Skoltech

Skolkovo Institute of Science and Technology

April 29, 2021

Certified Robustness in High Dimensions

Aleksandr Petiushko

Lomonosov MSU, PhD

Huawei, VI & Fundamental Research Team Leader

Intelligent Systems and Data Science Lab

Certified Robustness - definitions

- Suppose our NN function $f(x)$ is the **classifier** to K classes: $f: R^d \rightarrow Y, Y = \{1, \dots, K\}$
  - Usually we have NN $h(x)$: $R^d \rightarrow R^K$, and $f(x) = argmax_i h(x)_i$
  - Deterministic approach: we want to find provide the class of perturbation $S(x, f)$ so as the classifier output will not change, or more formally:
    - $f(x + \delta) = f(x)\ \forall \delta \in S(x, f)$
  - Probabilistic approach: having the probability of robustness P, find the class of input perturbations $S(x, f, P)$ s.t.:
    - $Prob_{\delta \in S(x,f,P)}\big(f(x + \delta) = f(x)\big) = P$

- If NN $f(x)$ is the **regressor**: $f: R^d \rightarrow R$
  - Having the upper and lower bounds on the output perturbation, find the class of input perturbations $S\big(x, f, f_{low}, f_{up}\big)$:
    - $f(x) - f_{low} \leq f(x + \delta) \leq f(x) + f_{up}, \forall \delta \in S(x, f, f_{low}, f_{up})$

Certified Robustness – definitions (2)

- Also, inverse tasks could considered

- If NN $f(x)$ is the **classifier** to K classes: $f: R^d \to Y, Y = \{1, \dots, K\}$
  - Probabilistic approach: we want to measure the probability of not changing the classifier output under some class of input perturbations S:
    - $Prob_{\delta \in S}(f(x+\delta) = f(x))$

- If NN $f(x)$ is the **regressor**: $f: R^d \to R$
  - We want to find the upper and lower bounds of the output perturbation under some class of input perturbations S in the analytical form:
    - $f(x) - f_{low}(f, x, S) \leq f(x+\delta) \leq f(x) + f_{up}(f, x, S), \forall \delta \in S$

- Can be measured just by analyzing the output of $f(x+\delta)$

Certified Robustness and Lipschitz Function

- Neural Net output is $h: R^d \to R^K$, and the classifier itself is $f: R^d \to Y, Y = \{1, \dots, K\}$, where $f(x) = argmax_{i \in Y} h(x)_i$
- Consider binary case ($K = 2$), and probabilistic output: $h(x)_1 + h(x)_2 = 1, h(x)_i \geq 0$
- **Lipschitz function** $f: R^d \to R$ with a Lipschitz constant L: $\forall x_1, x_2$ it is true that

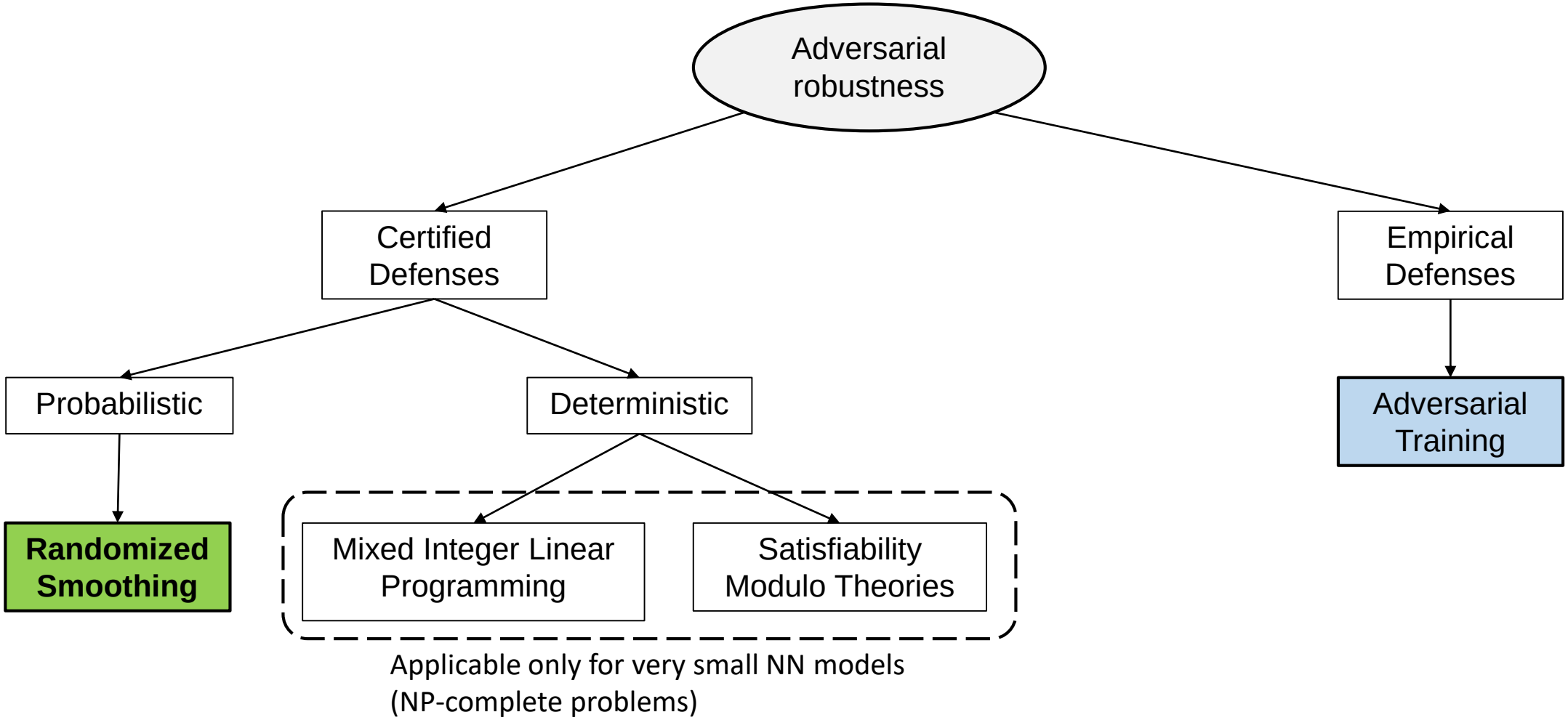
$$|f(x_1) - f(x_2)| \leq L||x_1 - x_2||$$

- **Local Lipschitz function** f with a Lipschitz constant $L(x_0)$: $\forall x \in S(x_0)$ it is true that

$$|f(x_0) - f(x)| \leq L(x_0)||x_0 - x||$$

- Let $j = argmax_{i \in Y} h(x_0)_i$, and $h(x_0)_j - h(x_0)_{i \neq j} \geq \epsilon \; \forall i \neq j$
- Let $h(x_0)_j$ - local Lipschitz function with a Lipschitz constant $L(x_0)$
- Then if $S(x_0) = \{x: \left|\left|x_0 - x\right|\right| \leq \frac{\epsilon}{2L(x_0)}\}$, we have $\left|h(x_0)_j - h(x)_j\right| \leq L(x_0) \frac{\epsilon}{2L(x_0)} = \frac{\epsilon}{2}$
- As a consequence we'll have $j = argmax_{i \in Y} h(x)_i$, and $f(x) = f(x_0) = j$ in the vicinity $S(x_0) = \{x: \left|\left|x_0 - x\right|\right| \leq \frac{\epsilon}{2L(x_0)}\}$, and certified robustness!
- True certified radius can be much bigger than Lipschitz vicinity $S(x_0)$

