Completing the Picture: Randomized Smoothing Suffers from the Curse of Dimensionality for a Large Family of Distributions

Yihan Wu **Aleksandar Bojchevski** **Aleksei Kuvshinov** **Stephan Günnemann**
Technical University of Munich, Germany

Abstract

Randomized smoothing is currently the most competitive technique for providing provable robustness guarantees. Since this approach is model-agnostic and inherently scalable we can certify arbitrary classifiers. Despite its success, recent works show that for a small class of i.i.d. distributions, the largest l_p radius that can be certified using randomized smoothing decreases as $O(1/d^{1/2-1/p})$ with dimension d for $p > 2$. We complete the picture and show that similar no-go results hold for the l_2 norm for a much more general family of distributions which are continuous and symmetric about the origin. Specifically, we calculate two different upper bounds of the l_2 certified radius which have a constant multiplier of order $\Theta(1/d^{1/2})$. Moreover, we extend our results to l_p $(p > 2)$ certification with spherical symmetric distributions solidifying the limitations of randomized smoothing. We discuss the implications of our results for how accuracy and robustness are related, and why robust training with noise augmentation can alleviate some of the limitations in practice. We also show that on real-world data the gap between the certified radius and our upper bounds is small.

1 INTRODUCTION

Most classifiers are vulnerable to adversarial examples (Akhtar and Mian, 2018; Hao-Chen et al., 2020). Slight perturbations of the data are often sufficient to manipulate their predictions. This lack of robustness is problematic, even in scenarios without adversaries, because real-world data can be noisy or anomalous. Since heuristic defenses can be easily broken (Carlini and Wagner, 2017; Tramer et al., 2020), a more promising direction is towards deriving adversarial robustness certificates which provide provable guarantees and are by definition unbreakable.

Probabilistic approaches based on randomized smoothing (Cohen et al., 2019; Li et al., 2019), inspired by connections to differential privacy (Lecuyer et al., 2019), tend to outperform deterministic approaches based on e.g. linear relaxations (Zhang et al., 2020b), MILPs (Tjeng et al., 2019), or Lipschitz constant estimation (Zhang et al., 2019). The biggest advantage of randomized smoothing techniques is that we can use them to certify arbitrary classifiers. These approaches are model-agnostic and scalable since they boil down to randomly perturbing the input and recording the class corresponding to the "majority vote" on the randomized samples. Given any base classifier $f(\cdot)$ we can build a "smoothed" classifier $g(\cdot)$ that has comparable accuracy to f but is amenable to (probabilistic) robustness guarantees.

Despite its success, randomized smoothing has several limitations. Since it does not make any assumptions about f it does not explicitly exploit any of f's properties such as smoothness. Moreover, we need a large number of samples (e.g. $\geq 10^5$) to provide any meaningful guarantees. An even more fundamental limitation is that this approach suffers from the curse of dimensionality: Kumar et al. (2020) have recently shown that for a small class of i.i.d. distributions, the largest l_p radius that can be certified using randomized smoothing decreases as $O(1/d^{1/2-1/p})$ with dimension d for $p > 2$. Here the l_p certified radius r with respect to a classifier g and a sample x is defined as the largest radius r such that for all perturbations δ with $||\delta||_p \leq r$ the prediction stays the same, i.e. $g(x + \delta) = g(x)$.

In this paper we show that similar no-go results hold more broadly, completing the picture in two different ways. First, we show that for the l_2 norm – which is commonly studied w.r.t. adversarial robustness – the curse of dimensional applies for a much more general

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

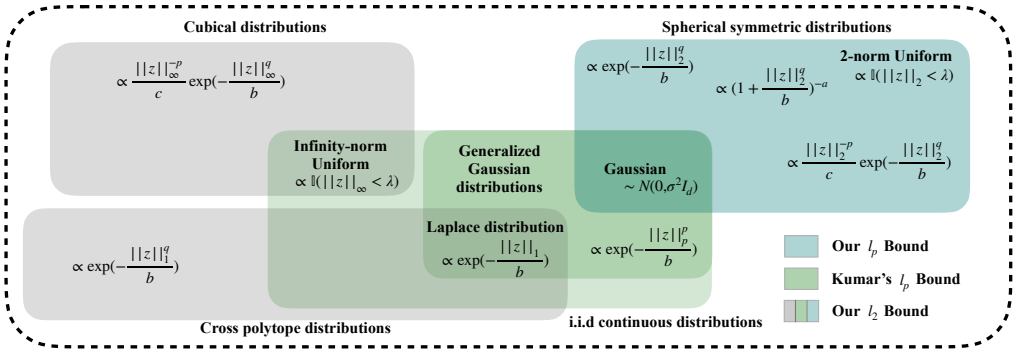


Figure 1: Popular families of smoothing distributions, each box refers to a certain family of distributions. All of these distributions are continuous and symmetric about the origin and thus covered by our (l_2 norm) bound.

family of smoothing distributions. Second, we extend our results to l_p norm ($p > 2$) certification with spherical symmetric distributions. On Fig. 1 we show a Venn diagram of the different families of distributions covered by our upper bounds, and the relations to previous work. Our main contributions are:

- We derive two upper bounds on the certified l_2 radius for randomized smoothing with any continuous and origin-symmetric distribution that have a constant multiplier of order $\Theta(1/d^{1/2})$.
- We derive an upper bound with a multiplier of order $\Theta(1/d^{1-1/p})$ on the l_p ($p > 2$) certified radius for smoothing with spherical symmetric distributions.

