Towards Better Understanding of Training Certifiably Robust Models against Adversarial Examples

Sungyoon Lee[1] Woojin Lee[2] Jinseong Park[3] Jaewook Lee[3]

[1]Korea Institute for Advanced Study (KIAS)

[2]Dongguk University-Seoul [3]Seoul National University

sungyoonlee@kias.re.kr

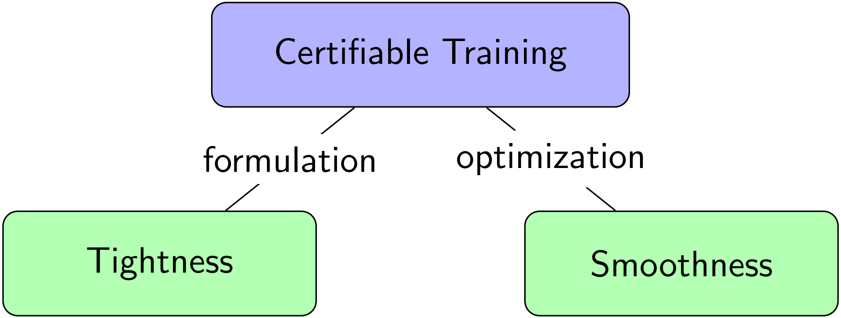
November 15, 2021

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

1. Introduction - Certifiable Training
2. Q. What is a key factor in certifiable training?
3. A. Smoothness
4. Experimental Results

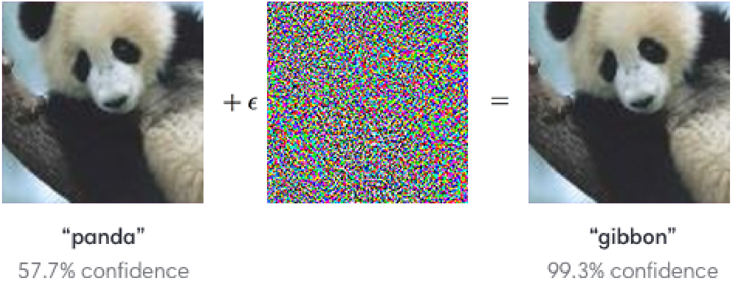
TL;DR



Certifiable Training
formulation
optimization
Tightness
Smoothness

Introduction - Certifiable Training

Adversarial Examples



Adversarial Example

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

Heuristic Defenses → Adaptive Attacks

To build a model that is robust to adversarial attacks,
many heuristic defenses are proposed, but broken by adaptive attacks.

- d → a (d is broken by a)
- Defensive distillation [Pap+16] → z/T [CW16], CW attack [CW17]
- ICLR 18 (preprocessing-based) → BPDA attack [ACW18]
- ICLR 18 (randomization-based) → EOT attack [Ath+18; ACW18]
- Many more → Adaptive attacks [Tra+20; CH20; Cro+20]
- ...

Heuristic Defenses → Adaptive Attacks

To build a model that is robust to adversarial attacks,
many heuristic defenses are proposed, but broken by adaptive attacks.

- d → a (d is broken by a)
- Defensive distillation [Pap+16] → z/T [CW16], CW attack [CW17]
- ICLR 18 (preprocessing-based) → BPDA attack [ACW18]
- ICLR 18 (randomization-based) → EOT attack [Ath+18; ACW18]
- Many more → Adaptive attacks [Tra+20; CH20; Cro+20]
- $\cdots$

To end this arms race of adversarial attack-defense,
certifiable training (certified defense) is proposed [HA17; RSL18; WK18; Won+18; Dvi+18; MGV18; Gow+18; Zha+19; BV19; LLP20].

