Trustworthy Machine Learning

Certified Adversarial Learning

Sangdon Park

POSTECH

Motivation

- Heuristic adversarial learning often fails against powerful adversaries.

CIFAR10

(a) Standard training

	Simple	Wide
Natural	92.7%	95.2%
FGSM	27.5%	32.7%
PGD	0.8%	3.5%

(b) FGSM training

	Simple	Wide
Natural	87.4%	90.3%
FGSM	90.9%	95.1%
PGD	0.0%	0.0%

(c) PGD training

	Simple	Wide
Natural	79.4%	87.3%
FGSM	51.7%	56.1%
PGD	43.7%	45.8%

  - FGSM training and FGSM attacks: 90.9% accuracy :)
  - FGSM training and PGD attacks: 0.0% accuracy :(

Motivation

- Heuristic adversarial learning often fails against powerful adversaries.

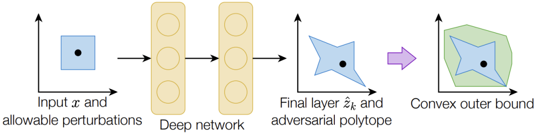
CIFAR10

	Simple	Wide	Simple	Wide	Simple	Wide
Natural	92.7%	95.2%	87.4%	90.3%	79.4%	87.3%
FGSM	27.5%	32.7%	90.9%	95.1%	51.7%	56.1%
PGD	0.8%	3.5%	0.0%	0.0%	43.7%	45.8%
	(a) Standard training		(b) FGSM training		(c) PGD training	

  - FGSM training and FGSM attacks: 90.9% accuracy :)
  - FGSM training and PGD attacks: 0.0% accuracy :(
- Can we learn a classifier robust to **any** small perturbations?

Certified Adversarial Learning

- Convex outer approximation [Kolter and Wong, 2017]



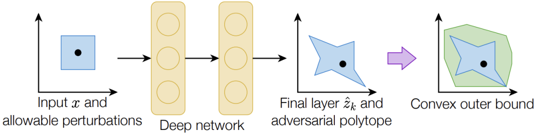
✓ Certified!

$$\max_{\|\delta\|_\infty \le \varepsilon} \ell(f, x+\delta, y) \le U(\varepsilon, f, x, y)$$

✗ Not scalable :(

Certified Adversarial Learning

- Convex outer approximation [Kolter and Wong, 2017]

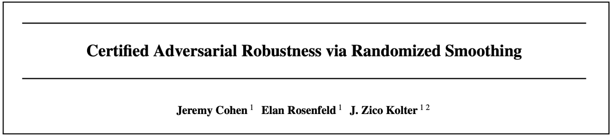


✓ Certified!

$$\max_{\|\delta\|_\infty \leq \varepsilon} \ell(f, x + \delta, y) \leq U(\varepsilon, f, x, y)$$

✗ Not scalable :(

- Randomized smoothing: a post-hoc method



Certified Adversarial Robustness via Randomized Smoothing

Jeremy Cohen [1] **Elan Rosenfeld** [1] **J. Zico Kolter** [1 2]

✓ (Probably) Certified!
✓ Scalable!

A Goodness Definition: Robustness

$$\max_{\|\delta\|_p \leq \varepsilon} f(x+\delta) = f(x)$$

- $f : \mathcal{X} \to \mathcal{Y}$: a classifier

A Goodness Definition: Robustness

$$\max_{\|\delta\|_p \leq \varepsilon} f(x+\delta) = f(x)$$

- $f : \mathcal{X} \to \mathcal{Y}$: a classifier
- The constraint on the perturbation δ can be more general.

A Goodness Definition: Robustness

$$\max_{\|\delta\|_p \leq \varepsilon} f(x+\delta) = f(x)$$

- $f : \mathcal{X} \to \mathcal{Y}$: a classifier
- The constraint on the perturbation δ can be more general.
- It does not matter whether $f(x)$ is correct.

A Certified Method: Randomized Smoothing

$$g(x) \coloneqq \arg\max_{c \in \mathcal{Y}} \mathbb{P}\left\{f(x+\delta) = c\right\} \quad \text{where} \quad \delta \sim \mathcal{N}(0, \sigma^2 I)$$

- $g : \mathcal{X} \to \mathcal{Y}$: a smoothed classifier
- σ is related to the maximum perturbation ε.

