_{j,i}^{ig(m-1)} &= \begin{cases} \gamma_{U,i}^{ig(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} \geq 0; \\ \gamma_{L,i}^{ig(m-1)} & \text{if } \Lambda_{\Delta,j,i}^{oc(m-1)} < 0; \end{cases} \\ \Lambda_{\lambda,j,i}^{fc(m-1)} &:= \Lambda_{\Delta,j,i}^{oc(m-1)} \lambda_{j,i}^{fc(m-1)}, \Lambda_{\Delta,j,i}^{fc(m-1)} := \Lambda_{\Delta,j,i}^{oc(m-1)} \Delta_{j,i}^{fc(m-1)}, \Lambda_{\varphi,j,i}^{fc(m-1)} := \Lambda_{\Delta,j,i}^{oc(m-1)} \varphi_{j,i}^{fc(m-1)}; \\ \Lambda_{\lambda,j,i}^{ig(m-1)} &:= \Lambda_{\Delta,j,i}^{oc(m-1)} \lambda_{j,i}^{ig(m-1)}, \Lambda_{\Delta,j,i}^{ig(m-1)} := \Lambda_{\Delta,j,i}^{oc(m-1)} \Delta_{j,i}^{ig(m-1)}, \Lambda_{\varphi,j,i}^{ig(m-1)} := \Lambda_{\Delta,j,i}^{oc(m-1)} \varphi_{j,i}^{ig(m-1)}; \\ \tilde{\mathbf{W}}_{U,j,q_{m-1}}^{x(m-1)} &= \Lambda_{\lambda,j,:}^{oc(m-1)} \mathbf{W}_{:,q_{m-1}}^{ox} + \Lambda_{\lambda,j,:}^{fc(m-1)} \mathbf{W}_{:,q_{m-1}}^{fx} + \Lambda_{\lambda,j,:}^{ig(m-1)} \mathbf{W}_{:,q_{m-1}}^{ix} + \Lambda_{\Delta,j,:}^{ig(m-1)} \mathbf{W}_{:,q_{m-1}}^{gx}; \\ \tilde{\mathbf{W}}_{U,j,r}^{a(m-1)} &= \Lambda_{\lambda,j,:}^{oc(m-1)} \mathbf{W}_{:,r}^{oa} + \Lambda_{\lambda,j,:}^{fc(m-1)} \mathbf{W}_{:,r}^{fa} + \Lambda_{\lambda,j,:}^{ig(m-1)} \mathbf{W}_{:,r}^{ia} + \Lambda_{\Delta,j,:}^{ig(m-1)} \mathbf{W}_{:,r}^{ga}; \\ \tilde{\mathbf{b}}_{U,j}^{(m-1)} &= \Lambda_{\lambda,j,:}^{oc(m-1)} \mathbf{b}^o + \Lambda_{\lambda,j,:}^{fc(m-1)} \mathbf{b}^f + \Lambda_{\lambda,j,:}^{ig(m-1)} \mathbf{b}^i + \Lambda_{\Delta,j,:}^{ig(m-1)} \mathbf{b}^g + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{oc(m-1)} + \Lambda_{\varphi,j,i}^{fc(m-1)} + \Lambda_{\varphi,j,i}^{ig(m-1)}). \end{aligned}$$

Then we can obtain equation similar to Equation (3):

$$\begin{aligned} F_j^{U,m}(\mathbf{X}) \leq F_j^{U,m-1}(\mathbf{X}) &= \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{W}}_{U,j,:}^{x(m-1)} \mathbf{x}^{(m-1)} + \tilde{\mathbf{W}}_{U,j,:}^{a(m-1)} \mathbf{a}^{(m-2)} \\ &\quad + \Lambda_{\Delta,j,:}^{fc(m-1)} \mathbf{c}^{(m-2)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m-1)} + \mathbf{b}_j^F. \end{aligned}$$

Thus, we can repeat the above procedure iteratively until we obtain the final upper bound $F_j^{U,1}(\mathbf{X})$, where $F_j(\mathbf{X}) \leq F_j^{U,m}(\mathbf{X}) \leq \dots \leq F_j^{U,1}(\mathbf{X})$. We let $F_j^U(\mathbf{X})$ denote the final upper bound $F_j^{U,1}(\mathbf{X})$, and we have

$$F_j^U(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \Lambda_{\Delta,j,:}^{fc(1)} \mathbf{c}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{U,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{U,j}^{(k)} + \mathbf{b}_j^F,$$

where

$$\begin{aligned} \tilde{\mathbf{W}}_{U,j,q_k}^{x(k)} &= \Lambda_{\lambda,j,:}^{oc(k)} \mathbf{W}_{:,q_k}^{ox} + \Lambda_{\lambda,j,:}^{fc(k)} \mathbf{W}_{:,q_k}^{fx} + \Lambda_{\lambda,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{ix} + \Lambda_{\Delta,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{gx}, \\ \tilde{\mathbf{W}}_{U,j,r}^{a(k)} &= \Lambda_{\lambda,j,:}^{oc(k)} \mathbf{W}_{:,r}^{oa} + \Lambda_{\lambda,j,:}^{fc(k)} \mathbf{W}_{:,r}^{fa} + \Lambda_{\lambda,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ia} + \Lambda_{\Delta,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ga}, \\ \tilde{\mathbf{b}}_{U,j}^{(k)} &= \Lambda_{\lambda,j,:}^{oc(k)} \mathbf{b}^o + \Lambda_{\lambda,j,:}^{fc(k)} \mathbf{b}^f + \Lambda_{\lambda,j,:}^{ig(k)} \mathbf{b}^i + \Lambda_{\Delta,j,:}^{ig(k)} \mathbf{b}^g + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{oc(k)} + \Lambda_{\varphi,j,i}^{fc(k)} + \Lambda_{\varphi,j,i}^{ig(k)}), \end{aligned}$$

and (\odot is the Hadamard product)

$$\begin{aligned} \Lambda_{\lambda,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \lambda_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases} \\ \Lambda_{\Delta,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Delta_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \Delta_{j,:}^{oc(k)} + \Lambda_{\Delta,j,:}^{fc(k+1)} & \text{if } k \in [m-1]; \end{cases} \\ \Lambda_{\varphi,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \varphi_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \varphi_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases} \end{aligned}$$

$$\begin{aligned}
\Lambda_{\lambda,j,:}^{fc(k)} &= \Lambda_{\Delta,j,:}^{oc(k)} \odot \lambda_{j,:}^{fc(k)} & \text{if } k \in [m]; & \quad \Lambda_{\lambda,j,:}^{ig(k)} = \Lambda_{\Delta,j,:}^{oc(k)} \odot \lambda_{j,:}^{ig(k)} & \text{if } k \in [m]; \\
\Lambda_{\Delta,j,:}^{fc(k)} &= \Lambda_{\Delta,j,:}^{oc(k)} \odot \Delta_{j,:}^{fc(k)} & \text{if } k \in [m]; & \quad \Lambda_{\Delta,j,:}^{ig(k)} = \Lambda_{\Delta,j,:}^{oc(k)} \odot \Delta_{j,:}^{ig(k)} & \text{if } k \in [m]; \\
\Lambda_{\varphi,j,:}^{fc(k)} &= \Lambda_{\Delta,j,:}^{oc(k)} \odot \varphi_{j,:}^{fc(k)} & \text{if } k \in [m]; & \quad \Lambda_{\varphi,j,:}^{ig(k)} = \Lambda_{\Delta,j,:}^{oc(k)} \odot \varphi_{j,:}^{ig(k)} & \text{if } k \in [m];
\end{aligned}$$

and $\forall r \in s$,

$$\lambda_{j,r}^{oc(m)} = \begin{cases} \alpha_{U,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \alpha_{L,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} \quad \Delta_{j,r}^{oc(m)} = \begin{cases} \beta_{U,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \beta_{L,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} \quad \varphi_{j,r}^{oc(m)} = \begin{cases} \gamma_{U,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \gamma_{L,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases}$$

and $\forall r \in s, k \in [m-1]$,

$$\begin{aligned}
\lambda_{j,r}^{oc(k)} &= \begin{cases} \alpha_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \alpha_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} \\
\Delta_{j,r}^{oc(k)} &= \begin{cases} \beta_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \beta_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} \\
\varphi_{j,r}^{oc(k)} &= \begin{cases} \gamma_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \gamma_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases}
\end{aligned}$$

and $\forall r \in s, k \in [m]$,

$$\begin{aligned}
\lambda_{j,r}^{fc(k)} &= \begin{cases} \alpha_{U,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \alpha_{L,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} \quad \lambda_{j,r}^{ig(k)} = \begin{cases} \alpha_{U,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \alpha_{L,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} \\
\Delta_{j,r}^{fc(k)} &= \begin{cases} \beta_{U,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \beta_{L,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} \quad \Delta_{j,r}^{ig(k)} = \begin{cases} \beta_{U,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \beta_{L,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} \\
\varphi_{j,r}^{fc(k)} &= \begin{cases} \gamma_{U,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \gamma_{L,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} \quad \varphi_{j,r}^{ig(k)} = \begin{cases} \gamma_{U,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \gamma_{L,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0. \end{cases}
\end{aligned}$$

Lower bound. The above derivations of upper bound can be applied similarly to derive lower bounds of $F_j(\mathbf{X})$, and the only difference is now we need to use the left-hand side of inequities (rather than right-hand side when deriving upper bound) to bound the activated terms. Thus, following the same procedure in deriving the upper bounds, we can iteratively unwrap the activation functions and obtain a final lower bound $F_j^{L,1}(\mathbf{X})$, where $F_j(\mathbf{X}) \geq F_j^{L,m}(\mathbf{X}) \geq \dots \geq F_j^{L,1}(\mathbf{X})$. We let $F_j^L(\mathbf{X})$ denote the final lower bound $F_j^{L,1}(\mathbf{X})$, and we have

$$F_j^L(\mathbf{X}) = \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} + \Omega_{\Delta,j,:}^{fc(1)} \mathbf{c}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{L,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{L,j}^{(k)} + \mathbf{b}_j^F,$$

where

$$\begin{aligned}
\tilde{\mathbf{W}}_{L,j,q_k}^{x(k)} &= \Omega_{\omega,j,:}^{oc(k)} \mathbf{W}_{:,q_k}^{ox} + \Omega_{\omega,j,:}^{fc(k)} \mathbf{W}_{:,q_k}^{fx} + \Omega_{\omega,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{ix} + \Omega_{\Theta,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{gx}, \\
\tilde{\mathbf{W}}_{L,j,r}^{a(k)} &= \Omega_{\omega,j,:}^{oc(k)} \mathbf{W}_{:,r}^{oa} + \Omega_{\omega,j,:}^{fc(k)} \mathbf{W}_{:,r}^{fa} + \Omega_{\omega,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ia} + \Omega_{\Theta,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ga}, \\
\tilde{\mathbf{b}}_{L,j}^{(k)} &= \Omega_{\omega,j,:}^{oc(k)} \mathbf{b}^o + \Omega_{\omega,j,:}^{fc(k)} \mathbf{b}^f + \Omega_{\omega,j,:}^{ig(k)} \mathbf{b}^i + \Omega_{\Theta,j,:}^{ig(k)} \mathbf{b}^g + \sum_{i=1}^s (\Omega_{\psi,j,i}^{oc(k)} + \Omega_{\psi,j,i}^{fc(k)} + \Omega_{\psi,j,i}^{ig(k)}),
\end{aligned}$$

and (\odot is the Hadamard product)

$$\begin{aligned}\Omega_{\omega,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \omega_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases} \\ \Omega_{\Theta,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Theta_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \Theta_{j,:}^{oc(k)} + \Omega_{\Theta,j,:}^{fc(k+1)} & \text{if } k \in [m-1]; \end{cases} \\ \Omega_{\psi,j,:}^{oc(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \psi_{j,:}^{oc(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \psi_{j,:}^{oc(k)} & \text{if } k \in [m-1]; \end{cases}\end{aligned}$$

$$\begin{aligned}\Omega_{\omega,j,:}^{fc(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \omega_{j,:}^{fc(k)} & \text{if } k \in [m]; & \quad \Omega_{\omega,j,:}^{ig(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \omega_{j,:}^{ig(k)} & \text{if } k \in [m]; \\ \Omega_{\Theta,j,:}^{fc(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \Theta_{j,:}^{fc(k)} & \text{if } k \in [m]; & \quad \Omega_{\Theta,j,:}^{ig(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \Theta_{j,:}^{ig(k)} & \text{if } k \in [m]; \\ \Omega_{\psi,j,:}^{fc(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \psi_{j,:}^{fc(k)} & \text{if } k \in [m]; & \quad \Omega_{\psi,j,:}^{ig(k)} &= \Omega_{\Theta,j,:}^{oc(k)} \odot \psi_{j,:}^{ig(k)} & \text{if } k \in [m];\end{aligned}$$

and $\forall r \in s$,

$$\omega_{j,r}^{oc(m)} = \begin{cases} \alpha_{L,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \alpha_{U,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} \quad \Theta_{j,r}^{oc(m)} = \begin{cases} \beta_{L,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \beta_{U,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} \quad \psi_{j,r}^{oc(m)} = \begin{cases} \gamma_{L,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \gamma_{U,r}^{oc(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases}$$

and $\forall r \in s, k \in [m-1]$,

$$\begin{aligned}\omega_{j,r}^{oc(k)} &= \begin{cases} \alpha_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \alpha_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\ \Theta_{j,r}^{oc(k)} &= \begin{cases} \beta_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \beta_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\ \psi_{j,r}^{oc(k)} &= \begin{cases} \gamma_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \gamma_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases}\end{aligned}$$

and $\forall r \in s, k \in [m]$,

$$\begin{aligned}\omega_{j,r}^{fc(k)} &= \begin{cases} \alpha_{L,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \alpha_{U,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} & \omega_{j,r}^{ig(k)} &= \begin{cases} \alpha_{L,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \alpha_{U,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \Theta_{j,r}^{fc(k)} &= \begin{cases} \beta_{L,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \beta_{U,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} & \Theta_{j,r}^{ig(k)} &= \begin{cases} \beta_{L,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \beta_{U,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \psi_{j,r}^{fc(k)} &= \begin{cases} \gamma_{L,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \gamma_{U,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} & \psi_{j,r}^{ig(k)} &= \begin{cases} \gamma_{L,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \gamma_{U,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0. \end{cases}\end{aligned}$$

Corollary A. 2 (Closed-form global bounds) Given data $\mathbf{X} \in \mathbb{R}^{n \times m}$, l_p ball parameters $p \geq 1$ and $\epsilon \geq 0$. For an m -layer long short-term memory network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$, there exists two fixed values γ_j^L and γ_j^U such that $\forall \mathbf{X} \in \mathbb{R}^{n \times m}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, and $\forall j \in [t]$, $1/q = 1 - 1/p$, the inequity $\gamma_j^L \leq F_j(\mathbf{X}) \leq \gamma_j^U$ holds true, where

$$\begin{aligned}\gamma_j^U &= \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \Lambda_{\Delta,j,:}^{fc(1)} \mathbf{c}^{(0)} + \sum_{k=1}^m \epsilon \|\tilde{\mathbf{W}}_{U,j,:}^{x(k)}\|_q + \sum_{k=1}^m \tilde{\mathbf{W}}_{U,j,:}^{x(k)} \mathbf{x}_0^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{U,j}^{(k)} + \mathbf{b}_j^F, \\ \gamma_j^L &= \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} + \Omega_{\Theta,j,:}^{fc(1)} \mathbf{c}^{(0)} - \sum_{k=1}^m \epsilon \|\tilde{\mathbf{W}}_{L,j,:}^{x(k)}\|_q + \sum_{k=1}^m \tilde{\mathbf{W}}_{L,j,:}^{x(k)} \mathbf{x}_0^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{L,j}^{(k)} + \mathbf{b}_j^F.\end{aligned}$$

Corollary A. 3 (Pre-activation bounds in LSTMs) *The pre-activation bounds of different gates can be obtained via simple modifications from Theorem A.3. Specifically, we recall the definitions of pre-activations in different gates:*

$$\begin{aligned}
\text{Input gate: } \quad \underline{\mathbf{y}}^{i(k)} &= \underline{\mathbf{W}^{ix} \mathbf{x}^{(k)}} + \mathbf{W}^{ia} \mathbf{a}^{(k-1)} + \mathbf{b}^i; \\
\text{Forget gate: } \quad \underline{\mathbf{y}}^{f(k)} &= \underline{\mathbf{W}^{fx} \mathbf{x}^{(k)}} + \mathbf{W}^{fa} \mathbf{a}^{(k-1)} + \mathbf{b}^f; \\
\text{Cell gate: } \quad \underline{\mathbf{y}}^{g(k)} &= \underline{\mathbf{W}^{gx} \mathbf{x}^{(k)}} + \mathbf{W}^{ga} \mathbf{a}^{(k-1)} + \mathbf{b}^g; \\
\text{Output gate: } \quad \underline{\mathbf{y}}^{o(k)} &= \underline{\mathbf{W}^{ox} \mathbf{x}^{(k)}} + \mathbf{W}^{oa} \mathbf{a}^{(k-1)} + \mathbf{b}^o.
\end{aligned}$$

For a given LSTM and input \mathbf{X}_0 , the underlined part of the equations are bounded by

$$\mathbf{W}^{gate} \mathbf{x}_0^{(k)} - \epsilon \|\mathbf{W}^{gate}\|_q \leq \underline{\mathbf{W}^{gate} \mathbf{x}^{(k)}} \leq \mathbf{W}^{gate} \mathbf{x}_0^{(k)} + \epsilon \|\mathbf{W}^{gate}\|_q,$$

where $gate = \{ix, fx, gx, ox\}$. Thus to derive pre-activation bounds, we only need to know the ranges of remaining parts in the equations above. These can be computed by replacing the output mapping \mathbf{W}^{Fa} in Theorem A.3 by $\mathbf{W}^{ia}, \mathbf{W}^{fa}, \mathbf{W}^{ga}, \mathbf{W}^{oa}$, respectively. The output bias \mathbf{b}^F is correspondingly substituted with $\mathbf{b}^i, \mathbf{b}^f, \mathbf{b}^g, \mathbf{b}^o$. For example, to derive the bounds for pre-activation bounds of input gates, we replace the output mapping matrix and bias from \mathbf{W}^{Fa} and \mathbf{b}^F by \mathbf{W}^{ia} and \mathbf{b}^i .

