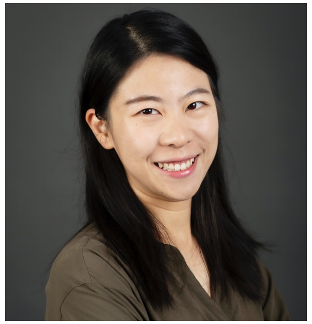
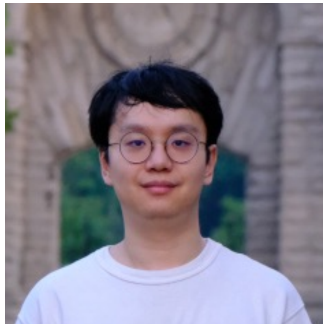
NEURAL INFORMATION PROCESSING SYSTEMS

Training Certifiably Robust Neural Networks with Efficient Local Lipschitz Bounds

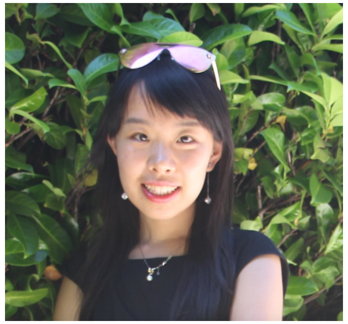
Yujia Huang, Huan Zhang, Yuanyuan Shi, J Zico Kolter, Anima Anandkumar



Caltech



CMU

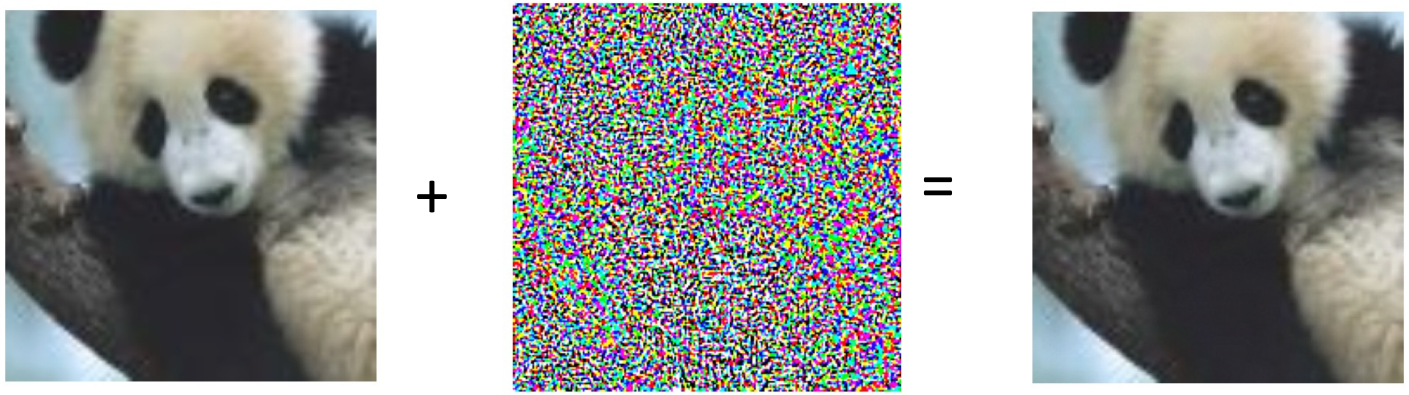


UCSD

CMU & Bosch

Caltech & NVIDIA

Adversarial Robustness

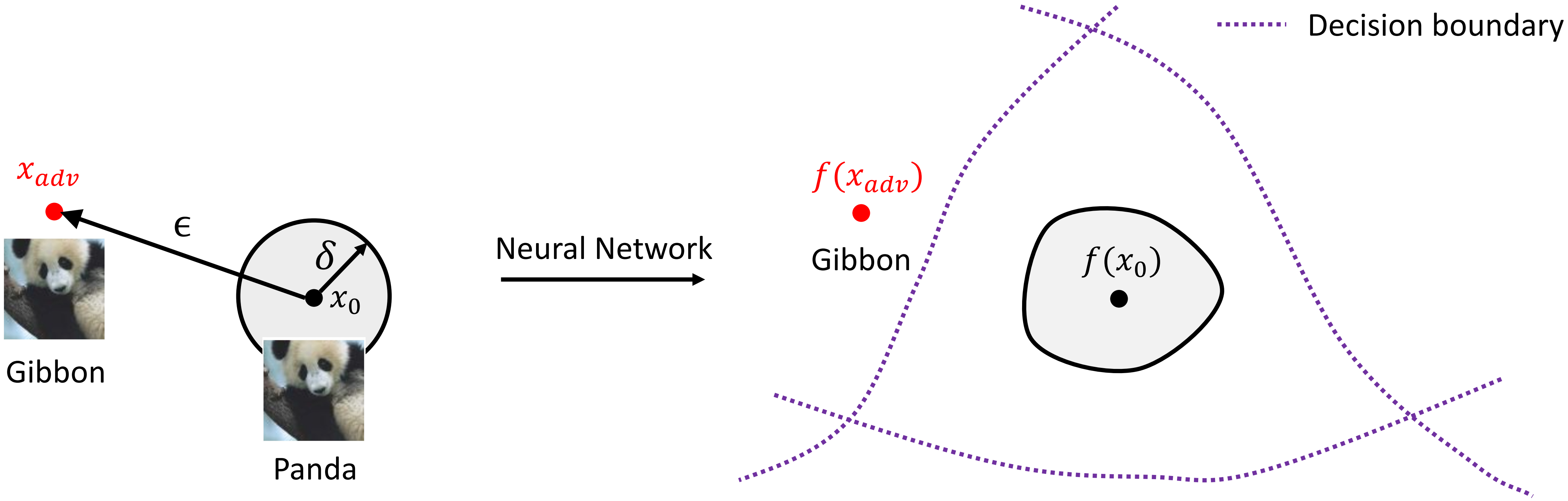


Clean image x_0 + Perturbation ϵ = Adversarial image x_{adv}

Panda (57.7% confidence) Gibbon (99.3% confidence)

[Goodfellow et. al., ICLR 2015]

