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.

- $f^{]}: \mathsf{R}^d \;!\; [0,1]$ a given binary classifier
  output the probability of "positive class"
- $f^{]}_{\mathbf{0}}(x_0) := \mathsf{E}_{z \;\; \mathbf{0}}\left[f^{]}(x_0 + z)\right]$
  randomized smoothed classifier
- $\Phi(\;)$ the cdf of standard Gaussian

Cohen [2]'s result

Theorem

For a randomized smoothing class er with Gaussian noise $z \quad \pi_{\mathbf{0}} = N\left(0, \sigma^2 I\right)$, *suppose* $f^{]}{}_{\mathbf{0}}(x_0) \quad p_0 \quad \frac{1}{2}$, *then* $f^{]}{}_{\mathbf{0}}(x_0 + \quad) \quad \frac{1}{2}$ *for all* $k \;\; k_2 \quad \sigma\Phi \;\; {}^{1}(p_0)$.

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 $f_{\theta}(x_0) \quad p_0$, which f achieves the lowest probability of $f_{\theta}(x_0 + \quad)$?

$$\min_{f \in F} \min_{\in B} f_{\theta}(x_0 + \quad)$$
$$\text{s.t.} \quad f_{\theta}(x_0) \quad f^{]}_{\theta}(x_0) := p_0$$

Functional Optimization

From

$$\min_{f \in F} \min_{\in B} f_{\theta}(x_0 + \quad)$$
$$\text{s.t.} \qquad f_{\theta}(x_0) \quad f^{]}_{\theta}(x_0) := p_0$$

we write out the Lagrangian:

$$V_{\theta}(F, B) = \min_{f \in F} \min_{\in B} \max_{\in \mathrm{R}} \left\{ f_{\theta}(x_0 + \quad) \quad \lambda(f_{\theta}(x_0) \quad p_0) \right\}$$

Define $\pi\ (z) = N(\ \ , \sigma^2 I)$

$$
\begin{aligned}
& V\ _{\mathbf{0}}(F, B) \\
&= \min_{f\,2F}\ \min_{2B}\ \max_{2\mathrm{R}}\ f f\ _{\mathbf{0}}(x_0 + \ \ ) \quad \lambda(f\ _{\mathbf{0}}(x_0) \quad p_0) g \\
&= \min_{2B}\ \min_{f\,2F}\ \max_{0}\ f \mathrm{E}\ _{\delta}[f(x_0 + z)] + \lambda(p_0 \quad \mathrm{E}\ _{\mathbf{0}}[f(x_0 + z)]) g \\
&= \min_{2B}\ \max_{0} \left\{ \lambda p_0\ + \ \min_{f\,2F}\ \mathrm{E}\ _{\delta}[f(x_0 + z)] \quad \lambda \mathrm{E}\ _{\mathbf{0}}[f(x_0 + z)]) \right\} \\
&= ?
\end{aligned}
$$

Specify our setting:

$$F_{[0;1]} = \left\{ f \ : f(z) \ 2 \ [0,1], 8z \ 2 \ \mathrm{R}^d \right\}$$

$$B = f \quad : k \ k_2 \quad rg$$

Thus our bound become

$$\lambda p_0 \qquad \min_{f \, 2 F_{[0,1]}} \left\{ \lambda \int f(x_0 + z) \pi_{\mathbf{0}}(z) dz \quad \int f(x_0 + z) \pi \ (z) dz \right\}$$

$$= \lambda p_0 \qquad \min_{f \, 2 F_{[0,1]}} \left\{ \int f(x_0 + z) \left(\lambda \pi_{\mathbf{0}}(z) \quad \pi \ (z) \right) dz \right\}$$

$$= \lambda p_0 \qquad \int \left(\lambda \pi_{\mathbf{0}}(z) \quad \pi \ (z) \right)_+ dz$$

Total Variation

Bonus

For two distribution q_1, q_2,

$$\int (q_1(z) \quad q_2(z))_+ \, dz = TV(q_1 jj q_2)$$

Thus $\int (\lambda \pi_{\mathbf{0}}(z) \quad \pi \ (z))_+ \, dz$ can also be seen as some sort of disrepancy.

Our Bound is Equivalent with Cohen's

Proposition

$$\max_{0\ k\ k_2\ r} \min \left\{ \lambda p_0 \quad \int (\lambda \pi_{\boldsymbol{\theta}}(z) \quad \pi\ (z))_+ dz \right\} = \Phi(\Phi\ ^1(p_0) \quad r/\sigma)$$

Thus
Confidence Lower Bound 0.5 , r $\sigma\Phi\ ^1(p_0)$!!

Remark

The min max can switch order.

Notation for ℓ_1 Setting

$$B = f \quad : k \;\; k_1 \quad\quad rg$$

$$\pi_{\mathbf{0}}(z) \;\; / \;\; \exp\left(\quad \frac{kzk_1}{\sigma} \right)$$

$$\pi \;\; (z) \;\; / \;\; \exp\left(\quad \frac{kz \qquad k_1}{\sigma} \right)$$

Optimization with Laplacian Smoothing

Still, we have

$$\min_{2B} \max_{0} \left\{ \lambda p_0 \quad \int (\lambda \pi_{\boldsymbol{\theta}}(z) \quad \pi \ (z))_+ \, dz \right\}$$

Proposition

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

Thus lower bound $\quad 0.5 \ , \ r \qquad \sigma \log[2(1 \quad p_0)]$

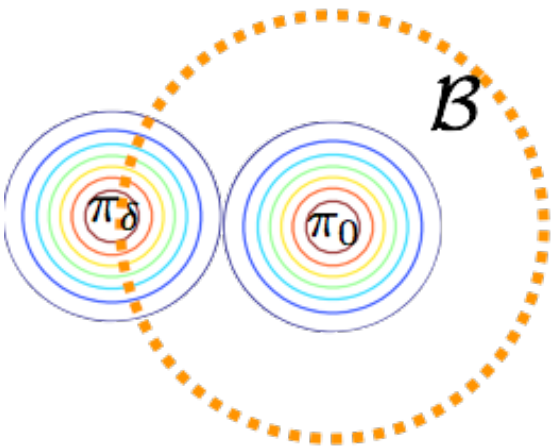
This is the same as [1].

