FILLING THE SOAP BUBBLES: EFFICIENT BLACK-BOX ADVERSARIAL CERTIFICATION WITH NON-GAUSSIAN SMOOTHING

Anonymous authors
Paper under double-blind review

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, helping to design two new families of non-Gaussian smoothing distributions that work more efficiently for ℓ_2 and ℓ_∞ attacks, respectively. Our proposed methods achieve better results than previous works and provide a new perspective on randomized smoothing certification.

1 INTRODUCTION

Deep neural networks have achieved state-of-the-art performance on many tasks such as image classification (He et al., 2016; Lu et al., 2018) and language modeling (Devlin et al., 2019). Nonetheless, modern deep learning models have been shown to be highly sensitive to small and adversarially crafted perturbations on the inputs (Goodfellow et al., 2015), which means a human-imperceptible changes on inputs could cause the model to make dramatically different predictions. Although many robust training algorithms have been developed to overcome adversarial attacking, most heuristically developed methods can be shown to be broken by more powerful adversaries eventually, (e.g., Athalye et al., 2018; Madry et al., 2018; Zhang et al., 2019; Wang et al., 2019). This casts an urgent demand for developing robust classifiers with provable worst case guarantees.

One promising approach for certifiably robustness is the recent *randomized smoothing method* (e.g., Cohen et al., 2019; Salman et al., 2019; Lee et al., 2019; Li et al., 2019; Lecuyer et al., 2018), which constructs smoothed classifiers with certifiable robustness by introducing noise on the inputs. Compared with the other more traditional verification approaches (e.g. Wong & Kolter, 2017; Jordan et al., 2019; Dvijotham et al., 2018) that exploits special structures of the neural networks (such as the properties of ReLU), the randomized smoothing methods 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 (Deng et al., 2009).

However, the existing randomized smoothing methods can only work against ℓ_2 attack, in which the perturbations are allowed within an ℓ_2 ball of certain radius. A stronger type of attack, such as the ℓ_∞ attacks, is much more challenging to defense and verify due to the larger set of perturbations, but is more relevant in practice.

In addition, all the existing randomized smoothing methods use Gaussian noise for smoothing. Although appearing to be a natural choice, one of our key observations is that Gaussian distributions is in fact a rather sub-optimal choice in high dimensional spaces, even for ℓ_2 attack. This is due to a counter-intuitive phenomenon in high dimensional spaces (Vershynin, 2018) that almost all of the probability mass of standard Gaussian distribution concentrates around the sphere of radius one (and hence "soap bubble" in the title), instead of the center point (which corresponds to the original input). As a result, the variance of the Gaussian noise needs to be sufficiently small to yield good approximation to the original classifiers (by squeezing the "soap bubble" towards the center point),

which, however, makes it difficult to verify due to the small noise. Further, for the more challenging ℓ_∞ attack, Gaussian smoothing provably degenerates in high dimensions.

Our contribution We propose a general framework of adversarial certification using non-Gaussian smoothing noises, based on a new perspective from functional optimization. Our framework re-derives the method of Cohen et al. (2019) as a special case, and is applicable to more general families of non-Gaussian smoothing distributions and more types of attacks beyond ℓ_2 norm. Importantly, our new framework reveals *a fundamental trade-off between accuracy and robustness* for guiding better choices of smoothing distributions. Leveraging our insight, we develop two new families of distributions for better certification results on ℓ_2 and ℓ_∞ attacks, respectively. Efficient computational approaches are developed to enable our method in practice. Empirical results show that our new framework and smoothing distributions significantly outperform the existing approaches for both ℓ_2 and ℓ_∞ attacking, on challenging datasets such as CIFAR-10 and ImageNet.

2 Related works

Empirical Defenses Since Szegedy et al. (2013) and Goodfellow et al. (2015), many previous works have focused on utilizing small perturbation δ under certain constraint, e.g. in a ℓ_p norm ball, to attack a neural network. Adversarial training (Madry et al., 2018) and its variants (Kannan et al., 2018; Zhang & Wang, 2019; Zhai et al., 2019) are the most successful defense methods to date, in which the network is forced to solve a mini-max game between the defender and attacker with adversarial examples as data augmentation. However, these empirical defense methods are still easy to be broken and cannot provide provable defense.

Certified Defenses Unlike the empirical defense methods, once a classifier can guarantee a constant classification within a local region, it is called a robust certificate. *Exact* certification methods provide the minimal perturbation condition which leads to a different classification result. This line of work focus 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 (Carlini et al., 2017; Ehlers, 2017) or mix integer linear programming (Cheng et al., 2017; Dutta et al., 2018). *Sufficient* certification methods take a conservative way and try to bound the Lipschitz constant or other information of the network (Jordan et al., 2019; Wong & Kolter, 2017; Raghunathan et al., 2018; Zhang et al., 2018). However, these certification strategies share a drawback that they are not feasible on large-scale scenarios, e.g. large enough practical networks, large enough datasets.

Randomized Smoothing To mitigate this limitation of previous certifiable defenses, improving network robustness via randomness has been recently discussed (Xie et al., 2018; Liu et al., 2018). In certification community, Lecuyer et al. (2018) first introduced randomization with technique in differential privacy. Li et al. (2019) improved their work with a bound given by Rényi divergence. In succession, Cohen et al. (2019) firstly provided a *tight* bound for *arbitrary* Gaussian smoothed classifiers based on previous theorems found by Li & Kuelbs (1998). Salman et al. (2019) combined the empirical and certification robustness, by applying adversarial training on randomized smoothed classifiers to achieve a higher certified accuracy. Lee et al. (2019) focused on ℓ_0 norm perturbation setting, and proposed a discrete smoothing distribution to beat the Gaussian distribution baseline. Similar to (Lee et al., 2019), we also focus on finding a suitable distribution to trade-off accuracy and robustness for different types of adversarial attacks, such as ℓ_2 and ℓ_∞.