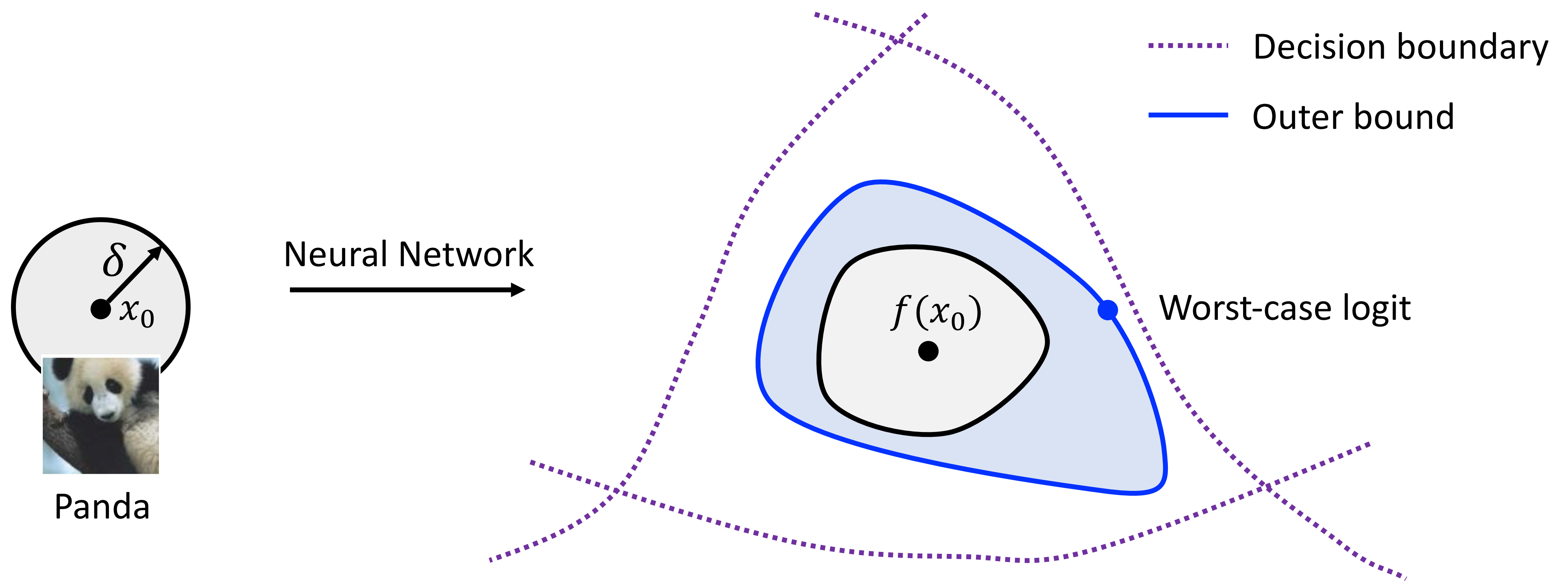
Certified Robustness



Certified Robustness

For $\forall x$ such that $\|x - x_0\|_p \leq \delta$, the neural network f outputs the same class.

Certifiable training

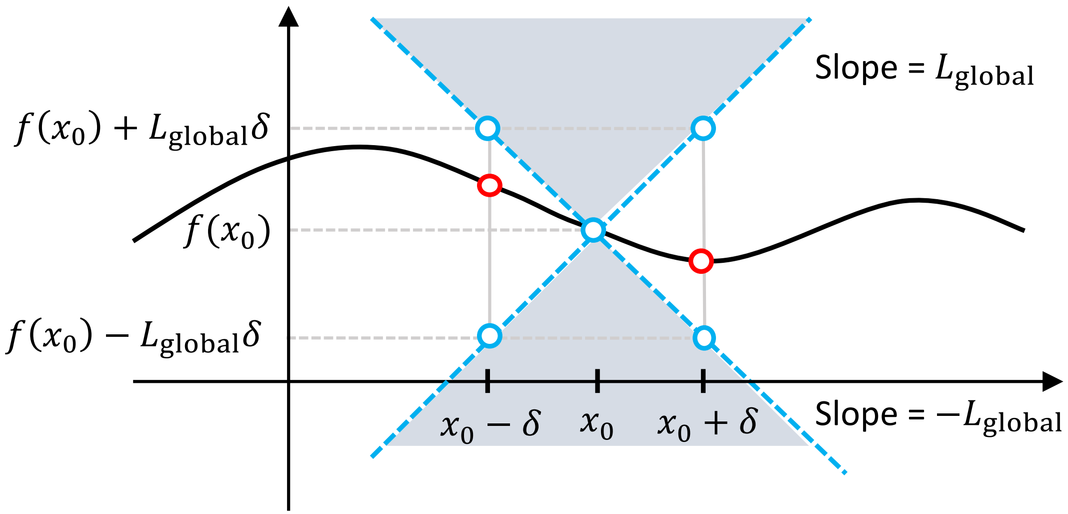


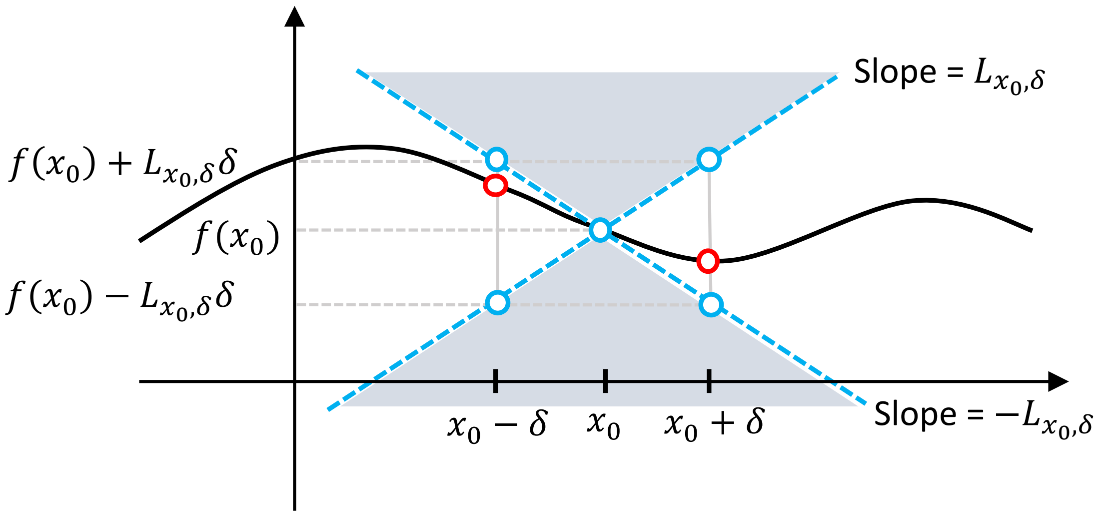
- Bound the neural network output given input perturbation
- Compute the worst-case logit over the bounded output region
- Train with the worst-case logit (using worst-case logit to replace normal logit in cross-entropy loss)

