Tight Certificates of Adversarial Robustness for Randomly Smoothened Classifiers

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

Summary

- A general approach to deriving tight certificates of robustness for randomly smoothed classifiers.
- We focus on \( \ell_0 \)-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.
- Ideally, we want a model without adversarial examples.
- If a heuristic search algorithm fails, there may still be adversarial examples.
- We need a certificate to show that no such example exists.
- We focus on \( \ell_0 \)-robustness in discrete spaces.

Our framework

- A tight point-wise robustness certificate for \( \mathcal{X} \):
  \[
  \rho_{x,a}(\phi) = \min_{f \in F: \Pr(f(\phi(x)) = y) = p} \Pr(f(\phi(x)) = y) 
  \leq \Pr(f(\phi(x)) = y) 
  \]
  - It can be solved by Neyman-Pearson lemma
- A regional certificate of robustness:
  - Define \( B_{\rho}(x) = \{ x \in \mathcal{X} : |x - \bar{x}| \leq \rho \} \)
  - \( R(x,p,q) = \sup_{p_0 \in B_{\rho}(x)} \min_{p_0 \rightarrow 0} \rho_{x,a}(p_0) > 0.5 \)
  - Implication: if \( \Pr(f(\phi(x)) = y) = p_0 \), then \( \forall x \in \mathcal{X} : |x - \bar{x}| < \rho \) \( R(p_0, q, \rho) \)
  - \( \Pr(f(\phi(x)) = y) > 0.5 \)

A warm-up example

- A uniform randomization scheme:
  \( \phi(x) = x + \epsilon_i, \epsilon_i \sim \mathcal{U}(-\gamma,\gamma) \)
- Illustration:
  \[
  \begin{pmatrix} x \end{pmatrix} 
  \]
- Randomization at \( x \) and \( \mathcal{X} \) divide the input space into non-overlapping regions \( C_1, \ldots, C_4 \) based on likelihood comparisons.
- For any \( f \) or \( \phi \), only the integral over a region matters; we search for \( f \) that assigns prob. \([0,1]\) integral value to each region.
- Worst case \( f \) assigns high values to \( L_0 \), low values to \( L_1, L_2 \) and \( L_4 \), subject to the constraint that the aggregate \( p \) across \( L_1 + L_2 \). \( \rho_{x,a}(p_0) = 0 \), if \( \rho(x) = 0 \), otherwise.
- Regional certificates from the worst \( x \in B_{\rho}(\phi(x)) \) such that \( L_0 \) is maximized.
- \( R(x,p,q) = \sup_{p_0 \rightarrow 0} \Pr(x, p, q) = 1 \)
- \( \Pr(x, p, q) = 1 \)
- \( \Pr(x, p, q) = 1 \)

A discrete distribution for \( \ell_0 \) robustness

- We consider the discrete space: \( x = (0, \frac{1}{k}, \ldots, \frac{1}{k}) \)
- A discrete randomization scheme:
  \[
  \Pr(f(\phi(x)) = y) = \frac{1}{(1 - \beta)K} \geq \beta (0,1/K), \quad y ; \quad \phi(x) = x \)
  \[
  \Pr(f(\phi(x)) = y) = \frac{1}{(1 - \beta)K} \geq \beta (0,1/K), \quad y ; \quad \phi(x) = x \)
- Key properties:
  1. For all \( x \) such that \( |x - \bar{x}| = r \), we have \( \rho_{x,a}(p) = \rho_{x,a}(p) \)
  2. \( \rho_{x,a}(p) \geq 0 \) is an increasing bijection

Implications:

- We can pre-compute \( \rho_{x,a}(p) \) (we have a \( \Theta(\alpha) \) algorithm).
- If \( p > \rho_{x,a}(p) \), the prediction is robust within \( B_{\rho}(x) \).
- \( R(x,p,q) \) is simply the max region \( p > \rho_{x,a}(p) \)

- For steps pre-computing \( \rho_{x,a}(p) \)
  - Similar to the uniform distribution, we partition the space into regions with constant likelihood ratio to simplify the problem.
  - Likelihood ratio: \( \Pr(f(\phi(x)) = y) / \Pr(f(\phi(x)) = y) \)
  - Assigning \( f \) to \( y \) in \( L_0 \) likelihood ratio computes \( \rho_{x,a}(p) \) (Neyman-Pearson lemma). It can be done in \( \Theta(\alpha) \).
- A large integer algorithm is needed for high dimension settings.

Towards tighter certification

- The certificates are tight w.r.t. measurable classifiers.
- More characterization of \( f \) always improves the pointwise (and regional) certificates: \( \min_{p \in F} \Pr(f(\phi(x)) = y) \geq \min_{p \in F} \Pr(f(\phi(x)) = y) \geq \Pr(f(\phi(x)) = y) \)

Example 1: when \( \phi(x) = x + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2) \)
- If \( x = (0, \frac{1}{k}) \), we can use (Cohen et al., 19) to derive \( f \) due to the bijection to \( L_0 \).
- If we apply denoising before feeding to model: \( \Pr(f(\phi(x)) = y) \)
- The result is 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/guangle/rand_smoothing)

- Evaluation metrics:
  - \( \mu(R) \): the average certified radius in testing set.
  - ACC\(\phi \), ACC\(\phi \), guaranteed accuracy within \( L_0 \) radius \( r \).
- Binarized MNIST (CNN model).
- Gain: \( \mu(R) \) at \( 1 \)
- ACC\(\phi \), guaranteed accuracy within \( L_0 \) radius \( r \).

Towards tighter certification

- The certificates are tight w.r.t. measurable classifiers.
- More characterization of \( f \) always improves the pointwise (and regional) certificates: \( \min_{p \in F} \Pr(f(\phi(x)) = y) \geq \min_{p \in F} \Pr(f(\phi(x)) = y) \geq \Pr(f(\phi(x)) = y) \)

Example 1: when \( \phi(x) = x + \epsilon_i, \epsilon_i \sim \mathcal{N}(0, \sigma^2) \)
- If \( x = (0, \frac{1}{k}) \), we can use (Cohen et al., 19) to derive \( f \) due to the bijection to \( L_0 \).
- If we apply denoising before feeding to model: \( \Pr(f(\phi(x)) = y) \)
- The result is 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/guangle/rand_smoothing)

- Evaluation metrics:
  - \( \mu(R) \): the average certified radius in testing set.
  - ACC\(\phi \), ACC\(\phi \), guaranteed accuracy within \( L_0 \) radius \( r \).
- Binarized MNIST (CNN model).
- Gain: \( \mu(R) \) at \( 1 \)
- ACC\(\phi \), guaranteed accuracy within \( L_0 \) radius \( r \).