ERM → ARM

Empirical Risk Minimization

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(f_\theta(x), y)] \qquad \text{(ERM)}$$

Adversarial Risk Minimization

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max_{x' \in \mathbb{B}(x,\epsilon)} \ell(f_\theta(x'), y)] \qquad \text{(ARM)}$$

Worst-case loss: $\max_{x' \in \mathbb{B}(x,\epsilon)} \ell(f_\theta(x'), y)$

Certifiable Training

Adversarial Risk Minimization

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\max_{x' \in \mathbb{B}(x,\epsilon)} \ell(f_\theta(x'), y)] \quad \text{(ARM)}$$

Upper Bound Approximation

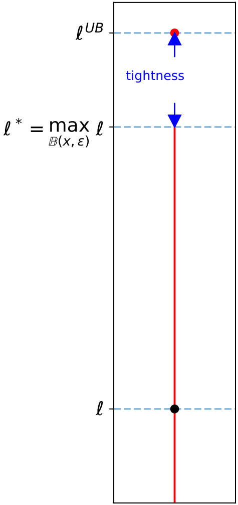
$$\max_{x' \in \mathbb{B}(x,\epsilon)} \ell(f_\theta(x'), y) \leq \ell^{UB}(x, y; \theta) \quad \text{(UB)}$$

Certifiable training minimizes the upper bound to build a "certifiably" robust model.

Certified Training

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell^{UB}(x, y; \theta)] \quad \text{(CT)}$$

Tightness



Interesting Observation

However, IBP [Gow+18] outperforms linear relaxation-based methods, especially when the perturbation is large, despite using much looser bounds.

	IBP		**CROWN-IBP** ($\beta = 1$)	**CAP**	**OURS**
train loss at the beginning	1.64	$>$	1.20	0.85	1.20
test error at the best checkpoint	73.19	$<$	75.82	73.91	70.92

Q. What is a key factor in certifiable training?

Questions

However, IBP outperforms linear relaxation-based methods, especially when the perturbation is large, despite using much looser bounds.

	IBP		CROWN-IBP ($\beta = 1$)	CAP	OURS
train loss at the beginning	1.64	$>$	1.20	0.85	1.20
test error at the best checkpoint	73.19	$<$	75.82	73.91	70.92

- **Q1.** Why does tighter bounds not result in a better performance?
- **Q2.** What other factors may influence the performance?

A. Smoothness

Certifiable Training from Optimization Perspectives

Total training loss: $\mathcal{L} = \mathbb{E}_{\mathcal{D}}[\ell]$

Certifiable Training

$$\min_{\theta \in \Theta} \mathcal{L}^*(\theta) \leq \min_{\theta \in \Theta} \mathcal{L}^{UB}(\theta) \qquad \text{(CD)}$$

- **Formulation**
  : tightness of **the upper bound** $\mathcal{L}^{UB}(\theta)$
- **Optimization**
  : smoothness of the landscape of **the objective function** $\mathcal{L}^{UB}(\theta)$

Smoothness of the loss landscape

Theorem (convergence rate of standard training)

Under some conditions,

$$\mathcal{L}(\theta_{t+1}) \leq \mathcal{L}(\theta_t)\left(1 - \alpha {\gamma_t}^{-1}\right) \tag{1}$$

for some $\alpha > 0$ *where* $\gamma_t = \frac{\|g_{t+1} - g_t\|}{\|g_t\|}$ *with* $g_t = \nabla_\theta \mathcal{L}(\theta_t)$.

Lower γ_t is favorable for the optimization.

Smoothness of the loss landscape

Theorem (convergence rate of certifiable training)

With gradient descent using a step size within an interval I_t during the ramp-up period ($0 \leq \epsilon_t \leq \epsilon$), the loss $\mathcal{L}^\epsilon$ for the target perturbation ϵ is reduced with

$$\mathcal{L}^\epsilon(\boldsymbol{\theta}_{t+1}) \leq \mathcal{L}^\epsilon(\boldsymbol{\theta}_t)\Big(1 - \frac{\mu}{2}\cos^2(\phi_t)\left\|\boldsymbol{H}_t^\epsilon \boldsymbol{u}_t\right\|^{-1}\Big) \tag{2}$$

