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

Dinghuai Zhang*, Mao Ye*, Chengyue Gong*,
Zhanxing Zhu, Qiang Liu
Peking University & University of Texas at Austin

Background on Randomized Smoothing

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. Below are three important notions we study:

- $f^\sharp : \mathbb{R}^d \to [0,1]$ a given binary classifier output the probability of "positive class"
- $f^\sharp_{\pi_0}(x_0) := \mathbb{E}_{z\sim\pi_0}\left[f^\sharp(x_0 + z)\right]$ randomized smoothed classifier
- $\Phi(\cdot)$ the cdf of standard Gaussian

For any testing data point $x_0 \in \mathbb{R}^d$ and the classifier predicts positively, i.e., $f^\sharp(x_0) > 1/2$, we then want to verify whether $f^\sharp(x_0+\delta) > 1/2$ still holds for any $\delta \in \mathcal{B}$. The mathematical formulation of certification in binary setting results in:

$$\min_{\delta\in\mathcal{B}} f^\sharp_{\pi_0}(x_0+\delta) = \min_{\delta\in\mathcal{B}} \mathbb{E}_{z\sim\pi_0}[f^\sharp(x_0+z+\delta)] > \frac{1}{2}$$

Compared to previous non-randomized certified defenses approaches including exact [2] or relaxed version [3] of certification, the randomized variants could significantly scale to larger settings [1]. We also discuss the pros and cons of our work compared to [6] in paper.

Constrained Adversarial Certification

We reformulate the original randomized smoothing certification problem as a functional optimization one.

$$\min_{\delta\in\mathcal{B}} f^\sharp_{\pi_0}(x_0+\delta) \geq \min_{f\in\mathcal{F}}\min_{\delta\in\mathcal{B}}\left\{f_{\pi_0}(x_0+\delta) \ \text{ s.t. } \ f_{\pi_0}(x_0) = f^\sharp_{\pi_0}(x_0)\right\}.$$

The Lagrangian function of this constrained optimization states

$$\mathcal{L}_{\pi_0}(\mathcal{F},\mathcal{B}) = \min_{f\in\mathcal{F}}\min_{\delta\in\mathcal{B}}\max_{\lambda\in\mathbb{R}} L(f,\delta,\lambda) \triangleq \min_{f\in\mathcal{F}}\min_{\delta\in\mathcal{B}}\max_{\lambda\in\mathbb{R}}\left\{f_{\pi_0}(x_0+\delta) - \lambda(f_{\pi_0}(x_0) - f^\sharp_{\pi_0}(x_0))\right\}$$

Then we can obtain our main theoretical argument:

Theorem 1. *I) (Dual Form) Denote by π_δ the distribution of $z+\delta$ when $z\sim\pi_0$. Assume $\mathcal{F}$ and $\mathcal{B}$ are compact set. We have the following lower bound of $\mathcal{L}_{\pi_0}(\mathcal{F},\mathcal{B})$:*

$$\mathcal{L}_{\pi_0}(\mathcal{F},\mathcal{B}) \geq \max_{\lambda\geq 0}\min_{f\in\mathcal{F}}\min_{\delta\in\mathcal{B}} L(f,\delta,\lambda) = \max_{\lambda\geq 0}\left\{\lambda f^\sharp_{\pi_0}(x_0) - \max_{\delta\in\mathcal{B}}\mathbb{D}_{\mathcal{F}}(\lambda\pi_0 \parallel \pi_\delta)\right\},$$

where we define the discrepancy term $\mathbb{D}_{\mathcal{F}}(\lambda\pi_0 \parallel \pi_\delta)$ as

$$\max_{f\in\mathcal{F}}\left\{\lambda\mathbb{E}_{z\sim\pi_0}[f(x_0+z)] - \mathbb{E}_{z\sim\pi_\delta}[f(x_0+z)]\right\},$$

which measures the difference of $\lambda\pi_0$ and π_δ by seeking the maximum discrepancy of the expectation for $f\in\mathcal{F}$. As we will show later, the bound in (1) is computationally tractable with proper $(\mathcal{F},\mathcal{B},\pi_0)$.