Global v.s. Local Lipschitz constant

Definition: Function $f(x)$ satisfies a Lipschitz condition over a set D if a constant $L > 0$ exists with $|f(x_1) - f(x_2)| \leq L|x_1 - x_2|, \forall x_1, x_2 \in D$. L_D is the Lipschitz constant over set D.

- If $D = \text{Domain}(f)$, L_D is called **global** Lipschitz constant.
- If $D = \{x \mid |x - x_0| \leq \delta\}$, L_D is called **local** Lipschitz constant.



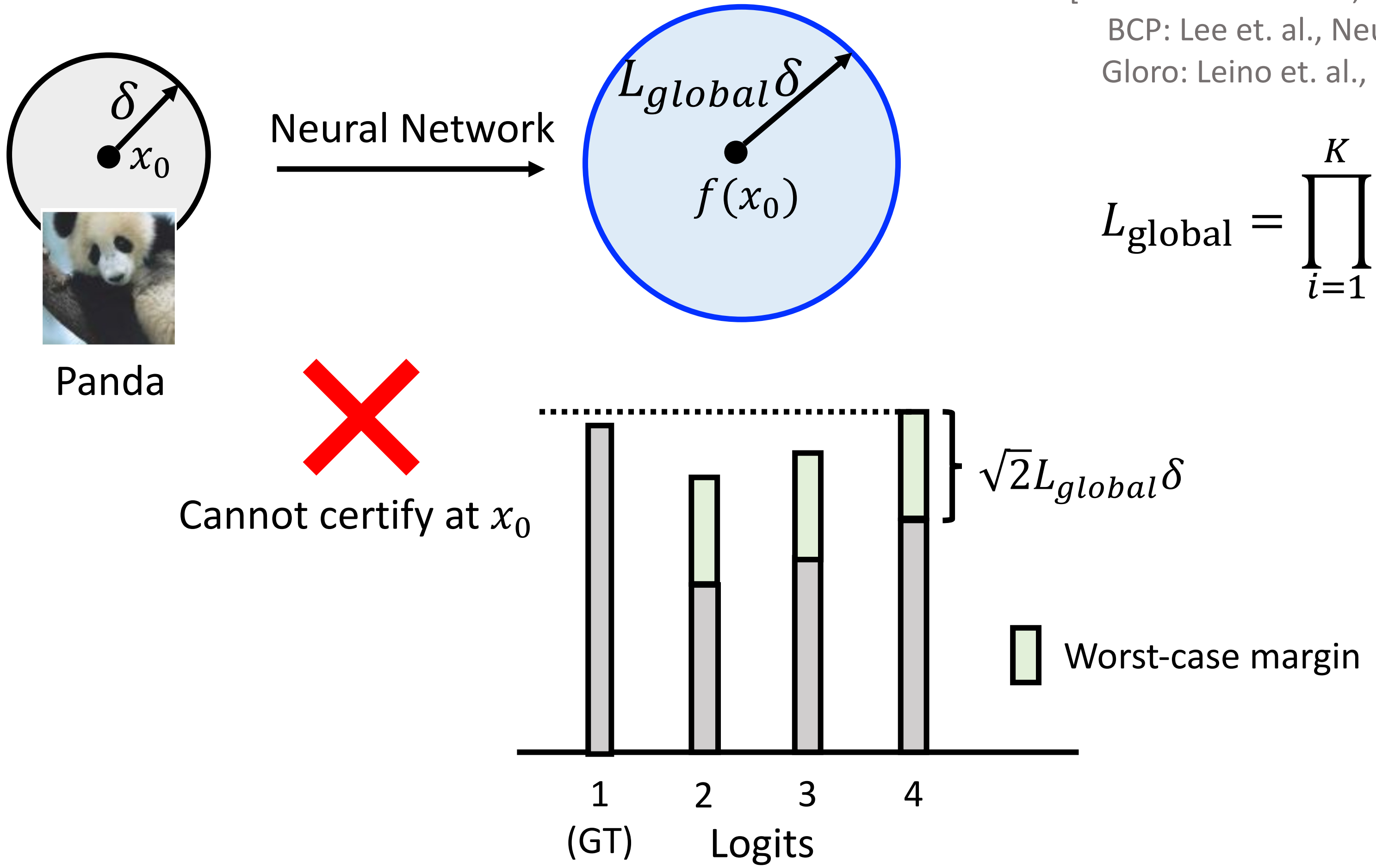


Local Lipschitz constant $\leq$ Global Lipschitz constant

Certified Defenses via Global Lipschitz bound

$$x \rightarrow W_1 \rightarrow \text{ReLU} \rightarrow W_2 \;\cdots\cdots\; \rightarrow \text{ReLU} \rightarrow W_K \rightarrow y$$