Adversarial Robustness



Empirical VS Certified

- **Empirical**
  - Upper bound on the true robustness accuracy
  - But only until the new stronger attack appears

- **Certified**
  - Lower bound on the true robustness accuracy
  - It is what has been theoretically proven, and no one attack can beat it

Empirical robustness: adversarial training

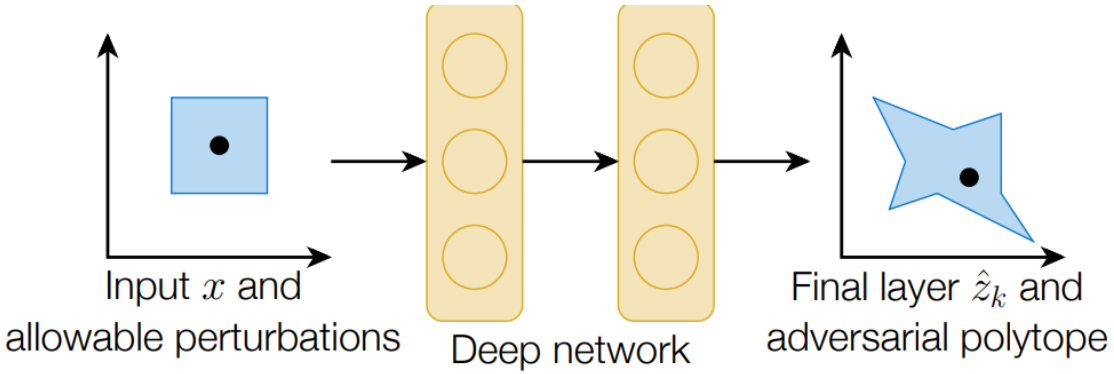
- **Main idea**:
  - train on the most hard examples using some class of perturbations $S(= \Delta)$ around training examples

$$\min_{\theta} \mathbf{E}_{x,y}[\text{Loss}(f_\theta(x), y)] \Longrightarrow \min_{\theta} \mathbf{E}_{x,y}\left[\max_{\delta \in \Delta} \text{Loss}\left(f_\theta(x+\delta), y\right)\right]$$

- **Drawbacks**:
  - Quite inefficient training (longer than usual because of finding of hard examples for every training sample for every iteration)
  - The accuracy on clean samples is lower than for usual training

The problem with l_p-balls

- Usually the robustness is studied under the l_p-balls perturbations of input (S: $\|\delta\|_p \leq \varepsilon$)
- The problem is while the input l_p-ball is convex, for the output it could be of any form and convexity
- That's why it is hard to prove anything



Convex relaxation[1]

- **Main idea**:
  - Let's make out regions convex by relaxation!

Corollary 1. *For a data point x, label $y^\star$ and $\epsilon > 0$, if*

$$J_\epsilon(x, g_\theta(e_{y^\star} 1^T - I)) \geq 0 \qquad (15)$$

(this quantity is a vector, so the inequality means that all elements must be greater than zero) then the model is guaranteed to be robust around this data point. Specifically, there does not exist an adversarial example $\tilde{x}$ such that $\|\tilde{x} - x\|_\infty \leq \epsilon$ and $f_\theta(\tilde{x}) \neq y^\star$.

  - This is the example of using MILP and cannot be generalized to ImageNet

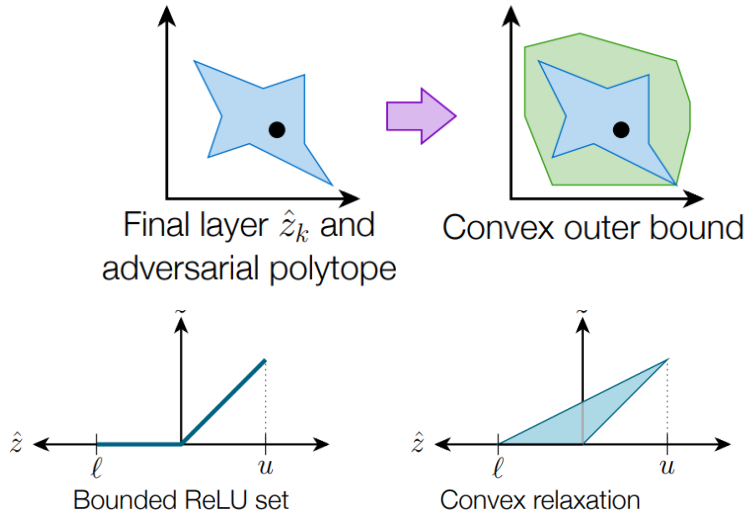


Figure 2. Illustration of the convex ReLU relaxation over the bounded set $[\ell, u]$.

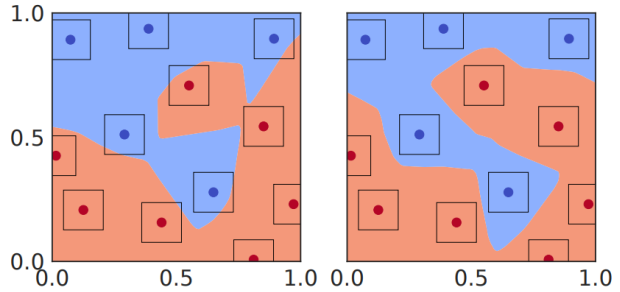
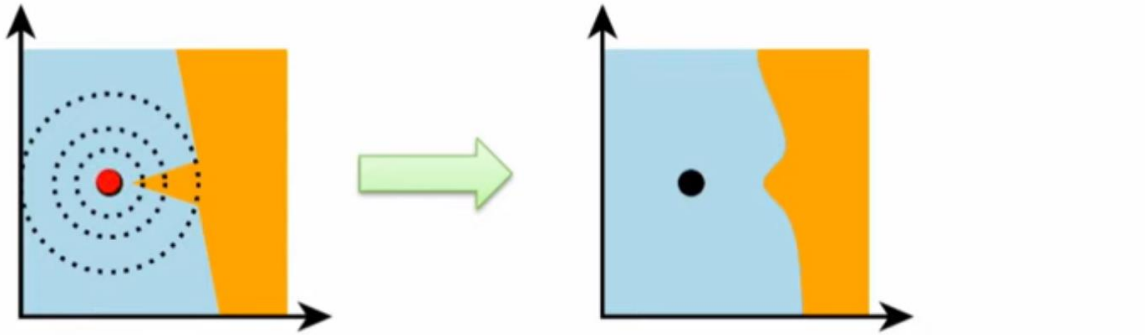


Figure 3. Illustration of classification boundaries resulting from standard training (left) and robust training (right) with ℓ_∞ balls of size $\epsilon = 0.08$ (shown in figure).

[1] Wong, Eric, and J. Zico Kolter. "Provable defenses against adversarial examples via the convex outer adversarial polytope."