for $\boldsymbol{u}_t = \frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon_t}(\boldsymbol{\theta}_t)}{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon_t}(\boldsymbol{\theta}_t)\|}$ *where* $0 < \mu \leq \frac{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}^\epsilon\|^2}{2\mathcal{L}^\epsilon}$, $\cos(\phi_t) = \frac{\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon\,T}\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon_t}}{\|\nabla_{\boldsymbol{\theta}}\mathcal{L}^\epsilon\|\|\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon_t}\|}$ *and* $\boldsymbol{H}_t^\epsilon$ *satisfies* $\mathcal{L}^\epsilon(\boldsymbol{\theta}_{t+1}) = \mathcal{L}^\epsilon(\boldsymbol{\theta}_t) + \nabla_{\boldsymbol{\theta}}\mathcal{L}^\epsilon(\boldsymbol{\theta}_t)^T\Delta_t + \frac{1}{2}\Delta_t^T\boldsymbol{H}_t^\epsilon\Delta_t$ *and* $\Delta_t^T\boldsymbol{H}_t^\epsilon\Delta_t > 0$ *with* $\Delta_t = \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t$.

Lower $\left\|\boldsymbol{H}_t^\epsilon \boldsymbol{u}_t\right\|$ is favorable for the optimization.

cf. $\left\|\boldsymbol{H}_t^\epsilon \boldsymbol{u}_t\right\| = \|\boldsymbol{H}_t^\epsilon g_t^{\epsilon_t}\|/\|g_t^{\epsilon_t}\| = \|\boldsymbol{H}_t^\epsilon\Delta_t\|/\|\Delta_t\| \approx \|g_{t+1}^\epsilon - g_t^\epsilon\|/\|\Delta_t\|$

Non-smoothness measures

We used the following **non-smoothness measures**:

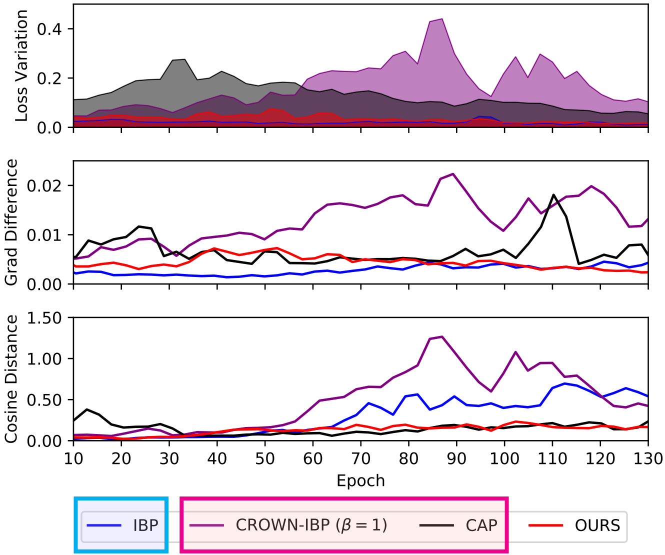
- Loss variation:
  $|\mathcal{L}^{\epsilon_t}(\boldsymbol{\theta}(\lambda)) - \mathcal{L}^{\epsilon_t}(\boldsymbol{\theta}(0))|$ for $\lambda \in [0, 5]$ where $\boldsymbol{\theta}(\lambda) \equiv \boldsymbol{\theta}_t - \lambda\eta\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon_t}(\boldsymbol{\theta}_t)$
- Grad Difference: $\|\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon_t}(\boldsymbol{\theta}_t) - \nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon_t}(\boldsymbol{\theta}_{t+1})\|$
- Cosine Distance: $1 - \cos(\nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon_t}(\boldsymbol{\theta}_t), \nabla_{\boldsymbol{\theta}}\mathcal{L}^{\epsilon_t}(\boldsymbol{\theta}_{t+1}))$

Higher non-smoothness measures indicate less smooth loss landscape

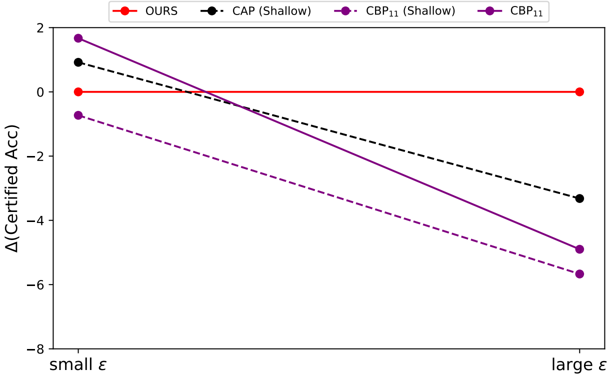
Experimental Results

Non-smoothness measures

Higher (non-smoothness) measures indicate less smooth loss landscape.



