Black-Box Certification with Randomized Smoothing: A Functional Optimization Based Framework

Dinghuai Zhang*, Mao Ye*, Chengyue Gong*, Zhanxing Zhu, Qiang Liu

Notation

Certification means a *guarantee* that a classifier won't change its prediction when perturbing input under some condition.

For simplicity, we consider a binary classification setting.

- $f^\sharp \colon \mathbb{R}^d \to [0,1]$ a given binary classifier
  output the probability of "positive class"
- $f^\sharp_{\pi_\mathbf{0}}(\boldsymbol{x}_0) := \mathbb{E}_{\boldsymbol{z} \sim \pi_\mathbf{0}} \left[f^\sharp(\boldsymbol{x}_0 + \boldsymbol{z}) \right]$
  randomized smoothed classifier
- $\Phi(\cdot)$ the cdf of standard Gaussian

Cohen [2]'s result

Theorem

For a randomized smoothing classfier with Gaussian noise $\boldsymbol{z} \sim \pi_{\mathbf{0}} = \mathcal{N}\left(0, \sigma^2 I\right)$, suppose $f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) \geq p_0 \geq \frac{1}{2}$, then $f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) \geq \frac{1}{2}$ for all $\|\boldsymbol{\delta}\|_2 \leq \sigma \Phi^{-1}(p_0)$.

Cohen et al. use some results from NP lemma to prove this bound.

Our Approach

The only thing we know about the given classfier is p_0. Naturally one would think: from all classfier f with $f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) \geq p_0$, which f^* achieves the lowest probability of $f^*_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta})$?

$$
\begin{aligned}
& \min_{f \in \mathcal{F}} \min_{\boldsymbol{\delta} \in \mathcal{B}} f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) \\
\text{s.t.} \quad & f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) \geq f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) := p_0
\end{aligned}
$$

Functional Optimization

From

$$\min_{f \in \mathcal{F}} \min_{\boldsymbol{\delta} \in \mathcal{B}} f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta})$$
$$\text{s.t.} \qquad f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) \geq f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) := p_0$$

we write out the Lagrangian:

$$V_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B}) = \min_{f \in \mathcal{F}} \min_{\boldsymbol{\delta} \in \mathcal{B}} \max_{\lambda \in \mathbb{R}} \left\{ f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) - \lambda \left(f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - p_0 \right) \right\}$$

Define $\pi_{\boldsymbol{\delta}}(z) = \mathcal{N}(\boldsymbol{\delta}, \sigma^2 I)$

$$
\begin{aligned}
& V_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B}) \\
= & \min_{f \in \mathcal{F}} \min_{\boldsymbol{\delta} \in \mathcal{B}} \max_{\lambda \in \mathbb{R}} \left\{ f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) - \lambda (f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - p_0) \right\} \\
= & \min_{\boldsymbol{\delta} \in \mathcal{B}} \min_{f \in \mathcal{F}} \max_{\lambda \geq 0} \left\{ \mathbb{E}_{\pi_{\boldsymbol{\delta}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})] + \lambda \left(p_0 - \mathbb{E}_{\pi_{\mathbf{0}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})] \right) \right\} \\
= & \min_{\boldsymbol{\delta} \in \mathcal{B}} \max_{\lambda \geq 0} \left\{ \lambda p_0 \; + \; \min_{f \in \mathcal{F}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})] - \lambda \mathbb{E}_{\pi_{\mathbf{0}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})]) \right\} \\
= & ?
\end{aligned}
$$

Specify our setting:

$$\mathcal{F}_{[0,1]} = \left\{ f : f(\boldsymbol{z}) \in [0,1], \forall \boldsymbol{z} \in \mathbb{R}^d \right\}$$
$$\mathcal{B} = \{ \boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq r \}$$

Thus our bound become

$$\begin{aligned} & \lambda p_0 \; - \; \min_{f \in \mathcal{F}_{[0,1]}} \left\{ \lambda \int f(\boldsymbol{x}_0 + \boldsymbol{z}) \pi_{\boldsymbol{0}}(\boldsymbol{z}) d\boldsymbol{z} - \int f(\boldsymbol{x}_0 + \boldsymbol{z}) \pi_{\boldsymbol{\delta}}(\boldsymbol{z}) d\boldsymbol{z} \right\} \\ = & \lambda p_0 \; - \; \min_{f \in \mathcal{F}_{[0,1]}} \left\{ \int f(\boldsymbol{x}_0 + \boldsymbol{z}) \left(\lambda \pi_{\boldsymbol{0}}(\boldsymbol{z}) - \pi_{\boldsymbol{\delta}}(\boldsymbol{z}) \right) d\boldsymbol{z} \right\} \\ = & \lambda p_0 - \int \left(\lambda \pi_{\boldsymbol{0}}(z) - \pi_{\boldsymbol{\delta}}(z) \right)_+ d\boldsymbol{z} \end{aligned}$$

Total Variation

Bonus

For two distribution q_1, q_2,

$$\int (q_1(\boldsymbol{z}) - q_2(\boldsymbol{z}))_+ \, d\boldsymbol{z} = TV(q_1 || q_2)$$

Thus $\int (\lambda \pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z))_+ \, d\boldsymbol{z}$ can also be seen as some sort of disrepancy.

