Lipschitz-Certifiable Training with a Tight Outer Bound

Sungyoon Lee*, Jaewook Lee*, Saerom Park**

* Statistical Learning & Computational Finace Lab

Seoul National University

** Department of Convergence Security Engineering

Sungshin Women's University

NEURAL INFORMATION PROCESSING SYSTEMS

Background and Motivation

- Deep neural networks are **vulnerable** to **small but adversarially designed perturbations** in the input which can mislead a network **to predict wrong label**.
- There are two different defense approaches:
  - **Heuristic Defense**: be designed to defend against specific predefined adversarial attacks → can be defeated by unseen stronger **adaptive adversaries** [Tra+20].
  - **Certified Defense**: minimize an upper bound on the worst-case logit over all possible perturbations within a given noise level [TSS18, Won+18, Gow+18].
- We build **a robust model** through **certifiable training method.**

Problem

- Certified Defenses minimize the following upper bound on the worst-case loss:

$$\max_{x' \in \mathbb{B}(x)} \mathcal{L}(z(x'), y) \le \mathcal{L}(\underline{z}(x), y)$$

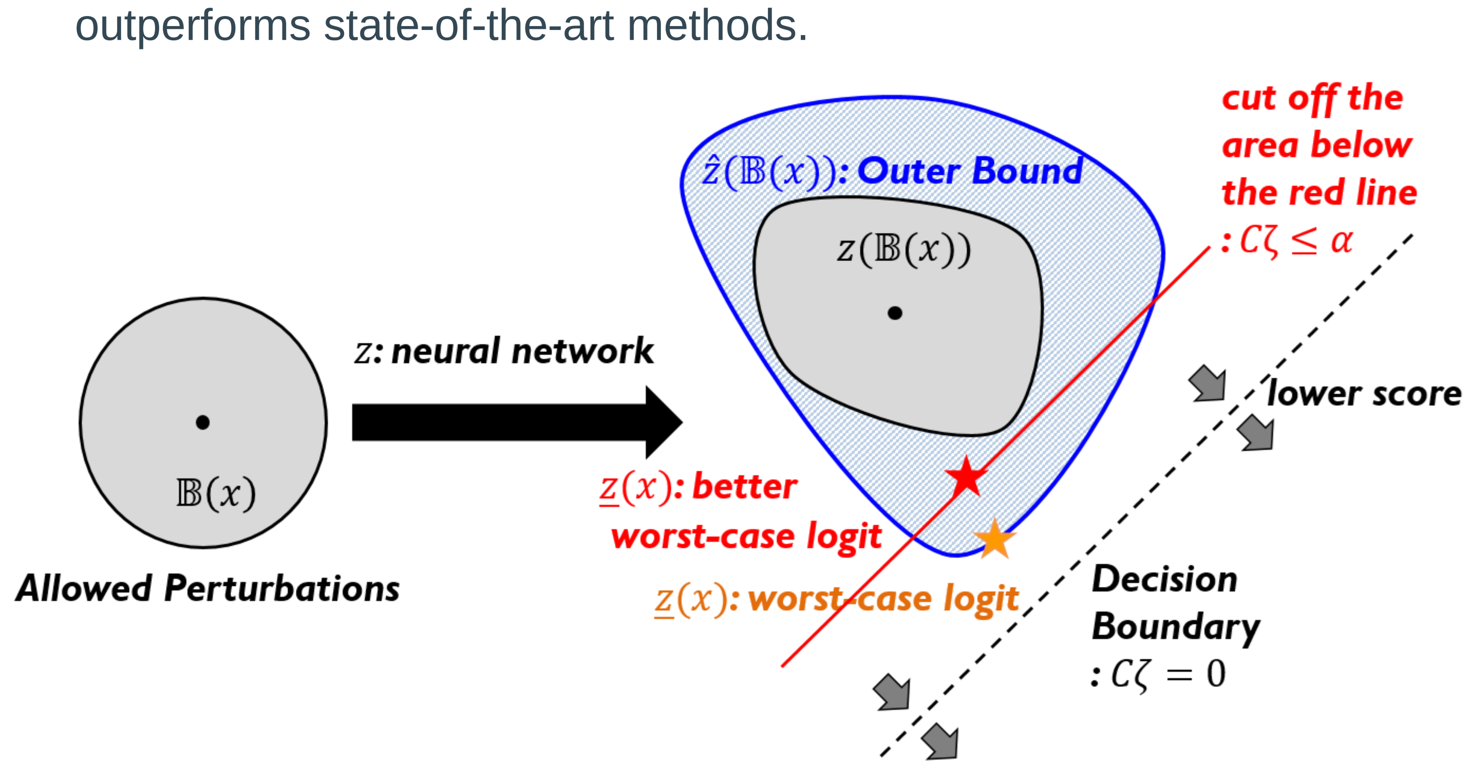
with **a worst-case logit** $\underline{z}(x) = \arg\min_{\zeta \in \hat{z}(\mathbb{B})} C\zeta$ where $C = \mathbf{1}e^{(y)T} - I$

