Extensions and limitations of randomized smoothing for robustness guarantees

Jamie Hayes
University College London
j.hayes@cs.ucl.ac.uk

Abstract

Randomized smoothing, a method to certify a classifier's decision on an input is invariant under adversarial noise, offers attractive advantages over other certification methods. It operates in a black-box and so certification is not constrained by the size of the classifier's architecture. Here, we extend the work of Li et al. [26], studying how the choice of divergence between smoothing measures affects the final robustness guarantee, and how the choice of smoothing measure itself can lead to guarantees in differing threat models. To this end, we develop a method to certify robustness against any ℓ_p ($p \in \mathbb{N}_{>0}$) minimized adversarial perturbation. We then demonstrate a negative result, that randomized smoothing suffers from the curse of dimensionality; as p increases, the effective radius around an input one can certify vanishes.

1. Introduction

Image classification is vulnerable to *adversarial examples*. Given an image classifier $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$ such that the decision function $F = \arg\max_i f_i$ classifies an input, x, correctly as $F(x) = y$, an adversarial example is an input, $x+\delta$, such that $F(x+\delta) \neq y$ where x and $x+\delta$ are assigned the same label by an oracle classifier, $\mathcal{O}$, which is usually taken to be the human vision system. To preserve oracle classification, it is common to minimize the perturbation, δ, with respect to an ℓ_p norm. Constructing a perturbation such that $\|\delta\|_p \ll \|x\|_p$, will result in an input such that $\|x + \delta\|_p \approx \|x\|_p$. With high likelihood x and $x + \delta$ will be visually similar and $\mathcal{O}$ will classify both correctly.

The vulnerability to adversarial examples requires a suitable defense. Many empirical defenses have been proposed and subsequently shown to be broken, implying more theoretically grounded techniques to measure robustness are required [1, 6, 7, 16, 34]. Recently, methods from verification literature have been used to provide guarantees of an inputs robustness to adversarial perturbations. These methods seek the minimum or a lower bound on the amount of noise required to cause a misclassification. These verification methods are most often tailored to a single ℓ_p norm for which the defense guarantees robustness. A number of defenses *certify* a neural network is robust to adversarial examples by propagating upper and lower input bounds throughout the network or by bounding the Lipschitz value of the network [4, 12, 17, 18, 27, 29, 33, 37].

Recently, *randomized smoothing* has been proposed to certify image classifiers to ℓ_0, ℓ_1, and ℓ_2 perturbations [10, 24, 25, 26]. By constructing a classifier that outputs a label based on a majority vote under repeated addition of Laplacian or Gaussian noise, Lecuyer *et al*. [24] found lower bounds to the amount of noise required for misclassification of an input in the ℓ_1 or ℓ_2 norm, respectively. Following this, Li *et al*. [26] and Cohen *et al*. [10] provided improved bounds in the ℓ_2 norm. As explained by Cohen *et al*. [10], randomized smoothing has attractive advantages over other certification methods: it is scalable to large classifiers and makes no assumption about the architecture. In this work, we extend the general framework for randomized smoothing as proposed by Li *et al*. [26]. Firstly, we study how the choice of divergence between inputs smoothed with noise affects the final certificate, and secondly, we study how the choice of smoothing measure itself can lead to guarantees for differing threat models. Concretely, we show how the choice of smoothing measure allows us to extend randomized smoothing to any ℓ_p norm ($p \in \mathbb{N}_{>0}$), showing we can certify inputs with non-vacuous bounds over a range of ℓ_p norms with small p values. We then show that randomized smoothing fails to certify meaningfully large radii around inputs as p increases.

2. Certified defenses

In this section, we discuss related work on certified defenses to adversarial examples, introduce extensions to randomized smoothing approaches to certified defenses, and provide a method to compute a certified robust area around an input under *any* ℓ_p norm attack, where $p \in \mathbb{N}_{>0}$.

2.1. Background on certified defenses

The vulnerability of empirical defenses to adversarial examples has driven the need for formal guarantees of robustness. We define *certified robustness* as a guarantee that the decision of a classifier is preserved within an ϵ-ball around an input, and we refer to size of this ϵ-ball as the *certified radius*. Formal methods can be separated into *complete* and *incomplete* methods. Complete methods such as Satisfiability Modulo Theory (SMT) [8, 15, 20] or Mixed-Integer Programming (MIP) [5, 9, 35] provide exact robustness bounds but are expensive to implement. Incomplete methods solve a convex relaxation of the verification problem. The bounds given by incomplete methods can be loose but are quicker to find than exact bounds [4, 12, 17, 18, 27, 29, 37].

Lecuyer *et al*. [24] developed the certification technique, referred to as *randomized smoothing*, by noticing a connection between differential privacy [14] and robustness, and show that robustness can be proven under concentration measures of classification under noise. This work was expanded upon by Lee *et al*. [25], Li *et al*. [26], and Cohen *et al*. [10], who found improved robustness guarantees in the ℓ_0, ℓ_1, and ℓ_2 norms, respectively. Similarly to this work, Dvijotham *et al*. [13] developed a general framework for randomized smoothing that can handle arbitrary smoothing measures and so find robustness guarantees in any ℓ_p norm. In concurrent work, Blum *et al*. [3], Kumar *et al*. [23], and Yang *et al*. [36] also show that randomized smoothing may be unable to find robustness guarantees in the ℓ_∞ norm. Most related to this work are the findings of Kumar *et al*. [23], who also use a generalized Gaussian distribution for smoothing and show that the certified radius in an ℓ_p norm decreases as $\mathcal{O}(1/d^{\frac{1}{2}-\frac{1}{p}})$, where d is the dimensionality of the data.

2.2. Certification via randomized smoothing

Here, we expand on how robustness guarantees can be found through randomized smoothing.

Problem statment. Given an input $x \in \mathcal{X}$ such that $\arg\max_i f_i(x) = y$, find the maximum ϵ such that $\forall x' \in \mathcal{X}$, $d(x, x') < \epsilon \implies \arg\max_i f_i(x') = y$, given a distance function $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$.

