

---

# Randomized Smoothing of All Shapes and Sizes

---

Greg Yang<sup>\*1</sup> Tony Duan<sup>\*12</sup> J. Edward Hu<sup>12</sup> Hadi Salman<sup>1</sup> Ilya Razenshteyn<sup>1</sup> Jerry Li<sup>1</sup>

## Abstract

Randomized smoothing is the current state-of-the-art defense with provable robustness against  $\ell_2$  adversarial attacks. Many works have devised new randomized smoothing schemes for other metrics, such as  $\ell_1$  or  $\ell_\infty$ ; 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  $\ell_1$  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  $\ell_p$ -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](https://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 per-

turbations, 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  $\ell_p$  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  $\ell_2$  (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.<sup>1</sup>

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  $\ell_1$  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  $\ell_1$  radius  $\geq 1.5$ . 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  $\Omega(\min(1, d^{\frac{1}{p}-\frac{1}{2}}))$   $\ell_p$ -radius, if all one uses are the probabilities of labels when classifying randomly perturbed input.

<sup>\*</sup>Equal contribution <sup>1</sup>Microsoft Research AI <sup>2</sup>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 <jerrli@microsoft.com>.

<sup>1</sup>V2 update: we added results using stability training, semi-supervised learning, and ImageNet pre-training. See Table 1.

**Randomized Smoothing of All Shapes and Sizes**

ImageNet	$\ell_1$ Radius	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
	Laplace, <a href="#">Teng et al. (2019)</a> (%)	48	40	31	26	22	19	17	14
	Uniform, Ours (%)	55	49	46	42	37	33	28	25
	+ Stability Training	<b>60</b>	<b>55</b>	<b>51</b>	<b>48</b>	<b>45</b>	<b>43</b>	<b>41</b>	<b>39</b>
CIFAR-10	$\ell_1$ Radius	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
	Laplace, <a href="#">Teng et al. (2019)</a> (%)	61	39	24	16	11	7	4	3
	Uniform, Ours (%)	70	59	51	43	33	27	22	18
	+ Stability Training	70	60	53	47	<b>43</b>	39	35	28
	+ Stability Training, Semi-supervision	<b>74</b>	<b>63</b>	54	<b>48</b>	<b>43</b>	38	34	31
	+ Stability Training, Pre-training	<b>74</b>	62	<b>55</b>	<b>48</b>	<b>43</b>	<b>40</b>	<b>37</b>	<b>33</b>

Table 1. Certified top-1 accuracies of our  $\ell_1$ -robust classifiers, vs previous state-of-the-art, at various radii, for ImageNet and CIFAR-10.<sup>3</sup>

## 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.

<sup>3</sup>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.

**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  $\ell_2$  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  $\ell_1$ -bounded ([Teng et al., 2019](#)),  $\ell_\infty$ -bounded ([Zhang\\* et al., 2020](#)),  $\ell_0$ -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  $\ell_\infty$ , then the  $\ell_2$  norm of the noise must be large on average. While this

provides indirect evidence for the hardness of certifying  $\ell_\infty$ , 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  $\ell_p$  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  $\ell_p$  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  $\ell_\infty$  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.,

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

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*

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

One can think of  $U$  as 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  $\ell_2$  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 <sup>4</sup>

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

### 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*

$$\mathcal{N}\mathcal{P}_\kappa \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : \kappa q(x - v) \geq q(x)\}. \quad (4)$$

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

$$\mathcal{G}_q(q(\mathcal{N}\mathcal{P}_\kappa), v) = q(\mathcal{N}\mathcal{P}_\kappa - v). \quad (\text{NP})$$

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  $\mathcal{N}\mathcal{P}_\kappa$  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

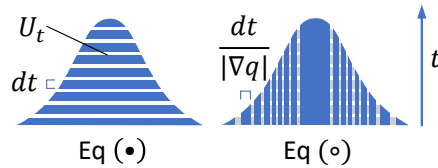
$$U_t \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : q(x) \geq t\}.$$

Then its boundary  $\partial U_t$  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:

$$1 = \int q(x) dx = \int_0^\infty \text{Vol}(U_t) dt. \quad (\bullet)$$

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

$$1 = \int_0^\infty \int_{\partial U_t} \frac{t}{\|\nabla q(x)\|_2} dx dt. \quad (\circ)$$



<sup>4</sup>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

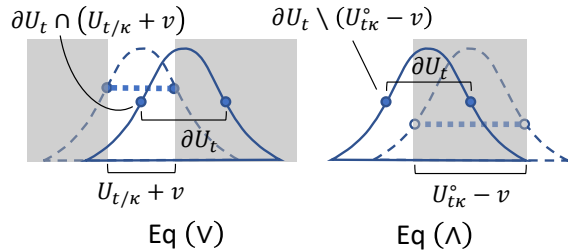
$$q(\mathcal{NP}_\kappa) = \int_0^\infty \int_{\partial U_t \cap (U_{t/\kappa} + v)} \frac{t}{\|\nabla q(x)\|_2} dx dt. \quad (\vee)$$

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

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



The graphics above illustrate the integration domains of  $x$  in Eqs. (V) and (A). 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. (V) and (A) can be calculated efficiently.

**Computing Robust Radius** Eqs. (V) and (A) 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  $\ell_2$  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 Against  $\ell_2$  Adversary

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

**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

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

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*

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

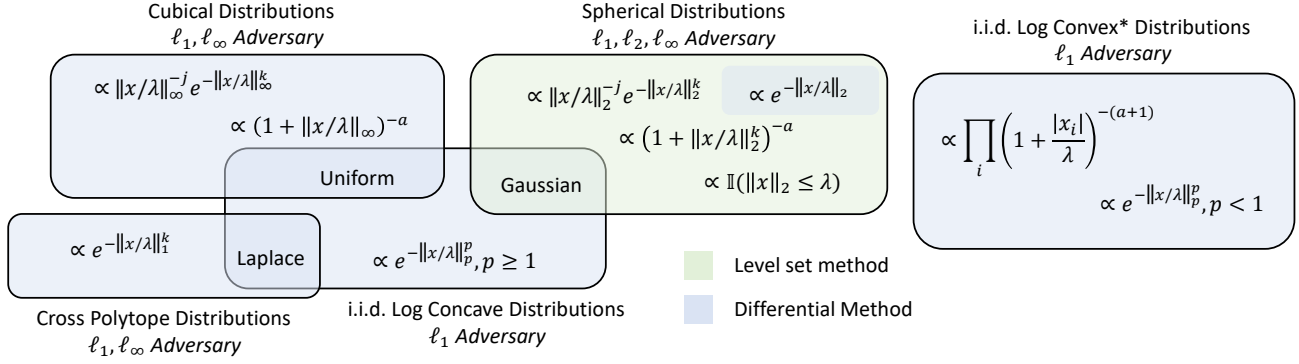
where  $\rho$  is the probability that the base classifier predicts the right label under random perturbation by  $q$ .

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  $\ell_1$  adversary is at least

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

when  $\rho$  is the probability of the correct class as in Theorem 4.1.



**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  $\ell_p$  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  $\lambda$ , we WLOG assume  $\lambda = 1$ . By [Theorem 4.1](#) and the monotonicity of  $\Phi$ , it suffices to show that  $\Phi(p) = 1/2d$  for  $p \geq 1/2d$ . For any fixed  $U$  with  $q(U) = p$ ,

$$\begin{aligned} \lim_{r \searrow 0} \frac{q(U - rv) - p}{r} &= \frac{d}{dr} \int_U q(x - rv) dx \Big|_{r=0} \\ &= \int_U \langle v, \nabla q(x) \rangle dx. \end{aligned}$$

Note  $\nabla q(x) = e_x q(x)$ , where  $e_x = \text{sgn}(x_{i^*}) e_{i^*}$ ,  $e_i$  is the  $i$ th unit vector, and  $i^* = \arg\max_i |x_i|$ . Additionally, the above integral is linear in  $v$ , so the supremum over  $\|v\|_1 = 1$  is achieved on one of the vertices of the  $\ell_1$  ball. So we may WLOG consider only  $v = \pm e_i$ ; furthermore, due to symmetry of  $\nabla q(x)$ , we can just assume  $v = e_1$ :

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

where  $U$  ranges over all  $q(U) = p$ . Note  $\langle e_1, e_x \rangle = 0$  if  $i^* \neq 1$ , and  $\text{sgn}(x_{i^*})$  otherwise. Thus, to maximize  $\lim_{r \searrow 0} \frac{q(U - r e_1) - p}{r}$  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.  $\square$

### 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  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be convex and even, and let  $\text{CDF}_\phi^{-1}$  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  $\rho$  is the probability of the correct class, then the robust radius in  $\ell_1$  is

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

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  $\ell_1$  adversary.

**Example 4.4 (Appendices I.2.1 and I.3.1).** Facing an  $\ell_1$  adversary, cubical distributions, like that in [Example 4.2](#), typically enjoy, via the differential method,  $\ell_1$  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  $\rho$ ; 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.

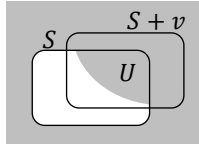
## 5. Wulff Crystals

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,  $\text{Vol}(S) = 1$ . Then for any  $v \in \mathbb{R}^d$  and any  $p \in [0, 1]$ ,

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

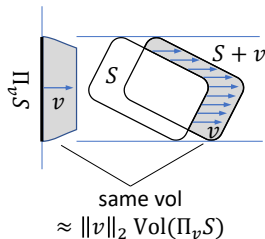
This can be seen easily by taking  $U$  in Eq. (2) to be a subset of  $(S + v) \cap S$  with volume  $p$  (or any set of volume  $p$  containing  $(S + v) \cap S$  if  $p \geq \text{Vol}((S + v) \cap S)$ ) unioned with the complement of  $S$ . For example, in the figure here,  $U$  would be the gray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

$$\begin{aligned} \lim_{r \rightarrow 0} \frac{\mathcal{G}_q(p, rv) - p}{r} &= \lim_{r \rightarrow 0} \frac{\text{Vol}((S + rv) \setminus S)}{r} \\ &= \|v\|_2 \text{Vol}(\Pi_v S) \end{aligned} \quad (6)$$

where  $\Pi_v S$  is the projection of  $S$  along the direction  $v/\|v\|_2$ , and  $\text{Vol}(\Pi_v S)$  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  $\frac{1}{\|v\|_2 \text{Vol}(\Pi_v S)^{\frac{1}{2}-p}} v$  when  $1/2 - p$  is small, and  $p$  is the probability the base classifier  $f$  misclassifies  $x + \delta$ ,  $\delta \sim q$ . Thus, for  $r$  small, we have

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

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

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

If  $\mathcal{B}$  is the  $\ell_p$  ball, the reader might guess  $S$  should either be the  $\ell_p$  ball or the  $\ell_r$  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  $\|x\|_* = \mathbb{E}_{y \sim \text{Vert}(\mathcal{B})} |\langle x, y \rangle|$  and  $y$  is sampled uniformly from the vertices of  $\mathcal{B}$ <sup>5</sup>.

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

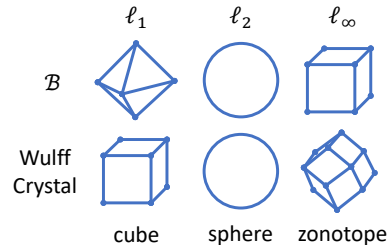
**Theorem 5.2 (Theorem G.7, informal).** *The Wulff Crystal w.r.t.  $\mathcal{B}$  minimizes*

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

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  $\text{Vert}(\mathcal{B})$  is a finite set, the Wulff Crystal has an elegant description as the *zonotope* of  $\text{Vert}(\mathcal{B})$ , 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.  $\ell_2$  ball is the  $\ell_2$  ball itself. The Wulff Crystal w.r.t.  $\ell_1$  ball is a cube ( $\ell_\infty$  ball). The Wulff Crystal w.r.t.  $\ell_\infty$  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  $\{\pm 1\}^d$ .



<sup>5</sup>When  $\mathcal{B}$  is the  $\ell_2$  ball,  $\text{Vert}(\mathcal{B})$  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”<sup>6</sup> 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  $\epsilon$ . 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  $\rho$  away from  $1/2$ , we find the Wulff Crystal distributions empirically to be highly effective, as we describe next in Section 6.

## 6. Experiments

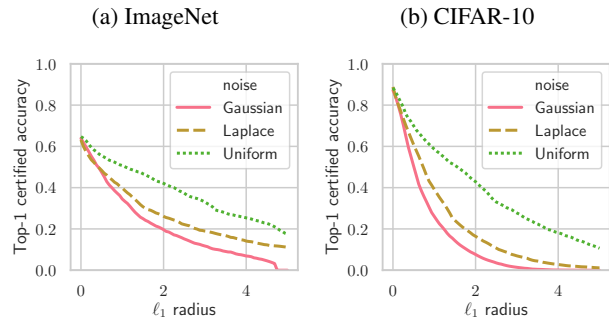
We empirically study the performance of different smoothing distributions on image classification datasets, using the bounds derived via the level set or the differential method, and verify predictions made by the Wulff Crystal theory. We follow the experimental procedure in Cohen et al. (2019) and further 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  $\epsilon$  is defined as the fraction of the test set for which the smoothed classifier  $g$  correctly classifies and certifies robust at an  $\ell_p$  radius of  $\epsilon$ . 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  $\lambda$  (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  $\epsilon$ . Finally, in our plots, we present, for each distribution, the upper envelopes of certified accuracies attained over the range of considered  $\sigma$ . 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

<sup>6</sup>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 2. **SOTA  $\ell_1$  Certified Accuracies**. Certified  $\ell_1$  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  $\ell_1$  adversarial perturbation radius  $\epsilon$  we report the best certified accuracy.



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.

### 6.1. $\ell_1$ Adversary

As previously mentioned, the Wulff Crystal for the  $\ell_1$  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  $\ell_1$  (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.

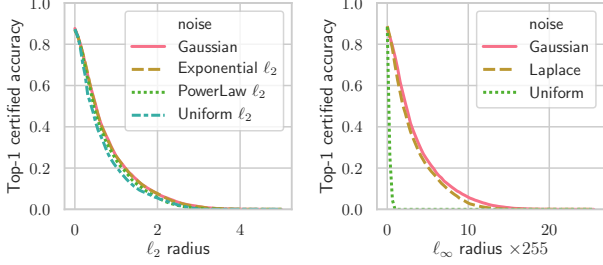
We obtain state-of-the-art robust certificates for ImageNet and CIFAR-10, finding that the Uniform distribution performs best, significantly better than the Gaussian and Laplace distributions (Table 1, Fig. 2). The other distributions with cubic level sets match but do not exceed the performance of Uniform distribution, after sweeping hyperparameters. This verifies that distributions with cubical level sets are significantly better for  $\ell_1$  certified accuracy than those with spherical or cross-polytope level sets. See results for other distributions in Appendix C.

### 6.2. $\ell_2$ Adversary

The Wulff Crystal w.r.t. the  $\ell_2$  ball is a sphere, so we explore distributions with spherical level sets (Table A.1):

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

Figure 3. CIFAR-10 certified accuracies for  $\ell_2$  (left) and  $\ell_\infty$  (right) adversaries. For each distribution  $q$  we train models across a range of  $\sigma^2 \stackrel{\text{def}}{=} \mathbb{E}[\frac{1}{d} \|\delta\|_2^2]$ , and at each level of  $\ell_p$  adversarial perturbation radius  $\epsilon$ , we pick the model that maximizes certified accuracy.



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

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

### 6.3. $\ell_\infty$ Adversary

The Wulff Crystal for the  $\ell_\infty$  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  $\Phi$  (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  $\ell_\infty$  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  $\ell_\infty$  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 with its decision region), the smoothed classifier outputs  $\text{sgn}(\rho - 1/2)$  where  $\rho = q(U - x)$  is the ‘‘confidence’’ of this prediction (Eq. (1)). Randomized smoothing (via Neyman-Pearson) tells us that, if  $\rho$  is large enough, then, no matter what  $U$  is, a small perturbation of  $x$  cannot decrease  $\rho$  too much to change  $\text{sgn}(\rho - 1/2)$  (Eq. (3)).

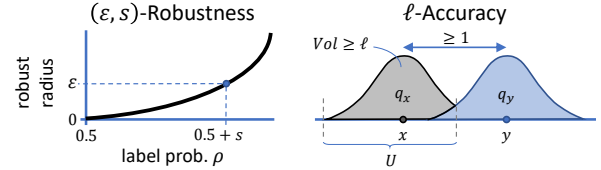
If all we care about is robustness, then the optimal strategy would set  $q$  to be an arbitrarily wide distribution (say, e.g. a wide Gaussian), and the resulting smoothed classifier is roughly constant. Of course, such a smoothed classifier can never achieve good clean accuracy, so it is not useful. 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  $\ell_p$  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  $(\epsilon, s, \ell)$ -useful smoothing with respect to  $\|\cdot\|$  if:

1.  **$(\epsilon, s)$ -Robustness** For all  $x, y$  with  $\|x - y\| < \epsilon$ , 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.  **$\ell$ -Accuracy** For all  $x, y$  with  $\|x - y\| \geq \ell$ , 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.  $(1/2 + s, \epsilon)$  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  $\epsilon$ , for every fixed  $s$ . In particular, observe that for  $\ell_2$ , Gaussian smoothing achieves dimension-independent  $\epsilon$ , 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  $\ell$ , any  $(\epsilon, s, \ell)$ -useful smoothing scheme must have  $\epsilon = o(1)$  for a number of norms, including  $\ell_\infty$ . 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  $2 \leq p \leq \infty$  if there exists  $C$  such that for all finite sequences  $x_1, \dots, x_n \in X$ , we have

$$\mathbb{E} \left[ \left\| \sum_{j=1}^n \sigma_j x_j \right\|^2 \right] \geq C^{-1} \left( \sum_{j=1}^n \|x_j\|^p \right)^{2/p},$$

where the  $\sigma_j$  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  $C_2(T) \leq O(d^{1/2})$ . Because  $C_2$  lower bounds the distortion of a metric embedding of  $T$ , by the aforementioned connection with randomized smoothing,  $C_2$  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  $(\varepsilon, s, \ell)$ -useful smoothing scheme for  $T$  with  $s/\ell < c$  must have*

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

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 immediate corollary, we get:

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

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

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  $\ell_2$ -robust classification problems, many anticipated that this can be extended to  $\ell_\infty$  as well. One might also hope that, even though it seems like we cannot certify  $\ell_2$  radius that grows with input dimension, we could do so for  $\ell_1$ . 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  $\rho$  of correctly classifying an input under random noise, no smoothing techniques can certify nontrivial robust accuracy at  $\ell_\infty$  radius  $\Omega(d^{-1/2})$ , or at  $\ell_2$  or  $\ell_1$  radius  $\Omega(1)$ .*

Indeed, the  $\ell_1$ -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  $\rho$ . It would be interesting to see whether this technique can be scaled up, and whether other methods can leverage more information<sup>7</sup>.

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  $\ell_p$  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  $\ell_1$  provable defense. At the same time, we have showed the limit current techniques face against  $\ell_p$  adversaries when  $p > 2$ , especially  $\ell_\infty$ . 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.

<sup>7</sup>[Lee et al. \(2019\)](#) also used the decision tree structure of their base classifier to improve  $\ell_0$  certification, but the  $\ell_0$ -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  $\ell_\infty$  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’Été 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 Down-sampled 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. Über einen Gometrischen Satz von Wulff für die Gleichgewichtsform von Kristallen. *Z. Kristallog*, 105:304, 1944.
- Dvijotham, K., Goyal, 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., Goyal, 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., Goyal, 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 Garipey, 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., Drachler-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 über 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. 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=HyGIIdiRqtm>.
- 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 Gleichgewichtsform 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 Kristallographie 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=Sk8gJBFvr>.

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>.

## A. Table of Robust Radii

Distribution	Density	Adv.	Certified radius	Reference
iid Log Concave	$\propto e^{-\sum_i \phi(x_i)}$	$\ell_1$	$\text{CDF}_{\phi}^{-1}(\rho)$	Theorem I.1
iid Log Convex*	$\propto e^{-\sum_i \phi( x_i )}$	$\ell_1$	$\int_{\varphi^{-1}(1-\rho)}^{\infty} \frac{1}{e^{\phi(c)-\phi(0)}-1} dc$	for $\varphi$ , see $\rightarrow$ Theorem I.3
Exp. $\ell_p, p \geq 1$	$\propto e^{-\ \frac{x}{\lambda}\ _p^p}$	$\ell_1$	$\lambda \sqrt[p]{\text{GammaCDF}^{-1}(2\rho-1; 1/p)}$	Corollary I.4
Exp. $\ell_p, p < 1$	$\propto e^{-\ \frac{x}{\lambda}\ _p^p}$	$\ell_1$	$\lambda \int_{\varphi^{-1}(1-\rho)}^{\infty} \frac{dc}{e^{c^p}-1}$	for $\varphi$ , see $\rightarrow$ Corollary I.5
Gaussian	$\propto e^{-\ \frac{x}{\lambda}\ _2^2/2}$	$\ell_2$	$\lambda \text{GaussianCDF}^{-1}(\rho; 0, 1)$	Theorem F.10 <sup>C</sup>
		$\ell_1$	$\lambda \text{GaussianCDF}^{-1}(\rho; 0, 1)$	Symmetry
		$\ell_{\infty}$	$\lambda \text{GaussianCDF}^{-1}(\rho; 0, 1)/\sqrt{d}$	Symmetry
Laplace	$\propto e^{-\ \frac{x}{\lambda}\ _1}$	$\ell_1$	$-\lambda \log(2(1-\rho))$	Theorem F.11 <sup>T</sup>
		$\ell_{\infty}$	$\approx \lambda \text{GaussianCDF}^{-1}(\rho; 0, 1)/\sqrt{d}$	see $\rightarrow$ Theorem I.15
Exp. $\ell_{\infty}$	$\propto e^{-\ \frac{x}{\lambda}\ _{\infty}}$	$\ell_1$	$2d\lambda(\rho - \frac{1}{2})$	Theorem I.6
		$\ell_{\infty}$	$\lambda \log \frac{1}{2(1-\rho)}$	Theorem I.9
Exp. $\ell_2$	$\propto e^{-\ \frac{x}{\lambda}\ _2}$	$\ell_2$	$\lambda(d-1) \operatorname{arctanh}(1 - 2\beta^{-1}(1-\rho; \frac{d-1}{2}, \frac{d-1}{2}))$	Theorem I.18
		$\ell_1$	$\lambda(d-1) \operatorname{arctanh}(1 - 2\beta^{-1}(1-\rho; \frac{d-1}{2}, \frac{d-1}{2}))$	Symmetry
		$\ell_{\infty}$	$\frac{\lambda(d-1)}{\sqrt{d}} \operatorname{arctanh}(1 - 2\beta^{-1}(1-\rho; \frac{d-1}{2}, \frac{d-1}{2}))$	Symmetry
Uniform $\ell_{\infty}$	$\propto \mathbb{I}(\ x\ _{\infty} \leq \lambda)$	$\ell_1$	$2\lambda(\rho - \frac{1}{2})$	Theorem I.8 <sup>L</sup>
		$\ell_{\infty}$	$2\lambda(1 - \sqrt{\frac{3}{2}} - \rho)$	Theorem I.11 <sup>L</sup>
Uniform $\ell_2$	$\propto \mathbb{I}(\ x\ _2 \leq \lambda)$	$\ell_2$	$\lambda(2 - 4\beta^{-1}(\frac{3}{4} - \frac{\rho}{2}; \frac{d+1}{2}, \frac{d+1}{2}))$	Theorem I.19
		$\ell_1$	$\lambda(2 - 4\beta^{-1}(\frac{3}{4} - \frac{\rho}{2}; \frac{d+1}{2}, \frac{d+1}{2}))$	Symmetry
		$\ell_{\infty}$	$\frac{\lambda}{\sqrt{d}}(2 - 4\beta^{-1}(\frac{3}{4} - \frac{\rho}{2}; \frac{d+1}{2}, \frac{d+1}{2}))$	Symmetry
General Exp. $\ell_{\infty}$	$\propto \ \frac{x}{\lambda}\ _{\infty}^{-j} e^{-\ \frac{x}{\lambda}\ _{\infty}^k}$	$\ell_1$	$\frac{2d\lambda}{d-1} \Gamma(\frac{d-j}{k}) / \Gamma(\frac{d-1-j}{k})(\rho - \frac{1}{2})$	Theorem I.7
		$\ell_{\infty}$	$\lambda \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp$	for $\Phi$ , see $\rightarrow$ Theorem I.10
General Exp. $\ell_2$	$\propto \ \frac{x}{\lambda}\ _2^{-j} e^{-\ \frac{x}{\lambda}\ _2^k}$	$\ell_2$	<i>level set method</i>	Example I.21
General Exp. $\ell_1$	$\propto e^{-\ \frac{x}{\lambda}\ _1^k}$	$\ell_1$	$\lambda \int_{1-\rho}^{1/2} \frac{R}{\Psi(p)} dp$	for $R, \Psi$ , see $\rightarrow$ Theorem I.14
		$\ell_{\infty}$	$\lambda \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp$	for $\Phi$ , see $\rightarrow$ Theorem I.16
Power Law $\ell_{\infty}$	$\propto \frac{1}{(1+\ \frac{x}{\lambda}\ _{\infty})^a}$	$\ell_1$	$\frac{2d\lambda}{a-d}(\rho - \frac{1}{2})$	Theorem I.12
		$\ell_{\infty}$	$\frac{2\lambda}{a-d} \int_{1-\rho}^{1/2} \frac{dp}{\Upsilon(\Upsilon^{-1}(2p; d, a-d); d, a+1-d)}$	Theorem I.13
Power Law $\ell_2$	$\propto \frac{1}{(1+\ \frac{x}{\lambda}\ _2)^a}$	$\ell_2$	<i>level set method</i>	Example I.22
Pareto (i.i.d.)	$\propto \frac{1}{\prod_i (1+\frac{ x_i }{\lambda})^{a+1}}$	$\ell_1$	$\lambda \frac{2\rho-1}{a} {}_2F_1\left(1, \frac{a}{a+1}, \frac{2a+1}{a+1}; (2\rho-1)^{1+1/a}\right)$	Theorem I.17

Table A.1. Distributions we derive robust radii for and assess experimentally. Here  $\rho$  is the probability the base classifier answers correctly when input is perturbed by the smoothing noise,  $d$  is the dimensionality of the noise,  $\text{CDF}_{\phi}^{-1}$  is the inverse CDF of the 1D random variable with density  $\propto e^{-\phi(x)}$ ,  $\beta^{-1}(\cdot; a, b)$  is the inverse Beta CDF function with shape parameters  $a$  and  $b$ ,  $\Upsilon(\cdot; a, b)$  (resp.  $\Upsilon^{-1}(\cdot; a, b)$ ) is the Beta Prime (resp. inverse) CDF function with shape parameters  $a$  and  $b$ ,  $\Gamma$  is the Gamma function, and  ${}_2F_1$  is the Gaussian hypergeometric function. Under *Reference*, superscript <sup>C</sup> refers to Cohen et al. (2019), superscript <sup>L</sup> refers to Lee et al. (2019), and superscript <sup>T</sup> refers to Teng et al. (2019).