over **an outer bound** $\hat{z}(\mathbb{B})$ in the logit space.

- Certified Defenses with worst-case logit consists of two stages:
  1. Propagate the input perturbation through the network and obtain an outer bound of the image → **the tightness of the outer bound**
  2. Solve an optimization problem to compute the worst-cae logit over the outer bound → **the efficient computation**
- We need an **efficient** certifiable training method with the **tight outer bound**.

Main Contribution

- **Efficiency**: We propose a fast certified defense method called Box Constraint Propagation (BCP).
- **Tightness**: By introducing an additional box constraint, we can obtain a tighter upper bound to be minimized.
- **Expressiveness/Robustness**: We can build a certifiably robust model outperforms state-of-the-art methods.



Proposed Method

Intuition Behind the Design

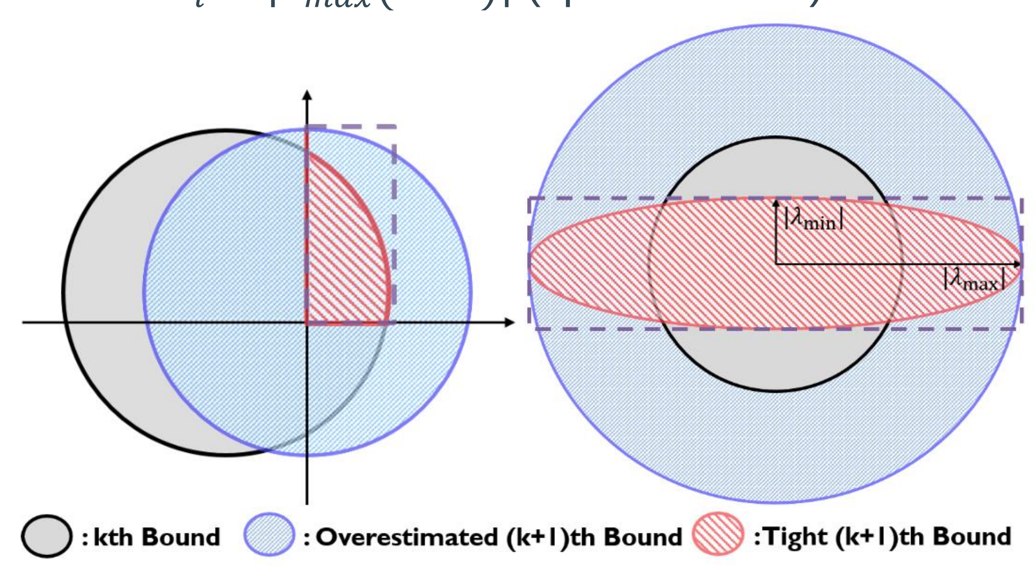
- LMT [TSS18]: Lipchitz outer bound is simple and efficient but **overestimates** the true image of the allowed perturbations:

  $\hat{z}(\mathbb{B}(x)) = \mathbb{B}_2(z(x), \epsilon L)$ with the Lipschitz constant L.

