Randomized Smoothing of All Shapes and Sizes

Greg Yang [*1] **Tony Duan** [*1 2] **J. Edward Hu** [1 2] **Hadi Salman** [1] **Ilya Razenshteyn** [1] **Jerry Li** [1]

Abstract

Randomized smoothing is the current state-of-the-art defense with provable robustness against ℓ_2 adversarial attacks. Many works have devised new randomized smoothing schemes for other metrics, such as ℓ_1 or ℓ_∞; however, substantial effort was needed to derive such new guarantees. This begs the question: can we find a general theory for randomized smoothing?

We propose a novel framework for devising and analyzing randomized smoothing schemes, and validate its effectiveness in practice. Our theoretical contributions are: (1) we show that for an appropriate notion of "optimal", the optimal smoothing distributions for any "nice" norms have level sets given by the norm's *Wulff Crystal*; (2) we propose two novel and complementary methods for deriving provably robust radii for any smoothing distribution; and, (3) we show fundamental limits to current randomized smoothing techniques via the theory of *Banach space cotypes*. By combining (1) and (2), we significantly improve the state-of-the-art certified accuracy in ℓ_1 on standard datasets. Meanwhile, we show using (3) that with only label statistics under random input perturbations, randomized smoothing cannot achieve nontrivial certified accuracy against perturbations of ℓ_p-norm $\Omega(\min(1, d^{\frac{1}{p}-\frac{1}{2}}))$, when the input dimension d is large. We provide code in `github.com/tonyduan/rs4a`.

1. Introduction

Deep learning models are vulnerable to adversarial examples – small imperceptible perturbations to their inputs that lead to misclassification (Goodfellow et al., 2015; Szegedy et al., 2014). To solve this problem, recent works proposed heuristic defenses that are robust to specific classes of perturbations, but many would later be broken by stronger attacking algorithms (Carlini & Wagner, 2017; Athalye et al., 2018; Uesato et al., 2018). This led the community to both strengthen empirical defenses (Kurakin et al., 2016; Madry et al., 2017) as well as build *certified* defenses that provide robustness guarantees, i.e., models whose predictions are constant within a neighborhood of their inputs (Wong & Kolter, 2018; Raghunathan et al., 2018a). In particular, *randomized smoothing* is a recent method that has achieved state-of-the-art provable robustness (Lecuyer et al., 2018; Li et al., 2019; Cohen et al., 2019). In short, given an input, it outputs the class most likely to be returned by a base classifier, typically a neural network, under random noise perturbation of the input. This mechanism confers stability of the output against ℓ_p perturbations, even if the base classifier itself is highly non-Lipschitz. Canonically, this noise has been Gaussian, and the adversarial perturbation it protects against has been ℓ_2 (Cohen et al., 2019; Salman et al., 2019a; Zhai et al., 2020), but some have explored other kinds of noises and adversaries as well (Lecuyer et al., 2018; Li et al., 2019; Dvijotham et al., 2019). In this paper, we seek to comprehensively understand the interaction between the choice of smoothing distribution and the perturbation norm.[1]

1. We propose two new methods to compute robust certificates for additive randomized smoothing against different norms.
2. We show that, for $\ell_1, \ell_2, \ell_\infty$ adversaries, the optimal smoothing distributions have level sets that are their respective *Wulff Crystals* — a kind of equilibrated crystal structure studied in physics since 1901 (Wulff).
3. Using the above advances, we obtain state-of-the-art ℓ_1 certified accuracy on CIFAR-10 and ImageNet. With stability training (Li et al., 2019), semi-supervised learning (Carmon et al., 2019), and pre-training in the fashion of Hendrycks et al. (2019), we further improve CIFAR-10 certified accuracies, with $> 30\%$ advantage over prior SOTA for ℓ_1 radius ≥ 1.5. See Table 1.
4. Finally, we leverage the classical theory of Banach space *cotypes* (Wojtaszczyk, 1996) to show that current techniques for randomized smoothing cannot certify nontrivial accuracy at more than $\Omega(\min(1, d^{\frac{1}{p}-\frac{1}{2}}))$ ℓ_p-radius, if all one uses are the probabilities of labels when classifying randomly perturbed input.

[*] Equal contribution [1] Microsoft Research AI [2] Work done as part of the Microsoft AI Residency Program. Correspondence to: Greg Yang <gregyang@microsoft.com>, Tony Duan <tony.duan@microsoft.com>, Jerry Li <jerrl@microsoft.com>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

[1] V2 update: we added results using stability training, semi-supervised learning, and ImageNet pre-training. See Table 1.

ImageNet	ℓ_1 Radius	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
	Laplace, Teng et al. (2019) (%)	48	40	31	26	22	19	17	14
	Uniform, Ours (%)	55	49	46	42	37	33	28	25
	+ Stability Training	**60**	**55**	**51**	**48**	**45**	**43**	**41**	**39**
CIFAR-10	ℓ_1 Radius	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
	Laplace, Teng et al. (2019) (%)	61	39	24	16	11	7	4	3
	Uniform, Ours (%)	70	59	51	43	33	27	22	18
	+ Stability Training	70	60	53	47	**43**	39	35	28
	+ Stability Training, Semi-supervision	**74**	**63**	54	**48**	**43**	38	34	31
	+ Stability Training, Pre-training	**74**	62	**55**	**48**	**43**	**40**	**37**	**33**

Table 1. Certified top-1 accuracies of our ℓ_1-robust classifiers, vs previous state-of-the-art, at various radii, for ImageNet and CIFAR-10.[3]

2. Related Works

Defences against adversarial examples are mainly divided into *empirical* defenses and *certified* defenses.

Empirical defenses are heuristics designed to make learned models empirically robust. An example of these are *adversarial training* based defenses (Kurakin et al., 2016; Madry et al., 2017) which optimize the parameters of a model by minimizing the worst-case loss over a neighborhood around the input to these models (Carlini & Wagner, 2017; Laidlaw & Feizi, 2019; Wong et al., 2019; Hu et al., 2020). Such defenses may seem powerful, but have no guarantees that they are not "breakable". In fact, the majority of the empirical defenses proposed in the literature were later "broken" by stronger attacks (Carlini & Wagner, 2017; Athalye et al., 2018; Uesato et al., 2018; Athalye & Carlini, 2018).

Certified defenses guarantee that for any input x, the classifier's output is constant within a small neighborhood of x. Such defenses are typically based on certification methods that are either *exact* or *conservative*. Exact methods include those based on Satisfiability Modulo Theories solvers (Katz et al., 2017; Ehlers, 2017) or mixed integer linear programming (Tjeng et al., 2019; Lomuscio & Maganti, 2017; Fischetti & Jo, 2017), which, although guaranteed to find adversarial examples if they exist, are unfortunately computationally inefficient. On the other hand, conservative methods are more computationally efficient, but might mistakenly flag a "safe" data point as vulnerable to adversarial examples (Wong & Kolter, 2018; Wang et al., 2018a;b; Raghunathan et al., 2018a;b; Wong et al., 2018; Dvijotham et al., 2018b;a; Croce et al., 2018; Salman et al., 2019b; Gehr et al., 2018; Mirman et al., 2018; Singh et al., 2018; Gowal et al., 2018; Weng et al., 2018; Zhang et al., 2018). However, none of these defenses scale to practical networks. Recently, a new method called randomized smoothing has been proposed as a *probabilistically* certified defense, whose architecture-independence makes it scalable.

Randomized smoothing Randomized smoothing was first proposed as a heuristic defense without any guarantees (Liu et al., 2018; Cao & Gong, 2017). Later on, Lecuyer et al. (2018) proved a robustness guarantee for smoothed classifiers from a differential privacy perspective. Subsequently, Li et al. (2019) gave a stronger robustness guarantee utilizing tools from information theory. Recently, Cohen et al. (2019) provided a tight ℓ_2 robustness guarantee for randomized smoothing, applied by Salman et al. (2020) to provably defend pre-trained models for the first time. Furthermore, a series of papers came out recently that developed robustness guarantees against other adversaries such as ℓ_1-bounded (Teng et al., 2019), ℓ_∞-bounded (Zhang* et al., 2020), ℓ_0-bounded (Levine & Feizi, 2019a; Lee et al., 2019), and Wasserstein attacks (Levine & Feizi, 2019b). In Section 4.3, we give a more in-depth comparison on how our techniques compare to their results.