3 Black-box Certification with Functional Optimization

We start with introducing the background of adversarial certification problem and and randomized smoothing method. We then introduce in Section 3.1 our general framework of adversarial certification using non-Gaussian smoothing noises, from a new functional optimization perspective. Our framework includes the method of Cohen et al. (2019) as a special case, and reveals *a critical trade-off between accuracy and robustness* that provides important guidance for better choices of smoothing distributions in Section 4.

Adversarial Certification We consider binary classification of predicting binary labels $y \in \{0, 1\}$ from feature vectors $x \in \mathbb{R}^d$ for simplicity. The extension to multi-class cases is straightforward, and is discussed in Appendix D. Assume $f^\sharp \colon \mathbb{R}^d \to [0, 1]$ is a pre-trained binary classifier ($\sharp$ means the classifier is *given*), which maps from the feature space $\mathbb{R}^d$ to either the class probability in interval $[0, 1]$ or the 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 we perturb the input $\boldsymbol{x}_0$ arbitrarily in $\mathcal{B}$, a given neighborhood of $\boldsymbol{x}_0$. Specifically, let $\mathcal{B}$ be a set of possible perturbation vectors, e.g., $\mathcal{B} = \{\boldsymbol{\delta} \in \mathbb{R}^d : \|\boldsymbol{\delta}\|_p \leq r\}$ 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 $f^\sharp(\boldsymbol{x}_0 + \boldsymbol{\delta}) > 1/2$ holds for any $\delta \in \mathcal{B}$. In this paper, we consider two types of attacks, including the ℓ_2 attack $\mathcal{B}_{\ell_2,r} \stackrel{\text{def}}{=} \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq r\}$, and the ℓ_∞ attack $\mathcal{B}_{\ell_\infty,r} \stackrel{\text{def}}{=} \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_\infty \leq r\}$. More general ℓ_p attack can also be handled by our framework but is left as future works.

Black-box Certification with Randomness Directly verifying $f^\sharp$ heavily relies on the smoothness of $f^\sharp$, which has been explored in a series of recent works (Lecuyer et al., 2018; Wong & Kolter, 2017). 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. We are instead interested in **black-box verification methods** that could work for arbitrary classifiers. One approach to enable this, as explored in recent works (Cohen et al., 2019; Lee et al., 2019), is to replace $f^\sharp$ with a smoothed classifier by convovling 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 verifying $f^\sharp_{\pi_{\mathbf{0}}}$ using its inherent smoothness. Specifically, if $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) > 1/2$, we want to verify that $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta}) > 1/2$ for every $\boldsymbol{\delta} \in \mathcal{B}$, that is,

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

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 Cohen et al. (2019), which shows that a lower bound of

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

where $\Phi(\cdot)$ is the cumulative density function (CDF) of standard Gaussian distribution, and $\Phi^{-1}(\cdot)$ represents its inverse function. The proof of this result in Cohen et al. (2019) uses Neyman-Pearson lemma (Li & Kuelbs, 1998), while in the following section another derivation using functional optimization is provided.

Note that the bound in Equation (2) is tractable since it only requires to evaluate the smoothed classifier $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)$ at the original image $\boldsymbol{x}_0$, instead of solving the difficult adversarial optimization over perturbation $\boldsymbol{z}$ in Equation (1). In practice, $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)$ is approximated by Monte Carlo approximation with a non-asymptotic confidence bound.

3.1 Certification via Functional Optimization

We propose a general framework for obtaining a guaranteed lower bound for Equation (1) based on functional optimization. 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,

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

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 f, and search for the minimum value of $f_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0 + \boldsymbol{\delta})$ for all classifiers in $\mathcal{F}$ and 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 Equation (3) computationally tractably. In this paper, we consider the set of all functions bounded in $[0, 1]$, that is,

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

which guarantees to include $f^\sharp$ by definition. There are other $\mathcal{F}$ that also yields computationally tractable bounds, including the L_p space $\mathcal{F} = \{f\colon \|f\|_{L_p} \leq v\}$, which we leave for future work.

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

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

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 dual form of $V_{\pi_{\mathbf{0}}}(\mathcal{F}, \mathcal{B})$ via strong duality:*

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

where we define

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

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 show later, the bound in (5) 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

$$\mathbb{D}_{\mathcal{F}_{[0,1]}} \left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}\right) = \int \left(\lambda\pi_{\mathbf{0}}(\boldsymbol{z}) - \pi_{\boldsymbol{\delta}}(\boldsymbol{z})\right)_+ d\boldsymbol{z}, \qquad \textit{where} \quad (t)_+ = \max(0, t).$$

In addition, we have $0 \leq \mathbb{D}_{\mathcal{F}_{[0,1]}} \left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}\right) \leq \lambda$ for any $\pi_{\mathbf{0}}$, $\pi_{\boldsymbol{\delta}}$ and $\lambda > 0$. Note that $\mathbb{D}_{\mathcal{F}_{[0,1]}} \left(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}\right)$ coincides with the total variation distance between $\pi_{\mathbf{0}}$ and $\pi_{\boldsymbol{\delta}}$ when $\lambda = 1$.