We emphasize that "origin-symmetric" and "continuous" are two common properties of all smoothing distributions that have been studied so far.[1]

We also discuss the implications of our theoretical results for how accuracy and robustness are related, and why robust training with noise augmentation (Salman et al., 2019) can alleviate some of the limitations in practice. On CIFAR-10 we show that the gap between the certified radius and our upper bounds is small.

Result Summary. We show that for an arbitrary continuous and origin-symmetric distribution q, i.e. $\forall z,\ q(z) = q(-z)$, the certified l_2 radius $r_g(x)$ of an arbitrary sample x is bounded by

$$r_g(x) < \frac{5}{\sqrt{d}}\Psi^{-1}\left(\frac{g(x)}{1-5*10^{-7}}; q\right),$$
$$\text{or} \quad r_g(x) < R_x\frac{5}{\sqrt{d}}$$

where $g(x) = \max_{c\in C}\mathbb{P}_{z\sim q}(f(x+z)=c)$ is the probability of the majority class when using the distribution q and the base classifier f, R_x is a dominating radius which we define as the minimum radius r s.t. $\mathbb{P}_{z\sim q}(f(x+z) = c_x \mid ||z||_2 > r) < 0.5$, and

$$\Psi(r; q) := \int_{||z||_2<r} q(z)dz.$$

Furthermore, for $q \sim \mathcal{N}(0, \sigma^2 I_d)$ we can compute in closed-form $\Psi^{-1}(x; q) = \sigma\sqrt{2\text{Gamma}^{-1}(x; \frac{d}{2}, 1)}$.

2 RELATED WORK

There are mainly three general approaches to calculate the certified radius for the smoothed classifier. The divergence-based method (Li et al., 2019; Dvijotham et al., 2020) provides a loose lower bound on the certified radius, while the method based on the Neyman-Pearson lemma (Cohen et al., 2019) and the method based on functional optimization (Zhang et al., 2020a) both certify a tight lower bound of the radius. For the latter two methods, the main idea is to relax the problem by optimizing over the set of all classifiers whose expectation under the smoothing distribution matches the expectation of the base classifier. Solving for the worst-case classifier yields a tight lower bound. For example, for Gaussian smoothing the worst-case classifier is linear with a decision boundary orthogonal to the adversarial perturbation.

In this paper we study the problem of finding the best possible radius which can be achieved using randomized smoothing, i.e. we derive an upper bound. Hayes (2020) and Zheng et al. (2020) upper bound the radius resulting from the divergence-based method. Zheng et al. (2020) study a non-standard notion of certified robustness and provide only indirect evidence for the hardness of the problem. Hayes (2020) prove that the certified radius $r < (-\sigma^p \log(2\sqrt{g(x)(1-g(x))}))^{\frac{1}{p}}$ for $p = 1, 2$ using generalized Gaussian distributions, given the prediction score $g(x)$ of a sample x. However,

[1] Using asymmetric smoothing distributions is sub-optimal since the adversary might exploit the asymmetry.

since the divergence-based radius is not tight, their upper bounds are actually below the certified radius, i.e. below the tight lower bound (see Fig. 4).

Kumar et al. (2020) and Yang et al. (2020a) upper bound the (tight) certified radius resulting from the worst-case classifier optimization problem. We also follow this approach. Kumar et al. (2020) show an upper bound of $r < \frac{\sigma}{\sqrt{2}d^{\frac{1}{2}-\frac{1}{p}}}\sqrt{\frac{1}{1-p_1}}, p \geq 2$ with i.i.d distributions and $r < \frac{4\sigma}{d^{\frac{1}{2}-\frac{1}{p}}}(\sqrt{\log\frac{1}{1-p_1}}), p \geq 2$ for generalized Gaussian distributions. To apply Yang et al. (2020a)'s upper bound, which is on the order of $O(d^{1/p-1/2})$, the smoothing distribution has to satisfy certain restrictive conditions which do not always hold, i.e. even simple Gaussian distributions with a large σ are excluded. Zhang (2020) and Blum et al. (2020) show that any smoothing distribution for l_p ($p > 2$) must have a large component-wise magnitude, which similarly shows the limitations of smoothing for certification.

Since most certificates focus on the common norms such as l_0, l_1, l_2, and l_∞, studying the best achievable performance of randomized smoothing on these norms is important. Moreover, the smoothing distributions in Kumar et al. (2020) and Hayes (2020) are limited to a small family, but we expect a more universal bound which can cover most smoothing distributions. We focus on defending against l_2 attack with a general family of smoothing distributions. Our upper bounds are tighter and more general. We further extend our results to l_p ($p > 2$) certification with spherical symmetric distributions which have not been covered before.

