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

Dinghuai Zhang*
Mila
dinghuai.zhang@mila.quebec

Mao Ye*, **Chengyue Gong***
Department of Computer Science
University of Texas at Austin
{my21, cygong}@cs.utexas.edu

Zhanxing Zhu
School of Mathematical Sciences
Peking University
zhanxing.zhu@pku.edu.cn

Qiang Liu
Department of Computer Science
University of Texas at Austin
lqiang@cs.utexas.edu

Abstract

Randomized classifiers have been shown to provide a promising approach for achieving certified robustness against adversarial attacks in deep learning. However, most existing methods only leverage Gaussian smoothing noise and only work for ℓ_2 perturbation. We propose a general framework of adversarial certification with non-Gaussian noise and for more general types of attacks, from a unified functional optimization perspective. Our new framework allows us to identify a key trade-off between accuracy and robustness via designing smoothing distributions and leverage it to design new families of non-Gaussian smoothing distributions that work more efficiently for different ℓ_p settings, including ℓ_1, ℓ_2 and ℓ_∞ attacks. Our proposed methods achieve better certification results than previous works and provide a new perspective on randomized smoothing certification.

1 Introduction

Although many robust training algorithms have been developed to overcome adversarial attacks [1–3], most heuristically developed methods can be shown to be broken by more powerful adversaries eventually (*e.g.,* [4–7]). This casts an urgent demand for developing robust classifiers with provable worst-case guarantees. One promising approach for certifiable robustness is the recent *randomized smoothing method* [8–15], which constructs smoothed classifiers with certifiable robustness by introducing noise on the inputs. Compared with the other more traditional certification approaches [16–18] that exploit special structures of the neural networks (such as the properties of ReLU), the randomized smoothing approaches work more flexibly on general black-box classifiers and is shown to be more scalable and provide tighter bounds on challenging datasets such as ImageNet [19].

Most existing methods use Gaussian noise for smoothing. Although appearing to be a natural choice, one of our key observations is that the Gaussian distribution is, in fact a sub-optimal choice in high dimensional spaces even for ℓ_2 attack. We observe that there is a counter-intuitive phenomenon in high dimensional spaces [20], that almost all of the probability mass of standard Gaussian distribution concentrates around the sphere surface of a certain radius. This makes tuning the variance of Gaussian distribution an inefficient way to trade off robustness and accuracy for randomized smoothing.

Our Contributions We propose a general framework of adversarial certification using non-Gaussian smoothing noises, based on a new functional optimization perspective. Our framework unifies the

*Equal contributions

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

methods of[9] and[14] as special cases, and is applicable to more general smoothing distributions and more types of attacks beyond ℓ_2-norm setting. Leveraging our insight, we develop a new family of distributions for better certification results on ℓ_1, ℓ_2 and ℓ_∞ attacks. An efficient computational approach is developed to enable our method in practice. Empirical results show that our new framework and smoothing distributions outperform existing approaches for ℓ_1, ℓ_2 and ℓ_∞ attacking, on datasets such as CIFAR-10 and ImageNet.

2 Related Works

Certified Defenses Unlike the empirical defense methods, once a classifier can guarantee a consistent prediction for input within a local region, it is called a certified-robustness classifier. *Exact* certification methods provide the minimal perturbation condition which leads to a different classification result. This line of work focuses on deep neural networks with ReLU-like activation that makes the classifier a piece-wise linear function. This enables researchers to introduce satisfiability modulo theories[21,22] or mix integer linear programming[23,24]. *Sufficient* certification methods take a conservative way and bound the Lipschitz constant or other information of the network[16,18,25,26]. However, these certification strategies share a drawback that they are not feasible on large-scale scenarios, *e.g.* large and deep networks and datasets.

Randomized Smoothing To mitigate this limitation of previous certifiable defenses, improving network robustness via randomness has been recently discussed[27,28]. [8] first introduced randomization with technique in differential privacy. [12] improved their work with a bound given by Rényi divergence. In succession, [9] firstly provided a *tight* bound for *arbitrary* Gaussian smoothed classifiers based on previous theorems found by[29]. [10] combined the empirical and certification robustness, by applying adversarial training on randomized smoothed classifiers to achieve a higher certified accuracy. [11] focused on ℓ_0 norm perturbation setting, and proposed a discrete smoothing distribution which can be shown perform better than the widely used Gaussian distribution. [14] took a similar statistical testing approach with[9], utilizing Laplacian smoothing to tackle ℓ_1 certification problem. [15] extended the approach of[9] to a top-k setting. [13] extends the total variant used by [9] to f-divergences. Recent works[30–32] discuss further problems about certification methods. We also focus on a generalization of randomized smoothing, but with a different view on loosing the constraint on classifier.

Noticeably, [30] also develops analysis on ℓ_1 setting and provide a thorough theoretical analysis on many kinds of randomized distribution. We believe the[30] and ours have different contributions and were developed concurrently. [30] derives the optimal shapes of level sets for ℓ_p attacks based on the Wulff Crystal theory, while our work, based on our functional-optimization framework and accuracy-robustness decomposition (Eq.9), proposes to use distribution that is more concentrated toward the center. Besides, we also consider a novel distribution using mixed ℓ_2 and ℓ_∞ norm for ℓ_∞ adversary, which hasn't been studied before and improve the empirical results.

3 Black-box Certification as Functional Optimization

3.1 Background

Adversarial Certification For simplicity, we consider binary classification of predicting binary labels $y \in \{0, 1\}$ given feature vectors $x \in \mathbb{R}^d$. The extension to multi-class cases is straightforward, and is discussed in Appendix C. We assume $f^\sharp : \mathbb{R}^d \to [0, 1]$ is a given binary classifier ($\sharp$ means the classifier is *given*), which maps from the input space $\mathbb{R}^d$ to either the positive class probability in interval $[0, 1]$ or binary labels in $\{0, 1\}$. In the robustness certification problem, a testing data point $\boldsymbol{x}_0 \in \mathbb{R}^d$ is given, and one is asked to verify if the classifier outputs the same prediction when the input $\boldsymbol{x}_0$ is perturbed arbitrarily in $\mathcal{B}$, a given neighborhood of $\boldsymbol{x}_0$. Specifically, let $\mathcal{B}$ be a set of possible perturbation vectors, *e.g.*, $\mathcal{B} = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_p \le r\}$ for ℓ_p norm with a radius r. If the classifier predicts $y = 1$ on $\boldsymbol{x}_0$, i.e. $f^\sharp(\boldsymbol{x}_0) > 1/2$, we want to verify if $f^\sharp(\boldsymbol{x}_0 + \boldsymbol{\delta}) > 1/2$ still holds for any $\delta \in \mathcal{B}$. Through this paper, we consider the most common adversarial settings: ℓ_1, ℓ_2 and ℓ_∞ attacks.

Black-box Randomized Smoothing Certification Directly certifying $f^\sharp$ heavily relies on the smooth property of $f^\sharp$, which has been explored in a series of prior works[8,16]. These methods typically depend on the special structure-property (*e.g.*, the use of ReLU units) of $f^\sharp$, and thus can not serve as general-purpose algorithms for any type of networks. Instead, We are interested in *black-box* verification methods that could work for *arbitrary* classifiers. One approach to enable this, as explored in recent works[9,11], is to replace $f^\sharp$ with a smoothed classifier by convolving it with Gaussian noise, and verify the *smoothed* classifier.

Specifically, assume $\pi_{\mathbf{0}}$ is a smoothing distribution with zero mean and bounded variance, e.g., $\pi_{\mathbf{0}} = \mathcal{N}(\mathbf{0}, \sigma^2)$. The randomized smoothed classifier is defined by

$$f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) := \mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}\left[f^\sharp(\boldsymbol{x}_0 + \boldsymbol{z})\right],$$

which returns the averaged probability of $\boldsymbol{x}_0 + \boldsymbol{z}$ under the perturbation of $\boldsymbol{z} \sim \pi_{\mathbf{0}}$. Assume we replace the original classifier with $f^\sharp_{\pi_{\mathbf{0}}}$, then the goal becomes certifying $f^\sharp_{\pi_{\mathbf{0}}}$ using its inherent smoothness. Specifically, if $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) > 1/2$, we want to certify that $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) > 1/2$ for every $\boldsymbol{\delta} \in \mathcal{B}$, that is, we want to certify that

$$\min_{\boldsymbol{\delta}\in\mathcal{B}} f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) = \min_{\boldsymbol{\delta}\in\mathcal{B}} \mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}[f^\sharp(\boldsymbol{x}_0 + \boldsymbol{z} + \boldsymbol{\delta})] > \frac{1}{2}. \tag{1}$$

In this case, it is sufficient to obtain a *guaranteed lower bound* of $\min_{\boldsymbol{\delta}\in\mathcal{B}} f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta})$ and check if it is larger than $1/2$. When $\pi_{\mathbf{0}}$ is Gaussian $\mathcal{N}(\mathbf{0},\ \sigma^2)$ and for ℓ_2 attack, this problem was studied in[9], which shows that a lower bound of

$$\min_{\boldsymbol{z}\in\mathcal{B}} \mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}[f^\sharp(\boldsymbol{x}_0 + \boldsymbol{z})] \geq \Phi(\Phi^{-1}(f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)) - \frac{r}{\sigma}), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative density function (CDF) of standard Gaussian distribution. The proof of this result in[9] uses Neyman-Pearson lemma[29]. In the following section, we will show that this bound is a special case of the proposed functional optimization framework for robustness certification.

3.2 Constrained Adversarial Certification

We propose a **constrained adversarial certification (CAC)** framework, which yields a guaranteed lower bound for Eq.1. The main idea is simple: assume $\mathcal{F}$ is a function class which is known to include $f^\sharp$, then the following optimization immediately yields a guaranteed lower bound

$$\min_{\boldsymbol{\delta}\in\mathcal{B}} f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) \geq \min_{f\in\mathcal{F}} \min_{\boldsymbol{\delta}\in\mathcal{B}} \left\{ f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) \ \text{ s.t. } \ f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) = f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) \right\}, \tag{3}$$

where we define $f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) = \mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})]$ for any given f. Then we need to search for the minimum value of $f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta})$ for all classifiers in $\mathcal{F}$ that satisfies $f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) = f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)$. This obviously yields a lower bound once $f^\sharp \in \mathcal{F}$. If $\mathcal{F}$ includes only $f^\sharp$, then the bound is exact, but is computationally prohibitive due to the difficulty of optimizing $\boldsymbol{\delta}$. The idea is then to choose $\mathcal{F}$ properly to incorporate rich information of $f^\sharp$, while allowing us to calculate the lower bound in Eq.3 computationally tractably. In this paper, we consider the set of all functions bounded in $[0, 1]$, namely

$$\mathcal{F}_{[0,1]} = \left\{ f : f(\boldsymbol{z}) \in [0, 1], \forall \boldsymbol{z} \in \mathbb{R}^d \right\}, \tag{4}$$

which guarantees to include all $f^\sharp$ by definition.

Denote by $\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B})$ the lower bound in Eq.3. We can rewrite it into the following minimax form using the Lagrangian function,

$$\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B}) = \min_{f\in\mathcal{F}} \min_{\boldsymbol{\delta}\in\mathcal{B}} \max_{\lambda\in\mathbb{R}} L(f, \boldsymbol{\delta}, \lambda) \triangleq \min_{f\in\mathcal{F}} \min_{\boldsymbol{\delta}\in\mathcal{B}} \max_{\lambda\in\mathbb{R}} \left\{ f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) - \lambda(f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)) \right\}, \tag{5}$$

where λ is the Lagrangian multiplier. Exchanging the $\min$ and $\max$ yields the following dual form.