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

$$\begin{aligned}
V(\mathcal{F}, \mathcal{B}) =& \min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}} \max_{\lambda\geq 0} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})] + \lambda \left(f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})]\right) \\
\geq& \max_{\lambda\geq 0} \min_{\boldsymbol{\delta}\in\mathcal{B}} \min_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})] + \lambda \left(f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{E}_{\pi_{\mathbf{0}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})]\right) \quad \text{//exchange min and max} \\
=& \max_{\lambda\geq 0} \min_{\boldsymbol{\delta}\in\mathcal{B}} \left\{ \lambda f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) + \min_{f\in\mathcal{F}} \mathbb{E}_{\pi_{\boldsymbol{\delta}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})] - \lambda \mathbb{E}_{\pi_{\mathbf{0}}}[f(\boldsymbol{x}_0 + \boldsymbol{z})]) \right\} \\
=& \max_{\lambda\geq 0} \min_{\boldsymbol{\delta}\in\mathcal{B}} \left\{ \lambda f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}}) \right\}
\end{aligned}$$

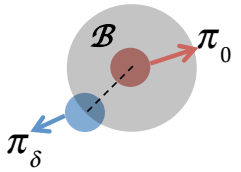
The proof of the strong duality is in Appendix A.1. II) follows a straightforward calculation. □

Although the lower bound in Equation (5) still involves an optimization on $\boldsymbol{\delta}$ and λ, both of them are much easier than the original adversarial optimization in Equation (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 a corollary of Theorem 1, we can exactly recover the bound derived by Cohen et al. (2019).

Corollary 1. *With isotropic Gaussian noise $\pi_{\mathbf{0}} = \mathcal{N}(\mathbf{0}, \sigma^2 I_{d\times d})$, ℓ_2 attack $\mathcal{B} = \{\boldsymbol{\delta}\colon \|\boldsymbol{\delta}\|_2 \leq r\}$ and $\mathcal{F} = \mathcal{F}_{[0,1]}$, the lower bound in Equation (5) equals the bound in Equation (2) by Cohen et al. (2019), that is,*

$$\max_{\lambda\geq 0} \left\{ \lambda f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) - \max_{\|\boldsymbol{\delta}\|_2\leq r} \mathbb{D}_{\mathcal{F}_{[0,1]}} \left(\lambda\pi_{\mathbf{0}} \| \pi_{\boldsymbol{\delta}}\right) \right\} = \Phi(\Phi^{-1}(f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)) - r/\sigma). \tag{6}$$

See Appendix A.2 for more details. A key step of the proof is to show that $\max_{\|\boldsymbol{\delta}\|_2 \le r} \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \| \pi_{\boldsymbol{\delta}})$ is achieved when $\boldsymbol{\delta}$ is on the boundary of the ℓ_2 ball $\mathcal{B} = \{\|\boldsymbol{\delta}\|_2 \le r\}$, which can be, for example, $\boldsymbol{\delta} = [r, 0, \ldots, 0]^\top$, due to symmetry of $\mathcal{B}$ (see figure on the right).



Trade-off between Accuracy and Robustness The lower bound in Equation (5) reflects an intuitive trade-off between the robustness and accuracy,

$$\max_{\lambda \ge 0} \Big[\underbrace{\lambda f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)}_{\text{Accuracy}} - \underbrace{\max_{\boldsymbol{\delta} \in \mathcal{B}} \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \,\|\, \pi_{\boldsymbol{\delta}})}_{\text{Robustness}} \Big], \tag{7}$$

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

More critically, different choices of smoothing distributions also yields a fundamental trade-off between accuracy and robustness in Equation (7). If $\pi_{\mathbf{0}}$ has large variance or high tail probability, the distance $\mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \,\|\, \pi_{\boldsymbol{\delta}})$ will tend to be large and the model is robust. However, if the variance of $\pi_{\mathbf{0}}$ is too large, we may obtain a low value of $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)$ and hence less accurate model. The optimal choice of the smoothing distribution should optimally balance the accuracy and robustness, by distribute its mass properly to yield small $\max_{\boldsymbol{\delta} \in \mathcal{B}} \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \,\|\, \pi_{\boldsymbol{\delta}})$ and large $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0)$ simultaneously.

4 Filling the Soap Bubbles: New Families of Non-Gaussian Smoothing Distributions

In this section, we identify a key problem of using Gaussian smoothing noise in high dimensional space, due to the "thin shell" phenomenon that the probability mass of Gaussian distributions concentrates on a sphere far away from the center points in high dimensional spaces. Motivated by this observation, we propose in Section 4.1 a new family of non-Gaussian smoothing distributions that alleviate this problem for ℓ_2 attack, and also in Section 4.2 another *mixed norm* smoothing distribution designed specifically for ℓ_∞ attack.

4.1 ℓ_2 Region Certification

Although isotropic Gaussian distributions appears to be a natural choice of smoothing distributions, they are in fact sub-optimal for trading-off accuracy and robustness in Equation (7), especially in high dimensions. The key problem is that, in high dimensional spaces, the probability mass of Gaussian distributions concentrates on a *thin shell* away from the center, and hence looks like "soap bubbles", instead of "solid balls" as what it appears in low dimension spaces.

Lemma 1 (Vershynin (2018), 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 $\delta \in (0, 1)$,*

$$\text{Prob}\left(\sqrt{d} - \sqrt{c\log(2/\delta)} \le \|\boldsymbol{z}\|_2 \le \sqrt{d} + \sqrt{c\log(2/\delta)}\right) \ge 1 - \delta.$$

This suggests that with high probability (e.g, $1-\delta = 0.99$), $\boldsymbol{z}$ takes values very close to the sphere of radius $\sqrt{d}$, within a constant distance from that sphere! See Vershynin (2018) for more discussion.