II) When $\mathcal{F} = \mathcal{F}_{[0,1]} := \{f\colon f(x)\in[0,1],\ x\in\mathbb{R}^d\}$, we have in particular

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

where $(t)_+ = \max(0,t)$. Furthermore, we have $0 \leq \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_0 \parallel \pi_\delta) \leq \lambda$ for any π_0, π_δ and $\lambda > 0$. Note that $\mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_0 \parallel \pi_\delta)$ coincides with the total variation distance between π_0 and π_δ when $\lambda = 1$.

III) (Strong duality) Suppose $\mathcal{F} = \mathcal{F}_{[0,1]}$ and suppose that for any $\lambda \geq 0$, $\min_{\delta\in\mathcal{B}}\min_{f\in\mathcal{F}_{[0,1]}} L(f,\delta,\lambda) = \min_{f\in\mathcal{F}_{[0,1]}} L(f,\delta^*,\lambda)$, for some $\delta^*\in\mathcal{B}$, we have

$$\mathcal{L}_{\pi_0}(\mathcal{F},\mathcal{B}) = \max_{\lambda\geq 0}\min_{\delta\in\mathcal{B}}\min_{f\in\mathcal{F}} L(f,\delta,\lambda).$$

Our theorem is applicable and flexible. When specified in ℓ_1 and ℓ_2 settings, we can exactly recover the bound derived by [4] and [1], different from their original Neyman-Pearson lemma approaches:

Corollary 1. *With Laplacian noise $\pi_0(\cdot) = \mathrm{Laplace}(\cdot; b)$, where $\mathrm{Laplace}(x; b) = \frac{1}{(2b)^d}\exp(-\frac{\|x\|_1}{b})$, ℓ_1 adversarial setting $\mathcal{B} = \{\delta\colon \|\delta\|_1 \leq r\}$ and $\mathcal{F} = \mathcal{F}_{[0,1]}$, the lower bound in Eq.1 becomes*

$$\max_{\lambda\geq 0}\left\{\lambda f^\sharp_{\pi_0}(x_0) - \max_{\|\delta\|_1\leq r}\mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_0\|\pi_\delta)\right\} = \begin{cases} 1 - e^{r/b}(1 - f^\sharp_{\pi_0}(x_0)), \text{when} f^\sharp_{\pi_0}(x_0) \geq 1 - \frac{1}{2}e^{-r/b} \\ \frac{1}{2}e^{-\frac{r}{b} - \log[2(1 - f^\sharp_{\pi_0}(x_0)]}, \text{when} f^\sharp_{\pi_0}(x_0) < 1 - \frac{1}{2}e^{-r/b} \end{cases}$$

Corollary 2. *With isotropic Gaussian noise $\pi_0 = \mathcal{N}(0, \sigma^2 I_{d\times d})$, ℓ_2 attack $\mathcal{B} = \{\delta\colon \|\delta\|_2 \leq r\}$ and $\mathcal{F} = \mathcal{F}_{[0,1]}$, the lower bound in Eq.1 becomes*

$$\max_{\lambda\geq 0}\left\{\lambda f^\sharp_{\pi_0}(x_0) - \max_{\|\delta\|_2\leq r}\mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_0\|\pi_\delta)\right\} = \Phi\left(\Phi^{-1}(f^\sharp_{\pi_0}(x_0)) - \frac{r}{\sigma}\right).$$

Improving Certification Bounds

We further demonstrate the effectiveness of our results by investigating more proper smoothing distribution for certification through its guide. An intuitive trade-off can be achieved from the confidence lower bound we obtained in Theorem 1:

$$\max_{\lambda\geq 0}\Bigg[\underbrace{\lambda f^\sharp_{\pi_0}(x_0)}_{\text{Accuracy}} + \underbrace{\left(-\max_{\delta\in\mathcal{B}}\mathbb{D}_{\mathcal{F}}(\lambda\pi_0 \parallel \pi_\delta)\right)}_{\text{Robustness}}\Bigg]$$