Robustness Guarantee

Binary Classification

Theorem

Suppose that $\underline{p_A} \in (0.5, 1]$ *satisfies*

$$\mathbb{P}\left\{f(x+\varepsilon) = c_A\right\} \geq \underline{p_A} \quad \textit{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Then, we have $g(x+\delta) = c_A$ *if*

$$\|\delta\|_2 < \sigma \Phi^{-1}(\underline{p_A}).$$

- c_A: the most probable class when f classifies $x + \varepsilon$
- p_A: the chance that f classifies $x + \varepsilon$ by c_A
- $\underline{p_A}$: the lower bound of p_A
- Φ^{-1}: the inverse of the standard Gaussian CDF

Robustness Guarantee

Binary Classification

Theorem

Suppose that $\underline{p_A} \in (0.5, 1]$ *satisfies*

$$\mathbb{P}\left\{f(x+\varepsilon) = c_A\right\} \geq \underline{p_A} \quad \textit{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Then, we have $g(x+\delta) = c_A$ *if*

$$\|\delta\|_2 < \sigma\Phi^{-1}(\underline{p_A}).$$

- c_A: the most probable class when f classifies $x+\varepsilon$
- p_A: the chance that f classifies $x+\varepsilon$ by c_A
- $\underline{p_A}$: the lower bound of p_A
- Φ^{-1}: the inverse of the standard Gaussian CDF
- Here, we assume that we can compute $\underline{p_A}$.

Robustness Guarantee

Binary Classification

Theorem

Suppose that $\underline{p_A} \in (0.5, 1]$ *satisfies*

$$\mathbb{P}\left\{f(x+\varepsilon) = c_A\right\} \geq \underline{p_A} \quad \textit{where} \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I).$$

Then, we have $g(x+\delta) = c_A$ *if*

$$\|\delta\|_2 < \sigma \Phi^{-1}(\underline{p_A}).$$

- c_A: the most probable class when f classifies $x + \varepsilon$
- p_A: the chance that f classifies $x + \varepsilon$ by c_A
- $\underline{p_A}$: the lower bound of p_A
- Φ^{-1}: the inverse of the standard Gaussian CDF
- Here, we assume that we can compute $\underline{p_A}$.
- Due to the Gaussian, we can compute the maximum perturbation to be robust!

Robustness Guarantee: A Proof Sketch (1/3)

Binary Classification

- Fix a perturbation δ.
- From the definition of g, we have

$$
\begin{aligned}
g(x+\delta) &:= \arg\max_c \mathbb{P}\left\{f(x+\varepsilon+\delta)=c\right\} \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0,\sigma^2 I) \\
&= \arg\max_c \mathbb{P}\left\{f(x+\varepsilon')=c\right\} \quad \text{where} \quad \varepsilon' \sim \mathcal{N}(\delta,\sigma^2 I) \\
&\stackrel{?}{=} c_A
\end{aligned} \tag{1}
$$

Robustness Guarantee: A Proof Sketch (1/3)

Binary Classification

- Fix a perturbation δ.
- From the definition of g, we have

$$\begin{aligned} g(x+\delta) &:= \arg\max_c \mathbb{P}\left\{f(x+\varepsilon+\delta)=c\right\} \quad \text{where} \quad \varepsilon \sim \mathcal{N}(0,\sigma^2 I) \\ &= \arg\max_c \mathbb{P}\left\{f(x+\varepsilon')=c\right\} \quad \text{where} \quad \varepsilon' \sim \mathcal{N}(\delta,\sigma^2 I) \\ &\overset{?}{=} c_A \end{aligned} \tag{1}$$

- We wish to prove (1). How?
  - f can be any classifier, which is not easy to analyze.
  - Consider a surrogate classifier that bounds the probability and is easier to analyze, *e.g.*,

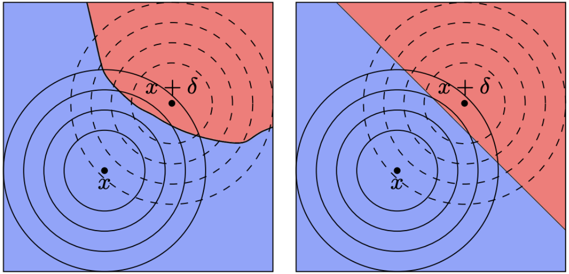
$$\mathbb{P}\left\{f(x+\varepsilon')=c_A\right\} \geq \min_{f':\mathbb{P}\{f(x+\varepsilon)=c_A\}\geq \underline{p_A}} \mathbb{P}\left\{f'(x+\varepsilon')=c_A\right\} > \frac{1}{2} \quad \Longrightarrow \quad g(x+\delta)=c_A.$$