This phenomenon makes it problematic to use standard Gaussian distribution 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 close to the original classifier (and hence accurate). To illustrate the problem, consider a simple example when the true classifier is $f^\sharp(\boldsymbol{x}) = \mathbb{I}(\|\boldsymbol{x} - \boldsymbol{x}_0\|_2 \le \epsilon\sqrt{d})$ for a constant $\epsilon < 1$, where $\mathbb{I}$ is the indicator function. Then when the dimension d is large, we would have $f^\sharp(\boldsymbol{x}_0) = 1$ while $f^\sharp_{\pi_{\mathbf{0}}}(\boldsymbol{x}_0) \approx 0$ when $\pi_{\mathbf{0}} = \mathcal{N}(\mathbf{0}, I_{d\times d})$. It is of course possible to decrease the variance of $\pi_{\mathbf{0}}$ to improve the accuracy of the smoothed classifier $f^\sharp_{\pi_{\mathbf{0}}}$. However, this would significantly improve the distance term in Equation (7) and does not yield an optimal trade-off on accuracy and robustness.

In this work, we introduce a new family of non-Gaussian distributions to address this curse of dimensionality. To motivate our method, it is useful to examine the density function of the distributions of the radius of spherical distributions in general.

Lemma 2. *Assume $\boldsymbol{z}$ is a spherically symmetric random variable on $\mathbb{R}^d$ with a probability density function (PDF) of form $\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \phi(\|\boldsymbol{z}\|_2)$, where $\phi\colon [0,\infty) \to [0,\infty)$ is a univariate function, then the PDF of the norm of $\boldsymbol{z}$ is $p_{\|\boldsymbol{z}\|_2}(r) \propto r^{d-1}\phi(r)$. The term r^{d-1} arises due to the integration on the sphere of radius r in $\mathbb{R}^d$.*

In particular, when $\boldsymbol{z} \sim \pi_{\mathbf{0}} = \mathcal{N}(0, \sigma^2 I_{d\times d})$, we have $\phi(r) = \exp(-r^2/(2\sigma^2))$ and hence $p_{\|\boldsymbol{z}\|_2}(r) \propto r^{d-1}\exp(-r^2/(2\sigma^2))$, which is a scaled Chi distribution, also known as Nakagami distribution. Examining this P.D.F., we can see that the concentration of the norm 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 have a way to cancel out the effect of r^{d-1}. This motivates the following family of smoothing distributions:

$$\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_2^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right), \qquad \text{and hence} \qquad p_{\|\boldsymbol{z}\|_2}(r) \propto r^{d-k-1}\exp\left(-\frac{r^2}{2\sigma^2}\right), \tag{8}$$

where we introduce a $\|\boldsymbol{z}\|_2^{-k}$ term in $\pi_{\mathbf{0}}$, with k a positive parameter, to make the radius distribution less concentrated when k is large.

The radius distribution in Equation (8) is controlled by two parameters $(\sigma,\ k)$, where σ controls the scale of the distribution (and is hence the *scale parameter*), while k controls the shape of the distribution (and hence the *shape parameter*). The key idea is that adjusting k allows us to trade-off the accuracy and robustness much more optimally. As shown in Figure 1, adjusting σ enables us to move the mean close to zero (hence yielding higher accuracy), but at cost of decreasing the variance quadratically (hence more less robust). In contrast, adjusting k allows us to decrease the mean without significantly impacting the variance, and hence yield a much better trade-off on the accuracy and robustness.

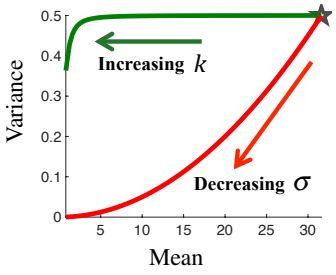


Figure 1: Starting from radius distribution in Equation (8) with $d = 100$ $\sigma = 1$ and $k = 0$ (black start), increasing k (green curve) allows us to move the mean towards zero *without significantly reducing the variance*. Decreasing σ (red curve) can also decrease the mean, but with a cost of decreasing the variance quadratically.

Computational Method With the more general non-Gaussian smoothing distribution, we no longer have the closed form solution of the bound like Equation (6). However, efficient computational methods can be still developed for calculating the bound in Equation (5) with $\pi_{\mathbf{0}}$ in Equation (8). The key is that the maximum of the distance term $\mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \,||\, \pi_{\boldsymbol{\delta}})$ over $\boldsymbol{\delta} \in \mathcal{B}$ is always achieved on the boundary of $\mathcal{B}$ as we show in the sequel, while the optimization on $\lambda \geq 0$ is one-dimensional and can be solved numerically efficiently.

Theorem 2. *Consider the ℓ_2 attack with $\boldsymbol{\mathcal{B}} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_2 \leq r\}$ 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 $k \geq 0$ and $\sigma > 0$. Define $\boldsymbol{\delta}^* = [r, 0, ..., 0]^\top$, we have*

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

With Theorem 2, we can compute Equation (5) with $\boldsymbol{\delta} = \boldsymbol{\delta}^*$. We then calculate $\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \,\|\, \pi_{\boldsymbol{\delta}^*}\right)$ using Monte Carlo approximation. Note that

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

which can be approximated with Monte Carlo method with Hoeffding concentration bound. Let $\{\boldsymbol{z}_i\}_{i=1}^n$ be an i.i.d. sample from $\pi_{\mathbf{0}}$, then we can approximate $\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \,\|\, \pi_{\boldsymbol{\delta}^*}\right)$ with $\hat{D} := n^{-1}\sum_{i=1}^n \left(\lambda - \pi_{\boldsymbol{\delta}^*}(\boldsymbol{z}_i)/\pi_{\mathbf{0}}(\boldsymbol{z}_i)\right)_+$. Because $0 \leq \left(\lambda - \pi_{\boldsymbol{\delta}^*}(\boldsymbol{z}_i)/\pi_{\mathbf{0}}(\boldsymbol{z}_i)\right)_+ \leq \lambda$, we have $\mathbb{D}_{\mathcal{F}_{[0,1]}}\left(\lambda\pi_{\mathbf{0}} \,\|\, \pi_{\boldsymbol{\delta}^*}\right) \in [\hat{D} - \lambda\sqrt{\log(2/\delta)/(2n)}, \hat{D} + \lambda\sqrt{\log(2/\delta)/(2n)}]$ with probability $1-\delta$ for $\delta \in (0,1)$. Drawing sufficiently large number of samples allows us to achieve approximation with arbitrary accuracy.