Theorem 1. *I) **(Dual Form)** Denote by $\pi_{\boldsymbol{\delta}}$ the distribution of $\boldsymbol{z} + \boldsymbol{\delta}$ when $\boldsymbol{z} \sim \pi_{\mathbf{0}}$. Assume $\mathcal{F}$ and $\mathcal{B}$ are compact set. We have the following lower bound of $\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B})$:*

$$\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B}) \geq \max_{\lambda\geq 0} \min_{f\in\mathcal{F}} \min_{\boldsymbol{\delta}\in\mathcal{B}} L(f, \boldsymbol{\delta}, \lambda) = \max_{\lambda\geq 0} \left\{ \lambda f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \max_{\boldsymbol{\delta}\in\mathcal{B}} \mathbb{D}_{\mathcal{F}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}\right) \right\}, \tag{6}$$

where we define the discrepancy term $\mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$ as

$$\max_{f\in\mathcal{F}} \Big\{\lambda\mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}[f(\boldsymbol{x}_0+\boldsymbol{z})] - \mathbb{E}_{\boldsymbol{z}\sim\pi_{\boldsymbol{\delta}}}[f(\boldsymbol{x}_0+\boldsymbol{z})]\Big\},$$

which measures the difference of $\lambda\pi_{\mathbf{0}}$ and $\pi_{\boldsymbol{\delta}}$ by seeking the maximum discrepancy of the expectation for $f \in \mathcal{F}$. As we will show later, the bound in (6) is computationally tractable with proper $(\mathcal{F}, \mathcal{B}, \pi_{\mathbf{0}})$.

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

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

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

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

$$\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B}) = \max_{\lambda\ge 0} \min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}} L(f, \boldsymbol{\delta}, \lambda).$$

Remark We will show later that the proposed methods and the cases we study satisfy the condition in part III of the theorem and thus all the lower bounds of the proposed method are tight.

Proof is deferred to Appendix A.1. Although the lower bound in Eq.6 still involves an optimization on $\boldsymbol{\delta}$ and λ, both of them are much easier than the original adversarial optimization in Eq.1. With proper choices of $\mathcal{F}$, $\mathcal{B}$ and $\pi_{\mathbf{0}}$, the optimization of $\boldsymbol{\delta}$ can be shown to provide simple closed-form solutions by exploiting the symmetry of $\mathcal{B}$, and the optimization of λ is a very simple one-dimensional searching problem.

As corollaries of Theorem 1, we can exactly recover the bound derived by[14] and[9] under our functional optimization framework, different from their original Neyman-Pearson lemma approaches.

Corollary 1. *With Laplacian noise $\pi_{\mathbf{0}}(\cdot) = \mathrm{Laplace}(\cdot; b)$, where $\mathrm{Laplace}(\boldsymbol{x}; b) = \frac{1}{(2b)^d}\exp(-\frac{\|\boldsymbol{x}\|_1}{b})$, ℓ_1 adversarial setting $\mathcal{B} = \{\boldsymbol{\delta}\colon \|\boldsymbol{\delta}\|_1 \le r\}$ and $\mathcal{F} = \mathcal{F}_{[0,1]}$, the lower bound in Eq.6 becomes*

$$\max_{\lambda\ge 0}\Big\{\lambda f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \max_{\|\boldsymbol{\delta}\|_1\le r} \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}}\|\pi_{\boldsymbol{\delta}})\Big\} = \begin{cases} 1 - e^{r/b}(1 - f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)), \textit{when} f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) \ge 1 - \frac{1}{2}e^{-r/b}, \\ \frac{1}{2}e^{-\frac{r}{b} - \log[2(1 - f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0))]}, \textit{when} f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) < 1 - \frac{1}{2}e^{-r/b}. \end{cases} \tag{7}$$

Thus, with our previous explanation, we obtain $\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B}) \ge \frac{1}{2} \iff r \le -b\log\big[2(1 - f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0))\big]$, which is exactly the ℓ_1 certification radius derived by[14]. See Appendix A.2 for proof details. For Gaussian noise setting which has been frequently adopted, we have

Corollary 2. *With isotropic Gaussian noise $\pi_{\mathbf{0}} = \mathcal{N}(\mathbf{0}, \sigma^2 I_{d\times d})$, ℓ_2 attack $\mathcal{B} = \{\boldsymbol{\delta}\colon \ \|\boldsymbol{\delta}\|_2 \le r\}$ and $\mathcal{F} = \mathcal{F}_{[0,1]}$, the lower bound in Eq.6 becomes*

$$\max_{\lambda\ge 0}\Big\{\lambda f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \max_{\|\boldsymbol{\delta}\|_2\le r} \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}}\|\pi_{\boldsymbol{\delta}})\Big\} = \Phi\left(\Phi^{-1}(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)) - \frac{r}{\sigma}\right). \tag{8}$$

Analogously, we can retrieve the main theoretical result of[9] :$\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B}) \ge \frac{1}{2} \iff r \le \sigma\Phi^{-1}(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0))$. See Appendix A.3 for proof details.

3.3 Trade-off Between Accuracy and Robustness

The lower bound in Eq.6 reflects an intuitive trade-off between the robustness and accuracy on the certification problem:

$$\max_{\lambda\ge 0}\Bigg[\lambda \underbrace{f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)}_{\text{Accuracy}} + \underbrace{\left(-\max_{\boldsymbol{\delta}\in\mathcal{B}} \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})\right)}_{\text{Robustness}}\Bigg], \tag{9}$$

where the first term reflects the accuracy of the smoothed classifier (assuming the true label is $y = 1$), while the second term $-\max_{\boldsymbol{\delta}\in\mathcal{B}} \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$ measures the robustness of the smoothing method, via the negative maximum discrepancy between the original smoothing distribution $\pi_{\mathbf{0}}$ and perturbed distribution $\pi_{\boldsymbol{\delta}}$ for $\boldsymbol{\delta} \in \mathcal{B}$. The maximization of dual coefficient λ can be viewed as searching for a best balance between these two terms to achieve the largest lower bound.

More critically, different choices of smoothing distributions yields a trade-off between accuracy and robustness in Eq.9. A good choice of the smoothing distribution should ① be centripetal enough to obtain a large $f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)$ and ② have large kurtosis or long tail to yield a small $\max_{\boldsymbol{\delta}\in\mathcal{B}} \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$ discrepancy term. In the next section, we'll show how to design a distribution that could improve both points.

4 Improving Certification Bounds with a New Distribution Family

4.1 *"Thin Shell"* Phenomenon and New Distribution Family

We first identify a key problem of the usage of Laplacian and Gaussian noise in high dimensional space, due to the "*thin shell*" phenomenon that the probability mass of them concentrates on a sphere far away from the center points [20].

Proposition 1 ([20], Section 3.1)**.** *Let $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0},\ I_{d\times d})$ be a d-dimensional standard Gaussian random variable. Then there exists a constant c, such that for any $\delta \in (0,1)$,* $\mathrm{Prob}\left(\sqrt{d} - \sqrt{c\log(2/\delta)} \leq \|\boldsymbol{z}\|_2 \leq \sqrt{d} + \sqrt{c\log(2/\delta)}\right) \geq 1-\delta$. *See [20] for more discussion.*

This suggests that with high probability, $\boldsymbol{z}$ takes values very close to the sphere of radius $\sqrt{d}$, within a constant distance from that sphere. There exists similar phenomenon for Laplacian distribution:

Proposition 2 (Chebyshev bound)**.** *Let $\boldsymbol{z}$ be a d-dimensional Laplacian random variable, $\boldsymbol{z} = (z_1, \cdots, z_d)$, where $z_i \sim \mathrm{Laplace}(1), i = 1, \cdots, d$. Then for any $\delta \in (0,1)$, we have* $\mathrm{Prob}\left(1 - 1/\sqrt{d\delta} \leq \|\boldsymbol{z}\|_1/d \leq 1 + 1/\sqrt{d\delta}\right) \geq 1-\delta$.

Although choosing isotropic Laplacian and Gaussian distribution appears to be natural, this "*thin shell*" phenomenon makes it sub-optimal to use them for adversarial certification, because one would expect that the smoothing distribution should concentrate around the center (the original image) in order to make the smoothed classifier accurate enough in trade-off of Eq.9.

Thus it's desirable to design a distribution more *concentrated* to center. To motivate our new distribution family, it's useful to examine the density function of the distributions of the radius of spherical distributions in general.

Proposition 3. *Assume $\boldsymbol{z}$ is a symmetric random variable on $\mathbb{R}^d$ with a probability density function (PDF) of form $\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \phi(\|\boldsymbol{z}\|)$, where $\phi\colon [0,\infty) \to [0,\infty)$ is a univariate function, then the PDF of the norm of $\boldsymbol{z}$ is $p_{\|\boldsymbol{z}\|}(r) \propto r^{d-1}\phi(r)$. The term r^{d-1} arises due to the integration on the surface of radius r norm ball in $\mathbb{R}^d$. Here $\|\cdot\|$ can be any L_p norm.*

In particular, for $\boldsymbol{z} \sim \mathcal{N}(0, \sigma^2 I_{d\times d})$, we have $\phi(r) \propto \exp(-r^2/(2\sigma^2))$ and hence $p_{\|\boldsymbol{z}\|_2}(r) \propto r^{d-1}\exp(-r^2/(2\sigma^2))$. We can see that the "*thin shell*" phenomenon is caused by the r^{d-1} term, which makes the density to be highly peaked when d is large. To alleviate the concentration phenomenon, we need to cancel out the effect of r^{d-1}, which motivates the following family of smoothing distributions:

$$\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_{n_1}^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_{n_2}^{p}}{b}\right),$$

where parameters $k, n_1, n_2, p \in \mathbb{N}$. Next we discuss how to choose suitable parameters depending on specific perturbation region.

4.2 ℓ_1 and ℓ_2 Region Certification

Based on original Laplacian and Gaussian distributions and above intuition, we propose:

$$\ell_1 : \pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_1^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_1}{b}\right) \tag{10}$$

$$\ell_2 : \pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_2^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right) \tag{11}$$

where we introduce the $\|\boldsymbol{z}\|^{-k}$ term in $\pi_{\mathbf{0}}$, with k a positive parameter, to make the radius distribution more concentrated when k is large.

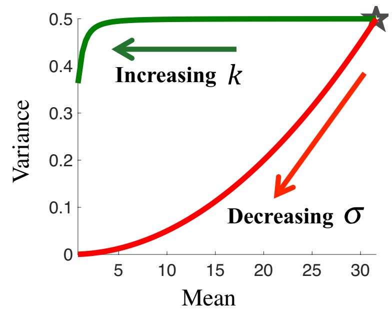

Figure 1: Starting from radius distribution in Eq.11 with $d = 100$ $\sigma = 1$ and $k = 0$ (black start), increasing k (green curve) moves the mean towards zero *without significantly reducing the variance*. Decreasing σ (red curve) can also decrease the mean, but with a cost of decreasing the variance quadratically.

The radius distribution in Eq.10 and Eq.11 is controlled by two parameters: σ (or b) and k, who control the scale and shape of the distribution, respectively. The key idea is that adjusting extra parameter k allows us to control the trade-off the accuracy and robustness more precisely. As shown in Fig.1, adjusting σ moves the mean close to zero (hence ① yielding higher accuracy), but at cost of decreasing the variance quadratically (hence ② less robust). In contrast, adjusting k decreases the mean without significantly impacting the variance, thus yield a much better trade-off on accuracy and robustness.

Computational Method Now we no longer have the closed-form solution of the bound like Eq.7 and Eq.8. However, efficient computational methods can still be developed for calculating the bound in Eq.6 with $\pi_{\mathbf{0}}$ in Eq.11 or Eq.11. The key is that the maximum of the distance term $\mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \mid\mid \pi_{\boldsymbol{\delta}})$ over $\boldsymbol{\delta} \in \mathcal{B}$ is always achieved on the boundary of $\mathcal{B}$:

Theorem 2. *Consider the ℓ_1 attack with $\mathcal{B} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_1 \le r\}$ and smoothing distribution $\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_1^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_1}{b}\right)$ with $k \ge 0$ and $b > 0$, or the ℓ_2 attack with $\mathcal{B} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \le r\}$ and smoothing distribution $\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_2^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$ with $k \ge 0$ and $\sigma > 0$. Define $\boldsymbol{\delta}^* = [r, 0, ..., 0]^\top$, we have*

$$\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}^*}\right) = \max_{\delta \in \mathcal{B}} \mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}\right)$$

for any positive λ.