3 PRELIMINARIES

A classification task consists of a sample space $X = \mathbb{R}^d$, a finite target space $C = \{c_1, c_2, ..., c_k\}$, and a classifier $f : X \to C$. In randomized smoothing we choose a distribution q with probability density function $q(z)$. For an arbitrary data point $x \in \mathbb{R}^d$, the prediction c_x of the smoothed classifier on x is

$$c_x = \arg\max_{c \in C} g(x, c) = \arg\max_{c \in C} \mathbb{P}_{z\sim q}(f(x + z) = c).$$

A smoothed classifier is l_p robust on sample x within radius r if and only if

$$\arg\max_{c \in C} g(x + \delta, c) = c_x, \quad \forall ||\delta||_p \leq r.$$

To reduce the complexity of the problem, we consider $g(x + \delta) := g(x + \delta, c_x) = \mathbb{P}_{z\sim q}(f(x + z + \delta) = c_x)$. In this setting, if $g(x + \delta) > \frac{1}{2}$, $(x + \delta)$ is correctly classified with the perturbation δ. This is tight for binary classification and sound for more than two classes.

Evaluating $g(x + \delta)$ for all δ is computationally intractable. So we find a worst-case classifier $f^* \in \mathcal{F}$, where $\mathcal{F}$ is the set of all functions bounded in $[0, 1]$, by solving the following optimization problem

$$\begin{aligned} f^* = \arg\min_{f' \in \mathcal{F}} & \mathbb{P}_{z\sim q}(f'(x + z + \delta) = c_x) \\ s.t. \quad & \mathbb{P}_{z\sim q}(f'(x + z) = c_x) = g(x). \end{aligned} \tag{1}$$

The minimum $\underline{g(x + \delta)} = \mathbb{P}_{z\sim q}(f^*(x + z + \delta) = c_x)$ is a lower bound of $g(x + \delta)$. By solving

$$r_g(x) = \max_{r>0} r \quad s.t. \quad \underline{g(x + \delta)} > \frac{1}{2}, \ \forall ||\delta||_p \leq r, \tag{2}$$

we obtain the l_p certified radius of sample point x.

Selecting any classifier $h \in \mathcal{F}$ which satisfies the constraint $\mathbb{P}_{z\sim q}(h(x+z) = c_x) = g(x)$ and a perturbation δ, if $\mathbb{P}_{z\sim q}(h(x + z + \delta) = c_x) < \frac{1}{2}$, the l_p certified radius $r_g(x)$ computed in Eq. 2 has to be less than $||\delta||_p$ which yields a valid upper bound (Kumar et al., 2020).

The idea is to select h and δ such that the upper bound can be easily evaluated for a large family of distributions. Next, we introduce the high-dimensional spherical sector and construct h and δ based on the special properties of hyperspherical sectors.

Missing proofs are delegated to Sec. 1 in the appendix.

4 BOUNDING THE l_2 RADIUS

Definition 1. *A hyperspherical sector is a part of an l_2 hypersphere defined by a conical boundary with apex at the center of the sphere.*

We denote the colatitude angle ϕ as the angle between the rim of the hyperspherical cap and the direction to the middle of the cap when viewed from the center of the hypersphere (Fig. 2 left).

Definition 2. *Define the probability mass of distribution q inside an l_2 ball with radius r by*

$$\Psi(r; q) := \int_{||z||_2 < r} q(z)dz, \tag{3}$$

the inverse function of $\Psi(r; q)$ is denoted by $\Psi^{-1}(\cdot; q)$.

Proposition 1 (Li (2011))**.** *The volume of a hyperspherical sector with colatitude angle ϕ in a d-dim l_2 ball with radius r is*

$$V_s(r) = \frac{1}{2} I_{sin^2(\phi)}(\frac{d-1}{2}, \frac{1}{2}) V_b(r), \tag{4}$$

where $V_b(r)$ is the volume of a l_2 ball with radius r and $I_x(a, b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}}{B(a,b)}$ is a regularized incomplete beta function.

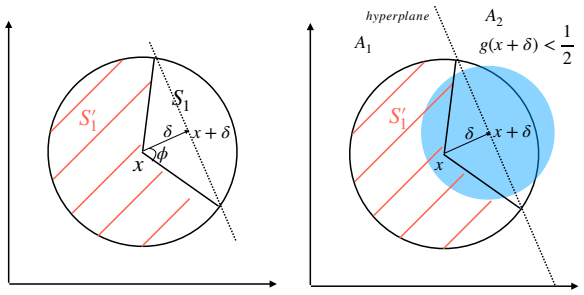


Figure 2: Illustration of the chosen perturbation δ, and the chosen classifier $h(x) = c_x 1_{(x \in S_1')}$.

Proposition 2. *When $R > \Psi^{-1}(g(x); q)$, there exists a perturbation δ and a hyperspherical sector S_1, whose symmetric axis is along the perturbation δ, in $B(R, x) := \{z \mid ||z - x||_2 < R\}$ (Fig. 2 left) and a classifier $h(x) = c_x 1_{(x \in S_1')}$ such that*

$$p(\{z|h(x+z) = c_x\}) = q(S_1') = g(x), \tag{5}$$

$$\frac{q(S_1)}{q(B(R,x))} \leq \frac{V_{S_1}(R)}{V_B(R)}. \tag{6}$$

For brevity we set $q(S) := \mathbb{P}_{z\sim q}(z + x \in S)$ for $S \subset X$.