Wulff Crystal We are the first to relate to adversarial robustness the theory of *Wulff Crystals*. Just as the round soap bubble minimizes surface tension for a given volume, the Wulff Crystal minimizes certain similar surface energy that arises when the crystal interfaces with another material. The Russian physicist George Wulff first proposed this shape via physical arguments in 1901 (Wulff, 1901), but its energy minimization property was not proven in full generality until relatively recently, building on a century worth of work (Gibbs, 1875; Wulff, 1901; Hilton, 1903; Liebmann, 1914; von Laue; Dinghas, 1944; Burton et al., 1951; Herring; Constable, 1968; Taylor, 1975; 1978; Fonseca & Müller, 1991; Brothers & Morgan, 1994; Cerf, 2006).

No-go theorems for randomized smoothing Prior to the initial submission of this manuscript, the only other no-go theorem for randomized smoothing in the context of adversarial robustness is Zheng et al. (2020). However, they are only concerned with a non-standard notion of certified robustness that does not imply anything for the original problem. Moreover, they show that, under this different notion of robustness, if they are robust for ℓ_∞, then the ℓ_2 norm of the noise must be large on average. While this

[3] Unless stated otherwise, these models were trained with noise augmentation. In our replication of Teng et al. (2019), our noise augmentation results matched their adversarial training results.

provides indirect evidence for the hardness of certifying ℓ_∞, it does not actually address the question. Our result, on the other hand, directly rules out a large suite of current techniques for deriving robust certificates for all ℓ_p norms for $p > 2$, for the standard notion of certified robustness.

After the initial submission of this manuscript, we became aware of two concurrent works (Blum et al., 2020; Kumar et al., 2020) that claim impossibility results for randomized smoothing. Blum et al. (2020) demonstrate that, under some mild conditions, any smoothing distribution for ℓ_p with $p > 2$ must have large component-wise magnitude. This gives indirect evidence for the hardness of the problem, but does not directly show a limit for the utility of randomized smoothness for the robust classification problem, which we do in this work. Kumar et al. (2020) demonstrate that certain classes of smoothing distributions cannot certify ℓ_∞ without losing dimension-dependent factors. Our result is more general, as it rules out *any* class of smoothing distributions, and in fact, any smoothing scheme that allows the distribution to vary arbitrarily with the input point.

3. Randomized Smoothing

Consider a classifier f from $\mathbb{R}^d$ to classes $\mathcal{Y}$ and a distribution q on $\mathbb{R}^d$. Randomized smoothing with q is a method that constructs a new, *smoothed* classifier g from the *base* classifier f. The smoothed classifier g assigns to a query point x the class which is most likely to be returned by the base classifier f when x is perturbed by a random noise sampled from q, i.e.,

$$g(x) \overset{\text{def}}{=} \operatorname*{argmax}_{c \in \mathcal{Y}} q(U_c - x) \tag{1}$$

where U_c is the decision region $\{x' \in \mathbb{R}^d : f(x') = c\}$, $U_c - x$ denotes the translation of U_c by $-x$, and $q(U)$ is the measure of U under q, i.e. $q(U) = \mathbb{P}_{\delta\sim q}(\delta \in U)$.

Robustness guarantee for smoothed classifiers For $p \in [0, 1]$, $v \in \mathbb{R}^d$, define the *growth function*

$$\mathcal{G}_q(p, v) \overset{\text{def}}{=} \sup_{U \subseteq \mathbb{R}^d : q(U) = p} q(U - v), \tag{2}$$

One can think of U has the decision region of some base classifier. Thus $\mathcal{G}_q(p, v)$ gives the maximal growth of measure of a set (i.e. decision region) when q is shifted by the vector v, if we only know the initial measure p of the set.

Consider an adversary that can perturb an input additively by any vector v inside an allowed set $\mathcal{B}$. In the case when $\mathcal{B}$ is the ℓ_2 ball and q is the Gaussian measure, Cohen et al. (2019) gave a simple expression for $\mathcal{G}_q$ involving the Gaussian CDF, derived via the Neyman-Pearson lemma, which is later rederived by Salman et al. (2019a) as a nonlinear Lipschitz property. Likewise, the expression for Laplace distributions was derived by Teng et al. (2019). (See Theorem F.10 and Theorem F.11 for their expressions.)

Suppose when the base classifier f classifies $x + \delta$, $\delta \sim q$, the class $c \in \mathcal{Y}$ is returned with probability $\rho = \mathbb{P}_{\delta\sim q}(f(x+\delta) = c) > 1/2$. Then the smoothed classifier g will not change its prediction under the adversary's perturbations if [4]

$$\sup_{v \in \mathcal{B}} \mathcal{G}_q(1 - \rho, v) < 1/2. \tag{3}$$

4. Methods for Deriving Robust Radii

Let q be a distribution with a density function, and we shall write $q(x), x \in \mathbb{R}^d$, for the value of the density function on x. Then, given a shift vector $v \in \mathbb{R}^d$ and a ratio $\kappa > 0$, define the *Neyman-Pearson set*

$$\mathcal{NP}_\kappa \overset{\text{def}}{=} \{x \in \mathbb{R}^d : \kappa q(x - v) \geq q(x)\}. \tag{4}$$

Then the Neyman-Pearson lemma tells us that (Neyman & Pearson, 1933; Cohen et al., 2019)

$$\mathcal{G}_q(q(\mathcal{NP}_\kappa), v) = q(\mathcal{NP}_\kappa - v). \tag{NP}$$

While this gives way to a simple expression for the growth function when q is Gaussian (Cohen et al., 2019), it is difficult for more general distributions as the geometry of $\mathcal{NP}_\kappa$ becomes hard to grasp. To overcome this difficulty, we propose the *level set method* that decomposes this geometry so as to compute the growth function exactly, and the *differential method* that upper bounds the growth function derivative, loosely speaking.

4.1. The Level Set Method

For each $t > 0$, let U_t be the superlevel set

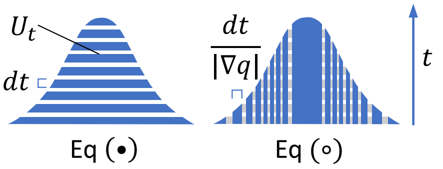
$$U_t \overset{\text{def}}{=} \{x \in \mathbb{R}^d : q(x) \geq t\}.$$

Then its boundary ∂U_t is the level set with $q(x) = t$ under regularity assumptions. The integral of q's density is of course 1, but this integral can be expressed as the integral of the volumes of its superlevel sets:

$$1 = \int q(x)\,\mathrm{d}x = \int_0^\infty \operatorname{Vol}(U_t)\,\mathrm{d}t. \tag{$\bullet$}$$

If q has a differentiable density, then we may rewrite this as an integral of *level* sets (Theorem E.3):

$$1 = \int_0^\infty \int_{\partial U_t} \frac{t}{\|\nabla q(x)\|_2}\,\mathrm{d}x\,\mathrm{d}t. \tag{$\circ$}$$



[4] Many earlier works state robustness guarantees in terms of estimates of $p_A = \rho$ of the top class and p_B of the runner up class; however, their implementations are all in the form provided here, as p_B is usually taken to be $1 - p_A$.