With Theorem 2, we can compute Eq.6 with $\boldsymbol{\delta} = \boldsymbol{\delta}^*$. We then calculate $\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}^*}\right) = \mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}\left[\left(\lambda - \frac{\pi_{\boldsymbol{\delta}^*}(\boldsymbol{z})}{\pi_{\mathbf{0}}(\boldsymbol{z})}\right)_+\right]$ using Monte Carlo approximation with i.i.d. samples $\{\boldsymbol{z}_i\}_{i=1}^n$ be i.i.d. samples from $\pi_{\mathbf{0}}$: $\hat{D} := \frac{1}{n}\sum_{i=1}^n \left(\lambda - \pi_{\boldsymbol{\delta}^*}(\boldsymbol{z}_i)/\pi_{\mathbf{0}}(\boldsymbol{z}_i)\right)_+$, which is bounded in the following confidence interval $[\hat{D} - \lambda\sqrt{\log(2/\delta)/(2n)}, \hat{D} + \lambda\sqrt{\log(2/\delta)/(2n)}]$ with confidence level $1 - \delta$ for $\delta \in (0, 1)$. What's more, the optimization on $\lambda \ge 0$ is one-dimensional and can be solved numerically efficiently (see Appendix for details).

4.3 ℓ_∞ Region Certification

Going further, we consider the more difficult ℓ_∞ attack whose attacking region is $\mathcal{B}_{\ell_\infty,r} = \{\boldsymbol{\delta}\colon \|\boldsymbol{\delta}\|_\infty \le r\}$. The commonly used Gaussian smoothing distribution, as well as our ℓ_2-based smoothing distribution in Eq.11, is unsuitable for this region:

Proposition 4. *With the smoothing distribution $\pi_{\mathbf{0}}$ in Eq.11 for $k \ge 0, \sigma > 0$, and $\mathcal{F} = \mathcal{F}_{[0,1]}$ shown in Eq.4, the bound we get for certifying the ℓ_∞ attack on $\mathcal{B}_{\ell_\infty,r} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_\infty \le r\}$ is equivalent to that for certifying the ℓ_2 attack on $\mathcal{B}_{\ell_2,\sqrt{d}r} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \le \sqrt{d}r\}$, that is,*

$$\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}_{[0,1]},\ \mathcal{B}_{\ell_\infty,r}) = \mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}_{[0,1]},\ \mathcal{B}_{\ell_2,\sqrt{d}r}).$$

As shown in this proposition, if we use ℓ_2 distribution in Eq.11 for certification, the bound we obtain is effectively the bound we get for verifying a ℓ_2 ball with radius $\sqrt{d}r$, which is too large to give meaningful results due to high dimension.

In order to address this problem, we extend our proposed distribution family with new distributions which are more suitable for ℓ_∞ certification setting:

$$\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_\infty^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_\infty^2}{2\sigma^2}\right), \quad (12)$$

$$\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_\infty^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right). \quad (13)$$

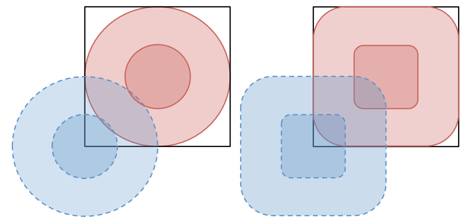

Figure 2: For ℓ_∞ attacking, compared with the distribution in Eq.11, the mixed norm distribution in Eq.13 (right) yields smaller discrepancy term (because of larger overlap areas), and hence higher robustness and better confidence bound. The distribution described in Eq.12 has the same impact.

The motivation is to allocate more probability mass along the "pointy" directions with larger ℓ_∞ norm, and hence decrease the maximum discrepancy term $\max_{\boldsymbol{\delta} \in \mathcal{B}_{\ell_\infty, r}} \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$, see Fig.2.

Computational Method In order to compute the lower bound with proposed distribution, we need to establish similar theoretical results as Theorem 2, showing the optimal $\boldsymbol{\delta}$ is achieved at one vertex (the pointy points) of ℓ_∞ ball.

Theorem 3. *Consider the ℓ_∞ attack with $\mathcal{B}_{\ell_\infty, r} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_\infty \leq r\}$ and the mixed norm smoothing distribution in Eq.13 with $k \geq 0$ and $\sigma > 0$. Define $\boldsymbol{\delta}^* = [r, r, ..., r]^\top$. We have for any $\lambda > 0$,*

$$\mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}^*}) = \max_{\boldsymbol{\delta} \in \mathcal{B}} \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}).$$

The proofs of Theorem 2 and 3 are non-trivial and deferred to Appendix. With the optimal $\boldsymbol{\delta}^*$ found above, we can calculate the bound with similar Monte Carlo approximation outlined in Section 4.2.

5 Experiments

We evaluate proposed certification bound and smoothing distributions for ℓ_1, ℓ_2 and ℓ_∞ attacks. We compare with the randomized smoothing method of[14] with Laplacian smoothing for ℓ_1 region cerification. For ℓ_2 and ℓ_∞ cases, we regard the method derived by[9] with Gaussian smoothing distribution as the baseline. For fair comparisons, we use the same model architecture and pre-trained models provided by[14],[9] and[10], which are ResNet-110 for CIFAR-10 and ResNet-50 for ImageNet. We use the official code[2] provided by [9] for all the following experiments. For all other details and parameter settings, we refer the readers to Appendix B.2.

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.

[2]`https://github.com/locuslab/smoothing`. Our results are slightly different with those in original paper due to the randomness of sampling.

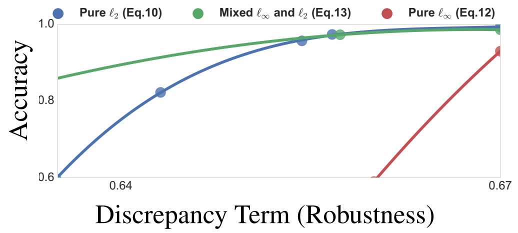

Figure 3: The Pareto frontier of accuracy and robustness (in the sense of Eq.9) of the three smoothing families in Eq.11, Eq.13, and Eq.12 for ℓ_∞ attacking, when we search for the best parameters (k, σ) for each of them. The mixed norm family Eq.13 yields the best trade-off than the other two. We assume $f^\sharp(\boldsymbol{x}) = \mathbb{I}(\|\boldsymbol{x}\|_2 \leq r)$ and dimension $d = 5$. The case when $f^\sharp(\boldsymbol{x}) = \mathbb{I}(\|\boldsymbol{x}\|_\infty \leq r)$ has similar result (not shown).

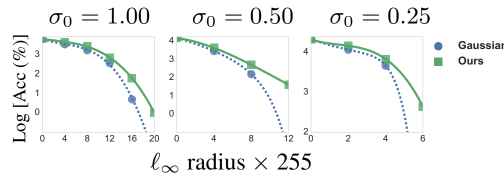

Figure 4: Results of ℓ_∞ verification on CIFAR-10, on models trained with Gaussian noise data augmentation with different variances σ_0. Our method obtains consistently better results.

Evaluation Metrics Methods are evaluated with the certified accuracy defined in [9]. Given an input image $\boldsymbol{x}$ and a perturbation region $\mathcal{B}$, the smoothed classifier certifies image $\boldsymbol{x}$ correctly if the prediction is correct and has a guaranteed confidence lower bound larger than $1/2$ for any $\boldsymbol{\delta} \in \mathcal{B}$. The certified accuracy is the percentage of images that are certified correctly. Following [10], we calculate the certified accuracy of all the classifiers in [9] or [10] for various radius, and report the best results over all of classifiers.

5.1 ℓ_1 & ℓ_2 Certification

For ℓ_1 certification, we compare our method with [14] on CIFAR-10 and ImageNet with the type 1 trained model in [14]. As shown in Table 1, our non-Laplacian centripetal distribution consistently outperforms the result of baseline for any 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.

For ℓ_2 certification, we compare our method with [9] on CIFAR-10 and ImageNet. For a fair comparison, we use the same pre-trained models as [9], which is trained with Gaussian noise on both CIFAR-10 and ImageNet dataset. Table 2 reports the certified accuracy of our method and the baseline on CIFAR-10 and ImageNet . We find that our method consistently outperforms the baseline. The readers are referred to the Appendix B.3 for detailed ablation studies.

5.2 ℓ_∞ Certification

Toy Example We first construct a simple toy example to verify the advantages of the distribution Eq.13 and Eq.12 over the ℓ_2 family in Eq.11. We set the true classifier to be $f^\sharp(\boldsymbol{x}) = \mathbb{I}(\|x\|_2 \leq r)$ in $r = 0.65$, $d = 5$ case and plot in Fig.3 the Pareto frontier of the accuracy and robustness terms in Eq.9 for the three families of smoothing distributions, as we search for the best combinations of parameters (k, σ). The mixed norm smoothing distribution clearly obtain the best trade-off on accuracy and robustness, and hence guarantees a tighter lower bound for certification. Fig.3 also shows that Eq.12 even performs worse than Eq.11. We further theoretically show that Eq.12 is provably not suitable for ℓ_∞ region certification in Appendix A.5.

CIFAR-10 Based on above results, we only compared the method defined by Eq.13 with[10] on CIFAR-10. The certified accuracy of our method and the baseline using Gaussian smoothing distribution and Proposition 4 are shown in Table 3. We can see that our method consistently outperforms the Gaussian baseline by a large margin. More clarification about ℓ_∞ experiments is in Appendix **??**.

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.

To further confirm the advantage of our method, we plot in Fig.4 the certified accuracy of our method and Gaussian baseline using models trained with Gaussian perturbation of different variances σ_0 under different ℓ_∞ radius. Our approach outperforms baseline consistently, especially when the ℓ_∞ radius is large. We also experimented our method and baseline on ImageNet but did not obtain non-trivial results. This is because ℓ_∞ verification is extremely hard with very large dimensions[31,32]. Future work will investigate how to obtain non-trivial bounds for ℓ_∞ attacking at ImageNet scales with smoothing classifiers.

6 Conclusion

We propose a general functional optimization based framework of adversarial certification with non-Gaussian smoothing distributions. Based on the insights from our new framework and high dimensional geometry, we propose a new family of non-Gaussian smoothing distributions, which outperform the Gaussian and Laplace smoothing for certifying ℓ_1, ℓ_2 and ℓ_∞ attacking. Our work provides a basis for a variety of future directions, including improved methods for ℓ_p attacks, and tighter bounds based on adding additional constraints to our optimization framework.

Broader Impact

Adversarial certification via randomized smoothing could achieve *guaranteed* robust machine learning models, thus has wide application on AI security. a & b) With our empirical results, security engineers could get better performance on defending against vicious attacks; With our theoretical results, it will be easier for following researchers to derive new bounds for different kinds of smoothing methods. We don't foresee the possibility that it could bring negative social impacts. c) Our framework is mathematically rigorous thus would never fail. d) Our method doesn't have bias in data as we provide a general certification method for all tasks and data, and our distribution is not adaptive towards data.

Acknowledgement

This work is supported by Beijing Nova Program (No. 202072) from Beijing Municipal Science & Technology Commission. This work is also supported in part by NSF CAREER 1846421. We would like to thank Tongzheng Ren, Jiaye Teng, Yang Yuan and the reviewers for helpful suggestions that improved the paper.

References

[1] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

[2] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training, 2019.

[3] Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

[4] Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. pages 274–283, 2018.

[5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. 2018.

[6] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. *Advances in neural information processing systems (NeurIPS)*, 2019.

[7] Dilin Wang, Chengyue Gong, and Qiang Liu. Improving neural language modeling via adversarial training. pages 6555–6565, 2019.

[8] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*, 2018.

[9] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

[10] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sebastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. *arXiv preprint arXiv:1906.04584*, 2019.

[11] Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi S Jaakkola. A stratified approach to robustness for randomly smoothed classifiers. *Advances in neural information processing systems (NeurIPS)*, 2019.