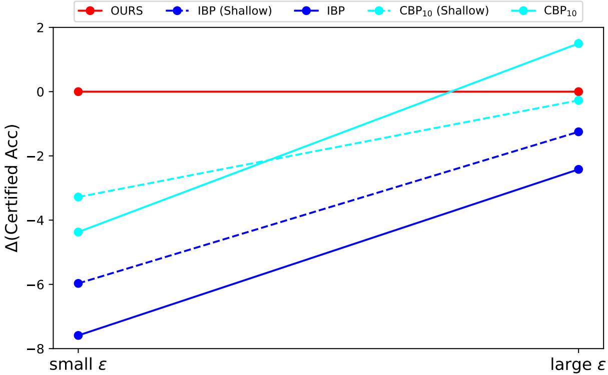
Tightness (small ϵ)



cf. $\text{CBP}_{11} = \text{CROWN-IBP } (\beta = 1)$

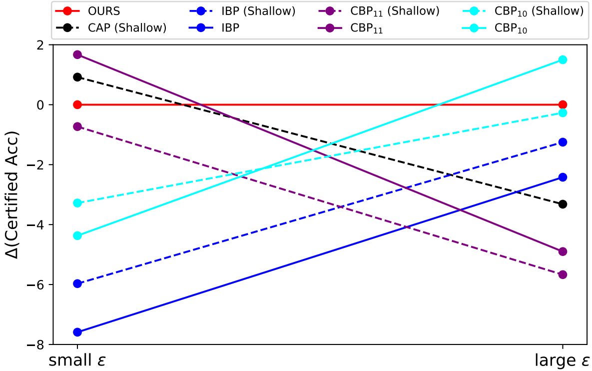
Δ(Certified Acc) indicates the difference of the certified accuracy with the proposed method when the same architecture is used.

Smoothness (large ϵ)



cf. $\text{CBP}_{10} = \text{CROWN-IBP} \ (\beta = 1 \rightarrow \beta = 0)$

Tightness (small ϵ) & Smoothness (large ϵ)



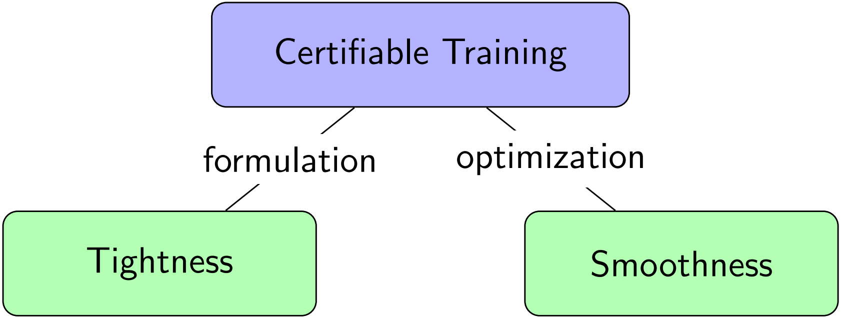
Performance

Table: Test errors (Standard / PGD / Verified error).
Bold and underline numbers are the **1st** and 2nd lowest verified error.