Distribution $q$	Density	$\lambda/\sigma = \lambda/\sqrt{\mathbb{E}_{\delta \sim q} \frac{1}{d} \ \delta\ ^2}$
Exp. $\ell_p$	$\propto e^{-\ \frac{x}{\lambda}\ _p^p}$	$\sqrt{\frac{\Gamma(1/p)}{\Gamma(3/p)}}$
Gaussian	$\propto e^{-\ \frac{x}{\lambda}\ _2^2/2}$	1
Laplace	$\propto e^{-\ \frac{x}{\lambda}\ _1}$	$1/\sqrt{2}$
Exp. $\ell_\infty$	$\propto e^{-\ \frac{x}{\lambda}\ _\infty}$	$\sqrt{\frac{1}{(d+1)((d-1)/3+1)}}$
Exp. $\ell_2$	$\propto e^{-\ \frac{x}{\lambda}\ _2}$	$\sqrt{\frac{1}{d+1}}$
Uniform $\ell_\infty$	$\propto \mathbb{I}(\ x\ _\infty \leq \lambda)$	$\sqrt{3}$
Uniform $\ell_2$	$\propto \mathbb{I}(\ x\ _2 \leq \lambda)$	$\sqrt{\frac{1}{d+2}}$
General Exp. $\ell_\infty$	$\propto \ \frac{x}{\lambda}\ _\infty^{-j} e^{-\ \frac{x}{\lambda}\ _\infty^k}$	$\sqrt{\frac{d\Gamma(\frac{d-j}{k})}{((d-1)/3+1)\Gamma(\frac{d+2-j}{k})}}$
General Exp. $\ell_2$	$\propto \ \frac{x}{\lambda}\ _2^{-j} e^{-\ \frac{x}{\lambda}\ _2^k}$	$\sqrt{\frac{d\Gamma(\frac{d-j}{k})}{\Gamma(\frac{d+2-j}{k})}}$
General Exp. $\ell_1$	$\propto e^{-\ \frac{x}{\lambda}\ _1^k}$	$\sqrt{\frac{d(d+1)\Gamma(\frac{d}{k})}{2\Gamma(\frac{d+2}{k})}}$
Power Law $\ell_\infty$	$\propto \frac{1}{(1+\ \frac{x}{\lambda}\ _\infty)^a}$	$\sqrt{\frac{(a-d-1)(a-d-2)}{(d+1)((d-1)/3+1)}}$
Power Law $\ell_2$	$\propto \frac{1}{(1+\ \frac{x}{\lambda}\ _2^k)^a}$	$\sqrt{\frac{\Gamma(\frac{d+2}{k})\Gamma(a-\frac{d+2}{k})}{d\Gamma(\frac{d}{k})\Gamma(a-\frac{d}{k})}}$
Pareto (i.i.d.)	$\propto \frac{1}{\prod_i (1+\frac{ x_i }{\lambda})^{a+1}}$	$\sqrt{\frac{1}{2}(a-1)(a-2)}$

Table A.2. Relation between the scale parameter  $\lambda$  and the variance  $\sigma^2 = \mathbb{E}_{\delta \sim q} \frac{1}{d} \|\delta\|^2$  of each distribution. This table is used to choose the correct  $\lambda$  to match  $\sigma$  across different distributions. All quantities can be computed easily using Lemmas I.25 and I.26. They are also tested to be numerically correct in the test suite of our code base [github.com/tonyduan/rs4a](https://github.com/tonyduan/rs4a).

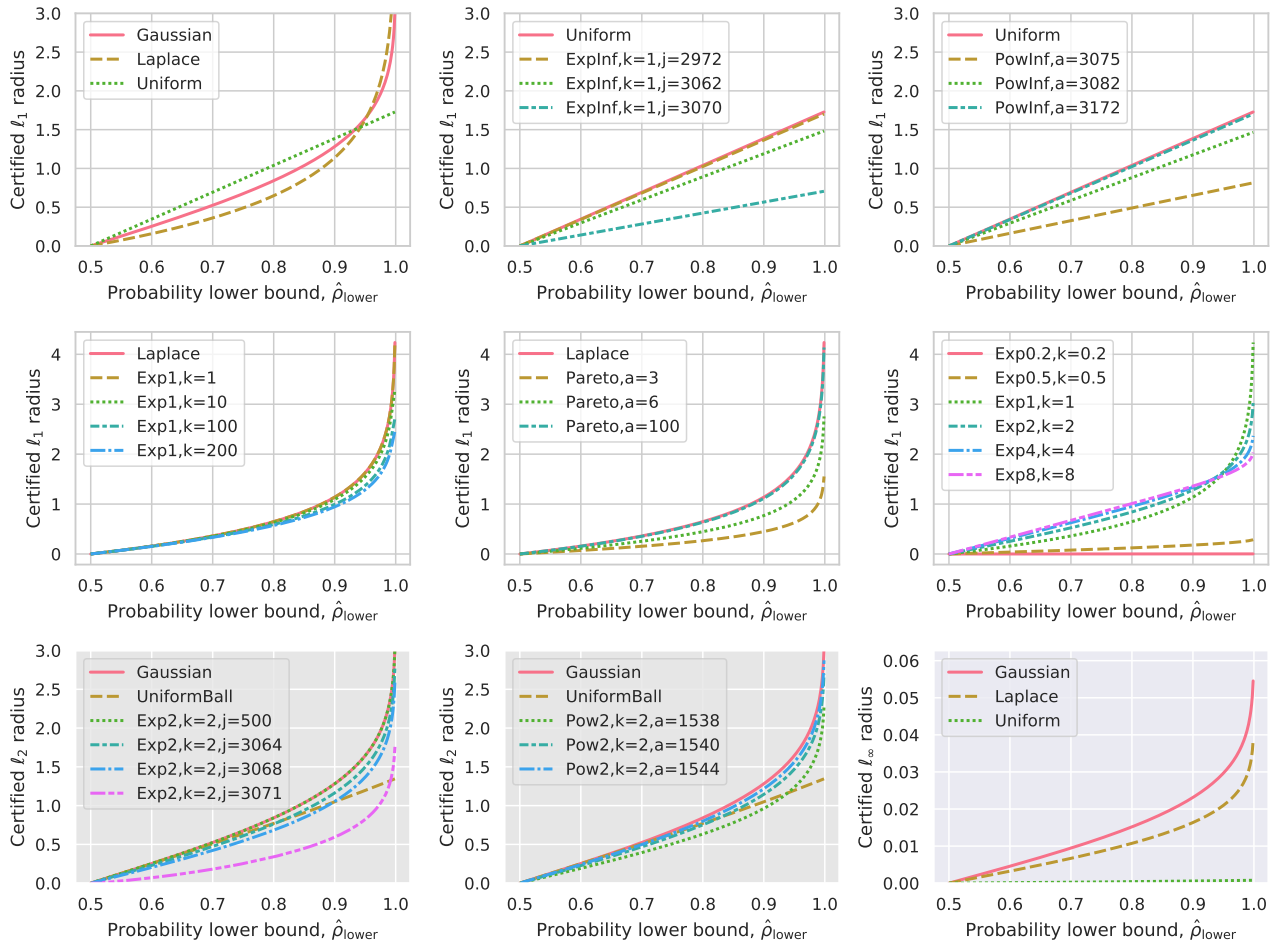


Figure A.1. Certified robust radii of a selection of the distributions in Table A.1, with input dimension  $d = 3072$  and normalized variance  $\sigma^2 = 1$ , across a range of  $\hat{\rho}_{\text{lower}}$ , the high probability lower bound of  $\rho$  (the probability that the base classifier answers correctly when perturbed by smoothing noise). The first two rows are for the  $\ell_1$  adversary while the last row is for the  $\ell_2$  and  $\ell_\infty$  adversaries.



## B. Analysis of Robust Radii

Here we make a few observations about the robust radii of the distributions studied in this paper.

**Distributions that concentrate around the same level set have similar robust radii.** This is evident, for example, in the top middle subplot of Fig. A.1, where the distribution  $\propto \|x\|_\infty^j e^{-\|x/\lambda\|_\infty}$  with “small”  $j = 2972$  has robust radii almost the same as those of  $\propto e^{-\|x/\lambda\|_\infty}$  (here  $\lambda$  for each distribution is the one that sets  $\sigma = 1$ ), and both distributions concentrate around the sphere of radius  $\sqrt{d}$ . We can also see this in the top right ( $\ell_\infty$ -based power law), middle left ( $\ell_1$ -based exponential law), bottom left ( $\ell_2$ -based exponential law), and bottom middle ( $\ell_2$ -based power law) subplots of Fig. A.1. This is also reflected in the center subplot of Fig. A.1, which shows that Pareto distribution with large power gets the same robust radii as Laplace. The reason is that such a high-power Pareto distribution concentrates around an  $\ell_1$ -ball in high dimension.

We can understand this phenomenon intuitively via the level set method: Two distributions concentrating around the same level set will have Eq. (V) and Eq. (A) evaluate to similar quantities.

**Among distributions concentrated around some level set, the shape of the level set is the biggest determinant of performance.** This is evident in the top left, middle right, and bottom right subplots of Fig. A.1.

**Distributions that don’t concentrate on a level set do worse than those that do.** This is evident, for example, in the top middle subplot of Fig. A.1, where the distribution  $\propto \|x\|_\infty^j e^{-\|x/\lambda\|_\infty}$  with “large”  $j = 3070$  has robust radii much smaller than those of  $\propto e^{-\|x/\lambda\|_\infty}$ . Same thing can be observed in the top right, center, middle right, bottom left, bottom middle subplots of Fig. A.1.

**Introducing a singularity at the origin only reduces robust radii.** The top middle and bottom left subplots of Fig. A.1 illustrate this point. Thus, we see no evidence for the “soap-bubble hypothesis” put forth by Zhang\* et al. (2020); see also Fig. C.8.

**Introducing a fatter tail yields larger robust radii for large  $\hat{\rho}_{\text{lower}}$ , as long as the level set concentration is not affected.** The middle left and bottom left subplots of Fig. A.1 demonstrate this behavior. The robust radii formulas for  $\exp(-\|x\|_\infty)$  (Theorem I.6) and for the uniform distribution (Theorem I.8) also reflect this, as the former has robust radius  $\rightarrow \infty$  as  $\rho \rightarrow 1$ , but the latter has a finite maximal robust radius.

### B.1. Level Set Method vs Differential Method

Here we concretely compare the robust radii obtained from the level set method and those obtained from the differential

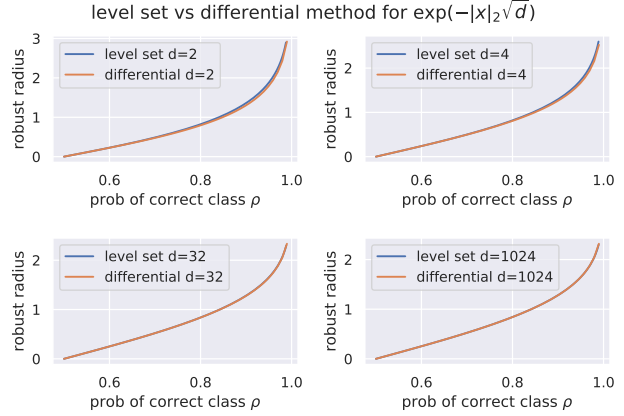


Figure B.1. **Differential Method is Tight** for practical purposes in high dimension  $d$ .

method for the distribution  $\exp(-\|x\|_2\sqrt{d})$ , for various input dimensions  $d$  (we scale the distributions this way so each coordinate has size  $\Theta(1)$ ). For convenience, here’s the robust radius from the differential method (Theorem I.18):

$$R = \frac{d-1}{\sqrt{d}} \operatorname{arctanh} \left( 1 - 2\operatorname{BetaCDF}^{-1} \left( 1 - \rho; \frac{d-1}{2}, \frac{d-1}{2} \right) \right).$$

The robust radii from level set method are computed as in Theorem I.20, and they are tight. As we see in Fig. B.1, the differential method is *very slightly* loose in low dimensions  $d = 2$  and  $4$ , but in high dimensions  $d = 32$  or  $1024$ , the robust radii obtained from both methods are indistinguishable.

### B.2. In-Depth Comparison with Dvijotham et al. (2019)

The information-limited certification algorithm in Dvijotham et al. (2019) relaxes the optimization problem

$$\begin{aligned} \sup_{v \in \mathcal{B}} \mathcal{G}_q(p, v) &= \sup_{q' \in \mathcal{Q}_B} \sup_{U: q(U)=p} q'(U) \\ &\leq \sup_{q' \in \mathcal{D}_F(q)} \sup_{U: q(U)=p} q'(U) \end{aligned} \quad (7)$$

enlarging the set of shifted distributions  $\mathcal{Q}_B \stackrel{\text{def}}{=} \{q(\cdot - v) : v \in \mathcal{B}\}$  to the set of distributions close to  $q$  in several  $f$ -divergences  $\mathcal{D}_F \stackrel{\text{def}}{=} \{q' : \mathcal{D}_f(q' \| q) \leq \epsilon_f, \forall f \in F\}$ , for a set of functions  $F$ . Dvijotham et al. (2019) showed that when  $F$  consists of all Hockey-Stick divergences, Eq. (7) becomes tight, but in practice this is not feasible. In fact, Dvijotham et al. (2019) admits themselves that

*It turns out that the Renyi and KL divergences are computationally attractive for a broad class of smoothing measures, while the Hockey-Stick divergences are theoretically attractive as they lead*

to optimal certificates in the information-limited setting. However, Hockey-Stick divergences are harder to estimate in general, so we only use them for Gaussian smoothing measures.

Concretely, the looseness of their relaxation can be observed when comparing our baseline Laplace results (Table 1) with theirs.

Operationally, their algorithm proceeds as follows

1. For each distribution  $q$  and function  $f$ , manually find the  $f$ -divergence “ball” that contains  $\{q(\cdot - v) : v \in \mathcal{B}\}$ , i.e. compute  $\{\epsilon_f\}_{f \in F}$  such that  $\{q(\cdot - v) : v \in \mathcal{B}\} \subseteq \{q' : \mathcal{D}_f(q' \| q) \leq \epsilon_f, \forall f \in F\}$ .
2. Then they relax the original certification problem to the certification of all  $q'$  close to  $q$  in  $f$ -divergence, i.e. they solve Eq. (7) for the  $\epsilon_f$  found in the previous step.

The 2nd step is a straightforward low-dimensional convex optimization problem, but the trickiness of the 1st step limits the distributions they can apply their technique to. For example, they only know how to do step 1 for  $\exp(-\|x\|_p)$  against  $\ell_p$  adversary, but not against  $\ell_r$  for  $r \neq p$ ; in contrast, our differential method computes robust radii for Laplace against  $\ell_\infty$  perturbation, for example.

## C. Additional Experimental Results

All results in this section are described for CIFAR-10.

**$\ell_1$  Adversary** In addition to the Gaussian, Laplace, and Uniform distributions, we considered an Exponential distribution with cubic level sets, an Exponential distribution with  $\ell_1$  level sets, a power law distribution with cubic level sets, and an i.i.d Pareto distribution.

$$\begin{aligned} q_{\text{Exp}_\infty}(x) &\propto \exp(-(\|x/\lambda\|_\infty)^k) \\ q_{\text{Exp}_1}(x) &\propto \exp(-(\|x/\lambda\|_1)^k) \\ q_{\text{Power}_\infty}(x) &\propto (1 + \|x/\lambda\|_\infty)^{-a} \\ q_{\text{Pareto}}(x) &\propto \prod_i \left(1 + \frac{|x_i|}{\lambda}\right)^{-(a+1)} \end{aligned}$$

Results for these experiments are shown in Fig. C.2. The suffix of the noises in the legend denotes the value of the shape parameter  $k$  or  $a$  that was chosen (whereas we fixed shape parameter  $j = 0$ ). We note that results for distributions with cubical level sets match but do not exceed that of the Uniform distribution. Meanwhile distributions without cubical level sets do not match performance of the Uniform distribution. This suggests that the tail behavior of the noise does not matter as much as the shape of level sets.

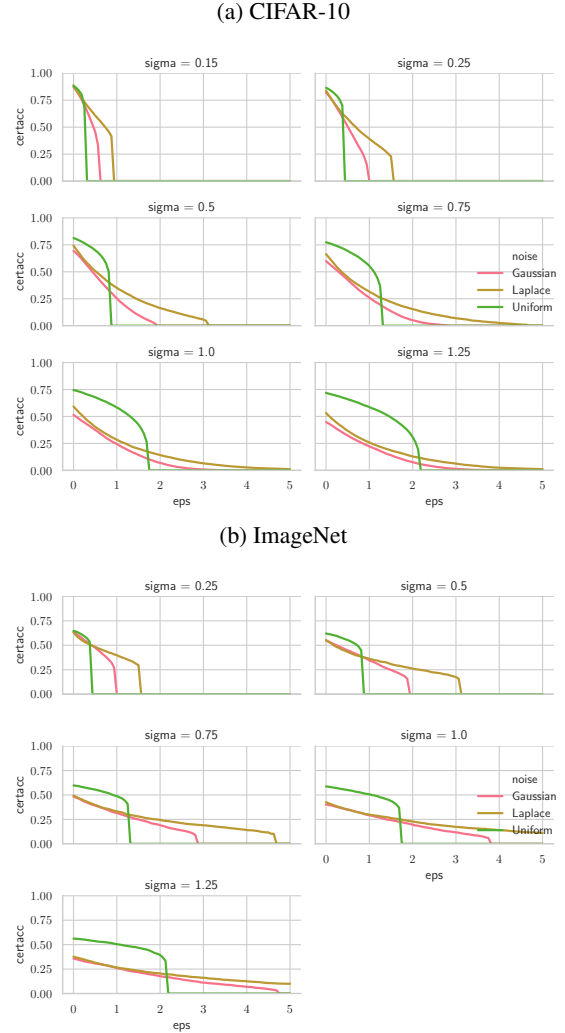
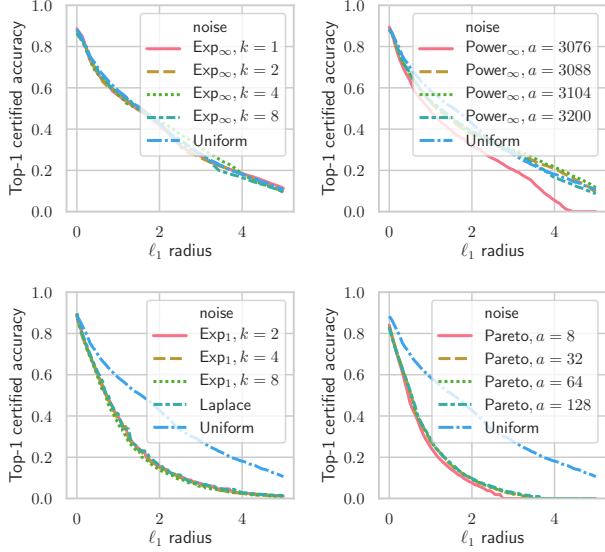


Figure C.1. **Certified Accuracy per  $\sigma$ .** Certified accuracies against an  $\ell_1$  adversary at each level of  $\epsilon$ , across the range of  $\sigma$  with which models were trained (we omit  $\sigma > 1.25$  for brevity). The upper envelope for each distribution is taken to be the maximum certified accuracy across values of  $\sigma$ .

### Ablation of Our $\ell_1$ Improvement over Previous SOTA

To understand how much of our  $\ell_1$  results come from improved certification vs improved training performance, we repeated our Wide ResNet experiments with a multi-layer perceptron (MLP) and AlexNet. We find that the Uniform distribution attains a higher upper envelope of certified accuracy than Gaussian or Laplace with this model (Fig. C.3), but the improvement is less dramatic compared to Table 1 and Fig. 2. Interestingly, the clean (i.e.  $\epsilon = 0$ ) training and testing accuracy of all three distributions are identical when fixed to the same level of  $\sigma$  for the fully-connected model, but for AlexNet, the Uniform noise allows much higher accuracies (Fig. C.4), and for Wide ResNet, even more so. This training improvement leads to substantial improvement in certified accuracies (Fig. C.5).



**Figure C.2. More Distributions for  $\ell_1$  Adversary.** CIFAR-10 certified top-1 accuracies of against the  $\ell_1$  adversary, on generalized exponential law (with  $\ell_\infty$  and  $\ell_1$  level sets), power law (with  $\ell_\infty$  level sets), and Pareto distributions. After appropriate hyperparameter search ( $k$  or  $a$ ), distributions with cubic level sets achieve performance roughly matching that of the Uniform distribution.

As an additional visualization, when we plot the certified accuracy at fixed  $\epsilon$ s versus the training accuracy of a Wide ResNet on noise-augmented CIFAR-10, the Uniform distribution can be seen to significantly outperform the Gaussian and Laplace distributions at all training accuracies except those very close to 1 (Fig. C.6).

So while some of the improvement in certified accuracy in Fig. 2 is due to improved certified radius per  $\rho$ , it seems much more of it is due to the difference in how well a classifier trains when smoothed by noise.

**Why Does Uniform Distribution Get Better Training Accuracy?** Here we further investigate why improvement in architecture seems to amplify the advantage of uniform distribution over others, in terms of training accuracy for each level of  $\sigma$ . Letting  $W \in \mathbb{R}^{d,d}$  denote a pre-specified rotation matrix fixed throughout training/testing, we consider:

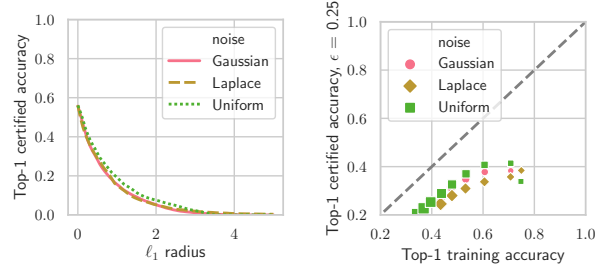
1. Smoothing with unmodified noise, rotated images:

$$x \leftarrow Wx + \delta, \quad \delta \sim q.$$

2. Smoothing with rotated noise, unmodified images:

$$x \leftarrow x + W\delta, \quad \delta \sim q.$$

Note that certification bounds are no longer necessarily applicable, so we only compare clean training accuracy i.e. whether  $\arg \max_y g(x) = y$ . Results for Wide ResNet are shown in Fig. C.7. We find that the difference in training



**Figure C.3. Multi-layer Perceptron.** (Left) CIFAR-10 certified top-1 accuracies for the  $\ell_1$  adversary, with a multi-layer perceptron. (Right) Certified accuracies at  $\ell_1$  perturbation  $\epsilon = 0.25$  plotted against training accuracy under smoothing noise.

performance still exists (but to a lesser degree) under alternative (1), smoothing with unmodified noise but rotated images. On the other hand, we find this difference vanishes under alternative (2), smoothing with rotated noise and unmodified images.

This suggests that the improvement of training accuracy under Uniform noise is due to some *synergy* of the model architecture with the data distribution and the smoothing noise. The choice of Uniform distribution induces some improvement in training accuracy but this is greatly amplified by the interaction between convolution layers and the image dataset. Thus, a good noise for randomized smoothing seems to be one that balances its robustness properties with its *compatibility* with the architecture and the data.

**$\ell_2$  Adversary** In addition to the Gaussian distribution, we considered an Exponential distribution with spherical level sets and a power law distribution with spherical level sets.

$$q_{\text{Exp}_2}(x) \propto (\|x\|_2/\lambda)^{-j} \exp(-(\|x\|_2^2/\lambda))$$

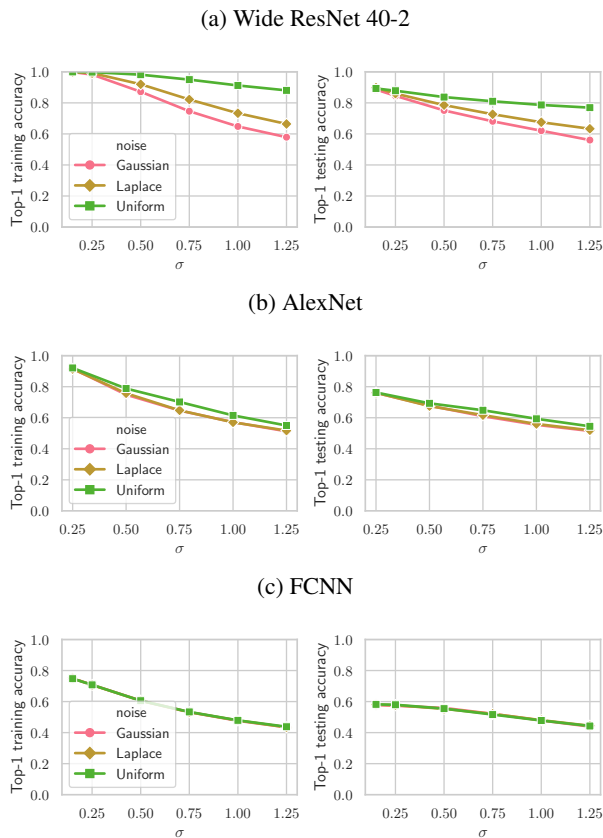
$$q_{\text{Power}_2}(x) \propto (1 + \|x\|_2^2/\lambda)^{-a}$$

Results for these experiments are shown in Fig. C.8. After appropriate hyperparameter search (of  $j$  and  $a$ ), performance for both distributions with spherical level sets matches that of the Gaussian.

**Does Training and Testing on Different Noises Help?**

One may hope that certifying with a different noise than what a model was trained on may improve performance of the classifier. For example,  $\text{Exp}_2$  noise with large  $j$  has more mass concentrated around zero compared to Gaussian noise, and may therefore be easier to “de-noise”. In this section we find that training and testing with different noises does *not* improve clean accuracy, when we compare noises at a fixed level of  $\sigma = 0.5$  (Figure Fig. C.9). For all the noises we considered, testing a model with the same noise it was trained upon results in the best clean accuracy. This suggests the classifier’s de-noising process is quite reliant on the properties of the noise to which it is exposed in the training process.

**Figure C.4. Effect of Architecture.** Clean CIFAR-10 training (left) and testing (right) accuracies for Wide ResNet, AlexNet, and a fully connected neural network, at fixed levels of  $\mathbb{E}[\frac{1}{d}\|\delta\|_2^2] \stackrel{\text{def}}{=} \sigma^2$ . For fixed  $\sigma$ , there is no difference between the distributions when smoothing a fully connected network, but differences arise when the architecture improves to AlexNet and ResNet.



## D. Experimental Details

**Training Methods** There are several methods of training a smoothed classifier. Let  $f$  denote the base classifier (up to the logit layer),  $q$  denote the smoothing distribution, and consider an observation  $(x, y)$ .

1. Noise augmentation as in [Cohen et al. \(2019\)](#),

$$\mathcal{L}(x, y) = -\log f(x + \delta)_y, \quad \delta \sim q.$$

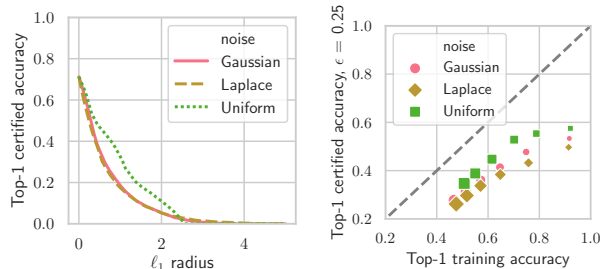
2. Directly training the smoothed classifier as described in [Salman et al. \(2019a\)](#) (without adversarial attacks),

$$\mathcal{L}(x, y) = -\log \mathbb{E}[f(x + \delta)]_y, \quad \delta \sim q.$$

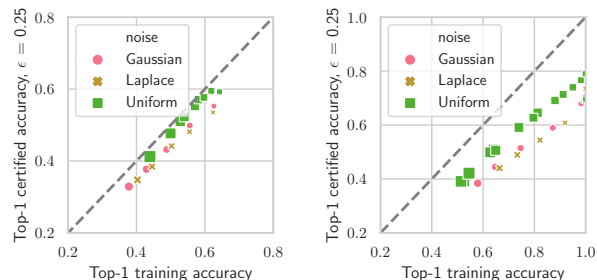
3. Adversarial training as in [Salman et al. \(2019a\)](#),

$$\mathcal{L}(x, y) = -\log \mathbb{E}[f(\tilde{x} + \delta)]_y, \quad \delta \sim q.$$

where  $\tilde{x}$  is found via PGD on the smoothed classifier and  $\delta$  noise samples are fixed throughout the PGD process.



**Figure C.5. AlexNet.** (Left) CIFAR-10 certified top-1 accuracies for the  $\ell_1$  adversary, with an AlexNet architecture. (Right) Certified accuracies at  $\epsilon = 0.25$ , plotted against training accuracy under noise.



**Figure C.6. Training Accuracy vs Certified Accuracy.** Top-1  $\ell_1$  certified accuracies for ImageNet (left) and CIFAR-10 (right) at pre-specified  $\epsilon = 0.25$ , controlling for fixed training accuracy. Larger sized points denote larger  $\sigma$ . Predictably, as  $\sigma$  increases, training and certified accuracy decreases. At fixed training accuracy, the Uniform distribution significantly outperforms Gaussian and Laplace.

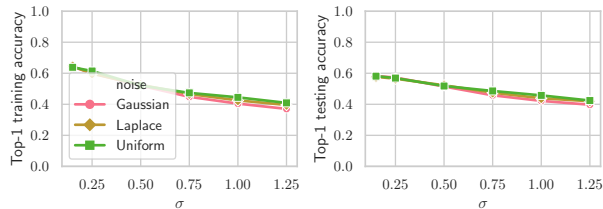
4. Stability training as in [Li et al. \(2019\)](#),

$$\mathcal{L}(x, y) = -\log f(x)_y + \gamma D_{\text{KL}}(\sigma(f(x)) \parallel \sigma(f(x + \delta)))$$

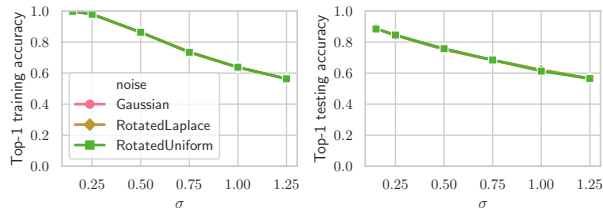
where  $\delta \sim q$ ,  $\sigma$  here denotes the softmax function, and  $\gamma$  is a hyper-parameter.