The graphics above illustrate the two integral expressions (best viewed on screen). In this level set perspective, the Neyman-Pearson set $\mathcal{NP}_\kappa$ (Eq. (4)) can be written as

$$\begin{aligned}\mathcal{NP}_\kappa &= \bigcup_{t>0} \{x : q(x) = t \text{ and } q(x-v) \geq t/\kappa\} \\ &= \bigcup_{t>0} \{\partial U_t \cap (U_{t/\kappa} + v)\}.\end{aligned}$$

Then naturally, its measure is calculated by

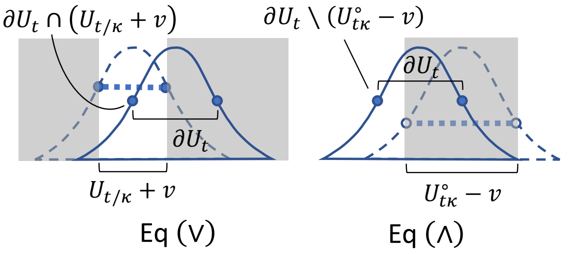
$$q(\mathcal{NP}_\kappa) = \int_0^\infty \int_{\partial U_t \cap (U_{t/\kappa}+v)} \frac{t}{\|\nabla q(x)\|_2} \,\mathrm{d}x \,\mathrm{d}t. \qquad (\vee)$$

Similarly, the Neyman-Pearson set can also be written from the perspective of $q(\cdot - v)$,

$$\begin{aligned}\mathcal{NP}_\kappa &= \bigcup_{t>0} \{x : q(x-v) = t \text{ and } q(x) \leq t\kappa\} \\ &= \bigcup_{t>0} \{(\partial U_t + v) \setminus \mathring{U}_{t\kappa}\},\end{aligned}$$

where $\mathring{U}$ is the interior of the closed set U. So its measure under $q(\cdot - v)$ is

$$q(\mathcal{NP}_\kappa - v) = \int_0^\infty \int_{\partial U_t \setminus (\mathring{U}_{t\kappa} - v)} \frac{t}{\|\nabla q(x)\|_2} \,\mathrm{d}x \,\mathrm{d}t. \qquad (\wedge)$$


$\partial U_t \cap (U_{t/\kappa} + v)$
$\partial U_t \setminus (U^\circ_{t\kappa} - v)$
∂U_t
∂U_t
$U_{t/\kappa} + v$
$U^\circ_{t\kappa} - v$
Eq (∨)
Eq (∧)

The graphics above illustrate the integration domains of x in Eqs. (∨) and (∧). In general, the geometry of $\partial U_t \cap (U_{t/\kappa} + v)$ or $\partial U_t \setminus (\mathring{U}_{t\kappa} - v)$ is still difficult to handle, but in highly symmetric cases when U_t are concentric balls or cubes, Eqs. (∨) and (∧) can be calculated efficiently.

Computing Robust Radius Eqs. (∨) and (∧) allow us to compute the growth function by Eq. (NP). In general, this yields an *upper bound* of the robust radius

$$\begin{aligned}&\sup\left\{r : \sup_{\|v\|_p \leq r} \mathcal{G}_q(1-\rho, v) < 1/2\right\} \\ &\leq \sup\{r : \mathcal{G}_q(1-\rho, ru) < 1/2\}\end{aligned}$$

for any particular u with $\|u\|_p = 1$. With sufficient symmetry, e.g. with ℓ_2 adversary and distributions with spherical level sets, this upper bound becomes *tight* for well-chosen u, and we can build a lookup table of certified radii. See Algorithms 1 and 2.

Algorithm 1 Pre-Computing Robust Radius Table via Level Set Method for Spherical Distributions Againt ℓ_2 Adversary

```
**Input:** Radii $r_1 < \ldots < r_N$
Initialize $u = (1, 0, \ldots, 0) \in \mathbb{R}^d$.
**for** $i = 1$ **to** $N$ **do**
    Find $\kappa$ s.t. $q(\mathcal{NP}_\kappa - r_i u) = 1/2$ (via Eq. (∧) or Theorem I.20) by binary search
    Compute $p_i \leftarrow q(\mathcal{NP}_\kappa)$ via Eq. (∨) or Theorem I.20
**end for**
**Output:** $p_1 > \cdots > p_N$
```

Algorithm 2 Certification with Table

```
**Input:** Probability of correct class $\rho$
**Output:** Look up $r_i$ where $p_i \geq 1 - \rho > p_{i+1}$
```

4.2. The Differential Method

To derive certification (robust radius *lower bounds*) for more general distributions, we propose a *differential method*, which can be thought of as a vast generalization of the proof in Salman et al. (2019a) of the Gaussian robust radius. The idea is to compute the largest possible *infinitesimal increase in q-measure* due to an *infinitesimal adversarial perturbation*. More precisely, given a norm $\|\cdot\|$, and a smoothing measure q, we define

$$\Phi(p) \overset{\text{def}}{=} \sup_{\|v\|=1} \sup_{U \subseteq \mathbb{R}^d : q(U) = p} \lim_{r \searrow 0} \frac{q(U - rv) - p}{r}. \qquad (5)$$

Intuitively, one can then think of $1/\Phi(p)$ as the *smallest* possible perturbation in $\|\cdot\|$ needed to effect a unit of infinitesimal increase in p. Therefore,

Theorem 4.1 (Theorem F.6)**.** *The robust radius in $\|\cdot\|$ is at least*

$$R \overset{\text{def}}{=} \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} \,\mathrm{d}p,$$

where ρ is the probability that the base classifier predicts the right label under random perturbation by q.

By exchanging differentiation and integration and applying a similar greedy reasoning as in the Neyman-Pearson lemma, $\Phi(p)$ can be derived for many distributions q and integrated symbolically to obtain expressions for R. We demonstrate the technique with a simple example below, but much of it can be automated; see Theorem F.6.

Example 4.2 (see Theorem I.6). If the smoothing distribution is $q(x) \propto \exp(-\|x\|_\infty/\lambda)$, then the robust radius against an ℓ_1 adversary is at least

$$R = 2d\lambda(\rho - 1/2),$$

when ρ is the probability of the correct class as in Theorem 4.1.

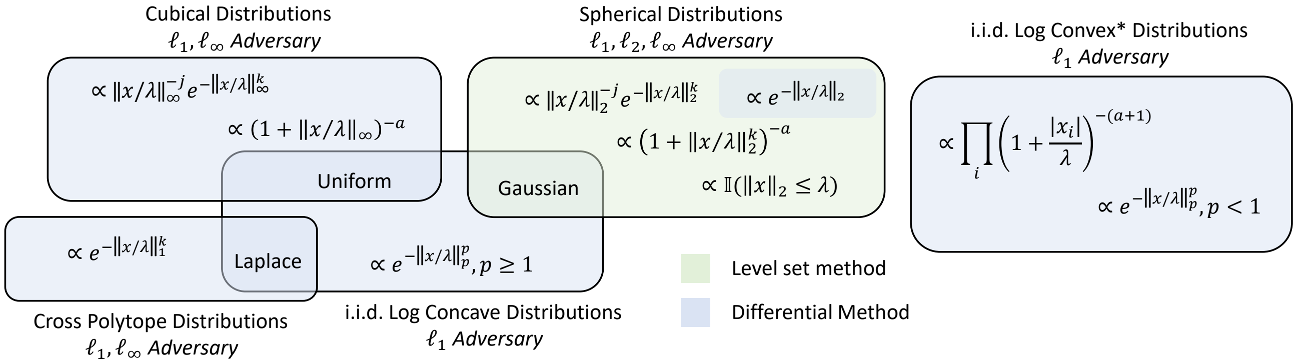

Cubical Distributions
ℓ_1, ℓ_∞ Adversary
$\propto \|x/\lambda\|_\infty^{-j} e^{-\|x/\lambda\|_\infty^k}$
$\propto (1 + \|x/\lambda\|_\infty)^{-a}$
Uniform
Spherical Distributions
$\ell_1, \ell_2, \ell_\infty$ Adversary
$\propto \|x/\lambda\|_2^{-j} e^{-\|x/\lambda\|_2^k}$
$\propto e^{-\|x/\lambda\|_2}$
$\propto (1 + \|x/\lambda\|_2^k)^{-a}$
Gaussian
$\propto \mathbb{I}(\|x\|_2 \leq \lambda)$
i.i.d. Log Convex* Distributions
ℓ_1 Adversary
$\propto \prod_i \left(1 + \frac{|x_i|}{\lambda}\right)^{-(a+1)}$
$\propto e^{-\|x/\lambda\|_p^p}, p < 1$
$\propto e^{-\|x/\lambda\|_1^k}$
Laplace
$\propto e^{-\|x/\lambda\|_p^p}, p \geq 1$
Cross Polytope Distributions
ℓ_1, ℓ_∞ Adversary
i.i.d. Log Concave Distributions
ℓ_1 Adversary
🟩 Level set method
🟦 Differential Method

Figure 1. **Smoothing distributions for which we derive robustness guarantees in this paper.** Each box represents a family of distributions that obtain guarantees through similar proofs. Text beside each box indicates the name of the family and the ℓ_p adversaries against which we have guarantees. *Log Convex** means log convex on the positive and negative half lines, but not necessarily on the whole line. The color indicates the basic technique used, among the two proposed techniques in this paper. We explicitly list example densities in each box. For the robust radii formulas, see Table A.1.