[12] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Second-order adversarial attack and certifiable robustness. *Advances in neural information processing systems (NeurIPS)*, 2019.

[13] Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020.

[14] Jiaye Teng, Guang-He Lee, and Yang Yuan. ℓ_1 adversarial robustness certificates: a randomized smoothing approach, 2020.

[15] Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *International Conference on Learning Representations*, 2020.

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

[17] Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy A. Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 550–559, 2018.

[18] Matt Jordan, Justin Lewis, and Alexandros G Dimakis. Provable certificates for adversarial examples: Fitting a ball in the union of polytopes. *arXiv preprint arXiv:1903.08778*, 2019.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

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

[21] Nicholas Carlini, Guy Katz, Clark Barrett, and David L. Dill. Provably minimally-distorted adversarial examples, 2017.

[22] Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 269–286. Springer, 2017.

[23] Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess. Maximum resilience of artificial neural networks. In *Automated Technology for Verification and Analysis - 15th International Symposium, ATVA 2017, Pune, India, October 3-6, 2017, Proceedings*, pages 251–268, 2017.

[24] Souradeep Dutta, Susmit Jha, Sriram Sankaranarayanan, and Ashish Tiwari. Output range analysis for deep feedforward neural networks, 2018.

[25] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples, 2018.

[26] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in neural information processing systems*, pages 4939–4948, 2018.

[27] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization, 2018.

[28] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–385, 2018.

[29] Wenbo V Li and James Kuelbs. Some shift inequalities for gaussian measures. In *High dimensional probability*, pages 233–243. Springer, 1998.

[30] Greg Yang, Tony Duan, Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. *arXiv preprint arXiv:2002.08118*, 2020.

[31] Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. Random smoothing might be unable to certify linf robustness for high-dimensional images. *arXiv preprint arXiv:2002.03517*, 2020.

[32] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. *arXiv preprint arXiv:2002.03239*, 2020.

[33] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

A Proofs

A.1 Proof for Theorem 1

A.1.1 Proof for (I) and (II)

First, observe that the constraint in Equation (3) can be equivalently replaced by an inequality constraint $f_{\pi_0}(\boldsymbol{x}_0) \geq f^\sharp_{\pi_0}(\boldsymbol{x}_0)$. Therefore, the Lagrangian multiplier can be restricted to be $\lambda \geq 0$. We have

$$\begin{aligned}
\mathcal{L}_{\pi_0}(\mathcal{F}, \mathcal{B}) &= \min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}} \max_{\lambda\geq 0} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}[f(\boldsymbol{x}_0+\boldsymbol{z})] + \lambda\left(f^\sharp_{\pi_0}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_0}[f(\boldsymbol{x}_0+\boldsymbol{z})]\right) \\
&\geq \max_{\lambda\geq 0} \min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}[f(\boldsymbol{x}_0+\boldsymbol{z})] + \lambda\left(f^\sharp_{\pi_0}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_0}[f(\boldsymbol{x}_0+\boldsymbol{z})]\right) \\
&= \max_{\lambda\geq 0} \min_{\boldsymbol{\delta}\in\mathcal{B}} \left\{\lambda f^\sharp_{\pi_0}(\boldsymbol{x}_0) \;+\; \min_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}[f(\boldsymbol{x}_0+\boldsymbol{z})] - \lambda\mathbb{E}_{\pi_0}[f(\boldsymbol{x}_0+\boldsymbol{z})])\right\} \\
&= \max_{\lambda\geq 0} \min_{\boldsymbol{\delta}\in\mathcal{B}} \left\{\lambda f^\sharp_{\pi_0}(\boldsymbol{x}_0) \;-\; \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})\right\}.
\end{aligned}$$

II) follows a straightforward calculation.

A.1.2 Proof for (III), the strong duality

We first introduce the following lemma, which is a straight forward generalization of the strong Lagrange duality to functional optimization case.

Lemma 1. *Given some $\boldsymbol{\delta}^*$, we have*

$$\begin{aligned}
&\max_{\lambda\in\mathbb{R}} \min_{f\in\mathcal{F}_{[0,1]}} \mathbb{E}_{\pi_{\boldsymbol{\delta}^*}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^\sharp_{\pi_0}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_0}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right) \\
= &\min_{f\in\mathcal{F}_{[0,1]}} \max_{\lambda\in\mathbb{R}} \mathbb{E}_{\pi_{\boldsymbol{\delta}^*}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^\sharp_{\pi_0}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_0}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right).
\end{aligned}$$

The proof of Lemma 1 is standard. However, for completeness, we include it here.

Proof. Without loss of generality, we assume $f^\sharp_{\pi_0}(\boldsymbol{x}_0) \in (0,1)$, otherwise the feasible set is trivial.

Let α^* be the value of the optimal solution of the primal problem. We define $f^\sharp_{\pi_0}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_0}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] = h[f]$ and $g[f] = \mathbb{E}_{\pi_{\boldsymbol{\delta}^*}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]$. We define the following two sets:

$$\begin{aligned}
\mathcal{A} &= \left\{(v,t)\in\mathbb{R}\times\mathbb{R} : \exists f\in\mathcal{F}_{[0,1]}, h[f]=v, g[f]\leq t\right\} \\
\mathcal{B} &= \left\{(0,s)\in\mathbb{R}\times\mathbb{R} : s<\alpha^*\right\}.
\end{aligned}$$

Notice that both sets $\mathcal{A}$ and $\mathcal{B}$ are convex. This is obvious for $\mathcal{B}$. For any $(v_1,t_1)\in\mathcal{A}$ and $(v_2,t_2)\in\mathcal{A}$, we define $f_1\in\mathcal{F}_{[0,1]}$ such that $h[f_1]=v_1, g[f_1]\leq t_1$ (and similarly we define f_2). Notice that for any $\gamma\in[0,1]$, we have

$$\begin{aligned}
\gamma f_1 + (1-\gamma) f_2 &\in \mathcal{F}_{[0,1]} \\
\gamma h[f_1] + (1-\gamma) h[f_2] &= \gamma v_1 + (1\text{-}\gamma) v_2 \\
\gamma g[f_1] + (1-\gamma) g[f_2] &\leq \gamma t_1 + (1-\gamma) t_2,
\end{aligned}$$

which implies that $\gamma(v_1,t_1)+(1-\gamma)(v_2,t_2)\in\mathcal{A}$ and thus $\mathcal{A}$ is convex. Also notice that by definition, $\mathcal{A}\cap\mathcal{B}=\emptyset$. Using separating hyperplane theorem, there exists a point $(q_1,q_2)\neq(0,0)$ and a value α such that for any $(v,t)\in\mathcal{A}$, $q_1 v + q_2 t \geq \alpha$ and for any $(0,s)\in\mathcal{B}$, $q_2 s\leq\alpha$. Notice that we must have $q_2\geq 0$, otherwise, for sufficient s, we will have $q_2 s>\alpha$. We thus have, for any $f\in\mathcal{F}_{[0,1]}$, we have

$$q_1 h[f] + q_2 g[f] \geq \alpha^* \geq q_2\alpha^*.$$

If $q_2>0$, we have

$$\max_{\lambda\in\mathbb{R}} \min_{f\in\mathcal{F}_{[0,1]}} g[f] + \lambda h[f] \geq \min_{f\in\mathcal{F}_{[0,1]}} g[f] + \frac{q_1}{q_2} h[f] \geq \alpha^*,$$

which gives the strong duality. If $q_2=0$, we have for any $f\in\mathcal{F}_{[0,1]}$, $q_1 h[f]\geq 0$ and by the separating hyperplane theorem, $q_1\neq 0$. However, this case is impossible: If $q_1>0$, choosing

$f \equiv 1$ gives $q_1 h[f] = q_1\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - 1\right) < 0$; If $q_1 < 0$, by choosing $f \equiv 0$, we have $q_1 h[f] = q_1\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - 0\right) < 0$. Both cases give contradiction.

□

Based on Lemma 1, we have the proof of the strong duality as follows.

Notice that by Lagrange multiplier method, our primal problem can be rewritten as follows:

$$\min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}_{[0,1]}} \max_{\lambda\in\mathbb{R}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right),$$

and the dual problem is

$$\begin{aligned}
&\max_{\lambda\in\mathbb{R}} \min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}_{[0,1]}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right)\\
=&\max_{\lambda\geq 0} \min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}_{[0,1]}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right).
\end{aligned}$$

By the assumption that for any $\lambda \geq 0$, we have

$$\begin{aligned}
&\max_{\lambda\geq 0} \min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}_{[0,1]}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right)\\
=&\max_{\lambda\geq 0} \min_{f\in\mathcal{F}_{[0,1]}} \mathbb{E}_{\pi_{\boldsymbol{\delta}^*}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right),
\end{aligned}$$

for some $\boldsymbol{\delta}^* \in \boldsymbol{\mathcal{B}}$. We have

$$\begin{aligned}
&\max_{\lambda\in\mathbb{R}} \min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}_{[0,1]}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right)\\
=&\max_{\lambda\geq 0} \min_{f\in\mathcal{F}_{[0,1]}} \mathbb{E}_{\pi_{\boldsymbol{\delta}^*}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right)\\
=&\max_{\lambda\in\mathbb{R}} \min_{f\in\mathcal{F}_{[0,1]}} \mathbb{E}_{\pi_{\boldsymbol{\delta}^*}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right)\\
\overset{*}{=}&\min_{f\in\mathcal{F}_{[0,1]}} \max_{\lambda\in\mathbb{R}} \mathbb{E}_{\pi_{\boldsymbol{\delta}^*}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right)\\
\geq&\min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}_{[0,1]}} \max_{\lambda\in\mathbb{R}} \mathbb{E}_{\pi_{\boldsymbol{\delta}^*}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right] + \lambda\left(f^{\sharp}_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}\left[f(\boldsymbol{x}_0+\boldsymbol{z})\right]\right),
\end{aligned}$$

where the second equality (*) is by Lemma 1.

A.2 Proof for Corollary 1

Proof. Given our confidence lower bound

$$\max_{\lambda\geq 0} \min_{\|\boldsymbol{\delta}\|_1\leq r} \left\{\lambda p_0 - \int \left(\lambda\pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z)\right)_+ dz\right\},$$

One can show that the worst case for $\boldsymbol{\delta}$ is obtained when $\boldsymbol{\delta}^* = (r, 0, \cdots, 0)$ (see following subsection), thus the bound is

$$\max_{\lambda\geq 0}\left\{\lambda p_0 - \int \frac{1}{2b}\exp\left(-\frac{|z_1|}{b}\right)\left[\lambda - \exp\left(\frac{|\ z_1\ | - |z_1+r|}{b}\right)\right]_+ dz_1\right\}.$$

Denote a to be the solution of $\lambda = \exp\left(\frac{|a|-|a+r|}{b}\right)$, then obviously we have

$$a = \begin{cases} -\infty, & b\log\lambda \geq r\\ -\frac{1}{2}\left(b\log\lambda + r\right), & -r < b\log\lambda < r\\ +\infty. & b\log\lambda \leq -r\end{cases}$$

So the bound above is

$$\lambda\int_{z_1>a} \frac{1}{2b}\exp\left(-\frac{|z_1|}{b}\right)dz_1 - \int_{z_1>a}\frac{1}{2b}\exp\left(-\frac{|z_1+r|}{b}\right)dz_1.$$

i) $b\log\lambda \geq r \Leftrightarrow \lambda \geq \exp\left(\frac{r}{b}\right)$
the bound is

$$\max_{\lambda \geq e^{r/b}} \{\lambda p_0 - (\lambda - 1)\} = 1 - \exp\left(\frac{r}{b}\right)(1 - p_0)\,.$$

ii) $-r < b\log\lambda < r \Leftrightarrow \exp\left(-\frac{r}{b}\right) < \lambda < \exp\left(\frac{r}{b}\right)$
the bound is

$$\begin{aligned}
&\max_{\lambda}\left\{\lambda p_0 - \lambda\left[1 - \frac{1}{2}\exp\left(-\frac{b\log\lambda + r}{2b}\right)\right] + \frac{1}{2}\exp\left(\frac{b\log\lambda - r}{2b}\right)\right\} \\
=&\max_{\lambda}\left\{\lambda(p_0 - 1) + \frac{\lambda}{2}\exp\left(-\frac{b\log\lambda + r}{2b}\right) + \frac{1}{2}\exp\left(\frac{b\log\lambda - r}{2b}\right)\right\} \\
=&\frac{1}{2}\exp\left(-\log\left[2(1-p_0)\right] - \frac{r}{b}\right).
\end{aligned}$$

the extremum is achieved when $\hat{\lambda} = \exp\left(-2\log\left[2(1-p_0)\right] - \frac{r}{b}\right)$. Notice that $\hat{\lambda}$ does not necessarily locate in $\left(e^{-r/b}, e^{r/b}\right)$, so the actual bound is always equal or less than $\frac{1}{2}\exp\left(-\log\left[2(1-p_0)\right] - \frac{r}{b}\right)$.