Theorem A. 4 (Closed-form cell state bounds of long short-term memory network F) *For the cell state $\mathbf{c}_j^{(v)}$ of the v -th layer in a given m -layer ($v \leq m$) long short-term memory network, there exists two explicit functions $g_j^{L,v} : \mathbb{R}^{n \times v} \rightarrow \mathbb{R}^s$ and $g_j^{U,v} : \mathbb{R}^{n \times v} \rightarrow \mathbb{R}^s$ and two fixed values $\mathbf{l}_j^{c(v)}$ and $\mathbf{u}_j^{c(v)}$ such that $\forall j \in [s]$ and $\forall \mathbf{X}^{(1:v)} \in \mathbb{R}^{n \times v}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$ and $1/q = 1 - 1/p$, the inequity $\mathbf{l}_j^{c(v)} \leq g_j^{L,v}(\mathbf{X}^{(1:v)}) \leq \mathbf{y}_j^{(v)} \leq g_j^{U,v}(\mathbf{X}^{(1:v)}) \leq \mathbf{u}_j^{c(v)}$ holds true, where*

$$\begin{aligned}
\mathbf{u}_j^{c(v)} &= \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \Lambda_{\Delta,j,:}^{fc(1)} \mathbf{c}^{(0)} + \sum_{k=1}^v \epsilon \|\tilde{\mathbf{W}}_{U,j,:}^{x(k)}\|_q + \sum_{k=1}^v \tilde{\mathbf{W}}_{U,j,:}^{x(k)} \mathbf{x}_0^{(k)} + \sum_{k=1}^v \tilde{\mathbf{b}}_{U,j}^{(k)}, \\
\mathbf{l}_j^{c(v)} &= \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} + \Omega_{\Theta,j,:}^{fc(1)} \mathbf{c}^{(0)} - \sum_{k=1}^v \epsilon \|\tilde{\mathbf{W}}_{L,j,:}^{x(k)}\|_q + \sum_{k=1}^v \tilde{\mathbf{W}}_{L,j,:}^{x(k)} \mathbf{x}_0^{(k)} + \sum_{k=1}^v \tilde{\mathbf{b}}_{L,j}^{(k)},
\end{aligned}$$

where

$$\begin{aligned}
\tilde{\mathbf{W}}_{U,j,q_v}^{x(v)} &= \Lambda_{\lambda,j,:}^{fc(v)} \mathbf{W}_{:,q_v}^{fx} + \Lambda_{\lambda,j,:}^{ig(v)} \mathbf{W}_{:,q_v}^{ix} + \Lambda_{\Delta,j,:}^{ig(v)} \mathbf{W}_{:,q_v}^{gx}, \\
\tilde{\mathbf{W}}_{U,j,r}^{a(v)} &= \Lambda_{\lambda,j,:}^{fc(v)} \mathbf{W}_{:,r}^{fa} + \Lambda_{\lambda,j,:}^{ig(v)} \mathbf{W}_{:,r}^{ia} + \Lambda_{\Delta,j,:}^{ig(v)} \mathbf{W}_{:,r}^{ga}, \\
\tilde{\mathbf{b}}_{U,j}^{(v)} &= \Lambda_{\lambda,j,:}^{fc(v)} \mathbf{b}^f + \Lambda_{\lambda,j,:}^{ig(v)} \mathbf{b}^i + \Lambda_{\Delta,j,:}^{ig(v)} \mathbf{b}^g + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{fc(v)} + \Lambda_{\varphi,j,i}^{ig(v)}), \\
\tilde{\mathbf{W}}_{L,j,q_v}^{x(v)} &= \Omega_{\omega,j,:}^{fc(v)} \mathbf{W}_{:,q_v}^{fx} + \Omega_{\omega,j,:}^{ig(v)} \mathbf{W}_{:,q_v}^{ix} + \Omega_{\Theta,j,:}^{ig(v)} \mathbf{W}_{:,q_v}^{gx}, \\
\tilde{\mathbf{W}}_{L,j,r}^{a(v)} &= \Omega_{\omega,j,:}^{fc(v)} \mathbf{W}_{:,r}^{fa} + \Omega_{\omega,j,:}^{ig(v)} \mathbf{W}_{:,r}^{ia} + \Omega_{\Theta,j,:}^{ig(v)} \mathbf{W}_{:,r}^{ga}, \\
\tilde{\mathbf{b}}_{L,j}^{(v)} &= \Omega_{\omega,j,:}^{fc(v)} \mathbf{b}^f + \Omega_{\omega,j,:}^{ig(v)} \mathbf{b}^i + \Omega_{\Theta,j,:}^{ig(v)} \mathbf{b}^g + \sum_{i=1}^s (\Omega_{\psi,j,i}^{fc(v)} + \Omega_{\psi,j,i}^{ig(v)}),
\end{aligned}$$

and $\forall k \in [v-1]$

$$\begin{aligned}
\tilde{\mathbf{W}}_{U,j,q_k}^{x(k)} &= \Lambda_{\lambda,j,:}^{oc(k)} \mathbf{W}_{:,q_k}^{ox} + \Lambda_{\lambda,j,:}^{fc(k)} \mathbf{W}_{:,q_k}^{fx} + \Lambda_{\lambda,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{ix} + \Lambda_{\Delta,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{gx}, \\
\tilde{\mathbf{W}}_{U,j,r}^{a(k)} &= \Lambda_{\lambda,j,:}^{oc(k)} \mathbf{W}_{:,r}^{oa} + \Lambda_{\lambda,j,:}^{fc(k)} \mathbf{W}_{:,r}^{fa} + \Lambda_{\lambda,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ia} + \Lambda_{\Delta,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ga}, \\
\tilde{\mathbf{b}}_{U,j}^{(k)} &= \Lambda_{\lambda,j,:}^{oc(k)} \mathbf{b}^o + \Lambda_{\lambda,j,:}^{fc(k)} \mathbf{b}^f + \Lambda_{\lambda,j,:}^{ig(k)} \mathbf{b}^i + \Lambda_{\Delta,j,:}^{ig(k)} \mathbf{b}^g + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{oc(k)} + \Lambda_{\varphi,j,i}^{fc(k)} + \Lambda_{\varphi,j,i}^{ig(k)}), \\
\tilde{\mathbf{W}}_{L,j,q_k}^{x(k)} &= \Omega_{\omega,j,:}^{oc(k)} \mathbf{W}_{:,q_k}^{ox} + \Omega_{\omega,j,:}^{fc(k)} \mathbf{W}_{:,q_k}^{fx} + \Omega_{\omega,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{ix} + \Omega_{\Theta,j,:}^{ig(k)} \mathbf{W}_{:,q_k}^{gx},
\end{aligned}$$

$$\begin{aligned}\tilde{\mathbf{W}}_{L,j,r}^{a(k)} &= \Omega_{\omega,j,:}^{oc(k)} \mathbf{W}_{:,r}^{oa} + \Omega_{\omega,j,:}^{fc(k)} \mathbf{W}_{:,r}^{fa} + \Omega_{\omega,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ia} + \Omega_{\Theta,j,:}^{ig(k)} \mathbf{W}_{:,r}^{ga}, \\ \tilde{\mathbf{b}}_{L,j}^{(k)} &= \Omega_{\omega,j,:}^{oc(k)} \mathbf{b}^o + \Omega_{\omega,j,:}^{fc(k)} \mathbf{b}^f + \Omega_{\omega,j,:}^{ig(k)} \mathbf{b}^i + \Omega_{\Theta,j,:}^{ig(k)} \mathbf{b}^g + \sum_{i=1}^s (\Omega_{\psi,j,i}^{oc(k)} + \Omega_{\psi,j,i}^{fc(k)} + \Omega_{\psi,j,i}^{ig(k)}),\end{aligned}$$

and

$$\begin{aligned}\Lambda_{\lambda,::}^{oc(k)} &= \tilde{\mathbf{W}}_{U,::}^{a(k+1)} \odot \lambda_{::}^{oc(k)} & \text{if } k \in [v-1]; & \quad \Omega_{\omega,::}^{oc(k)} = \tilde{\mathbf{W}}_{L,::}^{a(k+1)} \odot \omega_{::}^{oc(k)} & \text{if } k \in [v-1]; \\ \Lambda_{\Delta,::}^{oc(k)} &= \tilde{\mathbf{W}}_{U,::}^{a(k+1)} \odot \Delta_{::}^{oc(k)} + \Lambda_{\Delta,::}^{fc(k+1)} & \text{if } k \in [v-1]; & \quad \Omega_{\Theta,::}^{oc(k)} = \tilde{\mathbf{W}}_{L,::}^{a(k+1)} \odot \Theta_{::}^{oc(k)} + \Omega_{\Theta,::}^{fc(k+1)} & \text{if } k \in [v-1]; \\ \Lambda_{\varphi,::}^{oc(k)} &= \tilde{\mathbf{W}}_{U,::}^{a(k+1)} \odot \varphi_{::}^{oc(k)} & \text{if } k \in [v-1]; & \quad \Omega_{\psi,::}^{oc(k)} = \tilde{\mathbf{W}}_{L,::}^{a(k+1)} \odot \psi_{::}^{oc(k)} & \text{if } k \in [v-1]; \\ \Lambda_{\lambda,::}^{fc(k)} &= \begin{cases} \text{diag}(\alpha_{U,::}^{fc(k)}) & \text{if } k = v; \\ \Lambda_{\Delta,::}^{oc(k)} \odot \lambda_{::}^{fc(k)} & \text{if } k \in [v-1]; \end{cases} & \quad \Omega_{\omega,::}^{fc(k)} = \begin{cases} \text{diag}(\alpha_{L,::}^{fc(k)}) & \text{if } k = v; \\ \Omega_{\Theta,::}^{oc(k)} \odot \omega_{::}^{fc(k)} & \text{if } k \in [v-1]; \end{cases} \\ \Lambda_{\Delta,::}^{fc(k)} &= \begin{cases} \text{diag}(\beta_{U,::}^{fc(k)}) & \text{if } k = v; \\ \Lambda_{\Delta,::}^{oc(k)} \odot \Delta_{::}^{fc(k)} & \text{if } k \in [v-1]; \end{cases} & \quad \Omega_{\Theta,::}^{fc(k)} = \begin{cases} \text{diag}(\beta_{L,::}^{fc(k)}) & \text{if } k = v; \\ \Omega_{\Theta,::}^{oc(k)} \odot \Theta_{::}^{fc(k)} & \text{if } k \in [v-1]; \end{cases} \\ \Lambda_{\varphi,::}^{fc(k)} &= \begin{cases} \text{diag}(\gamma_{U,::}^{fc(k)}) & \text{if } k = v; \\ \Lambda_{\Delta,::}^{oc(k)} \odot \varphi_{::}^{fc(k)} & \text{if } k \in [v-1]; \end{cases} & \quad \Omega_{\psi,::}^{fc(k)} = \begin{cases} \text{diag}(\gamma_{L,::}^{fc(k)}) & \text{if } k = v; \\ \Omega_{\Theta,::}^{oc(k)} \odot \psi_{::}^{fc(k)} & \text{if } k \in [v-1]; \end{cases} \\ \Lambda_{\lambda,::}^{ig(k)} &= \begin{cases} \text{diag}(\alpha_{U,::}^{ig(k)}) & \text{if } k = v; \\ \Lambda_{\Delta,::}^{oc(k)} \odot \lambda_{::}^{ig(k)} & \text{if } k \in [v-1]; \end{cases} & \quad \Omega_{\omega,::}^{ig(k)} = \begin{cases} \text{diag}(\alpha_{L,::}^{ig(k)}) & \text{if } k = v; \\ \Omega_{\Theta,::}^{oc(k)} \odot \omega_{::}^{ig(k)} & \text{if } k \in [v-1]; \end{cases} \\ \Lambda_{\Delta,::}^{ig(k)} &= \begin{cases} \text{diag}(\beta_{U,::}^{ig(k)}) & \text{if } k = v; \\ \Lambda_{\Delta,::}^{oc(k)} \odot \Delta_{::}^{ig(k)} & \text{if } k \in [v-1]; \end{cases} & \quad \Omega_{\Theta,::}^{ig(k)} = \begin{cases} \text{diag}(\beta_{L,::}^{ig(k)}) & \text{if } k = v; \\ \Omega_{\Theta,::}^{oc(k)} \odot \Theta_{::}^{ig(k)} & \text{if } k \in [v-1]; \end{cases} \\ \Lambda_{\varphi,::}^{ig(k)} &= \begin{cases} \text{diag}(\gamma_{U,::}^{ig(k)}) & \text{if } k = v; \\ \Lambda_{\Delta,::}^{oc(k)} \odot \varphi_{::}^{ig(k)} & \text{if } k \in [v-1]; \end{cases} & \quad \Omega_{\psi,::}^{ig(k)} = \begin{cases} \text{diag}(\gamma_{L,::}^{ig(k)}) & \text{if } k = v; \\ \Omega_{\Theta,::}^{oc(k)} \odot \psi_{::}^{ig(k)} & \text{if } k \in [v-1]; \end{cases}\end{aligned}$$

and $\forall r \in s, k \in [v-1]$,

$$\begin{aligned}\lambda_{j,r}^{oc(k)} &= \begin{cases} \alpha_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \alpha_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \quad \omega_{j,r}^{oc(k)} &= \begin{cases} \alpha_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \alpha_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\ \Delta_{j,r}^{oc(k)} &= \begin{cases} \beta_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \beta_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \quad \Theta_{j,r}^{oc(k)} &= \begin{cases} \beta_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \beta_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\ \varphi_{j,r}^{oc(k)} &= \begin{cases} \gamma_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \gamma_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \quad \psi_{j,r}^{oc(k)} &= \begin{cases} \gamma_{L,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \gamma_{U,r}^{oc(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\ \lambda_{j,r}^{fc(k)} &= \begin{cases} \alpha_{U,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \alpha_{L,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \quad \omega_{j,r}^{fc(k)} &= \begin{cases} \alpha_{L,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \alpha_{U,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \Delta_{j,r}^{fc(k)} &= \begin{cases} \beta_{U,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \beta_{L,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \quad \Theta_{j,r}^{fc(k)} &= \begin{cases} \beta_{L,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \beta_{U,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \varphi_{j,r}^{fc(k)} &= \begin{cases} \gamma_{U,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \gamma_{L,r}^{fc(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \quad \psi_{j,r}^{fc(k)} &= \begin{cases} \gamma_{L,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \gamma_{U,r}^{fc(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \lambda_{j,r}^{ig(k)} &= \begin{cases} \alpha_{U,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \alpha_{L,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \quad \omega_{j,r}^{ig(k)} &= \begin{cases} \alpha_{L,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \alpha_{U,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases} \\ \Delta_{j,r}^{ig(k)} &= \begin{cases} \beta_{U,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} \geq 0; \\ \beta_{L,r}^{ig(k)} & \text{if } \Lambda_{\Delta,j,r}^{oc(k)} < 0; \end{cases} & \quad \Theta_{j,r}^{ig(k)} &= \begin{cases} \beta_{L,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} \geq 0; \\ \beta_{U,r}^{ig(k)} & \text{if } \Omega_{\Theta,j,r}^{oc(k)} < 0; \end{cases}\end{aligned}$$

$$\varphi_{j,r}^{ig(k)} = \begin{cases} \gamma_{U,r}^{ig(k)} & \text{if } \mathbf{\Lambda}_{\Delta,j,r}^{oc(k)} \geq 0; \\ \gamma_{L,r}^{ig(k)} & \text{if } \mathbf{\Lambda}_{\Delta,j,r}^{oc(k)} < 0; \end{cases} \quad \psi_{j,r}^{ig(k)} = \begin{cases} \gamma_{L,r}^{ig(k)} & \text{if } \mathbf{\Omega}_{\Theta,j,r}^{oc(k)} \geq 0; \\ \gamma_{U,r}^{ig(k)} & \text{if } \mathbf{\Omega}_{\Theta,j,r}^{oc(k)} < 0; \end{cases}$$

where \odot is the Hadamard product.

A.4 Bounding Planes for 2D Nonlinear Activation Functions

In this part, we elaborate on how bounding linear functions (planes) $h_{L,r}^{oc(k)}$, $h_{U,r}^{oc(k)}$, $h_{L,r}^{fc(k)}$, $h_{U,r}^{fc(k)}$, $h_{L,r}^{ig(k)}$, and $h_{U,r}^{ig(k)}$ mentioned in Definition 1 are computed. This problem corresponds to finding two bounding planes $h_{L,r}(\mathbf{v}, \mathbf{z}) = \alpha_{L,r}\mathbf{v} + \beta_{L,r}\mathbf{z} + \gamma_{L,r}$ and $h_{U,r}(\mathbf{v}, \mathbf{z}) = \alpha_{U,r}\mathbf{v} + \beta_{U,r}\mathbf{z} + \gamma_{U,r}$ for the two 2D nonlinear functions $\sigma(\mathbf{v})\mathbf{z}$ and $\sigma(\mathbf{v})\tanh(\mathbf{z})$ given the range of their inputs, namely, $\mathbf{v} \in [\mathbf{l}^v, \mathbf{u}^v]$ and $\mathbf{z} \in [\mathbf{l}^z, \mathbf{u}^z], \forall r \in [s]$.

As we will exemplify the procedures for one single neuron, the subscript r is omitted hereafter. We formulate the problem of finding 2D bounding planes as minimizing the volumes between upper-bounding/ lower-bounding planes and \mathbf{v} - \mathbf{z} plane (See Figure 3 in the main text.) for the given \mathbf{v} , \mathbf{z} intervals, under the constraints that the planes should be larger/ smaller than the nonlinear planes. Specifically, the two constrained optimization problems are formulated as

$$\begin{aligned} & \underset{\alpha_L, \beta_L, \gamma_L}{\operatorname{argmax}} && \iint_A (\alpha_L \mathbf{v} + \beta_L \mathbf{z} + \gamma_L) d\mathbf{v} d\mathbf{z} \\ & \text{subject to} && \alpha_L \mathbf{v} + \beta_L \mathbf{z} + \gamma_L \leq f(\mathbf{v}, \mathbf{z}), \forall (\mathbf{v}, \mathbf{z}) \in A \end{aligned} \quad (25)$$

and

$$\begin{aligned} & \underset{\alpha_U, \beta_U, \gamma_U}{\operatorname{argmin}} && \iint_A (\alpha_U \mathbf{v} + \beta_U \mathbf{z} + \gamma_U) d\mathbf{v} d\mathbf{z} \\ & \text{subject to} && \alpha_U \mathbf{v} + \beta_U \mathbf{z} + \gamma_U \geq f(\mathbf{v}, \mathbf{z}), \forall (\mathbf{v}, \mathbf{z}) \in A \end{aligned} \quad (26)$$

where $f(\mathbf{v}, \mathbf{z})$ is $\sigma(\mathbf{v})\mathbf{z}$ or $\sigma(\mathbf{v})\tanh(\mathbf{z})$, $A = \{(\mathbf{v}, \mathbf{z}) \mid \mathbf{v} \in [\mathbf{l}^v, \mathbf{u}^v], \mathbf{z} \in [\mathbf{l}^z, \mathbf{u}^z]\}$.