  ► Our proposed method aims to obtain a **tighter outer bound.**
- Our proposed method is effective in both nonlinear (ReLU) and linear operations.

  k-th Bound → (k+1)-th Bound (Tight bound ⊂ Overestimated bound)

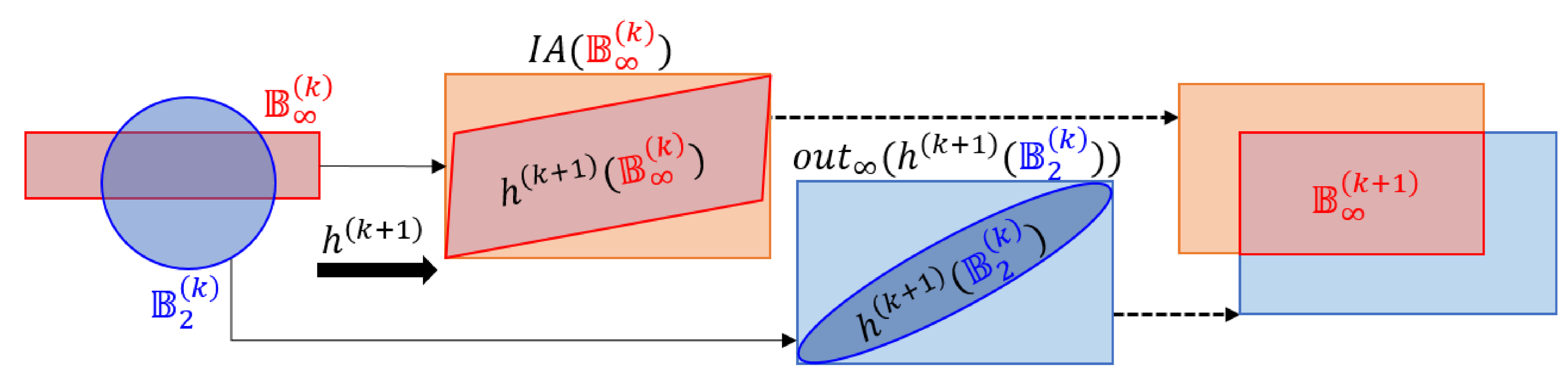
  ► Consider **element-wise bound (=Box Constraint)** propagation
  - Nonlinear operation (ReLU): $L_i = 1 \ge \frac{u^+}{u^+ - l^-}$
  - Linear operation: $L_i = |\lambda_{max}(W^{(i)})|$ (spectral norm)



Box Constraint Propagation (BCP)

- We introduce additional "Box Constraints" to further tighten the outer bound:
  - ① Circumscribing box of the propagated ℓ_2 ball $h^{(k+1)}\left(\mathbb{B}_2^{(k)}\right)$: $out_\infty(h^{(k+1)}\left(\mathbb{B}_2^{(k)}\right))$
  - ② Circumscribing box of the propagated ℓ_∞ box $h^{(k+1)}\left(\mathbb{B}_\infty^{(k)}\right)$: $IA\left(\mathbb{B}_\infty^{(k)}\right)$

  ► The (k+1)th box outer bound: $\mathbb{B}_\infty^{(k+1)}$ = ① ∩ ② = $out_\infty(h^{(k+1)}\left(\mathbb{B}_2^{(k)}\right)) \cap IA\left(\mathbb{B}_\infty^{(k)}\right)$



- Our certifiable training algorithm minimizes the following objective:
  - Training objective: $\mathcal{J}(f, \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathcal{L}(\underline{z}(x), y)]$
  - The worst-case logit: $\min_{\zeta'} c^T\zeta'$ s.t. $\left\|\zeta' - \mathbf{z}^{(K-1)}\right\|_2 \le \rho^{(K-1)}$,

    $\left|\zeta' - m^{(K-1)}\right| \le r^{K-1}$.
- To compute the worst-case logit, we propose an efficient algorithm (**Algorithm 1**) that terminates in finite iterations.