iii) $b\log\lambda \leq -r \Leftrightarrow \lambda \leq \exp\left(-\frac{r}{b}\right)$
the bound is

$$\max_{\lambda \leq \exp\left(-\frac{r}{b}\right)} \lambda \cdot p_0 = p_0 \exp\left(-\frac{r}{b}\right).$$

Since $\hat{\lambda} > e^{r/b} \Leftrightarrow p_0 > 1 - \frac{1}{2}\exp(-\frac{r}{b})$, notice that the lower bound is a concave function w.r.t. λ, making the final lower bound become

$$\begin{cases} 1 - \exp\left(\frac{r}{b}\right)(1-p_0), & \text{when} \quad p_0 > 1 - \frac{1}{2}\exp(-\frac{r}{b}) \\ \frac{1}{2}\exp\left(-\log\left[2(1-p_0)\right] - \frac{r}{b}\right). & \text{otherwise} \end{cases}$$

$\square$

Remark Actually, we have $1 - \exp\left(\frac{r}{b}\right)(1-p_0) \leq \frac{1}{2}\exp\left(-\log\left[2(1-p_0)\right] - \frac{r}{b}\right)$ all the time. Another interesting thing is that both the bound can lead to the same radius bound:

$$\begin{aligned}
1 - \exp\left(\frac{r}{b}\right)(1-p_0) > \frac{1}{2} &\Leftrightarrow r < -b\log\left[2(1-p_0)\right] \\
\frac{1}{2}\exp\left(-\log\left[2(1-p_0)\right] - \frac{r}{b}\right) > \frac{1}{2} &\Leftrightarrow r < -b\log\left[2(1-p_0)\right]
\end{aligned}$$

A.3 Proof for Corollary 2

Proof. With strong duality, our confidence lower bound is

$$\min_{\|\boldsymbol{\delta}\|_2 \leq r} \max_{\lambda \geq 0} \left\{\lambda p_0 - \int \left(\lambda \pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z)\right)_+ dz\right\},$$

define $C_\lambda = \{z : \lambda\pi_{\mathbf{0}}(z) \geq \pi_{\boldsymbol{\delta}}(z)\} = \{z : \boldsymbol{\delta}^\top z \leq \frac{\|\boldsymbol{\delta}\|^2}{2} + \sigma^2 \ln\lambda\}$ and $\Phi(\cdot)$ to be the cdf of standard gaussian distribution, then

$$\begin{aligned}
&\int \left(\lambda\pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z)\right)_+ dz \\
=&\int_{C_\lambda} \left(\lambda\pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z)\right) dz \\
=&\lambda \cdot \mathbb{P}\left(N(z; \mathbf{0}, \sigma^2\boldsymbol{I}) \in C_\lambda\right) - \mathbb{P}\left(N(z; \boldsymbol{\delta}, \sigma^2\boldsymbol{I}) \in C_\lambda\right) \\
=&\lambda \cdot \Phi\left(\frac{\|\boldsymbol{\delta}\|_2}{2\sigma} + \frac{\sigma\ln\lambda}{\|\boldsymbol{\delta}\|_2}\right) - \Phi\left(\frac{-\|\boldsymbol{\delta}\|_2}{2\sigma} + \frac{\sigma\ln\lambda}{\|\boldsymbol{\delta}\|_2}\right).
\end{aligned}$$

Define

$$F(\boldsymbol{\delta},\lambda) := \lambda p_0 - \int \left(\lambda\pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z)\right)_+ dz = \lambda p_0 - \lambda\cdot\Phi\left(\frac{\|\boldsymbol{\delta}\|_2}{2\sigma} + \frac{\sigma\ln\lambda}{\|\boldsymbol{\delta}\|_2}\right) + \Phi\left(\frac{-\|\boldsymbol{\delta}\|_2}{2\sigma} + \frac{\sigma\ln\lambda}{\|\boldsymbol{\delta}\|_2}\right).$$

For $\forall\boldsymbol{\delta}$, F is a concave function w.r.t. λ, as F is actually a summation of many concave piece wise linear function. See[33] for more discussions of properties of concave functions.

Define $\hat{\lambda}_{\boldsymbol{\delta}} = \exp\left(\frac{2\sigma\|\boldsymbol{\delta}\|_2\Phi^{-1}(p_0) - \|\boldsymbol{\delta}\|_2^2}{2\sigma^2}\right)$, simple calculation can show $\frac{\partial F(\boldsymbol{\delta},\lambda)}{\partial\lambda}|_{\lambda=\hat{\lambda}_{\boldsymbol{\delta}}} = 0$, which means

$$\begin{aligned}
\min_{\|\boldsymbol{\delta}\|_2\le r}\max_{\lambda\ge 0} F(\boldsymbol{\delta},\lambda) &= \min_{\|\boldsymbol{\delta}\|_2\le r} F(\boldsymbol{\delta},\lambda_{\boldsymbol{\delta}}) \\
&= \min_{\|\boldsymbol{\delta}\|_2\le r}\left\{0 + \Phi\left(\frac{-\|\boldsymbol{\delta}\|_2}{2\sigma} + \frac{\sigma\ln\hat{\lambda}_{\boldsymbol{\delta}}}{\|\boldsymbol{\delta}\|_2}\right)\right\} \\
&= \min_{\|\boldsymbol{\delta}\|_2\le r}\Phi\left(\Phi^{-1}(p_0) - \frac{\|\boldsymbol{\delta}\|_2}{\sigma}\right) \\
&= \Phi\left(\Phi^{-1}(p_0) - \frac{r}{\sigma}\right)
\end{aligned}$$

This tells us

$$\min_{\|\boldsymbol{\delta}\|_2\le r}\max_{\lambda\ge 0} F(\boldsymbol{\delta},\lambda) > 1/2 \Leftrightarrow \Phi\left(\Phi^{-1}(p_0) - \frac{r}{\sigma}\right) > 1/2 \Leftrightarrow r < \sigma\cdot\Phi^{-1}(p_0),$$

i.e. the certification radius is $\sigma\cdot\Phi^{-1}(p_0)$. This is exactly the core theoretical contribution of[9]. This bound has a straight forward expansion for multi-class classification situations, we refer interesting readers to Appendix C. □

A.4 Proof For Theorem 2 and 3

A.4.1 Proof for ℓ_2 and ℓ_∞ cases

Here we consider a more general smooth distribution $\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_\infty^{-k_1}\|\boldsymbol{z}\|_2^{-k_2}\exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$, for some $k_1, k_2 \ge 0$ and $\sigma > 0$. We first gives the following key theorem shows that $\mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$ increases as $|\delta_i|$ becomes larger for every dimension i.

Theorem 4. *Suppose $\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_\infty^{-k_1}\|\boldsymbol{z}\|_2^{-k_2}\exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right)$, for some $k_1, k_2 \ge 0$ and $\sigma > 0$, for any $\lambda \ge 0$ we have*

$$\operatorname{sgn}(\delta_i)\frac{\partial}{\partial\delta_i}\mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}) \ge 0,$$

for any $i \in \{1,2,...,d\}$.

Theorem 2 and 3 directly follows the above theorem. Notice that in Theorem 2, as our distribution is spherical symmetry, it is equivalent to set $\boldsymbol{\mathcal{B}} = \left\{\boldsymbol{\delta} : \boldsymbol{\delta} = [a,0,...,0]^\top, a \le r\right\}$ by rotating the axis.

Proof. Given λ, k_1 and k_2, we define $\phi_1(s) = s^{-k_1}$, $\phi_2(s) = s^{-k_2}e^{-\frac{s^2}{\sigma^2}}$. Notice that ϕ_1 and ϕ_2 are monotone decreasing for non-negative s. By the symmetry, without loss of generality, we assume $\boldsymbol{\delta} = [\delta_1,...,\delta_d]^\top$ for $\delta_i \ge 0$, $i \in [d]$. Notice that

$$\begin{aligned}
\frac{\partial}{\partial\delta_i}\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty &= \mathbb{I}\{\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty = |x_i - \delta_i|\}\frac{\partial}{\partial\delta_i}\sqrt{(x_i-\delta_i)^2} \\
&= \mathbb{I}\{\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty = |x_i - \delta_i|\}\frac{-(x_i-\delta_i)}{\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty}.
\end{aligned}$$

And also

$$\frac{\partial}{\partial \delta_i} \left\| \boldsymbol{x}_0 - \boldsymbol{\mu} \right\|_2 = \frac{\partial}{\partial \delta_i} \sqrt{\sum_i \left(x_i - \mu_i\right)^2}$$
$$= \frac{-\left(x_i - \mu_i\right)}{\left\| \boldsymbol{x}_0 - \boldsymbol{\mu} \right\|_2}.$$

We thus have

$$\begin{aligned}
&\frac{\partial}{\partial \delta_1} \int \left(\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) - \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right)_+ d\boldsymbol{x}_0 \\
&= - \int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right\} \frac{\partial}{\partial \delta_1} \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0) d\boldsymbol{x}_0 \\
&= \int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right\} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0 \\
&= \int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > \delta_1\right\} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0 \\
&+ \int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < \delta_1\right\} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0,
\end{aligned}$$

where we define

$$\begin{aligned}
&F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) \\
&= \phi_1'\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty\right) \phi_2\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) \mathbb{I}\{\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty = |x_1 - \delta_1|\} \frac{(x_1 - \delta_1)}{\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty} \\
&+ \phi_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty\right) \phi_2'\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) \frac{(x_1 - \delta_1)}{\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2}.
\end{aligned}$$

Notice that as $\phi_1' \leq 0$ and $\phi_2' \leq 0$ and we have

$$\int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > \delta_1\right\} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0 \leq 0$$
$$\int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < \delta_1\right\} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0 \geq 0.$$

Our target is to prove that $\frac{\partial}{\partial \delta_1} \int \left(\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) - \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right)_+ d\boldsymbol{x}_0 \geq 0$. Now define the set

$$\begin{aligned}
H_1 &= \left\{\boldsymbol{x}_0 : \lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > \delta_1\right\} \\
H_2 &= \left\{\left[2\delta_1 - x_1, x_2, ..., x_d\right]^\top : \boldsymbol{x}_0 = \left[x_1, ..., x_d\right]^\top \in H_1\right\}.
\end{aligned}$$

Here the set H_2 is defined as a image of a bijection

$$\operatorname{proj}(\boldsymbol{x}_0) = \left[2\delta_1 - x_1, x_2, ..., x_d\right]^\top = \tilde{\boldsymbol{x}_0},$$

that is constrained on the set H_1. Notice that under our definition,

$$\begin{aligned}
&\int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > \delta_1\right\} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0 \\
&= \int_{H_1} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0.
\end{aligned}$$

Now we prove that

$$\begin{aligned}
&\int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < \delta_1\right\} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0 \\
&\overset{(1)}{\geq} \int_{H_2} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0 \\
&\overset{(2)}{=} \left| \int_{H_1} F_1\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_\infty, \left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_2\right) d\boldsymbol{x}_0 \right|.
\end{aligned}$$

Property of the projection Before we prove the (1) and (2), we give the following property of the defined projection function. For any $\tilde{\boldsymbol{x}_0} = \mathrm{proj}(\boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in H_1$, we have

$$\begin{aligned} \|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty &= \|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_\infty \\ \|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_2 &= \|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_2 \\ \|\boldsymbol{x}_0\|_2 &\geq \|\tilde{\boldsymbol{x}_0}\|_2 \\ \|\boldsymbol{x}_0\|_\infty &\geq \|\tilde{\boldsymbol{x}_0}\|_\infty . \end{aligned}$$

This is because

$$\begin{aligned} \tilde{x}_i &= x_i, i \in [d] - \{1\} \\ \tilde{x}_1 &= 2\delta_1 - x_1, \end{aligned}$$

and by the fact that $x_1 \geq \delta_1 \geq 0$, we have $|\tilde{x}_1| \leq |x_1|$ and $|\tilde{x}_1 - \delta_1| \leq |x_1 - \delta_1|$.