Robustness Guarantee: A Proof Sketch (2/3)

Binary Classification

- Interestingly, f^* is linear (due to the Neyman-Perason lemma), where

$$f^* = \arg \min_{f' : \mathbb{P}\{f(x+\varepsilon)=c_A\} \geq \underline{p_A}} \mathbb{P}\left\{f'(x+\varepsilon') = c_A\right\}$$



Robustness Guarantee: A Proof Sketch (3/3)

Binary Classification

- We have a closed-form solution of f^*:

$$f^*(x') := \begin{cases} c_A & \text{if } \delta^T(x'-x) \leq \sigma\|\delta\|_2\Phi^{-1}(\underline{p_A}) \\ c_B & \text{otherwise} \end{cases} .$$

- This implies

$$\mathbb{P}\left\{f^*(x+\varepsilon') = c_A\right\} = \Phi\left(\Phi^{-1}(\underline{p_A}) - \frac{\|\delta\|_2}{\sigma}\right)$$

- The above probability should be larger than $\frac{1}{2}$, *i.e.*,

$$\Phi\left(\Phi^{-1}(\underline{p_A}) - \frac{\|\delta\|_2}{\sigma}\right) > \frac{1}{2} \quad \Longrightarrow \quad \|\delta\|_2 < \sigma\Phi^{-1}(\underline{p_A}).$$

Robustness Guarantee

Multi-class Classification

Theorem

Suppose that $\underline{p_A}, \overline{p_B} \in [0,1]$ *satisfies*

$$\mathbb{P}\left\{f(x+\varepsilon) = c_A\right\} \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}\left\{f(x+\varepsilon) = c\right\}.$$

Then, we have $g(x+\delta) = c_A$ *for all* $\|\delta\|_2 \leq R$*, where*

$$R \coloneqq \frac{\sigma}{2}\left(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})\right).$$

- c_A: the most probable label (with probability at least $\underline{p_A}$)
- $c_B \coloneqq \arg\max_{c \neq c_A} \mathbb{P}\left\{f(x+\varepsilon) = c\right\}$: the second-most probable label (with probability at most $\overline{p_B}$)

Prediction

function PREDICT(f, σ, x, n, α)
  `counts` $\leftarrow$ SAMPLEUNDERNOISE(f, x, n, σ)
  $\hat{c}_A, \hat{c}_B \leftarrow$ top two indices in `counts`
  $n_A, n_B \leftarrow$ `counts`$[\hat{c}_A]$, `counts`$[\hat{c}_B]$
  if BINOMPVALUE($n_A, n_A + n_B, 0.5$) $\leq \alpha$ **return** $\hat{c}_A$
  else return ABSTAIN

- Recall the randomized smoothing method:

$$g(x) := \arg\max_{c \in \mathcal{Y}} \mathbb{P}\left\{f(x+\delta) = c\right\} \quad \text{where} \quad \delta \sim \mathcal{N}(0, \sigma^2 I)$$

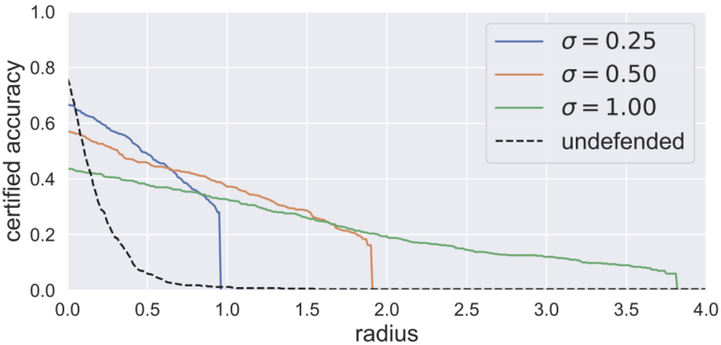
1. Draw n noisy perturbations $\delta_1, \ldots, \delta_n$.
2. Empirically compute the most probable and the second most probably labels, *i.e.*, $\hat{c}_A$ and $\hat{c}_B$.
3. If $\hat{c}_A$ is drawn from the binomial distribution with $p = 0.5$, return $\hat{c}_A$.