Proof Sketch. By linearity in λ, we WLOG assume $\lambda = 1$. By Theorem 4.1 and the monotonicity of Φ, it suffices to show that $\Phi(p) = 1/2d$ for $p \geq 1/2d$. For any fixed U with $q(U) = p$,

$$\lim_{r \searrow 0} \frac{q(U - rv) - p}{r} = \frac{d}{dr} \int_U q(x - rv)\, \mathrm{d}x \bigg|_{r=0}$$
$$= \int_U \langle v, \nabla q(x) \rangle \, \mathrm{d}x.$$

Note $\nabla q(x) = e_x q(x)$, where $e_x = \operatorname{sgn}(x_{i^*}) e_{i^*}$, e_i is the ith unit vector, and $i^* = \operatorname{argmax}_i |x_i|$. Additionally, the above integral is linear in v, so the supremum over $\|v\|_1 = 1$ is achieved on one of the vertices of the ℓ_1 ball. So we may WLOG consider only $v = \pm e_i$; furthermore, due to symmetry of $\nabla q(x)$, we can just assume $v = e_1$:

$$\Phi(p) = \sup_U \lim_{r \searrow 0} \frac{q(U - re_1) - p}{r} = \sup_U \int_U \langle e_1, e_x \rangle q(x)\, \mathrm{d}x,$$

where U ranges over all $q(U) = p$. Note $\langle e_1, e_x \rangle = 0$ if $i^* \neq 1$, and $\operatorname{sgn}(x_{i^*})$ otherwise. Thus, to maximize $\lim_{r \searrow 0} \frac{q(U - re_1) - p}{r}$ subject to the constraint that $q(U) = p$, we should put as much q-mass on those x with large $\langle e_1, e_x \rangle$. For $p \geq 1/2d$, we thus should occupy the entire region $\{x : \langle e_1, e_x \rangle = 1\}$, which has q-mass $1/2d$, and then assign the rest of the q-mass (amounting to $p - 1/2d$) to the region $\{x : \langle e_1, e_x \rangle = 0\}$, which has q-mass $1 - 1/d$. This shows that

$$\Phi(p) = 1/2d, \quad \forall p \in [1/2d, 1 - 1/2d]$$

as desired. □

4.3. Comparison of the Two Methods and Prior Works

We summarize the distributions our methods cover in Fig. 1 and the bounds we derive in Table A.1. We highlight a few broadly applicable robustness guarantees:

Example 4.3 (Theorem I.1). Let $\phi : \mathbb{R} \to \mathbb{R}$ be convex and even, and let CDF_ϕ^{-1} be the inverse CDF of the 1D random variable with density $\propto \exp(-\phi(x))$. If $q(x) \propto \prod_i e^{-\phi(x_i)}$, and ρ is the probability of the correct class, then the robust radius in ℓ_1 is

$$R = \mathrm{CDF}_\phi^{-1}(\rho)$$

and this radius is *tight*. This in particular recovers the Gaussian bound of Cohen et al. (2019), Laplace bound of Teng et al. (2019), and Uniform bound of Lee et al. (2019) in the setting of ℓ_1 adversary.

Example 4.4 (Appendices I.2.1 and I.3.1). Facing an ℓ_1 adversary, cubical distributions, like that in Example 4.2, typically enjoy, via the differential method, ℓ_1 robust radii of the form

$$R = c(\rho - 1/2)$$

for some constant c depending on the distribution.

In general, the level set method always gives certificate as tight as Neyman-Pearson, while the differential method is tight only for infinitesimal perturbations, but can be shown to be tight for certain families, like in Example 4.3 above. On the other hand, the latter will often give efficiently evaluable symbolic expressions and apply to more general distributions, while the former in general will only yield a table of robust radii, and only for distributions whose level sets are sufficiently symmetric (such as a sphere or cube).

For distributions that are covered by both methods, we compare the bounds obtained and note that the differential and level set methods yield almost identical robustness certificates in high dimensions (e.g. number of pixels in CIFAR-10 or ImageNet images). See Appendix B.1.

Many earlier works used differential privacy or f-divergence methods to compute robust radii of smoothed models

(Lecuyer et al., 2018; Li et al., 2019; Dvijotham et al., 2019). In particular, Dvijotham et al. (2019) proposed a general f-divergence framework that subsumed all such works. Our robust radii are computed only from ρ; Dvijotham et al. called this the "information-limited" setting, and we shall compare with their robustness guarantees of this type. While their algorithm in a certain limit becomes as good as Neyman-Pearson, in practice outside the Gaussian distribution, their robust radii are too loose. This is evident by comparing our baseline Laplace results in Table 1 with theirs, which are trained the same way. Additionally, our differential method often yields symbolic expressions for robust radii, making the certification algorithm easy to implement, verify, and run. Moreover, we derive robustness guarantees for many more (distributions, adversary) pairs (Fig. 1 and Table A.1). See Appendix B.2 for a more detailed comparison.

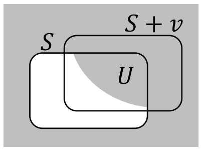
5. Wulff Crystals

A priori, it is a daunting task to understand the relationship between the adversary $\mathcal{B}$ and the smoothing distribution q. In this section, we shall begin our investigation by looking at uniform distributions, and then end with an optimality theorem for all "reasonable" distributions.

Let q be the uniform distribution supported on a measurable set $S \subseteq \mathbb{R}^d$. WLOG, assume S has (Lebesgue) volume 1, $\mathrm{Vol}(S) = 1$. Then for any $v \in \mathbb{R}^d$ and any $p \in [0, 1]$,

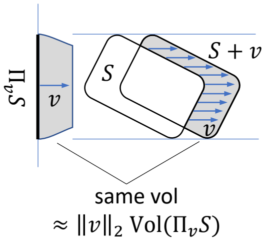
$$\mathcal{G}_q(p, v) = \min\left(1, p + \mathrm{Vol}((S + v) \setminus S)\right).$$

This can be seen easily by taking U in Eq. (2) to be a subset of $(S + v) \cap S$ with volume p (or any set of volume p containing $(S + v) \cap S$ if $p \geq \mathrm{Vol}((S + v) \cap S)$) unioned with the complement of S. For example, in the figure here, U would be the gray region, if $U \cap S$ has volume p.



If S is convex, and we take v to be an infinitesimal translation, then the RHS above is infinitesimally larger than p, as follows:

$$\lim_{r \to 0} \frac{\mathcal{G}_q(p, rv) - p}{r} = \lim_{r \to 0} \frac{\mathrm{Vol}((S + rv) \setminus S)}{r} = \|v\|_2 \mathrm{Vol}(\Pi_v S) \qquad (6)$$



where $\Pi_v S$ is the projection of S along the direction $v/\|v\|_2$, and $\mathrm{Vol}(\Pi_v S)$ is its $(d-1)$-dimensional Lebesgue measure. A similar formula holds when S is not convex as well (Eq. (13)). In the context of randomized smoothing, this means that the classifier g smoothed by q is robust at x under a perturbation $\frac{\frac{1}{2} - p}{\|v\|_2 \mathrm{Vol}(\Pi_v S)} v$ when $1/2 - p$ is small, and p is the probability the base classifier f *mis*classifies $x + \delta$, $\delta \sim q$. Thus, for r small, we have

$$\sup_{v \in r\mathcal{B}} \mathcal{G}_q(p, v) \approx p + r \sup_{v \in \mathcal{B}} \|v\|_2 \mathrm{Vol}(\Pi_v S) = p + r\Phi(p),$$

with Φ as in Eq. (5). The smaller $\sup_{v \in \mathcal{B}} \|v\|_2 \mathrm{Vol}(\Pi_v S)$ is, the more robust the smoothed classifier g is, for a fixed p. A natural question, then, is: among convex sets of volume 1,

$$\text{which set } S \text{ minimizes } \Phi = \sup_{v \in \mathcal{B}} \|v\|_2 \mathrm{Vol}(\Pi_v S)?$$

If $\mathcal{B}$ is the ℓ_p ball, the reader might guess S should either be the ℓ_p ball or the ℓ_r ball with $\frac{1}{r} + \frac{1}{p} = 1$. It turns out the correct answer, at least in the case when $\mathcal{B}$ is a highly symmetric polytope (e.g. $\ell_1, \ell_2, \ell_\infty$ balls), is a kind of *energy-minimizing* crystals studied in physics since 1901 (Wulff).