Proof of Equality (2) By the fact that proj is bijective constrained on the set H_1 and the property of proj, we have

$$\begin{aligned} &\int_{H_2} F_1\left(\|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_\infty, \|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_2\right) d\tilde{\boldsymbol{x}_0} \\ &= \int_{H_2} \phi_1'\left(\|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_\infty\right) \phi_2\left(\|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_2\right) \mathbb{I}\{\|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_\infty = |\tilde{x}_1 - \delta_1|\} \frac{(\tilde{x}_1 - \delta_1)}{\|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_\infty} d\tilde{\boldsymbol{x}_0} \\ &+ \int_{H_2} \phi_1\left(\|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_\infty\right) \phi_2'\left(\|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_2\right) \frac{(\tilde{x}_1 - \delta_1)}{\|\tilde{\boldsymbol{x}_0} - \boldsymbol{\delta}\|_2} d\tilde{\boldsymbol{x}_0} \\ &\overset{(*)}{=} \int_{H_1} \phi_1'\left(\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty\right) \phi_2\left(\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_2\right) \mathbb{I}\{\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty = |x_1 - \delta_1|\} \frac{(\delta_1 - x_1)}{\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty} \left|\det\left(\boldsymbol{J}\right)\right| d\boldsymbol{x}_0 \\ &+ \int_{H_1} \phi_1\left(\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty\right) \phi_2'\left(\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_2\right) \frac{(\delta_1 - x_1)}{\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_2} d\boldsymbol{x}_0 \\ &= -\int_{H_1} F_1\left(\|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_\infty, \|\boldsymbol{x}_0 - \boldsymbol{\delta}\|_2\right) d\boldsymbol{x}_0, \end{aligned}$$

where $(*)$ is by change of variable $\tilde{\boldsymbol{x}_0} = \mathrm{proj}(\boldsymbol{x}_0)$ and $\boldsymbol{J}$ is the Jacobian matrix $\boldsymbol{J} = \begin{bmatrix} -1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$ and here we have the fact that $\tilde{x}_1 - \delta_1 = (2\delta_1 - x_1) - \delta_1 = -(x_1 - \delta_1)$.

Proof of Inequality (1) This can be done by verifying that $H_2 \subseteq \{\boldsymbol{x}_0 : \lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < \delta_1\}$. By the property of the projection, for any $\boldsymbol{x}_0 \in H_1$, let $\tilde{\boldsymbol{x}_0} = \mathrm{proj}(\boldsymbol{x}_0)$, then $\lambda\pi_{\mathbf{0}}(\tilde{\boldsymbol{x}_0}) \geq \lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0) = \pi_{\boldsymbol{\delta}}(\tilde{\boldsymbol{x}_0})$ (by the fact that t ϕ_1 and ϕ_2 are monotone decreasing). It implies that for any $\tilde{\boldsymbol{x}_0} \in H_2$, we have $\lambda\pi_{\mathbf{0}}(\tilde{\boldsymbol{x}_0}) \geq \pi_{\boldsymbol{\delta}}(\tilde{\boldsymbol{x}_0})$ and thus $H_2 \subseteq \{\boldsymbol{x}_0 : \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < \delta_1\}$.

Final statement By the above result, we have

$$\frac{\partial}{\partial \delta_1} \int \left(\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) - \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right)_+ d\boldsymbol{x}_0 \geq 0,$$

and the same result holds for any $\frac{\partial}{\partial \delta_1} \int \left(\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) - \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right)_+ d\boldsymbol{x}_0, i \in [d]$, which implies our result. □

A.4.2 Proof for ℓ_1 case

Slightly different for former cases, apart from proving $\frac{\partial}{\partial \delta_i} \mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}\right) \geq 0$ for $\forall \delta_i \geq 0$, we also need to demonstrate

Theorem 5. *Suppose* $\pi_{\mathbf{0}}(\boldsymbol{x}_0) \propto \|\boldsymbol{x}_0\|^{-k} \exp\left(-\frac{\|\boldsymbol{x}_0\|_1}{b}\right)$, *then for* $\boldsymbol{\delta} = (r, d-r, \delta_3, \delta_4, \cdots)$ *and* $\tilde{\boldsymbol{\delta}} = (0, d, \delta_3, \delta_4, \cdots)$, $0 < r < d$, *we have*

$$\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}\right) \geq \mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\tilde{\boldsymbol{\delta}}}\right)$$

Proof. We turn to show that

$$\frac{\partial}{\partial r}\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}\right) \leq 0,$$

for $\boldsymbol{\delta} = (r, d-r, \delta_3, \delta_4, \cdots)$ and $r < d/2$. We define $\phi(s) = s^{-k}\exp(-\frac{s}{b})$. With

$$\frac{\partial}{\partial \delta_i}\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1 = \frac{\partial}{\partial \delta_i}|x_i - \delta_i| = -\mathrm{sgn}(x_i - \delta_i) = \frac{\delta_i - x_i}{|x_i - \delta_i|},$$

We have

$$\begin{aligned}
&\frac{\partial}{\partial r}\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}\right) \\
=& -\int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right\}\frac{\partial}{\partial r}\pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
=& \int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right\}F(\boldsymbol{x}_0)d\boldsymbol{x}_0,
\end{aligned}$$

where

$$\begin{aligned}
F(\boldsymbol{x}_0) =& -\frac{\partial}{\partial r}\phi\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1\right) = -\phi'\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1\right)\frac{\partial}{\partial r}\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1 \\
=& \phi'\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1\right)\frac{\partial}{\partial r}\left(|x_1 - r| + |x_2 - d + r|\right) \\
=& \phi'\left(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1\right)\cdot\left(\mathrm{sgn}(x_1 - r) + \mathrm{sgn}(d - x_2 - r)\right).
\end{aligned}$$

Thus the original derivative becomes

$$\begin{aligned}
=& \int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > r, x_2 < d - r\right\}F(\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
&+ \int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > r, x_2 > d - r\right\}F(\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
&+ \int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < r, x_2 > d - r\right\}F(\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
&+ \int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < r, x_2 < d - r\right\}F(\boldsymbol{x}_0)d\boldsymbol{x}_0 \\
=& 2\int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > r, x_2 < d - r\right\}\phi'(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1)d\boldsymbol{x}_0 \\
&- 2\int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < r, x_2 > d - r\right\}\phi'(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1)d\boldsymbol{x}_0
\end{aligned}$$

We only need to show that

$$\begin{aligned}
&\int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > r, x_2 < d - r\right\}\phi'(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1)d\boldsymbol{x}_0 \geq \\
&\quad\int \mathbb{I}\left\{\lambda\pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < r, x_2 > d - r\right\}\phi'(\left\|\boldsymbol{x}_0 - \boldsymbol{\delta}\right\|_1)d\boldsymbol{x}_0.
\end{aligned}$$

Notice that $r < d/2$, therefore this can be proved with a similar projection $\boldsymbol{x}_0 \mapsto \tilde{\boldsymbol{x}_0}$:

$$(x_1, x_2, x_3, x_4, \cdots) \mapsto (2r - x_1, 2d - 2r - x_2, x_3, x_4, \cdots)$$

and the similar deduction as previous theorem.

□

A.5 Theoretical Demonstration about the Ineffetivity of Equation (12)

Theorem 6. *Consider the adversarial attacks on the ℓ_∞ ball $\mathcal{B}_{\ell_\infty,r} = \{\boldsymbol{\delta} : \left\|\boldsymbol{\delta}\right\|_\infty \leq r\}$. Suppose we use the smoothing distribution $\pi_{\mathbf{0}}$ in Equation (12) and choose the parameters (k, σ) such that*

1) $\|\boldsymbol{z}\|_\infty$ is stochastic bounded when $\boldsymbol{z} \sim \pi_{\mathbf{0}}$, in that for any $\epsilon > 0$, there exists a finite $M > 0$ such that $\mathrm{P}_{\pi_{\mathbf{0}}}(|\boldsymbol{z}| > M) \le \epsilon$;

2) the mode of $\|\boldsymbol{z}\|_\infty$ under $\pi_{\mathbf{0}}$ equals Cr, where C is some fixed positive constant,

then for any $\epsilon \in (0, 1)$ and sufficiently large dimension d, there exists a constant $t > 1$, such that , we have

$$\max_{\boldsymbol{\delta} \in \mathcal{B}_{\ell_\infty, r}} \left\{ \mathbb{D}_{\mathcal{F}_{[0,1]}} \left(\lambda \pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}} \right) \right\} \ge (1 - \epsilon) \left(\lambda - \mathcal{O}(t^{-d}) \right).$$

This shows that, in very high dimensions, the maximum distance term is arbitrarily close to λ which is the maximum possible value of $\mathbb{D}_{\mathcal{F}_{[0,1]}} \left(\lambda \pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}} \right)$ (see Theorem 1). In particular, this implies that in high dimensional scenario, once $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) \le (1 - \epsilon)$ for some small ϵ, we have $\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}_{[0,1]}, \, \mathcal{B}_{\ell_\infty, r}) = \mathcal{O}(t^{-d})$ and thus fail to certify.

Remark The condition 1) and 2) in Theorem 6 are used to ensure that the magnitude of the random perturbations generated by $\pi_{\mathbf{0}}$ is within a reasonable range such that the value of $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)$ is not too small, in order to have a high accuracy in the trade-off in Equation (9). Note that the natural images are often contained in cube $[0, 1]^d$. If $\|\boldsymbol{z}\|_\infty$ is too large to exceed the region of natural images, the accuracy will be obviously rather poor. Note that if we use variants of Gaussian distribution, we only need $||\boldsymbol{z}||_2/\sqrt{d}$ to be not too large. Theorem 6 says that once $\|\boldsymbol{z}\|_\infty$ is in a reasonably small scale, the maximum distance term must be unreasonably large in high dimensions, yielding a vacuous lower bound.

Proof. First notice that the distribution of $\boldsymbol{z}$ can be factorized by the following hierarchical scheme:

$$\begin{aligned} & a \sim \pi_R(a) \propto a^{d-1-k} e^{-\frac{a^2}{2\sigma^2}} \mathbb{I}\{a \ge 0\} \\ & \boldsymbol{s} \sim \mathrm{Unif}^{\otimes d}(-1, 1) \\ & \boldsymbol{z} \leftarrow \frac{\boldsymbol{s}}{\|\boldsymbol{s}\|_\infty} a. \end{aligned}$$

Without loss of generality, we assume $\boldsymbol{\delta}^* = [r, ..., r]^\top$. (see Theorem 4)

$$\mathbb{D}_{\mathcal{F}_{[0,1]}} \left(\lambda \pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}^*} \right) = \mathbb{E}_{\boldsymbol{z} \sim \pi_{\mathbf{0}}} \left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}) \right)_+.$$

Notice that as the distribution is symmetry,

$$\mathrm{P}_{\pi_{\mathbf{0}}} \left(\|\boldsymbol{z} + \boldsymbol{\delta}^*\|_\infty = a + r \;\mid\; \|\boldsymbol{z}\|_\infty = a \right) = \frac{1}{2}.$$

Define $|z|^{(i)}$ is the i-th order statistics of $|z_j|$, $j = 1, ..., d$ conditioning on $\|\boldsymbol{z}\|_\infty = a$. By the factorization above and some algebra, we have, for any $\epsilon \in (0, 1)$,

$$\mathrm{P} \left(\frac{|z|^{(d-1)}}{|z|^{(d)}} > (1 - \epsilon) \mid \|\boldsymbol{z}\|_\infty = a \right) \ge 1 - (1 - \epsilon)^{d-1}.$$

And $\frac{|z|^{(d-1)}}{|z|^{(d)}} \perp |z|^{(d)}$. Now we estimate $\mathbb{D}_{\mathcal{F}_{[0,1]}} \left(\lambda \pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}^*} \right)$.