Adversarial examples and boundary curvature

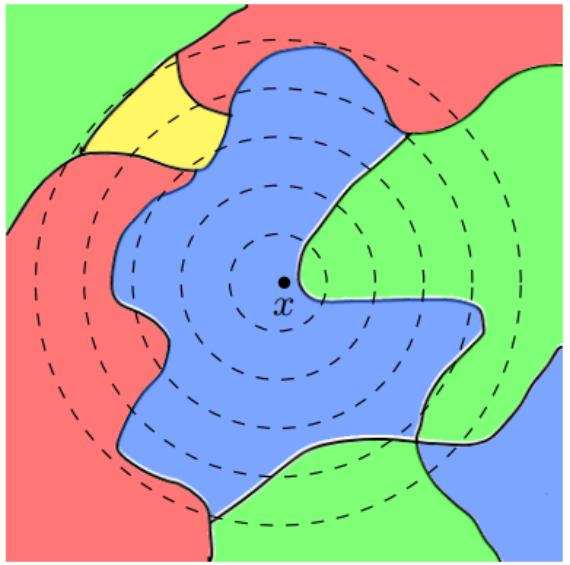
- Very curved boundary will lead to adversarial examples looking very similar to ones near the classification boundary
- So let's diminish this curvature spike influence!
  - Different approaches exist e.g. by *Lecuyer et al.*[1] *and Li et al.*[2], but the most famous one is by *Cohen et al*.

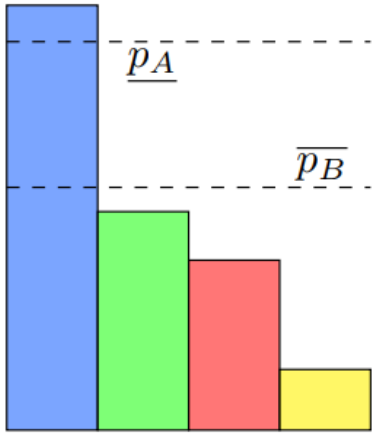


[1] Lecuyer, Mathias, et al. "Certified robustness to adversarial examples with differential privacy."
[2] Li, Bai, et al. "Certified Adversarial Robustness with Additive Noise."

Randomized Smoothing[1]

- **Main idea**:
  - Let's use another definition of classifier!
  - New classifier (in fact, sort of TTA):
    - $g(x) = \underset{c \in Y}{\operatorname{argmax}} P(f(x+\varepsilon) = c), \varepsilon \sim N(0, \sigma^2 I)$
- The main robustness result:
  - If $f(x)$ classifier is robust under Gaussian noise,
  - Then $g(x)$ classifier is robust under **ANY** noise
- The radius R in **Theorem 1** is tight: with the bigger radius there exists an adversarial example





Theorem 1. *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function, and let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let g be defined as in (1). Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy:*

$$\mathbb{P}(f(x+\varepsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c) \quad (2)$$

Then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < R$, where

$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \quad (3)$$

Theorem 2. *Assume $\underline{p_A} + \overline{p_B} \leq 1$. For any perturbation δ with $\|\delta\|_2 > R$, there exists a base classifier f consistent with the class probabilities (2) for which $g(x+\delta) \neq c_A$.*

[1] Cohen, Jeremy M., Elan Rosenfeld, and J. Zico Kolter. "Certified adversarial robustness via randomized smoothing."

Randomized smoothness: some interesting facts

- Some interesting details about linear classifier for two classes: $f(x) = sign(w^T x + b)$
  - It is the smoothed version of itself:

Proposition 3. *If f is a two-class linear classifier $f(x) = sign(w^T x + b)$, and g is the smoothed version of f with any σ, then $g(x) = f(x)$ for any x (where f is defined).*

  - Certified radius of it is the distance between the point x and the boundary:

Proposition 4. *If f is a two-class linear classifier $f(x) = sign(w^T x + b)$, and g is the smoothed version of f with any σ, then invoking Theorem 1 at any x (where f is defined) with $\underline{p_A} = p_A$ and $\overline{p_B} = p_B$ will yield the certified radius $R = \frac{|w^T x + b|}{\|w\|}$.*

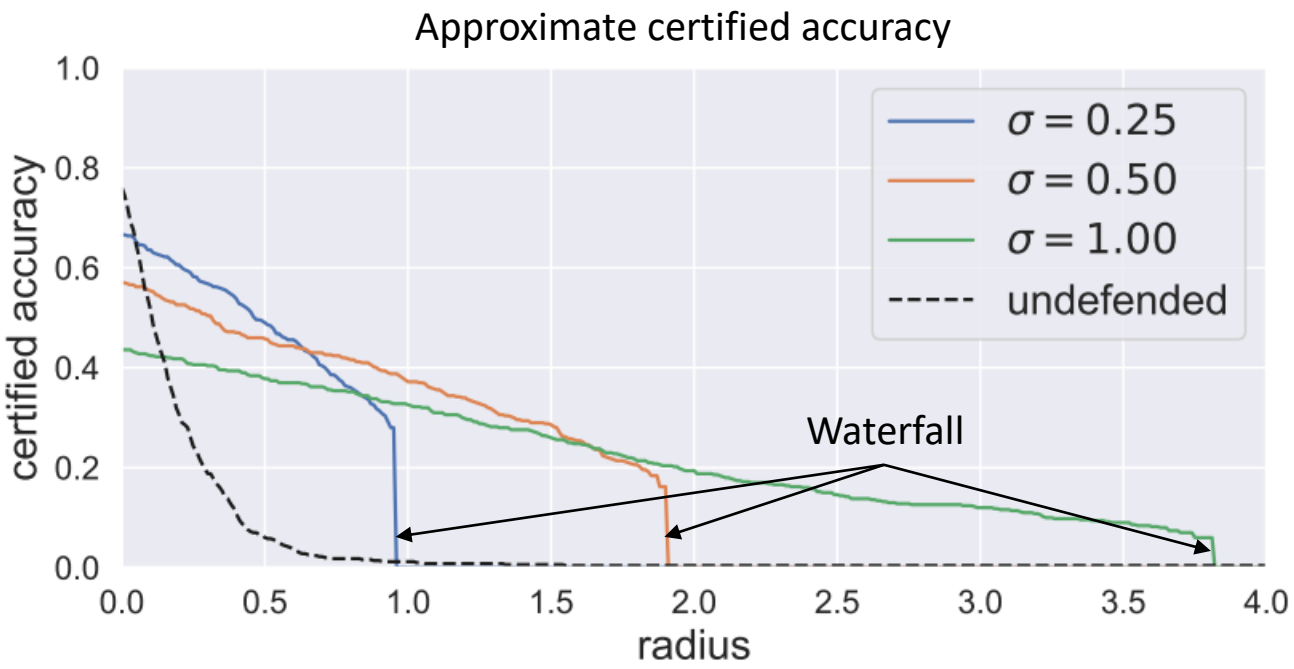
- But there always exists a classifier with real robustness radius more than R from **Theorem 1**:

Proposition 6. *For any $\tau > 0$, there exists a base classifier f and an input x_0 for which the corresponding smoothed classifier g is robust around x_0 at radius ∞, yet Theorem 1 only certifies a radius of τ around x_0.*

Randomized smoothness: results

- The authors propose the procedure to return the radius R and output class c based on input x and the deviance of noise σ
  - This procedure can even avoid to provide the answer with some probability α
- To certify the classifiers, authors **trained the base models with Gaussian noise from $N(\mathbf{0}, \boldsymbol{\sigma}^2 \boldsymbol{I})$** – in fact, to make the classifier $f(x)$ more robust to Gaussian noise
- Trained models are compared using "**approximate certified accuracy**":
  - For each test radius $\delta = r$ the fraction of examples is returned on which CERTIFY
    - Provides the answer
    - Returns the correct class
    - Returns a radius R so as $r \leq R$
- Also when estimating the $g(x)$ authors run Monte Carlo N times

Randomized smoothness: results on ImageNet (1)



Waterfall just because we the trained model is robust usually under some $r \leq R$

Table 1. Approximate certified accuracy on ImageNet. Each row shows a radius r, the best hyperparameter σ for that radius, the approximate certified accuracy at radius r of the corresponding smoothed classifier, and the standard accuracy of the corresponding smoothed classifier. To give a sense of scale, a perturbation with ℓ_2 radius 1.0 could change one pixel by 255, ten pixels by 80, 100 pixels by 25, or 1000 pixels by 8. Random guessing on ImageNet would attain 0.1% accuracy.