This can be cast as an optimization problem, given by

$$\begin{aligned} &\max_{x' \in \mathcal{X}} d(x, x') \\ &\text{subject to } \arg\max_i f_i(x') = y \end{aligned} \tag{1}$$

In general, solving the above formulation is difficult, however randomized smoothing, introduced by Lecuyer *et al*. [24], can be used to solve a relaxed version of this problem. Namely, the aim is to solve

$$\begin{aligned} &\max_{x' \in \mathcal{X}} d(x + \theta, x' + \theta) \\ &\text{subject to } \mathbb{E}[\arg\max_i f_i(x' + \theta)] = y, \end{aligned} \tag{2}$$

where θ is a sample from a smoothing measure, μ, and d is now taken to be a suitable divergence or distance measure between random variables. For example, Li *et al*. [26] take μ to be the centered Gaussian, $\mathcal{N}(0, \sigma^2)$. Since Gaussians belong to the location-scale family of distributions, we can treat x and x' as constants and so, $x + \theta$ and $x' + \theta$ can be treated as random variables from distributions $\mathcal{N}(x, \sigma^2)$ and $\mathcal{N}(x', \sigma^2)$, respectively. We can use well known properties of divergences of Gaussians to represent $d(x + \theta, x' + \theta)$ in terms of the ℓ_2 norm difference of their means. Specifically, $d(x + \theta, x' + \theta)$ can be represented as a function of $\|x - x'\|_2$ and σ, for common divergences such as the Rényi and KL divergences. However, we must still solve the problem of ensuring $\mathbb{E}[\arg\max_i f_i(x' + \theta)] = y$. Given a chosen divergence, Li *et al*. [26] approach this problem by finding a lower bound between two multinomial distributions, P and Q, in terms of the two largest probabilities of P, when $\arg\max_i P_i \neq \arg\max_i Q_i$. This shows that any distribution, Q, for which P and Q agree on the index of the top probability, the divergence between P and Q must be smaller than this lower bound. We denote this lower bound by $h(p_1, p_2)$, where p_1, p_2 represent the top two probabilities from P. Given this lower bound Li *et al*. [26], solve the following problem

$$\begin{aligned} &\max_{x' \in \mathcal{X}} d(f(x + \theta), f(x' + \theta)) \\ &\text{subject to } d(f(x + \theta), f(x' + \theta)) \leq h(p_1, p_2) \end{aligned} \tag{3}$$

This can be efficiently solved by finding an upper bound to the Lagrangian relaxed problem

$$\begin{aligned} \max_{\lambda \leq 0, x' \in \mathcal{X}} & d(f(x + \theta), f(x' + \theta)) \\ & + \lambda(h(p_1, p_2) - d(f(x + \theta), f(x' + \theta))) \end{aligned} \tag{4}$$

$$= \max_{\lambda \leq 0, x' \in \mathcal{X}} (1 - \lambda) d(f(x + \theta), f(x' + \theta)) + \lambda h(p_1, p_2) \tag{5}$$

$$= \max_{\lambda \geq 0, x' \in \mathcal{X}} (1 + \lambda) d(f(x + \theta), f(x' + \theta)) - \lambda h(p_1, p_2) \tag{6}$$

$$\leq \max_{\lambda \geq 0, x' \in \mathcal{X}} (1 + \lambda) d(x + \theta, x' + \theta) - \lambda h(p_1, p_2) \tag{7}$$

$$= \max_{\lambda \geq 0, x' \in \mathcal{X}} (1 + \lambda) g(\|x - x'\|_2, \sigma) - \lambda h(p_1, p_2), \tag{8}$$

where in eq. (7), we use the data processing inequality property of divergences, and in eq. (8), we use the fact that for many common divergences, we can represent the divergence between two Gaussians as a function of the ℓ_2 norm of their means and their standard deviation, which we denote by $g(\|x - x'\|_2, \sigma)$.

By choosing $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ to be the Rényi divergence,

Table 1: ℓ_2 certified radius when using different divergences.

Distance	$d(Q,P) \geq$ (when $\arg\max_i q_i \neq \arg\max_i p_i$)	$d(\mathcal{N}(x,\sigma^2),\mathcal{N}(x',\sigma^2))$	Certified radius (for $\|x-x'\|_2 < \epsilon$)
$d_{KL}(Q,P) = \sum_{i=1}^k q_i \log \frac{q_i}{p_i}$	$-\log(2\sqrt{p_1 p_2} + 1 - p_1 - p_2)$	$\frac{1}{\sigma^2}\|x-x'\|_2^2$	$\sqrt{-\sigma^2 \log(2\sqrt{p_1p_2}+1-p_1-p_2)}$
$d_{H^2}(Q,P) = \frac{1}{2}\sum_{i=1}^k (\sqrt{q_i}-\sqrt{p_i})^2$	$1-\sqrt{1-\frac{(\sqrt{p_1}-\sqrt{p_2})^2}{2}}$	$1-e^{-\frac{\|x-x'\|_2^2}{8\sigma^2}}$	$\sqrt{-8\sigma^2 \log(\sqrt{1-\frac{(\sqrt{p_1}-\sqrt{p_2})^2}{2}})}$
$d_{\chi^2}(Q,P) = \sum_{i=1}^k \frac{(q_i-p_i)^2}{p_i}$	$\frac{(p_1-p_2)^2}{(p_1+p_2)-(p_1-p_2)^2}$	$e^{\frac{\|x-x'\|_2^2}{\sigma^2}}-1$	$\sqrt{\sigma^2 \log(\frac{p_1+p_2}{(p_1+p_2)-(p_1-p_2)^2})}$
$d_B(Q,P) = -\log(\sum_{i=1}^k \sqrt{q_i p_i})$	$-\log\left(\frac{(\sqrt{p_1}+\sqrt{p_2})^2+2(1-p_1-p_2)}{\sqrt{2(2\sqrt{p_1p_2}+2-p_1-p_2)}}\right)$	$\frac{1}{8\sigma^2}\|x-x'\|_2^2$	$\sqrt{-8\sigma^2 \log\left(\frac{(\sqrt{p_1}+\sqrt{p_2})^2+2(1-p_1-p_2)}{\sqrt{2(2\sqrt{p_1p_2}+2-p_1-p_2)}}\right)}$
$d_{TV}(Q,P) = \frac{1}{2}\sum_{i=1}^k \lvert q_i - p_i\rvert$	$\frac{\lvert p_1-p_2\rvert}{2}$	$2\Phi(\frac{\|x-x'\|_2}{2\sigma})-1$	$2\sigma\Phi^{-1}(\frac{\lvert p_1-p_2\rvert}{2}+\frac{1}{2})$