Data	$\epsilon_{\text{test}}(l_\infty)$	IBP	CROWN-IBP ($\beta = 1$)	CAP	OURS
MNIST	0.1	1.18 / 2.16 / 3.52	1.07 / 1.69 / **2.10**	0.80 / 1.73 / 3.19	1.09 / 1.77 / 2.36
	0.2	2.00 / 3.29 / 6.31	2.99 / 5.50 / 7.97	3.22 / 6.72 / 11.06	1.70 / 3.44 / **4.34**
	0.3	3.50 / 5.85 / 10.45	5.73 / 10.76 / 16.28	19.19 / 35.84 / 47.85	3.49 / 5.59 / **9.79**
	0.4	3.50 / 7.30 / 17.96	5.73 / 14.63 / 23.80	-	3.49 / 6.77 / **15.42**
CIFAR-10 (Shallow)	2/255	37.98 / 49.40 / 55.39	32.48 / 42.77 / 50.15	28.80 / 38.95 / **48.50**	31.49 / 42.73 / 49.42
	4/255	46.42 / 57.42 / 62.80	45.56 / 58.24 / 64.47	40.78 / 52.62 / 61.88	42.53 / 55.55 / **61.52**
	6/255	52.84 / 63.92 / 68.79	54.72 / 65.28 / 71.04	49.20 / 60.85 / 69.03	50.19 / 61.88 / **66.90**
	8/255	55.71 / 66.79 / 70.95	61.37 / 70.66 / 75.37	56.77 / 66.78 / 73.02	56.01 / 66.17 / **69.70**
	16/255	67.10 / 75.12 / 78.26	76.65 / 81.90 / 84.42	75.11 / 80.67 / 82.07	65.93 / 75.39 / **77.87**
CIFAR-10 (Deep)	2/255	39.17 / 48.80 / 55.48	29.02 / 40.17 / **46.22**	-	31.48 / 42.52 / 47.89
	8/255	59.53 / 65.98 / 70.86	59.43 / 65.79 / 73.34	-	50.78 / 62.58 / **68.44**
SVHN	0.01	19.91 / 34.12 / 43.83	17.25 / 30.84 / 39.88	16.88 / 30.16 / **37.09**	16.41 / 30.43 / 39.44

cf. There are more comparison results (RS [Xia+18], DiffAI [MGV18], COLT [BV19], and CBP_{10} [Zha+19]) in the paper.

Summary



Certifiable Training
formulation
optimization
Tightness
Smoothness

Thank You

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

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.

Anish Athalye et al. "Synthesizing robust adversarial examples". In: *International conference on machine learning*. PMLR. 2018, pp. 284–293.

Mislav Balunovic and Martin Vechev. "Adversarial training and provable defenses: Bridging the gap". In: *International Conference on Learning Representations*. 2019.

Francesco Croce and Matthias Hein. "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks". In: *arXiv preprint arXiv:2003.01690* (2020).

Francesco Croce et al. "RobustBench: a standardized adversarial robustness benchmark". In: *arXiv preprint arXiv:2010.09670* (2020).

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.

Krishnamurthy Dvijotham et al. "Training verified learners with learned verifiers". In: *arXiv preprint arXiv:1805.10265* (2018).

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

Matthias Hein and Maksym Andriushchenko. "Formal guarantees on the robustness of a classifier against adversarial manipulation". In: *Advances in Neural Information Processing Systems*. 2017, pp. 2266–2276.

Sungyoon Lee, Jaewook Lee, and Saerom Park. "Lipschitz-Certifiable Training with a Tight Outer Bound". In: *Advances in Neural Information Processing Systems* 33 (2020).

Matthew Mirman, Timon Gehr, and Martin Vechev. "Differentiable abstract interpretation for provably robust neural networks". In: *International Conference on Machine Learning*. 2018, pp. 3575–3583.

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.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. "Semidefinite relaxations for certifying robustness to adversarial examples". In: *Advances in Neural Information Processing Systems*. 2018, pp. 10877–10887.

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).

Eric Wong and Zico Kolter. "Provable defenses against adversarial examples via the convex outer adversarial polytope". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 5286–5295.

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

Kai Y Xiao et al. "Training for faster adversarial robustness verification via inducing relu stability". In: *arXiv preprint arXiv:1809.03008* (2018).

Huan Zhang et al. "Towards Stable and Efficient Training of Verifiably Robust Neural Networks". In: *International Conference on Learning Representations*. 2019.