To solve Problem (25), we first construct a differentiable indicator function $I(\alpha_L, \beta_L, \gamma_L)$ such that

$$I(\alpha_L, \beta_L, \gamma_L) \begin{cases} \leq 0 & \text{if } h_L(\mathbf{v}, \mathbf{z}) \leq f(\mathbf{v}, \mathbf{z}), \forall (\mathbf{v}, \mathbf{z}) \in A; \\ > 0 & \text{otherwise.} \end{cases}$$

Then we define loss function as

$$\operatorname{loss}(\alpha_L, \beta_L, \gamma_L) = \begin{cases} -\iint_A (\alpha_L \mathbf{v} + \beta_L \mathbf{z} + \gamma_L) d\mathbf{v} d\mathbf{z} & \text{if } I(\alpha_L, \beta_L, \gamma_L) \leq 0; \\ -\theta * \iint_A (\alpha_L \mathbf{v} + \beta_L \mathbf{z} + \gamma_L) d\mathbf{v} d\mathbf{z} + I(\alpha_L, \beta_L, \gamma_L) & \text{otherwise,} \end{cases}$$

where θ is a small constant and we usually set it to 0.1. Then we can minimize the loss function using a standard gradient descent algorithm. We record the result of every gradient descend step and choose the $(\alpha_L, \beta_L, \gamma_L)$ with the maximum value of $\iint_A (\alpha_L \mathbf{v} + \beta_L \mathbf{z} + \gamma_L) d\mathbf{v} d\mathbf{z}$ and the indicator function value $I(\alpha_L, \beta_L, \gamma_L) \leq 0$ as the lower bounding plane. The process of solving Problem (26) is similar.

The key step in the above process is to construct the indicator function $I(\alpha_L, \beta_L, \gamma_L)$ for the functions $f(\mathbf{v}, \mathbf{z})$ and $h_L(\mathbf{v}, \mathbf{z}) = \alpha_L \mathbf{v} + \beta_L \mathbf{z} + \gamma_L$ in the area A . Here we use $f(\mathbf{v}, \mathbf{z}) = \sigma(\mathbf{v})\mathbf{z}$ (case $f(\mathbf{v}, \mathbf{z}) = \sigma(\mathbf{v})\tanh(\mathbf{z})$ can be handled similarly) as an example to elaborate on how to construct the indicator function. First we observe that

$$h_L(\mathbf{v}, \mathbf{z}) \leq \sigma(\mathbf{v})\mathbf{z}, \forall (\mathbf{v}, \mathbf{z}) \in A \Leftrightarrow \begin{cases} \alpha_L \mathbf{v} + \beta_L \mathbf{l}^z + \gamma_L \leq \mathbf{l}^z \sigma(\mathbf{v}), & \forall \mathbf{v} \in [\mathbf{l}^v, \mathbf{u}^v] \\ \alpha_L \mathbf{v} + \beta_L \mathbf{u}^z + \gamma_L \leq \mathbf{u}^z \sigma(\mathbf{v}), & \forall \mathbf{v} \in [\mathbf{l}^v, \mathbf{u}^v] \end{cases} \quad (27)$$

This greatly reduces the complexity to construct the indicator function $I(\alpha_L, \beta_L, \gamma_L)$ since the right hand side of Equation (27) are two inequalities of 1D functions. The indicator function I now can be constructed as

$$I(\alpha_L, \beta_L, \gamma_L) = \max(I_1(\alpha_L, \beta_L, \gamma_L), 0) + \max(I_2(\alpha_L, \beta_L, \gamma_L), 0) \quad (28)$$

where

$$\begin{aligned} I_1(\alpha_L, \beta_L, \gamma_L) & \begin{cases} \leq 0 & \text{if } \alpha_L \mathbf{v} + \beta_L \mathbf{l}^z + \gamma_L \leq \mathbf{l}^z \sigma(\mathbf{v}), \forall \mathbf{v} \in [\mathbf{l}^v, \mathbf{u}^v]; \\ > 0 & \text{otherwise;} \end{cases} \\ I_2(\alpha_L, \beta_L, \gamma_L) & \begin{cases} \leq 0 & \text{if } \alpha_L \mathbf{v} + \beta_L \mathbf{u}^z + \gamma_L \leq \mathbf{u}^z \sigma(\mathbf{v}), \forall \mathbf{v} \in [\mathbf{l}^v, \mathbf{u}^v]; \\ > 0 & \text{otherwise.} \end{cases} \end{aligned}$$

The construction of I_1 and I_2 corresponds to constructing an indicator function $I_3(\alpha, \beta)$ for two 1D functions $h(\mathbf{v}) = \alpha\mathbf{v} + \beta$ and $s\sigma(\mathbf{v})$ in the interval $[\mathbf{l}^v, \mathbf{u}^v]$, where

$$\begin{aligned}\alpha &= \alpha_L \\ \beta &= \beta_L \mathbf{l}^z (\text{or } \mathbf{u}^z) + \gamma_L \\ s &= \mathbf{l}^z (\text{or } \mathbf{u}^z)\end{aligned}$$

and

$$I_3(\alpha, \beta) \begin{cases} \leq 0 & \text{if } \alpha\mathbf{v} + \beta \leq s\sigma(\mathbf{v}), \forall \mathbf{v} \in [\mathbf{l}^v, \mathbf{u}^v]; \\ > 0 & \text{otherwise.} \end{cases}$$

The construction of I_3 can be easily achieved using the linearity of $h(\mathbf{v})$ and piecewise convexity of $s\sigma(\mathbf{v})$. Then we can use I_3 to construct I_1 and I_2 . Finally $I(\alpha_L, \beta_L, \gamma_L)$ can be obtained according to Equation (28).

A.5 Algorithms for Long Short-term Memory Networks

Algorithm 2 Compute POPQORN robustness bound of an m -layer LSTM (true label: c , target label: i)

Input: an m -layer LSTM, input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}]$, p -norm, ϵ_0 .

Output: a robustness bound $\tilde{\epsilon}$.

- 1: Replace the last layer weights $\bar{\mathbf{w}} \leftarrow \mathbf{W}_{c,:}^{Fa} - \mathbf{W}_{i,:}^{Fa}$
- 2: Initialize $\epsilon \leftarrow \epsilon_0$
- 3: **while** ϵ has not achieved a desired accuracy and iteration limit has not reached **do**
- 4: **for** $v=1, \dots, m$ **do**
- 5: **for** $j=1, \dots, s$ **do**
- 6: $\mathbf{l}_j^{i(v)}, \mathbf{l}_j^{f(v)}, \mathbf{l}_j^{g(v)}, \mathbf{l}_j^{o(v)}, \mathbf{u}_j^{i(v)}, \mathbf{u}_j^{f(v)}, \mathbf{u}_j^{g(v)}, \mathbf{u}_j^{o(v)} \leftarrow$ Corollary A.3
- 7: $h_{L,j}^{fc(v)}(\mathbf{y}_j^{f(v)}, \mathbf{c}_j^{(v-1)}), h_{U,j}^{fc(v)}(\mathbf{y}_j^{f(v)}, \mathbf{c}_j^{(v-1)}) \leftarrow$ compute bounding planes for $\sigma(\mathbf{y}_j^{f(v)})\mathbf{c}_j^{(v-1)}$ given $\mathbf{l}_j^{f(v)}, \mathbf{u}_j^{f(v)}, \mathbf{l}_j^{c(v-1)}, \mathbf{u}_j^{c(v-1)}$
- 8: $h_{L,j}^{ig(v)}(\mathbf{y}_j^{i(v)}, \mathbf{y}_j^{g(v)}), h_{U,j}^{ig(v)}(\mathbf{y}_j^{i(v)}, \mathbf{y}_j^{g(v)}) \leftarrow$ compute bounding planes for $\sigma(\mathbf{y}_j^{i(v)})\tanh(\mathbf{y}_j^{g(v)})$ given $\mathbf{l}_j^{i(v)}, \mathbf{u}_j^{i(v)}, \mathbf{l}_j^{g(v)}, \mathbf{u}_j^{g(v)}$
- 9: $\mathbf{l}_j^{c(v)}, \mathbf{u}_j^{c(v)} \leftarrow$ Theorem A.4
- 10: $h_{L,j}^{oc(v)}(\mathbf{y}_j^{o(v)}, \mathbf{c}_j^{(v)}), h_{U,j}^{oc(v)}(\mathbf{y}_j^{o(v)}, \mathbf{c}_j^{(v)}) \leftarrow$ compute bounding planes for $\sigma(\mathbf{y}_j^{o(v)})\tanh(\mathbf{c}_j^{(v)})$ given $\mathbf{l}_j^{o(v)}, \mathbf{u}_j^{o(v)}, \mathbf{l}_j^{c(v)}, \mathbf{u}_j^{c(v)}$
- 11: **end for**
- 12: **end for**
- 13: $\gamma^U, \gamma^L \leftarrow$ Corollary A.2
- 14: **if** $\gamma^L \geq 0$ **then**
- 15: ϵ is a lower bound; increase ϵ using a binary search procedure
- 16: **else**
- 17: ϵ is not a lower bound; decrease ϵ using a binary search procedure
- 18: **end if**
- 19: $\tilde{\epsilon} \leftarrow \epsilon$
- 20: **end while**

As in the above, we provide a succinct algorithm summary to compute a robustness bound for an m -layer LSTM, below we also list a more detailed version of algorithms to compute lower and upper bounds for the final output $\mathbf{W}^{Fa}\mathbf{a}^{(m)} + \mathbf{b}^F$ of the LSTM and then provide a robustness bound for it. We first introduce an auxiliary algorithm that is needed in the computation.

Algorithm 3 Compute lower and upper bounds of the term $\mathbf{W}\mathbf{a}^{(v)} + \mathbf{b}$

Input: layer index v (range from 0 to m), weight matrix \mathbf{W} and bias vector \mathbf{b} , bounding planes $[h_{L/U}^{oc(1)}, \dots, h_{L/U}^{oc(v)}]$, $[h_{L/U}^{fc(1)}, \dots, h_{L/U}^{fc(v)}]$ and $[h_{L/U}^{ig(1)}, \dots, h_{L/U}^{ig(v)}]$, an m -layer LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(v)}]$, p -norm, ϵ .

Output: lower bound $(\mathbf{W}\mathbf{a}^{(v)} + \mathbf{b})_{L,j}$ and upper bound $(\mathbf{W}\mathbf{a}^{(v)} + \mathbf{b})_{U,j}$, $j = 1, \dots, t$. (Here t is the number of rows in W .)

Implementation: As is shortly pointed out in Corollary A.3, this algorithm can be adapted from Theorem A.3 and Corollary A.2. Theorem A.3 is provided to compute the lower and upper bounds of the final output of an m -layer LSTM, which is $\mathbf{W}^{Fa}\mathbf{a}^{(m)} + \mathbf{b}^F$. Therefore, we only need to replace the output mapping \mathbf{W}^{Fa} and \mathbf{b}^F in Theorem A.3 by the input weight matrix \mathbf{W} and bias vector \mathbf{b} , respectively, to evaluate the bounds of $\mathbf{W}\mathbf{a}^{(m)} + \mathbf{b}$. Although Theorem A.3 is stated to compute bounds for the final layer (m -th layer) of the LSTM, it is also straightforward to generalize Theorem A.3 to compute bounds for an arbitrary layer v . Note that when $v = 0$, both the lower bound and upper bound of $\mathbf{W}\mathbf{a}^{(0)} + \mathbf{b}$ is just $\mathbf{W}\mathbf{a}^{(0)} + \mathbf{b}$ itself.

It may seem like one can directly use Algorithm 3 to compute lower and upper bounds of the final output $\mathbf{W}^{Fa}\mathbf{a}^{(m)} + \mathbf{b}^F$ of an m -layer LSTM, but this is in fact not feasible as Algorithm 3 requires the bounding planes $h_{L/U}^{oc}, h_{L/U}^{fc}, h_{L/U}^{ig}$, which can only be computed after we know the lower and upper bounds of the gate pre-activations $\mathbf{y}^i, \mathbf{y}^f, \mathbf{y}^g, \mathbf{y}^o$ and lower and upper bounds of the cell state \mathbf{c} . Therefore, we need to introduce two more auxiliary algorithms below to compute bounds for $\mathbf{y}^i, \mathbf{y}^f, \mathbf{y}^g, \mathbf{y}^o$ and \mathbf{c} . Note that Algorithm 3 will be used in Algorithm 4. That is why we adapt Theorem A.3 and Corollary A.2 to construct this more general Algorithm 3.

Algorithm 4 Compute pre-activation bounds for the 4 gates $\mathbf{y}^{i(v)}$, $\mathbf{y}^{f(v)}$, $\mathbf{y}^{g(v)}$ and $\mathbf{y}^{o(v)}$

Input: layer index v (range from 1 to m), bounding planes $[h_{L/U}^{oc(1)}, \dots, h_{L/U}^{oc(v-1)}]$, $[h_{L/U}^{fc(1)}, \dots, h_{L/U}^{fc(v-1)}]$ and $[h_{L/U}^{ig(1)}, \dots, h_{L/U}^{ig(v-1)}]$, an m -layer LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(v)}]$, p -norm, ϵ .

Output: lower bounds $\mathbf{l}_j^{i(v)}$, $\mathbf{l}_j^{f(v)}$, $\mathbf{l}_j^{g(v)}$, $\mathbf{l}_j^{o(v)}$ and upper bounds $\mathbf{u}_j^{i(v)}$, $\mathbf{u}_j^{f(v)}$, $\mathbf{u}_j^{g(v)}$, $\mathbf{u}_j^{o(v)}$, $j = 1, \dots, s$.

Implementation: As stated in Corollary A.3, pre-activation bounds of the 4 gates can be computed using Algorithm 3 (Inputs: layer index $v-1$, weight matrix $\mathbf{W}^{ia}/\mathbf{W}^{fa}/\mathbf{W}^{ga}/\mathbf{W}^{oa}$, bias vector $\mathbf{b}^i/\mathbf{b}^f/\mathbf{b}^g/\mathbf{b}^o$, bounding planes $[h_{L/U}^{oc(1)}, \dots, h_{L/U}^{oc(v-1)}]$, $[h_{L/U}^{fc(1)}, \dots, h_{L/U}^{fc(v-1)}]$ and $[h_{L/U}^{ig(1)}, \dots, h_{L/U}^{ig(v-1)}]$, the LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(v)}]$, p -norm, ϵ). Note that when $v = 1$, the upper and lower bounds can be directly computed according to the definitions of the 4 gates without using the bounding planes $h_{L/U}^{oc}$, $h_{L/U}^{fc}$, $h_{L/U}^{ig}$.

Algorithm 5 Compute lower and upper bounds for the cell state $\mathbf{c}^{(v)}$

Input: layer index v (range from 0 to m), bounding planes $[h_{L/U}^{oc(1)}, \dots, h_{L/U}^{oc(v-1)}]$, $[h_{L/U}^{fc(1)}, \dots, h_{L/U}^{fc(v)}]$ and $[h_{L/U}^{ig(1)}, \dots, h_{L/U}^{ig(v)}]$, an m -layer LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(v)}]$, p -norm, ϵ .

Output: lower bound $\mathbf{l}_j^{c(v)}$ and upper bound $\mathbf{u}_j^{c(v)}$, $j = 1, \dots, s$.

Implementation: Implementation details are stated in Theorem A.4. Note that when $v = 0$, both the lower bound and upper bound of $\mathbf{c}^{(0)}$ is just $\mathbf{c}^{(0)}$ itself. No bounding plane is needed in this case. And when $v = 1$, $h_{L/U}^{oc}$ is actually not needed either.

Now we have got all the necessary auxiliary algorithms to compute lower and upper bounds of the final output $\mathbf{W}^{Fa}\mathbf{a}^{(m)} + \mathbf{b}^F$ of an m -layer LSTM. Algorithm 3, Algorithm 4 and Algorithm 5 will be used iteratively to compute bounds for the final output of an LSTM. The resulting algorithm is stated below.

Algorithm 6 Compute lower and upper bounds for the final output $\mathbf{W}^{Fa}\mathbf{a}^{(m)} + \mathbf{b}^F$ of an m -layer LSTM

Input: an m -layer LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}]$, p -norm, ϵ .

Output: lower bound $(\mathbf{W}^{Fa}\mathbf{a}^{(m)} + \mathbf{b}^F)_{L,j}$ and upper bounds $(\mathbf{W}^{Fa}\mathbf{a}^{(m)} + \mathbf{b}^F)_{U,j}$, $j = 1, \dots, t$.