4.2 ℓ_∞ REGION CERTIFICATION

Going beyond the ℓ_2 attack, we consider the ℓ_∞ attack, whose attacking region is $\mathcal{B}_{\ell_\infty,r} = \{\boldsymbol{\delta}\colon \|\boldsymbol{\delta}\|_\infty \le r\}$. This is a far more difficult problem, because the ℓ_∞ ball is substantially larger than the ℓ_2 ball of the same radius in high dimensional space. This makes the Gaussian smoothing distribution, as well as our ℓ_2-based smoothing distribution in Equation (8), unsuitable for ℓ_∞ attack. In fact, as shown in the following negative result, if we use Equation (8) as the smoothing distribution for ℓ_∞ attack, the bound we obtain is effectively the bound we would get for verifying a ℓ_2 ball with radius $\sqrt{d}r$, which is too large to give meaningful results when the dimension is high.

Theorem 3. *With the smoothing distribution $\pi_{\mathbf{0}}$ in Equation (8) for $k \ge 0, \sigma > 0$, and $\mathcal{F} = \mathcal{F}_{[0,1]}$ shown in Equation (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,*

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

The key reason of this negative result is that the furthest points to the origin (vertexes) in $\mathcal{B}_{\ell_\infty,r}$ have an ℓ_2 radius of $\sqrt{d}r$, illustrated as the "pointy" points in Figure 2. Thus, the maximum distance $\max_{\boldsymbol{\delta}\in\mathcal{B}_{\ell_\infty,r}} \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$ is achieved at one of these pointy points, making it equivalent to optimizing in the ℓ_2 ball with radius $\sqrt{d}r$.

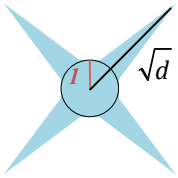


Figure 2: ℓ_∞ and ℓ_2 balls in high dimension.

In order to address this problem, we propose the following new *mixed norm* family of smoothing distribution that uses a mix of ℓ_2 and ℓ_∞ norms:

$$\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_\infty^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_2^2}{2\sigma^2}\right), \tag{9}$$

in which we replace the $\|\boldsymbol{z}\|_2^{-k}$ term in Equation (8) with $\|\boldsymbol{z}\|_\infty^{-k}$. The motivation is that this allows us to allocate more probability mass along the "pointy" directions with larger ℓ_∞ norm, and hence decrease the maximum distance term $\max_{\boldsymbol{\delta}\in\mathcal{B}_{\ell_\infty,r}} \mathbb{D}_{\mathcal{F}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$. See Figure 2 for an illustration. In practice, we find that this mixed norm smoothing distribution in Equation (9) work much more efficiently than the ℓ_2 norm-based family in Equation (8).

Given the difference of ℓ_2 and ℓ_∞ norms, it is also natural to consider the following *pure ℓ_∞ norm* distributions, which uses ℓ_∞ norm in both of the terms of the distribution,

$$\pi_{\mathbf{0}}(\boldsymbol{z}) \propto \|\boldsymbol{z}\|_\infty^{-k} \exp\left(-\frac{\|\boldsymbol{z}\|_\infty^2}{2\sigma^2}\right). \tag{10}$$

Unfortunately, this seemingly natural choice does not work efficiently for ℓ_∞ attacks (even worse than the ℓ_2 family Equation (8)). This is because the volume of the ℓ_∞ ball is in some sense "too large" (e.g., compared with the volume of ℓ_2 ball). As a result, in order to make the probability mass of Equation (10) in a reasonable scale, one has to choose a very small value of σ, which makes maximum distance term too large to be practically useful.

Theorem 4. *Consider the adversarial attacks on the ℓ_∞ ball $\mathcal{B}_{\ell_\infty,r} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_\infty \le r\}$. Suppose we use the smoothing distribution $\pi_{\mathbf{0}}$ in Equation (10) 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 $V_{\pi_{\mathbf{0}}}(\mathcal{F}_{[0,1]},\ \mathcal{B}_{\ell_\infty,r}) = \mathcal{O}(t^{-d})$ and thus fail to certify.

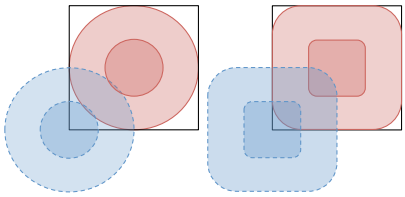


Figure 3: For ℓ_∞ attacking, the mixed norm distribution (right) yields smaller TV distances (larger overlap areas), and hence higher robustness.

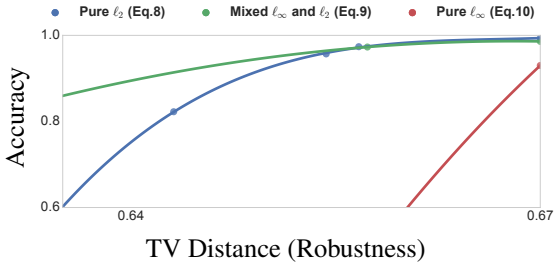


Figure 4: The Pareto frontier of accuracy and robustness (in the sense of Equation (7)) of the three smoothing families in Equation (8), Equation (9), and Equation (10) for ℓ_∞ attacking, when we search for the best parameters (k, σ) for each of them. The mixed norm family Equation (9) yields the best 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).

Remark The condition 1) and 2) in Theorem 4 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 (7). 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 4 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.