we recover the results of Li *et al*. [26] with

$$g(\|x-x'\|_2, \sigma) = \frac{\alpha\|x-x'\|_2^2}{2\sigma^2} \tag{9}$$

$$h(p_1,p_2) = -\log\left(1-p_1-p_2+2\big(\frac{1}{2}(p_1^{1-\alpha}+p_2^{1-\alpha})\big)^{\frac{1}{1-\alpha}}\right) \tag{10}$$

Thus, for any $x' \in \mathcal{X}$ with $\|x-x'\|_2 < \epsilon$ we can guarantee the classifier, f, will not change it's decision for any ϵ smaller than

$$\max_{\lambda\geq 0}\left(\sup_{\alpha>1}\left(-\frac{\lambda 2\sigma^2}{(1+\lambda)\alpha}\log\Big(1-p_1-p_2+ 2\big(\frac{1}{2}(p_1^{1-\alpha}+p_2^{1-\alpha})\big)^{\frac{1}{1-\alpha}}\Big)\right)\right)^{\frac{1}{2}} \tag{11}$$

$$= \left(\sup_{\alpha>1}\left(-\frac{2\sigma^2}{\alpha}\log\Big(1-p_1-p_2+ 2\big(\frac{1}{2}(p_1^{1-\alpha}+p_2^{1-\alpha})\big)^{\frac{1}{1-\alpha}}\Big)\right)\right)^{\frac{1}{2}} \tag{12}$$

Clearly, this framework for certifying inputs is general and extends to different choices of divergence. In the next section, we explore divergences beyond Rényi divergence and show this choice affects the certified radius, given a Gaussian smoothing measure.

2.3. Certification guarantees against ℓ_2 perturbations for common divergences

Li *et al*. [26] show that, given two distributions, P and Q, with different indexes for the top probability, a lower bound of the Rényi divergence (denoted by d_α) is given by eq. (10). We extend this line of reasoning to find lower bounds for the KL divergence (d_{KL}), Hellinger distance (d_{H^2}), (Neyman) chi-squared distance (d_{χ^2}), Bhattacharyya distance (d_B), and total variation distance (d_{TV}). Proofs of these lower bounds are given in appendix A. To find a certified radius of a classifier's decision around an input, we find the distances between Gaussian measures with respect to each of these divergences. These are both represented in table 1 along with the certification guarantee in the ℓ_2 norm. We visualize the trade-off in certified radius around an input in fig. 1 for a hypothetical binary classification task as a function of the classifier's top output probability, p_1. As well as including the certified radii derived from the aforementioned divergences, we include the certified radii for the ℓ_2 norm found by Lecuyer *et al*. [24] and Cohen *et al*. [10] approaches. Lecuyer *et al*. [24] find a certified radius against ℓ_2 perturbations given by $\sup_{0<\beta\leq\min(1,\frac{1}{2}\log\frac{p_1}{p_2})}\frac{\sigma\beta}{\sqrt{2\log\left(\frac{1.25(1+\exp(\beta))}{p_1-\exp(2\beta)p_2}\right)}}$, while Cohen *et al*. [10] give a tight robustness guarantee for ℓ_2 perturbations of the form $\frac{\sigma}{2}\left(\Phi^{-1}(p_1)-\Phi^{-1}(p_2)\right)$.

Clearly, all choices of distance metrics dominate the certificates found using the Lecuyer *et al*. [24] method, and for values of p_1 close to $1/2$, d_{TV} is approximately equal to the tight Cohen *et al*. [10] guarantee. However, the certified radius found using d_{TV} is linear with respect to the top predicted probability, and so becomes a weaker guarantee for larger probabilities. Robustness guarantees provided by Rényi and chi-squared divergences are approximately equal; a finer-grained visualization of the difference between these two divergences is given in appendix B.

We formalize the trade-offs between different choices of divergences with the following proposition.

Proposition 1. *Let $\epsilon_{d_{KL}}, \epsilon_{d_{\chi^2}}, \epsilon_{d_{H^2}}, \epsilon_{d_B}, \epsilon_{d_\alpha}$, and $\epsilon_{[24]}$, denote the certificates found using $d_{KL}, d_{\chi^2}, d_{H^2}, d_B, d_\alpha$, and*

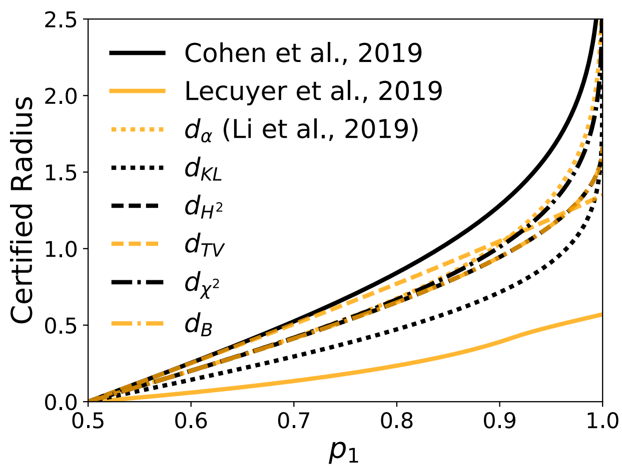


Figure 1: Comparison of the certified radius against perturbations targeting the ℓ_2 norm, for different divergences, as a function of the top predicted probability, p_1, with $\sigma = 1$.