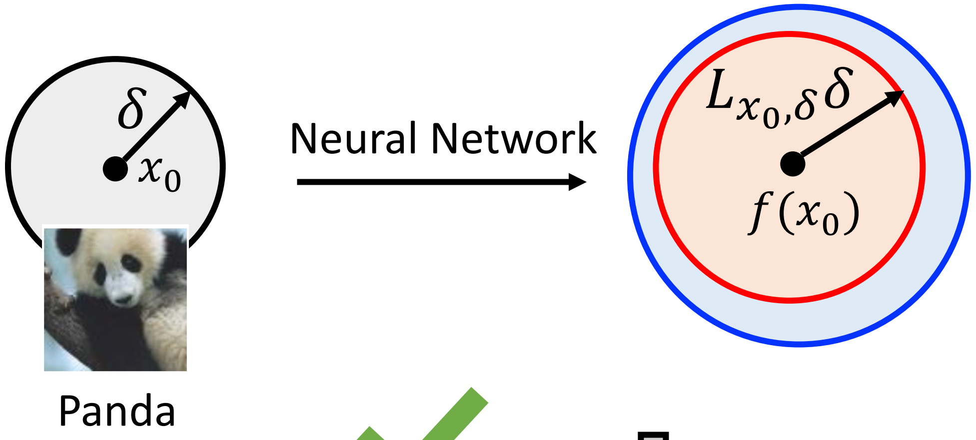
[LMT: Tsuzuku et. al., NeurIPS 2018,
BCP: Lee et. al., NeurIPS 2020,
Gloro: Leino et. al., ICML 2021]



$$L_{\text{global}} = \prod_{i=1}^{K} \|W_i\|_2$$

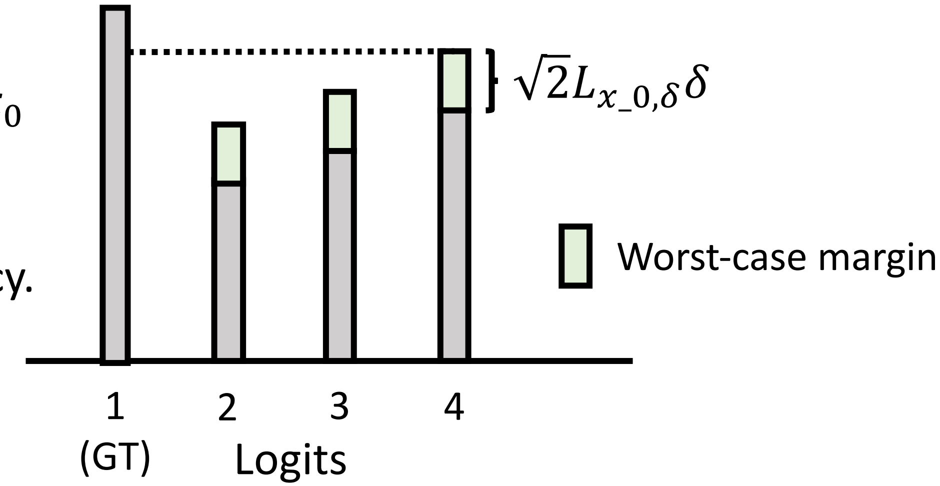
Certified Defenses via Local Lipschitz bound

$x \rightarrow W_1 \rightarrow \text{ReLU} \rightarrow W_2 \;\cdots\cdots \rightarrow \text{ReLU} \rightarrow W_K \rightarrow y$



$L_{x_0,\delta}$: Local Lipschitz bound
Varies for each data point

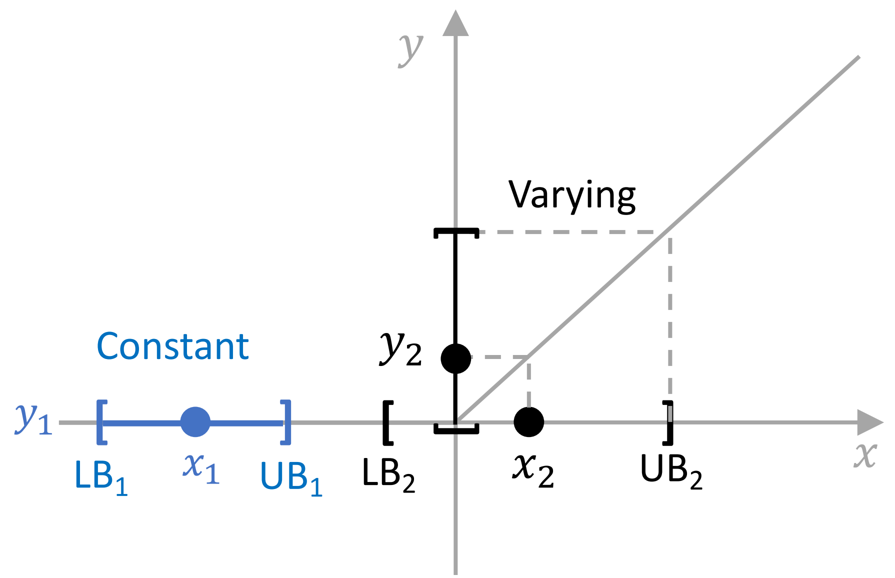
Can certify at x_0



- Local Lipchitz bound gives better certified accuracy.
- **None of existing approaches** [LMT, BCP, Gloro, etc.] **use local Lipschitz bound in their training!**

Our approach: An Efficient Local Lipschitz Bound

ReLU outputs under perturbation



$$L_{\mathrm{local}}(x) = \left\|W^L I_V^{L-1}\right\|_2 \left\|I_V^{L-1} W^{L-1} I_V^{L-1}\right\|_2 \dots \left\|I_V^1 W^1\right\|_2$$

Input	Output
$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$	$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathrm{ReLU}\left(\begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$

$\|x' - x\| \leq \delta$ $\Big\downarrow$ perturb

$$\begin{bmatrix} \mathrm{LB}_1 \leq x_1' \leq \mathrm{UB}_1 \leq 0 \\ \mathrm{LB}_2 \leq x_2' \leq \mathrm{UB}_2 \end{bmatrix} \qquad \begin{bmatrix} 0 \\ y_2' \end{bmatrix} = \mathrm{ReLU}\left(\begin{bmatrix} \cancel{W_{11}} & \cancel{W_{12}} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \end{bmatrix}\right)$$

Global Lipschitz bound	$\|\Delta y\| \leq \left\|\begin{matrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{matrix}\right\| \|\Delta x\|$
Local Lipschitz bound at x	$\|\Delta y\| \leq \|W_{21} \quad W_{22}\| \|\Delta x\|$