Definition 5.1. The *Wulff Crystal* (w.r.t. $\mathcal{B}$) is defined as the unit ball of the norm dual to $\|\cdot\|_*$, where $\|x\|_* = \mathbb{E}_{y \sim \mathrm{Vert}(\mathcal{B})} |\langle x, y \rangle|$ and y is sampled uniformly from the vertices of $\mathcal{B}$ [5].

In fact, Wulff Crystals solve the more general problem without convexity constraint.

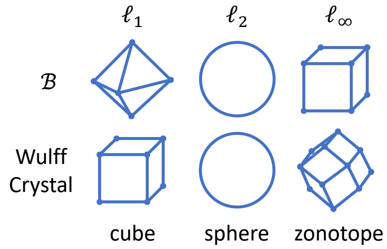
Theorem 5.2 (Theorem G.7, informal)**.** *The Wulff Crystal w.r.t. $\mathcal{B}$ minimizes*

$$\Phi = \sup_{v \in \mathcal{B}} \lim_{r \to 0} r^{-1} \mathrm{Vol}((S + rv) \setminus S)$$

among all measurable (not necessarily convex) sets S of the same volume, when $\mathcal{B}$ is sufficiently symmetric (e.g. $\ell_1, \ell_2, \ell_\infty$ balls).

When $\mathrm{Vert}(\mathcal{B})$ is a finite set, the Wulff Crystal has an elegant description as the *zonotope* of $\mathrm{Vert}(\mathcal{B})$, i.e. the Minkowski sum of the vertices of $\mathcal{B}$ as vectors (Proposition G.4), from which we can derive the following examples.

Example 5.3. The Wulff Crystal w.r.t. ℓ_2 ball is the ℓ_2 ball itself. The Wulff Crystal w.r.t. ℓ_1 ball is a cube (ℓ_∞ ball). The Wulff Crystal w.r.t. ℓ_∞ in 2 dimensions is a rhombus; in 3 dimensions, it is a rhombic dodecahedron; in higher dimension d, there is no simpler description of it other than the zonotope of the vectors $\{\pm 1\}^d$.



[5] When $\mathcal{B}$ is the ℓ_2 ball, $\mathrm{Vert}(\mathcal{B})$ is the entire boundary.

In fact, distributions with Wulff Crystal level sets more generally maximizes the robust radii for "hard" inputs.

Theorem 5.4 (Theorem G.20, informal)**.** *Let $\mathcal{B}$ be sufficiently symmetric. Let q_0 be any distribution with a "reasonable"*[6] *and even density function. Among all "reasonable" and even density functions q whose superlevel sets $\{x : q(x) \geq t\}$ have the same volumes as those of q_0, the quantity*

$$\Phi(1/2) = \sup_{v \in \mathcal{B}} \sup_{q(U)=1/2} \lim_{r \searrow 0} \frac{q(U - rv) - 1/2}{r}$$

is minimized by the unique distribution q^* whose superlevel sets are proportional to the Wulff Crystal w.r.t. $\mathcal{B}$.

This theorem implies that distributions with Wulff Crystal level sets give the best robust radii for those *hard* inputs x that a smooth classifier classifies correctly but only barely, in that the probability of the correct class $\rho = 1/2 + \epsilon$ for some small ϵ. The constraint on the volumes of superlevel sets indirectly controls the variance of the distribution. While this theorem says nothing about the robust radii for ρ away from $1/2$, we find the Wulff Crystal distributions empirically to be highly effective, as we describe next in Section 6.

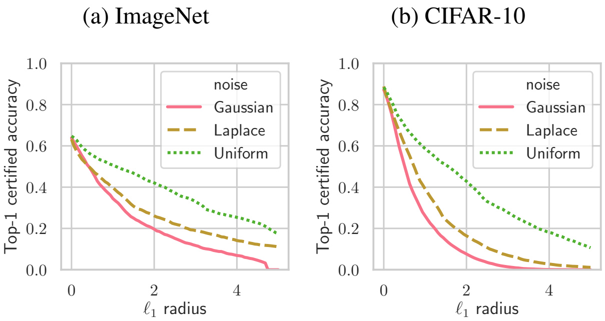
6. Experiments

We empirically study the performance of different smoothing distributions on image classification datasets, using the bounds derived via the level set or the differential method, and verify predictions made by the Wulff Crystal theory. We follow the experimental procedure in Cohen et al. (2019) and further works on randomized smoothing (Salman et al., 2019a; Li et al., 2019; Zhai et al., 2020) using ImageNet (Deng et al., 2009) and CIFAR-10 (Krizhevsky, 2009).

The certified accuracy at a radius ϵ is defined as the fraction of the test set for which the smoothed classifier g correctly classifies and certifies robust at an ℓ_p radius of ϵ. All results were certified with $N = 100,000$ samples and failure probability $\alpha = 0.001$. For each distribution q, we train models across a range of scale parameter λ (see Table A.1), corresponding to the same range of noise variances $\sigma^2 \stackrel{\text{def}}{=} \mathbb{E}_{\delta \sim q}[\frac{1}{d}\|\delta\|_2^2]$ across different distributions. Then we calculate for each model the certified accuracies across the range of considered ϵ. Finally, in our plots, we present, for each distribution, the upper envelopes of certified accuracies attained over the range of considered σ. Further details of experimental procedures are described in Appendix D.

We focus on the effect of the noise distribution in this section and only train models with noise augmentation. In Appendix D we also study (1) stability training, and (2) the use of more data through (a) pre-training on downsampled ImageNet (Hendrycks et al., 2019) and (b) semi-supervised self-training with data from 80 Million Tiny Images (Carmon et al., 2019). As shown in Table 1, these techniques further improve upon our results in this section.

Figure 2. **SOTA ℓ_1 Certified Accuracies.** Certified ℓ_1 top-1 accuracies for ImageNet (left) and CIFAR-10 (right). For each distribution q, we train models across a range of $\sigma^2 \stackrel{\text{def}}{=} \mathbb{E}_{\delta \sim q}[\frac{1}{d}\|\delta\|_2^2]$, and at each level of ℓ_1 adversarial perturbation radius ϵ we report the best certified accuracy.


(a) ImageNet

(b) CIFAR-10

6.1. ℓ_1 Adversary

As previously mentioned, the Wulff Crystal for the ℓ_1 ball is a cube. With this motivation, we explore certified accuracies attained by distributions with cubical level sets.

1. Uniform, $\propto \mathbb{I}(\|x\|_\infty \leq \lambda)$
2. Exponential, $\propto \|x\|_\infty^{-j} e^{-\|x/\lambda\|_\infty^k}$
3. Power law, $\propto (1 + \|x/\lambda\|_\infty)^{-a}$

We compare to previous state-of-the-art approaches using the Gaussian and Laplace distributions, as well as new non-cubical distributions.

4. Exponential ℓ_1 (non-cubical), $\propto \|x\|_1^{-j} e^{-\|x/\lambda\|_1^k}$
5. Pareto i.i.d. (non-cubical), $\propto \prod_i (1 + |x_i|/\lambda)^{-a}$.

The relevant certified bounds are given in Table A.1.