the Lecuyer et al. [24] approach, respectively. Then, the following holds

1. $\forall p_1 \in (\frac{1}{2}, 1)$, $\epsilon_{d_\alpha} > \epsilon_{d_{\chi^2}}$.
2. $\forall p_1 \in (\frac{1}{2}, 1)$, $\epsilon_{d_{\chi^2}} > \epsilon_{d_{KL}}$.
3. $\forall p_1 \in (\frac{1}{2}, 1)$, $\epsilon_{d_{\chi^2}} > \epsilon_{d_{H^2}}$.
4. $\forall p_1 \in [\frac{1}{2}, 1]$, $\epsilon_{d_B} = \epsilon_{d_{H^2}}$.
5. $\forall p_1 \in (\frac{1}{2}, 0.998)$, $\epsilon_{d_{H^2}} > \epsilon_{d_{KL}}$.
6. $\forall p_1 \in (\frac{1}{2}, 1)$, $\epsilon_{d_{KL}} > \epsilon_{[24]}$.

Proof. See appendix C. □

Proposition 1 defines a strict hierarchy, and so informs us of the best divergence one can use to certify an input against ℓ_2 perturbations using the Li *et al.* [26] approach.

2.4. Certification guarantees beyond the ℓ_2 based perturbations via different smoothing measures

The Gaussian distribution is a natural choice for the smoothing measure because it naturally leads to robustness guarantees in the ℓ_2 norm. However, it is also a convenient choice of smoothing measure because it is a member of the location-scale family of distributions. This means that, fixing $x \in \mathcal{X}$, sampling from $x + \mathcal{N}(0, \sigma^2)$ is equivalent to sampling from $\mathcal{N}(x, \sigma^2)$. Importantly, addition of a constant, x, does not change the family of the smoothing measure, and so we can use well known formula for the distances between two Gaussian distributions to derive robustness guarantees.

Unfortunately, not all distributions belong to the location-scale family, and so, in our formulation, we are not free to choose any distribution for smoothing. Another convenient choice of a location-scale distribution is the generalized Gaussian distribution [30], denoted $\mathcal{GN}(\mu, \sigma, s)$, whose density function is given by

$$p(x) = \frac{s}{2\sigma\Gamma(\frac{1}{s})} e^{-|\frac{x-\mu}{\sigma}|^s} \tag{13}$$

where μ is the mean, σ denotes a scaling factor and s denotes a shaping factor. The Laplacian distribution is recovered when $s = 1$, the Gaussian $\mathcal{N}(\mu, \frac{\sigma^2}{2})$ when $s = 2$, and the uniform distribution on $(\mu - \sigma, \mu + \sigma)$ as $s \to \infty$. We will show that by using this smoothing measure we can find robustness guarantees to ℓ_p perturbations, where $p \in \mathbb{N}_{>0}$.

We show in appendix D that given inputs x and x' the Kullback–Leibler (KL) divergence of $\mathcal{GN}(x, \sigma, s)$ and $\mathcal{GN}(x', \sigma, s)$ is given by

$$\sum_{k=1}^{s} \binom{s}{k} \frac{(1 + (-1)^{s-k})\Gamma(\frac{s-k+1}{s})\|x - x'\|_k^k}{2\sigma^k \Gamma(\frac{1}{s})} \tag{14}$$

We also show in appendix A that the KL divergence of two multinomial distributions P and Q (that disagree on the index of the top probability) is lower bounded by

$$d_{KL}(Q, P) \geq -\log(2\sqrt{p_1 p_2} + 1 - p_1 - p_2) \tag{15}$$

Then we use the data processing inequality to prove robustness up to $\|x - x'\|_p < \epsilon$ if the following holds

$$d_{KL}(f(x + \mathcal{GN}(0, \sigma, p)), f(x' + \mathcal{GN}(0, \sigma, p))) \tag{16}$$

$$\leq d_{KL}(x + \mathcal{GN}(0, \sigma, p), x' + \mathcal{GN}(0, \sigma, p)) \tag{17}$$

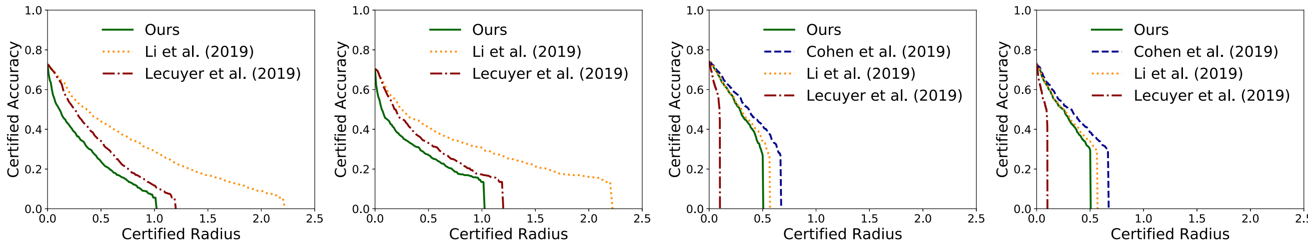
$$\leq \frac{\epsilon^p}{\sigma^p} + \sum_{k=1}^{p-1} \binom{p}{k} \frac{(1 + (-1)^{p-k})\Gamma(\frac{p-k+1}{p})\|x - x'\|_k^k}{2\sigma^k \Gamma(\frac{1}{p})} \tag{18}$$

$$\leq -\log(2\sqrt{p_1 p_2} + 1 - p_1 - p_2) \tag{19}$$

Table 2 gives examples of the KL-divergence of the generalized Gaussian distribution for small ℓ_p norms. For ℓ_p norms with $p = 1$ or $p = 2$, the upper bound to which an input is certifiably robust is given by

$$(-\sigma^p \log(2\sqrt{p_1 p_2} + 1 - p_1 - p_2))^{\frac{1}{p}} \tag{20}$$

For ℓ_p norms with $p > 2, p \in \mathbb{N}$, the upper bound to which an input is certifiably robust is given by ϵ satisfying



(a) CIFAR-10, ℓ_1 (b) ImageNet, ℓ_1 (c) CIFAR-10, ℓ_2 (d) ImageNet, ℓ_2

Figure 2: Certified accuracy against perturbations targeting the ℓ_1 and ℓ_2 norms. Given as a function of the certified radius, the radius around which an input is robust.

