Tight Certificates of Adversarial Robustness for Randomly Smoothed Classifiers

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi S. Jaakkola

Massachusetts Institute of Technology

CSAIL

Summary

- A general approach to deriving tight certificates of robustness for randomly smoothed classifiers.
- We focus on ℓ_0-robustness in discrete spaces.
- We show how certificates can be tightened with additional assumptions about the classifier.

Introduction

- Adversarial examples can be easily found on deep models

x x + Δx

Input

Accuracy can drop dramatically!

Predictions Rainbow Panda

  - Ideally, we want a model without adversarial example.

- If a heuristic search algorithm fails, there may still be adversarial examples.
- We need a certificate to show that no such example exists around a specified radius of the input example.
- Finding certificates is particularly challenging in discrete spaces as the problem is combinatorial.

Set-up & background

- Given an input $\boldsymbol{x} \in \mathcal{X}$, a randomization scheme ϕ assigns a distribution $\Pr(\phi(\boldsymbol{x}) = \boldsymbol{z})$ for each $\boldsymbol{z} \in \mathcal{X}$.
- We use a randomly smoothed classifier $f(\phi(\boldsymbol{x}))$.
  - f is a base classifier (e.g., a deep net / decision tree).
  - $\Pr(f(\phi(\boldsymbol{x})) = y)$ is abbreviated as p.
- Tight certificates exist with Gaussian randomization and ℓ_2 metric (Cohen et al., 19').

Our framework

- A tight point-wise robustness certificate for $\bar{\boldsymbol{x}}$:

$$\rho_{\boldsymbol{x},\bar{\boldsymbol{x}}}(p) \triangleq \min_{\bar{f} \in \mathcal{F}: \Pr(\bar{f}(\phi(\boldsymbol{x}))=y)=p} \Pr(\bar{f}(\phi(\bar{\boldsymbol{x}})) = y) \le \Pr(f(\phi(\bar{\boldsymbol{x}})) = y)$$

  - It can be solved by Neyman-Pearson lemma

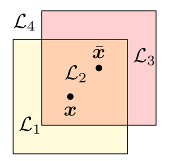
- A regional certificate of robustness:
  - Define $\mathcal{B}_{r,q}(\boldsymbol{x}) \triangleq \{\bar{\boldsymbol{x}} \in \mathcal{X} : \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_q \le r\}$
  - $R(\boldsymbol{x}, p, q) \triangleq \sup r \ s.t. \min_{\bar{\boldsymbol{x}} \in \mathcal{B}_{r,q}(\boldsymbol{x})} \rho_{\boldsymbol{x},\bar{\boldsymbol{x}}}(p) > 0.5$
  - Implication: if $\Pr(f(\phi(\boldsymbol{x})) = y) = p$, then $\forall \bar{\boldsymbol{x}} \in \mathcal{X} : \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_q < R(\boldsymbol{x}, p, q)$, $\Pr(f(\phi(\bar{\boldsymbol{x}})) = y) > 0.5$

A warm-up example

- A uniform randomization scheme:

$$\phi(\boldsymbol{x})_i = \boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \text{Uniform}([-\gamma, \gamma])$$

- Illustration:



- Randomization at $\boldsymbol{x}$ and $\bar{\boldsymbol{x}}$ divide the input space into non-overlapping regions $\mathcal{L}_1, \dots, \mathcal{L}_4$ based on likelihood comparisons
- For any f or $\bar{f}$, only the integral over a region matters; we search for $\bar{f}$ that assigns prob. [0,1] (integral value) to each region.
- Worst case $\bar{f}$ assigns high values to $\mathcal{L}_1$, low values to $\mathcal{L}_2$ and $\mathcal{L}_3$, subject to the constraint that the aggregate = p across $\mathcal{L}_1$ and $\mathcal{L}_2$.

⇨
$$\begin{cases} \rho_{\boldsymbol{x},\bar{\boldsymbol{x}}}(p) = 0, & \text{if } 0 \le p \le (2\gamma)^{-d}|\mathcal{L}_1|, \\ \rho_{\boldsymbol{x},\bar{\boldsymbol{x}}}(p) = p - (2\gamma)^{-d}|\mathcal{L}_1|, & \text{otherwise.} \end{cases}$$

- Regional certificate finds the worst $\bar{\boldsymbol{x}} \in \mathcal{B}_{r,q}(\boldsymbol{x})$ such that $|\mathcal{L}_1|$ is maximized.

⇨
$$R(\boldsymbol{x}, p, q=1) = 2p\gamma - \gamma$$
$$R(\boldsymbol{x}, p, q=\infty) = 2\gamma - 2\gamma(1.5 - p)^{1/d}$$

A discrete distribution for ℓ_0 robustness

- We consider the discrete space: $\mathcal{X} = \{0, \frac{1}{K}, \frac{2}{K}, \dots, 1\}^d$.
- A discrete randomization scheme:

$$\begin{cases} \Pr(\phi(\boldsymbol{x})_i = \boldsymbol{x}_i) = \alpha, \\ \Pr(\phi(\boldsymbol{x})_i = z) = (1-\alpha)/K \triangleq \beta \in (0, 1/K), & \text{if } z \in \{0, \frac{1}{K}, \frac{2}{K}, \dots, 1\} \text{ and } z \ne \boldsymbol{x}_i \end{cases}$$

- Key properties:

> 1. *for all $\boldsymbol{x}, \bar{\boldsymbol{x}}$ such that $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_0 = r$, we have $\rho_{\boldsymbol{x},\bar{\boldsymbol{x}}} = \rho_r$*
> 2. *$\rho_r : [0, 1] \to [0, 1]$ is an increasing bijection*

- Implications:
  - We can pre-compute $\rho_r^{-1}(0.5)$ (we have a $\Theta(d^3)$ algorithm).
  - If $p > \rho_r^{-1}(0.5)$, the prediction is robust within $\mathcal{B}_{r,0}(\boldsymbol{x})$.
  - $R(\boldsymbol{x}, p, q)$ is simply the maximum r s.t. $p > \rho_r^{-1}(0.5)$.