Note that, for the construction of h we assume w.l.o.g. that $c_x = 1$. From Eq. 6 we have

$$\begin{aligned} 1 - \frac{g(x)}{\Psi(R;q)} &= \frac{q(S_1)}{q(B(R,x))} \leq \frac{V_{S_1}(R)}{V_B(R)} \\ &= \frac{1}{2} I_{sin^2(\phi)}(\frac{d-1}{2}, \frac{1}{2}) = \frac{1}{2} I_{1-\frac{||\delta||_2^2}{R^2}}(\frac{d-1}{2}, \frac{1}{2}). \end{aligned} \tag{7}$$

Let $\mathcal{Q}_{\text{cs}} := \{q \mid q(z) = q(-z)\}$ be the set of continuous and origin-symmetric distributions. $\mathcal{Q}_{\text{cs}}$ includes all distributions studied so far in the context of randomized smoothing for continuous data. Randomized smoothing certificates for discrete data (Lee et al., 2019; Bojchevski et al., 2020) are out of our scope.

Proposition 3. *For any distribution $q \in \mathcal{Q}_{cs}$, $\mathbb{P}_{z\sim q}(h(x + z + \delta) = c_x) < \frac{1}{2}$ for a classifier h and a perturbation δ constructed as in Proposition 2.*

Proof. The blue region in Fig. 2 (right) illustrates the probability mass of $x + \delta + z, z \sim q$. According to $q(z) = q(-z)$, we have $\mathbb{P}_{z\sim q}(x+\delta+z) = \mathbb{P}_{z\sim q}(x+\delta-z)$, where $x + \delta + z$ and $x + \delta - z$ are in the two different subspaces A_1, A_2 generated by separating the sample space with a hyperplane $\delta^T z = \delta^T x + ||\delta||_2^2$.

Notice that $z \to 2(x + \delta) - z$ is a bijection between the two sample spaces, thus the probability mass of $x + \delta + z$ on A_1 and A_2 should be same. From the selection of classifier h, there exists one subspace A_2 in which all sample points are *not* classified as the class of x by h (right blue part in Fig. 2). Therefore $g(x + \delta) < 1 - \mathbb{P}_{x+\delta+z}(A_1) = \frac{1}{2}$. □

Theorem 1. *For any distribution $q \in \mathcal{Q}_{cs}$, when $R > \Psi^{-1}(g(x), q)$, the certified radius of any sample x w.r.t. the smoothed classifier is upper bounded by*

$$r_g(x) < ||\delta||_2 \leq R\sqrt{1 - I^{-1}_{2(1-\frac{g(x)}{\Psi(R;q)})}(\frac{d-1}{2}, \frac{1}{2})}, \tag{8}$$

where $I_*^{-1}(\frac{d-1}{2}, \frac{1}{2})$ is the inverse of $I_*(\frac{d-1}{2}, \frac{1}{2})$.

Proof. From Eq. 7 we have

$$\begin{aligned} &\frac{1}{2} I_{1-\frac{||\delta||_2^2}{R^2}}(\frac{d-1}{2}, \frac{1}{2}) \geq 1 - \frac{g(x)}{\Psi(R;q)} \\ \Leftrightarrow\ & 1 - \frac{||\delta||_2^2}{R^2} \geq I^{-1}_{2(1-\frac{g(x)}{\Psi(R;q)})}(\frac{d-1}{2}, \frac{1}{2}) \\ \Leftrightarrow\ & \frac{||\delta||_2^2}{R^2} \leq 1 - I^{-1}_{2(1-\frac{g(x)}{\Psi(R;q)})}(\frac{d-1}{2}, \frac{1}{2}) \\ \Leftrightarrow\ & ||\delta||_2 \leq R\sqrt{1 - I^{-1}_{2(1-\frac{g(x)}{\Psi(R;q)})}(\frac{d-1}{2}, \frac{1}{2})}. \end{aligned}$$ □

Proposition 4. *When $d \in [10^3, 10^7]$ and $\epsilon \in [10^{-6}, 1]$,*

$$\sqrt{1 - I_\epsilon^{-1}(\frac{d-1}{2}, \frac{1}{2})} < \frac{5}{\sqrt{d}} \tag{9}$$

Corollary 1. *Let $2(1 - \frac{g(x)}{\Psi(R;q)}) = 10^{-6}$, we have $R = \Psi^{-1}(\frac{g(x)}{1-5*10^{-7}}; q)$ and*

$$r_g(x) < \frac{5}{\sqrt{d}} \Psi^{-1}(\frac{g(x)}{1 - 5 * 10^{-7}}; q) \tag{10}$$

We call it "distribution-based" as Ψ is based on q.

Proof. From Eq. 8 we have

$$||\delta||_2 \leq R\sqrt{1 - I^{-1}_{2(1-\frac{g(x)}{\Psi(R;q)})}(\frac{d-1}{2}, \frac{1}{2})}.$$

Since $2(1 - \frac{g(x)}{\Psi(R;q)}) = 10^{-6}$ according to Proposition 4,

$$\sqrt{1 - I^{-1}_{2(1-\frac{g(x)}{\Psi(R;q)})}(\frac{d-1}{2}, \frac{1}{2})} < \frac{5}{\sqrt{d}}.$$

Therefore $r_g(x) < \frac{5}{\sqrt{d}} \Psi^{-1}(\frac{g(x)}{1-5*10^{-7}}; q)$. □

Trade-off Discussion. Whether there exists a fundamental trade-off between accuracy and robustness is an active open question (Tsipras et al., 2019; Yang et al.,

2020b). We discuss the influence of different smoothing distributions on robustness and accuracy. If we want to maintain the accuracy of our classifier, the distribution q should concentrate around the origin with a constant distance. According to Eq. 10 the certified radius will be of order $O(1/\sqrt{d})$. In order to improve the robustness of the smoothed classifier, $\Psi^{-1}(\cdot; q)$ should be $\Omega(\sqrt{d})$. In this case the probability mass of q will be pushed away from the origin, which might lead to poor accuracy of the smoothed classifier. For example, consider Gaussian smoothing, where the radius is given by $\sigma\Phi^{-1}(g(x))$. Here reducing σ tends to increase the prediction score $g(x)$ but does not necessarily improve the radius since we also multiply by σ. However, we cannot assert that there is a trade-off between accuracy and robustness. Robust training with noise augmentation (Cohen et al., 2019; Salman et al., 2019) can improve both clean accuracy and robustness at the same time (compared to standard training) since the classifier can learn to correctly classify noisy instances.