Table 2: Examples of the KL divergence between $\mathcal{GN}(\mu_1, \sigma, s)$ and $\mathcal{GN}(\mu_2, \sigma, s)$ for small s.

s	ℓ_s	$d_{KL}(p_1, p_2)$
1	ℓ_1	$\frac{1}{\sigma}\|\mu_1 - \mu_2\|_1$
2	ℓ_2	$\frac{1}{\sigma^2}\|\mu_1 - \mu_2\|_2^2$
3	ℓ_3	$\frac{1}{\sigma^3}\|\mu_1 - \mu_2\|_3^3 + \frac{3}{\sigma\Gamma(\frac{1}{3})}\|\mu_1 - \mu_2\|_1$
4	ℓ_4	$\frac{1}{\sigma^4}\|\mu_1 - \mu_2\|_4^4 + \frac{6\Gamma(\frac{3}{4})}{\sigma^2\Gamma(\frac{1}{4})}\|\mu_1 - \mu_2\|_2^2$

$$\frac{\epsilon^p}{\sigma^p} + \sum_{k=1}^{p-1} \binom{p}{k} \frac{(1 + (-1)^{p-k})\Gamma(\frac{p-k+1}{p}) d^{1-\frac{k}{p}} \epsilon^k}{2\sigma^k \Gamma(\frac{1}{p})} \leq -\log(2\sqrt{p_1 p_2} + 1 - p_1 - p_2) \tag{21}$$

The bound given by eq. (21) is found by noting that $\|x - x'\|_k \leq d^{\frac{1}{k} - \frac{1}{p}} \|x - x'\|_p$, where d is the dimensionality of the data. We can improve upon this naive bound to prove robustness for all norms smaller than p in parallel. Without loss of generality, assume p is even [1], then we can prove robustness for every $0 < k \leq p$, where k is even, up to $\|x - x'\|_k < \epsilon_k$ by solving the constrained problem

$$\max \quad \epsilon_2, \epsilon_4, ..., \epsilon_p \tag{22}$$

subject to

$$\sum_{k=1}^{p} \binom{p}{k} \frac{(1 + (-1)^{p-k})\Gamma(\frac{p-k+1}{p})\epsilon_k^k}{2\sigma^k \Gamma(\frac{1}{p})} \leq -\log(2\sqrt{p_1 p_2} + 1 - p_1 - p_2) \tag{23}$$

$$\epsilon_{i+2} \leq \epsilon_i \leq d^{\frac{1}{i} - \frac{1}{i+2}} \epsilon_{i+2} \tag{24}$$

$$\epsilon_i > 0, \quad 2 \leq i \leq p - 2, \quad i \equiv 0 \pmod 2 \tag{25}$$

Note that the certified radius of robustness around an input is probabilistic because we can only estimate p_1 and p_2, however, we can bound the probability of error to be arbitrarily small. In practice we follow the methods in [10, 24, 26] for estimating p_1 and p_2. Prediction error is bounded by collecting n samples of $f(x + \theta)$, where θ is sampled from a generalized Gaussian distribution, and using the Clopper-Pearson Bernoulli confidence interval to obtain a lower bound estimate of p_1 and an upper bound estimate of p_2, that holds with probability $1 - \gamma$ over the n samples, where $\gamma \ll 1$. Alternatively, we can use the Hoeffding inequality which gives a lower bound of prediction error of $1 - ce^{-2n\epsilon^2}$, where c is the number of classes $|P|$, n is the number of samples and ϵ is the perturbation size. Clearly the error becomes arbitrarily small as we increase the number of samples.

3. Discussion & experiments

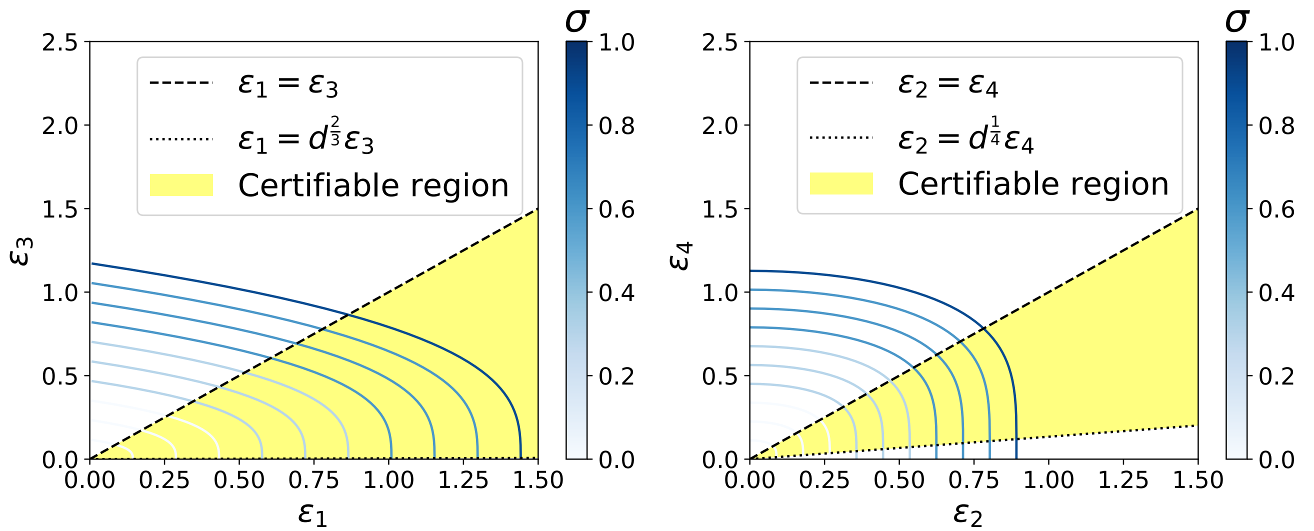
We experimentally validated the certification procedure on the CIFAR-10 [22] and ImageNet [11] datasets. The base classifier is ResNet-50 on ImageNet and ResNet-110 on CIFAR-10 [19]. Given an input x and a classifier f the certification procedure is as follows:

1. Collect n_0 Monte Carlo samples of $f(x + \theta_j)$ to estimate the true class y, where $\theta_j \sim \mathcal{GN}(0, \sigma, s)$ and $j \in [1, ..., n_0]$, with confidence $> 1 - \gamma_0$.
2. Use n_1 Monte Carlo samples to estimate, $\hat{p_1}$, a lower bound of the probability of the most-likely class with confidence $> 1 - \gamma_1$. We follow Cohen *et al.* [10] for estimating $\hat{p_2}$, an upper bound of the probability of the second most-likely class, who noticed nearly all probability mass on other classes is placed on the second most-likely class and so use $\hat{p_2} = 1 - \hat{p_1}$.
3. Use $\hat{p_1}$, $\hat{p_2}$ and eq. (20) or eq. (21) to find a certified radius around x.

For all experiments we use $n_0 = 100, n_1 = 100,000, \gamma_{\{0,1\}} = 0.001, \sigma = 0.25$ and certify 400 test set examples for both CIFAR-10 and ImageNet datasets [2]. The-

[1] A similar statement holds when p is not even.

[2] We perform experiments measuring the effect that various σ have on the certified radius in appendix E.



(a) Certified radius trade-off between ϵ_3 (ℓ_3 norm) and ϵ_1 (ℓ_1 norm). (b) Certified radius trade-off between ϵ_4 (ℓ_4 norm) and ϵ_2 (ℓ_2 norm).

Figure 3: Trade-off in adversarial robustness between different norms, as we vary the noise scale, σ. We plot for a data dimensionality, d, equal to $3 \times 32 \times 32$ (the dimension for CIFAR-10 inputs), and mark the region which gives valid certificates, assuming $\hat{p_1} = 0.99$ and $\hat{p_2} = 1 - \hat{p_1}$.

oretically, this procedure can certify any classifier, however in practice, image classifiers are not stable under noise and so we found it necessary to train classifiers with generalized Gaussian noise (using the same scale and shape parameters as is used during certification). Note that this has the same complexity as standard data augmentation during training and is less expensive than the Madry et al. [28] defense.

3.1. Comparison to related work

For both CIFAR-10 and ImageNet we certify inputs against perturbations in ℓ_1 and ℓ_2 norms and compare against [10, 24, 26]. Figure 2 shows certified accuracy as a function of the certified radius. In general, the largest certified regions come against perturbations targeting the ℓ_1 norm. In appendix F, we show qualitative examples of inputs smoothed with generalized Gaussian noise and the corresponding robustness guarantees in the ℓ_1, ℓ_2, and ℓ_3 norms.

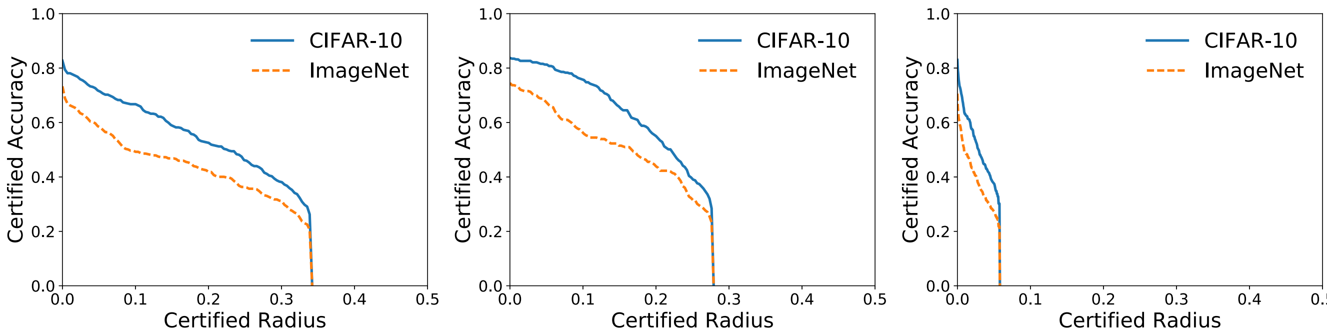
While the primary boon of our certification procedure is its ability to certify inputs to adversarial perturbations beyond ℓ_1 and ℓ_2 norms, the method is not substantially weaker than related work in either norm. In fig. 2a and fig. 2b, we compare with Lecuyer *et al*. [24] and Li *et al*. [26] for ℓ_1 norm certificates. Given estimates $\hat{p_1}$ and $\hat{p_2}$, Lecuyer *et al*. [24] find a certified radius against ℓ_1 perturbations given by $\frac{\sigma}{2}\log(\hat{p_1}/\hat{p_2})$, while Li *et al*. [26] find a certified radius against ℓ_1 perturbations given by $\sigma\log(1-\hat{p_1}+\hat{p_2})$. Li *et al*. [26] and Teng *et al*. [31] show that this robustness guarantee is tight for the ℓ_1 norm. Our ℓ_1 certificates are slightly weaker than Lecuyer *et al*. [24], and both are dominated by Li *et al*. [26] who obtain the tightest possible certificates.

In fig. 2c and fig. 2d, we compare with Lecuyer *et al*. [24], Li *et al*. [26], and Cohen *et al*. [10] for ℓ_2 norm certificates. Our ℓ_2 certificates strictly dominate Lecuyer *et al*. [24], and are approximately equivalent to Li *et al*. [26]. This equivalence is to be expected since our certificates are closely related to Li *et al*. [26] certificates, which are based on the Rényi divergence between two Gaussians, while ours are based on KL divergence. Clearly, we could improve upon this ℓ_2 guarantee if we used the chi-squared distance instead of KL divergence and a standard Gaussian smoothing measure, as proved by Proposition 1. However, our aim is to show the general capacity of the generalized Gaussian as a smoothing measure for certification.