We obtain state-of-the-art robust certificates for ImageNet and CIFAR-10, finding that the Uniform distribution performs best, significantly better than the Gaussian and Laplace distributions (Table 1, Fig. 2). The other distributions with cubic level sets match but do not exceed the performance of Uniform distribution, after sweeping hyperparameters. This verifies that distributions with cubical level sets are significantly better for ℓ_1 certified accuracy than those with spherical or cross-polytope level sets. See results for other distributions in Appendix C.

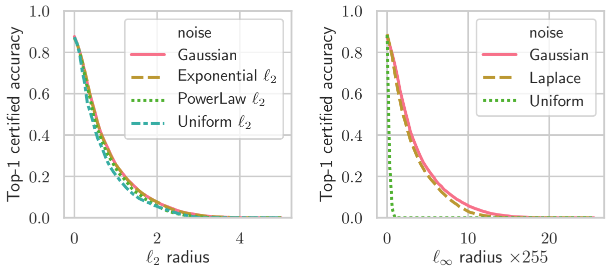
6.2. ℓ_2 Adversary

The Wulff Crystal w.r.t. the ℓ_2 ball is a sphere, so we explore distributions with spherical level sets (Table A.1):

1. Uniform, $\propto \mathbb{I}(\|x\|_2 \leq \lambda)$

[6] *Reasonable* here roughly means Sobolev, i.e. has weak derivative that is integrable, and this can be further relaxed to *bounded variations*; for details see Theorem G.20 and Theorem H.15.

Figure 3. CIFAR-10 certified accuracies for ℓ_2 (left) and ℓ_∞ (right) adversaries. For each distribution q we train models across a range of $\sigma^2 \overset{\text{def}}{=} \mathbb{E}[\frac{1}{d}\|\delta\|_2^2]$, and at each level of ℓ_p adversarial perturbation radius ϵ, we pick the model that maximizes certified accuracy.



2. Exponential, $\propto \|x\|_2^{-j} e^{-\|x/\lambda\|_2^k}$
3. Power law, $\propto (1 + \|x/\lambda\|_2)^{-a}$

We find these distributions perform similarly to, though do not surpass the Gaussian (Fig. 3, left).

6.3. ℓ_∞ Adversary

The Wulff Crystal for the ℓ_∞ ball is the zonotope of vectors $\{\pm 1\}^d$, which is a highly complex polytope hard to sample from and related to many open problems in polytope theory (Ziegler, 1995). However, we can note that it is approximated by a sphere with constant ratio (Proposition G.13), and in high dimension d, the sphere gets closer and closer to minimizing Φ (Theorem 5.2), but the cube and the cross polytope do not (Claim G.15). Accordingly, we find that distributions with spherical level sets outperform those with cubical or cross polytope level sets in certifying ℓ_∞ robustness (Fig. 3, right). In fact, in the next section we show that up to a dimension-independent factor, the Gaussian distribution is optimal for defending against ℓ_∞ adversary if we don't use a more powerful technique than Neyman-Pearson.

7. No-Go Results for Randomized Smoothing

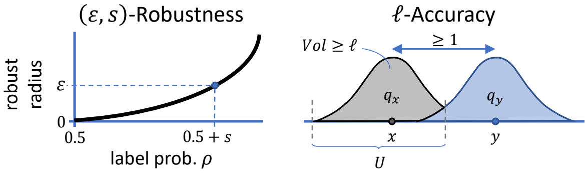
Recall that given a smoothing distribution q, a point $x \in \mathbb{R}^d$, and a binary base classifier $U \subseteq \mathbb{R}^d$ (identified wth its decision region), the smoothed classifier outputs $\operatorname{sgn}(\rho - 1/2)$ where $\rho = q(U - x)$ is the "confidence" of this prediction (Eq. (1)). Randomized smoothing (via Neyman-Pearson) tells us that, if ρ is large enough, then, no matter what U is, a small perturbation of x cannot decrease ρ too much to change $\operatorname{sgn}(\rho - 1/2)$ (Eq. (3)).

If all we care about is robustness, then the optimal strategy would set q to be an arbitrarily wide distribution (say, e.g. a wide Gaussian), and the resulting smoothed classifier is roughly constant. Of course, such a smoothed classifier can never achieve good clean accuracy, so it is not useful. Thus there is an inherent tension between 1) having to have large enough noise variance to be robust and 2) having to have small enough noise variance to avoid trivializing the smoothed classifier. In this section, we seek to formalize this tradeoff. As we'll show, even if we only assume a very weak condition on the accuracy, we can show strong upper bounds on the best robust radius for each ℓ_p norm.

In fact, our negative results below will hold for a more general class of smoothing schemes than those in our positive results in previous sections: In what follows, a *smoothing scheme* for $\mathbb{R}^d$ is any family of probability distributions $\mathcal{Q} = \{q_x\}_{x \in \mathbb{R}^d}$. In practice, including in our paper, almost all smoothing schemes are *translational*, that is, there is some base distribution q, and for every x, the smoothing distribution at x is defined by $q_x(U) = q(U - x)$, for all base classifiers $U \subseteq \mathbb{R}^d$. The above discussion then motivates the following

Definition 7.1. Let $\|\cdot\|$ be a norm over $\mathbb{R}^d$, and let $\mathcal{Q} = \{q_x\}_{x \in \mathbb{R}^d}$ be a smoothing scheme for $\mathbb{R}^d$. We say that $\mathcal{Q}$ satisfies (ε, s, ℓ)-*useful smoothing* with respect to $\|\cdot\|$ if:

1. **((ε, s)-Robustness)** For all x, y with $\|x - y\| < \varepsilon$, if $U \subseteq \mathbb{R}^d$ is any set (*read: base classifier*) satisfying $q_x(U) \geq 1/2 + s$, then $q_y(U) \geq 1/2$.
2. **(ℓ-Accuracy)** For all x, y with $\|x - y\| \geq 1$, there exists a set (*read: base classifier*) $U \subseteq \mathbb{R}^d$ so that $|q_x(U) - q_y(U)| \geq \ell$.



We pause to interpret this definition. Condition (1) indicates how large the certified radii can be for a classifier at any given point x, if the smoothed classifier assigns likelihood at least $1/2 + s$ to it; i.e. $(1/2 + s, \varepsilon)$ is a point on the robust radii curve in the style of Fig. A.1. The goal of the smoothing scheme is to achieve the largest possible ε, for every fixed s. In particular, observe that for ℓ_2, Gaussian smoothing achieves dimension-independent ε, for every fixed choice of s (Theorem F.10).

Condition (2) says that the resulting smoothing should not "collapse" points: in particular, if x, y are far in norm, then there should be some smoothed classifier that distinguishes them. We argue that this is a very mild assumption. For Condition (2) to be satisfied, the U which distinguishes these two points can be completely arbitrary. Thus, if it is violated for $\ell = o(1)$, the two distributions are indistinguishable by any statistical test in high dimension, implying the impossibility of classifying between x and y after smoothing.

We seek to show that, for constant s and l, any (ε, s, ℓ)-useful smoothing scheme must have $\varepsilon = o(1)$ for a number of norms, including ℓ_∞. This would imply that any smoothing scheme that satisfying our weak notion of accuracy can only certify a vanishingly small radius, even when the confidence of the classifier is strictly bounded away from $1/2$ by a constant.

Randomized Smoothing as Metric Embedding A smoothing scheme can be thought of as a mapping from a normed space supported on $\mathbb{R}^d$ to the space of distributions, e.g. each point x is mapped to the distribution q_x. We will show that Definition 7.1 is roughly equivalent to a bi-Lipschitz condition on this mapping, where the target distributions are equipped with the total variation distance. Then the existence of a *useful* smoothing scheme is equivalent to whether $(\mathbb{R}^d, \|\cdot\|)$ can be embedded *with low distortion* into the total variation space of distributions. Classical mathematics has a definitive answer to this question in the form of a geometric invariant, called the *cotype*.

Definition 7.2 (see e.g. Wojtaszczyk (1996))**.** A normed space $T = (X, \|\cdot\|)$ is said to have *cotype* p for $2 \le p \le \infty$ if there exists C such that for all finite sequences $x_1, \ldots, x_n \in X$, we have

$$\mathbb{E}\left[\left\|\sum_{j=1}^n \sigma_j x_j\right\|\right] \ge C^{-1}\left(\sum_{j=1}^n \|x_i\|^p\right)^{1/p},$$

where the σ_j are independent Rademacher random variables. The smallest such C is denoted $C_p(T)$.