ℓ_2 RADIUS	BEST σ	CERT. ACC (%)	STD. ACC(%)
0.5	0.25	49	67
1.0	0.50	37	57
2.0	0.50	19	57
3.0	1.00	12	44

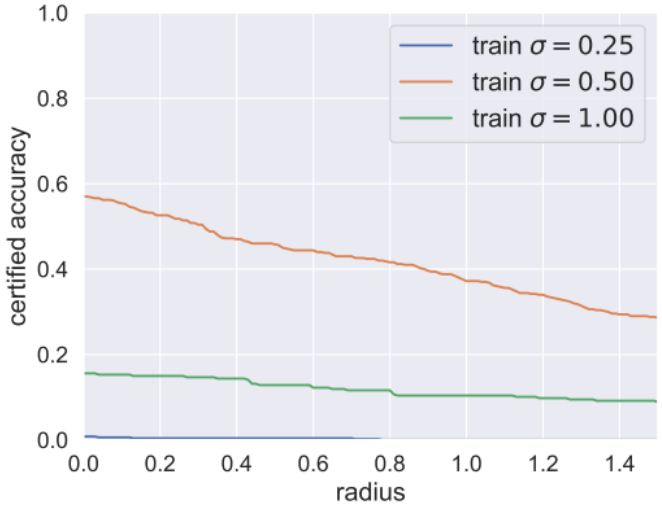
Randomized smoothness: results on ImageNet (2)

- In fact, with NN $f(x)$ we can have larger real robustness radius than R from **Theorem 1**:
  - Authors just tried to find the real adversaries under $r > R$ and measure the success of the attack
  - The lower success the more robust the model
  - $r = 1.5R$: 17% success
  - $r = 2R$: 53% success

Influence of N for Monte Carlo prediction

N	CORRECT, ACCURATE
100	0.65
1000	0.68
10000	0.69

Influence of training σ when testing with $\sigma = 0.5$



Improvement: Adversarial training for smoothed classifier[1]

- Instead of simple augmenting the training example with Gaussian noise, let's do in fact adversarial training using attacks on $g(x)$!

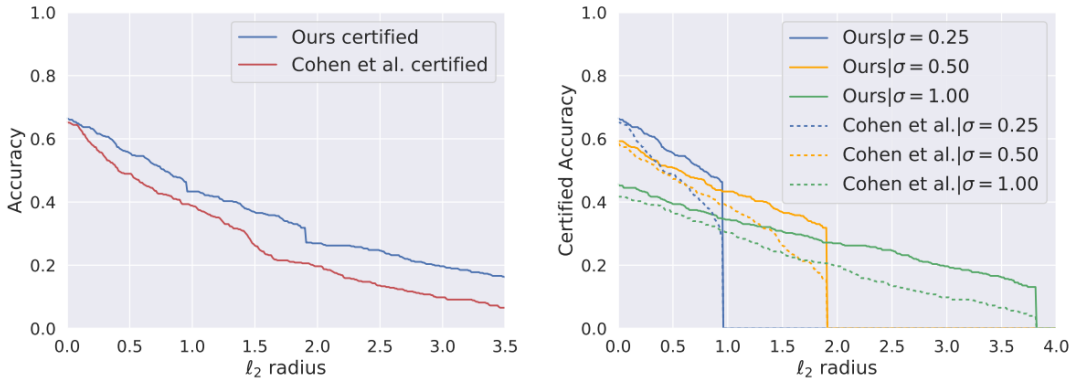
$$\hat{x} = \underset{\|x'-x\|_2 \leq \epsilon}{\arg\max}\ \ell_{\mathrm{CE}}(G(x'), y)$$

$$= \underset{\|x'-x\|_2 \leq \epsilon}{\arg\max} \left(-\log \underset{\delta \sim \mathcal{N}(0,\sigma^2 I)}{\mathbb{E}} \left[(F(x'+\delta))_y \right] \right)$$

$$\nabla_{x'} J(x') = \nabla_{x'} \left(-\log \underset{\delta \sim \mathcal{N}(0,\sigma^2 I)}{\mathbb{E}} [F(x'+\delta)_y] \right)$$

$$\nabla_{x'} J(x') \approx \nabla_{x'} \left(-\log \left(\frac{1}{m} \sum_{i=1}^{m} F(x'+\delta_i)_y \right) \right)$$

Comparison with the original on the ImageNet



[1] Salman, Hadi, et al. "Provably robust deep learning via adversarially trained smoothed classifiers."

Improvement: Certified Robustness for Top-k Predictions[1]

- The very same idea for Randomized Smoothing, but not only for the top class, but for Top-k classes:
  - $g_k(x) = \underset{c \in Y}{\text{argmax}_{1:k}}\, P(f(x+\varepsilon) = c)$, $\varepsilon \sim N(0, \sigma^2 I)$
- Needed to improve top-5 ImageNet:
  - Certified top-1/top-3/top-5 accuracies = 46.6% / 57.8% / 62.8% when $||\delta||_2 = 0.5$

Theorem 1 (Certified Radius for Top-k Predictions). *Suppose we are given an example* $\mathbf{x}$*, an arbitrary base classifier* f*,* $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$*, a smoothed classifier* g*, an arbitrary label* $l \in \{1, 2, \cdots, c\}$*, and* $\underline{p_l}, \overline{p}_1, \cdots, \overline{p}_{l-1}, \overline{p}_{l+1}, \cdots, \overline{p}_c \in [0, 1]$ *that satisfy the following conditions:*

$$Pr(f(\mathbf{x}+\epsilon) = l) \geq \underline{p_l} \text{ and } Pr(f(\mathbf{x}+\epsilon) = i) \leq \overline{p}_i, \forall i \neq l, \tag{1}$$

where $\underline{p}$ *and* $\overline{p}$ *indicate lower and upper bounds of* p*, respectively. Let* $\overline{p}_{b_k} \geq \overline{p}_{b_{k-1}} \geq \cdots \geq \overline{p}_{b_1}$ *be the* k *largest ones among* $\{\overline{p}_1, \cdots, \overline{p}_{l-1}, \overline{p}_{l+1}, \cdots, \overline{p}_c\}$*, where ties are broken uniformly at random. Moreover, we denote by* $S_t = \{b_1, b_2, \cdots, b_t\}$ *the set of* t *labels with the smallest probability upper bounds in the* k *largest ones and by* $\overline{p}_{S_t} = \sum_{j=1}^{t} \overline{p}_{b_j}$ *the sum of the* t *probability upper bounds, where* $t = 1, 2, \cdots, k$*. Then, we have:*

$$l \in g_k(\mathbf{x}+\delta), \forall ||\delta||_2 < R_l, \tag{2}$$

where R_l *is the unique solution to the following equation:*

$$\Phi(\Phi^{-1}(\underline{p_l}) - \frac{R_l}{\sigma})) - \min_{t=1}^{k} \frac{\Phi(\Phi^{-1}(\overline{p}_{S_t}) + \frac{R_l}{\sigma}))}{t} = 0, \tag{3}$$

where Φ *and* Φ^{-1} *are the cumulative distribution function and its inverse of the standard Gaussian distribution, respectively.*

Theorem 2 (Tightness of the Certified Radius). *Assuming we have* $\underline{p_l} + \sum_{j=1}^{k} \overline{p}_{b_j} \leq 1$ *and* $\underline{p_l} + \sum_{i=1,\cdots,l-1,l+1,\cdots,c} \overline{p}_i \geq 1$*. Then, for any perturbation* $||\delta||_2 > R_l$*, there exists a base classifier* f^* *consistent with (1) but we have* $l \notin g_k(\mathbf{x}+\delta)$*.*

[1] Jia, Jinyuan, et al. "Certified Robustness for Top-k Predictions against Adversarial Perturbations via Randomized Smoothing."