3.2. Robustness trade-offs between different ℓ_p norms.

As described by eq. (21), to obtain robustness guarantees in $\ell_{p>2}$ norms we must factor in required robustness guarantees in smaller ℓ_p norms. For example, to prove robustness up to $\|x - x'\|_3 < \epsilon_3$ and $\|x - x'\|_1 < \epsilon_1$ we find ϵ_1 and ϵ_3 satisfying

$$\begin{aligned}
&\frac{1}{\sigma^3}\epsilon_3^3 + \frac{3}{\sigma\Gamma(\frac{1}{3})}\epsilon_1 \leq -\log(2\sqrt{\hat{p_1}\hat{p_2}} + 1 - \hat{p_1} - \hat{p_2}) \\
&\wedge \\
&0 < \epsilon_3 \leq \epsilon_1 \leq d^{\frac{2}{3}}\epsilon_3,
\end{aligned} \tag{26}$$



(a) ℓ_3 (b) ℓ_4 (c) ℓ_5

Figure 4: Certified accuracy on 400 CIFAR-10 test set inputs and 400 ImageNet test set inputs against perturbations targeting the ℓ_3, ℓ_4, and ℓ_5 norms. Given as a function of the certified radius, the radius around which an input is robust. Inputs were smoothed under a generalized Gaussian distribution parameterized by $\mathcal{GN}(0, 0.25, p)$.

and to prove robustness up to $\|x - x'\|_4 < \epsilon_4$ and $\|x - x'\|_2 < \epsilon_2$ we find ϵ_2 and ϵ_4 satisfying

$$\begin{aligned} &\frac{1}{\sigma^4}\epsilon_4^4 + \frac{6\Gamma(\frac{3}{4})}{\sigma^2\Gamma(\frac{1}{4})}\epsilon_2^2 \leq -\log(2\sqrt{\hat{p_1}\hat{p_2}} + 1 - \hat{p_1} - \hat{p_2}) \\ &\wedge \\ &0 < \epsilon_4 \leq \epsilon_2 \leq d^{\frac{1}{4}}\epsilon_4, \end{aligned} \tag{27}$$

We visualize this trade-off in fig. 3 for ℓ_3 and ℓ_4 norms. That is, the trade-off in certified robustness between those norms and certified robustness in ℓ_1 and ℓ_2, respectively. We visualize the trade-off as we vary the noise scale σ, assuming a robust classifier that classifies inputs correctly with $\hat{p_1} = 0.99$ and $\hat{p_2} = 0.01$. We can smoothly exchange robustness in one norm for robustness in another norm. For example, given $\sigma = 1$ and a CIFAR-10 input, we can reduce the guaranteed robustness in the ℓ_3 norm from an approximate certified radius of 0.86 to approximately 0, and increase the guaranteed robustness in the ℓ_1 norm from a certified radius of 0.86 to 1.44. In fig. 4, we show certified accuracy as a function of certified radius in the ℓ_3, ℓ_4, and ℓ_5 norms on the CIFAR-10 and ImageNet datasets. To find the maximum ϵ_3 we solve eq. (26) such that $\epsilon_3 = \epsilon_1$. Similarly for ϵ_4 we solve eq. (27) such that $\epsilon_4 = \epsilon_2$, and exend this line of reasoning to find $\epsilon_5 = \epsilon_3 = \epsilon_1$ for the ℓ_5 norm. Clearly, we can find non-negligible certified radii in norms outside of ℓ_1 and ℓ_2.

3.3. Robustness guarantees as $\ell_{p\to\infty}$.

An immediate question arises when observing our certification procedure, can we find non-vacuous robustness guarantees for arbitrarily large ℓ_p norms, where p is even [3] [4]? Given eq. (23), note that $\binom{p}{k}(1+(-1)^{p-k})\Gamma(\frac{p-k+1}{p})/2\Gamma(\frac{1}{p}) \geq 1$, $\forall 1 \leq k \leq p$, where k is even, and as $p \to \infty$, $\exists k$ such that $\binom{p}{k}(1+(-1)^{p-k})\Gamma(\frac{p-k+1}{p})/2\Gamma(\frac{1}{p}) \to \infty$. We must therefore solve the problem given in eq. (22)-eq. (25), where eq. (23) is given by

$$\frac{c_2\epsilon_2^2}{\sigma^2} + \frac{c_4\epsilon_4^4}{\sigma^4} + ... + \frac{c_p\epsilon_p^p}{\sigma^p} \leq -\log(2\sqrt{p_1p_2} + 1 - p_1 - p_2) \tag{28}$$

$$\text{where } c_k \in \mathbb{R}_{>1}, 1 \leq k \leq p, k \equiv 0 \pmod 2 \tag{29}$$

To satisfy eq. (24), we can find $\epsilon_2, \epsilon_4, ..., \epsilon_p$ such that $\epsilon_2 = \epsilon_4 = ... = \epsilon_p$; we refer to this value as ϵ, and eq. (28) becomes

$$\begin{aligned} &c_2(\frac{\epsilon}{\sigma})^2 + c_4(\frac{\epsilon}{\sigma})^4 + ... + c_p(\frac{\epsilon}{\sigma})^p \\ &\leq -\log(2\sqrt{p_1p_2} + 1 - p_1 - p_2) \end{aligned} \tag{30}$$

$$\text{where } c_k \in \mathbb{R}_{>1}, 1 \leq k \leq p, k \equiv 0 \pmod 2 \tag{31}$$

For a fixed p_1, p_2, σ, since $\forall k, c_k \geq 1$, and $\exists k$ such that $c_k \to \infty$ when $p \to \infty$, to satisfy the inequality in eq. (30), we must have $\epsilon \to 0$. If we do not fix σ then we require $(\frac{\epsilon}{\sigma})^k \to 0$ as $c_k \to \infty$, and so to certify a non-negligible radius, ϵ, we require $\sigma \to \infty$. However, as $\sigma \to \infty$, the randomized smoothing will cause the input to become too noisy for any classifier to achieve low prediction error.

Clearly, as p grows the largest possible certified radius becomes smaller, because our bound requires this robustness

[3] Equivalent results for this section can be found when p is not even.