- 1: **for** $v=1, \dots, m$ **do**
- 2: **for** $j=1, \dots, s$ **do**
- 3: $\mathbf{l}_j^{i(v)}$, $\mathbf{l}_j^{f(v)}$, $\mathbf{l}_j^{g(v)}$, $\mathbf{l}_j^{o(v)}$, $\mathbf{u}_j^{i(v)}$, $\mathbf{u}_j^{f(v)}$, $\mathbf{u}_j^{g(v)}$, $\mathbf{u}_j^{o(v)} \leftarrow$ Algorithm 4 (Inputs: layer index v , bounding planes $[h_{L/U}^{oc(1)}, \dots, h_{L/U}^{oc(v-1)}]$, $[h_{L/U}^{fc(1)}, \dots, h_{L/U}^{fc(v-1)}]$ and $[h_{L/U}^{ig(1)}, \dots, h_{L/U}^{ig(v-1)}]$, the LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(v)}]$, p -norm, ϵ)
- 4: $h_{L,j}^{fc(v)}(\mathbf{y}_j^{f(v)}, \mathbf{c}_j^{(v-1)})$, $h_{U,j}^{fc(v)}(\mathbf{y}_j^{f(v)}, \mathbf{c}_j^{(v-1)}) \leftarrow$ compute bounding planes for $\sigma(\mathbf{y}_j^{f(v)})\mathbf{c}_j^{(v-1)}$ given $\mathbf{l}_j^{f(v)}$, $\mathbf{u}_j^{f(v)}$, $\mathbf{l}_j^{c(v-1)}$, $\mathbf{u}_j^{c(v-1)}$ (Note $\mathbf{l}_j^{c(0)} = \mathbf{u}_j^{c(0)} = \mathbf{c}_j^{(0)}$)
- 5: $h_{L,j}^{ig(v)}(\mathbf{y}_j^{i(v)}, \mathbf{y}_j^{g(v)})$, $h_{U,j}^{ig(v)}(\mathbf{y}_j^{i(v)}, \mathbf{y}_j^{g(v)}) \leftarrow$ compute bounding planes for $\sigma(\mathbf{y}_j^{i(v)})\tanh(\mathbf{y}_j^{g(v)})$ given $\mathbf{l}_j^{i(v)}$, $\mathbf{u}_j^{i(v)}$, $\mathbf{l}_j^{g(v)}$, $\mathbf{u}_j^{g(v)}$
- 6: $\mathbf{l}_j^{c(v)}$, $\mathbf{u}_j^{c(v)} \leftarrow$ Algorithm 5 (Inputs: layer index v , bounding planes $[h_{L/U}^{oc(1)}, \dots, h_{L/U}^{oc(v-1)}]$, $[h_{L/U}^{fc(1)}, \dots, h_{L/U}^{fc(v)}]$ and $[h_{L/U}^{ig(1)}, \dots, h_{L/U}^{ig(v)}]$, the LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(v)}]$, p -norm, ϵ)
- 7: $h_{L,j}^{oc(v)}(\mathbf{y}_j^{o(v)}, \mathbf{c}_j^{(v)})$, $h_{U,j}^{oc(v)}(\mathbf{y}_j^{o(v)}, \mathbf{c}_j^{(v)}) \leftarrow$ compute bounding planes for $\sigma(\mathbf{y}_j^{o(v)})\tanh(\mathbf{c}_j^{(v)})$ given $\mathbf{l}_j^{o(v)}$, $\mathbf{u}_j^{o(v)}$, $\mathbf{l}_j^{c(v)}$, $\mathbf{u}_j^{c(v)}$
- 8: **end for**
- 9: **end for**
- 10: $(\mathbf{W}^{Fa}\mathbf{a}^{(m)} + \mathbf{b}^F)_{L,j}$, $(\mathbf{W}^{Fa}\mathbf{a}^{(m)} + \mathbf{b}^F)_{U,j}$, $j = 1, \dots, t \leftarrow$ Algorithm 3 (Inputs: layer index m , weight matrix \mathbf{W}^{Fa} and bias vector \mathbf{b}^F , bounding planes $[h_{L/U}^{oc(1)}, \dots, h_{L/U}^{oc(v)}]$, $[h_{L/U}^{fc(1)}, \dots, h_{L/U}^{fc(v)}]$ and $[h_{L/U}^{ig(1)}, \dots, h_{L/U}^{ig(v)}]$, the LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}]$, p -norm, ϵ)

Finally, we can use Algorithm 6 to get a robustness bound for an LSTM through a binary search procedure. Assume the true class of the sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}]$ is c . The algorithm to certify the largest possible lower bound of targeted attacks (target class be i) is stated below.

Algorithm 7 Compute POPQORN robustness bound of an m -layer LSTM

Input: an m -layer LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}]$, p -norm, ϵ_0 .

Output: a robustness bound $\tilde{\epsilon}$.

- 1: Replace the last layer weights $\bar{\mathbf{w}} \leftarrow \mathbf{W}_{c,:}^{Fa} - \mathbf{W}_{i,:}^{Fa}$
- 2: Initialize $\epsilon \leftarrow \epsilon_0$
- 3: **while** ϵ has not achieved a desired accuracy and iteration limit has not reached **do**
- 4: $\gamma^L, \gamma^U \leftarrow$ Algorithm 6 (Inputs: the LSTM, LSTM input sequence $[\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}]$, p -norm, ϵ)
- 5: **if** $\gamma^L \geq 0$ **then**
- 6: ϵ is a lower bound; increase ϵ using a binary search procedure
- 7: **else**
- 8: ϵ is not a lower bound; decrease ϵ using a binary search procedure
- 9: **end if**
- 10: $\tilde{\epsilon} \leftarrow \epsilon$
- 11: **end while**

A.6 Long Short-term Memory Network (1D bounding lines)

Theorem A. 5 (Explicit output bound of long short-term memory network F with 1D bounding technique) Given an m -layer long short-term memory network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$, we can derive two explicit functions $F_j^L : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ and $F_j^U : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ such that $\forall j \in [t]$ and $\forall \mathbf{X} \in \mathbb{R}^{n \times m}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, the inequity $F_j^L(\mathbf{X}) \leq F_j(\mathbf{X}) \leq F_j^U(\mathbf{X})$ holds true. We can obtain them through following steps: Starting from $k = m, \forall r \in s$ we

1. define unified slopes and intercepts;

$$\lambda_{j,r}^{(k)} = \begin{cases} \alpha_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \alpha_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} \quad \omega_{j,r}^{(k)} = \begin{cases} \alpha_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \alpha_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases}$$

$$\Delta_{j,r}^{(k)} = \begin{cases} \beta_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \beta_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} \quad \Theta_{j,r}^{(k)} = \begin{cases} \beta_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \beta_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases}$$

where $\tilde{\mathbf{W}}_{U,j,r}^{a(m+1)} = \tilde{\mathbf{W}}_{L,j,r}^{a(m+1)} = \mathbf{W}_{j,r}^{Fa}$,

2. collect coefficients in front of pre-activation, and constants;

$$\Lambda_{\lambda,j,:}^{(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \lambda_{j,:}^{(k)} & \text{if } k \in [m-1]; \end{cases} \quad \Omega_{\omega,j,:}^{(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \omega_{j,:}^{(k)} & \text{if } k \in [m-1]; \end{cases}$$

$$\Lambda_{\Delta,j,:}^{(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Delta_{j,:}^{(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \Delta_{j,:}^{(k)} & \text{if } k \in [m-1]; \end{cases} \quad \Omega_{\Theta,j,:}^{(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Theta_{j,:}^{(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \Theta_{j,:}^{(k)} & \text{if } k \in [m-1]; \end{cases}$$

3. define equivalent weight matrices and biases;

$$\tilde{\mathbf{W}}_{U,j,q_k}^{x(k)} = \Lambda_{\lambda,j,:}^{(k)} \mathbf{W}_{:,q_k}^{ox}, \quad \tilde{\mathbf{W}}_{U,j,r}^{a(k)} = \Lambda_{\lambda,j,:}^{(k)} \mathbf{W}_{:,r}^{oa}, \quad \tilde{\mathbf{b}}_{U,j}^{(k)} = \Lambda_{\lambda,j,:}^{(k)} \mathbf{b}^o + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{(k)};$$

$$\tilde{\mathbf{W}}_{L,j,q_k}^{x(k)} = \Omega_{\omega,j,:}^{(k)} \mathbf{W}_{:,q_k}^{ox}, \quad \tilde{\mathbf{W}}_{L,j,r}^{a(k)} = \Omega_{\omega,j,:}^{(k)} \mathbf{W}_{:,r}^{oa}, \quad \tilde{\mathbf{b}}_{L,j}^{(k)} = \Omega_{\omega,j,:}^{(k)} \mathbf{b}^o + \sum_{i=1}^s \Omega_{\Theta,j,i}^{(k)};$$

4. After looping steps 1 to 3 from $k = m$ to $k = 1$, the bounds are given by

$$F_j^U(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{U,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{U,j}^{(k)} + \mathbf{b}_j^F,$$

$$F_j^L(\mathbf{X}) = \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{L,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{L,j}^{(k)} + \mathbf{b}_j^F.$$

Proof 4 Given an m -layer long short-term memory network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$ with pre-activation bounds $\mathbf{l}^{o(k)}, \mathbf{u}^{o(k)}, \mathbf{l}^{c(k)}, \mathbf{u}^{c(k)}$ for $\mathbf{X} \in \mathbb{R}^{n \times m}$, where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, let the pre-activation inputs for the j -th neuron at the output

layer be $F_j(\mathbf{X}) = \mathbf{W}_{j,:}^{Fa} \mathbf{a}^{(m)} + \mathbf{b}_j^F$. The j -th output of the recurrent neural network is the following:

$$\begin{aligned}
F_j(\mathbf{X}) &= \mathbf{W}_{j,:}^{Fa} \mathbf{a}^{(m)} + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} \mathbf{a}_i^{(m)} + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} [\mathbf{o}_i^{(m)} \tanh(\mathbf{c}_i^{(m)})] + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} [\sigma(\mathbf{y}_i^{o(m)}) \tanh(\mathbf{c}_i^{(m)})] + \mathbf{b}_j^F, \\
&= \sum_{\mathbf{W}_{j,i}^{Fa} \geq 0} \mathbf{W}_{j,i}^{Fa} [\sigma(\mathbf{y}_i^{o(m)}) \tanh(\mathbf{c}_i^{(m)})] + \sum_{\mathbf{W}_{j,i}^{Fa} < 0} \mathbf{W}_{j,i}^{Fa} [\sigma(\mathbf{y}_i^{o(m)}) \tanh(\mathbf{c}_i^{(m)})] + \mathbf{b}_j^F.
\end{aligned}$$

Assume $\mathbf{c}_i^{(m)}$ is bounded by $\mathbf{l}^{(m)}, \mathbf{u}^{(m)}$, $\tanh(\mathbf{c}_i^{(m)})$ can be bounded by $\mathbf{c}_{L,i}^{(m)}$ and $\mathbf{c}_{U,i}^{(m)}$, such that $\mathbf{c}_{L,i}^{(m)} \leq \tanh(\mathbf{c}_i^{(m)}) \leq \mathbf{c}_{U,i}^{(m)}$. We also assume the positive activation function $\sigma(\mathbf{y})$ is bounded by two linear functions $h_{L,i}^{(m)}(\mathbf{y})$, $h_{U,i}^{(m)}(\mathbf{y})$, which yield $H_{L,i}^{(m)}(\mathbf{y}) \leq \sigma(\mathbf{y}_i^{o(m)}) \tanh(\mathbf{c}_i^{(m)}) \leq H_{U,i}^{(m)}(\mathbf{y})$, where

$$\begin{aligned}
H_{L,i}^{(m)}(\mathbf{y}) &= \mathbb{1}_{\mathbf{c}_{L,i}^{(m)} \leq 0} h_{U,i}^{(m)}(\mathbf{y}_i^{o(m)}) \mathbf{c}_{L,i}^{(m)} + \mathbb{1}_{\mathbf{c}_{L,i}^{(m)} > 0} h_{L,i}^{(m)}(\mathbf{y}_i^{o(m)}) \mathbf{c}_{L,i}^{(m)}; \\
H_{U,i}^{(m)}(\mathbf{y}) &= \mathbb{1}_{\mathbf{c}_{U,i}^{(m)} \geq 0} h_{U,i}^{(m)}(\mathbf{y}_i^{o(m)}) \mathbf{c}_{U,i}^{(m)} + \mathbb{1}_{\mathbf{c}_{U,i}^{(m)} < 0} h_{L,i}^{(m)}(\mathbf{y}_i^{o(m)}) \mathbf{c}_{U,i}^{(m)}.
\end{aligned}$$

Upper bound. We can then obtain

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{\mathbf{W}_{j,i}^{Fa} \geq 0} \mathbf{W}_{j,i}^{Fa} H_{U,i}^{(m)}(\mathbf{y}_i^{o(m)}) + \sum_{\mathbf{W}_{j,i}^{Fa} < 0} \mathbf{W}_{j,i}^{Fa} H_{L,i}^{(m)}(\mathbf{y}_i^{o(m)}) + \mathbf{b}_j^F, \\
&= \sum_{\mathbf{W}_{j,i}^{Fa} \geq 0} \mathbf{W}_{j,i}^{Fa} (\alpha_{U,i}^{(m)} \mathbf{y}_i^{o(m)} + \beta_{U,i}^{(m)}) + \sum_{\mathbf{W}_{j,i}^{Fa} < 0} \mathbf{W}_{j,i}^{Fa} (\alpha_{L,i}^{(m)} \mathbf{y}_i^{o(m)} + \beta_{L,i}^{(m)}) + \mathbf{b}_j^F.
\end{aligned}$$

We then define $\lambda_{j,i}^{(m)}$ and $\Delta_{j,i}^{(m)}$ in the parentheses:

$$\lambda_{j,i}^{(m)} = \begin{cases} \alpha_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \alpha_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} \quad \Delta_{j,i}^{(m)} = \begin{cases} \beta_{U,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \beta_{L,i}^{(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}$$

and obtain

$$F_j(\mathbf{X}) \leq \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} (\lambda_{j,i}^{(m)} \mathbf{y}_i^{o(m)} + \Delta_{j,i}^{(m)}) + \mathbf{b}_j^F.$$

We further let $\Lambda_{\lambda,j,i}^{(m)} := \mathbf{W}_{j,i}^{Fa} \lambda_{j,i}^{(m)}$, $\Lambda_{\Delta,j,i}^{(m)} := \mathbf{W}_{j,i}^{Fa} \Delta_{j,i}^{(m)}$, and we have

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{(m)} \mathbf{y}_i^{o(m)} + \Lambda_{\Delta,j,i}^{(m)}) + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s [\Lambda_{\lambda,j,i}^{(m)} (\mathbf{W}_{i,:}^{ox} \mathbf{x}^{(m)} + \mathbf{W}_{i,:}^{oa} \mathbf{a}^{(m-1)} + \mathbf{b}_i^o) + \Lambda_{\Delta,j,i}^{(m)}] + \mathbf{b}_j^F, \\
&= \sum_{q_m=1}^n (\sum_{i=1}^s \Lambda_{\lambda,j,i}^{(m)} \mathbf{W}_{i,q_m}^{ox}) \mathbf{x}_{q_m}^{(m)} + \sum_{r=1}^s (\sum_{i=1}^s \Lambda_{\lambda,j,i}^{(m)} \mathbf{W}_{i,r}^{oa}) \mathbf{a}_r^{(m-1)} + \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{(m)} \mathbf{b}_i^o + \Lambda_{\Delta,j,i}^{(m)}) + \mathbf{b}_j^F, \\
&= \sum_{q_m=1}^n (\Lambda_{\lambda,j,:}^{(m)} \mathbf{W}_{:,q_m}^{ox}) \mathbf{x}_{q_m}^{(m)} + \sum_{r=1}^s (\Lambda_{\lambda,j,:}^{(m)} \mathbf{W}_{:,r}^{oa}) \mathbf{a}_r^{(m-1)} + \Lambda_{\lambda,j,:}^{(m)} \mathbf{b}^o + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{(m)} + \mathbf{b}_j^F.
\end{aligned}$$

We combine coefficients above into equivalent weights $\tilde{\mathbf{W}}_{U,j,q_m}^{x(m)}$, $\tilde{\mathbf{W}}_{U,j,r}^{a(m)}$ and bias $\tilde{\mathbf{b}}_{U,j}^{(m)}$ as defined by

$$\tilde{\mathbf{W}}_{U,j,q_m}^{x(m)} = \Lambda_{\lambda,j,:}^{(m)} \mathbf{W}_{:,q_m}^{ox}, \quad \tilde{\mathbf{W}}_{U,j,r}^{a(m)} = \Lambda_{\lambda,j,:}^{(m)} \mathbf{W}_{:,r}^{oa}, \quad \tilde{\mathbf{b}}_{U,j}^{(m)} = \Lambda_{\lambda,j,:}^{(m)} \mathbf{b}^o + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{(m)},$$

and obtain

$$\begin{aligned} F_j(\mathbf{X}) &\leq F_j^{U,m}(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{W}}_{U,j,:}^{a(m)} \mathbf{a}^{(m-1)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F, \\ &= \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F + \tilde{\mathbf{W}}_{U,j,:}^{a(m)} [\mathbf{o}^{(m-1)} \tanh(\mathbf{c}^{(m-1)})] \\ &= \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F \\ &+ \sum_{\tilde{\mathbf{W}}_{U,j,i}^{a(m)} \geq 0} \tilde{\mathbf{W}}_{U,j,i}^{a(m)} [\sigma(\mathbf{y}_i^{o(m-1)}) \tanh(\mathbf{c}_i^{(m-1)})] + \sum_{\tilde{\mathbf{W}}_{U,j,i}^{a(m)} < 0} \tilde{\mathbf{W}}_{U,j,i}^{a(m)} [\sigma(\mathbf{y}_i^{o(m-1)}) \tanh(\mathbf{c}_i^{(m-1)})]. \end{aligned}$$

Assume $\mathbf{c}_i^{(m-1)}$ is bounded by $\mathbf{I}^{c(m-1)}, \mathbf{u}^{c(m-1)}$, $\tanh(\mathbf{c}_i^{(m-1)})$ can be bounded by $\mathbf{c}_{U,i}^{(m-1)}$ and $\mathbf{c}_{L,i}^{(m-1)}$, such that $\mathbf{c}_{L,i}^{(m-1)} \leq \tanh(\mathbf{c}_i^{(m-1)}) \leq \mathbf{c}_{U,i}^{(m-1)}$. We also assume the positive activation function $\sigma(\mathbf{y})$ is bounded by two linear functions $h_{L,i}^{(m-1)}(\mathbf{y}), h_{U,i}^{(m-1)}(\mathbf{y})$, which yield $H_{L,i}^{(m-1)}(\mathbf{y}) \leq \sigma(\mathbf{y}_i^{o(m-1)}) \tanh(\mathbf{c}_i^{(m-1)}) \leq H_{U,i}^{(m-1)}(\mathbf{y})$, where

$$\begin{aligned} H_{L,i}^{(m-1)}(\mathbf{y}) &= \mathbb{1}_{\mathbf{c}_{L,i}^{(m-1)} \leq 0} h_{U,i}^{(m-1)}(\mathbf{y}_i^{o(m-1)}) \mathbf{c}_{L,i}^{(m-1)} + \mathbb{1}_{\mathbf{c}_{L,i}^{(m-1)} > 0} h_{L,i}^{(m-1)}(\mathbf{y}_i^{o(m-1)}) \mathbf{c}_{L,i}^{(m-1)}; \\ H_{U,i}^{(m-1)}(\mathbf{y}) &= \mathbb{1}_{\mathbf{c}_{U,i}^{(m-1)} \geq 0} h_{U,i}^{(m-1)}(\mathbf{y}_i^{o(m-1)}) \mathbf{c}_{U,i}^{(m-1)} + \mathbb{1}_{\mathbf{c}_{U,i}^{(m-1)} < 0} h_{L,i}^{(m-1)}(\mathbf{y}_i^{o(m-1)}) \mathbf{c}_{U,i}^{(m-1)}. \end{aligned}$$