Unless otherwise noted, in all experiments we trained with the first option, appropriate noise augmentation. We found that direct training was slower and did not yield superior performance in practice. Of these four options we found that stability training with  $\gamma = 6$  tended to produce the best results (our choice of  $\gamma$  follows [Carmon et al. \(2019\)](#)). Therefore, we re-trained our SOTA models with stability training and list results in [Table 1](#) and [Figure D.1](#).

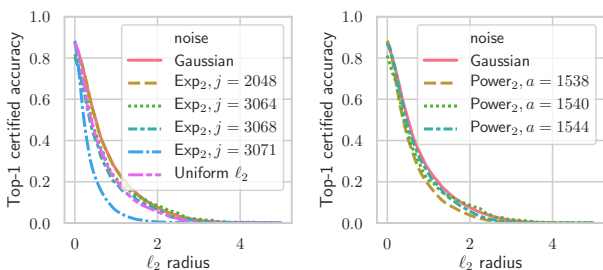
**Range of  $\sigma$**  Recall that  $\sigma^2 \stackrel{\text{def}}{=} \mathbb{E}[\frac{1}{d}\|\delta\|_2^2]$ . This is a fairly consistent measurement of noise level across different noise distributions, and is a natural control variate for comparing the effect (e.g. training, testing, and certified accuracies) of different noises. In addition, to obtain a good estimate of the upper envelope of certified accuracy, we need to take the pointwise maximum of the radius-vs-certified-accuracy



(a) Unmodified noise, rotated images.



(b) Rotated noise, unmodified images.

**Figure C.7. Rotation Experiments.** Wide ResNet clean training/testing accuracies in the two rotation experiments.

**Figure C.8. Distributions with Spherical Level Sets.** CIFAR-10 certified top-1 accuracies against the  $\ell_2$  adversary, on spherical level set exponential and power law distributions. After appropriate hyper-parameter search, performance matches that of the Gaussian distribution.

curve (such as those in Fig. 2) for many  $\sigma$ s. In this work, we swept over:

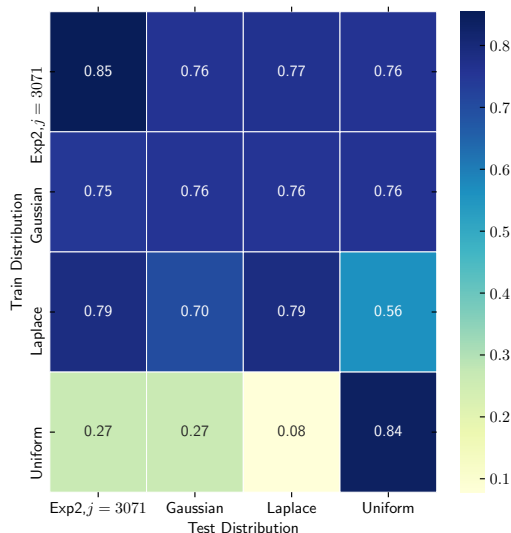
$$\sigma \in \{0.15, 0.25, 0.5, 0.75, 1.0, 1.25, 1.50\}.$$

For distributions with cubic level sets, we needed to sweep over larger  $\sigma$ s as well to estimate the large-radius portion of the upper envelope better:

$$\sigma \in \{1.75, 2.0, 2.25, 2.5, 2.75, 3.0, 3.25, 3.5\}.$$

Table A.2 lists for each distribution the conversion constant needed to obtain  $\lambda$  from  $\sigma = \sqrt{\mathbb{E}_{\delta \sim q} \frac{1}{q} \|\delta\|_2^2}$ .

**Certified Accuracy per  $\sigma$**  In Fig. C.1 we show the certified accuracies of Gaussian, Laplace and Uniform distributions, for each  $\sigma$ , for both ImageNet and CIFAR-10. The upper envelopes reported in the main text are defined as the maximum certified accuracies over  $\sigma$ .


**Figure C.9. Testing on a Different Noise than Trained For.** We compare clean testing accuracies of models (denoted by color) trained on one noise and tested on another, at fixed  $\sigma = 0.5$ . We find a model performs best when tested with the same noise for which it was trained.

**Experiment Hyperparameters** For all experiments we trained with a cosine-annealed learning rate of 0.1, optimized by stochastic gradient descent with momentum of 0.9 and weight decay of 0.0001.

For ImageNet experiments we used a ResNet-50 model and trained with a batch size of 64 for 30 epochs.

For CIFAR-10 experiments we used a Wide ResNet 40-2 model and trained with a batch size of 128 for 120 epochs.

Ablation studies with a fully connected neural network employed two hidden layers of 2048 and 512 nodes followed by ReLU activations, trained with a learning rate of 0.01.

To compute the top categories for certification (which used  $N = 100,000$  samples), we used 64 samples.

Our code is publicly available at: [github.com/tonyduan/rs4a](https://github.com/tonyduan/rs4a)

**More Data for Improved Robustness** We explore using more data to improve the robustness of our SOTA smoothed classifiers for CIFAR-10 in two ways: using *pre-training* as in Hendrycks et al. (2019), and *semi-supervised learning* as in Carmon et al. (2019). Results are listed in Table 1 and Figure D.1.

Pre-training is inspired by Hendrycks et al. (2019), who showed that pre-training on the large downsampled ImageNet dataset can improve empirical  $\ell_\infty$  robustness for

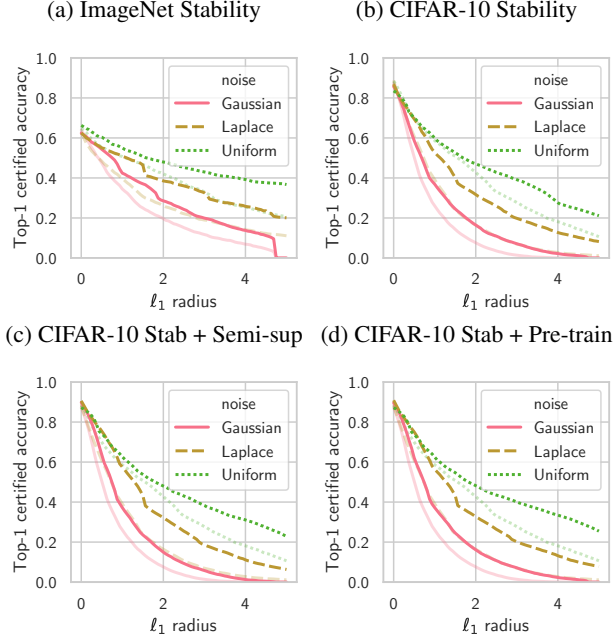


Figure D.1. **Improved Results with Stability Training and More Data.** Certified top-1 accuracies against the  $\ell_1$  adversary, showing improvements due to stability training, pre-training, and semi-supervised self-training. Certified accuracies yielded by the default noise augmentation are plotted in faint lines for ease of comparison.

CIFAR-10 and CIFAR-100 datasets. Similarly, our pre-trained models are initially trained on the 1000-class down-sampled ImageNet dataset (Chrabaszcz et al., 2017). We then re-initialize the final logit layers for the CIFAR-10 dataset and fine-tune with a learning rate of 0.001.

Semi-supervised learning is inspired by Carmon et al. (2019), who showed that self-training on the unlabeled 80 Million Tiny Images dataset can improve robustness of CIFAR-10 classifiers. We use their publicly released dataset of 500k images equipped with pseudo-labels generated by a network trained by CIFAR-10, and train on mini-batches from this dataset and CIFAR-10.

## E. Mathematical Preliminaries

In this section, we rigorously define several mathematical notions and their properties that will recurrent throughout what follows. We will be brief here, but readers can skip this on first reading and refer back when necessary.

**Note about Notation** We will use  $\text{Vol}$  to denote measure, typically Hausdorff measure with the dimension implicit from context. When integrating over a measurable set, the underlying measure is also typically the Hausdorff measure as well. By  $\partial U$  of a set  $U$ , we typically mean *reduced boundary* (when  $U$  has finite perimeter), especially in a

measure-theoretic context. Readers needing more background can consult Evans & Gariepy (2015).

**Sobolev Functions and Regular Functions** While many distributions like Gaussian have continuously differentiable densities, many others, like Laplace, only have “weak” derivatives. Thus, to cover all such distributions, we need to pin down a notion of “weakly differentiable.”

**Definition E.1.** Let  $\Omega \subseteq \mathbb{R}^d$ , and  $f : \Omega \rightarrow \mathbb{R}$ ,  $g : \Omega \rightarrow \mathbb{R}^d$ . We say  $g$  is a *weak derivative* of  $f$  if for every smooth function  $\phi : \Omega \rightarrow \mathbb{R}^d$  with compact support,

$$\int f \operatorname{div} \phi = - \int g \cdot \phi.$$

We write  $g = \nabla f$  in this case.

For any open set  $\Omega \subseteq \mathbb{R}^d$ , the *Sobolev space*  $W^{1,p}(\Omega)$  is defined as the functions  $f \in L^p(\Omega)$  whose weak derivative exists and is in  $L^p(\Omega; \mathbb{R}^d)$ , i.e.

$$W^{1,p}(\Omega) \stackrel{\text{def}}{=} \{f \in L^p(\Omega) : \nabla f \in L^p(\Omega; \mathbb{R}^d)\}.$$

**Definition E.2.** Let  $\Omega \subseteq \mathbb{R}^d$  be an open set. For the purpose of this paper, we say a function  $f : \Omega \rightarrow \mathbb{R}$  is *regular* if  $f \in W^{1,1}(\Omega)$ .

This means that  $f$  has a weak derivative  $\nabla f$  such that both  $f$  and  $\nabla f$  are integrable. For example, ReLU is not a continuously differentiable function, but it is regular since it has the Heavyside step function as its weak derivative.

### Coarea Formula and the Weak Sard’s Theorem

**Theorem E.3** (Coarea Formula (Federer, 2014; Evans & Gariepy, 2015)). Let  $\Omega \subseteq \mathbb{R}^d$  be an open set,  $g \in L^1(\Omega)$ , and  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  regular in the sense of Definition E.2.

Define  $U_t \stackrel{\text{def}}{=} \{x : f(x) \geq t\}$  to be  $f$ ’s superlevel sets. Then

$$\int_{\Omega} g(x) \|\nabla f(x)\|_2 dx = \int_{\mathbb{R}} \int_{\partial U_t} g(x) dx dt.$$

Here the integral in  $x$  in the RHS is over the  $(d-1)$ -dimensional Hausdorff measure of the reduced boundary of  $U_t$ , which we abuse notation and denote as  $\partial U_t$ .

The regularity of  $f$  can be replaced by weaker conditions (see Evans & Gariepy (2015)), but the statement here suffices for our purposes.

By setting  $g$  in Theorem E.3 to be the indicator function over the set where the gradient of  $f$  vanishes, we get

**Theorem E.4** (Weak Sard). For any  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  regular in the sense of Definition E.2, let  $Z \stackrel{\text{def}}{=} \{x \in \mathbb{R}^d : \nabla f(x) = 0\}$ . Let  $U_t \stackrel{\text{def}}{=} \{x : f(x) \geq t\}$  denote the superlevel set of  $f$  at level  $t$ . Then

$$\text{Vol}(Z \cap \partial U_t) = 0$$

for almost every  $t \in \mathbb{R}$ . Here  $\text{Vol}$  denote the Hausdorff measure of dimension  $d-1$ , and again  $\partial U_t$  denotes reduced boundary.

**Sobolev Functions and Absolute Continuity** Recall the standard definition of *absolute continuity*, which can be thought of as a more general notion of *differentiability*.

**Definition E.5.** A function  $f : \mathbb{R} \rightarrow \mathbb{R}$  is called *absolute continuous* if there exists a Lebesgue integrable function  $g : \mathbb{R} \rightarrow \mathbb{R}$  and some  $a \in \mathbb{R}$ , such that

$$f(x) = f(a) + \int_a^x g(t) dt.$$

Such an  $f$  has derivative  $f'$  almost everywhere, and  $f'$  coincides with  $g$  almost everywhere.

Sobolev functions are known to be absolutely continuous on every line, and this property roughly captures all Sobolev functions.

**Theorem E.6** (ACL Property of Sobolev Functions (Nikodym, 1933)). *Let  $\Omega \subseteq \mathbb{R}^d$  be an open set. The following statements hold.*

- *Let  $f : \Omega \rightarrow \mathbb{R}$  be Sobolev,  $f \in W^{1,p}(\Omega)$ . Then possibly after modifying  $f$  on a set of measure 0, for every  $u \in \mathbb{R}^d$ , the function  $t \mapsto f(x + tu)$  is absolutely continuous for almost every  $x$ . Furthermore, the (classical) directional derivative  $D_u f$  is in  $L^p(\Omega)$  for every  $u$ .*
- *Conversely, if the restriction of a function  $f : \Omega \rightarrow \mathbb{R}$  on almost every line parallel to the coordinate axes is absolutely continuous, then pointwise gradient  $\nabla f$  exists almost everywhere, and  $f \in W^{1,p}(\Omega)$  as long as  $f, \nabla f \in L^p(\Omega)$ .*

The ACL property of Sobolev functions yields the differentiability of the convolution of a  $L^\infty$  and a  $W^{1,1}$  function.

**Lemma E.7.** *If a function  $q$  is in  $W^{1,1}(\mathbb{R}^d)$ , then for every bounded measurable  $F : \mathbb{R}^d \rightarrow \mathbb{R}$ , the convolution  $F * q$  is continuously differentiable, and*

$$\nabla(F * q) = F * (\nabla q).$$

*Proof.* A function is differentiable if all of 1) its partial derivatives exist and 2) are continuous. First we show that, for any vector  $u$ ,  $F * (D_u q) = D_u(F * q)$ . We can compute

as follows.

$$\begin{aligned} & F * (D_u q)(x) \\ &= \int F(\hat{x}) D_u q(x - \hat{x}) d\hat{x} \\ &= \frac{d}{d\tau} \int_0^\tau \int F(\hat{x}) D_u q(x - \hat{x} + tu) d\hat{x} dt \Big|_{\tau=0} \end{aligned} \quad (8)$$

$$= \frac{d}{d\tau} \int F(\hat{x}) \int_0^\tau D_u q(x - \hat{x} + tu) dt d\hat{x} \Big|_{\tau=0} \quad (9)$$

$$= \frac{d}{d\tau} \int F(\hat{x}) [q(x - \hat{x} + \tau u) - q(x - \hat{x})] d\hat{x} \Big|_{\tau=0} \quad (10)$$

$$= D_u \int F(\hat{x}) q(x - \hat{x}) d\hat{x} = D_u(F * q)(x). \quad (11)$$

In these equations, first note that

$$\begin{aligned} & \int_0^\tau \int |F(\hat{x}) D_u q(x - \hat{x} + tu)| d\hat{x} dt \\ & \leq \tau \int |D_u q(\hat{x})| d\hat{x} < \infty, \end{aligned}$$

by the ACL property of  $q$  (Theorem E.6). Thus, in Eq. (8), we introduced  $\frac{d}{d\tau} \int_0^\tau$  innocuously by the fundamental theorem of calculus, since the inner integral is absolutely integrable in  $t$ . Then, in Eq. (9), we applied Fubini-Tonelli Theorem to swap the order of integration. In Eq. (10), we integrated out the directional derivative  $D_u q$  for almost every  $\hat{x}$  where  $t \mapsto q(x - \hat{x} + tu)$  is absolutely continuous. Finally, in Eq. (11), we simplified the integral by noting that  $q(x - \hat{x})$  does not depend on  $\tau$ , and  $q(x - \hat{x} + \tau u)$  is absolutely integrable in  $\hat{x}$ . This proves our claim that  $F * (D_u q) = D_u(F * q)$  for any  $u \in \mathbb{R}^d$ .

Note additionally that, since  $D_u q \in L^1$  (by assumption) and  $F \in L^\infty$ , their convolution  $F * (D_u q)$  is bounded and continuous.

Then, taking  $u$  to be the coordinate vectors, we see the partial derivatives of  $F * q$  all exist and are continuous, proving our lemma.  $\square$

## F. The Differential Method

We summarize the setup of this section in the following assumption. Here we use a notion called *regularity* introduced in Definition E.2 that roughly says that a function needs to be continuous almost everywhere and be “weakly” differentiable, and it and its gradient are both absolutely integrable. All concrete density functions we work with in this paper will be regular, with the exception of the uniform distribution.

**Assumption F.1.** *Let  $F : \mathbb{R}^d \rightarrow [0, 1]$  be a measurable function and let  $G : \mathbb{R}^d \rightarrow [0, 1]$  be the smoothing of  $F$  by the distribution  $q(x) \propto \exp(-\psi(x))$  for some  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$ ,*

such that  $q$  is regular in the sense of [Definition E.2](#). Formally,

$$\begin{aligned} G(x) &= \mathbb{E}_{\delta \sim q} F(x + \delta) = \int q(\delta) F(x + \delta) d\delta \\ &= \int q(\hat{x} - x) F(\hat{x}) d\hat{x}. \end{aligned}$$

Consider a norm  $\|\cdot\|$  with unit ball  $\mathcal{B}$  that is a convex body. Let  $\text{Vert}(\mathcal{B})$  be the set of its extremal points.

*Example F.2.* If  $\|\cdot\| = \|\cdot\|_1$  is the  $\ell_1$ -norm, then  $\mathcal{B}$  is what is called the cross-polytope, defined as the convex hull of the unit vectors and their negations. If  $\|\cdot\| = \|\cdot\|_\infty$  is the  $\ell_\infty$ -norm, then  $\mathcal{B}$  is the cube with vertices  $\{\pm 1\}^d$ . If  $\|\cdot\| = \|\cdot\|_2$  is the  $\ell_2$ -norm, then  $\mathcal{B}$  is the unit sphere, and  $\text{Vert}(\mathcal{B})$  is its entire boundary.

*Example F.3.* If  $\psi(x) = \|x\|_2^2$ , then  $q$  is the standard Gaussian distribution. If  $\psi(x) = \|x\|_1$ , then  $q$  is the Laplace distribution.

The following definition of  $\Phi$  turns out to be equivalent to [Eq. \(5\)](#), which will be apparent in the proof of [Theorem F.6](#). It gives a systematic way of computing  $\Phi$ .

**Definition F.4.** Let  $q(x) \propto \exp(-\psi(x))$  be a distribution over  $\mathbb{R}^d$  as in [Assumption F.1](#). For any vector  $u \in \mathbb{R}^d$ , let  $\gamma_u$  be the random variable  $\langle u, \nabla \psi(\delta) \rangle \in \mathbb{R}$  with  $\delta \sim q$ . Define  $\varphi_u$  to be the complementary CDF of  $\gamma_u$ ,

$$\varphi_u(c) \stackrel{\text{def}}{=} \mathbb{P}[\gamma_u > c],$$

and define the inverse complementary CDF  $\varphi_u^{-1}(p)$  of  $\gamma_u$  to be

$$\varphi_u^{-1}(p) \stackrel{\text{def}}{=} \inf\{c : \mathbb{P}[\gamma_u > c] \leq p\}.$$

For any  $p \in [0, 1]$ , define a new random variable  $\gamma_u^{(p)}$  by

$$\gamma_u^{(p)} = \begin{cases} \gamma_u|_{(c, \infty)} & \text{with probability } \varphi_u(c) \\ c & \text{with probability } p - \varphi_u(c) \\ 0 & \text{with probability } 1 - p, \end{cases}$$

where  $c \stackrel{\text{def}}{=} \varphi_u^{-1}(p)$  and  $\gamma_u|_{(c, \infty)}$  is the random variable  $\gamma_u$  conditioned on  $\gamma_u > c$ . Roughly speaking, the PDF of  $\gamma_u^{(p)}$  allocates probability  $p$  to the right portion of  $\gamma_u$ 's PDF, and puts the rest  $1 - p$  probability on 0. One just needs to be careful when  $\gamma_u$ 's measure has a singular point at  $\varphi_u^{-1}(p)$ , which is dealt with in the middle line above.

Let  $\mathcal{B}$  be the unit ball of  $\|\cdot\|$  as in [Assumption F.1](#). Then we define  $\Phi : [0, 1] \rightarrow \mathbb{R}$  by

$$\Phi(p) \stackrel{\text{def}}{=} \max_{u \in \text{Vert}(\mathcal{B})} \mathbb{E} \gamma_u^{(p)}.$$

*Remark F.5.* The function  $p \mapsto \mathbb{E} \gamma_u^{(p)}$  in [Definition F.4](#) is increasing on  $[0, \varphi_u(0)]$  and nonincreasing on  $[\varphi_u(0), 1]$ . Thus  $\Phi(p)$  is also increasing on  $[0, \inf_{u \in \text{Vert}(\mathcal{B})} \varphi_u(0)]$ .

The following theorem is the master theorem for applying the differential method. We illustrate its usage to recover the known Gaussian ([Cohen et al., 2019](#)) and Laplace ([Teng et al., 2019](#)) bounds as warmups in [Appendices F.1](#) and [F.2](#) before applying the technique at scale.

**Theorem F.6** (The Differential Method). *As in [Assumption F.1](#), fix any norm  $\|\cdot\|$  and let  $G : \mathbb{R}^d \rightarrow [0, 1]$  be the smoothing of any measurable  $F : \mathbb{R}^d \rightarrow [0, 1]$  by  $q(x) \propto \exp(-\psi(x))$ , such that  $q$  is regular in the sense of [Definition E.2](#). Let  $\Phi : [0, 1] \rightarrow \mathbb{R}$  be given as in [Definition F.4](#).*

*Then for any  $x$ , if  $G(x) < 1/2$ , then  $G(x + \delta) < 1/2$  for any*

$$\|\delta\| < \int_{G(x)}^{1/2} \frac{1}{\Phi(p)} dp. \quad (\star)$$

In [Theorem F.6](#), one should think of  $G(x)$  as the probability that the smoothed classifier assigns to *any class other than the correct one*. So [Theorem F.6](#) says that, if the smoothed classifier predicts the correct class ( $G(x) < 1/2$ ), then it continues to do so even when the input is perturbed by a noise with magnitude bounded by [Eq. \( \$\star\$ \)](#).

Sometimes, when  $\varphi_u$  is continuous for all  $u$ , for  $p \in [0, 1/2]$ , we can factor

$$\Phi(p) = \bar{\varphi}_u(\varphi_u^{-1}(p)), \quad \text{where } \bar{\varphi}_u(c) \stackrel{\text{def}}{=} \mathbb{E}_{\gamma_u} \mathbb{I}(\gamma_u > c),$$

for some specific  $u \in \text{Vert}(\mathcal{B})$ , either due to symmetry in the vertices of  $\mathcal{B}$  (so that it doesn't matter which  $u$  it is) or because a specific  $u$  maximizes the expression for all  $p \in [0, 1/2]$ . Then the following lemma is very useful for simplifying the integral in [Theorem F.6](#). It can be proved easily using change of coordinates.

**Lemma F.7.** *Suppose  $\Phi(p) = \bar{\varphi}(\varphi^{-1}(p))$  on  $p \in [0, 1/2]$ , where  $\varphi(p)$  is differentiable and both  $\varphi$  and  $\bar{\varphi}$  are nonincreasing. Then for any  $0 \leq p_0 \leq 1/2$ ,*

$$\int_{p_0}^{1/2} \frac{1}{\Phi(p)} dp = \int_{\varphi^{-1}(1/2)}^{\varphi^{-1}(p_0)} \frac{|\varphi'(c)|}{\bar{\varphi}(c)} dc$$

Finally, as mentioned before, the proof of [Theorem F.6](#) will show that

**Proposition F.8.** *The definition of  $\Phi$  in [Definition F.4](#) coincides with the definition [Eq. \(5\)](#) for any smoothing distribution  $q$  with regular density function supported everywhere in  $\mathbb{R}^d$ .*

*Proof of [Theorem F.6](#).* Consider a path  $\xi_t : [0, \|\delta\|] \rightarrow \mathbb{R}^d$  given by  $\xi_0 = x$ ,  $\xi_{\|\delta\|} = x + \delta$ , and  $\xi'_t = d\xi_t/dt = \delta/\|\delta\|$ . We will show

$$dG(\xi_t)/dt \leq \Phi(G(\xi_t))$$



and apply [Lemma F.9](#) to yield the desired result.

By chain rule,

$$dG(\xi_t)/dt = \xi_t' \cdot \nabla G(\xi_t) = \frac{\delta}{\|\delta\|} \cdot \nabla G(\xi_t).$$

To upper bound this quantity, we relax

$$\frac{\delta}{\|\delta\|} \cdot \nabla G(\xi_t) \leq \max_{u \in \mathcal{B}} u \cdot \nabla G(\xi_t) = \max_{u \in \text{Vert}(\mathcal{B})} u \cdot \nabla G(\xi_t)$$

where  $\mathcal{B}$  is the unit ball of the norm  $\|\cdot\|$ , and the equality is because  $u \cdot \nabla G(x)$  is linear in  $u$ , so optima are achieved on vertices. Therefore, it suffices to show that,

$$\forall u \in \text{Vert}(\mathcal{B}), x \in \mathbb{R}^d, \quad u \cdot \nabla G(x) \leq \Phi(G(x)). \quad (12)$$

Below, we let  $x$  be any vector in  $\mathbb{R}^d$  (not just those satisfying  $G(x) \leq 1/2$  as in the theorem statement). In general, for any vector  $u$  and any  $x \in \mathbb{R}^d$ , the directional derivative  $u \cdot \nabla G(x)$  of  $G(x)$  in the direction of  $u$  is given by

$$\begin{aligned} u \cdot \nabla G(x) &= u \cdot \int \nabla_x q(\hat{x} - x) F(\hat{x}) d\hat{x} \\ &= \int \langle u, \nabla \psi(\hat{x} - x) \rangle q(\hat{x} - x) F(\hat{x}) d\hat{x} \end{aligned}$$

where we used [Lemma E.7](#) and the assumption that  $q$  is regular. Then

$$\begin{aligned} u \cdot \nabla G(x) &= \mathbb{E}_{\delta \sim q} F(x + \delta) \langle u, \nabla \psi(\delta) \rangle \\ &\leq \sup_{\hat{F}: \hat{G}(x) = G(x)} \mathbb{E}_{\delta \sim q} \hat{F}(x + \delta) \langle u, \nabla \psi(\delta) \rangle, \end{aligned}$$

where we vary over all  $\hat{F}: \mathbb{R}^d \rightarrow [0, 1]$  such that its smoothing  $\hat{G}$  has the same value as  $G$  at  $x$ . While at first glance, this seems like a unwieldy quantity to maximize, there's a simple intuition to find the maximizing  $\hat{F}$ :

Imagine  $\hat{F}(x + \cdot)$  as some allocation of mass in  $\mathbb{R}^d$  that amounts to  $G(x)$  under the measure  $q$ . When we vary  $\hat{F}$ , we are allowed to shuffle this mass around while keeping its  $q$ -measure equal to  $G(x)$ , as long as  $0 \leq \hat{F} \leq 1$ . To maximize  $\mathbb{E}_{\delta \sim q} \hat{F}(x + \delta) \langle u, \nabla \psi(\delta) \rangle$ , we then need to allocate as much  $q$ -measure as possible toward regions where  $\langle u, \nabla \psi(\cdot) \rangle$  is large.