➤ **Theorem (Efficient Computation)** The while loop in Algorithm 1 finds the optimal solution ζ^* of the optimization problem $\min_{\zeta \in \mathbb{B}_2 \cap \mathbb{B}_\infty} c^T\zeta$ in a finite number of iterative steps less than the number of elements in c.

Algorithm 1 Box Constraint Propagation (BCP) Certifiable Training

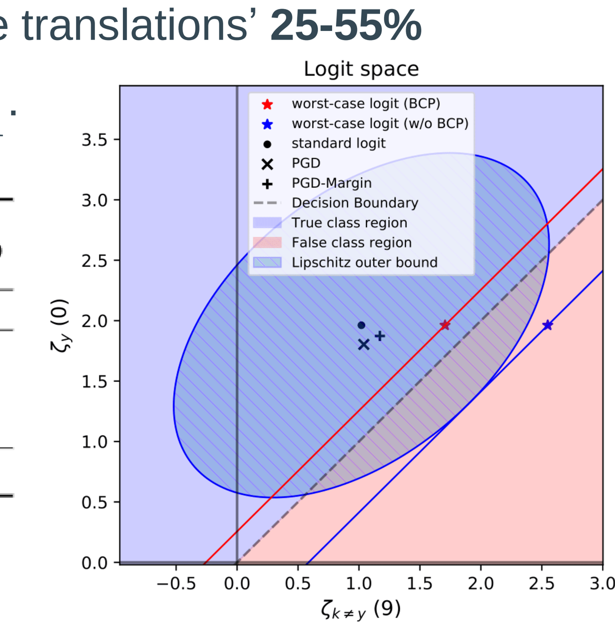
---

Input: training data $(x, y) \sim \mathcal{D}$, target perturbation size ϵ_{target}, network parameterized by θ
Output: Robust network f_θ
repeat
  Read mini-batch B from $\mathcal{D}$ and adjust ϵ and λ according to the schedule.
  // Compute the box outer bound and the ball outer bound //
  $\mathbb{B}_\infty^{(K-1)} = \text{midrad}(m^{(K-1)}, r^{(K-1)})$ where $m^{(K-1)}, r^{(K-1)} = \text{BCP}(x, \epsilon; \theta)$ ((6)-(8)).
  $\mathbb{B}_2^{(K-1)} = \mathbb{B}_2(\mathbf{z}^{(K-1)}, \rho^{(K-1)})$ where $\mathbf{z}^{(K-1)} = h^{(K-1)} \circ \cdots \circ h^{(1)}(x)$ and $\rho^{(K-1)} = \epsilon \prod_{i=1}^{K-1} L^{(i)}$
  // Solve the optimization in (11) for each $m \neq y$ in parallel //
  Initialize $p = \mathbf{z}^{(K-1)} - \rho^{(K-1)} \frac{\mathbf{c}}{\|\mathbf{c}\|}$.
  while not $|p - m^{(K-1)}| \le r^{(K-1)}$ **do**
    Decompose p into two parts: $p = p[I] + p[I^c]$, where $I \equiv \{l : |p_l - m_l^{(K-1)}| \ge r_l^{(K-1)}\}$.
    First phase Project $p[I]$ onto $\mathbb{B}_\infty^{(K-1)}$.
    Second phase With the scaling parameter η in (12), update $p \leftarrow \Pi_{\mathbb{B}_\infty^{(K-1)}} p[I] + \eta p[I^c]$.
  end while
  Calculate the worst-translated logit $\underline{z}(x) = z(x) + t(x)$ with (2) and (10):
    $t_m(x) = \mathbf{c}^T(\mathbf{z}^{(K-1)} - p)$.
  // Update Parameters //
  Update the parameter θ with the objective (13):
    $\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{J}(f_\theta, B; \lambda)$.
until training phase ends