In conclusion, we encourage to use a large value of σ, and develop more powerful training technique to maintain the clean accuracy. For example, using consistency regularization (Jeong and Shin, 2020) the clean accuracy on MNIST with $\sigma = 0.25$ or $\sigma = 0.50$ are both 99.2%. Relatedly, prepending a custom-trained denoiser (Salman et al., 2020) to the classifier can somewhat mitigate the effect of the input noise.

5 ALTERNATIVE BOUND FOR l_2

Next, we derive a data-dependent upper bound which accounts for the predictions around an instance x.

Definition 3. *Given a sample point x, a classifier f, and $r > 0$, the smoothed classifier can be written as $g(x) = g_{\leq r}(x) + g_{>r}(x)$, where*

$$g_{\leq r}(x) := \mathbb{P}_{z\sim q}(\{z|f(x+z) = c_x\} \cap \{||z||_2 \leq r\})$$
$$g_{>r}(x) := \mathbb{P}_{z\sim q}(\{z|f(x+z) = c_x\} \cap \{||z||_2 > r\})$$

Definition 4 (Dominating radius)**.** *Given a sample point x, a classifier f and a smoothing distribution q, we can calculate the dominating radius R_x of x*

$$R_x = \inf_{\frac{g_{>r}(x)}{1-\Psi(r;q)} < \frac{1}{2}} r. \tag{11}$$

The intuition behind this definition is: As the radius increases, the prediction of points outside the ball $B(x, r)$ will be less influenced by the prediction of x. If $\frac{g_{>r}(x)}{1-\Psi(r;p)} < \frac{1}{2}$, we can say that x can hardly influence the prediction of the points outside the ball. This radius depends mainly on the original classifier f and the smoothing distribution q via $g_{>r}(x)$.

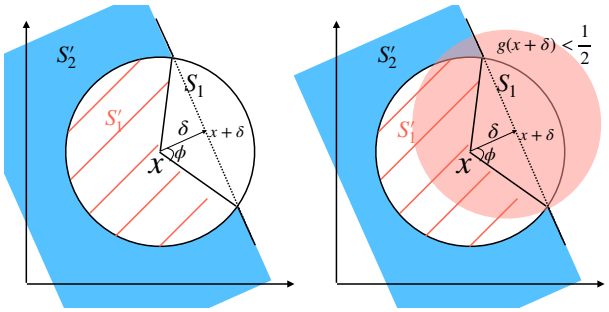


Figure 3: The chosen classifier $h(x) = c_x 1_{(x\in S_1')\vee(x\in S_2')}$ and perturbation δ.

Proposition 5. *For any distribution $q \in \mathcal{Q}_{cs}$, there exists a perturbation δ and a hyperspherical sector S_1, whose symmetric axis is along δ, in $B(R_x, x)$ (Fig. 3 left) and a classifier $h(x) = c_x 1_{(x\in S_1')\vee(x\in S_2')}$ such that*

$$q(S_1') = g_{\leq r}(x),$$
$$q(S_2') > \frac{1}{2}(1 - \Psi(R_x; q)) \geq g_{>r}(x),$$
$$q(\{z|h(x+z) = c_x\}) = q(S_1') + q(S_2') \geq g(x),$$

and

$$\frac{q(S_1)}{q(B(R_x, x))} \leq \frac{V_{S_1}(R)}{V_B(R)}$$

From here we have

$$\begin{aligned} 1 - \frac{g_{\leq R_x}(x)}{\Psi(R_x; q)} &= \frac{q(S_1)}{q(B(R_x, x))} \leq \frac{V_{S_1}(R)}{V_B(R)} \\ &= I_{sin^2(\phi)}(\frac{d-1}{2}, \frac{1}{2}) = I_{1-\frac{||\delta||_2^2}{R_x^2}}(\frac{d-1}{2}, \frac{1}{2}). \end{aligned} \tag{12}$$

Corollary 2. *For any distribution $q \in \mathcal{Q}_{cs}$, if $\frac{g_{\leq R_x}(x)}{\Psi(R_x;q)} < 1-5\cdot10^{-7}$, the certified radius of any sample x w.r.t. the smoothed classifier is bounded by*

$$r_g(x) < R_x \frac{5}{\sqrt{d}}. \tag{13}$$

Proof. From Eq. 12 we have

$$||\delta||_2 \leq R\sqrt{1 - I^{-1}_{2(1-\frac{g_{\leq R_x}(x)}{\Psi(R_x;q)})}(\frac{d-1}{2}, \frac{1}{2})}.$$

Since $2(1 - \frac{g_{\leq R_x}(x)}{\Psi(R_x;q)}) > 10^{-6}$, from Proposition 4

$$\sqrt{1 - I^{-1}_{2(1-\frac{g_{\leq R_x}(x)}{\Psi(R_x;q)})}(\frac{d-1}{2}, \frac{1}{2})} < \frac{5}{\sqrt{d}}. \tag{14}$$

Therefore $r_g(x) < R_x \frac{5}{\sqrt{d}}$. □

5.1 Estimating R_x

Since we cannot compute R_x in closed form, we design an algorithm for estimating it. Our Algorithm 1 mainly consists of two steps: a) calculating a table of $\frac{g_{>r}(x)}{1-\Psi(r,q)}$ values for different radii r, and b) selecting the smallest radius r with $\frac{g_{>r}(x)}{1-\Psi(r,q)} \leq 0.5$.