Empirical Justification We construct a simple toy example to verify the advantages of the mixed norm family Equation (9) overall the ℓ_2 family in Equation (8) and the ℓ_∞ family in Equation (10). We assume that the true classifier is $f^\sharp(\boldsymbol{x}) = \mathbb{I}(\|x\|_2 \leq r)$ in $r = 0.65$, $d = 5$ case and plot in Figure 4 the Pareto frontier of the accuracy and robustness terms in Equation (7) for the three families of smoothing distributions, as we search for the best combinations of parameters (k, σ). We can see that the mixed norm smoothing distribution clearly obtain the best trade-off on accuracy and robustness, and hence guarantees a tighter lower bound for certification.

Computational Method In order to compute the lower bound when using the mixed norm family Equation (9), we need to establish the closed form solution of the maximum distance term $\max_{\boldsymbol{\delta}\in\mathcal{B}} \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \,||\, \pi_{\boldsymbol{\delta}})$ similar to Theorem 2. The following result shows that the optimal $\boldsymbol{\delta}$ is achieved at one vertex (the pointy points) of the ℓ_∞ ball.

Theorem 5. *Consider the ℓ_∞ attack with $\mathcal{B}_{\ell_\infty, r} = \{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_\infty \leq r\}$ and the mixed norm smoothing distribution in Equation (9) with $k \geq 0$ and $\sigma > 0$. Define $\boldsymbol{\delta}^* = [r, r, ..., r]^\top$. We have for any $\lambda > 0$,*

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

The proof of Theorem 2 and 5 is non-trivial, thus we defer the details to Appendix A.3. With the optimal $\boldsymbol{\delta}^*$ found above, we can calculate the bound with similar Monte Carlo approximation outlined in Section 4.1.

5 Experiments

We evaluate our new bound and smoothing distributions for both ℓ_2 and ℓ_∞ attacks. We compare with the randomized smoothing method of Cohen et al. (2019) with Gaussian smoothing distribution. For fair comparisons, we use the same model architecture and pre-trained models provided by Cohen et al. (2019) and Salman et al. (2019), which are ResNet110 on CIFAR10 and ResNet50 on ImageNet.

ℓ_2 RADIUS (CIFAR-10)	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25
Cohen et al. (2019) (%)	60	43	34	23	17	14	12	10	8
OURS (%)	**61**	**46**	**37**	**25**	**19**	**16**	**14**	**11**	**9**

Table 1: Certified top-1 accuracy of the best classifiers with various ℓ_2 radius on CIFAR-10.

ℓ_2 RADIUS (ImageNet)	0.5	1.0	1.5	1.0	2.0	2.5	3.0
Cohen et al. (2019) (%)	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 on ImageNet.

Settings and Hyperparameters The details of our method are shown in Algorithm 2 in Appendix. 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 (7). The optimization on λ is solved using grid search. For ℓ_2 attacks, we set $k = 500$ for CIFAR10 and $k = 50000$ for ImageNet in our non-Gaussian smoothing distribution Equation (8). 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 (8) matches with the norm of $\mathcal{N}(0, \sigma_0^2)$. 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 (9). In both cases, the baseline algorithm uses a Gaussian smoothing distribution $\mathcal{N}(0, \sigma_0^2)$. More ablation study about k is deferred to Appendix C.

Evaluation Metrics The methods are evaluated using the certified accuracy defined in Cohen et al. (2019). Given an input image $\boldsymbol{x}$ and a perturbation region $\mathcal{B}$, the smoothed classifier is called certified correct if its prediction is correct and has a guaranteed lower bound larger than $1/2$ for $\boldsymbol{\delta} \in \mathcal{B}$. The certified accuracy is the percentage of images that are certified correct. Following Salman et al. (2019), we calculate the certified accuracy of all the classifiers in Cohen et al. (2019) or Salman et al. (2019) for various radius, and report the best results over all of classifiers.

5.1 ℓ_2 CERTIFICATION

We test our method on CIFAR10 and ImageNet for ℓ_2 certification. For fair comparison, we use the same pre-trained models as Cohen et al. (2019), which is trained with Gaussian noise on both CIFAR10 and ImageNet dataset. The readers are referred to Appendix C for detailed ablation studies.

Table 1 and Table 2 report the certified accuracy of our method with the non-Gaussian smoothing distribution in Equation (8) and the baseline on CIFAR10 and ImageNet, respectively. We find that our method consistently outperforms the baseline.

5.2 ℓ_∞ CERTIFICATION

We test our lower bound based on the mixed norm family in Equation (9) for verifying ℓ_∞ attacking on CIFAR10, using the models trained by Salman et al. (2019). The certified accuracy of our method and the baseline using Gaussian smoothing distribution are shown in Table 3. We can see that our method consistently outperforms the Gaussian distribution baseline by a large margin, which empirically shows our distribution is a more suitable distribution for ℓ_∞ perturbation.

To further confirm the advantage of our method, we plot in Figure 5 the certified accuracy of our method and Gaussian baseline using models trained with Gaussian perturbation of different variances σ_0, under different ℓ_∞ radius. We again find that our approach outperforms the 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. Future work will investigate how to obtain non-trivial bounds for ℓ_∞ attacking at ImageNet scales with smoothing classifiers.

l_∞ RADIUS (CIFAR-10)	$2/255$	$4/255$	$6/255$	$8/255$	$10/255$	$12/255$
Salman et al. (2019) (%)	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.