When the underlying space of the normed space T is $\mathbb{R}^d$, John's theorem (John, 1948) implies that any norm has cotype 2 with $C_2(T) \le O(d^{1/2})$. Because C_2 lower bounds the distortion of a metric embedding of T, by the aforementioned connection with randomized smothing, C_2 also limits the usefulness of any smoothing scheme of T:

Theorem 7.3. *Let T be any normed space over $\mathbb{R}^d$. There exist universal constants $c, K > 0$ so that any (ε, s, ℓ)-useful smoothing scheme for T with $s/\ell < c$ must have*

$$\varepsilon \le K\sqrt[4]{s/\ell} \cdot C_2(T)^{-1}.$$

In particular, it is well-known that $C_2((\mathbb{R}^d, \|\cdot\|_p)) = \Omega(\max(1, d^{1/2-1/p}))$, for all $p \in [1, \infty]$. Thus, as an immediately corollary, we get:

Corollary 7.4. *For the value of c in Theorem 7.3 and for $p \in [1, \infty]$, any (ε, s, ℓ)-useful smoothing scheme for $(\mathbb{R}^d, \|\cdot\|_p)$ with $s/\ell < c$ must have*

$$\varepsilon \le O(\min(1, d^{-1/2+1/p})).$$

It is easy to see that, up to constants, the Gaussian smoothing scheme achieves equality, and thus is optimal (in terms of dimension dependence), for all $p \in [1, \infty]$.

Discussion After Cohen et al. (2019) showed the surprising scalability of Gaussian randomized smoothing to high-dimensional ℓ_2-robust classification problems, many anticipated that this can be extended to ℓ_∞ as well. One might also hope that, even though it seems like we cannot certify ℓ_2 radius that grows with input dimension, we could do so for ℓ_1. But Theorem 7.3 and Corollary 7.4 present a strong barrier to such hopes. In words:

> *Without using more than the information of the probability ρ of correctly classifying an input under random noise, no smoothing techniques can certify nontrivial robust accuracy at ℓ_∞ radius $\Omega(d^{-1/2})$, or at ℓ_2 or ℓ_1 radius $\Omega(1)$.*

Indeed, the ℓ_1-radii we can obtain nontrivial certified accuracy at are on the same order between CIFAR10 and Imagenet (Fig. 2).

However, there are some ways to bypass this barrier. For one, more information about the base classifier can be collected to produce better robustness certificates. In fact, Dvijotham et al. (2019) proposed a "full-information" algorithm that computes many moments of the base classifier in a convex optimization procedure to improve certified radius, but it is 100 times slower than the "information-limited" algorithms we discuss here that use only ρ. It would be interesting to see whether this technique can be scaled up, and whether other methods can leverage more information[7].

Another route is to directly look for better randomized smoothing schemes for multi-class classification. We formulated our no-go result in the setting of binary classification, and it is not clear whether a similarly strong barrier applies for multi-class classification. However, current techniques for certification only look at the two most likely classes, and separately reason about how much each one can change by perturbing the input. Our no-go result then straightforwardly applies to this case as well.

8. Conclusion

In this work, we have showed how far we can push randomized smoothing with different smoothing distributions against different ℓ_p adversaries, by presenting two new techniques for deriving robustness guarantees, by elucidating the geometry connecting the noise and the norm, and by empirically achieving state-of-the-art in ℓ_1 provable defense. At the same time, we have showed the limit current techniques face against ℓ_p adversaries when $p > 2$, especially ℓ_∞. Our results point out ways to bypass this barrier, by either leveraging more information about the base classifier or by taking advantage of the multi-class problem structure better. We wish to investigate both directions in the future.

More broadly, randomized smoothing is a method for inducing stability in a mechanism while maintaining utility — precisely the bread and butter of differential privacy. We suspect our methods for deriving robustness guarantees here and for optimizing the noise distribution can be useful in that setting as well, where Laplace and Gaussian noise dominate the discussion. Whereas previous work Lecuyer et al. (2018) has applied differential privacy tools to randomized smoothing, we hope to go the other way around in the future.

[7] Lee et al. (2019) also used the decision tree structure of their base classifier to improve ℓ_0 certification, but the ℓ_0-adversary does not fall within our framework.

Acknowledgements

We thank Huan Zhang for brainstorming of ideas and performing a few experiments that unfortunately did not work out. We also thank Aleksandar Nikolov, Sebastien Bubeck, Aleksander Madry, Zico Kolter, Nicholas Carlini, Judy Shen, Pengchuan Zhang, and Maksim Andriushchenko for discussions and feedback.

References

Andoni, A., Krauthgamer, R., and Razenshteyn, I. Sketching and embedding are equivalent for norms. *SIAM Journal on Computing*, 47(3):890–916, 2018.

Athalye, A. and Carlini, N. On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*, 2018.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Blum, A., Dick, T., Manoj, N., and Zhang, H. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *arXiv preprint arXiv:2002.03517*, 2020.

Brothers, J. E. and Morgan, F. The isoperimetric theorem for general integrands. *The Michigan Mathematical Journal*, 41(3):419–431, 1994. doi: 10.1307/mmj/1029005070.

Burton, W. K., Cabrera, N., and Frank, F. C. The growth of crystals and the equilibrium structure of their surfaces. *Phil. Trans. Roy. Soc.*, 1951.

Cao, X. and Gong, N. Z. Mitigating evasion attacks to deep neural networks via region-based classification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pp. 278–287. ACM, 2017.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14. ACM, 2017.

Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P. S., and Duchi, J. C. Unlabeled Data Improves Adversarial Robustness. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11190–11201. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9298-unlabeled-data-improves-adversarial-robustness.pdf.

Cerf, R. *The Wulff Crystal in Ising and Percolation Models: Ecole d'Eté de Probabilités de Saint-Flour XXXIV - 2004*. École d'Été de Probabilités de Saint-Flour. Springer-Verlag, Berlin Heidelberg, 2006. ISBN 978-3-540-30988-8. doi: 10.1007/b128410.

Chrabaszcz, P., Loshchilov, I., and Hutter, F. A Downsampled Variant of ImageNet as an Alternative to the CIFAR datasets. Technical report, July 2017. URL https://arxiv.org/abs/1707.08819v3.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning*, pp. 1310–1320, May 2019. URL http://proceedings.mlr.press/v97/cohen19c.html.

Constable, R. F. S. *Kinetics and Mechanism of Crystallization*. Elsevier Science & Technology Books, 1968. ISBN 978-0-12-673550-5.

Croce, F., Andriushchenko, M., and Hein, M. Provable robustness of relu networks via maximization of linear regions. *arXiv preprint arXiv:1810.07481*, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. ISSN: 1063-6919.

Dinghas, A. Uber einen Gcometrischen Satz von Wulff fur die Gleichgewichtsform von Kristallen. *Z. Kristallog*, 105:304, 1944.

Dvijotham, K., Gowal, S., Stanforth, R., Arandjelovic, R., O'Donoghue, B., Uesato, J., and Kohli, P. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*, 2018a.

Dvijotham, K., Stanforth, R., Gowal, S., Mann, T., and Kohli, P. A dual approach to scalable verification of deep networks. *UAI*, 2018b.

Dvijotham, K. D., Hayes, J., Balle, B., Kolter, Z., Qin, C., Gyorgy, A., Xiao, K., Gowal, S., and Kohli, P. A Framework for Robustness Certification of Smoothed Classifiers Using F-Divergences. September 2019. URL https://openreview.net/forum?id=SJlKrkSFPH.

Ehlers, R. 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.

Evans, L. C. and Gariepy, R. F. *Measure theory and fine properties of functions*. Chapman and Hall/CRC, 2015.

Federer, H. *Geometric measure theory*. Springer, 2014.

Fischetti, M. and Jo, J. Deep neural networks as 0-1 mixed integer linear programs: A feasibility study. *arXiv preprint arXiv:1712.06174*, 2017.