Certification in Evaluation

```
# certify the robustness of g around x
function CERTIFY(f, σ, x, n0, n, α)
    counts0 ← SAMPLEUNDERNOISE(f, x, n0, σ)
    ĉA ← top index in counts0
    counts ← SAMPLEUNDERNOISE(f, x, n, σ)
    pA ← LOWERCONFBOUND(counts[ĉA], n, 1 − α)
    if pA > 1/2 return prediction ĉA and radius σ Φ^{-1}(pA)
    else return ABSTAIN
```

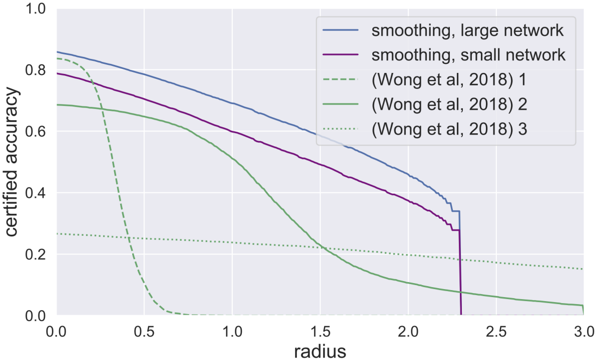
1. Compute $\underline{p_A}$ via the binomial tail bound.
2. Compute the robust radius, *i.e.*, $\sigma\Phi^{-1}(\underline{p_A})$.
3. If (a desired radius) $\leq \sigma\Phi^{-1}(\underline{p_A})$, then "certified".

Results: ImageNet



- Classifier: ResNet-50
- undefended: a classifier with heuristic adversarial training (using ℓ_2 adversarial attacks)
- perturbation: $\|\delta\|_2 \leq$ (radius)

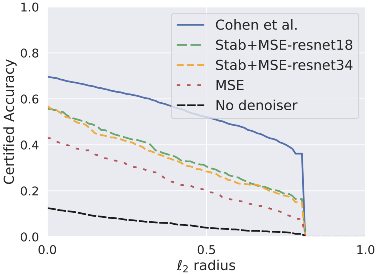
Results: Comparison



- (maybe) on MNIST
- Baseline: deterministic robustness guarantee
- randomized smoothing: high-probability guarantee

Limitation of Randomized Smoothing

- Randomized smoothing requires retraining (*e.g.*, Gaussian data augmentation).



  - Cohen et al.: Randomized smoothing with retraining
  - No denoiser: Randomized smoothing without retraining
- How to avoid retraining?

Denoise Gaussian Noise

Denoised Smoothing:
A Provable Defense for Pretrained Classifiers

Hadi Salman
hasalman@microsoft.com
Microsoft Research

Mingjie Sun
mingjies@cs.cmu.edu
CMU

Greg Yang
gragyang@microsoft.com
Microsoft Research

Ashish Kapoor
akapoor@microsoft.com
Microsoft Research

J. Zico Kolter
zkolter@cs.cmu.edu
CMU

- A classifier randomized smoothing needs to be robust to Gaussian noise for better certified robustness.
- How about denoise Gaussian noise and then use the randomized smoothing?

Denoised Smoothing

Randomized Smoothing:

$$g(x) := \arg\max_{c \in \mathcal{Y}} \mathbb{P}\left\{f(x+\delta) = c\right\} \quad \text{where} \quad \delta \sim \mathcal{N}(0, \sigma^2 I)$$

- Applicable for any classifier f

Denoised Smoothing:

$$g(x) := \arg\max_{c \in \mathcal{Y}} \mathbb{P}\left\{f(\mathcal{D}(x+\delta)) = c\right\} \quad \text{where} \quad \delta \sim \mathcal{N}(0, \sigma^2 I)$$

- $\mathcal{D} : \mathcal{X} \to \mathcal{X}$: a denoiser
- Consider a new classifier $f \circ \mathcal{D}$ and then enjoy randomized smoothing.

How to Train a Denoiser?

MSE objective:

$$L_{\mathsf{MSE}} := \mathop{\mathbb{E}}_{x,y,\delta} \|\mathcal{D}(x+\delta) - x\|_2^2$$

How to Train a Denoiser?