In step a) we first draw N samples from $\mathcal{N}(0, \sigma^2 I_d)$ and calculate the l_2 norm of each sample. Then we add the noise to the original sample and feed them into the original classifier f. The output will be an one-hot encoded vector. Next we create a linear search space for R_x, and a count table count$[l, c]$ where l is the length of the linear search space and c is the number of classes. For each sample, if its l_2 norm is larger than r of the search space, we add the one hot prediction of this sample to count$[r, :]$. In this way, $\frac{\text{count}[r,c_x]}{\sum(\text{count}[r,:])}$ and $\frac{\text{count}[0,c_x]-\text{count}[r,c_x]}{\sum(\text{count}[0,:]-\text{count}[r,:])}$ are the approximations of $\frac{g_{>r}(x)}{1-\Psi(r,q)}$ and $\frac{g_{\leq r}(x)}{\Psi(r,q)}$ respectively. We compute a probabilistic upper bound for these quantities using the Clopper-Pearson confidence interval (Clopper and Pearson, 1934) with a union confidence level $\alpha = 0.01$. Because of the dependency among the counts, we apply Bonferroni correction (Bonferroni, 1936) and use a confidence level of $\frac{\alpha}{2\cdot|S|}$ for each single count$[s, :]$. We need an upper bound for the approach to be sound since R_x is positively related to $\frac{g_{>r}(x)}{1-\Psi(r,q)}$.

In step b) we find the smallest r in the search space with $\text{CP}(\text{count}[s, y], \sum \text{count}[s, :], \alpha) \leq 0.5$ where $\text{CP}(\cdot, \cdot, \alpha)$ returns the Clopper-Pearson upper bound.

Improved Upper Bound. We also notice a gap between the second upper bound (Eq. 13) and the certified radius on Fig. 5 in our experiments. In Eq. 14, $\frac{5}{\sqrt{d}}$ is a loose bound of $\sqrt{1 - I^{-1}_{2(1-\frac{g_{\leq R_x}(x)}{\Psi(R_x;q)})}(\frac{d-1}{2}, \frac{1}{2})}$. It's possible to compute $\frac{g_{\leq R_x}(x)}{\Psi(R_x;q)}$ with our Algorithm 1. Therefore we will also use the following improved upper bound to compare with the certified radius

$$R_x \sqrt{1 - I^{-1}_{2\left(1-\frac{g_{\leq R_x}(x)}{\Psi(R_x;q)}\right)}(\frac{d-1}{2}, \frac{1}{2})}. \tag{15}$$

6 EXTENSION TO l_p $(p > 2)$

Similar to Kumar et al. (2020), we extend our result to l_p robustness certification. Unlike them however, our result applies to spherical symmetric distributions. The spherical symmetric family and the exponential family in Kumar et al. (2020) are two important distribution families, and the Gaussian distribution is the only intersection between them (see also Fig. 1).

Algorithm 1: Estimating R_x

Input: input point x; target y; original classifier f; max length L; number of samples N; standard deviation σ; weight function W; search step a; number of classes c; confidence level α;

```
sample z_1, z_2, ..., z_N from N(0, σ²I_d);
compute l_2 norm of samples r_i = ||z_i||_2;
compute search space for R_x: S = range(0, L, a);
create a count table: count[s, c] = 0;
for i = 1 : N do
    compute one-hot predictions p_i = f(x + z_i);
    for j,s in enumerate(S) do
        if r_i ≥ s then
            count[s, :] += p_i;
        end
    end
end
for s ∈ S do
    g_2 = g_{>r}(x)/(1−Ψ(s)) = CP(count[s, y],
          Σ count[s, :], α/(2|S|));
    g_1 = g_{≤r}(x)/Ψ(s) = CP(count[0, y] − count[s, y],
          Σ(count[0, :] − count[s, :])], α/(2|S|));
    if g_2 < 0.5 then
        return s, g_1;
    end
end
return L, g_1;
```