Fonseca, I. and Müller, S. A uniqueness proof for the Wulff Theorem. *Proceedings of the Royal Society of Edinburgh: Section A Mathematics*, 119(1-2):125–136, 1991. doi: 10.1017/S0308210500028365.

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., and Vechev, M. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18. IEEE, 2018.

Gibbs, W. On the Equilibrium of Heterogeneous Substances. *Transactions of the Connecticut Academy of Arts and Sciences*, 1875.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572. arXiv: 1412.6572.

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Mann, T., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Hendrycks, D., Lee, K., and Mazeika, M. Using Pre-Training Can Improve Model Robustness and Uncertainty. In *International Conference on Machine Learning*, pp. 2712–2721, May 2019. URL http://proceedings.mlr.press/v97/hendrycks19a.html.

Herring, C. Konferenz über Struktur und Eigenschaften fester Oberflächen Lake. Geneva (Wisconsin) USA, 29. September bis 1. Oktober 1952. *Angewandte Chemie*.

Hilton, H. *Mathematical Crystallography*. Oxford, 1903.

Hu, J. E., Swaminathan, A., Salman, H., and Yang, G. Improved image wasserstein attacks and defenses. *arXiv preprint arXiv:2004.12478*, 2020.

John, F. Extremum problems with inequalities as subsidiary conditions, studies and essays presented to r. courant on his 60th birthday, january 8, 1948, 1948.

Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pp. 97–117. Springer, 2017.

Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Technical report, 2009.

Kumar, A., Levine, A., Goldstein, T., and Feizi, S. Curse of dimensionality on randomized smoothing for certifiable robustness. *arXiv preprint arXiv:2002.03239*, 2020.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Laidlaw, C. and Feizi, S. Functional adversarial attacks. In *Advances in Neural Information Processing Systems*, pp. 10408–10418, 2019.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*, 2018.

Lee, G.-H., Yuan, Y., Chang, S., and Jaakkola, T. Tight Certificates of Adversarial Robustness for Randomly Smoothed Classifiers. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 4911–4922. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/8737-tight-certificates-of-adversarial-robustness-for-randomly-smoothed-classifiers.pdf.

Levine, A. and Feizi, S. Robustness Certificates for Sparse Adversarial Attacks by Randomized Ablation. Technical report, November 2019a. URL http://arxiv.org/abs/1911.09272. arXiv: 1911.09272.

Levine, A. and Feizi, S. Wasserstein Smoothing: Certified Robustness against Wasserstein Adversarial Attacks. Technical report, October 2019b. URL http://arxiv.org/abs/1910.10783. arXiv: 1910.10783.

Li, B., Chen, C., Wang, W., and Carin, L. Certified Adversarial Robustness with Additive Noise. In Wallach, H., Larochelle, H., Beygelzimer, A., Alché-Buc, F. d., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 9459–9469. Curran Associates, Inc., 2019. URL http://papers.nips.cc/paper/9143-certified-adversarial-robustness-with-additive-noise.pdf.

Liebmann, H. Der Curie-Wulff'sche Satz uber Combinationsformen von Krystallen. *Z. Kristallog*, 53, 1914.

Liu, X., Cheng, M., Zhang, H., and Hsieh, C.-J. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 369–385, 2018.

Lomuscio, A. and Maganti, L. An approach to reachability analysis for feed-forward relu neural networks. *arXiv preprint arXiv:1706.07351*, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Matoušek, J. Lecture notes on metric embeddings. Technical report, Technical report, ETH Zürich, 2013.

McMullen, P. On zonotopes. *Transactions of the American Mathematical Society*, 159:91–109, 1971.

Mirman, M., Gehr, T., and Vechev, M. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pp. 3575–3583, 2018.

Morgan, F. *Geometric measure theory: a beginner's guide*. Academic press, 2016.

Neyman, J. and Pearson, E. IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, February 1933. ISSN 0264-3952, 2053-9258. doi: 10.1098/rsta.1933.0009. URL https://royalsocietypublishing.org/doi/10.1098/rsta.1933.0009.

Nguyen, H. H., Vu, V., et al. Random matrices: Law of the determinant. *The Annals of Probability*, 42(1):146–167, 2014.

Nikodym, O. Sur une classe de fonctions considérée dans l'étude du problème de Dirichlet. *Fund. Math.*, 21:129–150, 1933. URL http://matwbn.icm.edu.pl/ksiazki/fm/fm21/fm21119.pdf.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. *International Conference on Learning Representations (ICLR), arXiv preprint arXiv:1801.09344*, 2018a.

Raghunathan, A., Steinhardt, J., and Liang, P. S. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pp. 10877–10887, 2018b.

Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pp. 11289–11300, 2019a.

Salman, H., Yang, G., Zhang, H., Hsieh, C.-J., and Zhang, P. A convex relaxation barrier to tight robustness verification of neural networks. In *Advances in Neural Information Processing Systems*, pp. 9832–9842, 2019b.

Salman, H., Sun, M., Yang, G., Kapoor, A., and Kolter, J. Z. Black-box smoothing: A provable defense for pretrained classifiers, 2020.

Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems*, pp. 10825–10836, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

Taylor, J. Unique structure of solutions to a class of nonelliptic variational problems. *Proc. Sympos. Pure Math.*, 27:419–427, 1975.

Taylor, J. E. Crystalline variational problems. *Bulletin of the American Mathematical Society*, 84(4):568–588, July 1978. ISSN 0002-9904, 1936-881X.

Teng, J., Lee, G.-H., and Yuan, Y. ℓ_1 Adversarial Robustness Certificates: a Randomized Smoothing Approach. Technical report, September 2019. URL https://openreview.net/forum?id=H1lQIgrFDS.

Tjeng, V., Xiao, K. Y., and Tedrake, R. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyGIdiRqtm.

Uesato, J., O'Donoghue, B., Oord, A. v. d., and Kohli, P. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

von Laue, M. Der Wulffsche Satz für die Gleidigewichtsform von Kristallen. *Zeitschrift für Kristallographie Crystalline Materials*, 105.

Wang, S., Chen, Y., Abdou, A., and Jana, S. Mixtrain: Scalable training of formally robust neural networks. *arXiv preprint arXiv:1811.02625*, 2018a.

Wang, S., Pei, K., Whitehouse, J., Yang, J., and Jana, S. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems*, pp. 6369–6379, 2018b.

Weng, T.-W., Zhang, H., Chen, H., Song, Z., Hsieh, C.-J., Boning, D., Dhillon, I. S., and Daniel, L. Towards fast computation of certified robustness for ReLU networks. In *International Conference on Machine Learning*, 2018.

Wojtaszczyk, P. *Banach spaces for analysts*, volume 25. Cambridge University Press, 1996.

Wong, E. and Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, pp. 5283–5292, 2018.

Wong, E., Schmidt, F., Metzen, J. H., and Kolter, J. Z. Scaling provable adversarial defenses. *Advances in Neural Information Processing Systems (NIPS)*, 2018.

Wong, E., Schmidt, F. R., and Kolter, J. Z. Wasserstein adversarial examples via projected sinkhorn iterations. *arXiv preprint arXiv:1902.07906*, 2019.

Wulff, G. Zur Frage der Geschwindigkeit des Wachstums und der Auflösung der Krystallflagen. *Zeitschrift für Krystallographie und Mineralogie*, 34:449–530, 1901.

Zhai, R., Dan, C., He, D., Zhang, H., Gong, B., Ravikumar, P., Hsieh, C.-J., and Wang, L. Macer: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJx1Na4Fwr.

Zhang*, D., Ye*, M., Gong*, C., Zhu, Z., and Liu, Q. Filling the soap bubbles: Efficient black-box adversarial certification with non-gaussian smoothing, 2020. URL https://openreview.net/forum?id=Skg8gJBFvr.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems*, pp. 4939–4948, 2018.

Zheng, T., Wang, D., Li, B., and Xu, J. A unified framework for randomized smoothing based certified defenses, 2020. URL https://openreview.net/forum?id=ryl71a4YPB.

Ziegler, G. M. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer New York, New York, NY, 1995. ISBN 978-0-387-94365-7 978-1-4613-8431-1. URL http://link.springer.com/10.1007/978-1-4613-8431-1.