[4] The subject of simultaneous robustness over every ℓ_p norm is expanded upon in appendix G.

guarantee holds for every norm smaller than p. One may wonder if we can find an ℓ_p norm in which we can certify a non-vacuous radius that approximates the ℓ_∞ norm arbitrarily well. The difference in volume between a unit ball in the ℓ_p norm and ℓ_∞ norm is given by $\Gamma(1+1/p)^d / \Gamma(1+d/p)$, where d is the data dimensionality. Unfortunately, the error in the approximation is dependent on the data dimensionality. For example, for an ImageNet input where $d = 3 \times 224 \times 224$, if we require the ratio of volumes between an ℓ_p unit ball and ℓ_∞ unit ball to be larger than 0.99, we must take $p = 9 \times 3 \times 224 \times 224$.

3.4. How tight is the bound?

The difference between the certified area and the size of an adversarial perturbation gives a tightness estimate. If the certified radius is close to the size of an adversarial perturbation this implies the bound is close to optimal. To check how tight our bound is we ran the PGD attack [28] minimizing perturbations in the ℓ_2 norm. Because the certification procedure requires the addition of generalized Gaussian noise to the input, the gradient is highly stochastic, leading to extremely slow convergence of the PGD attack. We circumvent this stochasticity by optimizing using the Expectation Over Transformation [2] – we use 1000 Monte Carlo samples to estimate the gradient of an input during the attack. Figure 5 gives attack results on CIFAR-10 along with the certified radius of 400 inputs. We find adversarial examples with norms within $2 - 2.5\times$ the certified radius. Unfortunately, this does not inform us if our bound is loose or if the attack is sub-optimal. We leave a more rigorous investigation of assessing the tightness of our bound for future work.

4. Conclusion

Randomized smoothing has offered a promising approach to scaling robustness guarantees to large architectures. By extending the framework developed by Li *et al*. [26], we showed how different choices of divergences affects the certified radius of robustness around an input. We verified that Rényi divergence is superior to other common f-divergences in this framework, for certifying an input against ℓ_2 perturbations. We then showed that a generalized Gaussian smoothing measure leads to robustness guarantees against any ℓ_p ($p \in \mathbb{N}_{>0}$) minimized adversarial perturbation, however, non-negligible certified radii are only available for small ℓ_p norms.

Acknowledgements

Jamie Hayes is funded by a Google PhD Fellowship in Machine Learning.

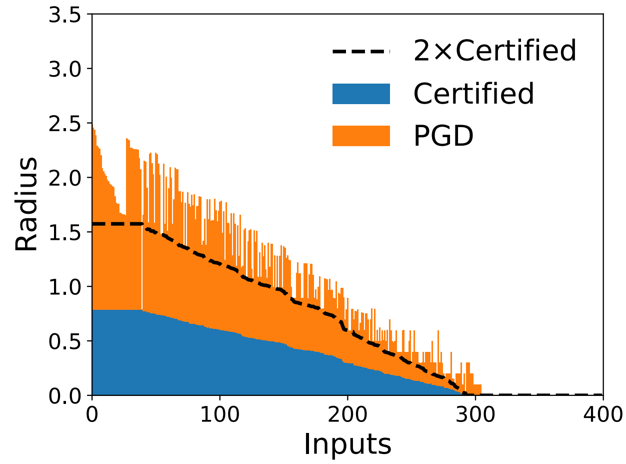


Figure 5: The certified radius and size of adversarial perturbations for 400 CIFAR-10 test inputs using a PGD attack optimizing the ℓ_2 norm. As a guide to assess how close the certified radius is to adversarial perturbation size, we also display $2\times$ the certified radius of an input.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. *arXiv preprint arXiv:1707.07397*, 2017.

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

[4] Akhilan Boopathy, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Cnn-cert: An efficient framework for certifying robustness of convolutional neural networks. *arXiv preprint arXiv:1811.12395*, 2018.

[5] Rudy Bunel, Ilker Turkaslan, Philip HS Torr, Pushmeet Kohli, and M Pawan Kumar. Piecewise linear neural networks verification: A comparative study. 2018.

[6] Nicholas Carlini. Is ami (attacks meet interpretability) robust to adversarial examples? *arXiv preprint arXiv:1902.02322*, 2019.

[7] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[8] Nicholas Carlini, Guy Katz, Clark Barrett, and David L Dill. Ground-truth adversarial examples. 2018.

[9] Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess. Maximum resilience of artificial neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 251–268. Springer, 2017.

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

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[12] Krishnamurthy Dvijotham, Sven Gowal, Robert Stanforth, Relja Arandjelovic, Brendan O'Donoghue, Jonathan Uesato, and Pushmeet Kohli. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*, 2018.

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

[14] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

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

[16] Logan Engstrom, Andrew Ilyas, and Anish Athalye. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1807.10272*, 2018.

[17] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2018.

[18] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

[19] 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*, pages 770–778, 2016.

[20] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.

[21] Marc Khoury and Dylan Hadfield-Menell. On the geometry of adversarial examples. *arXiv preprint arXiv:1811.00525*, 2018.

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

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

[24] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

[25] Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi S Jaakkola. A stratified approach to robustness for randomly smoothed classifiers. *arXiv preprint arXiv:1906.04948*, 2019.

[26] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, pages 9459–9469, 2019.

[27] Chen Liu, Ryota Tomioka, and Volkan Cevher. On certifying non-uniform bound against adversarial attacks. *arXiv preprint arXiv:1903.06603*, 2019.

[28] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[29] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3575–3583, 2018.

[30] Saralees Nadarajah. A generalized normal distribution. *Journal of Applied Statistics*, 32(7):685–694, 2005.

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

[32] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. There is no free lunch in adversarial robustness (but there are unexpected benefits). *arXiv preprint arXiv:1805.12152*, 2018.

[33] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 6541–6550, 2018.

[34] Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

[35] Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S Dhillon, and Luca Daniel. Towards fast computation of certified robustness for relu networks. *arXiv preprint arXiv:1804.09699*, 2018.

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

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