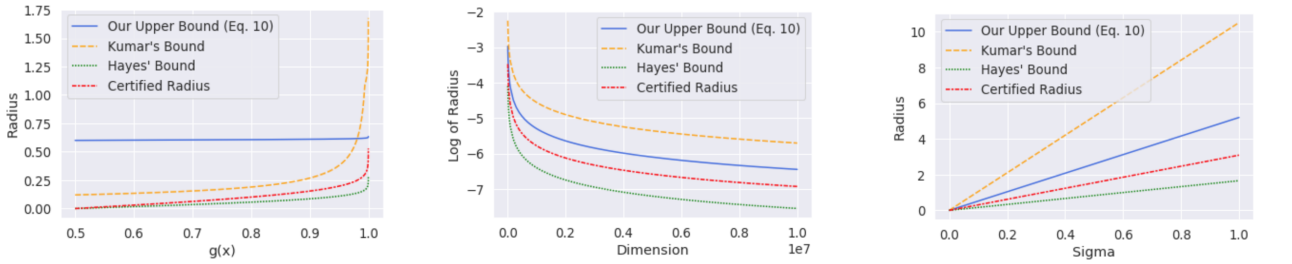
Corollary 3. *If q is a spherical distribution i.e. $q(z) = q(||z||_2)$, Proposition 2 and Proposition 5 hold for any perturbation δ. Therefore, we can choose $\delta = (\frac{b}{\sqrt{d}}, \frac{b}{\sqrt{d}}, \frac{b}{\sqrt{d}}, ..., \frac{b}{\sqrt{d}})$ and classifier $h = c_x 1_{(x \in S'_1)}$ or $h = c_x 1_{x \in (S'_1 \cup S'_2)}$ such that the l_p certified radius is*

$$r_g(x) < \frac{5}{d^{1-\frac{1}{p}}} \Psi^{-1}\left(\frac{g(x)}{1 - 5 * 10^{-7}}; q\right),$$
$$\text{or} \quad r_g(x) < R_x \frac{5}{d^{1-\frac{1}{p}}}$$

However, if q is not spherical symmetric distributed, $\frac{q(S_1)}{q(B(R,x))} \leq \frac{V_{S_1}(R)}{V_B(R)}$ might not hold for this choice of δ and the radius is not necessarily of order $O(1/d^{1-\frac{1}{p}})$.

7 LIMITATIONS

Our results apply to randomized smoothing certificates which are functions of only $g(x)$. Better certificates may be possible if more information is available to the certification algorithm (Dvijotham et al., 2020). However, we suspect that similar curse of dimensionality effects are still present for high dimensional problems.



(a) $\sigma = 1$, $d = 3 \cdot 32 \cdot 32$, $g(x)$ varies (b) $\sigma = 1/\sqrt{d}$, $g(x) = 0.999$, d varies (c) $g(x) = 0.999$, $d = 3 \cdot 32 \cdot 32$, σ varies

Figure 4: Our first upper bound (blue), Kumar et al. (2020)'s bound (orange), Hayes (2020)'s bound (green), and the certified radius by Cohen et al. (2019) (red). We choose $g(x) = 0.999$ for the second and the third case because in practice, most $g(x)$ values are close to 1. Our upper bound is better in most cases. Hayes (2020)'s upper bound is below the certified radius since it bounds the divergence-based radius which is not tight.

8 EXPERIMENTS

Since there is a closed form of the certified radius $r_g(x) = \sigma\Phi^{-1}(g(x))$ for smoothing with a Gaussian, $q \sim \mathcal{N}(0, \sigma^2 I_d)$ (Cohen et al., 2019), we conduct three experiments with Gaussian smoothing to compare our bounds with the existing methods. Since both the i.i.d and the generalized Gaussian bounds in Kumar et al. (2020) are applicable for the Gaussian distribution, for a fair comparison we show the minimum of their two bounds in each case. In Sec. 3 of the appendix we plot the two bounds by Kumar et al. (2020) separately as a reference in relation to the rest.

8.1 Evaluating our first upper bound for different parameters

In this experiment we evaluate our first upper bound derived in Sec. 4, together with Kumar et al. (2020)'s and Hayes (2020)'s bounds, and the certified radius.

For $q \sim \mathcal{N}(0, \sigma^2 I_d)$ we have that

$$\Psi(R; q) = \int_{||z||_2 < R} q(z)dz = \text{Gamma}(\frac{R^2}{2\sigma^2}; \frac{d}{2}, 1),$$

where $\text{Gamma}(\cdot; \frac{d}{2}, 1)$ is the cumulative density function of a Gamma distribution with shape $\frac{d}{2}$ and rate 1. See Lemma 2 of Sec. 1 in the appendix. Therefore, $\Psi^{-1}(x; q) = \sigma\sqrt{2\text{Gamma}^{-1}(x; \frac{d}{2}, 1)}$.

There are three parameters σ, d, and $g(x)$ which control the value of the first upper bound. We mainly focus on the l_2 radius, because the l_p bound of Kumar et al. (2020) and ours are exactly given by the respective l_2 bounds multiplied by the same factor $d^{1/p-1/2}$. Since the results for l_p radius look the same as the results for l_2 radius, we show experiments only for l_2.

Fig. 4 shows our first upper bound and the certified radius $r_g(x)$. In Fig. 4(a), if we keep σ and d as constants, our upper bound is strictly larger than the real radius, and the difference between them decreases when $g(x)$ increases. Note for values of $g(x) \approx 1$, which are most relevant in practice, the gap is relatively small.

Next, we evaluate how large dimensions affects the certified radius while maintaining high clean accuracy. In this case the smoothing distribution should be concentrated around the origin. We set $\sigma = 1/\sqrt{d}$, so the probability mass is concentrated around the $\mathbb{R}^d$ unit ball. Fig. 4(b) shows that both the upper bound and real radius decrease quickly as the dimension increases.