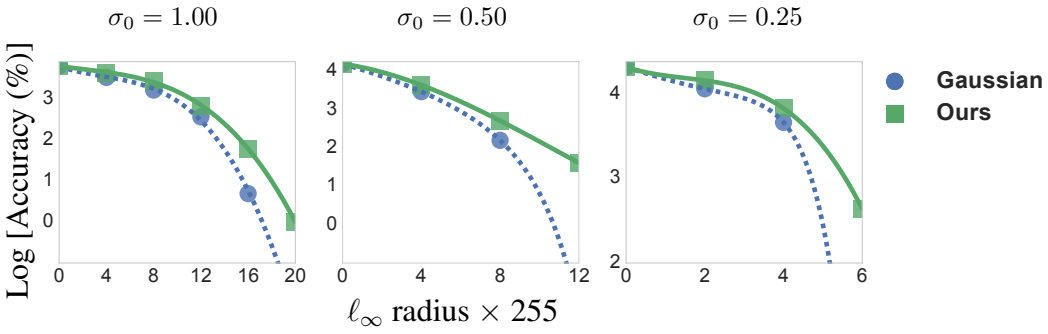


Figure 5: Results of ℓ_∞ verification on CIFAR10, on models trained with Gaussian noise data augmentation with different variances σ_0. Our method obtains consistently better results.

6 Conclusions

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 two new families of non-Gaussian smoothing distributions, which significantly outperform the Gaussian-based smoothing for ℓ_2 and ℓ_∞ attacking, respectively. Our work provides 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.

References

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

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. ISBN 0521833787.

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

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*, pp. 251–268, 2017.

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

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*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. URL `https://www.aclweb.org/anthology/N19-1423/`.

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

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*, pp. 550–559, 2018. URL http://auai.org/uai2018/proceedings/papers/204.pdf.

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

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

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.

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

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.

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.

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

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

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)*, pp. 369–385, 2018.

Yiping Lu, Aoxiao Zhong, Quanzheng Li, and Bin Dong. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. 2018.

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

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

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.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

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

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

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

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

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.

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.

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

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*, pp. 4939–4948, 2018.

A PROOFS

A.1 PROOF FOR THE STRONG DUALITY IN THEOREM 1

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

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

and the dual problem is

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

Our strategy is to verify that the solution of the dual problem is also a solution of the primal problem. Notice that for both primal and dual problem, it is equivalent to change the maximum over λ from $\max_{\lambda\in\mathbb{R}}$ to $\max_{\lambda\geq 0}$. Suppose that $(\boldsymbol{\delta}^*, f^*, \lambda^*)$ is the solution of the dual problem. Our proof is by verifying $(\boldsymbol{\delta}^*, f^*, \lambda^*) = \left(\bar{\boldsymbol{\delta}}, \bar{f}, \bar{\lambda}\right)$, where $\left(\bar{\boldsymbol{\delta}}, \bar{f}, \bar{\lambda}\right)$ is some optimal solution of primal problem.

Claim 1 $\bar{f}$ satisfies the constraint condition, i.e, $\mathbb{E}_{\pi_{\mathbf{0}}}\bar{f} = f^{\sharp}_{\pi_{\mathbf{0}}}$.

Proof of Claim 1: Otherwise, for example $\mathbb{E}_{\pi_{\mathbf{0}}}\bar{f} < p_0$, by letting $\lambda\to\infty$, the minimization problem has optimal value $+\infty$. However, by letting $\bar{f}$ be the classifier $f^{\sharp}$ we use, we achieve a better solution, which makes contradiction.

Claim 2 $\left(\bar{\boldsymbol{\delta}}, \bar{f}, \lambda^*\right)$ is the optimal solution for the primal problem.

Proof of Claim 2: This is obvious as

$$\begin{aligned}\mathbb{E}_{\pi_{\bar{\boldsymbol{\delta}}}}\bar{f}+\bar{\lambda}\left(f^{\sharp}_{\pi_{\mathbf{0}}}-\mathbb{E}_{\pi_{\mathbf{0}}}\bar{f}\right) &= \mathbb{E}_{\pi_{\bar{\boldsymbol{\delta}}}}\bar{f}\\ &= \mathbb{E}_{\pi_{\bar{\boldsymbol{\delta}}}}\bar{f}+\lambda^*\left(f^{\sharp}_{\pi_{\mathbf{0}}}-\mathbb{E}_{\pi_{\mathbf{0}}}\bar{f}\right).\end{aligned}$$

Claim 3 Define a new constraint optimization problem Q:

$$\min_{\boldsymbol{\delta}\in\mathcal{B}}\min_{f\in\mathcal{F}}\mathbb{E}_{\pi_{\boldsymbol{\delta}}}f+\lambda^*\left(f^{\sharp}_{\pi_{\mathbf{0}}}-\mathbb{E}_{\pi_{\mathbf{0}}}f\right).$$

And both $(\boldsymbol{\delta}^*, f^*)$ and $\left(\bar{\boldsymbol{\delta}}, \bar{f}\right)$ is the optimal solution for the problem Q.

Proof of Claim 3: We first prove that $\left(\bar{\boldsymbol{\delta}}, \bar{f}\right)$ is the optimal solution for the problem Q. If not, then there exists $(', f')$ such that

$$\begin{aligned}&\mathbb{E}_{\pi_{\boldsymbol{\delta}'}}f'+\lambda^*\left(f^{\sharp}_{\pi_{\mathbf{0}}}-\mathbb{E}_{\pi_{\mathbf{0}}}f'\right)\\ <&\mathbb{E}_{\pi_{\bar{\boldsymbol{\delta}}}}\bar{f}+\lambda^*\left(f^{\sharp}_{\pi_{\mathbf{0}}}-\mathbb{E}_{\pi_{\mathbf{0}}}\bar{f}\right).\end{aligned}$$

Then $(\boldsymbol{\delta}', f', \lambda^*)$ becomes a better solution for the primal problem, which makes contradiction. And by definition of $(\boldsymbol{\delta}^*, f^*, \lambda^*)$, $(\boldsymbol{\delta}^*, f^*)$ is the minimizer of problem Q.