In other words, the maximizing  $\hat{F}$ , which we denote as  $\hat{F}^*$ , is

$$\hat{F}^*(x + \delta) = \begin{cases} 1 & \text{if } \langle u, \nabla \psi(\delta) \rangle > \varphi_u^{-1}(G(x)) \\ 0 & \text{else,} \end{cases}$$

if  $\mathbb{P}[\langle u, \nabla \psi(\delta) \rangle = \varphi_u^{-1}(G(x))] = 0$ , where  $\varphi_u^{-1}$  is the inverse complementary CDF of the random variable  $\gamma_u = \langle u, \nabla \psi(\delta) \rangle$  (with randomness induced by  $\delta \sim q$ ), as defined

in [Definition F.4](#). If there is a singular point at  $\varphi_u^{-1}(G(x))$ , i.e.  $\mathbb{P}[\langle u, \nabla \psi(\delta) \rangle = \varphi_u^{-1}(G(x))] > 0$ , then we choose a subset of  $U \subseteq \{\delta : \langle u, \nabla \psi(\delta) \rangle = \varphi_u^{-1}(G(x))\}$  with  $q$ -measure  $\mathbb{P}[\delta \in U] = G(x) - \varphi_u(\varphi_u^{-1}(G(x)))$ , and define  $\hat{F}^*$  as

$$\hat{F}^*(x + \delta) = \begin{cases} 1 & \text{if } \langle u, \nabla \psi(\delta) \rangle > \varphi_u^{-1}(G(x)) \text{ or } \delta \in U \\ 0 & \text{else.} \end{cases}$$

Then

$$\begin{aligned} &\mathbb{E}_{\delta \sim q} \hat{F}^*(x + \delta) \langle u, \nabla \psi(\delta) \rangle \\ &\leq \mathbb{E}_{\delta \sim q} \hat{F}^*(x + \delta) \langle u, \nabla \psi(\delta) \rangle = \mathbb{E} \gamma_u^{G(x)}, \end{aligned}$$

where  $\gamma_u^{(p)}$  is the random variable defined in [Definition F.4](#). Finally, putting everything together,

$$\begin{aligned} &\max_{u \in \text{Vert}(\mathcal{B})} u \cdot \nabla G(x) \\ &\leq \max_{u \in \text{Vert}(\mathcal{B})} \sup_{\hat{F}: \hat{G}(x) = G(x)} \mathbb{E} \hat{F}(x + \delta) \langle u, \nabla \psi(\delta) \rangle \\ &= \max_{u \in \text{Vert}(\mathcal{B})} \mathbb{E} \gamma_u^{G(x)} \\ &= \Phi(G(x)) \end{aligned}$$

by the definition of  $\Phi$  in [Definition F.4](#). This shows [Eq. \(12\)](#) and consequently the theorem as well.  $\square$

**Lemma F.9.** Consider a function  $p_t$  differentiable in  $t \in [0, \infty)$ . Suppose  $0 < p_0 \leq 1/2$ , and

$$dp_t/dt \leq \Phi(p_t)$$

for some function  $\Phi: (0, \infty) \rightarrow \mathbb{R}^+$  taking only positive values, Then  $p_T < 1/2$  as well for any

$$T < \int_{p_0}^{1/2} \frac{1}{\Phi(p)} dp.$$

*Proof.* WLOG, we can assume that  $dp_t/dt > 0$  for all  $t \in [0, \infty)$ . Thus,  $p_t$  is increasing in  $t$ , and there exists a differentiable inverse function  $t(p)$  that expresses the time  $t$  that  $p_t = p$ . We then have  $dt(p)/dp = \frac{1}{dp_t/dt} = \frac{1}{\Phi(p)}$ , and for any  $\epsilon \geq 0$ ,

$$\begin{aligned} t(1/2 - \epsilon) &= t(1/2 - \epsilon) - t(p_0) \\ &= \int_{p_0}^{1/2 - \epsilon} \frac{dt(p)}{dp} dp = \int_{p_0}^{1/2 - \epsilon} \frac{1}{\Phi(p)} dp. \end{aligned}$$

Since this integral is continuous in  $\epsilon$ , there is an  $\epsilon^* > 0$  such that

$$t(1/2 - \epsilon^*) = \int_{p_0}^{1/2 - \epsilon^*} \frac{1}{\Phi(p)} dp = T.$$

Therefore  $p_T = 1/2 - \epsilon^* < 1/2$ , as desired.  $\square$

**F.1. Example: Gaussian against  $\ell_2$  Adversary**

We give a quick example of recovering the tight Gaussian bound of [Cohen et al. \(2019\)](#) using the differential method.

In this section, we set the norm  $\|\cdot\|$  to be the  $\ell_2$  norm  $\|x\|_2 = \sqrt{\sum_{i=1}^d x_i^2}$ . Then  $\mathcal{B}$  is just the unit ball, and its ‘‘vertices’’ are just the points on the unit sphere. Additionally, we let  $q$  be the Gaussian measure

$$q(x) \propto \exp(-\|x\|_2^2/2)$$

$$\text{so that } \psi(x) = \|x\|_2^2/2 \text{ and } \nabla\psi(x) = x.$$

Below, let `GaussianCDF` be the CDF of the standard Gaussian in 1D.

**Theorem F.10.** *Suppose  $H$  is a smoothed classifier smoothed by the Gaussian distribution*

$$q(x) \propto \exp(-\|x\|_2^2/2\sigma^2),$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_2 < \sigma \text{GaussianCDF}^{-1}(\rho).$$

*Proof.* By linearity in  $\sigma$ , it suffices to show this for  $\sigma = 1$ . For brevity, let us denote `GaussianCDF` in this proof by  $\Psi$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

For any  $u \in \text{Vert}(\mathcal{B})$  (i.e. any unit vector  $u$ ),  $\gamma_u = \langle u, \nabla\psi(\delta) \rangle = \langle u, \delta \rangle$ ,  $\delta \sim q$ , is a standard Gaussian random variable (in  $\mathbb{R}$ ). Therefore, for  $p \in [0, 1]$ , the random variable  $\gamma_u^{(p)}$  defined in [Definition F.4](#) is just

$$\gamma_u^{(p)} = \begin{cases} 0 & \text{with prob. } 1 - p \\ \mathcal{N}(0, 1)|_{[c, \infty)} & \text{with prob. } p, \end{cases}$$

where  $c \stackrel{\text{def}}{=} \Psi^{-1}(1 - p)$ , and  $\mathcal{N}(0, 1)|_{[c, \infty)}$  is a standard Gaussian  $z$  conditioned on  $z \geq c$ . Thus, for any  $u \in \text{Vert}(\mathcal{B})$ ,

$$\begin{aligned} \Phi(p) &= \mathbb{E} \gamma_u^{(p)} = \mathbb{E}_{z \sim \mathcal{N}(0, 1)} z \mathbb{I}(z \geq c) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \Big|_c^\infty \\ &= \Psi'(c) = \Psi'(\Psi^{-1}(1 - p)). \end{aligned}$$

Then, by setting  $G(x)$  in [Theorem F.6](#) to be  $1 - H(x)_y = 1 - \rho$ , we get the provably robust radius of

$$\begin{aligned} \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp &= \int_{1-\rho}^{1/2} \frac{1}{\Psi'(\Psi^{-1}(1 - p))} dp \\ &= \int_0^{\Psi^{-1}(\rho)} dc = \Psi^{-1}(\rho), \end{aligned}$$

as desired.  $\square$

**F.2. Example: Laplace against  $\ell_1$  Adversary**

Let us give another quick example of recovering the tight Laplace bound of [Teng et al. \(2019\)](#) using the differential method.

In this section, we set the norm  $\|\cdot\|$  to be the  $\ell_1$  norm  $\|x\|_1 = \sum_{i=1}^d |x_i|$ . Then the unit ball  $\mathcal{B}$  is the convex hull of its vertices which are the coordinates vectors and their negations:

$$\text{Vert}(\mathcal{B}) = \{\pm e_i : i \in [d]\}.$$

Consider the Laplace distribution

$$q(x) \propto \exp(-\|x\|_1) \text{ so that}$$

$$\psi(x) = \|x\|_1 \text{ and } \nabla\psi(x) = (\text{sgn}(x_1), \dots, \text{sgn}(x_d)),$$

with  $\nabla\psi(x)$  defined whenever all  $x_i$ s are nonzero.

**Theorem F.11.** *Suppose  $H$  is a smoothed classifier smoothed by the Laplace distribution*

$$q(x) \propto \exp(-\|x\|_1/\lambda),$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \lambda \log \frac{1}{2(1 - \rho)}.$$

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

For any  $u \in \text{Vert}(\mathcal{B})$  (i.e.  $u = \pm e_i$ ),  $\gamma_u = \langle u, \nabla\psi(\delta) \rangle$ ,  $\delta \sim q$ , is a Rademacher random variable that takes values 1 and  $-1$  with equal probability. Therefore, for  $p \in [0, 1/2]$ , the random variable  $\gamma_u^{(p)}$  defined in [Definition F.4](#) is

$$\gamma_u^{(p)} = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1 - p. \end{cases}$$

Thus, for any  $u \in \text{Vert}(\mathcal{B})$ ,

$$\Phi(p) = \mathbb{E} \gamma_u^{(p)} = p.$$

Then, by setting  $G(x)$  in [Theorem F.6](#) to be  $1 - H(x)_y = 1 - \rho$ , we get the provably robust radius of

$$\int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp = \int_{1-\rho}^{1/2} \frac{1}{p} dp = \log \frac{1}{2(1 - \rho)}.$$

$\square$

## G. Wulff Crystal

The following is an intuitive statement of the main isoperimetric property of Wulff Crystals.

**Theorem G.1** (Isoperimetric property of Wulff Crystals (Brothers & Morgan, 1994), informal statement). *Let  $\|\cdot\|$  be any norm on  $\mathbb{R}^n$ . Let  $Z$  be the Wulff Crystal of  $\|\cdot\|$ , i.e. the unit ball of the norm  $\|\cdot\|_*$  dual to  $\|\cdot\|$ .*

$$Z = \{x : \|x\|_* \leq 1\}.$$

Let  $\Omega$  be any measurable subset of  $\mathbb{R}^n$  with finite perimeter and of the same volume as  $Z$ . Then with  $\mathbf{n}(\Omega, x)$  denoting the normal vector at  $x$  with respect to  $\Omega$ , normalized to have  $\ell_2$  norm 1,

$$\int_{\partial\Omega} \|\mathbf{n}(\Omega, x)\| dx \geq \int_{\partial Z} \|\mathbf{n}(Z, x)\| dx$$

with equality holding if and only if  $\Omega$  differs from a translate of  $Z$  by a set of volume zero.

This statement carries across the core essence of the isoperimetry, and is a rigorous statement if  $\Omega$  is restricted to have smooth boundary, but care needs to be taken to explain the concept of “finite perimeter,” “normal vector,” the “boundary  $\partial\Omega$ ,” and the boundary measure on  $\partial\Omega$ , in the context of general, measurable  $\Omega$ . These quantities are defined in Appendix H, but we also refer the interested reader to (Brothers & Morgan, 1994) for more mathematical details.

### G.1. Wulff Crystals are Zonotopes

In this paper, the norm  $\|\cdot\|$  referred to in Theorem G.1 will take the form of  $\|x\| = \mathbb{E}_v |\langle x, v \rangle|$  where  $v$  is sampled from some distribution, as in Definition 5.1. When  $v$  is sampled uniformly from some finite set of vectors  $\mathcal{S}$ , then the Wulff Crystal  $B = \{x : \|x\|_* \leq 1\}$  is proportional to the Zonotope of  $\mathcal{S}$  (McMullen, 1971).

**Definition G.2** (Zonotope). Given a finite collection of vectors  $\mathcal{S}$ , the zonotope  $\text{Zon}(\mathcal{S})$  is defined as the Minkowski sum of the vectors of  $\mathcal{S}$ , i.e.

$$\text{Zon}(\mathcal{S}) \stackrel{\text{def}}{=} \left\{ \sum_{v \in \mathcal{S}} a_v v : a_v \in [0, 1], \forall v \in \mathcal{S} \right\}.$$

The zonotope can be viewed as a linear projection of the cube  $[0, 1]^{\mathcal{S}}$  sending each unit vector to a vector of  $\mathcal{S}$ .

*Example G.3.* If  $\mathcal{B}$  is the  $\ell_1$  unit ball, then  $\text{Zon}(\text{Vert}(\mathcal{B}))$  is a cube. If  $\mathcal{B}$  is the  $\ell_\infty$  unit ball, then in 2 dimensions,  $\text{Zon}(\text{Vert}(\mathcal{B}))$  is a rhombus; in 3 dimensions, it is rhombic dodecahedron; in higher dimensions, there is no simpler description of the resulting polytope.

**Proposition G.4.** *The Wulff Crystal w.r.t.  $\mathcal{B}$  is equal to the zonotope  $\frac{2}{|\text{Vert}(\mathcal{B})|} \text{Zon}(\text{Vert}(\mathcal{B}))$ .*

The volume of a zonotope, and thus of Wulff Crystals, can be computed easily using the following formula.

**Proposition G.5.** *Let  $\mathcal{S}$  be a finite set of vectors in  $\mathbb{R}^d$ . Then the  $d$ -dimensional volume of  $\text{Zon}(\mathcal{S})$  is given by*

$$\sum_{\mathcal{T} \subseteq \mathcal{S}: |\mathcal{T}|=d} |\text{Vol}(\text{Zon}(\mathcal{T}))| = \sum_{\mathcal{T} \subseteq \mathcal{S}: |\mathcal{T}|=d} |\det \mathcal{T}|,$$

where  $\det \mathcal{T}$  is the determinant of the square matrix with vectors of  $\mathcal{T}$  as columns.

### G.2. Wulff Crystals Yield Optimal Uniform Distributions for Randomized Smoothing

In this section, we will formulate Theorem 5.2 rigorously and prove it.

**Definition G.6.** Let  $\mathcal{S}$  be a finite set of vectors in  $\mathbb{R}^d$  and let  $G$  be the group of linear transformations that permute  $\mathcal{S}$  (i.e.  $G$  is  $\mathcal{S}$ 's linear symmetry group). We say  $\mathcal{S}$  is *symmetric* if  $G$  acts on  $\mathcal{S}$  transitively, i.e. for any two elements  $v, w \in \mathcal{S}$ , there is a group element  $g \in G$  such that  $g \cdot v = w$ .

For example, the boolean cube  $\{\pm 1\}^d$  is symmetric, and so is the set of coordinate vectors and their negations. The following is the main theorem of this section, stating the optimality of uniform distributions supported Wulff Crystals.

**Theorem G.7.** *If  $\mathcal{B}$  is a full-dimensional polytope in  $\mathbb{R}^d$  symmetric around the origin, and whose vertices form a symmetric set, then the Wulff Crystal w.r.t.  $\mathcal{B}$  minimizes*

$$\sup_{v \in \mathcal{B}} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S)$$

among all measurable, not necessarily convex, sets  $S \subseteq \mathbb{R}^d$  of the same volume and of finite perimeter. In other words, among uniform distributions supported on measurable sets of volume  $1$  and finite perimeter, the one supported on the Wulff Crystal minimizes the maximal instantaneous growth  $\Phi(p)$  in the measure of a set due to an instantaneous perturbation from  $\mathcal{B}$ .

The condition “finite perimeter” can be interpreted intuitively here, but a formal definition is given in Definition H.4. This condition is necessary because otherwise the limit in question does not exist.

*Proof.* By Lemma G.18, we have

$$\lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) = \int_{\partial S} \Theta(\langle \mathbf{n}(x), v \rangle) dx \quad (13)$$

where  $\mathbf{n}(x)$  is the normal at  $x$  w.r.t.  $S$ , and  $\Theta(x) = \max(0, x)$ . Note this quantity is convex in  $v$  because  $\Theta$

is convex. Then

$$\begin{aligned}
 & \sup_{v \in \mathcal{B}} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) \\
 &= \sup_{v \in \text{Vert}(\mathcal{B})} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) \\
 &\geq \mathbb{E}_{v \sim \text{Vert}(\mathcal{B})} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) \quad (14) \\
 &= \mathbb{E}_{v \sim \text{Vert}(\mathcal{B})} \int_{\partial S} \Theta(\langle \mathbf{n}(x), v \rangle) dx \\
 &= \int_{\partial S} \mathbb{E}_{v \sim \text{Vert}(\mathcal{B})} \Theta(\langle \mathbf{n}(x), v \rangle) dx
 \end{aligned}$$

Since  $\mathcal{B} = -\mathcal{B}$  and thus  $\text{Vert}(\mathcal{B}) = -\text{Vert}(\mathcal{B})$ ,

$$\|w\| \stackrel{\text{def}}{=} \mathbb{E}_{v \sim \text{Vert}(\mathcal{B})} \Theta(\langle w, v \rangle) = \frac{1}{2} \mathbb{E}_{v \sim \text{Vert}(\mathcal{B})} |\langle w, v \rangle|$$

is a seminorm. This is in fact a norm because  $\text{Vert}(\mathcal{B})$  spans  $\mathbb{R}^d$ , by the assumption that  $\mathcal{B}$  is full-dimensional. Then, plugging  $\|\cdot\|$  into  $\|\cdot\|$  in [Theorem G.1](#), we get that the Wulff Crystal  $Z$  w.r.t.  $\mathcal{B}$  minimizes

$$Z = \underset{S: \text{Vol}(S) = \text{Vol}(Z)}{\text{argmin}} \mathbb{E}_{v \sim \text{Vert}(\mathcal{B})} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S).$$

Now note that the norm above is invariant under the transpose of  $\mathcal{B}$ 's symmetry group: for any linear symmetry  $g$  of  $\text{Vert}(\mathcal{B})$ ,

$$\|g^\top w\| = \frac{1}{2} \mathbb{E}_v |\langle g^\top w, v \rangle| = \frac{1}{2} \mathbb{E}_v |\langle w, gv \rangle| = \|w\|.$$

This invariance translates to the dual norm  $\|\cdot\|_*$ 's invariance under the symmetry group itself. Thus the Wulff Crystal  $Z$ , being the unit ball of the dual norm, is itself invariant under the symmetry group of  $\text{Vert}(\mathcal{B})$ . By assumption, this symmetry group acts transitively on  $\text{Vert}(\mathcal{B})$ , so  $Z$  ‘‘looks the same’’ from the angle of every  $v \in \text{Vert}(\mathcal{B})$ , i.e.

$$\text{Vol}((Z + rv) \setminus Z) = \text{Vol}((Z + rv) \setminus Z)$$

for any  $w, v \in \text{Vert}(\mathcal{B})$ . Consequently, [Eq. \(14\)](#) holds with equality, and  $Z$  minimizes the supremum in question as well.  $\square$

### G.2.1. GROWTH CALCULATIONS FOR STANDARD SHAPES

Using the fact that the volume of the  $d$ -dimensional unit ball is  $\pi^{d/2} \Gamma(d/2 + 1)^{-1}$ , and the volume of the standard  $d$ -dimensional cross polytope is  $2^d/d!$ , as well as the identity

$$\lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) = \|v\|_2 \text{Vol}(\Pi_v S)$$

if  $S$  is convex, we can derive the following facts easily.

**Theorem G.8.** *If  $S \subseteq \mathbb{R}^d$  is an axis-parallel unit cube and  $e_1$  is the first unit vector, then*

$$\lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + re_1) \setminus S) = 1.$$

**Theorem G.9.** *If  $S \subseteq \mathbb{R}^d$  is a  $(\ell_2)$ -ball of volume 1 and  $v$  is any  $(\ell_2)$ -unit vector, then*

$$\lim_{d \rightarrow \infty} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) = \sqrt{e}.$$

**Theorem G.10.** *If  $S \subseteq \mathbb{R}^d$  is the cross polytope (i.e.  $\ell_1$  ball) of volume 1, and  $e_1$  is the first unit vector, then*

$$\lim_{d \rightarrow \infty} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + re_1) \setminus S) = e.$$

**Theorem G.11.** *If  $S \subseteq \mathbb{R}^d$  is an axis-parallel unit cube and  $v = (1, \dots, 1)$ , then*

$$\lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) = d.$$

**Theorem G.12.** *If  $S \subseteq \mathbb{R}^d$  is the cross polytope (i.e.  $\ell_1$  ball) of volume 1, and  $v = (1, \dots, 1)$ , then*

$$\lim_{d \rightarrow \infty} d^{-1/2} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) = e\sqrt{2/\pi}.$$

*Proof.* It is equivalent to take  $S$  to be the standard  $\ell_1$  ball, and to calculate

$$\lim_{d \rightarrow \infty} \frac{\lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S)}{\text{Vol}(S)^{\frac{d-1}{d}} \sqrt{d}}, \quad (15)$$

and confirm it equals  $e\sqrt{2/\pi}$ . Note that the unit surface normals of  $S$  are  $\{\pm 1\}^d/\sqrt{d}$ , occurring with equal probability over the surface measure of  $S$ . Using [Eq. \(13\)](#), we then see that

$$\lim_{d \rightarrow \infty} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) = W \sum_{i=0}^{\lfloor d/2 \rfloor} \binom{d}{i} \frac{d-2i}{\sqrt{d}}$$

where  $W = \frac{\sqrt{d}}{(d-1)!}$  is the volume of the simplex  $\{x : \sum_i x_i = 1, x \geq 0\}$ . This evaluates to

$$\frac{1}{(d-1)!} \times \begin{cases} \frac{d+2}{2} \binom{d}{\frac{d}{2}+1} & \text{if } d \text{ is even} \\ \frac{d+1}{2} \binom{d}{\frac{d+1}{2}} & \text{if } d \text{ is odd.} \end{cases}$$

Finally, since the volume of  $S$  is  $2^d/d!$ , we can calculate [Eq. \(15\)](#) directly and obtain the desired result.  $\square$

### G.2.2. WULFF CRYSTAL OF THE $\ell_\infty$ BALL

In this section, let  $Z$  be the Wulff Crystal ([Definition 5.1](#)) w.r.t.  $\mathcal{B} = \{x : \|x\|_\infty \leq 1\}$ , i.e.  $Z$  is the unit ball of the norm dual to  $\|x\|_*$   $\stackrel{\text{def}}{=} \mathbb{E}_{v \sim \{\pm 1\}^d} |\langle x, v \rangle|$ . By [Proposition G.4](#),  $Z$  can also be described as the zonotope of  $2^{-d+1} \{\pm 1\}^d$ . From these descriptions, we can straightforwardly see the following properties of  $Z$ .

**Proposition G.13.** *The vertices of  $Z$  farthest from the origin are coordinate vectors and their negations. The facets of  $Z$  closest to the origin are of the form  $\{x : \pm x_i \pm x_j \leq 1\}$ . Therefore, with  $B$  denoting the  $\ell_2$  unit ball,*

$$\frac{1}{\sqrt{2}}B \subseteq Z \subseteq B.$$

In general, the properties of  $Z$  are elusive, and tied to many open problems in combinatorics and polytope theory (Ziegler, 1995). But we may heuristically compute  $\lim_{d \rightarrow \infty} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((Z + rv) \setminus Z) = \lim_{d \rightarrow \infty} \|v\|_2 \Pi_v Z$  when  $v = (1, \dots, 1)$ , as follows. (Because our computation is heuristic, we phrase the following as a claim, and not a theorem)

**Claim G.14.** *If  $S \subseteq \mathbb{R}^d$  is the Wulff Crystal w.r.t. the  $\ell_\infty$  unit ball, scaled to have volume 1, and  $v = (1, \dots, 1)$ , then*

$$\lim_{d \rightarrow \infty} d^{-1/2} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) = \sqrt{e}.$$

*Derivation.* Since  $2^{d-1}Z$  is  $\text{Zon}(\{\pm 1\}^d)$ , we have  $\Pi_v 2^{d-1}Z = \text{Zon}(\Pi_v \{\pm 1\}^d)$ , the zonotope of the set of vectors  $\left\{x - \frac{\langle x, v \rangle}{\|v\|_2} v : v \in \{\pm 1\}^d\right\}$ . By Eq. (6), we then have

$$\lim_{r \rightarrow 0} \frac{\text{Vol}((S + rv) \setminus S)}{r\sqrt{d}} = \frac{\text{Vol}(\text{Zon}(\Pi_v \{\pm 1\}^d))}{\text{Vol}(\text{Zon}(\{\pm 1\}^d))^{\frac{d-1}{d}}}. \quad (16)$$

Now Lemma G.16 tells us that the  $\text{Vol}(\text{Zon}(\{\pm 1\}^d))$  is a multiple of the expected determinants of all  $d \times d$  matrices with entries  $\pm 1$ . By a result of Nguyen et al. (2014) (Theorem G.17), the determinant of a random  $d \times d$  matrix with iid  $\pm 1$  entries is distributed in high dimension  $d$  roughly as  $\sqrt{(d-1)!} e^{z\sqrt{\frac{1}{2}\log d}}$  where  $z \sim \mathcal{N}(0, 1)$ . Thus, by Lemma G.16, we should expect (this is the first place where we argue heuristically)

$$\begin{aligned} \text{Vol}(\text{Zon}(\{\pm 1\}^d)) &\approx \frac{1}{d!} 2^{d^2} \mathbb{E}_z \sqrt{(d-1)!} e^{z\sqrt{\frac{1}{2}\log d}} \\ &= \frac{1}{d!} 2^{d^2} \sqrt{(d-1)!} d^{\frac{1}{4}}. \end{aligned} \quad (17)$$

We verify this approximation to be correct numerically for moderately large  $d$ . Similarly, the uniform distribution over  $\{\pm 1\}^d$  is close to a standard Gaussian when  $d \gg 1$ , so that  $\Pi_v \{\pm 1\}^d$  is close to a  $(d-1)$ -dimensional standard Gaussian. Therefore, we should expect that

$$\begin{aligned} &\text{Vol}(\text{Zon}(\Pi_v \{\pm 1\}^d)) \\ &\approx \frac{1}{(d-1)!} (2^d - 1)^{d-1} \mathbb{E} |\det Y| \\ &\approx \frac{1}{(d-1)!} 2^{d(d-1)} \sqrt{(d-2)!} (d-1)^{\frac{1}{4}}, \end{aligned} \quad (18)$$

where  $Y$  is a  $(d-1) \times (d-1)$  Gaussian matrix, and where in Eq. (18), we used the heuristic that for large  $d$ ,  $|\det Y|$

is lognormal (Theorem G.17). Again, we verify these approximations numerically. Plugging in Eqs. (17) and (18) into Eq. (16) and taking the  $d \rightarrow \infty$  limit yields the desired result.  $\square$

Since the sphere has the same large  $d$  limit (Theorem G.9), we can say that

**Claim G.15.** *For every  $\epsilon > 0$ ,  $S$  = the  $\ell_2$ -ball of volume 1 achieves within  $\epsilon$  of*

$$\min_{\text{Vol}(S)=1} \sup_{\|v\|_\infty \leq 1} d^{-1/2} \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S),$$

*for sufficiently large  $d$ . This is not true for  $S$  = the  $\ell_\infty$ - or the  $\ell_1$ -ball.*

**Lemma G.16.** *The volume of  $\text{Zon}(\{\pm 1\}^d)$  is*

$$\frac{1}{d!} 2^{d^2} \mathbb{E}_X |\det X|$$

*where  $X \in \{\pm 1\}^{d \times d}$  is a random  $d \times d$  matrix whose coordinates are iid Rademacher variables (i.e. 1 or -1 with equal probability).*

*Proof.* The above expression can be rewritten as

$$\frac{1}{d!} \sum_{X \in \{\pm 1\}^{d \times d}} |\det X|.$$

Because  $\det X = 0$  if any two columns are equal, so this is equivalent to summing over all  $X$  with distinct columns.

$$\frac{1}{d!} \sum_{\substack{X \in \{\pm 1\}^{d \times d} \\ X \text{ has distinct columns}}} |\det X|.$$

Finally, any given set of  $d$  distinct column vectors is represented  $d!$  times in the sum through its  $d!$  permutations, so this is equal to

$$\sum_{T \subseteq \{\pm 1\}^d: |T|=d} |\det T|,$$

which by Proposition G.5 is the volume of the zonotope in question.  $\square$

**Theorem G.17** (Nguyen et al. (2014)). *Let  $A_n$  be an  $n \times n$  random matrix whose entries are independent real random variables with mean zero, variance one and with subexponential tail. Then with  $\mu_n = \frac{1}{2} \log(n-1)!$  and*

$$\sigma_n = \sqrt{\frac{1}{2} \log n},$$

$$\sup_{x \in \mathbb{R}} \left| \mathbb{P} \left( \frac{\log(|\det A_n|) - \mu_n}{\sigma_n} \leq x \right) - \mathbb{P}(\mathcal{N}(0, 1) \leq x) \right| \leq \log^{-1/3+o(1)} n.$$

In other words,  $\det A_n \approx \sqrt{(n-1)!} e^{z\sqrt{\frac{1}{2}\log n}}$  where  $z \sim \mathcal{N}(0, 1)$ .

## G.2.3. GROWTH FORMULA OF A SET

**Lemma G.18.** *Let  $S \subseteq \mathbb{R}^d$  be a set of finite perimeter and  $v \in \mathbb{R}^d$  be any vector. Then*

$$\lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) = \int_{\partial S} \Theta(\langle \mathbf{n}(x), v \rangle) dx,$$

where  $\mathbf{n}(x)$  is the normal at  $x$  w.r.t.  $S$ , and  $\Theta(x) = \max(0, x)$ .