Provable tightness of our Local Lipschitz Bound

- **Global** Lipschitz bound: $L_{\text{global}} = \prod_{i=1}^{K} \|W_i\|_2$ (1)
- **Local** Lipschitz bound: $L_{\text{local}}(x) = \left\|W^L I_V^{L-1}\right\|_2 \left\|I_V^{L-1} W^{L-1} I_V^{L-1}\right\|_2 \dots \left\|I_V^1 W^1\right\|_2$ (2)

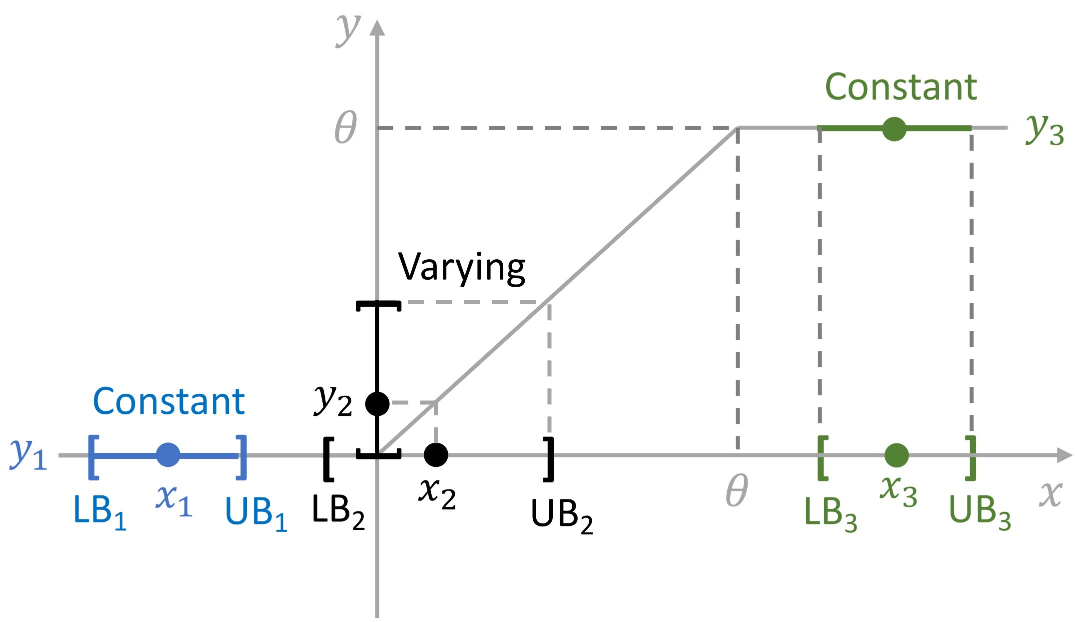
I_V: Indicator matrix for varying ReLU outputs

Theorem: For any x and L-layer ReLU neural network, the local Lipchitz bound calculated via (2) is no larger than the global Lipschitz bound in (1), i.e.

$$\boldsymbol{L_{local}(x) \leq L_{global}}$$

A new activation function for tighter local Lipschitz bound

ReLUθ outputs under perturbation



$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \mathrm{ReLU}\left(\begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right)$$

$$\downarrow$$

$$\begin{bmatrix} 0 \\ y_2' \\ \theta \end{bmatrix} = \mathrm{ReLU}\left(\begin{bmatrix} \cancel{W_{11}} & \cancel{W_{12}} & \cancel{W_{13}} \\ W_{21} & W_{22} & W_{23} \\ \cancel{W_{31}} & \cancel{W_{32}} & \cancel{W_{33}} \end{bmatrix} \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} \right) + \begin{pmatrix} 0 \\ 0 \\ \theta \end{pmatrix}$$

Local Lipschitz bound at x $\quad \|\Delta y\| \le \|W_{21} \quad W_{22} \quad W_{23}\|\|\Delta x\|$

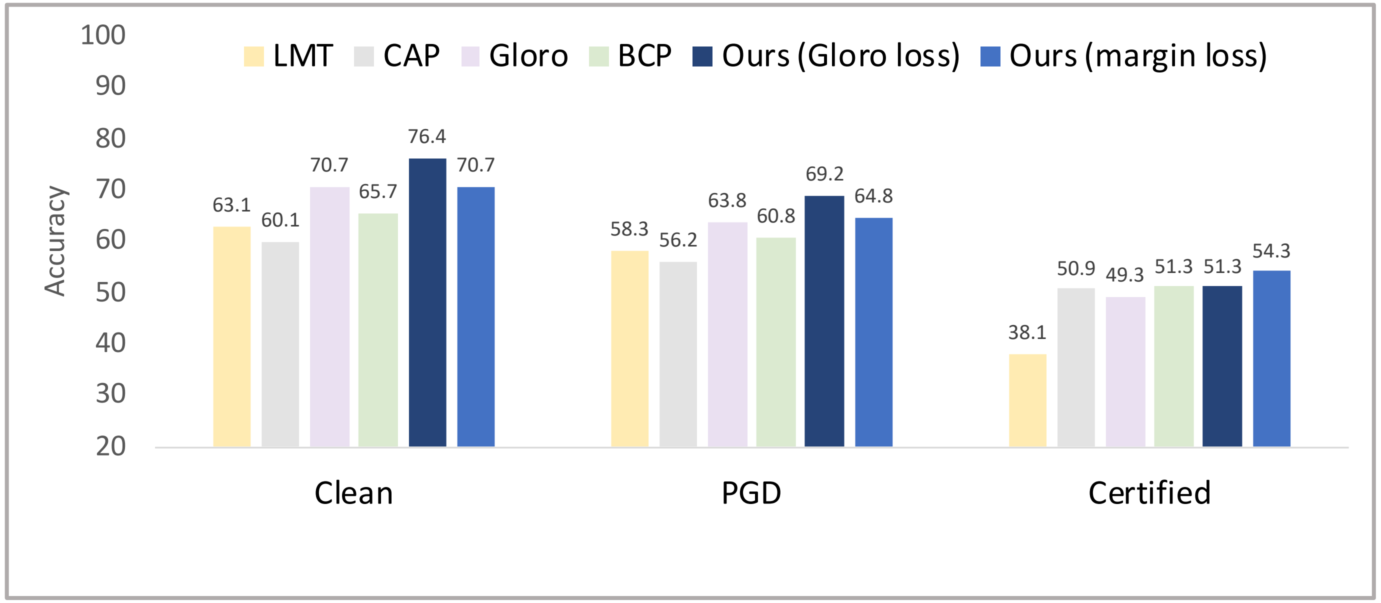
- This trick can be applied to other activation functions such as MaxMin [Anil et. al., ICML 2019].