Experiments

Efficiency & Tightness

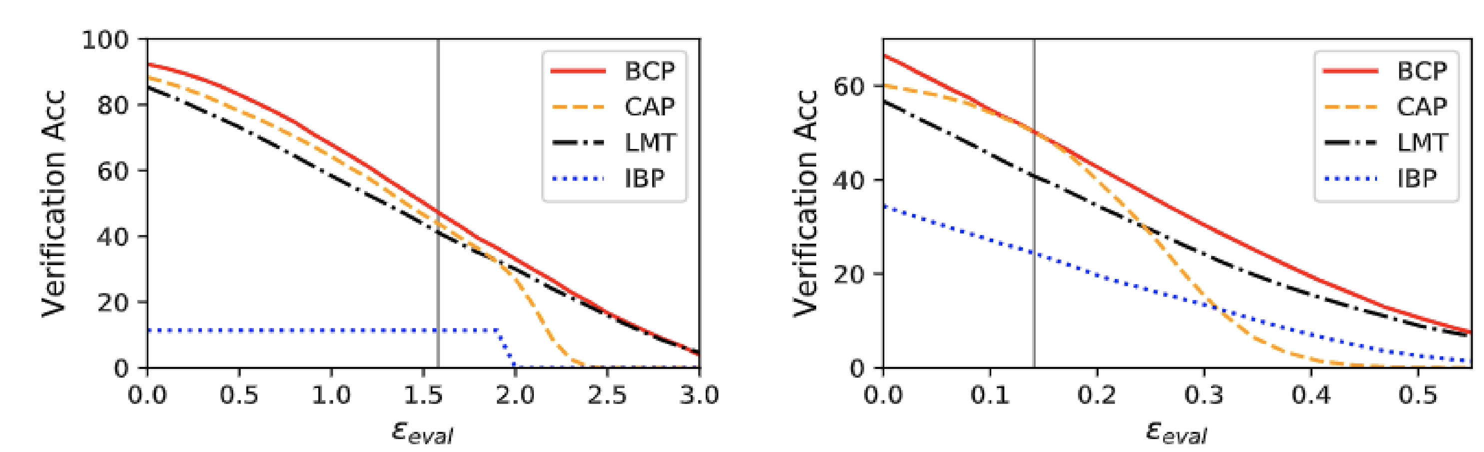
- Our proposed method is **over 12 times faster** than CAP [Won+18].
- The additional box constraint makes 'wost-case translations' **25-55% tighter** in terms of translation ∝ $\left\|\underline{z}(x) - z(x)\right\|_1$.



Data	Structure	Computation time (sec/epoch)		Speed up
		CAP	BCP	
MNIST	4C3F	689	**57.5**	×12.0
CIFAR-10	4C3F	645	**53.0**	×12.2
	6C2F	1,369	**56.5**	×24.2
	WRN	1,121 (2 GPUs)	**89.5**	×12.5
Tiny ImageNet	8C2F	-	**3,268**	-

Expressiveness/Robustness

- Our proposed method **outperforms state-of-the-art methods** (CAP [Won+18], LMT [TSS18], IBP [Gow+18]).



(a) MNIST (b) CIFAR-10 (36/255)

References

[Gow+18] S. Gowal, K. et al. (2018). "On the effectiveness of interval bound propagation for training verifiably robust models." In: arXivpreprint arXiv:1810.12715

[Tra+20] Florian Tramer et al. (2020). "On adaptive attacks to adversarial example defenses." In: arXiv preprint arXiv:2002.08347

[TSS18] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. (2018) "Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks". In: Advances in Neural Information Processing Systems 31. pp. 6541–6550

[Won+18] Eric Wong et al. (2018). "Scaling provable adversarial defenses". In: Advances in Neural Information Processing Systems 31. pp. 8400–8409