*Proof.* Let  $\partial S_v \stackrel{\text{def}}{=} \{x \in \partial S : \langle \mathbf{n}(x), v \rangle > 0\}$  be the part of  $S$ 's boundary whose surface normal aligns with  $v$ . For any vector  $w$ , let  $\partial S_v + [0, w] = \{x + rw : x \in \partial S_v, r \in [0, 1]\}$  be the Minkowski sum of  $\partial S_v$  and the segment  $[0, w]$ . Then it's clear that  $\text{Vol}(\partial S_v + [0, rv]) \leq r \int_{\partial S} \Theta(\langle \mathbf{n}(x), v \rangle) dx$ , and that

$$(S + rv) \setminus S \subseteq \partial S_v + [0, rv].$$

Thus,

$$\begin{aligned} & \lim_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) \\ & \leq \lim_{r \rightarrow 0} r^{-1} r \int_{\partial S} \Theta(\langle \mathbf{n}(x), v \rangle) dx \\ & = \int_{\partial S} \Theta(\langle \mathbf{n}(x), v \rangle) dx. \end{aligned}$$

Now for the other direction, observe that the signed measure  $r^{-1}(\mathbb{I}(x \in S + rv) - \mathbb{I}(x \in S))$  converges weakly to the (singular) signed measure  $\langle \mathbf{n}(x), v \rangle$  supported on  $\partial S$ . Indeed, for any compactly supported  $C^1$  function  $f$ , we have

$$\begin{aligned} & \lim_{r \rightarrow 0} r^{-1} \int f(x)(\mathbb{I}(x \in S + rv) - \mathbb{I}(x \in S)) dx \\ & = \lim_{r \rightarrow 0} r^{-1} \int (f(x + rv) - f(x)) \mathbb{I}(x \in S) dx \\ & = \int_S D_v f(x) dx = \int_{\partial S} \langle \mathbf{n}(x), v f(x) \rangle dx. \end{aligned}$$

Now, taking the supremum of the RHS over all compactly supported  $C^1$  function  $|f| \leq 1$ , we get

$$\begin{aligned} & \int_{\partial S} \Theta(\langle \mathbf{n}(x), v \rangle) dx \\ & = \frac{1}{2} \int_{\partial S} |\langle \mathbf{n}(x), v \rangle| dx \\ & = \frac{1}{2} \sup_f \int_{\partial S} \langle \mathbf{n}(x), v f(x) \rangle dx \\ & = \frac{1}{2} \sup_f \lim_{r \rightarrow 0} r^{-1} \int f(x)(\mathbb{I}(x \in S + rv) - \mathbb{I}(x \in S)) dx \\ & \leq \frac{1}{2} \liminf_{r \rightarrow 0} r^{-1} \sup_f \int f(x)(\mathbb{I}(x \in S + rv) - \mathbb{I}(x \in S)) dx \\ & = \liminf_{r \rightarrow 0} r^{-1} \text{Vol}((S + rv) \setminus S) \end{aligned}$$

as desired.  $\square$

## G.3. Optimal Smoothing Distributions Have Wulff Crystal Level Sets

**Definition G.19** (Level-Equivalence). Call two distribution  $q_1$  and  $q_2$  level-equivalent if their superlevel sets have the same volumes:

$$\text{Vol}\{x : q_1(x) \geq t\} = \text{Vol}\{x : q_2(x) \geq t\}, \quad \forall t \in (0, \infty).$$

**Theorem G.20.** *Let  $\mathcal{B}$  be a full-dimensional polytope in  $\mathbb{R}^d$  symmetric around the origin, and whose vertices form a symmetric set. Let  $Z$  be the Wulff Crystal w.r.t.  $\mathcal{B}$ . Let  $q_0$  be a regular (Definition E.2) probability density function. Among all probability distributions  $q$  with regular (Definition E.2) and even density function that is level equivalent to  $q_0$ , the probability density function with concentric superlevel sets proportional to  $Z$  minimize*

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

where  $\|\cdot\|$  is the norm defined by  $\mathcal{B}$ .

Note that this theorem does not imply Theorem G.7 since uniform distributions do not have regular densities. However, this can be generalized to bounded-variation densities (Theorem H.15) which subsume both Theorem G.7 and Theorem G.20.

*Proof.* Consider any distribution  $q$  level-equivalent to  $q_0$ . Let  $U_t$  be its superlevel sets.

Expanding the definition of  $\Phi$  in terms of  $\gamma_u^{(p)}$  (see Definition F.4), and exchanging maximization and expectation, we get

$$\begin{aligned} \Phi(1/2) & = \max_{u \in \text{Vert}(\mathcal{B})} \mathbb{E} \gamma_u^{(1/2)} \geq \mathbb{E}_u \mathbb{E} \gamma_u^{(1/2)} \\ & = \mathbb{E}_u \int \max(\nabla q(x) \cdot u, 0) dx \\ & = \int \mathbb{E}_u \max(\nabla q(x) \cdot u, 0) dx. \end{aligned}$$

Since  $\mathcal{B} = -\mathcal{B}$  and thus  $\text{Vert}(\mathcal{B}) = -\text{Vert}(\mathcal{B})$ ,

$$\|w\| \stackrel{\text{def}}{=} \mathbb{E}_{u \sim \text{Vert}(\mathcal{B})} \max(0, \langle w, u \rangle) = \frac{1}{2} \mathbb{E}_{u \sim \text{Vert}(\mathcal{B})} |\langle w, u \rangle|$$

is a seminorm. This is in fact a norm because  $\text{Vert}(\mathcal{B})$  spans  $\mathbb{R}^d$ , by the assumption that  $\mathcal{B}$  is full-dimensional. Thus

$$\Phi(1/2) \geq \int \|\nabla q(x)\| dx.$$

Define  $g(x) \stackrel{\text{def}}{=} \frac{\|\nabla q(x)\|}{\|\nabla q(x)\|_2}$  if  $\nabla q(x) \neq 0$ , and  $g(x) = 0$  otherwise. Then by Theorem E.3,

$$\begin{aligned} \int \|\nabla q(x)\| dx & = \int g(x) \|\nabla q(x)\|_2 dx \\ & = \int_0^\infty \int_{\partial U_t} g(x) dx dt. \end{aligned}$$

By the Weak Sard's theorem (Theorem E.4), we may ignore the places where  $\nabla q(x) = 0$ , and this integral is the same as

$$\int_0^\infty \int_{\partial U_t} g(x) dx dt = \int_0^\infty \int_{\partial U_t} \frac{\|\nabla q(x)\|}{\|\nabla q(x)\|_2} dx dt. \quad (19)$$

Now, the surface normal at  $x$  w.r.t.  $U_t$  is proportional to  $\nabla q(x)$ . Thus, the  $(\ell_2)$ -unit normal  $\mathbf{n}(x)$  at  $x$  w.r.t.  $U_t$  is given by  $\frac{-\nabla q(x)}{\|\nabla q(x)\|_2}$ , so  $\frac{\|\nabla q(x)\|}{\|\nabla q(x)\|_2}$  is the  $\|\cdot\|$ -norm of  $\mathbf{n}(x)$ . Therefore, the inner integral is

$$\int_{\partial U_t} \frac{\|\nabla q(x)\|}{\|\nabla q(x)\|_2} dx = \int_{\partial U_t} \|\mathbf{n}(x)\| dx.$$

By Theorem G.7, this is minimized for fixed  $\text{Vol}(U_t)$  by  $U_t \propto$  the Wulff Crystal w.r.t.  $\mathcal{B}$ . Thus, the unique distribution  $q^*$  level equivalent to  $q_0$  and with concentric Wulff Crystal superlevel sets (all centered at 0) minimizes Eq. (19). But since

$$\Phi(1/2) = \max_u \int \max(\nabla q(x) \cdot u, 0) dx$$

and the inner integral here is invariant in  $u$  when  $q = q^*$  by the symmetry of the Wulff Crystal, as in the proof of Theorem G.7, Eq. (19) in fact holds with equality for  $q^*$ , so that  $q^*$  minimizes  $\Phi(1/2)$  as well.  $\square$

#### G.4. Optimality among Wulff Crystal Distributions

Given the optimality of Wulff Crystal distributions among level-equivalent distributions, one may wonder, among Wulff Crystal distributions themselves, which one minimizes  $\Phi(1/2)$ ? Because no two such distributions are level-equivalent, we need to fix another notion of the spread of the distribution. The below theorem answers this question, controlling for the expected Wulff Crystal norm.

**Theorem G.21.** *Let  $\mathcal{B}$  be a full-dimensional polytope in  $\mathbb{R}^d$  symmetric around the origin, and whose vertices form a symmetric set. Let  $Z$  be the Wulff Crystal w.r.t.  $\mathcal{B}$ , and let  $\|\cdot\|$  denote the norm with  $Z$  as its unit ball. Consider a probability distribution  $q(x) \propto \exp(-\psi(\|x\|))$  on  $\mathbb{R}^d$ ,  $d \geq 2$ , for some regular, even  $\psi$ . Then with  $\Phi$  defined against the adversary  $\mathcal{B}$ , we have, for any  $k > 0$ ,*

$$\Phi(1/2) \geq \frac{(d-1)C}{\sqrt[k]{\mathbb{E}_{\delta \sim q} \|\delta\|^k}},$$

where  $C \stackrel{\text{def}}{=} \frac{1}{2} \mathbb{E}_{x \sim \partial Z} |\langle \nabla \|x\|, u \rangle|$ , for any vertex  $u$  of  $\mathcal{B}$ , is a constant that depends only on  $Z$ .

**Remark G.22.** Note that for any  $p > 0, k > 0$ , there is a constant  $T_{p,k}$  depending only on  $Z$  such that

$$\sqrt[k]{\mathbb{E}_{\delta \sim q} \|\delta\|^k} = T_p \sqrt[k]{\mathbb{E}_{\delta \sim q} \|\delta\|^p}$$

for any  $q$  of the form in Theorem G.21. So this theorem applies when we want to fix most measures of spread.

*Proof.* By the symmetry of the Wulff Crystal w.r.t. symmetry group of  $\text{Vert}(\mathcal{B})$ , we have  $\Phi(1/2) = \mathbb{E} \gamma_u^{(1/2)}$  for any  $u \in \text{Vert}(\mathcal{B})$ . Henceforth, we fix  $u$  to be one such vertex of  $\mathcal{B}$ . Note that  $\nabla q(x) = -q(x)\psi'(\|x\|)\nabla \|x\|$ . Thus  $\nabla q(x) = -\nabla q(-x)$ , and

$$\begin{aligned} \Phi(1/2) &= \mathbb{E} \gamma_u^{(1/2)} = \frac{1}{2} \int q(x) |\psi'(\|x\|) \langle \nabla \|x\|, u \rangle| dx \\ &= \frac{1}{2} \mathbb{E}_{x \sim q} |\psi'(\|x\|) \langle \nabla \|x\|, u \rangle| \end{aligned}$$

Note that a sample from  $q$  can be obtained by first sampling  $v \sim \partial Z$  from the (uniform distribution on the) boundary of  $Z$  and  $r \sim q_r$ , where  $q_r(r) \propto r^{d-1} e^{-\psi(r)}$ , and finally returning their product  $rv$ . Therefore, because  $\nabla \|x\|$  doesn't depend on  $\|x\|$ , we can continue the above equations as follows.

$$\begin{aligned} \Phi(1/2) &= \frac{1}{2} \mathbb{E}_{x \sim \partial Z} |\langle \nabla \|x\|, u \rangle| \times \mathbb{E}_{r \sim q_r} |\psi'(r)| \\ &= C \mathbb{E}_{r \sim q_r} |\psi'(r)|. \end{aligned}$$

Now notice that, because  $d \geq 2$ , with  $R \stackrel{\text{def}}{=} \int_0^\infty r^{d-1} e^{-\psi(r)} dr$ ,

$$\begin{aligned} &R \mathbb{E}_{r \sim q_r} |\psi'(r)| \\ &= \int_0^\infty r^{d-1} e^{-\psi(r)} |\psi'(r)| dr \\ &\geq \int_0^\infty r^{d-1} e^{-\psi(r)} \psi'(r) dr \quad (20) \\ &= -r^{d-1} e^{-\psi(r)} \Big|_0^\infty (d-1) \int_0^\infty r^{d-2} e^{-\psi(r)} dr \\ &= (d-1) \int_0^\infty r^{d-2} e^{-\psi(r)} dr \\ &= (d-1) R \mathbb{E}_{r \sim q_r} r^{-1}. \quad (21) \end{aligned}$$

Then, by Holder's inequality, for any  $k > 0$ ,

$$\begin{aligned} \Phi(1/2) \sqrt[k]{\mathbb{E}_{x \sim q} \|x\|^k} &= (d-1) C \mathbb{E}_{r \sim q_r} r^{-1} \sqrt[k]{\mathbb{E}_{r \sim q_r} r^k} \\ &\geq (d-1) C \left( \mathbb{E}_{r \sim q_r} 1 \right)^{1+\frac{1}{k}} \quad (22) \\ &= (d-1) C. \end{aligned}$$

$\square$

**Remark G.23.** Let us comment briefly on the equality case of Theorem G.21, or the lack thereof. There are two inequalities used in the proof above, namely Eqs. (20) and (22). For Eq. (20) to hold with equality, we just need  $\psi'(r) \geq 0$  for

all  $r \geq 0$ . On the other hand, it is impossible for Eq. (22) to hold with equality when  $\psi$  is not allowed to be a delta function on  $r = 1$  (and if that were the case, then Eq. (20) cannot hold with equality). However, as long as the radial distribution  $q_r$  concentrates around its mean value sufficiently well, the inequality should be approximately tight. This is typically the case for high dimensional distributions.

For example, in the Gaussian case with  $\psi(r) = e^{-r^2}$ , we have  $\int_0^\infty r^c e^{-\psi(r)} dr = \frac{1}{2} \Gamma\left(\frac{c+1}{2}\right)$  for any  $c > -1$ , so that

$$\begin{aligned} \mathbb{E}_{r \sim q_r} r^{-1} & \mathbb{E}_{r \sim q_r} r = \frac{\Gamma\left(\frac{d-1}{2}\right) \Gamma\left(\frac{d+1}{2}\right)}{\Gamma\left(\frac{d}{2}\right)^2} \\ & = 1 + \frac{1}{2d} + O(d^{-3/2}), \quad \text{as } d \rightarrow \infty. \end{aligned}$$

Concretely, when  $d = 3 \times 1024$  as in the case of CIFAR10, this quantity is 1.00016, so Eq. (22) is quite close to being tight here.

## H. Generalization of Differential Method and Wulff Crystal Optimality Results to Bounded Variation Densities

While the regularity condition Definition E.2 covers most distributions we care about, we still have to reason separately about, e.g. uniform distributions on sets, or mixture of such distributions and regular distribution. However, regularity can be weakened to the notion of *bounded variation* to cover all such cases. Bounded variation (BV) is “essentially the weakest measure theoretic sense in which a function can be differentiable” (Evans & Gariepy, 2015). BV functions include the usual continuously differentiable functions as well as indicator functions of “finite perimeter” sets. More generally, the notion of BV allows a “controlled” amount of jump-type discontinuities. Our differential method and our Wulff Crystal optimality results can be generalized to distributions with BV densities.

Readers exposed to the notion of bounded variation for the first time might find it helpful to mentally substitute “BV” in our results below with “differentiable” or with “indicator function” on the first read through. All probability distribution densities we work with concretely in this paper have bounded variation.

**Definition H.1** (Bounded Variation). Let  $\Omega \subseteq \mathbb{R}^d$  be an open set. A function  $f \in L^1(\Omega)$  is said to be of *bounded variation* (or BV), written  $f \in \text{BV}(\Omega)$ , if there exists a finite Radon measure  $|Df|$  on  $\mathbb{R}^d$  along with a vector function  $\mathbf{n} : \mathbb{R}^d \rightarrow \mathbb{R}^d$  with  $\|\mathbf{n}(x)\|_2 = 1$  almost everywhere w.r.t.  $|Df|$ , such that, for every compactly supported, continuous differentiable  $\phi : \Omega \rightarrow \mathbb{R}^d$ , we have

$$\int_{\Omega} f(x) \operatorname{div} \phi(x) dx = - \int_{\Omega} \langle \phi, \mathbf{n}(x) \rangle |Df|(x). \quad (23)$$

We denote by  $Df$  the vector measure  $\mathbf{n}|Df|$ .

*Example H.2.* If  $u$  is differentiable, then  $Du(x)$  is just the vector measure  $\nabla u(x) dx$  and  $|Du|(x)$  is  $\|\nabla u(x)\|_2 dx$ , and Eq. (23) follows just by ordinary integration by parts. Same thing holds for Sobolev (i.e. weakly differentiable) functions.

*Example H.3.* If  $u$  is the indicator function of, for example, a ball, then  $|Du|$  is the  $(d-1)$ -dimensional Hausdorff measure supported on its boundary (a sphere), and  $\mathbf{n}(x)$  is the unit normal at  $x$  pointing inward. More generally, this characterization of  $Du$  as the boundary measure with unit normals holds when  $u$  is the indicator function of a set of *finite perimeter*.

**Definition H.4** (Sets of Finite Perimeter). A set  $U$  has *finite perimeter* if its indicator function  $\chi$  is a BV function. In this case, we write

$$\int_{\partial U} g(x) dx \stackrel{\text{def}}{=} \int g(x) |D\chi|(x) \quad (24)$$

for any Borel function  $g$ .

For sufficiently smooth sets  $U$  (like a sphere), the LHS of Eq. (24) can be interpreted as an integral over the Hausdorff measure of the topological boundary  $\partial U$ , and Eq. (24) still holds. More generally, there is a subset of the topological boundary, called the *reduced boundary* of  $U$ , containing “almost every point” of  $\partial U$ , such that the LHS of Eq. (24) can be interpreted as an integral over the Hausdorff measure of the reduced boundary. See (Evans & Gariepy, 2015) for more details.

**Coarea Formula and Weak Sard for BV Functions**  
Coarea formula also holds for BV functions.

**Theorem H.5** (Coarea Formula (Federer, 2014; Morgan, 2016)). Let  $\Omega \subseteq \mathbb{R}^d$  be an open set,  $g \in L^1(\Omega)$  be Borel, and  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  have bounded variation. Let  $U_t \stackrel{\text{def}}{=} \{x : f(x) \geq t\}$  denote the superlevel set of  $f$  at level  $t$ . Then for almost every  $t$ ,  $U_t$  has finite perimeter, and we have

$$\int g(x) |Df|(x) = \int_{\mathbb{R}} \int_{\partial U_t} g(x) dx dt. \quad (25)$$

*Example H.6.* If  $f$  is differentiable, then Eq. (25) reduces to

$$\int g(x) \|\nabla f(x)\|_2 dx = \int_{\mathbb{R}} \int_{\partial U_t} g(x) dx dt. \quad (26)$$

*Example H.7.* If  $f$  is the indicator function of a set  $U$  of finite perimeter, then both sides of Eq. (25) reduce to  $\int_{\partial U} g(x) dx$ .

We also have a converse that tells us a function is BV if almost all of its superlevel sets have finite perimeter.

**Theorem H.8** (c.f. Evans & Gariepy (2015)). Let  $\Omega \subseteq \mathbb{R}^d$  be an open set and  $f \in L^1(\Omega)$ . Let  $U_t \stackrel{\text{def}}{=} \{x : f(x) \geq t\}$  denote the superlevel set of  $f$  at level  $t$ . If almost every  $U_t$



has finite perimeter and

$$\int_{-\infty}^{\infty} \int_{\partial U_t} dx dt = \int_{-\infty}^{\infty} \text{Vol}(\partial U_t) dt < \infty,$$

then  $f$  is BV (where  $\text{Vol}(\partial U_t)$  denotes  $(d-1)$ -Hausdorff measure of the reduced boundary of  $U_t$ ).

By setting  $g$  in [Theorem H.5](#) to be the indicator function over the complement of the support of  $|Df|$ , we get

**Theorem H.9** (Weak Sard). *For any BV  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ , let  $Z$  denote the complement of the support of  $|Df|$ . Let  $U_t \stackrel{\text{def}}{=} \{x : f(x) \geq t\}$  denote the superlevel set of  $f$  at level  $t$ . Then*

$$\text{Vol}(Z \cap \partial U_t) = 0$$

for almost every  $t \in \mathbb{R}$ . Here  $\text{Vol}$  denote the Hausdorff measure of dimension  $d-1$ , and again  $\partial U_t$  denotes reduced boundary.

**Bounded Variation on Almost Every Line** Like how Sobolev functions has the ACL property, a BV function on  $\mathbb{R}^d$  is also BV on almost every line parallel to a given direction.

**Theorem H.10** (c.f. Thm 5.22 of [Evans & Gariepy \(2015\)](#)). *Let  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  be BV, and let  $u \in \mathbb{R}^d$  be some vector. Then for almost every line parallel to  $u$ , the restriction of  $f$  to that line is BV, possibly after changing values on a Lebesgue measure 0 (on that line).*

This allows to show convolution with BV functions yields a.e. differentiability.

**Lemma H.11.** *If a function  $q$  is in  $\text{BV}(\mathbb{R}^d)$ , then for every bounded measurable  $F : \mathbb{R}^d \rightarrow \mathbb{R}$ , the convolution  $F * q$  is absolutely continuous on every line, and for every vector  $u \neq 0$ ,*

$$D_u(F * q) = F * (D_u q), \quad \text{a.e.},$$

where  $D_u$  on the LHS denotes ordinary directional derivative, and  $D_u q$  on the RHS denotes the measure  $\langle \mathbf{n}, u \rangle |Dq|$ , with  $\mathbf{n}$  as in [Definition H.1](#).

Note that  $F * q$  is already bounded and continuous as the convolution of a  $L^\infty$  and a  $L^1$  function.

*Proof.* It suffices to show that  $(F * q)(x + \tau u) - (F * q)(x) = \int_0^\tau (F * D_u q)(x + tu) dt$  for every  $x$  and every unit vector

$u$ .

$$\begin{aligned} & \int_0^\tau (F * D_u q)(x + tu) dt \\ &= \int_0^\tau \int F(\hat{x}) D_u q(x - \hat{x} + tu) d\hat{x} dt \\ &= \int F(\hat{x}) \int_0^\tau D_u q(x - \hat{x} + tu) dt d\hat{x} \end{aligned} \quad (27)$$

$$\begin{aligned} &= \int F(\hat{x}) [q(x - \hat{x} + \tau u) - q(x - \hat{x})] d\hat{x} \quad (28) \\ &= F * q(x + \tau u) - F * q(x). \end{aligned}$$

In [Eq. \(27\)](#), we applied Fubini's theorem after noting that

$$\begin{aligned} & \int_0^\tau \int |F(\hat{x})| |D_u q|(x - \hat{x} + tu) d\hat{x} dt \\ & \leq \|F\|_{L^\infty} \int_0^\tau \int |Dq|(x - \hat{x} + tu) d\hat{x} dt \\ & \leq \tau \|F\|_{L^\infty} \text{TV}(|Dq|) < \infty, \end{aligned}$$

where  $\text{TV}$  denotes total variation. In [Eq. \(28\)](#), we applied the linewise BV property ([Theorem H.10](#)) of  $q$  to modify  $q$  on a null set to obtain a version  $\tilde{q}$  that is BV and right continuous on almost every line parallel to  $u$ . Then [Eq. \(28\)](#) expands as

$$\begin{aligned} & \int F(\hat{x}) \int_0^\tau D_u q(x - \hat{x} + tu) dt d\hat{x} \\ &= \int F(\hat{x}) \int_0^\tau D_u \tilde{q}(x - \hat{x} + tu) dt d\hat{x} \end{aligned} \quad (29)$$

$$= \int F(\hat{x}) [\tilde{q}(x - \hat{x} + \tau u) - \tilde{q}_-(x - \hat{x})] d\hat{x} \quad (30)$$

$$\text{where } \tilde{q}_-(x) = \lim_{t \nearrow 0} \tilde{q}(x + tu)$$

$$= \int F(\hat{x}) [\tilde{q}(x - \hat{x} + \tau u) - \tilde{q}(x - \hat{x})] d\hat{x} \quad (31)$$

$$= \int F(\hat{x}) [q(x - \hat{x} + \tau u) - q(x - \hat{x})] d\hat{x}. \quad (32)$$

Here in [Eq. \(30\)](#), we use the fact that the Lebesgue integral in  $t$  in [Eq. \(29\)](#) is equal to the Lebesgue-Stieltjes integral with integrator  $t \mapsto \tilde{q}(x - \hat{x} + tu)$ . Because on almost every line,  $\tilde{q}_-$  differs from  $\tilde{q}$  only at the points of discontinuity, of which there are only countably many,  $\tilde{q}_-$  differs from  $\tilde{q}$  on  $\mathbb{R}^d$  in a null set; this is [Eq. \(31\)](#). Finally,  $q$  differs from  $\tilde{q}$  on a null set, so [Eq. \(32\)](#) holds.  $\square$

## H.1. Differential Method for BV Densities

To generalize the differential method to distribution with BV densities, we need to define  $\Phi$  differently.

**Definition H.12.** Let  $q(x)$  be a distribution with BV density, which we also denote as  $q$ . Then  $|Dq|$  is a finite Radon measure. By Lebesgue Decomposition Theorem,  $|Dq|$  is the sum of two measures  $|Dq|_{ac}$  and  $|Dq|_s$  which are resp.

absolutely continuous and singular w.r.t.  $q$ . Thus, there is some set of  $q$ -measure 0 that has full measure under  $|Dq|_s$ .

For any vector  $u \in \mathbb{R}^d$ , let  $\gamma_u$  be the random variable given by

$$\gamma_u \stackrel{\text{def}}{=} \langle u, \mathbf{n}(\delta) \rangle \frac{d|Dq|_{ac}(\delta)}{dq(\delta)}, \delta \sim q,$$

where  $\frac{d|Dq|_{ac}(\delta)}{dq(\delta)}$  is the Radon-Nikodym derivative of  $|Dq|_{ac}$  against  $q$ , and  $\mathbf{n}$  is the vector component of  $Dq$  as in [Definition H.1](#). Define  $\varphi_u$  to be the complementary CDF of  $\gamma_u$ ,

$$\varphi_u(c) \stackrel{\text{def}}{=} \mathbb{P}[\gamma_u > c],$$

and define the inverse complementary CDF  $\varphi_u^{-1}(p)$  of  $\gamma_u$  to be

$$\varphi_u^{-1}(p) \stackrel{\text{def}}{=} \inf\{c : \mathbb{P}[\gamma_u > c] \leq p\}.$$

For any  $p \in [0, 1]$ , define a new random variable  $\gamma_u^{(p)}$  by

$$\gamma_u^{(p)} = \begin{cases} \gamma_u|_{(c, \infty)} & \text{with probability } \varphi_u(c) \\ c & \text{with probability } p - \varphi_u(c) \\ 0 & \text{with probability } 1 - p, \end{cases}$$

where  $c \stackrel{\text{def}}{=} \varphi_u^{-1}(p)$  and  $\gamma_u|_{(c, \infty)}$  is the random variable  $\gamma_u$  conditioned on  $\gamma_u > c$ .

Define

$$\vartheta_u \stackrel{\text{def}}{=} \int \max(0, \langle \mathbf{n}(x), u \rangle) |Dq|_s(x).$$

Let  $\mathcal{B}$  be the unit ball of  $\|\cdot\|$  as in [Assumption F.1](#). Then we define  $\Phi : [0, 1] \rightarrow \mathbb{R}$  by

$$\Phi(p) \stackrel{\text{def}}{=} \max_{u \in \text{Vert}(\mathcal{B})} \mathbb{E} \gamma_u^{(p)} + \vartheta_u.$$

Here,  $\vartheta_u$  represents the instantaneous growth in measure when the set has maximal allocation toward the support of  $|Dq|_s$ .

*Example H.13.* Let  $q$  be the uniform distribution on  $[0, 1]^d$ . Then  $|Dq|$  is the Hausdorff measure on the surface of the cube, which is purely singular w.r.t.  $q$ . Thus,  $|Dq| = |Dq|_s$ ,  $|Dq|_{ac} = 0$ , and  $\gamma_u = 0$ . On the other hand, for  $x$  on the boundary of the cube,  $\mathbf{n}(x)$  is the unit normal pointing into the cube, and

$$\vartheta_u = \int \max(0, \langle \mathbf{n}(x), u \rangle) |Dq|_s(x) = \text{Vol}(\Pi_u[0, 1]^d),$$

for any  $\ell_2$  unit vector  $u$ .

This example generalizes to any uniform distribution on any set  $S$  of finite perimeter, except that the last equality needs not hold if  $S$  is not convex.

With this definition of  $\Phi$ , the proof of the differential method goes through if we swap usage of [Lemma E.7](#) with [Lemma H.11](#).