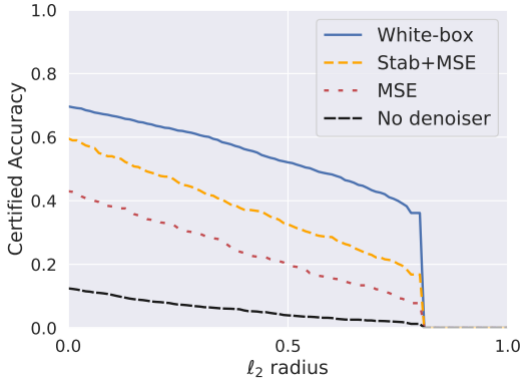
BlackBox for Randomized Smoothing[1]

- What if we cannot change pretrained classifier, but want to increase its certified robustness?
- Let's train denoiser D used after we add Gaussian noise!
- And then simply apply majority rule
- Denoiser: trained with two losses for every Gaussian σ
  - MSE
  - Stability (CE loss)

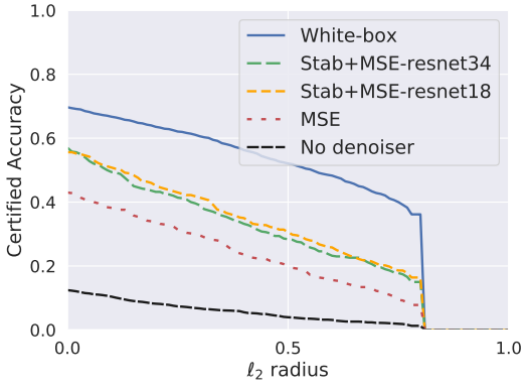
Table 1. Certified top-1 accuracy of ResNet-50 on **ImageNet** at various ℓ_2 radii (Standard accuracy is in parenthesis).

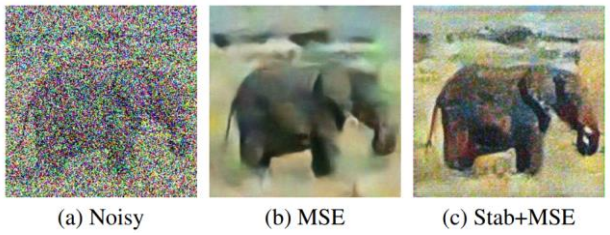
ℓ_2 RADIUS (IMAGENET)	0.25	0.5	0.75	1.0	1.25	1.5
WHITE-BOX SMOOTHING (COHEN ET AL., 2019) (%)	$^{(70)}62$	$^{(70)}52$	$^{(62)}45$	$^{(62)}39$	$^{(62)}34$	$^{(50)}29$
NO DENOISER (BASELINE) (%)	$^{(49)}32$	$^{(12)}4$	$^{(12)}2$	$^{(0)}0$	$^{(0)}0$	$^{(0)}0$
BLACK-BOX SMOOTHING (QUERY ACCESS) (%)	$^{(69)}48$	$^{(56)}31$	$^{(56)}19$	$^{(34)}12$	$^{(34)}7$	$^{(30)}4$
BLACK-BOX SMOOTHING (FULL ACCESS) (%)	$^{(67)}50$	$^{(60)}33$	$^{(60)}20$	$^{(38)}14$	$^{(38)}11$	$^{(38)}6$

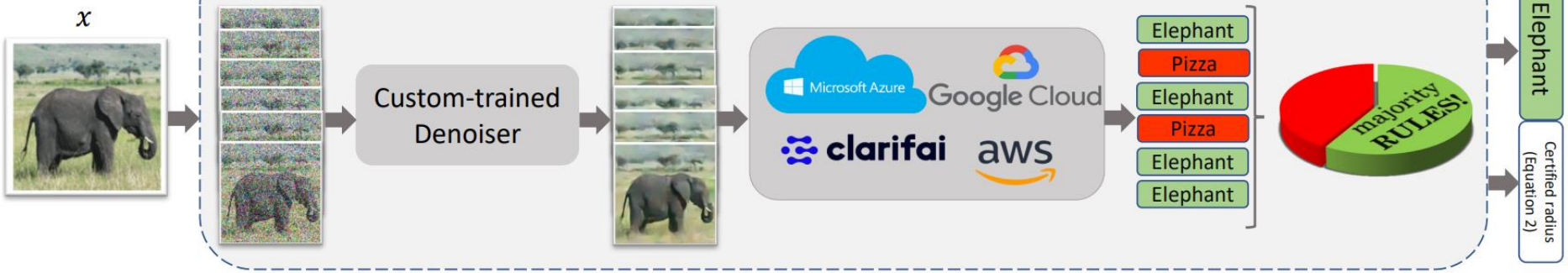
ImageNet, ResNet-50, Full-access



Query-access







[1] Salman, Hadi, et al. "Black-box Smoothing: A Provable Defense for Pretrained Classifiers."

High Dimension case: noise variance[1]

- d: input dimension ($d = h \times w$)
- η: input noise
- ϵ: robustness radius
- δ: diff between top-1 and top-2 class scores

Table 1: Certified ℓ_∞ robustness at a radius of 2/255 on the CIFAR-10 dataset (without extra unlabelled data or pre-trained model).

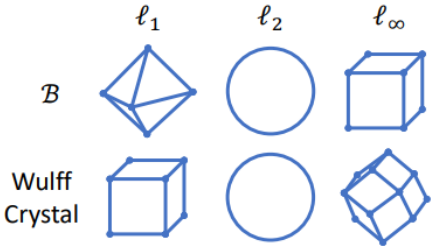
Method	Certified Robust Accuracy	Natural Accuracy
TRADES + Random Smoothing	62.6%	78.8%
Salman et al. (2019)	60.8%	82.1%
Zhang et al. (2020)	54.0%	72.0%
Wong et al. (2018)	53.9%	68.3%
Mirman et al. (2018)	52.2%	62.0%
Gowal et al. (2018)	50.0%	70.2%
Xiao et al. (2019)	45.9%	61.1%

- **Main result**: any noise distribution that provides ℓ_p robustness for all base classifiers with p >= 2 for 99% of the features (pixels) must satisfy $\mathbb{E}\,\eta_i^2 = \Omega(d^{1-2/p}\epsilon^2(1-\delta)/\delta^2)$
  - **Corollary 1**: for high-dimensional images the required noise will eventually dominate the useful information in the images, leading to trivial smoothed classifier. For $p = \infty$, noise variance grows linearly with dimension
  - **Corollary 2**: for $p = 2$ noise variance is independent on the dimension => non-trivial smoothed classifiers
  - **Corollary 3**: defending against adversarial attacks in the ℓ_p ball of radius ϵ by random smoothing is almost as hard as defending against attacks in the ℓ_2 ball of radius $\epsilon d^{\frac{1}{2}-\frac{1}{p}}$
- **Experiments**: for ℓ_∞ hard to achieve promising robust accuracy (≥ 70%) even when the perturbation radius is as small as 2 pixels
- **Proposal**: to use dimension-reduction techniques