We can then obtain

$$\begin{aligned} F_j^{U,m}(\mathbf{X}) &\leq \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F \\ &+ \sum_{\tilde{\mathbf{W}}_{U,j,i}^{a(m)} \geq 0} \tilde{\mathbf{W}}_{U,j,i}^{a(m)} (\alpha_{U,i}^{(m-1)} \mathbf{y}_i^{o(m-1)} + \beta_{U,i}^{(m-1)}) + \sum_{\tilde{\mathbf{W}}_{U,j,i}^{a(m)} < 0} \tilde{\mathbf{W}}_{U,j,i}^{a(m)} (\alpha_{L,i}^{(m-1)} \mathbf{y}_i^{o(m-1)} + \beta_{L,i}^{(m-1)}). \end{aligned}$$

We then define $\lambda_{j,i}^{(m-1)}$ and $\Delta_{j,i}^{(m-1)}$ in the parentheses:

$$\lambda_{j,i}^{(m-1)} = \begin{cases} \alpha_{U,i}^{(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \geq 0; \\ \alpha_{L,i}^{(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} < 0; \end{cases} \quad \Delta_{j,i}^{(m-1)} = \begin{cases} \beta_{U,i}^{(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \geq 0; \\ \beta_{L,i}^{(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,i}^{a(m)} < 0; \end{cases}$$

and obtain

$$F_j^{U,m}(\mathbf{X}) \leq \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F + \sum_{i=1}^s \tilde{\mathbf{W}}_{U,j,i}^{a(m)} (\lambda_{j,i}^{(m-1)} \mathbf{y}_i^{o(m-1)} + \Delta_{j,i}^{(m-1)}).$$

We further let $\Lambda_{\lambda,j,i}^{(m-1)} := \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \lambda_{j,i}^{(m-1)}$, $\Lambda_{\Delta,j,i}^{(m-1)} := \tilde{\mathbf{W}}_{U,j,i}^{a(m)} \Delta_{j,i}^{(m-1)}$ and we have

$$F_j^{U,m}(\mathbf{X}) \leq \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{(m-1)} \mathbf{y}_i^{o(m-1)} + \Lambda_{\Delta,j,i}^{(m-1)}) + \mathbf{b}_j^F.$$

Similarly, $\forall r \in s$ we can define variables as follows:

$$\tilde{\mathbf{W}}_{U,j,q_{m-1}}^{x(m-1)} = \Lambda_{\lambda,j,:}^{(m-1)} \mathbf{W}_{:,q_{m-1}}^{ox}; \quad \tilde{\mathbf{W}}_{U,j,r}^{a(m-1)} = \Lambda_{\lambda,j,:}^{(m-1)} \mathbf{W}_{:,r}^{oa}; \quad \tilde{\mathbf{b}}_{U,j}^{(m-1)} = \Lambda_{\lambda,j,:}^{(m-1)} \mathbf{b}^o + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{(m-1)}.$$

Then we can obtain

$$\begin{aligned} F_j^{U,m}(\mathbf{X}) &\leq F_j^{U,m-1}(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{W}}_{U,j,:}^{x(m-1)} \mathbf{x}^{(m-1)} + \tilde{\mathbf{W}}_{U,j,:}^{a(m-1)} \mathbf{a}^{(m-2)} \\ &+ \tilde{\mathbf{b}}_{U,j}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m-1)} + \mathbf{b}_j^F. \end{aligned}$$

Thus, we can repeat the above procedure iteratively until we obtain the final upper bound $F_j^{U,1}(\mathbf{X})$, where $F_j(\mathbf{X}) \leq F_j^{U,m}(\mathbf{X}) \leq \dots \leq F_j^{U,1}(\mathbf{X})$. We let $F_j^U(\mathbf{X})$ denote the final upper bound $F_j^{U,1}(\mathbf{X})$, and we have

$$F_j^U(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{U,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{U,j}^{(k)} + \mathbf{b}_j^F,$$

where

$$\tilde{\mathbf{W}}_{U,j,qk}^{x(k)} = \Lambda_{\lambda,j,:}^{(k)} \mathbf{W}_{:,qk}^{ox}, \quad \tilde{\mathbf{W}}_{U,j,r}^{a(k)} = \Lambda_{\lambda,j,:}^{(k)} \mathbf{W}_{:,r}^{oa}, \quad \tilde{\mathbf{b}}_{U,j}^{(k)} = \Lambda_{\lambda,j,:}^{(k)} \mathbf{b}^o + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{(k)},$$

and (\odot is the Hadamard product)

$$\Lambda_{\lambda,j,:}^{(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \lambda_{j,:}^{(k)} & \text{if } k \in [m-1]; \end{cases} \quad \Lambda_{\Delta,j,:}^{(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Delta_{j,:}^{(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \Delta_{j,:}^{(k)} & \text{if } k \in [m-1]; \end{cases}$$

and $\forall r \in s$,

$$\lambda_{j,r}^{(m)} = \begin{cases} \alpha_{U,r}^{(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \alpha_{L,r}^{(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} \quad \Delta_{j,r}^{(m)} = \begin{cases} \beta_{U,r}^{(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \beta_{L,r}^{(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases}$$

and $\forall r \in s, k \in [m-1]$,

$$\lambda_{j,r}^{(k)} = \begin{cases} \alpha_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \alpha_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} \quad \Delta_{j,r}^{(k)} = \begin{cases} \beta_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \beta_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0. \end{cases}$$

Lower bound. The above derivations of upper bound can be applied similarly to derive lower bounds of $F_j(\mathbf{X})$, and the only difference is now we need to use the left-hand side of inequities (rather than right-hand side when deriving upper bound) to bound the activated terms. Thus, following the same procedure in deriving the upper bounds, we can iteratively unwrap the activation functions and obtain a final lower bound $F_j^{L,1}(\mathbf{X})$, where $F_j(\mathbf{X}) \geq F_j^{L,m}(\mathbf{X}) \geq \dots \geq F_j^{L,1}(\mathbf{X})$. We let $F_j^L(\mathbf{X})$ denote the final lower bound $F_j^{L,1}(\mathbf{X})$, and we have

$$F_j^L(\mathbf{X}) = \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{L,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{L,j}^{(k)} + \mathbf{b}_j^F,$$

where

$$\tilde{\mathbf{W}}_{L,j,qk}^{x(k)} = \Omega_{\omega,j,:}^{(k)} \mathbf{W}_{:,qk}^{ox}, \quad \tilde{\mathbf{W}}_{L,j,r}^{a(k)} = \Omega_{\omega,j,:}^{(k)} \mathbf{W}_{:,r}^{oa}, \quad \tilde{\mathbf{b}}_{L,j}^{(k)} = \Omega_{\omega,j,:}^{(k)} \mathbf{b}^o + \sum_{i=1}^s \Omega_{\Theta,j,i}^{(k)},$$

and (\odot is the Hadamard product)

$$\Omega_{\omega,j,:}^{(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \omega_{j,:}^{(k)} & \text{if } k \in [m-1]; \end{cases} \quad \Omega_{\Theta,j,:}^{(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Theta_{j,:}^{(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \Theta_{j,:}^{(k)} & \text{if } k \in [m-1]; \end{cases}$$

and $\forall r \in s$,

$$\omega_{j,r}^{(m)} = \begin{cases} \alpha_{L,r}^{(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \alpha_{U,r}^{(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} \quad \Theta_{j,r}^{(m)} = \begin{cases} \beta_{L,r}^{(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \beta_{U,r}^{(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases}$$

and $\forall r \in s, k \in [m-1]$,

$$\omega_{j,r}^{(k)} = \begin{cases} \alpha_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \alpha_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \quad \Theta_{j,r}^{(k)} = \begin{cases} \beta_{L,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \beta_{U,r}^{(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0. \end{cases}$$

Corollary A. 4 (Closed-form global bounds) Given data $\mathbf{X} \in \mathbb{R}^{n \times m}$, l_p ball parameters $p \geq 1$ and $\epsilon \geq 0$. For an m -layer long short-term memory network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$, there exists two fixed values γ_j^L and γ_j^U such that $\forall \mathbf{X} \in \mathbb{R}^{n \times m}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, and $\forall j \in [t]$, $1/q = 1 - 1/p$, the inequality $\gamma_j^L \leq F_j(\mathbf{X}) \leq \gamma_j^U$ holds true, where

$$\begin{aligned}\gamma_j^U &= \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \epsilon \|\tilde{\mathbf{W}}_{U,j,:}^{x(k)}\|_q + \sum_{k=1}^m \tilde{\mathbf{W}}_{U,j,:}^{x(k)} \mathbf{x}_0^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{U,j}^{(k)} + \mathbf{b}_j^F, \\ \gamma_j^L &= \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} - \sum_{k=1}^m \epsilon \|\tilde{\mathbf{W}}_{L,j,:}^{x(k)}\|_q + \sum_{k=1}^m \tilde{\mathbf{W}}_{L,j,:}^{x(k)} \mathbf{x}_0^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{L,j}^{(k)} + \mathbf{b}_j^F.\end{aligned}$$

A.7 Gated Recurrent Unit Network

Notations In the derivation of gated recurrent unit network (GRU), we define the equations as follows:

$$\begin{aligned}
\text{Reset gate: } \mathbf{r}^{(k)} &= \sigma(\mathbf{y}^{r(k)}) = \sigma(\mathbf{W}^{rx} \mathbf{x}^{(k)} + \mathbf{W}^{ra} \mathbf{a}^{(k-1)} + \mathbf{b}^r); \\
\text{Update gate: } \mathbf{z}^{(k)} &= \sigma(\mathbf{y}^{z(k)}) = \sigma(\mathbf{W}^{zx} \mathbf{x}^{(k)} + \mathbf{W}^{za} \mathbf{a}^{(k-1)} + \mathbf{b}^z); \\
\text{New gate: } \mathbf{n}^{(k)} &= \tanh(\mathbf{y}^{n(k)}) = \tanh(\mathbf{W}^{nx} \mathbf{x}^{(k)} + \mathbf{b}^{nx} + \mathbf{r}^{(k)} \odot (\mathbf{W}^{na} \mathbf{a}^{(k-1)} + \mathbf{b}^{na})); \\
\text{Hidden state: } \mathbf{a}^{(k)} &= (1 - \mathbf{z}^{(k)}) \odot \mathbf{n}^{(k)} + \mathbf{z}^{(k)} \odot \mathbf{a}^{(k-1)}.
\end{aligned}$$

Theorem A. 6 (Explicit output bound of gated recurrent unit network F) *Given an m -layer gated recurrent unit network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$, we can derive two explicit functions $F_j^L : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ and $F_j^U : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$ such that $\forall j \in [t]$ and $\forall \mathbf{X} \in \mathbb{R}^{n \times m}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, the inequity $F_j^L(\mathbf{X}) \leq F_j(\mathbf{X}) \leq F_j^U(\mathbf{X})$ holds true. We can obtain them through following steps: Starting from $k = m$, $\forall r \in s$ we*

1. define unified slopes and intercepts of cross terms of the update gate and the new gate, the update gate and the hidden state;

$$\begin{aligned}
\lambda_{j,r}^{zn(k)} &= \begin{cases} \alpha_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \alpha_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \omega_{j,r}^{zn(k)} &= \begin{cases} \alpha_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \alpha_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\
\Delta_{j,r}^{zn(k)} &= \begin{cases} \beta_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \beta_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \Theta_{j,r}^{zn(k)} &= \begin{cases} \beta_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \beta_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\
\varphi_{j,r}^{zn(k)} &= \begin{cases} \gamma_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \gamma_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \psi_{j,r}^{zn(k)} &= \begin{cases} \gamma_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \gamma_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\
\lambda_{j,r}^{za(k)} &= \begin{cases} \alpha_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \alpha_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \omega_{j,r}^{za(k)} &= \begin{cases} \alpha_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \alpha_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\
\Delta_{j,r}^{za(k)} &= \begin{cases} \beta_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \beta_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \Theta_{j,r}^{za(k)} &= \begin{cases} \beta_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \beta_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\
\varphi_{j,r}^{za(k)} &= \begin{cases} \gamma_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \gamma_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \psi_{j,r}^{za(k)} &= \begin{cases} \gamma_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \gamma_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases}
\end{aligned}$$

where $\tilde{\mathbf{W}}_{U,j,r}^{a(m+1)} = \tilde{\mathbf{W}}_{L,j,r}^{a(m+1)} = \mathbf{W}_{j,r}^{Fa}$;

2. collect coefficients in front of the update gate *pre-activations*, the new gate *pre-activation*, the hidden state, and constants of the two couplings;

$$\begin{aligned}
\Lambda_{\lambda,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \lambda_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\omega,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \omega_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} \\
\Lambda_{\Delta,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Delta_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \Delta_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\Theta,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Theta_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \Theta_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} \\
\Lambda_{\varphi,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \varphi_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \varphi_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\psi,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \psi_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \psi_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} \\
\Lambda_{\lambda,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \lambda_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\omega,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \omega_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases} \\
\Lambda_{\Delta,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Delta_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \Delta_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\Theta,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Theta_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \Theta_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases}
\end{aligned}$$

$$\Lambda_{\varphi,j,:}^{za(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \varphi_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \varphi_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases} \quad \Omega_{\psi,j,:}^{za(k)} = \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \psi_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \psi_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases}$$

3. define unified slopes and intercepts of cross terms of the reset gate and the hidden state;

$$\lambda_{j,r}^{ra(k)} = \begin{cases} \alpha_{U,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} \geq 0; \\ \alpha_{L,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} < 0; \end{cases} \quad \omega_{j,r}^{ra(k)} = \begin{cases} \alpha_{L,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} \geq 0; \\ \alpha_{U,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} < 0; \end{cases}$$

$$\Delta_{j,r}^{ra(k)} = \begin{cases} \beta_{U,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} \geq 0; \\ \beta_{L,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} < 0; \end{cases} \quad \Theta_{j,r}^{ra(k)} = \begin{cases} \beta_{L,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} \geq 0; \\ \beta_{U,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} < 0; \end{cases}$$

$$\varphi_{j,r}^{ra(k)} = \begin{cases} \gamma_{U,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} \geq 0; \\ \gamma_{L,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} < 0; \end{cases} \quad \psi_{j,r}^{ra(k)} = \begin{cases} \gamma_{L,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} \geq 0; \\ \gamma_{U,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} < 0; \end{cases}$$

4. collect coefficients in front of the reset gate *pre-activation* and the hidden state, and constants;

$$\Lambda_{\lambda,j,:}^{ra(k)} = \Lambda_{\Delta,j,:}^{zn(k)} \odot \lambda_{j,:}^{ra(k)} \quad \text{if } k \in [m]; \quad \Omega_{\omega,j,:}^{ra(k)} = \Omega_{\Theta,j,:}^{zn(k)} \odot \omega_{j,:}^{ra(k)} \quad \text{if } k \in [m];$$

$$\Lambda_{\Delta,j,:}^{ra(k)} = \Lambda_{\Delta,j,:}^{zn(k)} \odot \Delta_{j,:}^{ra(k)} \quad \text{if } k \in [m]; \quad \Omega_{\Theta,j,:}^{ra(k)} = \Omega_{\Theta,j,:}^{zn(k)} \odot \Theta_{j,:}^{ra(k)} \quad \text{if } k \in [m];$$

$$\Lambda_{\varphi,j,:}^{ra(k)} = \Lambda_{\Delta,j,:}^{zn(k)} \odot \varphi_{j,:}^{ra(k)} \quad \text{if } k \in [m]; \quad \Omega_{\psi,j,:}^{ra(k)} = \Omega_{\Theta,j,:}^{zn(k)} \odot \psi_{j,:}^{ra(k)} \quad \text{if } k \in [m];$$

5. define equivalent weight matrices and biases.

$$\tilde{\mathbf{W}}_{U,j,qk}^x(k) = \Lambda_{\lambda,j,:}^{zn(k)} \mathbf{W}_{:,qk}^{zx} + \Lambda_{\lambda,j,:}^{za(k)} \mathbf{W}_{:,qk}^{zx} + \Lambda_{\Delta,j,:}^{zn(k)} \mathbf{W}_{:,qk}^{nx} + \Lambda_{\lambda,j,:}^{ra(k)} \mathbf{W}_{:,qk}^{rx},$$

$$\tilde{\mathbf{W}}_{U,j,r}^a(k) = \Lambda_{\lambda,j,:}^{zn(k)} \mathbf{W}_{:,r}^{za} + \Lambda_{\lambda,j,:}^{za(k)} \mathbf{W}_{:,r}^{za} + \Lambda_{\Delta,j,r}^{za(k)} + \Lambda_{\lambda,j,:}^{ra(k)} \mathbf{W}_{:,r}^{ra} + \Lambda_{\Delta,j,:}^{ra(k)} \mathbf{W}_{:,r}^{na},$$

$$\tilde{\mathbf{b}}_{U,j}^{(k)} = \Lambda_{\lambda,j,:}^{zn(k)} \mathbf{b}^z + \Lambda_{\lambda,j,:}^{za(k)} \mathbf{b}^z + \Lambda_{\Delta,j,:}^{zn(k)} \mathbf{b}^{nx} + \Lambda_{\lambda,j,:}^{ra(k)} \mathbf{b}^r + \Lambda_{\Delta,j,:}^{ra(k)} \mathbf{b}^{na} + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{zn(k)} + \Lambda_{\varphi,j,i}^{za(k)} + \Lambda_{\varphi,j,i}^{ra(k)}),$$

$$\tilde{\mathbf{W}}_{L,j,qk}^x(k) = \Omega_{\omega,j,:}^{zn(k)} \mathbf{W}_{:,qk}^{zx} + \Omega_{\omega,j,:}^{za(k)} \mathbf{W}_{:,qk}^{zx} + \Omega_{\Theta,j,:}^{zn(k)} \mathbf{W}_{:,qk}^{nx} + \Omega_{\omega,j,:}^{ra(k)} \mathbf{W}_{:,qk}^{rx},$$

$$\tilde{\mathbf{W}}_{L,j,r}^a(k) = \Omega_{\omega,j,:}^{zn(k)} \mathbf{W}_{:,r}^{za} + \Omega_{\omega,j,:}^{za(k)} \mathbf{W}_{:,r}^{za} + \Omega_{\Theta,j,r}^{za(k)} + \Omega_{\omega,j,:}^{ra(k)} \mathbf{W}_{:,r}^{ra} + \Omega_{\Theta,j,:}^{ra(k)} \mathbf{W}_{:,r}^{na},$$

$$\tilde{\mathbf{b}}_{L,j}^{(k)} = \Omega_{\omega,j,:}^{zn(k)} \mathbf{b}^z + \Omega_{\omega,j,:}^{za(k)} \mathbf{b}^z + \Omega_{\Theta,j,:}^{zn(k)} \mathbf{b}^{nx} + \Omega_{\omega,j,:}^{ra(k)} \mathbf{b}^r + \Omega_{\Theta,j,:}^{ra(k)} \mathbf{b}^{na} + \sum_{i=1}^s (\Omega_{\psi,j,i}^{zn(k)} + \Omega_{\psi,j,i}^{za(k)} + \Omega_{\psi,j,i}^{ra(k)}),$$