**Theorem H.14** (The Differential Method for BV Densities). *Fix any norm  $\|\cdot\|$  and let  $G : \mathbb{R}^d \rightarrow [0, 1]$  be the smoothing of any measurable  $F : \mathbb{R}^d \rightarrow [0, 1]$  by  $q(x)$  with BV density. Let  $\Phi : [0, 1] \rightarrow \mathbb{R}$  be given as in [Definition H.12](#).*

*Then for any  $x$ , if  $G(x) < 1/2$ , then  $G(x + \delta) < 1/2$  for any*

$$\|\delta\| < \int_{G(x)}^{1/2} \frac{1}{\Phi(p)} dp.$$

## H.2. Wulff Crystal Optimality for BV Densities

Similarly, if we swap out usage of [Theorem E.3](#) with [Theorem H.5](#) and the usage of [Theorem E.4](#) with [Theorem H.9](#), then we generalize [Theorem G.20](#) to distributions with BV densities.

**Theorem H.15.** *Let  $\mathcal{B}$  be a full-dimensional polytope in  $\mathbb{R}^d$  symmetric around the origin, and whose vertices form a symmetric set. Let  $Z$  be the Wulff Crystal w.r.t.  $\mathcal{B}$ . Let  $q_0$  be a BV probability density function. Among all probability distributions  $q$  with BV and even density function that is level equivalent to  $q_0$ , the probability density function with concentric superlevel sets proportional to  $Z$  minimize*

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

where  $\|\cdot\|$  is the norm defined by  $\mathcal{B}$ .

Likewise, [Theorem G.21](#) generalizes similarly to BV densities.

**Theorem H.16.** *Let  $\mathcal{B}$  be a full-dimensional polytope in  $\mathbb{R}^d$  symmetric around the origin, and whose vertices form a symmetric set. Let  $Z$  be the Wulff Crystal w.r.t.  $\mathcal{B}$ , and let  $\|\cdot\|$  denote the norm with  $Z$  as its unit ball. Consider a probability distribution  $q(x)$  on  $\mathbb{R}^d$ ,  $d \geq 2$  with even, BV density that depends only on the norm  $\|x\|$ . Then with  $\Phi$  defined against the adversary  $\mathcal{B}$ , we have, for any  $k > 0$ ,*

$$\Phi(1/2) \geq \frac{(d-1)C}{\sqrt[k]{\mathbb{E}_{\delta \sim q} \|\delta\|^k}},$$

where  $C \stackrel{\text{def}}{=} \frac{1}{2} \mathbb{E}_{x \sim \partial Z} |\langle \nabla \|x\|, u \rangle|$ , for any vertex  $u$  of  $\mathcal{B}$ , is a constant that depends only on  $Z$ .

## I. Robust Radii Derivations

### I.1. IID Distributions

In this section, we study smoothing distributions that have i.i.d. coordinates.

#### I.1.1. $\ell_1$ ADVERSARY

#### IID Log Concave Distributions

**Theorem I.1.** Let  $\phi : \mathbb{R} \rightarrow \mathbb{R}$  be absolutely continuous, even, and convex such that  $\exp(-\phi(x))$  is integrable. Suppose  $H$  is a smoothed classifier smoothed by

$$q(x) \propto \prod_{i=1}^d e^{-\phi(x_i)},$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \text{CDF}_{\phi}^{-1}(\rho),$$

where  $\text{CDF}_{\phi}^{-1}$  is the inverse CDF of the 1D random variable with density  $\propto \exp(-\phi(x))$ . This robust radius is tight.

*Proof.* We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

WLOG, assume  $u \in \text{Vert}(\mathcal{B})$  is  $e_1$ . Then  $\gamma_u = \langle u, \nabla \log q(\delta) \rangle = \phi'(\delta_1)$ , for  $\delta \sim q$ . Let  $X$  be the random variable whose density function is  $\propto e^{-\phi(x)}$ , and denote  $\varphi(c) \stackrel{\text{def}}{=} \mathbb{P}[X > c]$ . Thus  $\gamma_u$  is distributed as  $\phi'(X)$ . Since  $\phi$  is convex,  $\phi'$  is nondecreasing, so that  $\gamma_u^{(p)}$  is distributed like  $\phi'(X) \mathbb{I}(X > \varphi^{-1}(p))$  (using the fact that  $X$  has an atomless measure). Then with  $C = \int_{-\infty}^{\infty} e^{-\phi(t)} dt$ , we have

$$\begin{aligned} \Phi(p) &= \mathbb{E} \gamma_u^{(p)} = C^{-1} \int_{\varphi^{-1}(p)}^{\infty} e^{-\phi(t)} \phi'(t) dt \\ &= C^{-1} e^{-\phi(\varphi^{-1}(p))}. \end{aligned}$$

Then with  $p_0 = 1 - \rho$ , and by reparametrization the integral ([Lemma F.7](#)), the certified radius is

$$\begin{aligned} \int_{p_0}^{1/2} \frac{1}{\Phi(p)} dp &= \int_{\varphi^{-1}(1/2)}^{\varphi^{-1}(p_0)} \frac{|\varphi'(c)|}{C^{-1} e^{-\phi(c)}} dc \\ &= \int_{\varphi^{-1}(1/2)}^{\varphi^{-1}(p_0)} dc \\ &= \varphi^{-1}(p_0) - \varphi^{-1}(1/2). \end{aligned}$$

Simplifying this in terms of the CDF, and noting that  $\varphi^{-1}(1/2) = 0$  because  $\phi$  is even, yields the expression in the theorem statement.

This robust radius is tight, as can be seen from the case when a half-plane  $\{x : x_1 \geq s\}$  is the set of inputs that the base classifier assigns the label  $y$ .

The same proof can be generalized straightforwardly to distributions with BV densities by using [Theorem H.14](#) (this, for example, yields another proof of the robust radii of uniform distribution against  $\ell_1$  adversary).

**Theorem I.2.** Let  $q_1 : \mathbb{R} \rightarrow \mathbb{R}$  be an even and convex function and assume it has bounded variations. Suppose  $H$  is a smoothed classifier smoothed by

$$q(x) \propto \prod_{i=1}^d q_1(x_i),$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \text{CDF}_{q_1}^{-1}(\rho),$$

where  $\text{CDF}_{q_1}^{-1}$  is the inverse CDF of the 1D random variable with density  $\propto q_1(x)$ . This robust radius is tight.

### IID Log Convex\* Distributions

**Theorem I.3.** Let  $\phi : [0, \infty) \rightarrow \mathbb{R}$  be absolutely continuous, concave, and nondecreasing, such that  $\exp(-\phi(|x|))$  is integrable. Suppose  $H$  is a smoothed classifier smoothed by

$$q(x) \propto \prod_{i=1}^d e^{-\phi(|x_i|)}$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\begin{aligned} \|\eta\|_1 &< \int_{\varphi^{-1}(1-\rho)}^{\infty} \frac{dc}{e^{\phi(c)-\phi(0)} - 1} \\ &= \int_{1-\rho}^{1/2} \frac{C dp}{e^{-\phi(0)} - e^{-\phi(\varphi^{-1}(p))}} \end{aligned}$$

where  $\varphi^{-1}$  is the inverse function of

$$\varphi(c) \stackrel{\text{def}}{=} \mathbb{P}_{z \sim q}[0 \leq z_1 \leq c],$$

and

$$C = \int_{-\infty}^{\infty} e^{-\phi(|t|)} dt.$$

*Proof.* We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

□

WLOG, assume  $u \in \text{Vert}(\mathcal{B})$  is  $e_1$ . Then  $\gamma_u$  is the random variable  $\langle u, \nabla \log q(\delta) \rangle = \phi'(|\delta_1|) \text{sgn}(\delta_1)$  where  $\delta \sim q$ . Let  $X \in \mathbb{R}$  be the random variable whose density function is  $\propto e^{-\phi(|x|)}$ , and so  $\varphi(c) = \mathbb{P}[0 \leq X \leq c]$ . Thus  $\gamma_u$  is distributed as  $\phi'(X) \text{sgn}(X)$ . Since  $\phi$  is concave,  $\phi'$  is nonincreasing, so that for  $p < 1/2$ ,  $\gamma_u^{(p)}$  is distributed as  $\phi'(X) \mathbb{I}(\varphi^{-1}(p) \geq X \geq 0)$  (using the fact that  $X$  has an atomless measure). Then with  $C = \int_{-\infty}^{\infty} e^{-\phi(|t|)} dt$ ,

$$\begin{aligned} \Phi(p) &= \mathbb{E} \gamma_u^{(p)} = C^{-1} \int_0^{\varphi^{-1}(p)} e^{-\phi(t)} \phi'(x) dt \\ &= C^{-1} (e^{-\phi(0)} - e^{-\phi(\varphi^{-1}(p))}). \end{aligned}$$

Then by change of variables  $c = \varphi^{-1}(p)$  and with  $p_0 = 1 - \rho$ , the certified radius is

$$\begin{aligned} \int_{p_0}^{1/2} \frac{1}{\Phi(p)} dp &= \int_{\varphi^{-1}(p_0)}^{\varphi^{-1}(1/2)} \frac{|\varphi'(c)|}{C^{-1}(e^{-\phi(0)} - e^{-\phi(c)})} dc \\ &= \int_{\varphi^{-1}(p_0)}^{\varphi^{-1}(1/2)} \frac{e^{-\phi(c)}}{e^{-\phi(0)} - e^{-\phi(c)}} dc \\ &= \int_{\varphi^{-1}(p_0)}^{\varphi^{-1}(1/2)} \frac{1}{e^{\phi(c)-\phi(0)} - 1} dc. \end{aligned}$$

Since  $\phi$  is even,  $\varphi^{-1}(1/2) = \infty$ , yielding the desired result.  $\square$

**Corollaries** The  $\ell_p$  based exponential distribution  $\propto e^{-\|x\|_p^p}$  has each coordinate is distributed as Rademacher  $\cdot \sqrt[p]{\text{Gamma}(1/p)}$ . When  $p \geq 1$ , it satisfies [Theorem I.1](#), so we obtain

**Corollary I.4.** *Suppose  $H$  is a smoothed classifier smoothed by*

$$q(x) \propto e^{-\|x/\lambda\|_p^p}, p \geq 1,$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \lambda \sqrt[p]{\text{GammaCDF}^{-1}(2\rho - 1; 1/p)},$$

where  $\text{CDF}_\phi^{-1}$  is the inverse CDF of the 1D random variable with density  $\propto \exp(-\phi(x))$ . This robust radius is tight.

When  $p < 1$ , it satisfies [Theorem I.3](#), so we obtain

**Corollary I.5.** *Suppose  $H$  is a smoothed classifier smoothed by*

$$q(x) \propto e^{-\|x/\lambda\|_p^p}, p < 1,$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\begin{aligned} \|\eta\|_1 &< \lambda \int_{\varphi^{-1}(1-\rho)}^{\infty} \frac{dc}{e^{c^p} - 1} \\ &= \lambda \int_{1-\rho}^{1/2} \frac{2\Gamma(1 + \frac{1}{p}) dp_0}{1 - e^{-|\varphi^{-1}(p_0)|^p}}, \end{aligned}$$

where  $\varphi^{-1}(p_0) \stackrel{\text{def}}{=} \text{GammaCDF}^{-1}(2p_0; 1/p)^{1/p}$ .

The integral above can be evaluated explicitly for inverse integer  $p = 1/k$ . We show a few examples below:

$$\begin{aligned} p = 1/2: \quad R &= 2\lambda(-c \log(1 - e^{-c}) + \text{polylog}(2, e^{-c})) \\ p = 1/3: \quad R &= 3\lambda \left( -c^2 \log(1 - e^{-c}) \right. \\ &\quad \left. + 2c \text{polylog}(2, e^{-c}) + 2 \text{polylog}(3, e^{-c}) \right) \\ p = 1/4: \quad R &= 4\lambda \left( -c^3 \log(1 - e^{-c}) \right. \\ &\quad \left. + 3c^2 \text{polylog}(2, e^{-c}) + 6c \text{polylog}(3, e^{-c}) \right. \\ &\quad \left. + 6 \text{polylog}(4, e^{-c}) \right) \end{aligned}$$

where  $c = \text{GammaCDF}^{-1}(2(1 - \rho); 1/p)$  for each  $p$ , and  $\text{polylog}$  is the function defined as

$$\text{polylog}(n, z) = \sum_{k=1}^{\infty} z^k / k^n.$$

## I.2. $\ell_\infty$ Norm-Based Exponential Law

In this section, we derive robustness guarantees for distributions of the form  $q(x) \propto \|x\|_\infty^{-j} \exp(-\|x\|_\infty^k)$ .

### I.2.1. $\ell_1$ ADVERSARY

In this section, we set the norm  $\|\cdot\|$  in [Assumption F.1](#) to be the  $\ell_1$  norm  $\|x\|_1 = \sum_{i=1}^d |x_i|$ . Then the unit ball  $\mathcal{B}$  in [Assumption F.1](#) is the convex hull of its vertices which are the coordinates vectors and their negations:

$$\text{Vert}(\mathcal{B}) = \{\pm e_i : i \in [d]\}.$$

**Overview**  $\ell_\infty$  norm-based distributions will in general have certified radius that is linear in  $\rho - 1/2$ , where  $\rho$  is the probability that the *smoothed* classifier assigns to the correct class.

We first demonstrate the differential method on  $q(x) \propto \exp(-\|x\|_\infty)$  as a warmup before stating the more general result.

**Theorem I.6.** Suppose  $H$  is a smoothed classifier smoothed by

$$q(x) \propto \exp(-\|x\|_\infty/\lambda),$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \begin{cases} 2d\lambda(\rho - \frac{1}{2}) & \text{if } \rho \leq 1 - \frac{1}{2d} \\ \lambda \log \frac{1}{2d(1-\rho)} + \lambda(d-1) & \text{if } \rho > 1 - \frac{1}{2d}. \end{cases}$$

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ . Here, we have  $q(x) \propto \exp(\psi(x))$  with

$$\psi(x) = \|x\|_\infty \quad \text{and} \quad \nabla\psi(x) = \text{sgn}(x_{i^*})e_{i^*},$$

where  $i^* = \text{argmax}_i |x_i|$ , and  $e_{i^*}$  is the  $i^*$ th coordinate vector, with  $\nabla\psi(x)$  defined whenever  $i^*$  is the unique argmax.

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

For any  $u \in \text{Vert}(\mathcal{B})$  (i.e.  $u = \pm e_i$ ), the random variable  $\gamma_u = \langle u, \nabla\psi(\delta) \rangle = \langle u, \text{sgn}(\delta_{i^*})e_{i^*} \rangle$ ,  $\delta \sim q$ , is given by

$$\gamma_u = \begin{cases} 0 & \text{with prob. } 1 - \frac{1}{d} \\ 1 & \text{with prob. } \frac{1}{2d} \\ -1 & \text{with prob. } \frac{1}{2d}. \end{cases}$$

Therefore, for  $p \in [0, 1/2]$ , the random variable  $\gamma_u^{(p)}$  defined in [Definition F.4](#) is

$$\begin{cases} \gamma_u^{(p)} = \begin{cases} 1 & \text{with prob. } \frac{1}{2d} \\ 0 & \text{with prob. } 1 - \frac{1}{2d} \end{cases} & \text{if } p \in [\frac{1}{2d}, \frac{1}{2}], \\ \gamma_u^{(p)} = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1 - p \end{cases} & \text{if } p \in [0, \frac{1}{2d}]. \end{cases}$$

Thus, for any  $u \in \text{Vert}(\mathcal{B})$ ,

$$\Phi(p) = \mathbb{E} \gamma_u^{(p)} = \begin{cases} \frac{1}{2d} & \text{if } p \in [\frac{1}{2d}, \frac{1}{2}] \\ p & \text{if } p \in [0, \frac{1}{2d}]. \end{cases}$$

Then, by setting  $G(x)$  in [Theorem F.6](#) to be  $1 - H(x)_y = 1 - \rho \stackrel{\text{def}}{=} p_0$ , we get the provably robust radius of

$$\begin{aligned} & \int_{p_0}^{1/2} \frac{1}{\Phi(p)} dp \\ &= \begin{cases} \int_{p_0}^{1/2} 2d dp = 2d(\frac{1}{2} - p_0) & \text{if } p_0 \geq \frac{1}{2d} \\ \int_{p_0}^{1/2d} p^{-1} dp + \int_{1/2d}^{1/2} 2d dp & \text{if } p_0 \leq \frac{1}{2d}. \end{cases} \end{aligned}$$

Simplifying the arithmetics yields the desired claim.  $\square$

Now we tackle the general case.

**Theorem I.7.** Suppose  $H$  is a smoothed classifier smoothed by

$$q(x) \propto (\|x\|_\infty/\lambda)^{-j} \exp(-(\|x\|_\infty/\lambda)^k), \quad k \geq 1,$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y \in (1/2, 1 - \frac{1}{2d}]$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \frac{2d\lambda}{d-1} \frac{\Gamma\left(\frac{d-j}{k}\right)}{\Gamma\left(\frac{d-1-j}{k}\right)} \left(\rho - \frac{1}{2}\right).$$

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ . Here, we have

$$q(x) \propto \exp(-\|x\|_\infty^k - j \log \|x\|_\infty) \quad \text{so that}$$

$$\psi(x) = \|x\|_\infty^k + j \log \|x\|_\infty$$

$$\nabla\psi(x) = (k\|x\|_\infty^{k-1} + j\|x\|_\infty^{-1}) \text{sgn}(x_{i^*})e_{i^*},$$

where  $i^* = \text{argmax}_i |x_i|$ , and  $e_{i^*}$  is the  $i^*$ th coordinate vector, with  $\nabla\psi(x)$  defined whenever  $i^*$  is the unique argmax.

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

WLOG among  $\text{Vert}(\mathcal{B})$ , let's assume  $u = e_1$ . Then the random variable  $\gamma_u = \langle u, \nabla\psi(\delta) \rangle$ ,  $\delta \sim q$ , is 0 with probability  $1 - \frac{1}{d}$ , when  $i^* \neq 1$ . When  $i^* = 1$  and  $\text{sgn}(x_{i^*}) = 1$  (which happens with probability  $\frac{1}{2d}$ ),  $\gamma_u$  is  $k\|x\|_\infty^{k-1} + j\|x\|_\infty^{-1}$ , where  $x \sim q$ . By [Lemma I.25](#), this is just the random variable  $kz^{\frac{k-1}{k}} + jz^{-1}$ , where  $z \sim \text{Gamma}(d/k)$ . Likewise, with probability  $\frac{1}{2d}$ ,  $\gamma_u$  is the random variable  $-kz^{\frac{k-1}{k}} - jz^{-1}$ . This can be summarized below.

$$\gamma_u = \begin{cases} 0 & \text{with prob. } 1 - \frac{1}{d} \\ kz^{\frac{k-1}{k}} + jz^{-1} & \text{with prob. } \frac{1}{2d} \\ -kz^{\frac{k-1}{k}} - jz^{-1} & \text{with prob. } \frac{1}{2d}, \end{cases}$$

where  $z \sim \text{Gamma}(d/k)$ .

Therefore, for  $p \in [\frac{1}{2d}, \frac{1}{2}]$ , the random variable  $\gamma_u^{(p)}$  defined in [Definition F.4](#) is

$$\gamma_u^{(p)} = \begin{cases} kz^{\frac{k-1}{k}} + jz^{-1} & \text{with prob. } \frac{1}{2d} \\ 0 & \text{with prob. } 1 - \frac{1}{2d} \end{cases}$$

where  $z$  is sampled from  $\text{Gamma}(d/k)$ .

Thus, for any  $u \in \text{Vert}(\mathcal{B})$ , by [Lemma I.26](#),

$$\Phi(p) = \mathbb{E} \gamma_u^{(p)} = \frac{1}{2d} \mathbb{E} kz^{\frac{k-1}{k}} = \frac{d-1}{2d} \frac{\Gamma\left(\frac{d-1-j}{k}\right)}{\Gamma\left(\frac{d-j}{k}\right)}$$

which does not depend on  $p$ .

Then, by setting  $G(x)$  in [Theorem F.6](#) to be  $1 - H(x)_y = 1 - \rho$ , we get the provably robust radius of

$$\int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp = \frac{2d}{d-1} \frac{\Gamma\left(\frac{d-j}{k}\right)}{\Gamma\left(\frac{d-1-j}{k}\right)} \left(\rho - \frac{1}{2}\right)$$

as desired.  $\square$

As  $j = 0$  and  $k \rightarrow \infty$ , the distribution above converges to the uniform distribution, and the robust certificate converges likewise to the one computed previous for uniform distribution.

**Theorem I.8** (Lee et al. (2019)). *Suppose  $H$  is a smoothed classifier smoothed by the uniform distribution on the cube  $[-\lambda, \lambda]^d$ , such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with*

$$\|\eta\|_1 < 2\lambda \left(\rho - \frac{1}{2}\right).$$

### I.2.2. $\ell_\infty$ ADVERSARY

In this section, we set the norm  $\|\cdot\|$  in [Assumption F.1](#) to be the  $\ell_\infty$  norm  $\|x\|_\infty = \max_{i=1}^d |x_i|$ . Then the unit ball  $\mathcal{B}$  in [Assumption F.1](#) is the convex hull of its vertices which are points in the Boolean cube:

$$\text{Vert}(\mathcal{B}) = \{\pm 1\}^d.$$

**Theorem I.9.** *Suppose  $H$  is a smoothed classifier smoothed by*

$$q(x) \propto \exp(-\|x\|_\infty/\lambda),$$

*such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with*

$$\|\eta\|_\infty < \lambda \log \frac{1}{2(1-\rho)}.$$

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

For any  $u \in \text{Vert}(\mathcal{B})$ , the random variable  $\gamma_u = \langle u, \nabla \psi(\delta) \rangle = \langle u, \text{sgn}(\delta_{i^*}) e_{i^*} \rangle$ ,  $\delta \sim q$ , is a Rademacher

random variable taking values  $\pm 1$  with equal probability. Therefore, for  $p \in [0, 1/2]$ , the random variable  $\gamma_u^{(p)}$  defined in [Definition F.4](#) is

$$\gamma_u^{(p)} = \begin{cases} 1 & \text{with prob. } p \\ 0 & \text{with prob. } 1 - p. \end{cases}$$

Thus, for any  $u \in \text{Vert}(\mathcal{B})$ ,

$$\Phi(p) = \mathbb{E} \gamma_u^{(p)} = p.$$

Then, by setting  $G(x)$  in [Theorem F.6](#) to be  $1 - H(x)_y = 1 - \rho$ , we get the provably robust radius of

$$\int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp = \int_{1-\rho}^{1/2} \frac{1}{p} dp = \log \frac{1}{2(1-\rho)}.$$

$\square$

**Theorem I.10.** *Suppose  $H$  is a smoothed classifier smoothed by*

$$q(x) \propto \exp(-\|x/\lambda\|_\infty^k), k \geq 1$$

*such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with*

$$\|\eta\|_\infty < \lambda \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp, \quad (33)$$

in which

$$\Phi(p) \stackrel{\text{def}}{=} C \left( 1 - \text{GammaCDF} \left( c^*(p); \frac{d+k-1}{k} \right) \right),$$

$$\text{where } c^*(p) \stackrel{\text{def}}{=} \text{GammaCDF}^{-1} \left( 1 - 2p; \frac{d}{k} \right),$$

$$C \stackrel{\text{def}}{=} \frac{k \Gamma\left(\frac{d+k-1}{k}\right)}{2 \Gamma\left(\frac{d}{k}\right)}.$$

More generally, if the smoothing distribution is

$$q(x) \propto \|x/\lambda\|_\infty^{-j} \exp(-\|x/\lambda\|_\infty^k), k \geq 1, j < d-1,$$

then  $H$  is robust against  $\ell_\infty$  perturbation

$$\|\eta\|_\infty < \lambda \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp, \quad (34)$$

where

$$\Phi(p) \stackrel{\text{def}}{=} \frac{1}{2} \bar{\phi}(\phi^{-1}(2p)), \quad \text{where}$$

$$\phi(c) \stackrel{\text{def}}{=} \mathbb{P}[\gamma > c]$$

$$\bar{\phi}(c) \stackrel{\text{def}}{=} \mathbb{E} \gamma \mathbb{I}(\gamma > c)$$

and  $\gamma \stackrel{\text{def}}{=} (k-1)\xi^{\frac{k-1}{k}} + j\xi^{-\frac{1}{k}}$ ,  $\xi \sim \text{Gamma}\left(\frac{d}{k} - \frac{j}{k}\right)$ .

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

For any  $u \in \text{Vert}(\mathcal{B})$ , we have

$$\begin{aligned}\gamma_u &= \langle u, -\nabla \log q(\delta) \rangle \\ &= \langle u, (k \|\delta\|_\infty^{k-1} + j \|\delta\|_\infty^{-1}) \text{sgn}(\delta_{i^*}) e_{i^*} \rangle, \delta \sim q.\end{aligned}$$

Since  $\|\delta\|_\infty$  is distributed as  $\sqrt[k]{\text{Gamma}(\frac{d}{k} - \frac{j}{k})}$  and  $\langle u, \text{sgn}(\delta_{i^*}) e_{i^*} \rangle$  is  $\pm 1$  with equal probability,  $\gamma_u$  is distributed as the product of random variables

$$\begin{aligned}\gamma_u &= \zeta (k \xi^{\frac{k-1}{k}} + j \xi^{-\frac{1}{k}}), \\ \zeta &\sim \text{Rademacher}, \xi \sim \text{Gamma}\left(\frac{d}{k} - \frac{j}{k}\right).\end{aligned}$$

Let  $\varphi(c) \stackrel{\text{def}}{=} \mathbb{P}[\gamma_u > c]$ . Then for  $p < 1/2$ ,  $\varphi^{-1}(2p) = \varphi^{-1}(p)$ . Since  $\gamma_u$  has an absolutely continuous distribution, the variable  $\gamma_u^{(p)} = \gamma_u |_{(\varphi^{-1}(p), \infty)}$  with probability  $p$ , and 0 otherwise. Thus

$$\begin{aligned}\Phi(p) &= \mathbb{E} \gamma_u^{(p)} = \bar{\varphi}(\varphi^{-1}(p)), \quad \text{where} \\ \bar{\varphi}(c) &= \mathbb{E} \gamma_u \mathbb{I}(\gamma_u > c).\end{aligned}$$

Note that  $\bar{\varphi}(c) = \frac{1}{2} \bar{\phi}(c)$ . Plugging into [Theorem F.6](#) yields [Eq. \(34\)](#).

**Assuming  $j = 0$**  If  $j = 0$ ,  $\gamma_u = \zeta k \xi^{\frac{k-1}{k}}$ ,  $\zeta \sim \text{Rademacher}$ ,  $\xi \sim \text{Gamma}(\frac{d}{k})$ . Then for  $p < 1/2$ ,

$$\begin{aligned}\bar{\varphi}(c) &= \frac{k}{2} \mathbb{E} \xi^{\frac{k-1}{k}} \mathbb{I}(\xi > c^*), \\ \text{where } c^* &= \text{GammaCDF}^{-1}\left(1 - 2p; \frac{d}{k}\right) \\ &= C \left(1 - \text{GammaCDF}\left(c^*; \frac{d+k-1}{k}\right)\right),\end{aligned}$$

where  $C = \frac{k}{2} \frac{\Gamma(\frac{d+k-1}{k})}{\Gamma(\frac{d}{k})}$ . Plugging into [Theorem F.6](#) yields [Eq. \(33\)](#).  $\square$

Compare this with the uniform case below.

**Theorem I.11** ([Lee et al. \(2019\)](#)). *Suppose  $H$  is a smoothed classifier smoothed by the uniform distribution on the cube  $[-\lambda, \lambda]^d$ , such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with*

$$\|\eta\|_\infty < 2\lambda \left(1 - \sqrt{\frac{3}{2} - \rho}\right).$$

When  $d \rightarrow \infty$ , this robust radius is roughly

$$\begin{aligned}2\lambda \left(1 - e^{\frac{1}{d} \log[1 - (\rho - \frac{1}{2})]}\right) \\ \approx 2\lambda \left(1 - \left(1 + \frac{1}{d} \log\left[1 - \left(\rho - \frac{1}{2}\right)\right]\right)\right) \\ \approx \frac{2\lambda}{d} \left(\rho - \frac{1}{2}\right).\end{aligned}$$

On the other hand, when  $k \rightarrow \infty$  in [Eq. \(33\)](#), we see that

1.  $d/k \rightarrow 0$  while  $\frac{d+k-1}{k} \rightarrow 1$
2.  $c^* \rightarrow 0$  for any  $p < 1/2$
3.  $\text{GammaCDF}(c^*; \frac{d+k-1}{k}) \rightarrow \text{GammaCDF}(0; 1) = 1$  consequently
4. by simple calculation  $k \frac{\Gamma(\frac{d+k-1}{k})}{\Gamma(\frac{d}{k})} \rightarrow d$
5. so  $\Phi(p) \rightarrow d/2$  for any  $p < 1/2$ .