Outline

Motivation

Bound Decomposition

$$\max_{0}\Bigg[\underbrace{\lambda p_0}_{\text{Accuracy}}\underbrace{\max_{2B}\int(\lambda\pi_{\mathbf{0}}(z)\quad\pi\ (z))_+\,dz}_{\text{Robustness}}\Bigg]$$



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":

$$\|z\|_2^2 = d\frac{\sum_i z_i^2}{d} \approx d\sigma^2$$

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

Filling the Soap Bubbles

We propose a new Centripetal distribution family:

$$\pi_{\boldsymbol{\theta}}(z) \;/\; kzk_2{}^{k} \exp\left(\frac{kzk_2^2}{2\sigma^2} \right)$$

The distribution of its raidus is

$$p_{kzk_2}(r) \;/\; r^{d \;\; k \;\; 1} \exp\left(\frac{r^2}{2\sigma^2} \right)$$

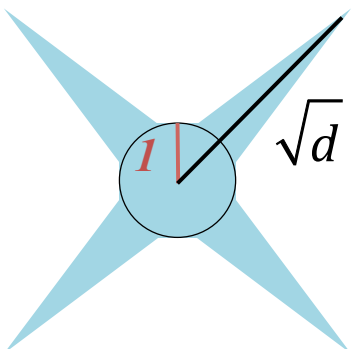
ℓ_2 RADIUS	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25
baseline (%)	60	43	34	23	17	14	12	10	8
OURS (%)	**61**	**46**	**37**	**25**	**19**	**16**	**14**	**11**	**9**

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_1

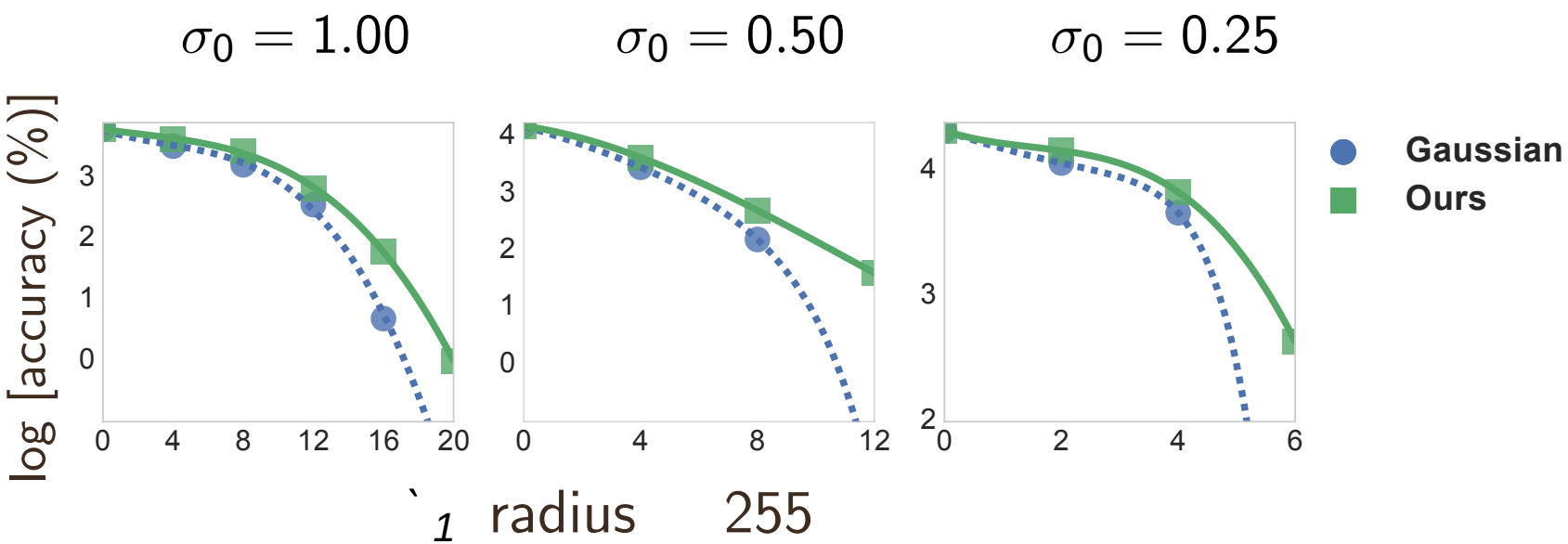


Infinite norm setting is very challenging, we propose the following centripetal distribution:

$$\pi_{\mathbf{0}}(z) \propto \|z\|_1^k \exp\left(-\frac{\|z\|_2^2}{2\sigma^2}\right)$$

l_1 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, *$nell_1$ adversarial robustness certi cates: a randomized smoothing approach*, in Submitted to International Conference on Learning Representations, 2020.
  under review.
- J. M. Cohen, E. Rosenfeld, and J. Z. Kolter, *Certi ed adversarial robustness via randomized smoothing*, arXiv preprint arXiv:1902.02918, (2019).