6. After looping steps 1 to 5 from $k = m$ to $k = 1$, the bounds are given by

$$F_j^U(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{U,j,:}^x(k) \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{U,j}^{(k)} + \mathbf{b}_j^F,$$

$$F_j^L(\mathbf{X}) = \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{L,j,:}^x(k) \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{L,j}^{(k)} + \mathbf{b}_j^F.$$

Proof 5 Given an m -layer gated recurrent unit network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$ with pre-activation bounds $\mathbf{l}^{z(k)}, \mathbf{u}^{z(k)}, \mathbf{l}^{n(k)}, \mathbf{u}^{n(k)}, \mathbf{l}^a(k), \mathbf{u}^a(k), \mathbf{l}^r(k), \mathbf{u}^r(k), \mathbf{l}^{na(k)}$ and $\mathbf{u}^{na(k)}$ for $\mathbf{X} \in \mathbb{R}^{n \times m}$, where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, let the pre-activation inputs for the j -th neuron at the output layer be $F_j(\mathbf{X}) = \mathbf{W}_{j,:}^{Fa} \mathbf{a}^{(m)} + \mathbf{b}_j^F$. The j -th output of the recurrent neural

network is the following:

$$\begin{aligned}
F_j(\mathbf{X}) &= \mathbf{W}_{j,:}^{Fa(m)} \mathbf{a}^{(m)} + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} \mathbf{a}_i^{(m)} + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} [(1 - \mathbf{z}_i^{(m)}) \mathbf{n}_i^{(m)} + \mathbf{z}_i^{(m)} \mathbf{a}_i^{(m-1)}] + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} [(1 - \sigma(\mathbf{y}_i^{z(m)})) \tanh(\mathbf{y}_i^{n(m)}) + \sigma(\mathbf{y}_i^{z(m)}) \mathbf{a}_i^{(m-1)}] + \mathbf{b}_j^F, \\
&= \sum_{\mathbf{W}_{j,i}^{Fa} \geq 0} \mathbf{W}_{j,i}^{Fa} [(1 - \sigma(\mathbf{y}_i^{z(m)})) \tanh(\mathbf{y}_i^{n(m)}) + \sigma(\mathbf{y}_i^{z(m)}) \mathbf{a}_i^{(m-1)}] \\
&\quad + \sum_{\mathbf{W}_{j,i}^{Fa} < 0} \mathbf{W}_{j,i}^{Fa} [(1 - \sigma(\mathbf{y}_i^{z(m)})) \tanh(\mathbf{y}_i^{n(m)}) + \sigma(\mathbf{y}_i^{z(m)}) \mathbf{a}_i^{(m-1)}] + \mathbf{b}_j^F.
\end{aligned} \tag{29}$$

Assume the activation function $(1 - \sigma(\mathbf{y})) \tanh(\mathbf{x})$ is bounded by two linear functions $h_{L,i}^{zn(m)}(\mathbf{y}, \mathbf{x})$, $h_{U,i}^{zn(m)}(\mathbf{y}, \mathbf{x})$ and $\sigma(\mathbf{y})\mathbf{x}$ is bounded by two linear functions $h_{L,i}^{za(m)}(\mathbf{y}, \mathbf{x})$, $h_{U,i}^{za(m)}(\mathbf{y}, \mathbf{x})$, we have

$$\begin{aligned}
h_{L,i}^{zn(m)}(\mathbf{y}_i^{z(m)}, \mathbf{y}_i^{n(m)}) &\leq (1 - \sigma(\mathbf{y}_i^{z(m)})) \tanh(\mathbf{y}_i^{n(m)}) \leq h_{U,i}^{zn(m)}(\mathbf{y}_i^{z(m)}, \mathbf{y}_i^{n(m)}), \\
h_{L,i}^{za(m)}(\mathbf{y}_i^{z(m)}, \mathbf{a}_i^{(m-1)}) &\leq \sigma(\mathbf{y}_i^{z(m)}) \mathbf{a}_i^{(m-1)} \leq h_{U,i}^{za(m)}(\mathbf{y}_i^{z(m)}, \mathbf{a}_i^{(m-1)}).
\end{aligned}$$

Upper bound. We can then obtain

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{\mathbf{W}_{j,i}^{Fa} \geq 0} \mathbf{W}_{j,i}^{Fa} (\alpha_{U,i}^{zn(m)} \mathbf{y}_i^{z(m)} + \beta_{U,i}^{zn(m)} \mathbf{y}_i^{n(m)} + \gamma_{U,i}^{zn(m)}) \\
&\quad + \alpha_{U,i}^{za(m)} \mathbf{y}_i^{z(m)} + \beta_{U,i}^{za(m)} \mathbf{a}_i^{(m-1)} + \gamma_{U,i}^{za(m)} \\
&\quad + \sum_{\mathbf{W}_{j,i}^{Fa} < 0} \mathbf{W}_{j,i}^{Fa} (\alpha_{L,i}^{zn(m)} \mathbf{y}_i^{z(m)} + \beta_{L,i}^{zn(m)} \mathbf{y}_i^{n(m)} + \gamma_{L,i}^{zn(m)}) \\
&\quad + \alpha_{L,i}^{za(m)} \mathbf{y}_i^{z(m)} + \beta_{L,i}^{za(m)} \mathbf{a}_i^{(m-1)} + \gamma_{L,i}^{za(m)} + \mathbf{b}_j^F.
\end{aligned}$$

We then define $\lambda_{j,i}^{zn(m)}$, $\Delta_{j,i}^{zn(m)}$, $\varphi_{j,i}^{zn(m)}$, $\lambda_{j,i}^{za(m)}$, $\Delta_{j,i}^{za(m)}$ and $\varphi_{j,i}^{za(m)}$ in the parentheses:

$$\begin{aligned}
\lambda_{j,i}^{zn(m)} &= \begin{cases} \alpha_{U,i}^{zn(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \alpha_{L,i}^{zn(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} & \lambda_{j,i}^{za(m)} &= \begin{cases} \alpha_{U,i}^{za(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \alpha_{L,i}^{za(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} \\
\Delta_{j,i}^{zn(m)} &= \begin{cases} \beta_{U,i}^{zn(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \beta_{L,i}^{zn(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} & \Delta_{j,i}^{za(m)} &= \begin{cases} \beta_{U,i}^{za(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \beta_{L,i}^{za(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} \\
\varphi_{j,i}^{zn(m)} &= \begin{cases} \gamma_{U,i}^{zn(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \gamma_{L,i}^{zn(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases} & \varphi_{j,i}^{za(m)} &= \begin{cases} \gamma_{U,i}^{za(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} \geq 0; \\ \gamma_{L,i}^{za(m)} & \text{if } \mathbf{W}_{j,i}^{Fa} < 0; \end{cases}
\end{aligned}$$

and obtain

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{i=1}^s \mathbf{W}_{j,i}^{Fa} (\lambda_{j,i}^{zn(m)} \mathbf{y}_i^{z(m)} + \Delta_{j,i}^{zn(m)} \mathbf{y}_i^{n(m)} + \varphi_{j,i}^{zn(m)}) \\
&\quad + \lambda_{j,i}^{za(m)} \mathbf{y}_i^{z(m)} + \Delta_{j,i}^{za(m)} \mathbf{a}_i^{(m-1)} + \varphi_{j,i}^{za(m)} + \mathbf{b}_j^F.
\end{aligned}$$

We further let

$$\begin{aligned}
\Lambda_{\lambda,j,i}^{zn(m)} &:= \mathbf{W}_{j,i}^{Fa} \lambda_{j,i}^{zn(m)}, \Lambda_{\Delta,j,i}^{zn(m)} := \mathbf{W}_{j,i}^{Fa} \Delta_{j,i}^{zn(m)}, \Lambda_{\varphi,j,i}^{zn(m)} := \mathbf{W}_{j,i}^{Fa} \varphi_{j,i}^{zn(m)}, \\
\Lambda_{\lambda,j,i}^{za(m)} &:= \mathbf{W}_{j,i}^{Fa} \lambda_{j,i}^{za(m)}, \Lambda_{\Delta,j,i}^{za(m)} := \mathbf{W}_{j,i}^{Fa} \Delta_{j,i}^{za(m)}, \Lambda_{\varphi,j,i}^{za(m)} := \mathbf{W}_{j,i}^{Fa} \varphi_{j,i}^{za(m)},
\end{aligned}$$

and we have

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{i=1}^s (\Lambda_{\lambda,j,i}^{zn(m)} \mathbf{y}_i^{z(m)} + \Lambda_{\Delta,j,i}^{zn(m)} \mathbf{y}_i^{n(m)} + \Lambda_{\varphi,j,i}^{zn(m)} + \Lambda_{\lambda,j,i}^{za(m)} \mathbf{y}_i^{z(m)} + \Lambda_{\Delta,j,i}^{za(m)} \mathbf{a}_i^{(m-1)} + \Lambda_{\varphi,j,i}^{za(m)}) + \mathbf{b}_j^F, \\
&= \sum_{i=1}^s [(\Lambda_{\lambda,j,i}^{zn(m)} + \Lambda_{\lambda,j,i}^{za(m)}) (\mathbf{W}_{i,:}^{zx} \mathbf{x}^{(m)} + \mathbf{W}_{i,:}^{za} \mathbf{a}^{(m-1)} + \mathbf{b}_i^z) + \Lambda_{\Delta,j,i}^{zn(m)} (\mathbf{W}_{i,:}^{nx} \mathbf{x}^{(m)} + \mathbf{b}_i^{nx}) \\
&\quad + \mathbf{r}_i^{(m)} (\mathbf{W}_{i,:}^{na} \mathbf{a}^{(m-1)} + \mathbf{b}_i^{na}) + \Lambda_{\varphi,j,i}^{zn(m)} + \Lambda_{\varphi,j,i}^{za(m)} + \Lambda_{\Delta,j,i}^{za(m)} \mathbf{a}_i^{(m-1)}] + \mathbf{b}_j^F, \\
&= \sum_{q_m=1}^n (\sum_{i=1}^s (\Lambda_{\lambda,j,i}^{zn(m)} + \Lambda_{\lambda,j,i}^{za(m)}) \mathbf{W}_{i,q_m}^{zx} + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{zn(m)} \mathbf{W}_{i,q_m}^{nx}) \mathbf{x}_{q_m}^{(m)} \\
&\quad + \sum_{r=1}^s (\sum_{i=1}^s (\Lambda_{\lambda,j,i}^{zn(m)} + \Lambda_{\lambda,j,i}^{za(m)}) \mathbf{W}_{i,r}^{za} + \Lambda_{\Delta,j,r}^{za(m)}) \mathbf{a}_r^{(m-1)} \\
&\quad + \sum_{i=1}^s \Lambda_{\lambda,j,i}^{zn(m)} \mathbf{b}_i^z + \sum_{i=1}^s \Lambda_{\lambda,j,i}^{za(m)} \mathbf{b}_i^z + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{zn(m)} \mathbf{b}_i^{nx} + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{zn(m)} + \Lambda_{\varphi,j,i}^{za(m)}) \\
&\quad + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{zn(m)} (\sigma(\mathbf{y}_i^{r(m)}) (\mathbf{W}_{i,:}^{na} \mathbf{a}^{(m-1)} + \mathbf{b}_i^{na})) + \mathbf{b}_j^F.
\end{aligned}$$

Assume the activation function $\sigma(\mathbf{y})\mathbf{x}$ is bounded by two linear functions $h_{L,i}^{ra(m)}(\mathbf{y}, \mathbf{x})$, $h_{U,i}^{ra(m)}(\mathbf{y}, \mathbf{x})$, we have

$$h_{L,i}^{ra(m)}(\mathbf{y}_i^{r(m)}, \mathbf{W}_{i,:}^{na} \mathbf{a}^{(m-1)} + \mathbf{b}_i^{na}) \leq \sigma(\mathbf{y}_i^{r(m)}) (\mathbf{W}_{i,:}^{na} \mathbf{a}^{(m-1)} + \mathbf{b}_i^{na}) \leq h_{U,i}^{ra(m)}(\mathbf{y}_i^{r(m)}, \mathbf{W}_{i,:}^{na} \mathbf{a}^{(m-1)} + \mathbf{b}_i^{na}).$$

We then have

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{q_m=1}^n ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,q_m}^{zx} + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{W}_{:,q_m}^{nx}) \mathbf{x}_{q_m}^{(m)} \\
&\quad + \sum_{r=1}^s ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,r}^{za} + \Lambda_{\Delta,j,r}^{za(m)}) \mathbf{a}_r^{(m-1)} \\
&\quad + \Lambda_{\lambda,j,:}^{zn(m)} \mathbf{b}^z + \Lambda_{\lambda,j,:}^{za(m)} \mathbf{b}^z + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{b}^{nx} + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{zn(m)} + \Lambda_{\varphi,j,i}^{za(m)}) \\
&\quad + \sum_{\Lambda_{\Delta,j,i}^{zn(m)} \geq 0} \Lambda_{\Delta,j,i}^{zn(m)} [\alpha_{U,i}^{ra(m)} \mathbf{y}_i^{r(m)} + \beta_{U,i}^{ra(m)} (\mathbf{W}_{i,:}^{na} \mathbf{a}^{(m-1)} + \mathbf{b}_i^{na}) + \gamma_{U,i}^{ra(m)}] \\
&\quad + \sum_{\Lambda_{\Delta,j,i}^{zn(m)} < 0} \Lambda_{\Delta,j,i}^{zn(m)} [\alpha_{L,i}^{ra(m)} \mathbf{y}_i^{r(m)} + \beta_{L,i}^{ra(m)} (\mathbf{W}_{i,:}^{na} \mathbf{a}^{(m-1)} + \mathbf{b}_i^{na}) + \gamma_{L,i}^{ra(m)}] + \mathbf{b}_j^F.
\end{aligned}$$

We then define $\lambda_{j,i}^{ra(m)}$, $\Delta_{j,i}^{ra(m)}$, $\varphi_{j,i}^{ra(m)}$, $\Lambda_{\lambda,j,i}^{ra(m)}$, $\Lambda_{\Delta,j,i}^{ra(m)}$ and $\Lambda_{\varphi,j,i}^{ra(m)}$ in the parentheses:

$$\begin{aligned}
\lambda_{j,i}^{ra(m)} &= \begin{cases} \alpha_{U,i}^{ra(m)} & \text{if } \Lambda_{\Delta,j,i}^{zn(m)} \geq 0; \\ \alpha_{L,i}^{ra(m)} & \text{if } \Lambda_{\Delta,j,i}^{zn(m)} < 0; \end{cases} \quad \Lambda_{\lambda,j,i}^{ra(m)} := \Lambda_{\Delta,j,i}^{zn(m)} \lambda_{j,i}^{ra(m)}; \\
\Delta_{j,i}^{ra(m)} &= \begin{cases} \beta_{U,i}^{ra(m)} & \text{if } \Lambda_{\Delta,j,i}^{zn(m)} \geq 0; \\ \beta_{L,i}^{ra(m)} & \text{if } \Lambda_{\Delta,j,i}^{zn(m)} < 0; \end{cases} \quad \Lambda_{\Delta,j,i}^{ra(m)} := \Lambda_{\Delta,j,i}^{zn(m)} \Delta_{j,i}^{ra(m)}; \\
\varphi_{j,i}^{ra(m)} &= \begin{cases} \gamma_{U,i}^{ra(m)} & \text{if } \Lambda_{\Delta,j,i}^{zn(m)} \geq 0; \\ \gamma_{L,i}^{ra(m)} & \text{if } \Lambda_{\Delta,j,i}^{zn(m)} < 0; \end{cases} \quad \Lambda_{\varphi,j,i}^{ra(m)} := \Lambda_{\Delta,j,i}^{zn(m)} \varphi_{j,i}^{ra(m)};
\end{aligned}$$