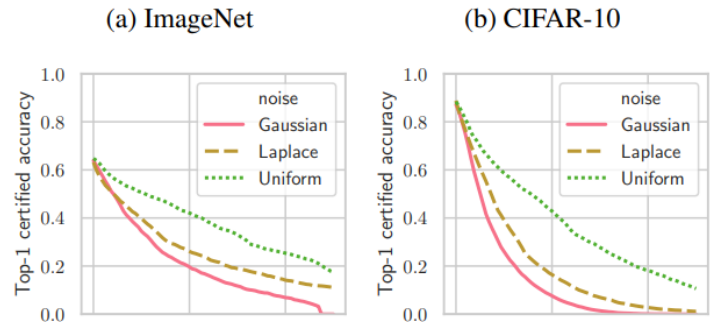
[1] Blum A., et al. "Random Smoothing Might be Unable to Certify ℓ∞ Robustness for High-Dimensional Images"

High Dimension case: Wulff Crystals[1]

- **Main results**:
  - SotA result for ℓ_1-certified radius
  - **NO-GO** Theorem:
  - *Without using more than the information of the probability ρ of correctly classifying an input under random noise, no smoothing techniques can certify nontrivial robust accuracy at ℓ_∞-radius $\Omega(d^{-1/2})$, or at ℓ_1 or ℓ_2-radius $\Omega(1)$*
  - *Still not clear about NO-GO theorem for multiclass case*
- **Proposal**:
  - Better usage of base classifier or multi-class structure



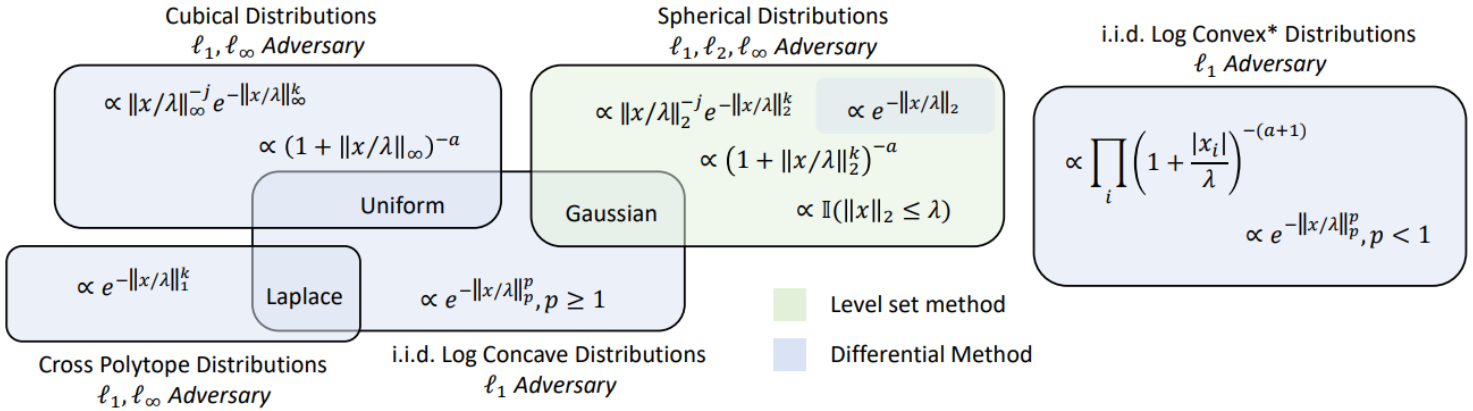
Definition 5.1. The *Wulff Crystal* (w.r.t. $\mathcal{B}$) is defined as the unit ball of the norm dual to $\|\cdot\|_*$, where $\|x\|_* = \mathbb{E}_{y\sim \mathrm{Vert}(\mathcal{B})} |\langle x, y\rangle|$ and y is sampled uniformly from the vertices of $\mathcal{B}$ [5].



Randomized Smoothing of All Shapes and Sizes

ImageNet	ℓ_1 Radius	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
	Laplace, Teng et al. (2019) (%)	48	40	31	26	22	19	17	14
	Uniform, Ours (%)	55	49	46	42	37	33	28	25
	+ Stability Training	**60**	**55**	**51**	**48**	**45**	**43**	**41**	**39**
CIFAR-10	ℓ_1 Radius	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
	Laplace, Teng et al. (2019) (%)	61	39	24	16	11	7	4	3
	Uniform, Ours (%)	70	59	51	43	33	27	22	18
	+ Stability Training	70	60	53	47	**43**	39	35	28
	+ Stability Training, Semi-supervision	**74**	**63**	54	**48**	**43**	38	34	31
	+ Stability Training, Pre-training	**74**	62	**55**	**48**	**43**	**40**	**37**	**33**

Table 1. Certified top-1 accuracies of our ℓ_1-robust classifiers, vs previous state-of-the-art, at various radii, for ImageNet and CIFAR-10.[3]



[1] Yang G. et al. "Randomized smoothing of all shapes and sizes"

High Dimension case: randomized smoothing[1]

- r_p^*: largest ℓ_p-radius that can be certified
- **Main results**:
  - General smoothing noise with σ^2-variance: $r_p^* \leq \frac{\sigma}{2\sqrt{2}d^{\frac{1}{2}-\frac{1}{p}}}\left(\frac{1}{\sqrt{1-p_1(x)}}+\frac{1}{\sqrt{p_2(x)}}\right)$
  - Generalized Gaussian distribution (including Laplacian, Gaussian, uniform) with σ^2-variance $r_p^* \leq \frac{2\sigma}{d^{\frac{1}{2}-\frac{1}{p}}}\left(\sqrt{\log\frac{1}{1-p_1(x)}}+\sqrt{\log\frac{1}{p_2(x)}}\right)$
  - Smoothing inside ℓ_∞-ball of radius b: $r_p^* < \frac{2b}{d^{1-\frac{1}{p}}}$
  - Smoothing inside ℓ_1-ball of radius b: $r_p^* < \frac{2b}{d}$
- **Corollary**: for p>=2 exact estimation

$$r_p = \frac{\sigma}{2d^{\frac{1}{2}-\frac{1}{p}}}\left(\Phi^{-1}\left(p_1(x)\right)-\Phi^{-1}\left(p_2(x)\right)\right)$$