Certified Robustness

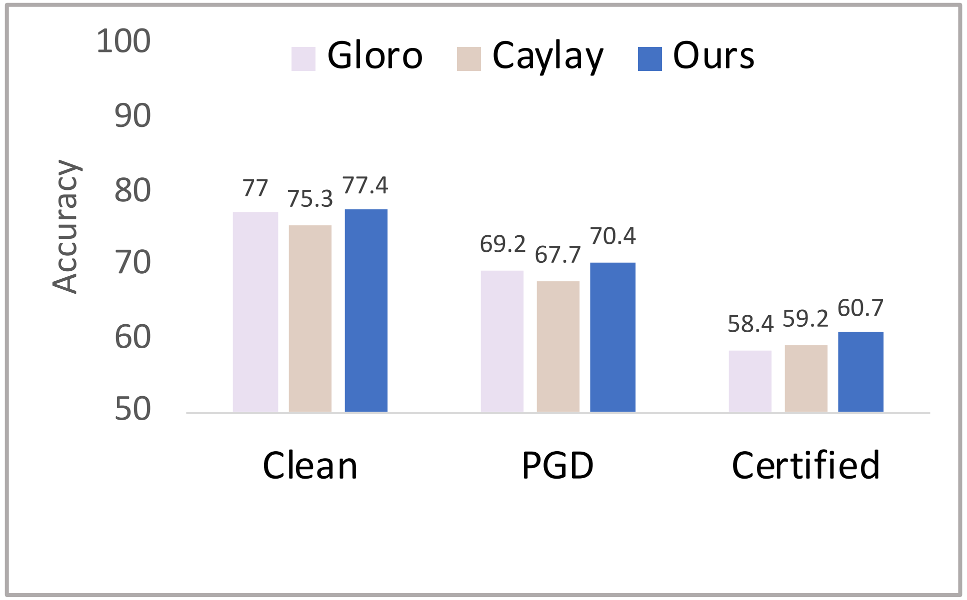
Our method (Local Lipschitz bound) outperforms state-of-the-art methods

- On Various datasets: MNIST, CIFAR-10 and Tiny-imagenet
- With different activation functions: ReLU or clipped MaxMin

CIFAR-10, ReLU activations, $\epsilon = 36/255$



CIFAR-10, MaxMin activations, $\epsilon = 36/255$

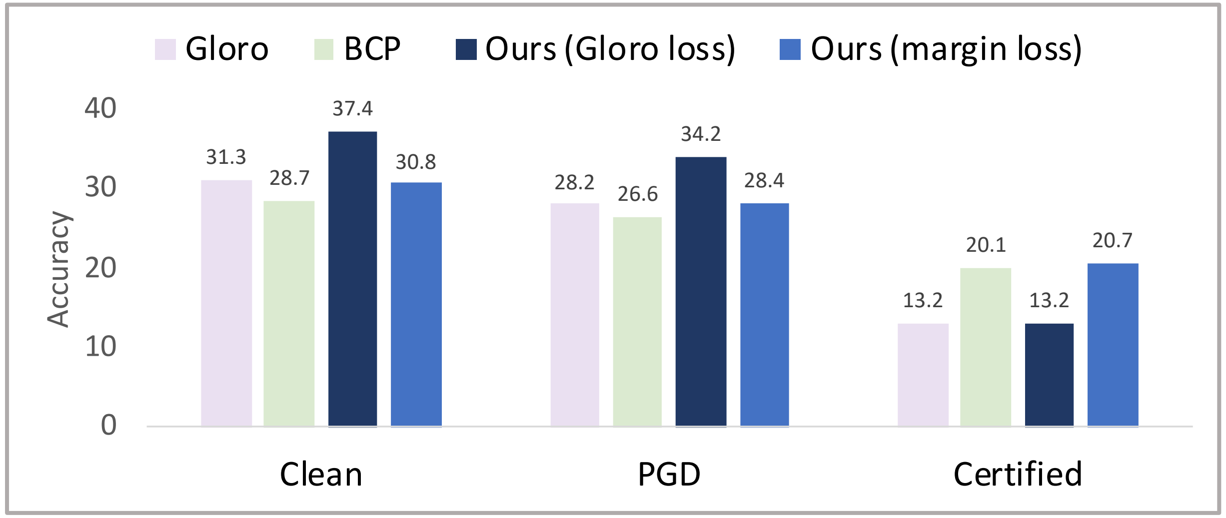


Certified Robustness - continued

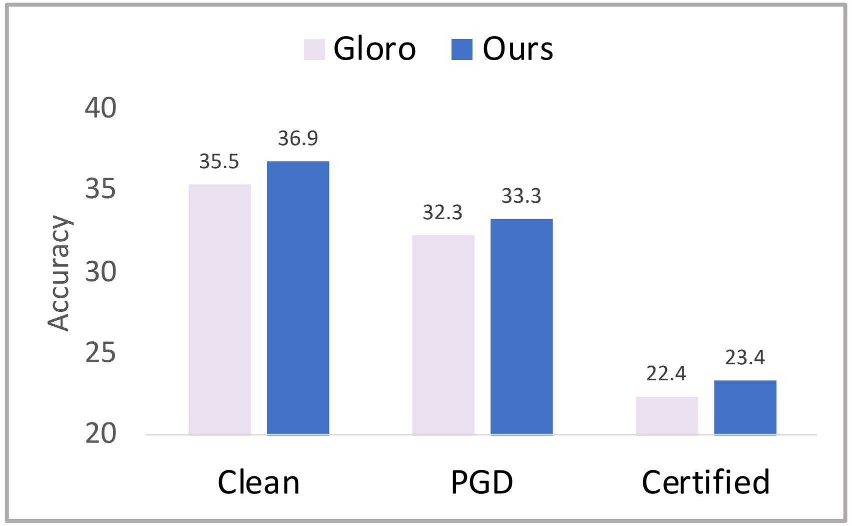
Our method (Local Lipschitz bound) outperforms state-of-the-art methods