and obtain

$$\begin{aligned}
F_j(\mathbf{X}) &\leq \sum_{q_m=1}^n ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,q_m}^{zx} + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{W}_{:,q_m}^{nx}) \mathbf{x}_{q_m}^{(m)} \\
&+ \sum_{r=1}^s ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,r}^{za} + \Lambda_{\Delta,j,r}^{za(m)}) \mathbf{a}_r^{(m-1)} \\
&+ \Lambda_{\lambda,j,:}^{zn(m)} \mathbf{b}^z + \Lambda_{\lambda,j,:}^{za(m)} \mathbf{b}^z + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{b}^{nx} + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{zn(m)} + \Lambda_{\varphi,j,i}^{za(m)}) \\
&+ \sum_{i=1}^s [\Lambda_{\lambda,j,i}^{ra(m)} \mathbf{y}_i^{r(m)} + \Lambda_{\Delta,j,i}^{ra(m)} (\mathbf{W}_{i,:}^{na} \mathbf{a}^{(m-1)} + \mathbf{b}_i^{na}) + \Lambda_{\varphi,j,i}^{ra(m)}] + \mathbf{b}_j^F \\
&= \sum_{q_m=1}^n ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,q_m}^{zx} + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{W}_{:,q_m}^{nx}) \mathbf{x}_{q_m}^{(m)} \\
&+ \sum_{r=1}^s ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,r}^{za} + \Lambda_{\Delta,j,r}^{za(m)}) \mathbf{a}_r^{(m-1)} \\
&+ \Lambda_{\lambda,j,:}^{zn(m)} \mathbf{b}^z + \Lambda_{\lambda,j,:}^{za(m)} \mathbf{b}^z + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{b}^{nx} + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{zn(m)} + \Lambda_{\varphi,j,i}^{za(m)}) \\
&+ \sum_{i=1}^s \Lambda_{\lambda,j,i}^{ra(m)} (\mathbf{W}_{i,:}^{rx} \mathbf{x}^{(m)} + \mathbf{W}_{i,:}^{ra} \mathbf{a}^{(m-1)} + \mathbf{b}_i^r) + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{ra(m)} (\mathbf{W}_{i,:}^{na} \mathbf{a}^{(m-1)} + \mathbf{b}_i^{na}) + \sum_{i=1}^s \Lambda_{\varphi,j,i}^{ra(m)} + \mathbf{b}_j^F \\
&= \sum_{q_m=1}^n ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,q_m}^{zx} + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{W}_{:,q_m}^{nx}) \mathbf{x}_{q_m}^{(m)} \\
&+ \sum_{r=1}^s ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,r}^{za} + \Lambda_{\Delta,j,r}^{za(m)}) \mathbf{a}_r^{(m-1)} \\
&+ \Lambda_{\lambda,j,:}^{zn(m)} \mathbf{b}^z + \Lambda_{\lambda,j,:}^{za(m)} \mathbf{b}^z + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{b}^{nx} + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{zn(m)} + \Lambda_{\varphi,j,i}^{za(m)}) \\
&+ \sum_{q_m=1}^n (\sum_{i=1}^s \Lambda_{\lambda,j,i}^{ra(m)} \mathbf{W}_{i,q_m}^{rx}) \mathbf{x}_{q_m}^{(m)} + \sum_{r=1}^s (\sum_{i=1}^s \Lambda_{\lambda,j,i}^{ra(m)} \mathbf{W}_{i,r}^{ra}) \mathbf{a}_r^{(m-1)} + \sum_{i=1}^s \Lambda_{\lambda,j,i}^{ra(m)} \mathbf{b}_i^r \\
&+ \sum_{r=1}^s (\sum_{i=1}^s \Lambda_{\Delta,j,i}^{ra(m)} \mathbf{W}_{i,r}^{na}) \mathbf{a}_r^{(m-1)} + \sum_{i=1}^s \Lambda_{\Delta,j,i}^{ra(m)} \mathbf{b}_i^{na} + \sum_{i=1}^s \Lambda_{\varphi,j,i}^{ra(m)} + \mathbf{b}_j^F \\
&= \sum_{q_m=1}^n ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,q_m}^{zx} + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{W}_{:,q_m}^{nx} + \Lambda_{\lambda,j,:}^{ra(m)} \mathbf{W}_{:,q_m}^{rx}) \mathbf{x}_{q_m}^{(m)} \\
&+ \sum_{r=1}^s ((\Lambda_{\lambda,j,:}^{zn(m)} + \Lambda_{\lambda,j,:}^{za(m)}) \mathbf{W}_{:,r}^{za} + \Lambda_{\Delta,j,r}^{za(m)} + \Lambda_{\lambda,j,:}^{ra(m)} \mathbf{W}_{:,r}^{ra} + \Lambda_{\Delta,j,:}^{ra(m)} \mathbf{W}_{:,r}^{na}) \mathbf{a}_r^{(m-1)} \\
&+ \Lambda_{\lambda,j,:}^{zn(m)} \mathbf{b}^z + \Lambda_{\lambda,j,:}^{za(m)} \mathbf{b}^z + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{b}^{nx} + \Lambda_{\lambda,j,:}^{ra(m)} \mathbf{b}^r + \Lambda_{\Delta,j,:}^{ra(m)} \mathbf{b}^{na} + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{zn(m)} + \Lambda_{\varphi,j,i}^{za(m)} + \Lambda_{\varphi,j,i}^{ra(m)}) \\
&+ \mathbf{b}_j^F. \tag{30}
\end{aligned}$$

We combine coefficients in Equation (30) into equivalent weights $\tilde{\mathbf{W}}_{U,j,q_m}^{x(m)}$, $\tilde{\mathbf{W}}_{U,j,r}^{a(m)}$ and bias $\tilde{\mathbf{b}}_{U,j}^{(m)}$ as defined by

$$\begin{aligned}
\tilde{\mathbf{W}}_{U,j,q_m}^{x(m)} &= \Lambda_{\lambda,j,:}^{zn(m)} \mathbf{W}_{:,q_m}^{zx} + \Lambda_{\lambda,j,:}^{za(m)} \mathbf{W}_{:,q_m}^{zx} + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{W}_{:,q_m}^{nx} + \Lambda_{\lambda,j,:}^{ra(m)} \mathbf{W}_{:,q_m}^{rx}, \\
\tilde{\mathbf{W}}_{U,j,r}^{a(m)} &= \Lambda_{\lambda,j,:}^{zn(m)} \mathbf{W}_{:,r}^{za} + \Lambda_{\lambda,j,:}^{za(m)} \mathbf{W}_{:,r}^{za} + \Lambda_{\Delta,j,r}^{za(m)} + \Lambda_{\lambda,j,:}^{ra(m)} \mathbf{W}_{:,r}^{ra} + \Lambda_{\Delta,j,:}^{ra(m)} \mathbf{W}_{:,r}^{na},
\end{aligned}$$

$$\tilde{\mathbf{b}}_{U,j}^{(m)} = \Lambda_{\lambda,j,:}^{zn(m)} \mathbf{b}^z + \Lambda_{\lambda,j,:}^{za(m)} \mathbf{b}^z + \Lambda_{\Delta,j,:}^{zn(m)} \mathbf{b}^{nx} + \Lambda_{\lambda,j,:}^{ra(m)} \mathbf{b}^r + \Lambda_{\Delta,j,:}^{ra(m)} \mathbf{b}^{na} + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{zn(m)} + \Lambda_{\varphi,j,i}^{za(m)} + \Lambda_{\varphi,j,i}^{ra(m)}),$$

and obtain

$$F_j(\mathbf{X}) \leq F_j^{U,m}(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{W}}_{U,j,:}^{a(m)} \mathbf{a}^{(m-1)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \mathbf{b}_j^F. \quad (31)$$

Notice that Equation (29) and Equation (31) are in similar forms. Thus, $\forall r \in s$ we can define variables as follows:

$$\begin{aligned} \lambda_{j,r}^{zn(m-1)} &= \begin{cases} \alpha_{U,r}^{zn(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} \geq 0; \\ \alpha_{L,r}^{zn(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} < 0; \end{cases} & \lambda_{j,r}^{za(m-1)} &= \begin{cases} \alpha_{U,r}^{za(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} \geq 0; \\ \alpha_{L,r}^{za(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} < 0; \end{cases} \\ \Delta_{j,r}^{zn(m-1)} &= \begin{cases} \beta_{U,r}^{zn(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} \geq 0; \\ \beta_{L,r}^{zn(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} < 0; \end{cases} & \Delta_{j,r}^{za(m-1)} &= \begin{cases} \beta_{U,r}^{za(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} \geq 0; \\ \beta_{L,r}^{za(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} < 0; \end{cases} \\ \varphi_{j,r}^{zn(m-1)} &= \begin{cases} \gamma_{U,r}^{zn(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} \geq 0; \\ \gamma_{L,r}^{zn(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} < 0; \end{cases} & \varphi_{j,r}^{za(m-1)} &= \begin{cases} \gamma_{U,r}^{za(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} \geq 0; \\ \gamma_{L,r}^{za(m-1)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(m)} < 0; \end{cases} \\ \Lambda_{\lambda,j,:}^{zn(m-1)} &= \tilde{\mathbf{W}}_{U,j,:}^{a(m)} \odot \lambda_{j,:}^{zn(m-1)}; & \Lambda_{\lambda,j,:}^{za(m-1)} &= \tilde{\mathbf{W}}_{U,j,:}^{a(m)} \odot \lambda_{j,:}^{za(m-1)}; \\ \Lambda_{\Delta,j,:}^{zn(m-1)} &= \tilde{\mathbf{W}}_{U,j,:}^{a(m)} \odot \Delta_{j,:}^{zn(m-1)}; & \Lambda_{\Delta,j,:}^{za(m-1)} &= \tilde{\mathbf{W}}_{U,j,:}^{a(m)} \odot \Delta_{j,:}^{za(m-1)}; \\ \Lambda_{\varphi,j,:}^{zn(m-1)} &= \tilde{\mathbf{W}}_{U,j,:}^{a(m)} \odot \varphi_{j,:}^{zn(m-1)}; & \Lambda_{\varphi,j,:}^{za(m-1)} &= \tilde{\mathbf{W}}_{U,j,:}^{a(m)} \odot \varphi_{j,:}^{za(m-1)}; \\ \lambda_{j,r}^{ra(m-1)} &= \begin{cases} \alpha_{U,r}^{ra(m-1)} & \text{if } \Lambda_{\Delta,j,r}^{zn(m-1)} \geq 0; \\ \alpha_{L,r}^{ra(m-1)} & \text{if } \Lambda_{\Delta,j,r}^{zn(m-1)} < 0; \end{cases} & \Lambda_{\lambda,j,:}^{ra(m-1)} &= \Lambda_{\Delta,j,:}^{zn(m-1)} \odot \lambda_{j,:}^{ra(m-1)}; \\ \Delta_{j,r}^{ra(m-1)} &= \begin{cases} \beta_{U,r}^{ra(m-1)} & \text{if } \Lambda_{\Delta,j,r}^{zn(m-1)} \geq 0; \\ \beta_{L,r}^{ra(m-1)} & \text{if } \Lambda_{\Delta,j,r}^{zn(m-1)} < 0; \end{cases} & \Lambda_{\Delta,j,:}^{ra(m-1)} &= \Lambda_{\Delta,j,:}^{zn(m-1)} \odot \Delta_{j,:}^{ra(m-1)}; \\ \varphi_{j,r}^{ra(m-1)} &= \begin{cases} \gamma_{U,r}^{ra(m-1)} & \text{if } \Lambda_{\Delta,j,r}^{zn(m-1)} \geq 0; \\ \gamma_{L,r}^{ra(m-1)} & \text{if } \Lambda_{\Delta,j,r}^{zn(m-1)} < 0; \end{cases} & \Lambda_{\varphi,j,:}^{ra(m-1)} &= \Lambda_{\Delta,j,:}^{zn(m-1)} \odot \varphi_{j,:}^{ra(m-1)}. \end{aligned}$$

Then we have

$$F_j^{U,m}(\mathbf{X}) \leq F_j^{U,m-1}(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{x(m)} \mathbf{x}^{(m)} + \tilde{\mathbf{W}}_{U,j,:}^{x(m-1)} \mathbf{x}^{(m-1)} + \tilde{\mathbf{W}}_{U,j,:}^{a(m-1)} \mathbf{a}^{(m-2)} + \tilde{\mathbf{b}}_{U,j}^{(m)} + \tilde{\mathbf{b}}_{U,j}^{(m-1)} + \mathbf{b}_j^F.$$

Thus, we can repeat the above procedure iteratively until we obtain the final upper bound $F_j^{U,1}(\mathbf{X})$, where $F_j(\mathbf{X}) \leq F_j^{U,m}(\mathbf{X}) \leq \dots \leq F_j^{U,1}(\mathbf{X})$. We let $F_j^U(\mathbf{X})$ denote the final upper bound $F_j^{U,1}(\mathbf{X})$, and we have

$$F_j^U(\mathbf{X}) = \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{U,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{U,j}^{(k)} + \mathbf{b}_j^F,$$

where

$$\begin{aligned} \tilde{\mathbf{W}}_{U,j,qk}^{x(k)} &= \Lambda_{\lambda,j,:}^{zn(k)} \mathbf{W}_{:,qk}^{zx} + \Lambda_{\lambda,j,:}^{za(k)} \mathbf{W}_{:,qk}^{zx} + \Lambda_{\Delta,j,:}^{zn(k)} \mathbf{W}_{:,qk}^{nx} + \Lambda_{\lambda,j,:}^{ra(k)} \mathbf{W}_{:,qk}^{rx}, \\ \tilde{\mathbf{W}}_{U,j,r}^{a(k)} &= \Lambda_{\lambda,j,:}^{zn(k)} \mathbf{W}_{:,r}^{za} + \Lambda_{\lambda,j,:}^{za(k)} \mathbf{W}_{:,r}^{za} + \Lambda_{\Delta,j,r}^{za(k)} + \Lambda_{\lambda,j,:}^{ra(k)} \mathbf{W}_{:,r}^{ra} + \Lambda_{\Delta,j,:}^{ra(k)} \mathbf{W}_{:,r}^{na}, \\ \tilde{\mathbf{b}}_{U,j}^{(k)} &= \Lambda_{\lambda,j,:}^{zn(k)} \mathbf{b}^z + \Lambda_{\lambda,j,:}^{za(k)} \mathbf{b}^z + \Lambda_{\Delta,j,:}^{zn(k)} \mathbf{b}^{nx} + \Lambda_{\lambda,j,:}^{ra(k)} \mathbf{b}^r + \Lambda_{\Delta,j,:}^{ra(k)} \mathbf{b}^{na} + \sum_{i=1}^s (\Lambda_{\varphi,j,i}^{zn(k)} + \Lambda_{\varphi,j,i}^{za(k)} + \Lambda_{\varphi,j,i}^{ra(k)}), \end{aligned}$$

and (\odot is the Hadamard product)

$$\begin{aligned} \Lambda_{\lambda,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \lambda_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} & \Lambda_{\lambda,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \lambda_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \lambda_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases} \\ \Lambda_{\Delta,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Delta_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \Delta_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} & \Lambda_{\Delta,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Delta_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \Delta_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases} \\ \Lambda_{\varphi,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \varphi_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \varphi_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} & \Lambda_{\varphi,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \varphi_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{U,j,:}^{a(k+1)} \odot \varphi_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases} \end{aligned}$$

$$\begin{aligned}
\Lambda_{\lambda,j,:}^{ra(k)} &= \Lambda_{\Delta,j,:}^{zn(k)} \odot \lambda_{j,:}^{ra(k)} & \text{if } k \in [m]; \\
\Lambda_{\Delta,j,:}^{ra(k)} &= \Lambda_{\Delta,j,:}^{zn(k)} \odot \Delta_{j,:}^{ra(k)} & \text{if } k \in [m]; \\
\Lambda_{\varphi,j,:}^{ra(k)} &= \Lambda_{\Delta,j,:}^{zn(k)} \odot \varphi_{j,:}^{ra(k)} & \text{if } k \in [m];
\end{aligned}$$

and $\forall r \in s$,

$$\begin{aligned}
\lambda_{j,r}^{zn(m)} &= \begin{cases} \alpha_{U,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \alpha_{L,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} & \Delta_{j,r}^{zn(m)} &= \begin{cases} \beta_{U,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \beta_{L,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} & \varphi_{j,r}^{zn(m)} &= \begin{cases} \gamma_{U,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \gamma_{L,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} \\
\lambda_{j,r}^{za(m)} &= \begin{cases} \alpha_{U,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \alpha_{L,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} & \Delta_{j,r}^{za(m)} &= \begin{cases} \beta_{U,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \beta_{L,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} & \varphi_{j,r}^{za(m)} &= \begin{cases} \gamma_{U,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \gamma_{L,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases}
\end{aligned}$$

and $\forall r \in s, k \in [m-1]$,

$$\begin{aligned}
\lambda_{j,r}^{zn(k)} &= \begin{cases} \alpha_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \alpha_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \lambda_{j,r}^{za(k)} &= \begin{cases} \alpha_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \alpha_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} \\
\Delta_{j,r}^{zn(k)} &= \begin{cases} \beta_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \beta_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \Delta_{j,r}^{za(k)} &= \begin{cases} \beta_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \beta_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} \\
\varphi_{j,r}^{zn(k)} &= \begin{cases} \gamma_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \gamma_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases} & \varphi_{j,r}^{za(k)} &= \begin{cases} \gamma_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} \geq 0; \\ \gamma_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{U,j,r}^{a(k+1)} < 0; \end{cases}
\end{aligned}$$

and $\forall r \in s, k \in [m]$,

$$\begin{aligned}
\lambda_{j,r}^{ra(k)} &= \begin{cases} \alpha_{U,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} \geq 0; \\ \alpha_{L,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} < 0; \end{cases} \\
\Delta_{j,r}^{ra(k)} &= \begin{cases} \beta_{U,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} \geq 0; \\ \beta_{L,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} < 0; \end{cases} \\
\varphi_{j,r}^{ra(k)} &= \begin{cases} \gamma_{U,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} \geq 0; \\ \gamma_{L,r}^{ra(k)} & \text{if } \Lambda_{\Delta,j,r}^{zn(k)} < 0. \end{cases}
\end{aligned}$$