Our Bound is Equivalent with Cohen's

Proposition

$$\max_{\lambda \geq 0} \min_{\|\boldsymbol{\delta}\|_2 \leq r} \left\{ \lambda p_0 - \int \left(\lambda \pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z) \right)_+ d\boldsymbol{z} \right\} = \Phi(\Phi^{-1}(p_0) - r/\sigma)$$

Thus
Confidence Lower Bound $\geq 0.5 \Leftrightarrow r \leq \sigma \Phi^{-1}(p_0)$!!

Remark

The min max can switch order.

Notation for ℓ_1 Setting

$$\mathcal{B} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1 \leq r\}$$

$$\pi_{\mathbf{0}}(z) \propto \exp\left(-\frac{\|z\|_1}{\sigma}\right)$$

$$\pi_{\boldsymbol{\delta}}(z) \propto \exp\left(-\frac{\|z - \boldsymbol{\delta}\|_1}{\sigma}\right)$$

Optimization with Laplacian Smoothing

Still, we have

$$\min_{\boldsymbol{\delta}\in\mathcal{B}} \max_{\lambda\geq 0} \left\{ \lambda p_0 - \int \left(\lambda\pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z)\right)_+ d\boldsymbol{z} \right\}$$

Proposition

the bound $= \frac{1}{2}\exp(-\log[2(1-p_0)] - \frac{r}{\sigma})$

Thus lower bound $\geq 0.5 \Leftrightarrow r \leq -\sigma\log[2(1-p_0)]$

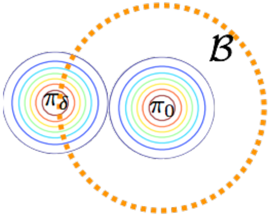
This is the same as [1].

Outline

Motivation

Bound Decomposition

$$\max_{\lambda \geq 0} \left[\underbrace{\lambda p_0}_{\text{Accuracy}} \underbrace{- \max_{\boldsymbol{\delta} \in \mathcal{B}} \int \left(\lambda \pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z) \right)_+ d\boldsymbol{z}}_{\text{Robustness}} \right]$$



New distribution can improve:

- More "center-massed" distribution can boost the accuracy term
- A heavy tail can boost the robustness term

Gaussian Issues

For high dimensional Gaussian distribution, the samples will concentrate on a "soap bubble":

$$\|z\|_2^2 = d\frac{\sum_i z_i^2}{d} \text{"} \rightarrow \text{"} d\sigma^2$$

Samples far from center will cause a small "accuracy" term!

Filling the Soap Bubbles

We propose a new Centripetal distribution family:

$$\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_2^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$$

The distribution of its raidus is

$$p_{\|\boldsymbol{z}\|_2}(r) \propto r^{d-k-1} \exp\left(-\frac{r^2}{2\sigma^2}\right)$$

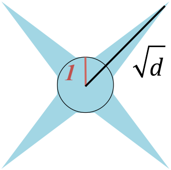
ℓ_2 RADIUS	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25
baseline (%)	60	43	34	23	17	14	12	10	8
OURS (%)	**61**	**46**	**37**	**25**	**19**	**16**	**14**	**11**	**9**

Table: Certified top-1 accuracy with various ℓ_2 radius on CIFAR-10.

ℓ_2 RADIUS	0.5	1.0	1.5	1.0	2.0	2.5	3.0
baseline (%)	49	37	29	19	15	12	9
OURS (%)	**50**	**39**	**31**	**21**	**17**	**13**	**10**

Table: Certified top-1 accuracy with various ℓ_2 radius on ImageNet.

Cracking ℓ_∞

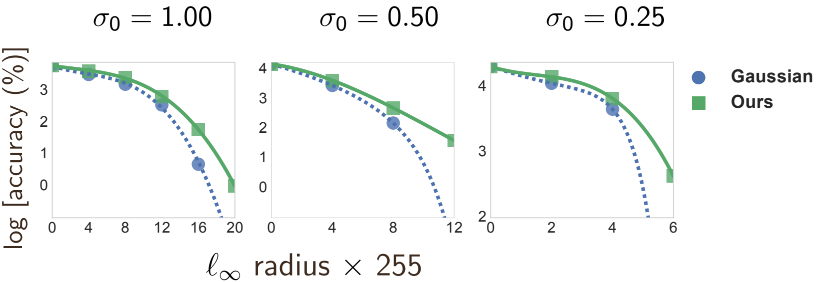


Infinite norm setting is very challenging, we propose the following centripetal distribution:

$$\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_\infty^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$$

l_∞ RADIUS	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**

Table: Certified top-1 accuracy with various l_∞ radius on CIFAR-10.



Thank you for listening!

References I

Anonymous, *ℓ_1 adversarial robustness certificates: a randomized smoothing approach*, in Submitted to International Conference on Learning Representations, 2020.
under review.

J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, *Certified adversarial robustness via randomized smoothing*, arXiv preprint arXiv:1902.02918, (2019).