- On Various datasets: MNIST, CIFAR-10 and Tiny-imagenet
- With different activation functions: ReLU or clipped MaxMin

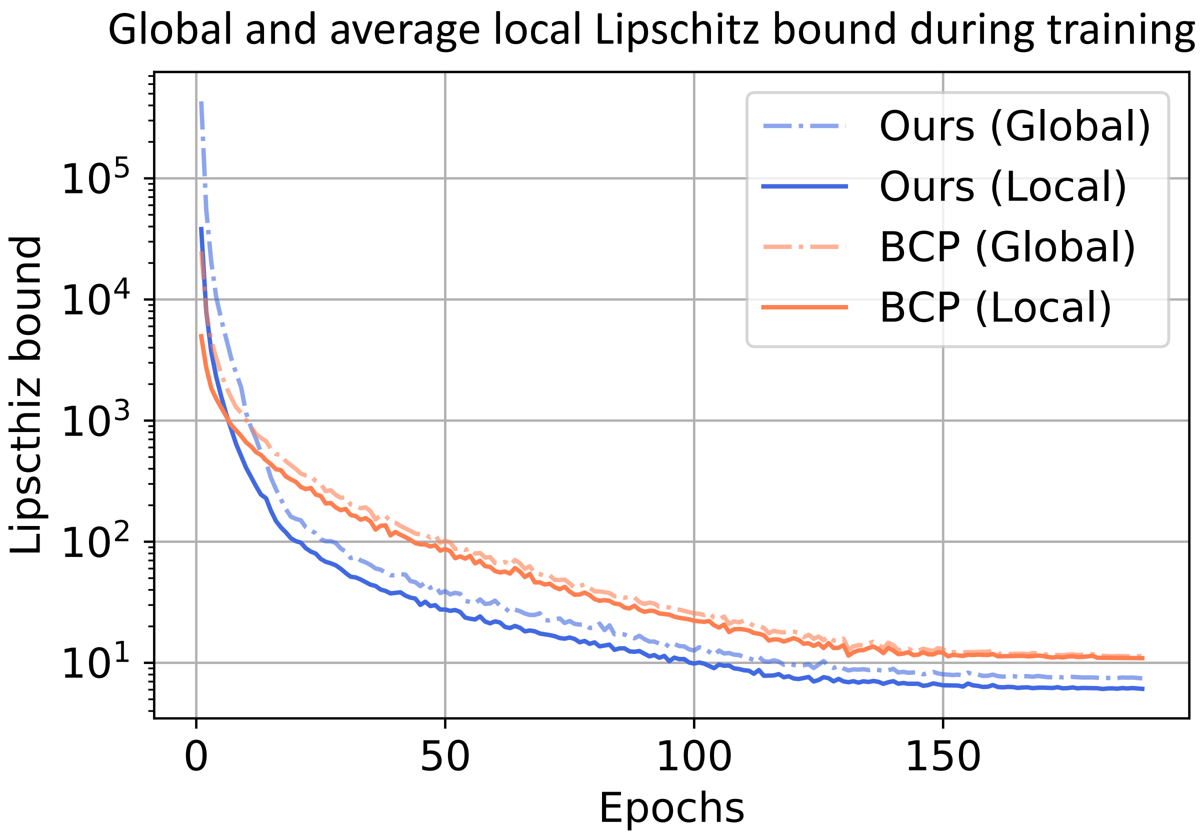
Tiny-Imagenet, ReLU activations, $\epsilon = 36/255$