In our paper, we analyze this insightful decomposition and diagnosing what properties a good certification distribution should possess. We find that the smoothing distribution should avoid so-called "*then shell*" phenomenon [5] and hence more concentrated. Henceforth, we propose new distribution family to achieve the goal for :

$$\ell_1 : \pi_0(z) \propto \|z\|_1^{-k}\exp\left(-\frac{\|z\|_1}{b}\right) \qquad \ell_2 : \pi_0(z) \propto \|z\|_2^{-k}\exp\left(-\frac{\|z\|_2^2}{2\sigma^2}\right)$$

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

Experimental Results

Results for ℓ_1 and ℓ_2 certification

ℓ_1 Radius (CIFAR-10)	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25
Baseline (%)	62	49	38	30	23	19	17	14	12
Ours (%)	**64**	**51**	**41**	**34**	**27**	**22**	**18**	**17**	**14**

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**

Table 1: Certified top-1 accuracy of the best classifiers with various ℓ_1 radius.

ℓ_2 Radius (CIFAR-10)	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**

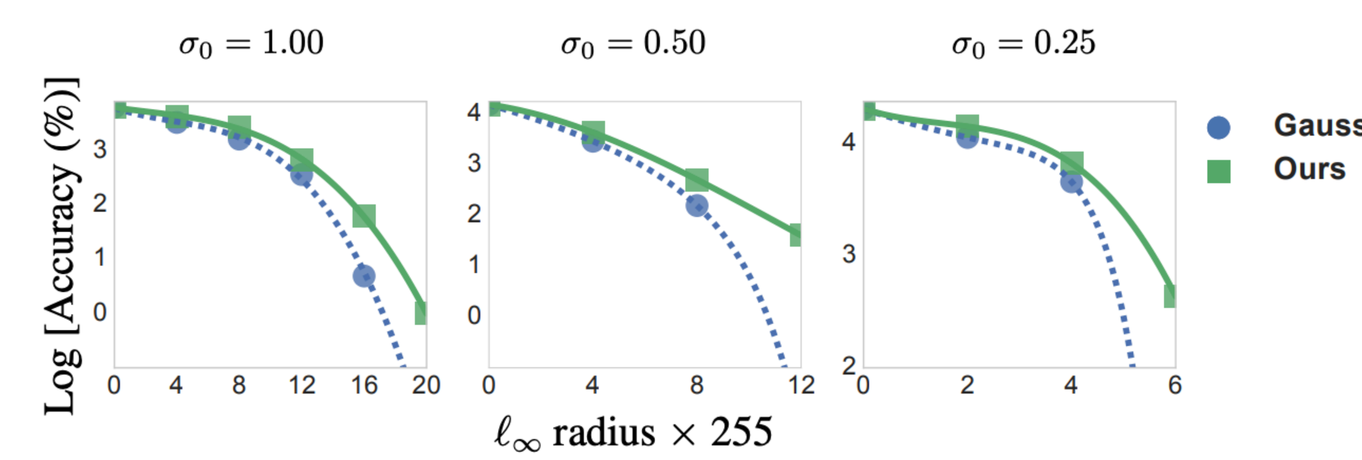
ℓ_2 Radius (ImageNet)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
Baseline (%)	49	37	29	19	15	12	9
Ours (%)	**50**	**39**	**31**	**21**	**17**	**13**	**10**

Table 2: Certified top-1 accuracy of the best classifiers with various ℓ_2 radius.

Results for ℓ_∞ certification

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 3: Certified top-1 accuracy of the best classifiers with various l_∞ radius on CIFAR-10.



Results of l_inf verification on CIFAR-10, on models trained with Gaussian noise data augmentation with different variances σ0. Our method obtains consistently better results

References

[1] Jeremy M Cohen, et al. Certified adversarial robustness via randomized smoothing. arXiv preprint arXiv:1902.02918, 2019.

[2] Chih-Hong Cheng, et al. Maximum resilience of artificial neural networks. In Automated Technology for Verification and Analysis 2017.

[3] Eric Wong and J Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. arXiv preprint arXiv:1711.00851, 2017.

[4] Jiaye Teng, et al. ℓ_1 adversarial robustness certificates: a randomized smoothing approach, 2020.

[5] Roman Vershynin. High-dimensional probability: An introduction with applications in data science, volume 47. Cambridge University Press, 2018

[6] Greg Yang et al. Randomized smoothing of all shapes and sizes. arXiv preprint arXiv:2002.08118, 2020.