Lipschitz-Certifiable Training with a Tight Outer Bound

Sungyoon Lee[1] Jaewook Lee[1] Saerom Park[2]

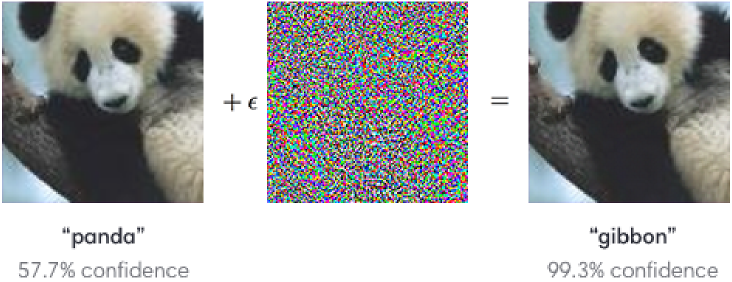
[1]Seoul National University
Statistical Learning & Computational Finance Lab

[2]Sungshin Women's University
Department of Convergence Security Engineering

goman1934@snu.ac.kr, psr6275@sungshin.ac.kr

December 9, 2020

Adversarial Examples



An input perturbed with a small adversarially designed perturbation that can change the network's prediction [Sze+13].

Heuristic Defenses ➔ Adaptive Attacks

Many heuristic defenses are proposed, but broken by adaptive attacks.

- Defensive distillation [Pap+16] ➔ $\boldsymbol{z}/T$ [CW16], CW attack [CW17]
- ICLR 18 ➔ EOT, BPDA attack [ACW18]
- Many more ➔ Adaptive attacks [Tra+20]
- ...

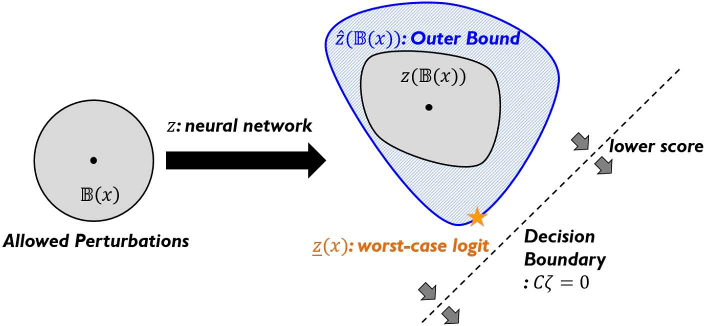
To this end, **certified defenses** are proposed.

Certified Defenses

Certified defenses minimize an upper bound on the worst-case loss over all possible perturbations $\mathbb{B}(\boldsymbol{x})$ as follows:

$$\max_{\boldsymbol{x}' \in \mathbb{B}(\boldsymbol{x})} \mathcal{L}(z(\boldsymbol{x}'), y) \leq \mathcal{L}(\underline{z}(\boldsymbol{x}), y) \tag{1}$$

with a worst-case logit $\underline{z}(\boldsymbol{x}) = \arg\min_{\zeta \in \hat{z}(\mathbb{B}(\boldsymbol{x}))} \boldsymbol{C}\zeta$.



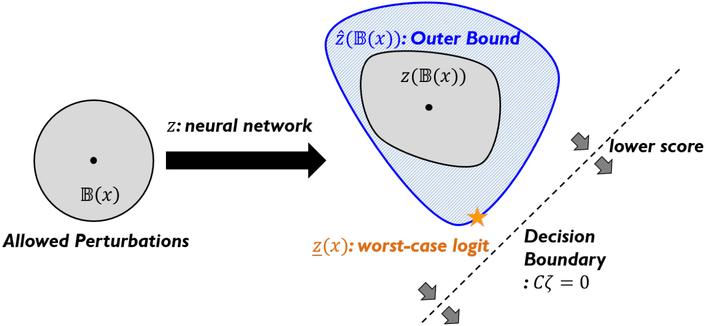
Certified Defenses

A worst-case logit $\underline{z}(\boldsymbol{x})$ can be obtained via the following optimization over an outer bound $\hat{z}(\mathbb{B}(\boldsymbol{x})) \supset z(\mathbb{B}(\boldsymbol{x}))$ where $\boldsymbol{C} = \mathbf{1}\boldsymbol{e}^{(y)T} - I$:

$$\underline{z}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\zeta \in \hat{z}(\mathbb{B}(\boldsymbol{x}))} \boldsymbol{C}\zeta \tag{2}$$

- LMT [TSS18]: $\hat{z}(\mathbb{B}(\boldsymbol{x})) = \mathbb{B}_2(z(\boldsymbol{x}), \epsilon L)$ with the Lipschitz constant L

Lipschitz outer bound



Intuition behind the Design

Overestimation problem of Lipschitz outer bound $\mathbb{B}(z(\boldsymbol{x}), L)$.

kth Bound → $(k+1)$th Bound (Tight bound $\subset$ Overestimated bound)

- Nonlinear operation (ReLU): $L_i = 1 \geq \frac{u^+}{u^+ - l^-}$
- Linear operation: $L_i = |\lambda_{max}(\boldsymbol{W}^{(i)})|$ (spectral norm)