Therefore, when  $k \rightarrow \infty$ , the  $\ell_\infty$  robust radius in [Eq. \(33\)](#) converges to

$$\frac{2\lambda}{d} \left(\rho - \frac{1}{2}\right)$$

as well.

### I.3. $\ell_\infty$ Norm-Based Power Law

Now consider a power law of the  $\ell_\infty$  norm: For  $a > d$ ,

$$\begin{aligned}q(x) &\propto (1 + \|x\|_\infty)^{-a} && \text{so that} \\ \psi(x) &= a \log(1 + \|x\|_\infty) \\ \nabla \psi(x) &= a(1 + \|x\|_\infty)^{-1} \text{sgn}(x_{i^*}) e_{i^*},\end{aligned}$$

where  $i^* = \text{argmax}_i |x_i|$ , and  $e_{i^*}$  is the  $i^*$ th coordinate vector, with  $\nabla \psi(x)$  defined whenever  $i^*$  is the unique argmax. Note that the  $\ell_\infty$  norm of vector sampled from  $q$  has distribution with CDF

$$\mathbb{P}_{\delta \sim q}[\|\delta\|_\infty \leq c] = \frac{\Gamma(a)}{\Gamma(a-d)\Gamma(d)} \int_0^c \frac{r^{d-1}}{(1+r)^a} dr. \quad (35)$$

This is known as the *Beta prime or Beta distribution of the second kind*, with shape parameters  $\alpha = d, \beta = a - d$ , which we will denote by  $\text{BetaPrime}(d, a - d)$ . If  $a > d + 1$ , this distribution has mean

$$\mathbb{E}_{\delta \sim q} \|\delta\|_\infty = \frac{d}{a - d - 1}.$$

#### I.3.1. $\ell_1$ ADVERSARY

**Theorem I.12.** *Suppose  $H$  is a smoothed classifier smoothed by*

$$q(x) \propto (1 + \|x\|_\infty / \lambda)^{-a}, a > d,$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \lambda \frac{2d}{a-d} \left( \rho - \frac{1}{2} \right).$$

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

WLOG among  $\text{Vert}(\mathcal{B})$ , let's assume  $u = e_1$ . Then the random variable  $\gamma_u = \langle u, \nabla \psi(\delta) \rangle, \delta \sim q$ , is 0 with probability  $1 - \frac{1}{d}$ , when  $i^* \neq 1$ . When  $i^* = 1$  and  $\text{sgn}(x_{i^*}) = 1$  (which happens with probability  $\frac{1}{2d}$ ),  $\gamma_u$  is distributed as  $a(1+r)^{-1}$ , where  $r$  has CDF [Eq. \(35\)](#). Likewise, with probability  $\frac{1}{2d}$ ,  $\gamma_u$  is distributed as  $-a(1+r)^{-1}$ .

This can be summarized below.

$$\gamma_u = \begin{cases} 0 & \text{with prob. } 1 - \frac{1}{d} \\ a(1+r)^{-1} & \text{with prob. } \frac{1}{2d} \\ -a(1+r)^{-1} & \text{with prob. } \frac{1}{2d}, \end{cases}$$

where  $r$  is a random variable with CDF [Eq. \(35\)](#).

Therefore, for  $p \in [\frac{1}{2d}, \frac{1}{2}]$ , the random variable  $\gamma_u^{(p)}$  defined in [Definition F.4](#) is

$$\gamma_u^{(p)} = \begin{cases} a(1+r)^{-1} & \text{with prob. } \frac{1}{2d} \\ 0 & \text{with prob. } 1 - \frac{1}{2d}. \end{cases}$$

Thus, for any  $u \in \text{Vert}(\mathcal{B})$ , by [Lemma I.26](#),

$$\begin{aligned} \Phi(p) &= \mathbb{E} \gamma_u^{(p)} = \frac{1}{2d} \mathbb{E} a(1+r)^{-1} \\ &= \frac{1}{2d} \frac{a\Gamma(a)}{\Gamma(a-d)\Gamma(d)} \int_0^\infty a \frac{r^{d-1}}{(1+r)^{a+1}} dr \\ &= \frac{1}{2d} \frac{\Gamma(a+1)}{\Gamma(a-d)\Gamma(d)} \frac{\Gamma(a+1-d)\Gamma(d)}{\Gamma(a+1)} \\ &= \frac{a-d}{2d}. \end{aligned}$$

which does not depend on  $p$ .

Then, by setting  $G(x)$  in [Theorem F.6](#) to be  $1 - H(x)_y = 1 - \rho$ , we get the provably robust radius of

$$\int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp = \frac{2d}{a-d} \left( \rho - \frac{1}{2} \right)$$

as desired.  $\square$

### I.3.2. $\ell_\infty$ ADVERSARY

**Theorem I.13.** Suppose  $H$  is a smoothed classifier smoothed by

$$q(x) \propto (1 + \|x\|_\infty / \lambda)^{-a}, a > d,$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \frac{2\lambda}{a-d} \int_{1-\rho}^{1/2} \frac{dp}{\Upsilon(\Upsilon^{-1}(2p; d, a-d); d, a+1-d)},$$

where  $\Upsilon = \text{BetaPrimeCDF}$ .

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

For any  $u \in \text{Vert}(\mathcal{B})$ , we have

$$\begin{aligned} \gamma_u &= \langle u, -\nabla \log q(\delta) \rangle \\ &= \langle u, a(1 + \|\delta\|_\infty)^{-1} \text{sgn}(\delta_{i^*}) e_{i^*} \rangle, \delta \sim q. \end{aligned}$$

Since  $\|\delta\|_\infty$  is distributed as  $\text{BetaPrime}(d, a-d)$  and  $\langle u, \text{sgn}(\delta_{i^*}) e_{i^*} \rangle$  is  $\pm 1$  with equal probability,  $\gamma_u$  is distributed as the product of random variables

$$\begin{aligned} \gamma_u &= \zeta a(1 + \xi)^{-1}, \\ \zeta &\sim \text{Rademacher}, \xi \sim \text{BetaPrime}(d, a-d). \end{aligned}$$

Since  $r \mapsto (1+r)^{-1}$  is a decreasing function on  $r \in [0, \infty)$ , we have, for  $p < 1/2$ ,

$$\begin{aligned} \Phi(p) &= \mathbb{E} \gamma_u^{(p)} = \frac{1}{2} \mathbb{E}_\xi a(1 + \xi)^{-1} \mathbb{I}(\xi < c(p)), \\ &\text{where } c(p) = \text{BetaPrimeCDF}^{-1}(2p; d, a-d). \end{aligned}$$

Of course, we can simplify

$$\begin{aligned} &\mathbb{E}_\xi (1 + \xi)^{-1} \mathbb{I}(\xi < c) \\ &= \frac{\Gamma(a)}{\Gamma(a-d)\Gamma(d)} \int_0^c \frac{r^{d-1}}{(1+r)^{a+1}} dr \\ &= \frac{\Gamma(a)}{\Gamma(a-d)\Gamma(d)} \frac{\Gamma(a+1-d)\Gamma(d)}{\Gamma(a+1)} \times \\ &\quad \text{BetaPrimeCDF}(c; d, a+1-d) \\ &= \frac{(a-d)}{a} \text{BetaPrimeCDF}(c; d, a+1-d). \end{aligned}$$

Therefore,

$$\Phi(p) = \frac{a-d}{2} \text{BetaPrimeCDF}(c(p); d, a+1-d).$$

Plugging into [Theorem F.6](#) yields the desired result.  $\square$



#### I.4. $\ell_1$ Norm-Based Exponential Law

Consider the following generalization of the Laplace distribution

$$\begin{aligned} q(x) &\propto \exp(-\|x\|_1^k) \quad \text{so that} \\ \psi(x) &= \|x\|_1^k \\ \nabla\psi(x) &= k\|x\|_1^{k-1}(\text{sgn}(x_1), \dots, \text{sgn}(x_d)), \end{aligned}$$

with  $\nabla\psi(x)$  defined whenever all  $x_i$ s are nonzero.

##### I.4.1. $\ell_1$ ADVERSARY

**Theorem I.14.** *Suppose  $H$  is a smoothed classifier smoothed by*

$$q(x) \propto \exp(-(\|x\|_1/\lambda)^k),$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \lambda \int_{1-\rho}^{1/2} \frac{R}{\Psi(p)} dp.$$

Here  $R = \frac{2\Gamma(\frac{d}{k})}{k\Gamma(\frac{d+k-1}{k})}$ , and

$$\Psi(p) \stackrel{\text{def}}{=} \begin{cases} 1 - \Upsilon(\Upsilon^{-1}(1-2p; \frac{d}{k}); \frac{d+k-1}{k}) & \text{if } k \geq 1 \\ \Upsilon(\Upsilon^{-1}(2p; \frac{d}{k}); \frac{d+k-1}{k}) & \text{if } k \in (0, 1). \end{cases}$$

where  $\Upsilon = \text{GammaCDF}$ .

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

For any  $u \in \text{Vert}(\mathcal{B})$  (i.e.  $u = \pm e_i$ ),  $\gamma_u = \langle u, \nabla\psi(\delta) \rangle = \pm k\|\delta\|_1^{k-1}$ ,  $\delta \sim q$ , takes positive or negative sign with equal probability. By [Lemma I.25](#),  $\|\delta\|_1$  is the random variable  $\Gamma(d/k)^{1/k}$ . Therefore,  $\gamma_u$  is distributed as  $k\text{Gamma}(d/k)^{\frac{k-1}{k}} \text{Rademacher}(1/2)$ .

Therefore, for  $p \in [0, 1/2]$ , the random variable  $\gamma_u^{(p)}$  defined in [Definition F.4](#) is

$$\gamma_u^{(p)} = \begin{cases} kz_p^{\frac{k-1}{k}} & \text{with prob. } p \\ 0 & \text{with prob. } 1-p, \end{cases}$$

where, if  $k \geq 1$ ,  $z_p$  is sampled from  $\text{Gamma}(d/k)$  conditioned on  $z_p > \Upsilon^{-1}(1-2p; d/k)$  (because  $z^{\frac{k-1}{k}}$  is increasing in  $z$ ), but if  $k \in (0, 1)$ , then  $z_p$  is sampled

from  $\text{Gamma}(d/k)$  conditioned on  $z_p < \Upsilon^{-1}(2p; d/k)$  (because  $z^{\frac{k-1}{k}}$  is decreasing in  $z$ ).

Thus, for any  $u \in \text{Vert}(\mathcal{B})$ ,

$$\Phi(p) = \mathbb{E} \gamma_u^{(p)} = \begin{cases} \frac{k}{2} \mathbb{E} z^{\frac{k-1}{k}} \mathbb{I}(z > \Upsilon^{-1}(1-2p)) & \text{if } k \geq 1 \\ \frac{k}{2} \mathbb{E} z^{\frac{k-1}{k}} \mathbb{I}(z < \Upsilon^{-1}(2p)) & \text{if } k < 1 \end{cases}$$

where  $z \sim \text{Gamma}(d/k)$ . This integral simplifies to  $R^{-1}\Psi(p)$  (with  $\Psi$  taking different forms depending on  $k$ ) by [Lemma I.26](#).

Then, by setting  $G(x)$  in [Theorem F.6](#) to be  $1 - H(x)_y = 1 - \rho$ , we get the provably robust radius of

$$\int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp = \int_{1-\rho}^{1/2} \frac{R}{\Psi(p)} dp.$$

□

##### I.4.2. $\ell_\infty$ ADVERSARY

We first start with the Laplace distribution to highlight the basic logic behind the  $\ell_\infty$  radius derivation.

**Theorem I.15.** *Suppose  $H$  is a smoothed classifier smoothed by the Laplace distribution*

$$q(x) \propto \exp(-\|x\|_1/\lambda),$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_\infty < \lambda \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp,$$

where

$$\Phi(p) = c(p - \phi_d(c)) + d\phi_{d-1}\left(c - \frac{1}{2}\right) - d\phi_d(c),$$

in which  $c = \phi_d^{-1}(p)$ , and

$$\begin{aligned} \phi_d(c) &\stackrel{\text{def}}{=} 2^{-d} \sum_{i=\frac{c+d}{2}+1}^d \binom{d}{i} \\ &= 1 - \text{BinomCDF}\left(\frac{c+d}{2}; d\right) \end{aligned}$$

for any  $c \equiv d \pmod{2}$

$$\begin{aligned} \phi_d^{-1}(p) &\stackrel{\text{def}}{=} \inf\{c : \phi_d(c) \leq p\} \\ &= 2\text{BinomCDF}^{-1}(1-p) - d. \end{aligned}$$

Note that when  $d$  is large,

$$\Phi(p) \approx \text{GaussianCDF}'(\text{GaussianCDF}^{-1}(p))\sqrt{d},$$

so that the the bound above is roughly

$$\|\delta\|_\infty < \lambda \text{GaussianCDF}^{-1}(\rho)/\sqrt{d}.$$

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

WLOG, let  $u \in \text{Vert}(\mathcal{B})$  be  $u = (1, \dots, 1)$ ; arguments for other  $u \in \text{Vert}(\mathcal{B})$  proceeds similarly. For this  $u$ ,  $\gamma_u = \langle u, -\nabla \log q(\delta) \rangle$ ,  $\delta \sim q$ , is a sum of independent Rademacher random variable  $\gamma_u = \sum_{i=1}^d R_i$ , where each  $R_i$  independently takes values 1 and  $-1$  with equal probability. Thus  $\gamma_u$  is distributed like  $2B_d - d$ , where  $B_d$  is the binomial random variable corresponding to the number of heads in  $d$  coin tosses. Then for any integer  $c$  with the same parity as  $d$ ,  $\phi_d(c)$  is the complementary CDF of  $\gamma_u$  and  $\phi_d^{-1}$  is the corresponding inverse CDF. Then, for  $p \in [0, 1/2]$ , we have

$$\gamma_u^{(p)} = \begin{cases} \gamma_u|_{(c, \infty)} & \text{with probability } \phi_d(c) \\ c & \text{with probability } p - \phi_d(c) \\ 0 & \text{with probability } 1 - p, \end{cases}$$

where  $c \stackrel{\text{def}}{=} \phi_d^{-1}(p)$  and  $\gamma_u|_{(c, \infty)}$  is the random variable  $\gamma_u$  conditioned on  $\gamma_u > c$ . Therefore,

$$\begin{aligned} \Phi(p) &= \mathbb{E} \gamma_u^{(p)} = c(p - \phi_d(c)) + 2^{-d} \sum_{i=\frac{c+d}{2}+1}^d (2i-d) \binom{d}{i} \\ &= c(p - \phi_d(c)) + 2^{-d} \sum_{i=\frac{c+d}{2}+1}^d 2d \binom{d-1}{i-1} - d \binom{d}{i} \\ &= c(p - \phi_d(c)) + d\phi_{d-1}\left(c - \frac{1}{2}\right) - d\phi_d(c) \end{aligned}$$

Then, by setting  $G(x)$  in [Theorem F.6](#) to be  $1 - H(x)_y = 1 - \rho$ , we get the desired robust radius.  $\square$

**Theorem I.16.** *Suppose  $H$  is a smoothed classifier smoothed by*

$$q(x) \propto \exp(-\|x/\lambda\|_1^k), k > 1,$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_\infty < \lambda \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp,$$

where

$$\Phi(p) = \mathbb{E} \gamma \mathbb{I}(\gamma > \varphi^{-1}(p)), \quad \text{with}$$

$$\gamma = \left( \sum_{i=1}^d \zeta_i \right) k \xi^{\frac{k-1}{k}},$$

$$\zeta_i \sim \text{Rademacher}, \xi \sim \text{Gamma}(d/k)$$

$$\varphi(c) = \mathbb{P}[\gamma > c].$$

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

WLOG, let  $u \in \text{Vert}(\mathcal{B})$  be  $u = (1, \dots, 1)$ ; arguments for other  $u \in \text{Vert}(\mathcal{B})$  proceeds similarly. As in the proof of [Theorem I.15](#), we find  $\gamma_u = \langle u, -\nabla \log q(\delta) \rangle = \langle u, k \|\delta\|_1^{k-1} \text{sgn}(\delta) \rangle$ ,  $\delta \sim q$ , is distributed like  $\gamma$  in the theorem statement — a product of sum of Rademacher variables (coming from  $\langle u, \text{sgn}(\delta) \rangle$ ) and  $k \xi^{\frac{k-1}{k}}$ ,  $\xi \sim \Gamma(d/k)$  (coming from  $k \|\delta\|_1^{k-1}$ ). Because  $k > 1$ ,  $\gamma$ 's distribution is absolutely continuous (as it's a mixture of scaled versions of  $\xi^{\frac{k-1}{k}}$ 's distribution, which is absolutely continuous). Therefore, the random variable  $\gamma_u^{(p)} = \gamma \mathbb{I}(\gamma > \varphi^{-1}(p))$ . Then the theorem statement follows straightforwardly from [Theorem F.6](#).  $\square$

## I.5. Pareto Distribution

For  $a, \lambda > 0$  and  $x \in \mathbb{R}$ , define the 0-centered, symmetric Pareto distribution by its PDF

$$\begin{aligned} \text{Pareto}(x; a, \lambda) &= \frac{a}{2\lambda} \left(1 + \left|\frac{x}{\lambda}\right|\right)^{-1-a} \\ &= \frac{a}{2\lambda} \exp\left[-(1+a) \log\left(1 + \left|\frac{x}{\lambda}\right|\right)\right]. \end{aligned}$$

Its CDF is given by

$$\text{ParetoCDF}(x; a, \lambda) = \begin{cases} 1 - \frac{1}{2} \left(1 + \left|\frac{x}{\lambda}\right|\right)^{-a} & \text{if } x > 0 \\ \frac{1}{2} \left(1 + \left|\frac{x}{\lambda}\right|\right)^{-a} & \text{else.} \end{cases}$$

Consider smoothing distributions of the form

$$q(x) = \prod_{i=1}^d \text{Pareto}(x_i; a, 1), \quad \text{so that}$$

$$\psi(x) = (1+a) \sum_{i=1}^d \log(1 + |x_i|)$$

$$\nabla \psi(x) = \left\{ (1+a) \frac{\text{sgn}(x_i)}{1 + |x_i|} \right\}_{i=1}^d,$$

with  $\nabla \psi(x)$  defined when all coordinates  $x_i$ s are nonzero.

### I.5.1. $\ell_1$ ADVERSARY

**Theorem I.17.** *Suppose  $H$  is a smoothed classifier smoothed by*

$$q(x) \propto \prod_{i=1}^d \text{Pareto}(x_i; a, \lambda),$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_1 < \lambda \frac{2\rho - 1}{a} {}_2F_1\left(1, \frac{a}{a+1}; \frac{a}{a+1} + 1; (2\rho - 1)^{1+\frac{1}{a}}\right),$$

where  ${}_2F_1$  is the hypergeometric function.

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

WLOG, assume  $u \in \text{Vert}(\mathcal{B})$  is  $e_1$ . Then  $\gamma_u = \langle u, \nabla\psi(\delta) \rangle = (1+a) \frac{\text{sgn}(\delta_1)}{1+|\delta_1|}$ ,  $\delta \sim q$ , is distributed as  $(1+a) \frac{\text{sgn}(z)}{1+|z|}$  where  $z \sim \text{Pareto}(a, 1)$ . Therefore, for  $p \in [0, 1/2]$ , the random variable  $\gamma_u^{(p)}$  defined in [Definition F.4](#) is

$$\gamma_u^{(p)} = \begin{cases} \frac{1+a}{1+z_p} & \text{with prob. } p \\ 0 & \text{with prob. } 1-p, \end{cases}$$

where  $z_p$  is sampled from  $\text{Pareto}(a, 1)$  conditioned on the interval  $[0, \text{ParetoCDF}^{-1}(p + 1/2; a, 1)]$ .

Thus, for any  $u \in \text{Vert}(\mathcal{B})$ ,

$$\Phi(p) = \mathbb{E} \gamma_u^{(p)} = \mathbb{E} \frac{1+a}{1+z} \mathbb{I}(z \in [0, c]),$$

where  $z \sim \text{Pareto}(a, 1)$  and  $c = \text{ParetoCDF}^{-1}(p + 1/2; a, 1)$ . This can be simplified as follows:

$$\begin{aligned} \Phi(p) &= \int_0^c \text{Pareto}(r; a, 1) \frac{1+a}{1+r} dr \\ &= - \int_0^c \text{Pareto}'(r; a, 1) dr \\ &= \text{Pareto}(0; a, 1) - \text{Pareto}(c; a, 1) \\ &= \frac{a}{2} \left(1 - (1+c)^{-1-a}\right). \end{aligned}$$

Note that for  $p \in [1/2, 1]$ ,

$$\text{ParetoCDF}^{-1}(p + 1/2; a, 1) = (1 - 2p)^{-1/a} - 1,$$

so  $\Phi(p)$  can be further simplified:

$$\Phi(p) = \frac{a}{2} \left(1 - (1 - 2p)^{\frac{a+1}{a}}\right).$$

Then, by setting  $G(x)$  in [Theorem F.6](#) to be  $1 - H(x)_y = 1 - \rho$ , we get the provably robust radius of

$$\begin{aligned} \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp &= \int_{1-\rho}^{1/2} \frac{2}{a} \left(1 - (1 - 2p)^{\frac{a+1}{a}}\right)^{-1} dp \\ &= \frac{2\rho - 1}{a} {}_2F_1\left(1, \frac{a}{a+1}; \frac{a}{a+1} + 1; (1 - 2\rho)^{1+\frac{1}{a}}\right) \Big|_{1-\rho}^{1/2} \\ &= \frac{2\rho - 1}{a} {}_2F_1\left(1, \frac{a}{a+1}; \frac{a}{a+1} + 1; (2\rho - 1)^{1+\frac{1}{a}}\right). \end{aligned}$$

□

## I.6. $\ell_2$ -Norm Based Exponential Law

In this section we consider

$$\begin{aligned} q(x) &\propto \exp(-\|x\|_2) \quad \text{so that} \\ \psi(x) &= \|x\|_2 \quad \text{and} \quad \nabla\psi(x) = x/\|x\|_2, \end{aligned}$$

defined as long as  $x \neq 0$ .

### I.6.1. $\ell_2$ ADVERSARY

**Theorem I.18.** *Suppose  $H$  is a smoothed classifier smoothed by*

$$q(x) \propto \exp(-\|x\|_2/\lambda),$$

such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with

$$\|\eta\|_2 < \lambda(d-1) \operatorname{arctanh}\left(1 - 2\text{BetaCDF}^{-1}\left(1 - \rho; \frac{d-1}{2}, \frac{d-1}{2}\right)\right).$$

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

We seek to apply [Theorem F.6](#) to  $G(x) = 1 - H(x)_y$ , for which we need to derive random variables  $\gamma_u$  and  $\gamma_u^{(p)}$ , and most importantly, the function  $\Phi$ .

For any  $u \in \text{Vert}(\mathcal{B})$  (i.e. any unit vector  $u$ ),  $\gamma_u = \langle u, \nabla\psi(\delta) \rangle = \langle u, \frac{\delta}{\|\delta\|_2} \rangle$ ,  $\delta \sim q$ , is distributed like  $2\text{Beta}(\frac{d-1}{2}, \frac{d-1}{2}) - 1$  by [Lemma I.23](#). Its complementary CDF is given by

$$\begin{aligned} \mathbb{P}[\gamma_u > c] &= R \int_c^1 (1-t^2)^{\frac{d-3}{2}} dt \\ &= \text{BetaCDF}\left(\frac{1-c}{2}; \frac{d-1}{2}, \frac{d-1}{2}\right) \stackrel{\text{def}}{=} \varphi(c), \end{aligned}$$

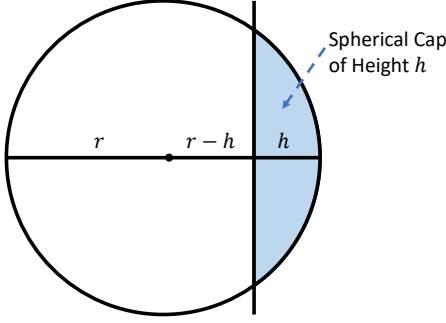


Figure I.1. Spherical Cap

where  $R \stackrel{\text{def}}{=} \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}$ . Therefore, for  $p \in [0, 1]$ , the random variable  $\gamma_u^{(p)}$  defined in Definition F.4 is given by

$$\gamma_u^{(p)} = \begin{cases} \gamma_u|_{(\varphi^{-1}(p), \infty)} & \text{with probability } p \\ 0 & \text{with probability } 1 - p, \end{cases}$$

Thus, for any  $u \in \text{Vert}(\mathcal{B})$ ,

$$\begin{aligned} \Phi(p) &= \mathbb{E} \gamma_u^{(p)} = R \int_{\varphi^{-1}(p)}^1 t(1-t^2)^{\frac{d-3}{2}} dt \\ &= \frac{R}{d-1} (1 - \varphi^{-1}(p)^2)^{\frac{d-1}{2}}. \end{aligned}$$

Then, by setting  $G(x)$  in Theorem F.6 to be  $1 - H(x)_y = 1 - \rho$ , we get the provably robust radius of

$$\begin{aligned} & \int_{1-\rho}^{1/2} \frac{1}{\Phi(p)} dp \\ &= \frac{d-1}{R} \int_{1-\rho}^{1/2} (1 - \varphi^{-1}(p)^2)^{-\frac{d-1}{2}} dp \\ &= (d-1) \int_0^{\varphi^{-1}(1-\rho)} (1-c^2)^{-\frac{d-1}{2}} (1-c^2)^{\frac{d-3}{2}} dc \\ &= (d-1) \int_0^{\varphi^{-1}(1-\rho)} (1-c^2)^{-1} dc \\ &= (d-1) \operatorname{arctanh}(\varphi^{-1}(1-\rho)). \end{aligned}$$

Unpacking the definition of  $\varphi$  yields the result.  $\square$

## I.7. Uniform Distribution over a Sphere

### I.7.1. $\ell_2$ ADVERSARY

Consider the distribution that is uniform on the  $\ell_2$  unit ball  $\{x : \|x\|_2 \leq 1\}$ . The *spherical cap* of height  $h \leq 1$  in this unit ball is the portion of the ball that is cut away by a hyperplane of distance  $1-h$  from the origin; see Fig. I.1. By Lemma I.24, this spherical cap has volume

$$V_d^h \stackrel{\text{def}}{=} V_d \operatorname{BetaCDF} \left( \frac{h}{2}; \frac{d+1}{2}, \frac{d+1}{2} \right).$$

where  $V_d$  is the volume of the unit sphere in  $\mathbb{R}^d$ .

Two unit radius spheres with centers  $\epsilon$  apart intersects in a region that is the union of two spherical caps of height  $1 - \epsilon/2$ . This intersection thus has volume  $2V_d^{1-\epsilon/2}$ , and the volume of one of the spheres outside this intersection is  $V_d - 2V_d^{1-\epsilon/2} = V_d \left( 1 - 2\operatorname{BetaCDF} \left( \frac{1-\epsilon/2}{2}; \frac{d+1}{2}, \frac{d+1}{2} \right) \right)$ .

**Theorem I.19.** *Suppose  $H$  is a smoothed classifier smoothed by the uniform distribution  $q$  over a ball of radius  $\lambda$  centered at the origin, such that  $H(x) = (H(x)_1, \dots, H(x)_C)$  is a vector of probabilities that  $H$  assigns to each class  $1, \dots, C$ . If  $H$  correctly predicts the class  $y$  on input  $x$ , and the probability of the correct class is  $\rho \stackrel{\text{def}}{=} H(x)_y > 1/2$ , then  $H$  continues to predict the correct class when  $x$  is perturbed by any  $\eta$  with*

$$\|\eta\|_2 < \lambda \left( 2 - 4\operatorname{BetaCDF}^{-1} \left( \frac{3}{4} - \frac{\rho}{2}; \frac{d+1}{2}, \frac{d+1}{2} \right) \right).$$

*Proof.* By linearity in  $\lambda$ , it suffices to show this for  $\lambda = 1$ .