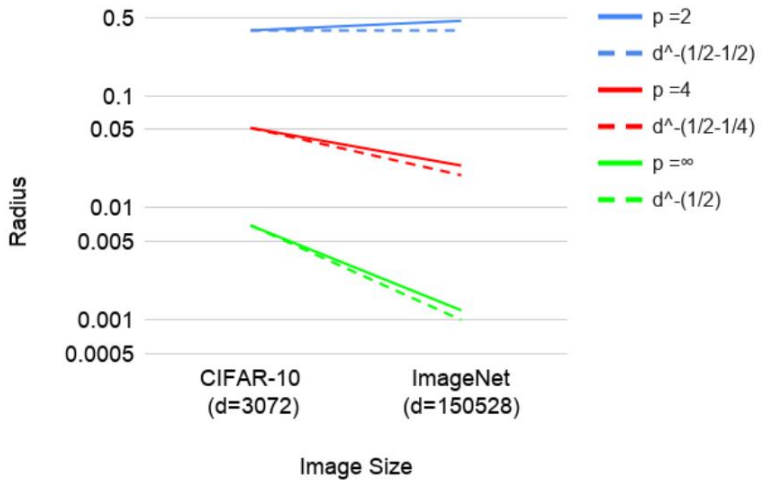
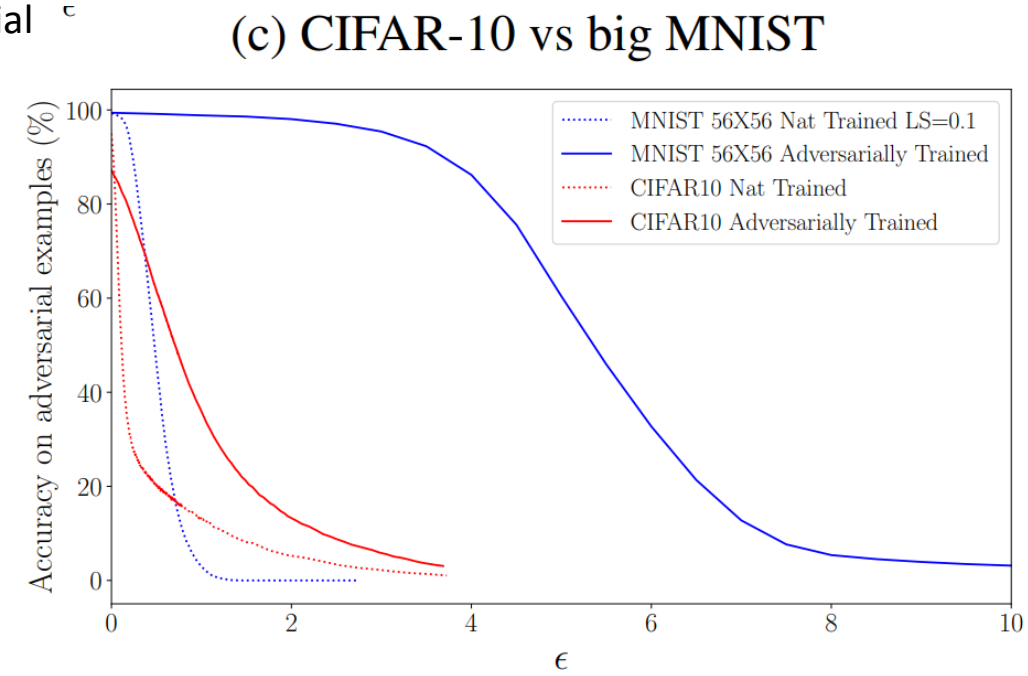


Figure 9. Certified Radius using Gaussian noise ($\sigma = .25$), for datasets of different image resolutions. We see that for $p > 2$, the certificates (solid lines) decrease with higher dimensionality almost as quickly as one would expect from the explicit dependence on d in Equation 1 (dashed lines).

[1] Kumar A. et al. "Curse of dimensionality on randomized smoothing for certifiable robustness"

High Dimension case: dataset complexity[1]

- **Main results**:
  - No fundamental link between dimensionality and robustness
  - Data distribution, and not dimensionality, is the primary cause of adversarial susceptibility
  - P* = min(2,p)
  - On (n-1)-dim unit sphere: certified radius ϵ exist with probability $C \times (\exp(-\frac{n-1}{2}\epsilon^2))$
    - [so it is VERY small]
  - On n-dim hypercube: certified radius ϵ exist with probability $C \times O(\exp(-2\pi n^{1-\frac{2}{p*}}\epsilon^2)/2\pi\epsilon n^{\frac{1}{2}-\frac{1}{p*}})$
    - In case of p>=2 we have $C \times O(\exp(-2\pi\epsilon^2)/2\pi\epsilon)$
    - C is dependent on the 1/variance of some class
    - [so it can be quite high]
  - For ℓ_0-ball: certified radius ϵ exist with probability $C \times O(\exp(-\epsilon^2/n))$
- **Corollary**:
  - Highly concentrated datasets (with big C) can be relatively safe from adversarial examples



[1] Shafahi A. et al. "Are adversarial examples inevitable?"

Functional approach to randomized smoothing[1]

- **Notations:**

$$f^{\sharp}_{\pi_0}(\boldsymbol{x}_0) := \mathbb{E}_{\boldsymbol{z}\sim\pi_0}\left[f^{\sharp}(\boldsymbol{x}_0+\boldsymbol{z})\right] \quad \mathcal{F}_{[0,1]} = \left\{f : f(\boldsymbol{z}) \in [0,1], \forall \boldsymbol{z} \in \mathbb{R}^d\right\} \quad \pi_{\boldsymbol{\delta}} \text{ the distribution of } \boldsymbol{z}+\boldsymbol{\delta} \text{ when } \boldsymbol{z}\sim\pi_0$$

  - Certification as optimization task: $\min_{\boldsymbol{\delta}\in\mathcal{B}} f^{\sharp}_{\pi_0}(\boldsymbol{x}_0+\boldsymbol{\delta}) \geq \min_{f\in\mathcal{F}}\min_{\boldsymbol{\delta}\in\mathcal{B}}\left\{f_{\pi_0}(\boldsymbol{x}_0+\boldsymbol{\delta}) \text{ s.t. } f_{\pi_0}(\boldsymbol{x}_0) = f^{\sharp}_{\pi_0}(\boldsymbol{x}_0)\right\}$
  - Lagrangian: $\mathcal{L}_{\pi_0}(\mathcal{F},\mathcal{B}) = \min_{f\in\mathcal{F}}\min_{\boldsymbol{\delta}\in\mathcal{B}}\max_{\lambda\in\mathbb{R}}\left\{f_{\pi_0}(\boldsymbol{x}_0+\boldsymbol{\delta}) - \lambda(f_{\pi_0}(\boldsymbol{x}_0) - f^{\sharp}_{\pi_0}(\boldsymbol{x}_0))\right\}$
- **Main results**:
  - Unified functional optimization perspective for different smoothing distribution
  - Theorem:

$$\mathcal{L}_{\pi_0}(\mathcal{F},\mathcal{B}) \geq \max_{\lambda\geq 0}\left\{\lambda f^{\sharp}_{\pi_0}(\boldsymbol{x}_0) - \max_{\boldsymbol{\delta}\in\mathcal{B}}\mathbb{D}_{\mathcal{F}}\left(\lambda\pi_0 \parallel \pi_{\boldsymbol{\delta}}\right)\right\}$$

$$\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_0 \parallel \pi_{\boldsymbol{\delta}}\right) = \int \left(\lambda\pi_0(\boldsymbol{z}) - \pi_{\boldsymbol{\delta}}(\boldsymbol{z})\right)_+ d\boldsymbol{z},$$

  - For specific types of smoothing distributions we can calculate it more analytically

ℓ_1 Radius (ImageNet)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
Baseline (%)	50	41	33	29	25	18	15
Ours (%)	**51**	**42**	**36**	**30**	**26**	**22**	**16**

ℓ_2 Radius (ImageNet)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
Baseline (%)	49	37	29	19	15	12	9
Ours (%)	**50**	**39**	**31**	**21**	**17**	**13**	**10**

l_∞ Radius CIFAR	2/255	4/255	6/255	8/255	10/255	12/255
Baseline (%)	58	42	31	25	18	13
Ours (%)	**60**	**47**	**38**	**32**	**23**	**17**

$$\ell_1 : \pi_0(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_1^{-k}\exp\left(-\frac{\|\boldsymbol{z}\|_1}{b}\right)$$

$$\ell_2 : \pi_0(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_2^{-k}\exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$$

$$\ell_0 : \pi_0(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_\infty^{-k}\exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$$