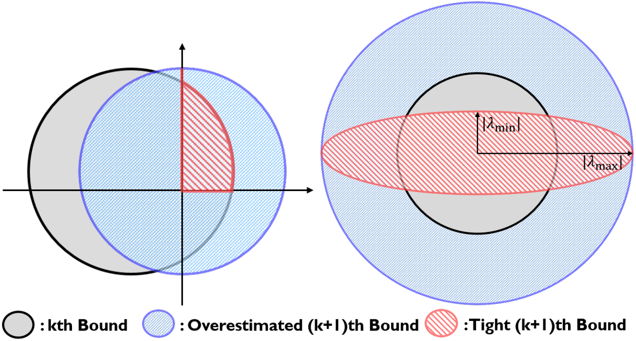


Figure: Overestimation in nonlinear (LEFT) and linear operation (RIGHT)

Intuition behind the Design

To address the overestimation problem of Lipschitz outer bound
→ Consider **element-wise bound (= Box Constraint)** propagation

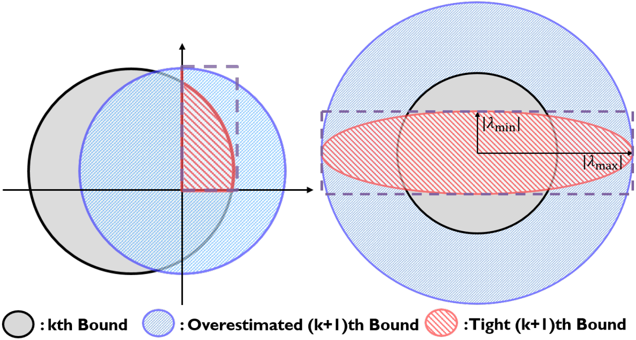


Figure: Overestimation in nonlinear (LEFT) and linear operation (RIGHT)

Proposed Method: Box Constraint Propagation

By introducing **an additional "Box Constraint ($\mathbb{B}_\infty$)"**, we can further tighten the worst-case bound as follows:

$$\mathcal{L}(\underset{\zeta \in \hat{z}(\mathbb{B}(\boldsymbol{x}))}{\arg\min}\ \boldsymbol{C}\zeta, y)$$

$$\downarrow$$

$$\max_{\boldsymbol{x}' \in \mathbb{B}(\boldsymbol{x})} \mathcal{L}(z(\boldsymbol{x}'), y) \leq \mathcal{L}(\underset{\zeta \in \mathbb{B}_2 \cap \mathbb{B}_\infty}{\arg\min}\ \boldsymbol{C}\zeta, y) \leq \mathcal{L}(\underset{\zeta \in \mathbb{B}_2}{\arg\min}\ \boldsymbol{C}\zeta, y)$$

"optimal (infeasible)" "tight" "loose"

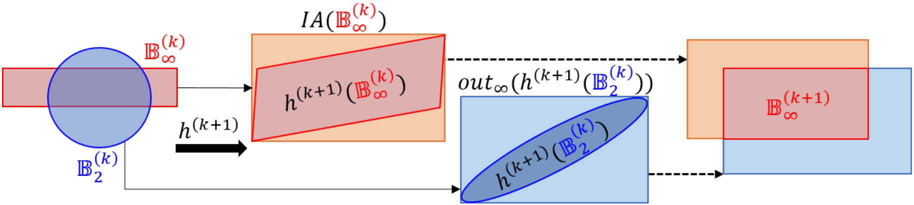


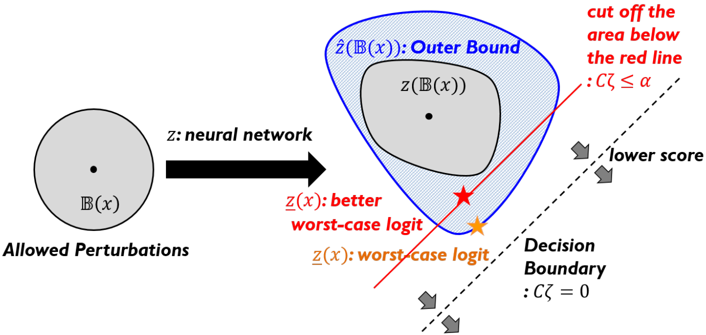
Figure: Box Constraint Propagation

Proposed Method: Box Constraint Propagation

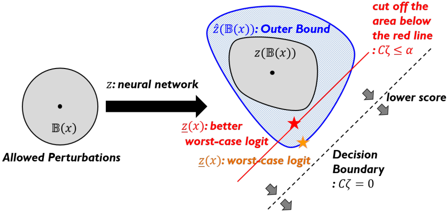
By introducing an additional "**Box Constraint** ($\mathbb{B}_\infty$)", we can further tighten the worst-case bound as follows:

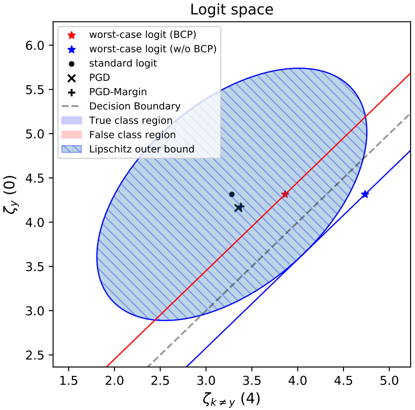
$$\max_{\boldsymbol{x}' \in \mathbb{B}(\boldsymbol{x})} \mathcal{L}(z(\boldsymbol{x}'), y) \leq \mathcal{L}(\operatorname*{arg\,min}_{\zeta \in \mathbb{B}_2 \cap \mathbb{B}_\infty} \boldsymbol{C}\zeta, y) \leq \mathcal{L}(\operatorname*{arg\,min}_{\zeta \in \mathbb{B}_2} \boldsymbol{C}\zeta, y)$$

"optimal (infeasible)" "tight" "loose"



Visualization





- CIFAR-10
- multi-class classification (10 classes)
- Lipschitz outer bound: blue ellipse
- worst-case logit with BCP: ⋆
- worst-case logit w/o BCP: ⋆

Contributions1 - Efficiency

It is **over 12 times faster** than CAP [Won+18].

Theorem (Efficient Computation)

We can find the optimal solution ζ^* of $\min_{\zeta \in \mathbb{B}_2 \cap \mathbb{B}_\infty} \mathbf{c}^T \zeta$ in a finite number of iterative steps less than the number of elements in $\mathbf{c}$.

Table: Computation time compared to CAP [Won+18].

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

Contributions2 - Tightness

The additional box constraint makes 'worst-case translations' **25-55% tighter** in terms of

$$\text{translation} \propto ||\underline{z}(\boldsymbol{x}) - z(\boldsymbol{x})||_1$$

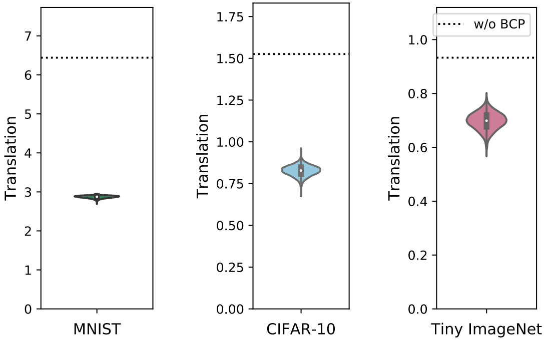
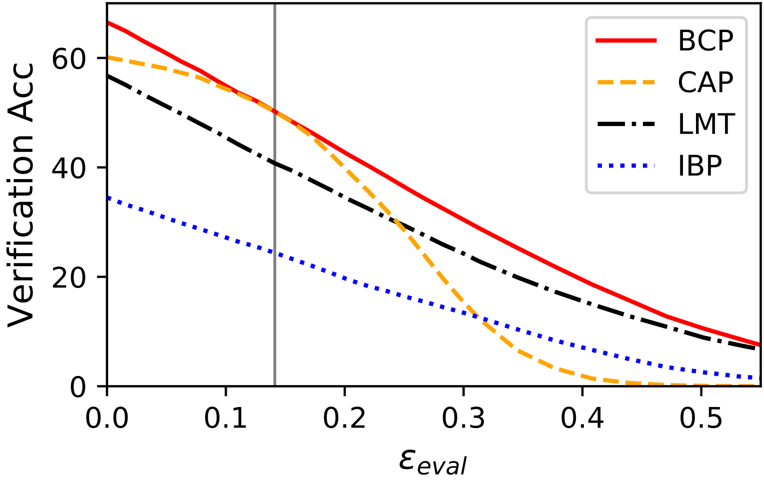


Figure: Tightness of the outer bounds. The dotted lines indicate the tightness without BCP. A smaller value indicates a better tightness.

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



Summary

- **Efficiency**: We propose a fast certified defense method called Box Constraint Propagation (BCP).

Summary

- **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.

Summary

- **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**: Therefore, we can build a certifiably robust model outperforms state-of-the-art methods.

*Focus: ℓ_2-norm bounded perturbations,
but applicable to any ℓ_p-cases $(p > 0)$.

Thank You

https://github.com/sungyoon-lee/bcp

Figure: Code & Paper

References

Anish Athalye, Nicholas Carlini, and David Wagner. "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples". In: *International Conference on Machine Learning*. 2018, pp. 274–283.

Nicholas Carlini and David Wagner. "Defensive distillation is not robust to adversarial examples". In: *arXiv preprint arXiv:1607.04311* (2016).

Nicholas Carlini and David Wagner. "Towards evaluating the robustness of neural networks". In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 39–57.

Sven Gowal et al. "On the effectiveness of interval bound propagation for training verifiably robust models". In: *arXiv preprint arXiv:1810.12715* (2018).

Nicolas Papernot et al. "Distillation as a defense to adversarial perturbations against deep neural networks". In: *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2016, pp. 582–597.

Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

Florian Tramer et al. "On adaptive attacks to adversarial example defenses". In: *arXiv preprint arXiv:2002.08347* (2020).

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

Eric Wong et al. "Scaling provable adversarial defenses". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8400–8409.