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{z} \sim \pi_{\mathbf{0}}} \left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}) \right)_+ \\ = & \mathbb{E}_a \mathbb{E}_{\boldsymbol{z} \sim \pi_{\mathbf{0}}} \left[\left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}) \right)_+ \mid \|\boldsymbol{z}\|_\infty = a \right] \\ = & \frac{1}{2} \mathbb{E}_a \mathbb{E}_{\boldsymbol{z} \sim \pi_{\mathbf{0}}} \left[\left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}) \right)_+ \mid \|\boldsymbol{z}\|_\infty = a, \|\boldsymbol{z} + \boldsymbol{\delta}^*\|_\infty = a + r \right] \\ + & \frac{1}{2} \mathbb{E}_a \mathbb{E}_{\boldsymbol{z} \sim \pi_{\mathbf{0}}} \left[\left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}) \right)_+ \mid \|\boldsymbol{z}\|_\infty = a, \|\boldsymbol{z} + \boldsymbol{\delta}^*\|_\infty \ne a + r \right]. \end{aligned}$$

Conditioning on $\|\boldsymbol{z}\|_\infty = a, \|\boldsymbol{z}+\boldsymbol{\delta}^*\|_\infty = a+r$, we have

$$\begin{aligned}\frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}) &= \left(\frac{1}{1+\frac{r}{a}}\right)^k e^{-\frac{1}{2\sigma^2}\left(2ra+r^2\right)}\\ &= \left(\frac{1}{1+\frac{r}{a}}\right)^k e^{-\frac{d-1-k}{2C^2}\left(2\frac{a}{r}+1\right)}.\end{aligned}$$

Here the second equality is because we choose $\mathrm{mode}(\|\boldsymbol{z}\|_\infty) = Cr$, which implies that $\sqrt{d-1-k}\sigma = Cr$. And thus we have

$$\begin{aligned}&\mathbb{E}_a\mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}\left[\left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z})\right)_+ \mid \|\boldsymbol{z}\|_\infty = a, \|\boldsymbol{z}+\boldsymbol{\delta}^*\|_\infty = a+r\right]\\ =&\int \left(\lambda - \left(\frac{1}{1+\frac{r}{a}}\right)^k e^{-\frac{d-1-k}{2C^2}\left(2\frac{a}{r}+1\right)}\right)_+ \pi(a)da\\ =&\int \left(\lambda - \left(1+\frac{r}{a}\right)^{-k}\left(e^{\frac{2a/r+1}{2C^2}}\right)^{-(d-1-k)}\right)_+ \pi(a)da\\ =&\lambda - \mathcal{O}(t^{-d}),\end{aligned}$$

for some $t>1$. Here the last equality is by the assumption that $\|\boldsymbol{z}\|_\infty = \mathcal{O}_p(1)$.

Next we bound the second term $\mathbb{E}_a\mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}\left[\left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z})\right)_+ \mid \|\boldsymbol{z}\|_\infty = a, \|\boldsymbol{z}+\boldsymbol{\delta}^*\|_\infty \neq a+r\right]$. By the property of uniform distribution, we have

$$\begin{aligned}&\mathrm{P}\left(\frac{|z|^{(d-1)}}{|z|^{(d)}} > (1-\epsilon) \mid \|\boldsymbol{z}\|_\infty = a, \|\boldsymbol{z}+\boldsymbol{\delta}^*\|_\infty \neq a+r\right)\\ =&\mathrm{P}\left(\frac{|z|^{(d-1)}}{|z|^{(d)}} > (1-\epsilon) \mid \|\boldsymbol{z}\|_\infty = a\right)\\ \geq& 1-(1-\epsilon)^{d-1}.\end{aligned}$$

And thus, for any $\epsilon \in [0,1)$,

$$\mathrm{P}\left(\|\boldsymbol{z}+\boldsymbol{\delta}^*\|_\infty \geq ((1-\epsilon)a+r)^2 \mid \|\boldsymbol{z}\|_\infty = a, \|\boldsymbol{z}+\boldsymbol{\delta}^*\|_\infty \neq a+r\right) \geq \frac{1}{2}\left(1-(1-\epsilon)^{d-1}\right).$$

It implies that

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}\left[\left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z})\right)_+ \mid \|\boldsymbol{z}\|_\infty = a, \|\boldsymbol{z}+\boldsymbol{\delta}^*\|_\infty = a+r\right]\\ \geq&\frac{1}{2}\left(1-(1-\epsilon)^{d-1}\right)\left(\lambda - \left(1-\epsilon+\frac{r}{a}\right)^{-k} e^{-\frac{1}{2\sigma^2}\left(\epsilon(\epsilon-2)a^2+2r(1-\epsilon)a+r^2\right)}\right)_+\\ =&\frac{1}{2}\left(1-(1-\epsilon)^{d-1}\right)\left(\lambda - \left(1-\epsilon+\frac{r}{a}\right)^{-k} e^{-\frac{d-1-k}{2C^2}\left(\epsilon(\epsilon-2)a^2/r^2+2(1-\epsilon)a/r+1\right)}\right)_+.\end{aligned}$$

For any $\epsilon' \in (0,1)$, by choosing $\epsilon = \frac{\log(2/\epsilon')}{d-1}$, for large enough d, we have

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}\left[\left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z})\right)_+ \mid \|\boldsymbol{z}\|_\infty = a, \|\boldsymbol{z}+\boldsymbol{\delta}^*\|_\infty = a+r\right]\\ \geq&\frac{1}{2}\left(1-(1-\epsilon)^{d-1}\right)\left(\lambda - \left(1-\epsilon+\frac{r}{a}\right)^{-k} e^{-\frac{d-1-k}{2C^2}(2(1-\epsilon)a/r+1)}e^{\frac{a^2\log(2/\epsilon')}{C^2r^2}}\right)_+\\ =&\frac{1}{2}\left(1-(1-\frac{\log(2/\epsilon')}{d-1})^{d-1}\right)\left(\lambda - \left(1-\frac{\log(2/\epsilon')}{d-1}+\frac{r}{a}\right)^{-k} e^{-\frac{d-1-k}{2C^2}(2(1-\epsilon)a/r+1)}e^{\frac{a^2\log(2/\epsilon')}{C^2r^2}}\right)_+\\ \geq&\frac{1}{2}\left(1-\epsilon'\right)\left(\lambda - \left(1-\epsilon+\frac{r}{a}\right)^{-k} e^{-\frac{d-1-k}{2C^2}(2(1-\epsilon)a/r+1)}e^{\frac{a^2\log(2/\epsilon')}{C^2r^2}}\right)_+.\end{aligned}$$

Thus we have

$$
\begin{aligned}
&\frac{1}{2}\mathbb{E}_a\mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}\left[\left(\lambda-\frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z})\right)_+ \mid \|\boldsymbol{z}\|_\infty=a, \|\boldsymbol{z}+\boldsymbol{\delta}^*\|_\infty\neq a+r\right]\\
=&\frac{1}{2}\left(1-\epsilon'\right)\left(\lambda-\mathcal{O}(t^{-d})\right).
\end{aligned}
$$

Combine the bounds, for large d, we have

$$
\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}^*}\right)=\left(1-\epsilon'\right)\left(\lambda-\mathcal{O}(t^{-d})\right).
$$

□

B More about Experiments

B.1 Practical Algorithm

In this section, we give our algorithm for certification. Our target is to give a high probability bound for the solution of

$$
\mathcal{L}_{\pi_{\mathbf{0}}}(\mathcal{F}_{[0,1]},\,\mathcal{B}_{\ell_\infty,r})=\max_{\lambda\geq 0}\left\{\lambda f^\sharp_{\pi_{\mathbf{0}}}-\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}}\parallel\pi_{\boldsymbol{\delta}}\right)\right\}
$$

given some classifier $f^\sharp$. Following [9], the given classifier here has a binary output $\{0,1\}$. Computing the above quantity requires us to evaluate both $f^\sharp_{\pi_{\mathbf{0}}}$ and $\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}}\parallel\pi_{\boldsymbol{\delta}}\right)$. A lower bound $\hat{p}_0$ of the former term is obtained through binominal test as [9] do, while the second term can be estimated with arbitrary accuracy using Monte Carlo samples. We perform grid search to optimize λ and given λ, we draw N i.i.d. samples from the proposed smoothing distribution $\pi_{\mathbf{0}}$ to estimate $\lambda f^\sharp_{\pi_{\mathbf{0}}}-\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}}\parallel\pi_{\boldsymbol{\delta}}\right)$. This can be achieved by the following importance sampling manner:

$$
\begin{aligned}
&\lambda f^\sharp_{\pi_{\mathbf{0}}}-\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}}\parallel\pi_{\boldsymbol{\delta}}\right)\\
\geq\ &\lambda\hat{p}_0-\int\left(\lambda-\frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z})\right)_+\pi_{\mathbf{0}}(\boldsymbol{z})d\boldsymbol{z}\\
\geq\ &\lambda\hat{p}_0-\frac{1}{N}\sum_{i=1}^{N}\left(\lambda-\frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}_i)\right)_+-\epsilon.
\end{aligned}
$$

And we use reject sampling to obtain samples from $\pi_{\mathbf{0}}$. Notice that, we restrict the search space of λ to a finite compact set so the importance samples is bounded. Since the Monte Carlo estimation is not exact with an error ϵ, we give a high probability concentration lower bound of the estimator. Algorithm 1 summarized our algorithm.

Algorithm 1 Certification algorithm

Input: input image $\boldsymbol{x}_0$; original classifier: $f^\sharp$; smoothing distribution $\pi_{\mathbf{0}}$; radius r; search interval $[\lambda_{\text{start}}, \lambda_{\text{end}}]$ of λ; search precision h for optimizing λ; number of samples N_1 for testing p_0; pre-defined error threshold ϵ; significant level α;

compute search space for $\lambda : \Lambda =$range$(\lambda_{\text{start}}, \lambda_{\text{end}}, h)$

compute N_2: number of Monte Carlo estimation given ϵ, α and Λ

compute optimal disturb: $\boldsymbol{\delta}$ depends on specific setting

for λ *in* Λ **do**

sample $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N_1} \sim \pi_{\mathbf{0}}$

compute $n_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} f^\sharp(\boldsymbol{x}_0 + \boldsymbol{z}_i)$

compute $\hat{p}_0 =$LowerConfBound$(n_1, N_1, 1-\alpha)$

sample $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N_2} \sim \pi_{\mathbf{0}}$

compute $\hat{\mathbb{D}}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}) = \frac{1}{N_2}\sum_{i=1}^{N_2}\left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}_i)\right)_+$

compute confidence lower bound $b_\lambda = \lambda\hat{p}_0 - \hat{\mathbb{D}}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}) - \epsilon$

end

if $\max_{\lambda\in\Lambda} b_\lambda \geq 1/2$ **then**

$\boldsymbol{x}_0$ can be certified

else

$\boldsymbol{x}_0$ cannot be certified

end

The LowerConfBound function performs a binominal test as described in[9]. The ϵ in Algorithm 1 is given by concentration inequality.