- Key steps for pre-computing $\rho_r^{-1}(0.5)$
  - Similar to the uniform distribution, we partition the space into regions with constant likelihood ratio to simplify the problem.
    - Likelihood ratio: $\Pr(\phi(\boldsymbol{x}) = \boldsymbol{z}) / \Pr(\phi(\bar{\boldsymbol{x}}) = \boldsymbol{z})$.
  - Assigning $\bar{f}(\boldsymbol{z})$ to y in ↓ likelihood ratio computes $\rho_r^{-1}(0.5)$ (Neyman-Pearson lemma. It can be done in $\Theta(d^3)$).
  - A large integer algorithm is needed for high dimension setting.

Towards tighter certification

- The certificates are tight w.r.t. measurable classifiers.
- More characterization of f always improves the point-wise (and regional) certificates: if $f \in \mathcal{F}_\zeta \subset \mathcal{F}$,

$$\min_{\bar{f} \in \mathcal{F}_\zeta: \Pr(\bar{f}(\phi(\boldsymbol{x}))=y)=p} \Pr(\bar{f}(\phi(\bar{\boldsymbol{x}})) = y) \ge \min_{\bar{f} \in \mathcal{F}: \Pr(\bar{f}(\phi(\boldsymbol{x}))=y)=p} \Pr(\bar{f}(\phi(\bar{\boldsymbol{x}})) = y)$$

- Example 1: when $\phi(\boldsymbol{x})_i = \boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$
  - If $\mathcal{X} = \{0, 1\}^d$, we can use (Cohen et al., 19) to derive ℓ_0 certificates due to the bijection to ℓ_2.
  - If we apply denoising before feeding to model:

$$\zeta(\phi(\boldsymbol{x}))_i = \mathbb{I}\{\phi(\boldsymbol{x})_i > 0.5\}, \forall i \in [d]$$

    - The resulting input is equivalent to our discrete randomization scheme.
  - Our certificate is always tighter than using the one derived from the Gaussian distribution in this case.

- Example 2: when f is a decision tree:
  - The randomization can be expressed as a probabilistic routing scheme for each decision node.
  - The exact certificate of robustness can be computed using dynamic programming over tree nodes.

Experiment (project page: http://people.csail.mit.edu/guanghe/randomized_smoothing)

- Evaluation metrics:
  - $\mu(R)$: the average certified radius in testing set.
  - ACC@r: guaranteed accuracy within ℓ_0 radius r.

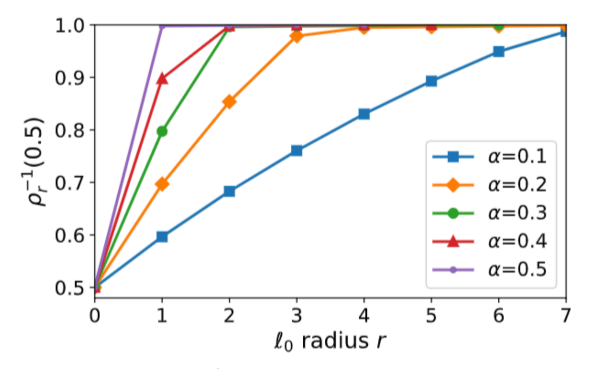
- Binarized MNIST (CNN model).

ϕ	Certificate	$\mu(R)$	ACC@r						
			$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$
Discrete	Discrete	**3.456**	**0.921**	**0.774**	**0.539**	**0.524**	**0.357**	**0.202**	**0.097**
Discrete	Gaussian	1.799	0.830	0.557	0.272	0.119	0.021	0.000	0.000
Gaussian	Gaussian	2.378	0.884	0.701	0.464	0.252	0.078	0.000	0.000

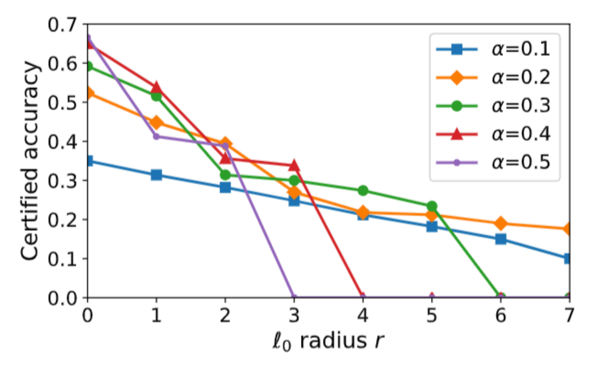
- (Discrete) Exact ACC@1 = 0.954, ACC@2 = 0.926.

- ImageNet (ResNet50 model).

ϕ and certificate	ACC@r						
	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$	$r=6$	$r=7$
Discrete	**0.538**	**0.394**	**0.338**	**0.274**	**0.234**	**0.190**	**0.176**
Gaussian	0.372	0.292	0.226	0.194	0.170	0.154	0.138



$\rho_r^{-1}(0.5)$ for an α



The certified accuracy for an α