MSE objective:

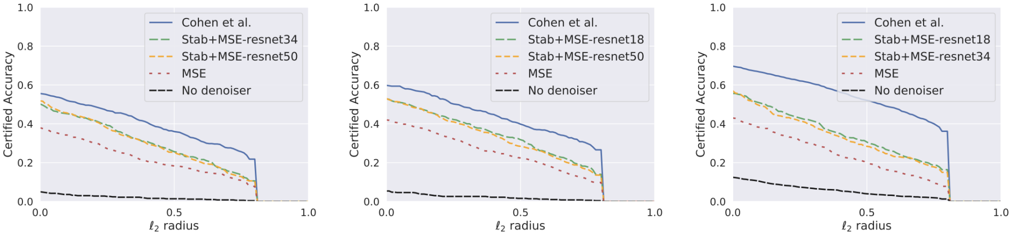
$$L_{\mathsf{MSE}} := \underset{x,y,\delta}{\mathbb{E}} \|\mathcal{D}(x+\delta) - x\|_2^2$$

✗ Does not consider the accuracy of a classifier.

Stability objective:

$$L_{\mathsf{Stab}} := \underset{x,y,\delta}{\mathbb{E}} \ell(f, \mathcal{D}(x+\delta), f(x)) \quad \text{where} \quad \delta \sim \mathcal{N}(0, \sigma^2 I)$$

Results



(a) ResNet-18 (b) ResNet-34 (c) ResNet-50

- The denoised smoothing without retraining is quite similar to the randomized smoothing with retraining.

Conclusion

- Randomized smoothing provides a simple defense mechanism.
- Desnoised smoothing does not require to retrain a classifier (but still requires training the denoiser).
- Recently, the denoised smoothing was improved via denoising diffusion probabilistic Models [Carlini et al., 2023].

			Certified Accuracy at ε (%)				
Method	Off-the-shelf	Extra data	0.5	1.0	1.5	2.0	3.0
PixelDP (Lecuyer et al., 2019)	○	✗	$^{(33.0)}$16.0	-	-		
RS (Cohen et al., 2019)	○	✗	$^{(67.0)}$49.0	$^{(57.0)}$37.0	$^{(57.0)}$29.0	$^{(44.0)}$19.0	$^{(44.0)}$12.0
SmoothAdv (Salman et al., 2019)	○	✗	$^{(65.0)}$56.0	$^{(54.0)}$43.0	$^{(54.0)}$37.0	$^{(40.0)}$27.0	$^{(40.0)}$20.0
Consistency (Jeong & Shin, 2020)	○	✗	$^{(55.0)}$50.0	$^{(55.0)}$44.0	$^{(55.0)}$34.0	$^{(41.0)}$24.0	$^{(41.0)}$17.0
MACER (Zhai et al., 2020)	○	✗	$^{(68.0)}$57.0	$^{(64.0)}$43.0	$^{(64.0)}$31.0	$^{(48.0)}$25.0	$^{(48.0)}$14.0
Boosting (Horváth et al., 2022a)	○	✗	$^{(65.6)}$57.0	$^{(57.0)}$44.6	$^{(57.0)}$**38.4**	$^{(44.6)}$**28.6**	$^{(38.6)}$**21.2**
DRT (Yang et al., 2021)	○	✗	$^{(52.2)}$46.8	$^{(55.2)}$44.4	$^{(49.8)}$**39.8**	$^{(49.8)}$**30.4**	$^{(49.8)}$**23.4**
SmoothMix (Jeong et al., 2021)	○	✗	$^{(55.0)}$50.0	$^{(55.0)}$43.0	$^{(55.0)}$**38.0**	$^{(40.0)}$26.0	$^{(40.0)}$20.0
ACES (Horváth et al., 2022b)	◐	✗	$^{(63.8)}$54.0	$^{(57.2)}$42.2	$^{(55.6)}$35.6	$^{(39.8)}$25.6	$^{(44.0)}$19.8
Denoised (Salman et al., 2020)	◐	✗	$^{(60.0)}$33.0	$^{(38.0)}$14.0	$^{(38.0)}$6.0	-	-
Lee (Lee, 2021)	●	✗	41.0	24.0	11.0	-	-
Ours	●	✓	$^{(82.8)}$**71.1**	$^{(77.1)}$**54.3**	$^{(77.1)}$**38.1**	$^{(60.0)}$**29.5**	$^{(60.0)}$13.1

Reference I

N. Carlini, F. Tramer, K. D. Dvijotham, L. Rice, M. Sun, and J. Z. Kolter. (certified!!) adversarial robustness for free!, 2023.

J. Z. Kolter and E. Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017.