Lower bound. The above derivations of upper bound can be applied similarly to derive lower bounds of $F_j(\mathbf{X})$, and the only difference is now we need to use the left-hand side of inequities (rather than right-hand side when deriving upper bound) to bound the activated terms. Thus, following the same procedure in deriving the upper bounds, we can iteratively unwrap the activation functions and obtain a final lower bound $F_j^{L,1}(\mathbf{X})$, where $F_j(\mathbf{X}) \geq F_j^{L,m}(\mathbf{X}) \geq \dots \geq F_j^{L,1}(\mathbf{X})$. We let $F_j^L(\mathbf{X})$ denote the final lower bound $F_j^{L,1}(\mathbf{X})$, and we have

$$F_j^L(\mathbf{X}) = \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \tilde{\mathbf{W}}_{L,j,:}^{x(k)} \mathbf{x}^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{L,j}^{(k)} + \mathbf{b}_j^F,$$

where

$$\begin{aligned}
\tilde{\mathbf{W}}_{L,j,q_k}^{x(k)} &= \Omega_{\omega,j,:}^{zn(k)} \mathbf{W}_{:,q_k}^{zx} + \Omega_{\omega,j,:}^{za(k)} \mathbf{W}_{:,q_k}^{zx} + \Omega_{\Theta,j,:}^{zn(k)} \mathbf{W}_{:,q_k}^{nx} + \Omega_{\omega,j,:}^{ra(k)} \mathbf{W}_{:,q_k}^{rx}, \\
\tilde{\mathbf{W}}_{L,j,r}^{a(k)} &= \Omega_{\omega,j,:}^{zn(k)} \mathbf{W}_{:,r}^{za} + \Omega_{\omega,j,:}^{za(k)} \mathbf{W}_{:,r}^{za} + \Omega_{\Theta,j,r}^{za(k)} + \Omega_{\omega,j,:}^{ra(k)} \mathbf{W}_{:,r}^{ra} + \Omega_{\Theta,j,:}^{ra(k)} \mathbf{W}_{:,r}^{na}, \\
\tilde{\mathbf{b}}_{L,j}^{(k)} &= \Omega_{\omega,j,:}^{zn(k)} \mathbf{b}^z + \Omega_{\omega,j,:}^{za(k)} \mathbf{b}^z + \Omega_{\Theta,j,:}^{zn(k)} \mathbf{b}^{nx} + \Omega_{\omega,j,:}^{ra(k)} \mathbf{b}^r + \Omega_{\Theta,j,:}^{ra(k)} \mathbf{b}^{na} + \sum_{i=1}^s (\Omega_{\psi,j,i}^{zn(k)} + \Omega_{\psi,j,i}^{za(k)} + \Omega_{\psi,j,i}^{ra(k)}),
\end{aligned}$$

and (\odot is the Hadamard product)

$$\begin{aligned}\Omega_{\omega,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \omega_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\omega,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \omega_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \omega_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases} \\ \Omega_{\Theta,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Theta_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \Theta_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\Theta,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \Theta_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \Theta_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases} \\ \Omega_{\psi,j,:}^{zn(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \psi_{j,:}^{zn(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \psi_{j,:}^{zn(k)} & \text{if } k \in [m-1]; \end{cases} & \Omega_{\psi,j,:}^{za(k)} &= \begin{cases} \mathbf{W}_{j,:}^{Fa} \odot \psi_{j,:}^{za(k)} & \text{if } k = m; \\ \tilde{\mathbf{W}}_{L,j,:}^{a(k+1)} \odot \psi_{j,:}^{za(k)} & \text{if } k \in [m-1]; \end{cases}\end{aligned}$$

$$\begin{aligned}\Omega_{\omega,j,:}^{ra(k)} &= \Omega_{\Theta,j,:}^{zn(k)} \odot \omega_{j,:}^{ra(k)} & \text{if } k \in [m]; \\ \Omega_{\Theta,j,:}^{ra(k)} &= \Omega_{\Theta,j,:}^{zn(k)} \odot \Theta_{j,:}^{ra(k)} & \text{if } k \in [m]; \\ \Omega_{\psi,j,:}^{ra(k)} &= \Omega_{\Theta,j,:}^{zn(k)} \odot \psi_{j,:}^{ra(k)} & \text{if } k \in [m];\end{aligned}$$

and $\forall r \in s$,

$$\begin{aligned}\omega_{j,r}^{zn(m)} &= \begin{cases} \alpha_{L,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \alpha_{U,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} & \Theta_{j,r}^{zn(m)} &= \begin{cases} \beta_{L,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \beta_{U,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} & \psi_{j,r}^{zn(m)} &= \begin{cases} \gamma_{L,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \gamma_{U,r}^{zn(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} \\ \omega_{j,r}^{za(m)} &= \begin{cases} \alpha_{L,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \alpha_{U,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} & \Theta_{j,r}^{za(m)} &= \begin{cases} \beta_{L,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \beta_{U,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases} & \psi_{j,r}^{za(m)} &= \begin{cases} \gamma_{L,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} \geq 0; \\ \gamma_{U,r}^{za(m)} & \text{if } \mathbf{W}_{j,r}^{Fa} < 0; \end{cases}\end{aligned}$$

and $\forall r \in s, k \in [m-1]$,

$$\begin{aligned}\omega_{j,r}^{zn(k)} &= \begin{cases} \alpha_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \alpha_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} & \omega_{j,r}^{za(k)} &= \begin{cases} \alpha_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \alpha_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\ \Theta_{j,r}^{zn(k)} &= \begin{cases} \beta_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \beta_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} & \Theta_{j,r}^{za(k)} &= \begin{cases} \beta_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \beta_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} \\ \psi_{j,r}^{zn(k)} &= \begin{cases} \gamma_{L,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \gamma_{U,r}^{zn(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases} & \psi_{j,r}^{za(k)} &= \begin{cases} \gamma_{L,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} \geq 0; \\ \gamma_{U,r}^{za(k)} & \text{if } \tilde{\mathbf{W}}_{L,j,r}^{a(k+1)} < 0; \end{cases}\end{aligned}$$

and $\forall r \in s, k \in [m]$,

$$\begin{aligned}\omega_{j,r}^{ra(k)} &= \begin{cases} \alpha_{L,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} \geq 0; \\ \alpha_{U,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} < 0; \end{cases} \\ \Theta_{j,r}^{ra(k)} &= \begin{cases} \beta_{L,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} \geq 0; \\ \beta_{U,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} < 0; \end{cases} \\ \psi_{j,r}^{ra(k)} &= \begin{cases} \gamma_{L,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} \geq 0; \\ \gamma_{U,r}^{ra(k)} & \text{if } \Omega_{\Theta,j,r}^{zn(k)} < 0. \end{cases}\end{aligned}$$

Corollary A.5 (Closed-form global bounds) Given data $\mathbf{X} \in \mathbb{R}^{n \times m}$, l_p ball parameters $p \geq 1$ and $\epsilon \geq 0$. For an m -layer gated recurrent unit network $F : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}^t$, there exists two fixed values γ_j^L and γ_j^U such that $\forall \mathbf{X} \in \mathbb{R}^{n \times m}$ where $\mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon)$, and $\forall j \in [t]$, $1/q = 1 - 1/p$, the inequity $\gamma_j^L \leq F_j(\mathbf{X}) \leq \gamma_j^U$ holds true, where

$$\begin{aligned}\gamma_j^U &= \tilde{\mathbf{W}}_{U,j,:}^{a(1)} \mathbf{a}^{(0)} + \sum_{k=1}^m \epsilon \|\tilde{\mathbf{W}}_{U,j,:}^{x(k)}\|_q + \sum_{k=1}^m \tilde{\mathbf{W}}_{U,j,:}^{x(k)} \mathbf{x}_0^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{U,j}^{(k)} + \mathbf{b}_j^F, \\ \gamma_j^L &= \tilde{\mathbf{W}}_{L,j,:}^{a(1)} \mathbf{a}^{(0)} - \sum_{k=1}^m \epsilon \|\tilde{\mathbf{W}}_{L,j,:}^{x(k)}\|_q + \sum_{k=1}^m \tilde{\mathbf{W}}_{L,j,:}^{x(k)} \mathbf{x}_0^{(k)} + \sum_{k=1}^m \tilde{\mathbf{b}}_{L,j}^{(k)} + \mathbf{b}_j^F.\end{aligned}$$

Corollary A. 6 (Pre-activation bounds in GRUs) *The pre-activation bounds of different gates can be obtained via simple modifications from Theorem A.6. Specifically, we recall the definitions of pre-activations in different gates:*

$$\begin{aligned}
\text{Reset gate: } \quad \mathbf{y}^{r(k)} &= \underline{\mathbf{W}^{rx} \mathbf{x}^{(k)}} + \mathbf{W}^{ra} \mathbf{a}^{(k-1)} + \mathbf{b}^r; \\
\text{Update gate: } \quad \mathbf{y}^{z(k)} &= \underline{\mathbf{W}^{zx} \mathbf{x}^{(k)}} + \mathbf{W}^{za} \mathbf{a}^{(k-1)} + \mathbf{b}^z; \\
\text{New gate: } \quad \mathbf{y}^{n(k)} &= \underline{\mathbf{W}^{nx} \mathbf{x}^{(k)}} + \mathbf{b}^{nx} + \mathbf{r}^{(k)} \odot (\mathbf{W}^{na} \mathbf{a}^{(k-1)} + \mathbf{b}^{na}).
\end{aligned}$$

For a given GRU and input \mathbf{X}_0 , the underlined part of the equations are bounded by

$$\mathbf{W}^{gate} \mathbf{x}_0^{(k)} - \epsilon \|\mathbf{W}^{gate}\|_q \leq \mathbf{W}^{gate} \mathbf{x}^{(k)} \leq \mathbf{W}^{gate} \mathbf{x}_0^{(k)} + \epsilon \|\mathbf{W}^{gate}\|_q,$$

where $gate = \{rx, zx, nx\}$. Thus to derive pre-activation bounds, we only need to know the ranges of remaining parts in the equations above. These can be computed by replacing the output mapping \mathbf{W}^{Fa} in Theorem A.6 by $\mathbf{W}^{ra}, \mathbf{W}^{za}, \mathbf{W}^{na}$, respectively. The output bias \mathbf{b}^F is correspondingly substituted with $\mathbf{b}^r, \mathbf{b}^z, \mathbf{b}^{na}$. For example, to derive the bounds for pre-activation bounds of reset gates, we replace the output mapping matrix and bias from \mathbf{W}^{Fa} and \mathbf{b}^F by \mathbf{W}^{ra} and \mathbf{b}^r . For the pre-activation bound of new gate, we need to first derive the bound of reset gate. Considering that there is a cross-nonlinear term of reset gate and the previous hidden state, therefore we need to further use 2D bounding planes as introduced in Section 3.2 to bound the term. Lastly, the pre-activation bound is found by combining the linear bounds with $\mathbf{W}^{nx} \mathbf{x}_0^{(k)} \pm \epsilon \|\mathbf{W}^{nx}\|_q + \mathbf{b}^{nx}$.

B. Experiments

B.1 Significance of plane (2D)-bounding POPQORN certificates

Recalling that when analyzing the bounds for an LSTM, we extend the ideas of bounding lines to bounding planes. In what follows, as proofs and formulas for POPQORN certificates with both 2D bounding planes (A.3) and 1D bounding lines (A.6) are provided, we justify our proposals of using 2D bounding planes instead of 1D or even constant bounding² alternatives from two angles:

1. In Experiment (I) - MNIST handwritten digit classification task, constant and 1D bounding approaches give much smaller (3-6 times smaller for 4-slice) certificates than those obtained by 2D bounding, and they become even looser as the LSTM goes deeper (400-3000 times smaller for 14-slice). Specifically,
 - LSTM 4-slice: average bound for (2D, 1D, constant) bound = (0.232, 0.071, 0.041)
 - LSTM 7-slice: average bound for (2D, 1D, constant) bound = (0.140, 0.015, 0.012)
 - LSTM 14-slice: average bound for (2D, 1D, constant) bound = (0.059, 1.5e-4, 2e-5)
2. In Experiment (III) - question classification task, robustness bounds obtained are only meaningful when the bounds are large enough to cover word pairs in the word embedding space. Here, we show that the robustness certificates obtained via constant and 1D bounding contain significantly fewer word embeddings compared with the proposed 2D bounding. Specifically, we use 2 different groups of word embeddings in a length-10K vocabulary and show that our 2D bounding technique indeed contains non-trivial numbers (19, 535635) of word embedding pairs for the 1st LSTM and the 2nd LSTM. Meanwhile, 1D and constant bounding have issues of robustness certificates being smaller than the minimum distance of word embeddings (MDWE), where their certificates contain (9, 0) and (0, 0) word embedding pairs.
 - LSTM 1: MDWE = 0.120, average bound for (2D, 1D, constant) bound = (0.138,0.067,0.039)
 - LSTM 2: MDWE = 0.012, average bound for (2D, 1D, constant) bound = (0.027,0.013,0.008)

The above results demonstrate the necessity and significance of the proposed 2D gate-bounding planes, in order to obtain *non-trivial* robustness certificates.

²Modifying the results of 2D and 1D bounding to constant bounding means that pre-activation bounds are directly combined and considered when evaluating ranges of activated states. This is rather a trivial implementation, whose details are therefore omitted herein.

B.2 CLEVER-RNN Score

Assume for a given data sequence \mathbf{X} , j is the true label, and i is the target label. Let $g(\mathbf{X}) = F_j(\mathbf{X}) - F_i(\mathbf{X})$, is a Lipschitz function that has the property below:

$$\begin{aligned} |g(\mathbf{X}) - g(\mathbf{Y})| &= |g(\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}) - g(\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \dots, \mathbf{y}^{(m)})| \\ &\leq \sum_{t=1}^m L_q^t \|\mathbf{x}^{(t)} - \mathbf{y}^{(t)}\|_p, \end{aligned}$$

where $L_q^t = \max \{ \|\nabla_t g(\mathbf{X})\|_q : \mathbf{x}^{(k)} \in \mathbb{B}_p(\mathbf{x}_0^{(k)}, \epsilon_0), \forall k \in [m] \}$ and $\nabla_t g(\mathbf{X}) = (\frac{\partial g(\mathbf{X})}{\partial \mathbf{x}_1^{(t)}}, \dots, \frac{\partial g(\mathbf{X})}{\partial \mathbf{x}_n^{(t)}})^T$. We then have

$$|g(\mathbf{X}) - g(\mathbf{X}_0)| \leq \sum_{t=1}^m L_q^t \|\mathbf{x}^{(t)} - \mathbf{x}_0^t\|_p,$$

which can be rearranged into the following form

$$g(\mathbf{X}_0) - \sum_{t=1}^m L_q^t \|\mathbf{x}^{(t)} - \mathbf{x}_0^t\|_p \leq g(\mathbf{X}) \leq g(\mathbf{X}_0) + \sum_{t=1}^m L_q^t \|\mathbf{x}^{(t)} - \mathbf{x}_0^t\|_p.$$

Recalling that $g(\mathbf{X}) = 0$ when $F_j(\mathbf{X}) = F_i(\mathbf{X})$, which indicates an adversarial example is found. That is to say, we want to ensure $g(\mathbf{X}) \geq 0$. This is assured by utilizing its lower bound:

$$g(\mathbf{X}_0) - \sum_{t=1}^m L_q^t \|\mathbf{x}^{(t)} - \mathbf{x}_0^t\|_p \geq 0 \Rightarrow \epsilon = \|\mathbf{x}^{(t)} - \mathbf{x}_0^t\|_p \leq \frac{g(\mathbf{X}_0)}{\sum_{t=1}^m L_q^t}.$$

In summary, CLEVER-Ada score is given as $\epsilon = \min \{ \frac{g(\mathbf{X}_0)}{\sum_{t=1}^m L_q^t}, \epsilon_0 \}$.

B.3 Adapted C&W Attack

In the original C&W attack, the loss function is designed with a loss function as $loss = c * q + \|\mathbf{X} - \mathbf{X}_0\|_p$, where $q = \max(\max_{j \neq i}(F_j(\mathbf{X}) - F_i(\mathbf{X})), 0)$ and c is a positive constant. In practice, we use binary search procedure to find a c such that an attack is obtained with a small magnitude. In essence, this corresponds to finding a targeted attack for class i , and the perturbation is measured by computing the norm of difference between \mathbf{X} and \mathbf{X}_0 . Here, we adapt the loss function to be $loss = q + (v + c) * \max_k \|\mathbf{x}^{(k)} - \mathbf{x}_0^{(k)}\|_p$, where $q = \max(F_j(\mathbf{X}) - F_i(\mathbf{X}), 0)$ and v is defined by indicator function $\mathbb{1}_{q \leq 0}(q)$. In our formulation, the c is a pre-defined small constant, which does not update during the process. More concretely, C&W-Ada puts higher weights on minimizing the q until an attack is found, and prioritizes the minimization of distortion magnitude when an attack has been found ($q = 0, v = 1$). We use the maximum perturbation of all frames as the C&W score.

B.4 Experimental Details

We evaluate POPQORN and other baselines on vanilla RNNs and LSTMs trained on MNIST dataset, MNIST sequence dataset, and Classifier for the question classification (UIUC’s CogComp QC) Dataset. We train vanilla RNN and LSTM models on MNIST dataset, and evaluate by randomly selecting 1000 test images as samples. For these 1000 test images, we randomly generate a set of adversarial target labels, and compute their average targeted POPQORN bounds (the true label output is only guaranteed to be bigger than the target label output) and standard deviation, CLEVER-RNN scores, and compare them with the magnitude of attacks found by C&W-Ada. Different from MNIST dataset, MNIST sequence dataset records handwritten numbers as sequential data of line segment sequences. Due to the poor performance of vanilla RNNs on MNIST sequence dataset, only LSTM classifiers are trained for robustness evaluations. In this example, we compute the untargeted POPQORN bound on only one single stroke of a digit sequence. After calculating bounds of all strokes, we can identify the strokes with minimal bounds in a digit sequence. We call the strokes *sensitive strokes*. Question classification task is to classify a question into several pre-defined categories. For example, “What is a mirror made out of?” corresponds to “ENTY” (entity), “What is the temperature at the center of the earth?” corresponds to “NUM” (number), “What is the abbreviation for Texas?” corresponds to “ABBR” (abbreviation), “Who discovered x-rays?” corresponds to “HUM” (human), “What is the capital of Yugoslavia?” corresponds to “LOC” (location), and “What is caffeine?” corresponds to “DESC” (description). In this example, we also compute the untargeted POPQORN bound on one single input frame, and call the words with minimal bounds *sensitive words*.

Word embedding is used in question classifications. In the task of certifying bounds for question classification networks, we study how much we can perturb a single frame in a sequence while not causing any misclassification. However, different from the former tasks, the space of words is discrete. Thus instead of dealing with words in a discrete space, we evaluate their word embedding. Specifically, we use pretrained word embedding from “glove.6B.100d”³.

Additional NLP Experiment: News Title Classification. We further exemplify the usefulness of POPQORN by identifying key words (quantifying robustness) for TagMyNews [1] LSTMs. TagMyNews is a dataset consisting of 32,567 English news items grouped into 7 categories: Sport, Business, U.S., Health, Sci&Tech, World, and Entertainment. Each news item has a news title and a short description. We train an LSTM to classify the news items into the 7 categories according to their news titles. Then we use POPQORN to identify keywords in the news titles by selecting the words with the smallest POPQORN robustness quantification. Three examples are given in Figure 3, where we can see that the keywords (underlined) identified by POPQORN are indeed more closely tied to the category of each news.

Example	<u>Samsung</u>	to	launch	<u>galaxy</u>	<u>s</u>	sequel	In	south	korea	in	late	april
Sci&Tech	0.4219	0.7344	0.5469	0.4648	0.5234	0.5664	0.8047	0.6641	0.6680	0.8516	0.7188	0.8125

Example	3	<u>journalists</u>	<u>kidnapped</u>	in	<u>afghanistan</u>	are	set	free
World	0.4492	0.4258	0.4219	0.7305	0.3945	0.6484	0.5977	0.5547

Example	<u>actress</u>	<u>evan</u>	<u>rachel</u>	wood	dates	both	men	and	women
Entertainment	0.4062	0.3906	0.4609	0.4805	0.6211	0.6875	0.9688	1.0625	0.8672

Figure 3: Three examples in the news title classification task. The upper row gives the sample sentence; the lower row shows the POPQORN (2-norm) lower bounds of individual words. “Sci&Tech”, “World”, and “Entertainment” are the corresponding categories these news belong to.

³<https://nlp.stanford.edu/projects/glove/>

References

- [1] D. Vitale, P. Ferragina, and U. Scaiella. Classification of short texts by deploying topical annotations. In *Advances in Information Retrieval*, pages 376–387, 2012.