Theorem 2. *Consider the ℓ_1 attack with $\mathcal{B} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1 \leq r\}$ and smoothing distribution $\pi_0(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_1^{-k}\exp\left(-\frac{\|\boldsymbol{z}\|_1}{b}\right)$ with $k \geq 0$ and $b > 0$, or the ℓ_2 attack with $\mathcal{B} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq r\}$ and smoothing distribution $\pi_0(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_2^{-k}\exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$ with $k \geq 0$ and $\sigma > 0$. Define $\boldsymbol{\delta}^* = [r, 0, ..., 0]^\top$, we have*

$$\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_0 \parallel \pi_{\boldsymbol{\delta}^*}\right) = \max_{\delta\in\mathcal{B}}\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_0 \parallel \pi_{\boldsymbol{\delta}}\right)$$

Theorem 3. *Consider the ℓ_∞ attack with $\mathcal{B}_{\ell_\infty,r} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_\infty \leq r\}$ and the mixed norm smoothing distribution in Eq.13 with $k \geq 0$ and $\sigma > 0$. Define $\boldsymbol{\delta}^* = [r, r, ..., r]^\top$. We have for any $\lambda > 0$,*

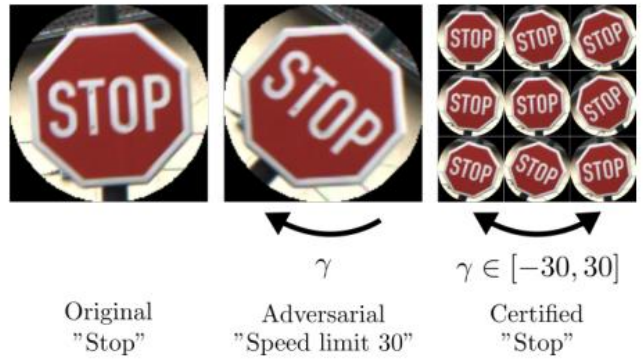
$$\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_0 \parallel \pi_{\boldsymbol{\delta}^*}\right) = \max_{\delta\in\mathcal{B}}\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_0 \parallel \pi_{\boldsymbol{\delta}}\right).$$

- **Corollary**:
  - Trade-off between accuracy and robustness:

$$\max_{\lambda\geq 0}\left[\underbrace{\lambda f^{\sharp}_{\pi_0}(\boldsymbol{x}_0)}_{\text{Accuracy}} + \underbrace{\left(-\max_{\boldsymbol{\delta}\in\mathcal{B}}\mathbb{D}_{\mathcal{F}}\left(\lambda\pi_0 \parallel \pi_{\boldsymbol{\delta}}\right)\right)}_{\text{Robustness}}\right]$$

[1] Zhang D. et al. “Black-box certification with randomized smoothing: A functional optimization based framework”

Semantic perturbations certified robustness



- Let's certify semantic perturbations!
  - In fact, rotations and translations are studied: $\psi_\beta : \mathbb{R}^n \to \mathbb{R}^n$
  - Smoothed classifier: $g(\boldsymbol{x}) = \arg\max_c \mathbb{P}_{\beta \sim \mathcal{N}(0,\sigma^2\mathbb{1})} (f \circ \psi_\beta(\boldsymbol{x}) = c)$
  - Also interpolation procedure is taken into account because after rotation we need to interpolate anyway

Theorem 4.2. *Let $\boldsymbol{x} \in \mathbb{R}^n$, $f : \mathbb{R}^m \to \mathcal{Y}$ be a classifier and $\psi_\beta : \mathbb{R}^n \to \mathbb{R}^m$ be a composable transformation as above. If*

$$\mathbb{P}_\beta(f \circ \psi_\beta(x) = c_A) = p_A \geq \underline{p_A} \geq \overline{p_B} \geq p_B = \max_{c_B \neq c_A} \mathbb{P}_\beta(f \circ \psi_\beta(x) = c_B),$$

then $g \circ \psi_\gamma(\boldsymbol{x}) = c_A$ for all γ satisfying

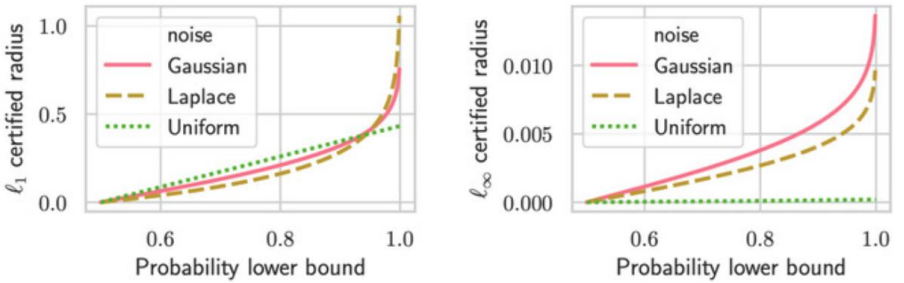
$$\|\gamma\|_2 < \tfrac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) =: r_\gamma.$$

Rotation						r_γ percentile		
Dataset	$\mathcal{I}$	σ_γ	α_γ	f Acc.	g Acc.	25^{th}	50^{th}	75^{th}
ImageNet	bil.	10	0.001	0.39	0.29	10.81	10.81	10.81
ImageNet	bil.	10	0.001	0.39	0.29	18.29	18.29	18.29
ImageNet	bil.	30	0.001	0.39	0.28	9.09	16.59	28.60
ImageNet	bil.	30	0.001	0.39	0.28	20.22	25.36	$30^\dagger$
ImageNet	bic.	10	0.001	0.39	0.29	10.40	10.40	10.40
ImageNet	bic.	30	0.001	0.39	0.27	9.33	17.00	28.74
ImageNet	near.	10	0.001	0.39	0.29	9.62	9.62	9.62
ImageNet	near.	30	0.001	0.39	0.26	7.38	16.63	27.72

Translation						r_γ percentile		
Dataset	$\mathcal{I}$	σ_γ	α_γ	f Acc.	g Acc.	25^{th}	50^{th}	75^{th}
ImageNet	bil.	50	0.001	0.48	0.36	2.4%	2.4%	2.4%
ImageNet	bic.	50	0.001	0.48	0.36	2.4%	2.4%	2.4%

[1] Fischer, Marc, Maximilian Baader, and Martin Vechev. "Certification of Semantic Perturbations via Randomized Smoothing."

Takeaway



- Certification is only for much smaller regions than humans can do
- Certified robustness is better than empirical adversarial training in certification, but worse than clean performance (and too much time to train)
- Using l_p-balls is neither necessary nor sufficient for perceptual robustness
- Other types of randomized smoothing could be taking into account: e.g. *Uniform*[1] or *Laplacian*[2]
- Randomized smoothing requires multiple inferences ☹
- High dimensionality <-> complex datasets <-> ℓ_∞-ball influence
- BTW some note about physical nature of l_p-balls:
  - l_2: corresponds to the power of signals
  - l_1: corresponds to the pixel mass
  - l_∞: corresponds to the noise in camera sensors
  - l_0: corresponds to the practical robustness

[1] Lee, Guang-He, et al. "Tight certificates of adversarial robustness for randomly smoothed classifiers."
[2] Teng, Jiaye, et al. "l_1 Adversarial Robustness Certificates: a Randomized Smoothing Approach"

Thank you!

(need to certify everything)