By assumption, there is a region of probability  $\rho$  under the uniform distribution  $q(x + \cdot)$  centered at  $x$  that the base classifier classifies as  $y$ . The intersection between the support of  $q(x + \cdot)$  and  $q(x + \delta + \cdot)$  for any  $\|\delta\|_2 \leq \epsilon$  contains a region of probability at least

$$\rho - \left( 1 - 2\operatorname{BetaCDF} \left( \frac{1-\epsilon/2}{2}; \frac{d+1}{2}, \frac{d+1}{2} \right) \right)$$

that the base classifier classifies as  $y$ . For this probability to be at least  $1/2$ , we require

$$\begin{aligned} \frac{1}{2} &\leq \rho - (1 - 2\operatorname{BetaCDF} \left( \frac{1-\epsilon/2}{2}; \frac{d+1}{2}, \frac{d+1}{2} \right)) \\ \frac{3}{4} - \frac{\rho}{2} &\leq \operatorname{BetaCDF} \left( \frac{1-\epsilon/2}{2}; \frac{d+1}{2}, \frac{d+1}{2} \right) \\ 1 - \epsilon/2 &\geq 2\operatorname{BetaCDF}^{-1} \left( \frac{3}{4} - \frac{\rho}{2}; \frac{d+1}{2}, \frac{d+1}{2} \right) \\ \epsilon &\leq 2 - 4\operatorname{BetaCDF}^{-1} \left( \frac{3}{4} - \frac{\rho}{2}; \frac{d+1}{2}, \frac{d+1}{2} \right), \end{aligned}$$

as desired.  $\square$

## I.8. General $\ell_2$ -Norm Based Distributions via the Level Set Method

### I.8.1. $\ell_2$ ADVERSARY

Define  $W_d(r, s, \epsilon)$  to be the probability a point sampled from the *surface* of a ball of radius  $r$  centered at the origin is outside a ball of radius  $s$  with center  $\epsilon$  away from the origin. By Lemma I.23, we have

$$W_d(r, s, \epsilon) = \operatorname{BetaCDF} \left( \frac{(r+\epsilon)^2 - s^2}{4\epsilon r}; \frac{d-1}{2}, \frac{d-1}{2} \right). \quad (36)$$

Note that  $W_d$  can be evaluated quickly using standard `scipy` functions.

**Theorem I.20.** *Suppose that the density of a distribution satisfies  $q(x) = \bar{q}(\|x\|_2)$  for some differentiable, decreasing function  $\bar{q} : \mathbb{R}^{\geq 0} \rightarrow \mathbb{R}^{\geq 0}$ . Then for any  $\kappa > 0$  and any  $v \in \mathbb{R}^d$ , the growth function satisfies*

$$\mathcal{G}_q(p_0, v) = p_1$$

where

$$\begin{aligned} p_0 &= 1 - \mathbb{E}_r W_d(r, \bar{q}^{-1}(\bar{q}(r)/\kappa), \|v\|_2) \\ p_1 &= \mathbb{E}_r W_d(r, \bar{q}^{-1}(\bar{q}(r)\kappa), \|v\|_2), \\ r &\sim \text{distribution with density } \propto r^{d-1} \bar{q}(r). \end{aligned}$$

For most  $\bar{q}$ ,  $p_0$  and  $p_1$  can be evaluated numerically and quickly for each  $\kappa$  and  $\|v\|_2$  using 1-dimensional integrals.

*Proof.* Let  $r_t \stackrel{\text{def}}{=} \bar{q}^{-1}(t)$ . Then the superlevel set  $U_t = \{x : q(x) \geq t\}$  is a ball with radius  $r_t$ . Furthermore,

$$\begin{aligned} \nabla q(x) &= \bar{q}'(\|x\|_2) \frac{x}{\|x\|_2} \\ \|\nabla q(x)\|_2^{-1} &= -\bar{q}'(\|x\|_2)^{-1} = -r'_{\bar{q}(\|x\|_2)}. \end{aligned}$$

Let  $SA_d$  be the surface area of the unit sphere in  $\mathbb{R}^d$ . Then

$$\begin{aligned} q(\mathcal{NP}_\kappa) &= -SA_d \int_0^\infty r'_t t r_t^{d-1} (1 - W_d(r_t, r_t/\kappa, \|v\|_2)) dt \\ q(\mathcal{NP}_\kappa - v) &= -SA_d \int_0^\infty r'_t t r_t^{d-1} W_d(r_t, r_t\kappa, \|v\|_2) dt. \end{aligned}$$

If we change coordinates from  $t$  to  $r$ , then

$$\begin{aligned} q(\mathcal{NP}_\kappa) &= SA_d \times \\ &\int_0^\infty \bar{q}(r) r^{d-1} (1 - W_d(r, \bar{q}^{-1}(\bar{q}(r)/\kappa), \|v\|_2)) dr \\ q(\mathcal{NP}_\kappa - v) &= SA_d \times \\ &\int_0^\infty \bar{q}(r) r^{d-1} W_d(r, \bar{q}^{-1}(\bar{q}(r)\kappa), \|v\|_2) dr. \end{aligned}$$

Since  $\bar{q}(r)r^{d-1}$  is proportional to the density of  $\|x\|_2$ ,  $x \sim q$ , we can also write this as

$$\begin{aligned} q(\mathcal{NP}_\kappa) &= \mathbb{E}_{r=\|x\|_2, x \sim q} (1 - W_d(r, \bar{q}^{-1}(\bar{q}(r)/\kappa), \|v\|_2)) \\ q(\mathcal{NP}_\kappa - v) &= \mathbb{E}_{r=\|x\|_2, x \sim q} W_d(r, \bar{q}^{-1}(\bar{q}(r)\kappa), \|v\|_2). \end{aligned} \quad (37)$$

The distribution of  $r$  here has density  $\propto r^{d-1} \bar{q}(r)$ . Then setting  $p_0 = q(\mathcal{NP}_\kappa)$ ,  $p_1 = q(\mathcal{NP}_\kappa - v)$  yields the desired result by Eq. (NP).  $\square$

*Example I.21.* If  $q(x) \propto \|x\|_2^{-j} \exp(-\|x\|_2^k)$ , then  $\bar{q}(r) \propto r^{-j} \exp(-r^k)$ , and the radius is distributed as  $\sqrt[k]{\text{Gamma}(d/k - j/k)}$  by Lemma I.25. A table of robust radii can then be built according to Algorithm 1, and certification can be done via Algorithm 2.

*Example I.22.* If  $q(x) \propto (1 + \|x\|_2^k)^{-a}$ , then  $\bar{q}(r) \propto (1 + r^k)^{-a}$ , and the radius is distributed as  $\sqrt[k]{\text{BetaPrime}(d/k, a - d/k)}$ . A table of robust radii can then be built according to Algorithm 1, and certification can be done via Algorithm 2.

## I.9. Basic Facts about Probability Distributions

**Lemma I.23.** *If  $(x_1, \dots, x_d)$  is sampled uniformly from the unit sphere  $S^{d-1} \subseteq \mathbb{R}^d$ , then*

$$\frac{1 + x_1}{2} \text{ is distributed as } \text{Beta}\left(\frac{d-1}{2}, \frac{d-1}{2}\right),$$

i.e.

$$\begin{aligned} \mathbb{P}[x_1 \geq c] &= \text{BetaCDF}\left(\frac{1-c}{2}; \frac{d-1}{2}, \frac{d-1}{2}\right) \\ &= 1 - \text{BetaCDF}\left(\frac{1+c}{2}; \frac{d-1}{2}, \frac{d-1}{2}\right). \end{aligned}$$

*Proof.* From simple geometric reasoning, we get

$$\begin{aligned} \mathbb{P}[x_1 \geq c] &\propto \int_c^1 (1-t^2)^{\frac{d-3}{2}} dt \\ &= \int_c^1 (1-t)^{\frac{d-3}{2}} (1+t)^{\frac{d-3}{2}} dt \\ &= \int_0^{\frac{1-c}{2}} (2x)^{\frac{d-3}{2}} (2(1-x))^{\frac{d-3}{2}} dx. \end{aligned}$$

$\square$

**Lemma I.24.** *If  $(x_1, \dots, x_d)$  is sampled uniformly from the ball  $\{y : \|y\|_2 \leq 1\} \subseteq \mathbb{R}^d$ , then*

$$\frac{1 + x_1}{2} \text{ is distributed as } \text{Beta}\left(\frac{d+1}{2}, \frac{d+1}{2}\right),$$

i.e.

$$\begin{aligned} \mathbb{P}[x_1 \geq c] &= \text{BetaCDF}\left(\frac{1-c}{2}; \frac{d+1}{2}, \frac{d+1}{2}\right) \\ &= 1 - \text{BetaCDF}\left(\frac{1+c}{2}; \frac{d+1}{2}, \frac{d+1}{2}\right). \end{aligned}$$

*Proof.* Similar to Lemma I.23.  $\square$

**Lemma I.25.** *For any norm  $\|\cdot\|$  on  $\mathbb{R}^d$ , the distribution*

$$q(x) \propto \|x\|^{-j} \exp(-\|x\|^k),$$

with  $j < d$ , can be sampled as follows:

1. Sample the radius  $r \sim \sqrt[k]{\text{Gamma}(\frac{d}{k} - \frac{j}{k})}$
2. Sample a point  $v$  from the unit sphere of  $\|\cdot\|$
3. return  $rv$

**Lemma I.26.** For any  $c, s \geq 0$  and  $r > 0$ ,

$$\begin{aligned} & \mathbb{E}_{z \sim \text{Gamma}(r)} z^s \mathbb{I}(z > c) \\ &= \frac{\Gamma(r+s)}{\Gamma(r)} (1 - \text{GammaCDF}(c; r+s)). \end{aligned}$$

## J. Proof of Theorem 7.3

In this section we prove our main impossibility result, [Theorem 7.3](#). We will assume throughout this proof that the reader is familiar with standard notions in functional analysis. Our proof will proceed in two steps.

For all  $p \in (0, 2]$  and  $d' \geq 1$ , we let  $\ell_p^{d'}$  denote  $\mathbb{R}^{d'}$  equipped with the  $p$ -quasinorm. First, we show that if there exists a useful smoothing scheme, this implies a low embedding distortion of our normed space into  $\ell_{0.99}^{d'}$ , for some  $d'$ .

Formally, let  $(X, d_X)$  and  $(Y, d_Y)$  be two metric spaces. We say an embedding  $f : X \rightarrow Y$  has distortion  $D$  if there exist positive constants  $\alpha < 1 < \beta$  so that

$$\alpha d_Y(f(x_1), f(x_2)) \leq d_X(x_1, x_2) \leq \beta d_Y(f(x_1), f(x_2))$$

for all  $x_1, x_2 \in X$ , where  $\beta/\alpha \leq D$ . We will first show:

**Lemma J.1.** Suppose there exists an  $(\varepsilon, s, \ell)$ -useful smoothing scheme for  $\|\cdot\|$ , and  $s/\ell \leq 1/162$ . Then, there exists  $d'$  and a linear embedding from  $(\mathbb{R}^d, \|\cdot\|)$  into  $\ell_{0.99}^{d'}$  with distortion at most

$$O\left(\left(\frac{s}{\ell}\right)^{1/4} \cdot \frac{1}{\varepsilon}\right).$$

Next we show that any linear embedding into  $\ell_{0.99}^{d'}$  will suffer distortion which is at least  $C_2((\mathbb{R}^d, \|\cdot\|))$ :

**Lemma J.2.** Any linear embedding from  $(\mathbb{R}^d, \|\cdot\|)$  into  $\ell_{0.99}^{d'}$  must have distortion  $\Omega(C_2((\mathbb{R}^d, \|\cdot\|)))$ , where  $\Omega$  hides constant independent of  $d$  and  $d'$ .

This result is essentially folklore in the metric embedding community, but we include a proof for completeness.

These two lemmas together immediately imply [Theorem 7.3](#). The rest of this section is dedicated to proofs of these two lemmas. To do so, it will first be useful to establish some regularity conditions on a variant of the growth function considered previously in this paper.

### J.1. The Pairwise Growth Function

For any two probability densities  $q_1, q_2$ , over  $\mathbb{R}^d$ , define the *pairwise growth function* between  $q_1$  and  $q_2$ , denoted

$\mathcal{G}_{q_1, q_2}$ , to be

$$\mathcal{G}_{q_1, q_2}(p) = \sup_{U: q_1(U)=p} q_2(U).$$

We will assume throughout this proof that  $q_2$  is absolutely continuous with respect to  $q_1$ . The more general case can be easily handled by the theory of Radon-Nikodym derivatives and Lebesgue's decomposition theorem. This growth function satisfies the following, basic properties, whose proofs are easy and are omitted.

**Fact J.3.** Let  $q_1, q_2, \mathcal{G}_{q_1, q_2}$  be above. Then:

- $\mathcal{G}_{q_1, q_2}(p)$  is monotonically increasing.
- $\mathcal{G}_{q_1, q_2}(1) = 1$ , and  $\mathcal{G}_{q_1, q_2}(0) \geq 0$ .
- $d_{\text{TV}}(q_1, q_2) = \sup_{p \in [0, 1]} (\mathcal{G}_{q_1, q_2}(p) - p)$ .

Then, we have:

**Lemma J.4.** For any  $q_1, q_2 \in \Delta_d$ , the function  $\mathcal{G}_{q_1, q_2}$  is concave.

*Proof.* For clarity, since  $q_1, q_2$  will be fixed throughout this proof, we will omit the subscripts in the definition of  $\mathcal{G}$ .

For any  $t > 0$ , define the set

$$S_t = \left\{ x \in \mathbb{R}^d : \frac{dq_2}{dq_1}(x) \geq t \right\},$$

which we can think of as a generalized Neyman-Pearson set, for the two distributions  $q_1, q_2$ .

Then, by classical arguments, for every  $p$ , the set which obtains the supremum in the definition of the growth function for that value of  $p$  is given by

$$S_{K(p)} = \operatorname{argmax}_{U: q_1(U)=p} q_2(U),$$

where  $K(p)$  is defined so that  $q_1(S_{K(p)}) = p$ . Therefore, for all  $p$ , we have that  $\mathcal{G}(p) = q_2(S_{K(p)})$ .

We will show that for all  $p < p' < p''$ , the growth function satisfies

$$\frac{\mathcal{G}(p') - \mathcal{G}(p)}{p' - p} \geq \frac{\mathcal{G}(p'') - \mathcal{G}(p')}{p'' - p'}, \quad (38)$$

which is equivalent to the claim.

Note that for any  $0 \leq r \leq r'$ , we have that  $\mathcal{G}(r') - \mathcal{G}(r) = q_2(\Delta_{r', r})$ , where  $\Delta_{r', r} = S_{K(r')} \setminus S_{K(r)}$ . However, observe that for  $p < p' < p''$ , we have that  $\frac{dq_2}{dq_1}(x) \geq \frac{dq_2}{dq_1}(x')$  for all  $x \in \Delta_{p', p}$  and  $x' \in \Delta_{p'', p'}$ . But we also have

$$\begin{aligned} \mathcal{G}(p') - \mathcal{G}(p) &\geq q_1(\Delta_{p', p}) \cdot \min_{x \in \Delta_{p', p}} \frac{dq_2}{dq_1}(x) \\ &= (p' - p) \min_{x \in \Delta_{p', p}} \frac{dq_2}{dq_1}(x), \end{aligned}$$

and similarly

$$\mathcal{G}(p'') - \mathcal{G}(p') \leq (p'' - p') \max_{x \in \Delta_{p'', p'}} \frac{dq_2}{dq_1}(x),$$

which implies Eq. (38).  $\square$

## J.2. Proof of Lemma J.1

Let  $\mathcal{Q} = \{q_x\}_{x \in \mathbb{R}^d}$  be an  $(\varepsilon, r, \ell)$ -useful smoothing scheme for  $\|\cdot\|$ . For simplicity, throughout this proof, we will assume that  $q_x$  has a probability density function, denoted  $Q_x$ , for all  $x$ , that is, the distributions are absolutely continuous with respect to the Lebesgue measure. It is not hard to generalize this proof to handle general probability distributions by using Lebesgue decomposition and taking the appropriate Radon-Nikodym derivatives.

First, we demonstrate that a useful smoothing scheme actually implies an embedding of the norm  $\|\cdot\|$  into an infinite dimensional  $L_1$  space, namely, the space of all distributions with distance given by total variation distance. Recall the total variation distance between two distributions  $q_1, q_2$ , denoted  $d_{\text{TV}}(q_1, q_2)$ , is given by

$$d_{\text{TV}}(q_1, q_2) = \sup_{U \subseteq \mathbb{R}^d} |q_1(U) - q_2(U)| = \frac{1}{2} \|Q_1 - Q_2\|_1.$$

We denote the space of probability distributions over  $\mathbb{R}^d$  by  $\Delta_d$ . Note that the metric space  $(\Delta_d, d_{\text{TV}})$  is an infinite dimensional  $L_1$  space. Thus, classical results yield:

**Fact J.5** (see e.g. Wojtaszczyk (1996)). *For all  $d \geq 1$ , we have  $C_2((\Delta_d, d_{\text{TV}})) = \Theta(1)$ .*

We now need another notion, introduced in Andoni et al. (2018).

**Definition J.6** (Andoni et al. (2018)). A map  $f : X \rightarrow Y$  between two metric spaces  $(X, d_X)$  and  $(Y, d_Y)$  is an  $(s_1, s_2, \tau_1, \tau_2)$ -threshold map if it satisfies:

- If  $d_X(x_1, x_2) \leq s_1$ , then  $d_Y(f(x_1), f(x_2)) \leq \tau_1$ .
- If  $d_X(x_1, x_2) \geq s_2$ , then  $d_Y(f(x_1), f(x_2)) \geq \tau_2$ .

Any smoothing scheme  $\mathcal{Q} = \{q_x\}_{x \in \mathbb{R}^d}$  can be viewed as a map  $\mathbb{R}^d \rightarrow \Delta_d, x \mapsto q_x$  that takes a point in  $\mathbb{R}^d$  and maps it to its associated distribution after smoothing. Our main technical work will be to demonstrate the following lemma:

**Lemma J.7.** *Let  $q$  be a  $(\varepsilon, s, \ell)$ -useful smoothing distribution for  $\|\cdot\|$ . Then  $\mathcal{Q}$  is a  $(\varepsilon, 1, 2s, \ell)$ -threshold map between  $(\mathbb{R}^d, \|\cdot\|)$  and  $(\Delta, d_{\text{TV}})$ .*

Our first observation is that Lemma J.4 allows us to relate the usefulness of the smoothing scheme to total variation distance:

**Corollary J.8.** *For any  $q_1, q_2$ , we have that*

$$\mathcal{G}_{q_1, q_2}(1/2) - 1/2 \geq \frac{1}{2} d_{\text{TV}}(q_1, q_2).$$

*Proof.* As before, for conciseness we will drop the subscripts in the definition of  $\mathcal{G}$ . We will show that for all  $p \in [0, 1]$ , we have that

$$\mathcal{G}(1/2) - 1/2 \geq \frac{1}{2} (\mathcal{G}(p) - p),$$

which by Fact J.3 implies the lemma.

First, consider the case where  $p \geq 1/2$ . Then, by concavity of  $\mathcal{G}(p)$ , we have that

$$\begin{aligned} \mathcal{G}(1/2) &\geq \frac{1}{2p} \mathcal{G}(p) + \frac{2p-1}{2p} \mathcal{G}(0) \\ &\geq \frac{1}{2p} \mathcal{G}(p), \end{aligned}$$

since  $\mathcal{G}(0) \geq 0$  by Fact J.3. Therefore, we have that

$$\mathcal{G}(1/2) - \frac{1}{2} \geq \frac{\mathcal{G}(p) - p}{2p} \geq \frac{1}{2} (\mathcal{G}(p) - p).$$

The case where  $p < 1/2$  follows symmetrically by considering the line segment between  $p$  and 1.  $\square$

From this, the proof of Lemma J.7 is simple.

*Proof of Lemma J.7.* We first prove that it satisfies the first condition. Let  $x, y$  be so that  $\|x - y\| \leq \varepsilon$ . Then, the robustness condition implies that

$$\mathcal{G}_{q_2, q_1}(1/2) - 1/2 \leq r.$$

By Corollary J.8 this implies that  $d_{\text{TV}}(q_x, q_y) \leq 2r$ .

We now prove it satisfies the second condition. But, the accuracy condition immediately implies that if  $x, y$  satisfy  $\|x - y\| \geq 1$ , we must have  $d_{\text{TV}}(q_x, q_y) \geq \ell$ . This proves the claim.  $\square$

With Lemma J.7 in hand, we can now invoke a number of classical results from the theory of metric embeddings to obtain our desired result. We first use the following fact, which follows since  $L_1$  embeds isometrically into squared- $L_2$ .

**Fact J.9** (see e.g. Matoušek (2013)). *Suppose there exists an  $(s_1, s_2, \tau_1, \tau_2)$ -threshold map from  $(\mathbb{R}^d, \|\cdot\|)$  to  $(\Delta_d, d_{\text{TV}})$ . Then there exists an  $(s_1, s_2, \sqrt{\tau_1}, \sqrt{\tau_2})$ -threshold map from  $(\mathbb{R}^d, \|\cdot\|)$  to a Hilbert space  $H$ .*

This implies:

**Corollary J.10.** *Suppose there exists an  $(\varepsilon, s, \ell)$ -useful smoothing distribution for  $\|\cdot\|$ . Then there exists a  $(\varepsilon, 1, \sqrt{2s}, \sqrt{\ell})$ -threshold map between  $(\mathbb{R}^d, \|\cdot\|)$  and a Hilbert space  $H$ .*

We now require the following theorem, first proven in Andoni et al. (2018), which we reproduce below in a slightly simplified form:

**Theorem J.11** (Theorem 4.12 in Andoni et al. (2018)). Suppose there exists a  $(\varepsilon, 1, \tau_1, \tau_2)$ -threshold map from  $(\mathbb{R}^d, \|\cdot\|)$  to a Hilbert space, for  $\tau_2 \geq 9\tau_1$ . Then there exists a map  $h$  from  $\mathbb{R}^d$  into a Hilbert space with induced norm  $\|\cdot\|_H$  such that for every  $x_1, x_2 \in \mathbb{R}^d$ , we have:

$$\begin{aligned} \sqrt{\tau_2} \cdot \min(1, \varepsilon \|x_1 - x_2\|) &\leq \|h(x_1) - h(x_2)\|_H \\ &\leq 10 \cdot \sqrt{2\tau_1 \|x_1 - x_2\|}. \end{aligned}$$

Combining Corollary J.10 and Theorem J.11, we obtain:

**Corollary J.12.** Suppose there exists an  $(\varepsilon, s, \ell)$ -useful smoothing distribution for  $\|\cdot\|$ , and suppose that  $s/\ell \leq 1/162$ . Then, there exists a Hilbert space  $H$  with induced norm  $\|\cdot\|_H$  and a map  $h : \mathbb{R}^d \rightarrow H$  so that for all  $x, y \in \mathbb{R}^d$ , we have

$$\begin{aligned} \min(1, \varepsilon \|x_1 - x_2\|) &\leq \|h(x_1) - h(x_2)\|_H \\ &\leq 10 \cdot 2^{3/4} \cdot \left(\frac{s}{\ell}\right)^{1/4} \sqrt{\|x_1 - x_2\|}. \end{aligned}$$

Finally, we require the following theorem from Andoni et al. (2018), which we reproduce for completeness:

**Theorem J.13** (Theorem 5.1 in Andoni et al. (2018)). Let  $X$  be a finite-dimensional normed space with norm  $\|\cdot\|$ , and let  $\Delta > 0$ . Let  $H$  be a Hilbert space with associated norm  $\|\cdot\|_H$ . Assume we have a map  $f : X \rightarrow H$ , such that, for some absolute constant  $K > 0$ , and for all  $x, y \in X$ , we have:

- $\|f(x_1) - f(x_2)\|_H \leq K \cdot \sqrt{\|x_1 - x_2\|}$ , and
- if  $\|x - y\| \geq \Delta$ , then  $\|f(x_1) - f(x_2)\|_H \geq 1$ .

Then, for any  $\xi \in (0, 1/3)$ , the space  $X$  linearly embeds into  $\ell_{1-\xi}^{d'}$  with distortion  $O(\Delta/\xi)$ , for some finite  $d'$ .

Combining Corollary J.12 and Theorem J.13 immediately yields Lemma J.1.

### J.3. Proof of Lemma J.2

We now turn to the proof of Lemma J.2. The only reason why this is slightly non-standard is that  $\ell_p^{d'}$  for  $p < 1$  are not norms, as they do not satisfy the triangle inequality. Despite this, we show that the standard results that the cotype constant is a lower bound on distortion of any linear embedding still holds in these spaces.

**Fact J.14** (Khinchine's Inequality). For any  $p \in (0, \infty)$  there exist constants  $A_p, B_p$  such that for any  $x_1, \dots, x_n \in \mathbb{R}$

$$A_p \sqrt{\sum_{i=1}^n x_i^2} \leq \left( \mathbb{E} \left| \sum_{i=1}^n \sigma_i x_i \right|^p \right)^{1/p} \leq B_p \sqrt{\sum_{i=1}^n x_i^2}.$$

Here  $\sigma_1, \dots, \sigma_n$  are independent Rademacher random variables. In particular, for  $0 < p \leq p_0 \approx 1.8$ ,  $A_p = 2^{1/2-1/p}$ .

This implies:

**Lemma J.15** (Cotype Estimate). For any  $p \in (0, 1]$ ,  $\ell_p^{d'}$  has cotype 2 with cotype constant  $1/A_p$  with  $A_p$  as in Fact J.14, i.e. for any  $x_1, \dots, x_n \in \mathbb{R}^{d'}$ ,

$$A_p \sqrt{\sum_{i=1}^n \|x_i\|_p^2} \leq \mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i \right\|_p,$$

where  $\sigma_1, \dots, \sigma_n$  are independent Rademacher random variables.

*Proof.* Let  $x_{ij}$  denote coordinate  $j$  of  $x_i$ , i.e.,  $X$  is the  $n \times d$  matrix whose rows are  $x_1, \dots, x_n$ . By Khinchine's inequality,

$$\begin{aligned} \mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i \right\|_p^p &= \sum_{j=1}^d \mathbb{E} \left| \sum_{i=1}^n \sigma_i x_{ij} \right|^p \\ &\geq A_p^p \sum_{j=1}^d \left( \sum_{i=1}^n x_{ij}^2 \right)^{p/2}. \end{aligned}$$

Let us now consider the case  $p \leq 2$ . By the triangle inequality for  $\|\cdot\|_q$ , where  $q = 2/p \geq 1$ , applied to the vectors  $(|x_{1j}|^p, |x_{2j}|^p, \dots, |x_{nj}|^p)$ ,  $j \in [d]$ ,

$$\begin{aligned} \sum_{j=1}^d \left( \sum_{i=1}^n x_{ij}^2 \right)^{p/2} &= \sum_{j=1}^d \left( \sum_{i=1}^n (|x_{ij}|^p)^{2/p} \right)^{p/2} \\ &\geq \left( \sum_{i=1}^n \left| \sum_{j=1}^d |x_{ij}|^p \right|^{2/p} \right)^{p/2} \\ &= \left( \sum_{i=1}^n \|x_i\|_p^2 \right)^{p/2}. \end{aligned}$$

Finally, by the concavity of the function  $x \mapsto x^p$  for  $p \in (0, 1]$ , we have

$$\left( \mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i \right\|_p \right)^p \geq \mathbb{E} \left\| \sum_{i=1}^n \sigma_i x_i \right\|_p^p.$$

□

We now have all the tools we need to prove Lemma J.2:

*Proof of Lemma J.2.* For brevity, let  $p = 0.99$  in this proof. Let  $T : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$  be any linear map satisfying

$$\alpha \|Tx\|_p \leq \|x\| \leq \beta \|Tx\|_p.$$



Let  $C_2 = C_2((\mathbb{R}^d, \|\cdot\|))$ , and let  $x_1, \dots, x_n$  be a sequence in  $(\mathbb{R}^d, \|\cdot\|)$  satisfying

$$\mathbb{E} \left[ \left\| \sum_{j=1}^n \sigma_j x_j \right\|^2 \right] = C_2 \sqrt{\sum_{j=1}^n \|x_j\|^2}.$$

However, we have that

$$\alpha \mathbb{E} \left[ \left\| \sum_{j=1}^n \sigma_j T x_j \right\|_p \right] \leq \mathbb{E} \left[ \left\| \sum_{j=1}^n \sigma_j x_j \right\|^2 \right],$$

and simultaneously, we have

$$\sqrt{\sum_{j=1}^n \|x_j\|^2} \leq \beta \sqrt{\sum_{j=1}^n \|T x_j\|_p^2}.$$

Combining these facts and [Lemma J.15](#), we obtain that  $\beta/\alpha \geq A_p C_2 = \Omega(C_2)$ , as claimed, where  $A_p$  is as in [Lemma J.15](#) and [Fact J.14](#).  $\square$