Theorem 7. *Let* $h(z_1, \cdots, z_N) = \frac{1}{N}\sum_{i=1}^{N}\left(\lambda - \frac{\pi_{\boldsymbol{\delta}}(z_i)}{\pi_{\mathbf{0}}(z_i)}\right)_+$*, we yield*

$$\Pr\{|h(z_1, \cdots, z_N) - \int (\lambda\pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z))_+ \, dz| \geq \varepsilon\} \leq \exp\left(\frac{-2N\varepsilon^2}{\lambda^2}\right).$$

Proof. Given *McDiarmid's Inequality*, which says

$$\sup_{x_1, x_2, \ldots, x_n, \hat{x}_i} |h(x_1, x_2, \ldots, x_n) - h(x_1, x_2, \ldots, x_{i-1}, \hat{x}_i, x_{i+1}, \ldots, x_n)| \leq c_i \quad \text{for} \quad 1 \leq i \leq n,$$

we have $c_i = \frac{\lambda}{N}$, and then obtain

$$\Pr\{|h(z_1, \cdots, z_N) - \int (\lambda\pi_{\mathbf{0}}(z) - \pi_{\boldsymbol{\delta}}(z))_+ \, dz| \geq \varepsilon\} \leq \exp\left(\frac{-2N\varepsilon^2}{\lambda^2}\right).$$

□

The above theorem tells us that, once ϵ, λ, N is given, we can yield a bound with high-probability $1-\alpha$. One can also get N when $\epsilon, \lambda, \alpha$ is provided. Note that this is the same as the Hoeffding bound mentioned in Section 4.2 as Micdiarmid bound is a generalization of Hoeffding bound.

However, in practice we can use a small trick as below to certify with much less comupation:

Algorithm 2 Practical certification algorithm

Input: input image $\boldsymbol{x}_0$; original classifier: $f^\sharp$; smoothing distribution $\pi_{\mathbf{0}}$; radius r; search interval for λ: $[\lambda_{\text{start}}, \lambda_{\text{end}}]$; search precision h for optimizing λ; number of Monte Carlo for first estimation: N_1^0, N_2^0; number of samples N_1 for a second test of p_0; pre-defined error threshold ϵ; significant level α; optimal perturbation $\boldsymbol{\delta}$ ($\boldsymbol{\delta} = [r, 0, \ldots, 0]^\top$ for ℓ_2 attacking and $\boldsymbol{\delta} = [r, \ldots, r]^\top$ for ℓ_∞ attacking).

for λ *in* Λ **do**

  sample $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N_1^0} \sim \pi_{\mathbf{0}}$

  compute $n_1^0 = \frac{1}{N_1^0} \sum_{i=1}^{N_1^0} f^\sharp(\boldsymbol{x}_0 + \boldsymbol{z}_i)$

  compute $\hat{p}_0$ =LowerConfBound$(n_1^0, N_1^0, 1-\alpha)$

  sample $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N_2^0} \sim \pi_{\mathbf{0}}$

  compute $\hat{\mathbb{D}}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}) = \frac{1}{N_2^0} \sum_{i=1}^{N_2^0} \left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}_i)\right)_+$

  compute confidence lower bound $b_\lambda = \lambda\hat{p}_0 - \hat{\mathbb{D}}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$

end

compute $\hat{\lambda} = \arg\max_{\lambda\in\Lambda} b_\lambda$

compute N_2: number of Monte Carlo estimation given ϵ, α and $\hat{\lambda}$

sample $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N_1} \sim \pi_{\mathbf{0}}$

compute $n_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} f^\sharp(\boldsymbol{x}_0 + \boldsymbol{z}_i)$

compute $\hat{p}_0$ =LowerConfBound$(n_1, N_1, 1-\alpha)$

sample $\boldsymbol{z}_1, \cdots, \boldsymbol{z}_{N_2} \sim \pi_{\mathbf{0}}$

compute $\hat{\mathbb{D}}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}) = \frac{1}{N_2} \sum_{i=1}^{N_2} \left(\lambda - \frac{\pi_{\boldsymbol{\delta}}}{\pi_{\mathbf{0}}}(\boldsymbol{z}_i)\right)_+$

compute $b = \hat{\lambda}\hat{p}_0 - \hat{\mathbb{D}}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}) - \epsilon$

if $b \geq 1/2$ **then**

  $\boldsymbol{x}_0$ can be certified

else

  $\boldsymbol{x}_0$ cannot be certified

end

Algorithm 2 allow one to begin with small N_1^0, N_2^0 to obtain the first estimation and choose a $\hat{\lambda}$. Then a rigorous lower bound can be achieved with $\hat{\lambda}$ with enough (i.e. N_1, N_2) Monte Carlo samples.

B.2 Experiment Settings

The details of our method are shown in the supplementary material. Since our method requires Monte Carlo approximation, we draw $0.1M$ samples from $\pi_{\mathbf{0}}$ and construct $\alpha = 99.9\%$ confidence lower bounds of that in Equation (9). The optimization on λ is solved using grid search. For ℓ_2 attacks, we set $k = 500$ for CIFAR-10 and $k = 50000$ for ImageNet in our non-Gaussian smoothing distribution Equation (11). If the used model was trained with a Gaussian perturbation noise of $\mathcal{N}(0, \sigma_0^2)$, then the σ parameter of our smoothing distribution is set to be $\sqrt{(d-1)/(d-1-k)}\sigma_0$, such that the expectation of the norm $\|\boldsymbol{z}\|_2$ under our non-Gaussian distribution Equation (11) matches with the norm of $\mathcal{N}(0, \sigma_0^2)$. For ℓ_1 situation, we keep the same rule for hyperparameter selection as ℓ_2 case, in order to make the norm of proposed distribution has the same mean with original distribution. For ℓ_∞ situation, we set $k = 250$ and σ also equals to $\sqrt{(d-1)/(d-1-k)}\sigma_0$ for the mixed norm smoothing distribution Equation (13) just for consistency. More ablation study about k is deferred to Appendix B.3.

B.3 Abalation Study

On CIFAR10, we also do ablation study to show the influence of different k for the ℓ_2 certification case as shown in Table 4.

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
$k = 100$ (%)	60	43	34	23	18	15	12	10	8
$k = 200$ (%)	60	44	36	24	18	15	13	10	8
$k = 500$ (%)	**61**	**46**	**37**	**25**	**19**	**16**	14	11	**9**
$k = 1000$ (%)	59	44	36	**25**	**19**	**16**	14	11	**9**
$k = 2000$ (%)	56	41	35	24	**19**	**16**	**15**	**12**	**9**

Table 4: Certified top-1 accuracy of the best classifiers on cifar10 at various ℓ_2 radius. We use the same model as[9] and do not train any new models.

C Illumination about Bilateral Condition[3]

The results in the main context is obtained under binary classfication setting. Here we show it has a natural generalization to multi-class classification setting. Suppose the given classifier $f^\sharp$ classifies an input $\boldsymbol{x}_0$ correctly to class A, i.e.,

$$f_A^\sharp(\boldsymbol{x}_0) > \max_{B \neq A} f_B^\sharp(\boldsymbol{x}_0) \tag{14}$$

where $f_B^\sharp(\boldsymbol{x}_0)$ denotes the prediction confidence of any class B different from ground truth label A. Notice that $f_A^\sharp(\boldsymbol{x}_0) + \sum_{B\neq A} f_B^\sharp(\boldsymbol{x}_0) = 1$, so the necessary and sufficient condition for correct binary classification $f_A^\sharp(\boldsymbol{x}_0) > 1/2$ becomes a *sufficient* condition for multi-class prediction.

Similarly, the necessary and sufficient condition for correct classification of the *smoothed* classifier is

$$\begin{aligned} &\min_{f\in\mathcal{F}}\left\{\mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}[f_A(\boldsymbol{x}_0+\boldsymbol{\delta}+\boldsymbol{z})] \quad \text{s.t.} \quad \mathbb{E}_{\pi_{\mathbf{0}}}[f_A(\boldsymbol{x}_0)] = f^\sharp_{\pi_{\mathbf{0}},A}(\boldsymbol{x}_0)\right\} > \\ &\max_{f\in\mathcal{F}}\left\{\mathbb{E}_{\boldsymbol{z}\sim\pi_{\mathbf{0}}}[f_B(\boldsymbol{x}_0+\boldsymbol{\delta}+\boldsymbol{z})] \quad \text{s.t.} \quad \mathbb{E}_{\pi_{\mathbf{0}}}[f_B(\boldsymbol{x}_0)] = f^\sharp_{\pi_{\mathbf{0}},B}(\boldsymbol{x}_0)\right\}\end{aligned}$$

for $\forall B \neq A$ and any perturbation $\boldsymbol{\delta} \in \mathcal{B}$. Writing out their Langragian forms makes things clear:

$$\max_\lambda \lambda f^\sharp_{\pi_{\mathbf{0}},A}(\boldsymbol{x}_0) - \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}) > \min_\lambda \max_{B\neq A} \lambda f^\sharp_{\pi_{\mathbf{0}},B}(\boldsymbol{x}_0) + \mathbb{D}_{\mathcal{F}_{[0,1]}}(\pi_{\boldsymbol{\delta}} \parallel \lambda\pi_{\mathbf{0}})$$

Thus the overall necessary and sufficient condition is

$$\min_{\boldsymbol{\delta}\in\mathcal{B}}\left\{\max_\lambda\left(\lambda f^\sharp_{\pi_{\mathbf{0}},A}(\boldsymbol{x}_0) - \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})\right) - \max_{B\neq A}\min_\lambda\left(\lambda f^\sharp_{\pi_{\mathbf{0}},B}(\boldsymbol{x}_0) + \mathbb{D}_{\mathcal{F}_{[0,1]}}(\pi_{\boldsymbol{\delta}} \parallel \lambda\pi_{\mathbf{0}})\right)\right\} > 0$$

Optimizing this *bilateral* object will *theoretically give a better certification result* than our method in main context, especially when the number of classes is large. But we do not use this bilateral formulation as reasons stated below.

When both $\pi_{\mathbf{0}}$ and $\pi_{\boldsymbol{\delta}}$ are gaussian, which is[9]'s setting, this condition is equivalent to:

$$\begin{aligned} &\min_{\boldsymbol{\delta}\in\mathcal{B}}\left\{\Phi\left(\Phi^{-1}(f^\sharp_{\pi_{\mathbf{0}},A}(\boldsymbol{x}_0)) - \frac{\|\boldsymbol{\delta}\|_2}{\sigma}\right) - \max_{B\neq A}\Phi\left(\Phi^{-1}(f^\sharp_{\pi_{\mathbf{0}},B}(\boldsymbol{x}_0)) + \frac{\|\boldsymbol{\delta}\|_2}{\sigma}\right)\right\} > 0 \\ \Leftrightarrow \quad &\Phi^{-1}(f^\sharp_{\pi_{\mathbf{0}},A}(\boldsymbol{x}_0)) - \frac{r}{\sigma} > \Phi^{-1}(f^\sharp_{\pi_{\mathbf{0}},B}(\boldsymbol{x}_0)) + \frac{r}{\sigma}, \quad \forall B \neq A \\ \Leftrightarrow \quad &r < \frac{\sigma}{2}\left(\Phi^{-1}(f^\sharp_{\pi_{\mathbf{0}},A}(\boldsymbol{x}_0)) - \Phi^{-1}(f^\sharp_{\pi_{\mathbf{0}},B}(\boldsymbol{x}_0))\right), \forall B \neq A\end{aligned}$$

with a similar derivation process like Appendix A.3. This is exactly the same bound in the (restated) theorem 1 of[9].

[9] use $1 - \underline{p_A}$ as a naive estimate of the upper bound of $f^\sharp_{\pi_{\mathbf{0}},B}(\boldsymbol{x}_0)$, where $\underline{p_A}$ is a lower bound of $f^\sharp_{\pi_{\mathbf{0}},A}(\boldsymbol{x}_0)$. This leads the confidence bound decay to the bound one can get in binary case, i.e., $r \leq \sigma\Phi^{-1}(f^\sharp_{\pi_{\mathbf{0}},A}(\boldsymbol{x}_0))$.

As the two important baselines[9,10] do not take the bilateral form, we also do not use this form in experiments for fairness.

[3]In fact, the theoretical part of[15] share some similar discussion with this section.