Tiny-Imagenet, MaxMin activations, $\epsilon = 36/255$

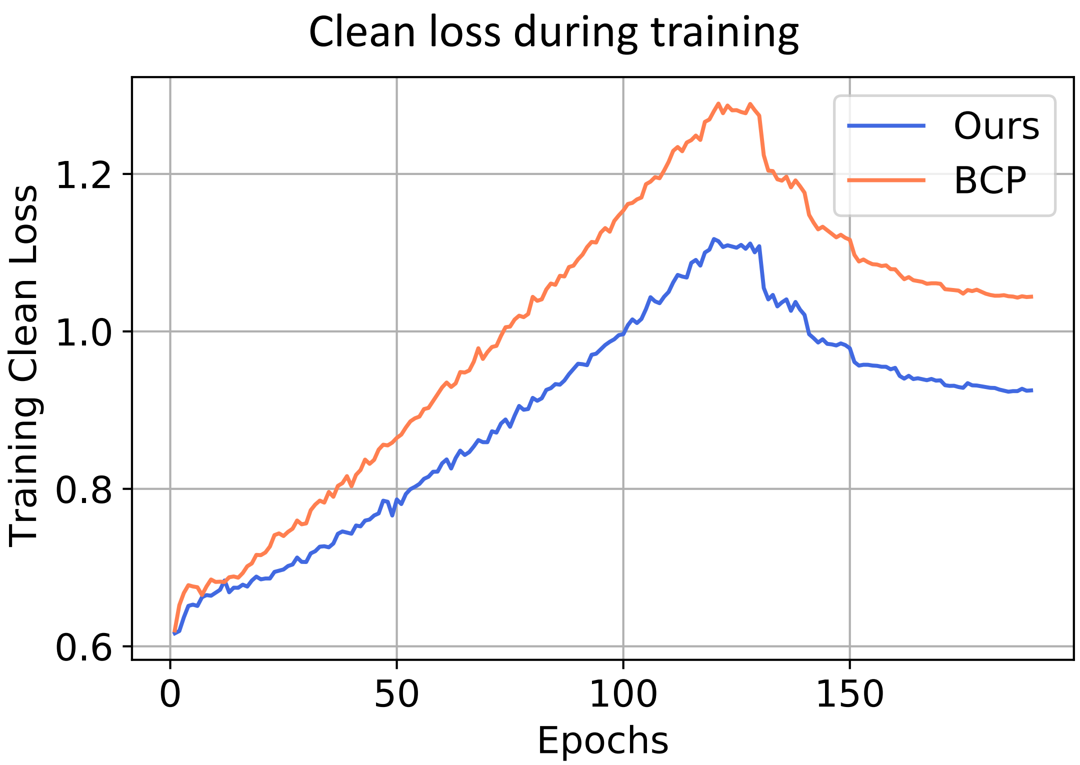


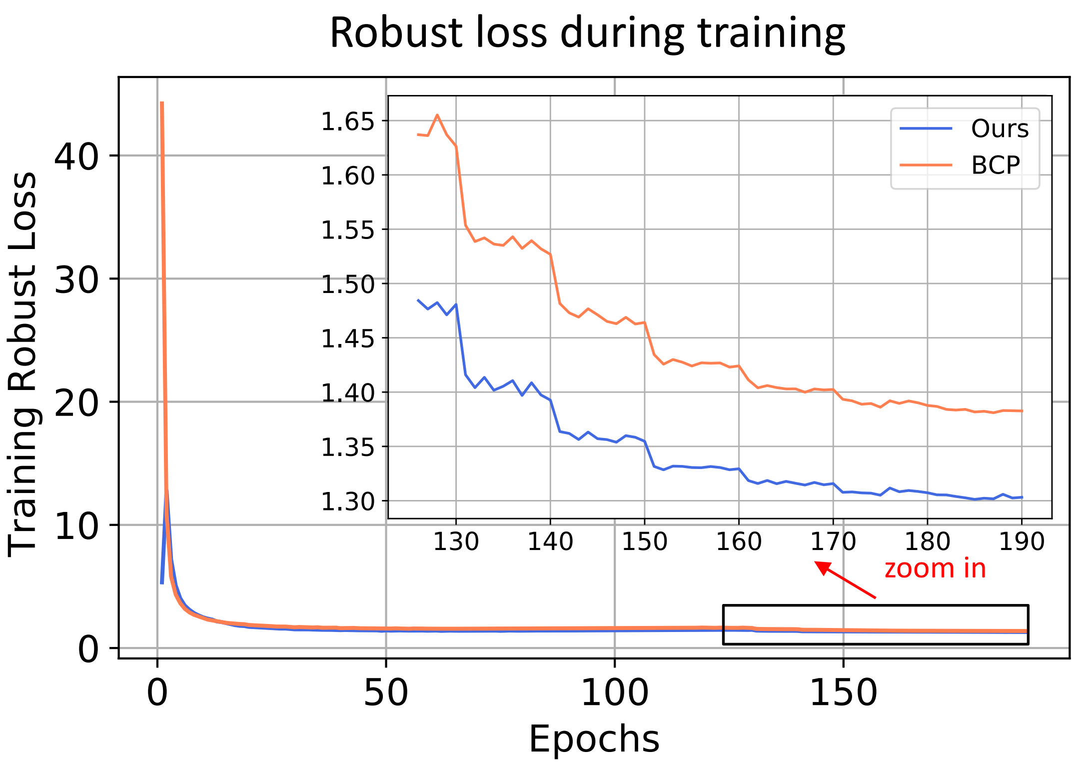
Tightness of our local Lipschitz bound



- Our local Lipschitz bound is always tighter than the global Lipschitz bound.
- Directly applying our bound on a trained network has much less improvement (red curves).
- **It is crucial to incorporate Local Lipschitz bound during training.**

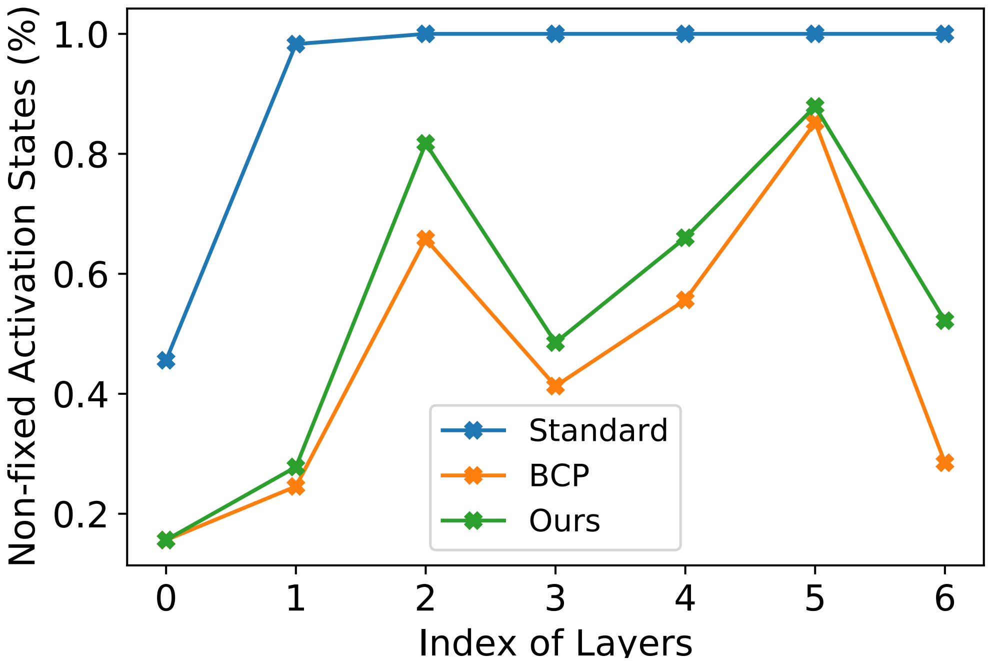
Better clean loss and robust loss





- Larger global Lipschitz bound in the beginning of training.
- Larger model capacity and easier training in the early stage.
- **Improvement of both clean loss and robust loss.**

Sparsity of varying ReLU outputs



- Dense varying ReLU neurons leads to better clean accuracy but worse robustness.
- **Incorporate local Lipschitz bound during training to allow for denser varying ReLU neurons without hurting robustness.**

Summary

- We propose an **efficient and trainable** Local Lipschitz bound.
- The proposed local bound is **provably tighter** than the global Lipschitz bound.
- Our method outperforms state-of-the-art methods on L2 certified robustness.

Our code is available at https://github.com/yjhuangcd/local-lipschitz.

Thank You!