Claim 4 Define

$$L[\boldsymbol{\delta}, f, \lambda] = \mathbb{E}_{\pi_{\boldsymbol{\delta}}}f+\lambda\left(-\mathbb{E}_{\pi_{\mathbf{0}}}f\right).$$

By claim 1 and 3, we have $L[\bar{\boldsymbol{\delta}}, \bar{f}, \bar{\lambda}] = L[\bar{\boldsymbol{\delta}}, \bar{f}, \lambda^*] = L[\boldsymbol{\delta}^*, f^*, \lambda^*]$, which implies that $(\boldsymbol{\delta}^*, f^*, \lambda^*)$ is the optimal solution of primal solution.

A.2 PROOF FOR COROLLARY 1

Proof. Given our confidence lower bound

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

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

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

Define

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

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 Boyd & Vandenberghe (2004) 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

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

This tells us

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

, i.e. the certification radius is $\sigma\cdot\Phi^{-1}(p_0)$. This is exactly the core theoretical contribution of Cohen et al. (2019). This bound has a straight forward expansion for multi-class classification situations, we refer interesting readers to Appendix D. □

A.3 Proof For Theorem 2 and 5

In this subsection, we give the proof of theorem 2 and 5. 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 $k_1, k_2 \geq 0$ 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 6. *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 $k_1, k_2 \geq 0$ and $\sigma > 0$, for any $\lambda \geq 0$ we have*

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

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

Theorem 2 and 5 directly follows the above theorem. Notice that in Theorem 2, as our distribution is spherical symmetry, it is equivalent to set $\boldsymbol{\mathcal{B}} = \left\{\boldsymbol{\delta} : \boldsymbol{\delta} = [a, 0, ..., 0]^\top, a \leq r\right\}$ 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 $\delta_i \geq 0$, $i \in [d]$. Notice that

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

And also

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

We thus have

$$\begin{aligned}&\frac{\partial}{\partial \delta_1} \int \left(\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) - \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right)_+ d\boldsymbol{x}_0 \\ =& - \int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right\} \frac{\partial}{\partial \delta_1} \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0) d\boldsymbol{x}_0 \\ =& \int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0)\right\} F_1\left(\left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_\infty, \left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_2\right) d\boldsymbol{x}_0 \\ =& \int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > \mu_1\right\} F_1\left(\left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_\infty, \left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_2\right) d\boldsymbol{x}_0 \\ &+ \int \mathbb{I}\left\{\lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 < \mu_1\right\} F_1\left(\left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_\infty, \left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_2\right) d\boldsymbol{x}_0,\end{aligned}$$

where we define

$$\begin{aligned}&F_1\left(\left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_\infty, \left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_2\right) \\ &= \phi_1'\left(\left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_\infty\right) \phi_2\left(\left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_2\right) \mathbb{I}\{\left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_\infty = |x_1 - \delta_1|\} \frac{(x_1 - \delta_1)}{\left\| \boldsymbol{x}_0 - \boldsymbol{\delta} \right\|_\infty} \\ &+ \phi_1\left(\left\| \boldsymbol{x}_0 - \boldsymbol{\mu} \right\|_\infty\right) \phi_2'\left(\left\| \boldsymbol{x}_0 - \boldsymbol{\mu} \right\|_2\right) \frac{(x_1 - \mu_1)}{\left\| \boldsymbol{x}_0 - \boldsymbol{\mu} \right\|_2}.\end{aligned}$$

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

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

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

$$\begin{aligned}H_1 &= \{\boldsymbol{x}_0 : \lambda \pi_{\mathbf{0}}(\boldsymbol{x}_0) \geq \pi_{\boldsymbol{\delta}}(\boldsymbol{x}_0), x_1 > \mu_1\} \\ H_2 &= \left\{[2\delta_1 - x_1, x_2, ..., x_d]^\top : \boldsymbol{x}_0 = [x_1, ..., x_d]^\top \in H_1\right\}.\end{aligned}$$

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

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

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

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

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

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

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

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

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

A.3.1 PROOF FOR THEOREM 4

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

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

Without loss of generality, we assume $\boldsymbol{\delta}^* = [r, ..., r]^\top$. (see Theorem 6)

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

Notice that as the distribution is symmetry,

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

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 $\epsilon \in (0, 1)$,

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

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

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

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

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

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 PRACTICAL ALGORITHM

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

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

given some classifier $f^\sharp$. Following Cohen et al. (2019), 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 $\mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$. A lower bound $\hat{p}_0$ of the former term is obtained through binominal test as Cohen et al. (2019) 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 $\lambda f^\sharp_{\pi_{\mathbf{0}}} - \mathbb{D}_{\mathcal{F}_{[0,1]}}(\lambda\pi_{\mathbf{0}} \parallel \pi_{\boldsymbol{\delta}})$. This can be achieved by the following importance sampling manner:

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

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 λ : Λ =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 Cohen et al. (2019). 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 4.1 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.

C 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.

D ILLUMINATION ABOUT BILATERAL CONDITION

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.,

$$f_A^\sharp(\boldsymbol{x}_0) > \max_{B \neq A} f_B^\sharp(\boldsymbol{x}_0) \tag{11}$$

ℓ_2 RADIUS (CIFAR-10)	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25
Cohen et al. (2019) (%)	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 at various ℓ_2 radius. We use the same model as Cohen et al. (2019) and do not train any new models.

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

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

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 Cohen et al. (2019)'s setting, this condition is equivalent to:

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

with a similar derivation process like Appendix A.2. This is exactly the same bound in the (restated) theorem 1 of Cohen et al. (2019).

Cohen et al. (2019) 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 (Cohen et al., 2019; Salman et al., 2019) do not take the bilateral form, we also do not use this form in experiments for fairness.