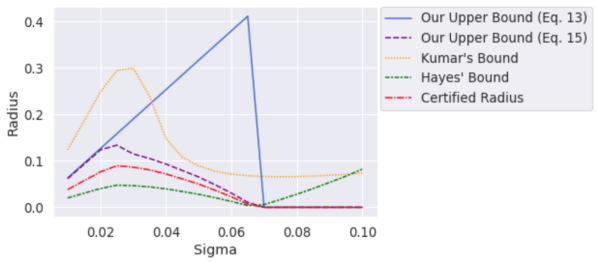
Finally, in Fig. 4(c) we show that if we push the probability mass away from the origin, while keeping $g(x)$ and d invariant, the certified radius grows linearly with σ. This is actually the effect of robust training: maintaining large $g(x)$ while increasing σ. However, for a large enough σ, the smoothed classifier starts behaving like a constant classifier. This means that in practice we cannot maintain a large value for $g(x)$ for all instances, even with robust training. This is further evidence for the curse of dimensionality, since σ has to scale with d to obtain meaningful guarantees.

In all plots on Fig. 4 Hayes (2020)'s bound is below the certified radius since it bounds the divergence-based radius which is not tight. Our bound outperforms Kumar et al. (2020)'s bound in most cases.

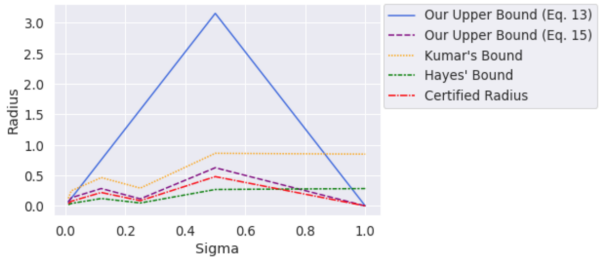
8.2 Our second upper bound on real data

Since our second upper bound is data dependent, via the R_x term (Eq. 13, Eq. 15), we evaluate it on real data. Specifically, in this experiment we train a ResNet110 on CIFAR10 with and without noise augmentation. See Sec. 2 in the appendix for more details about the training and the hyperparameters.

We calculate R_x for different test instances with different values of σ on both ResNet models. We com-

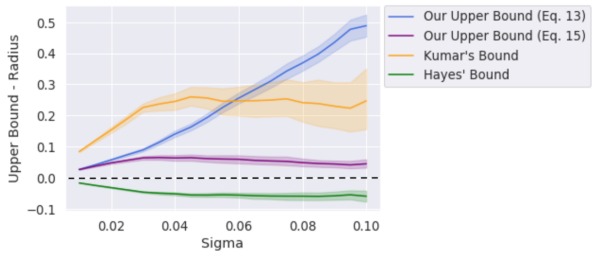


(a) ResNet110 without robust training on CIFAR10

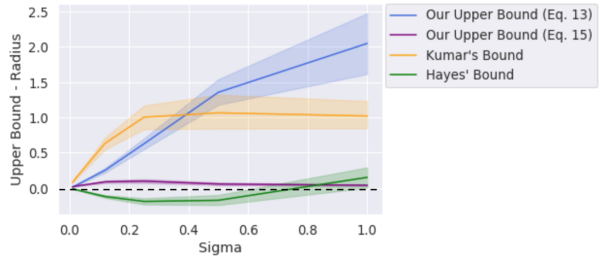


(b) ResNet110 with robust training on CIFAR10

Figure 5: Comparing different upper bounds and the certified radius w.r.t. σ for a random CIFAR10 test sample, and ResNet110 models without (a) and with (b) robust training. With robust training we can certify larger radii since the model remains accurate for larger σ. Our upper bounds have the smallest gap to the certified radius.



(a) ResNet110 without robust training on CIFAR10



(b) ResNet110 with robust training on CIFAR10

Figure 6: The mean and standard deviation of the difference between different upper bounds and the certified radius over 100 random samples from CIFAR10 with respect to σ. In (a) σ is from a linear search space of $[0.01, 0.1]$, while in (b) $\sigma \in \{0.01, 0.02, 0.12, 0.25, 0.5, 1.0\}$. Both of our upper bounds outperform the competitors.

pare the second upper bound with the certified radius from Cohen et al. (2019). We draw 10^5 samples from $\mathcal{N}(0, \sigma^2 I_d)$ and calculate $g_{\leq r}(x)$ and $g_{>r}(x)$ with the Monte Carlo method. We then compute their Clopper-Pearson upper bounds with confidence level $\alpha = 0.01$ and use these value for computing R_x. This means that these upper bounds are probabilistic and hold with probability $1 - \alpha$. The pseudo-code is shown in Algorithm 1. Our code is available at `https://github.com/YihanWu95/smoothing`.

One Random Test Sample. We first use one random instance from the test set to compare different upper bounds. In this way we can closely examine how the certified radius and the upper bounds change with different values of the variance σ.

Fig. 5 illustrates the relation of σ and the certified radius on models with and without robust training. For the normally trained model we use different σ from 0.01 to 0.1. As Fig. 5(a) shows, when σ increases, the radius will firstly increase and then drop down to 0. Zero radius indicates $g(x) < 0.5$, i.e. x is not correctly classified. If we train the model with noise augmentation, we are able to increase the certified radius while keeping x corrected classified for larger σ (Fig. 5(b)).

We also observe that Hayes (2020)'s bound is again lower than the certified radius in both plots, and our second bound (Eq. 13) is comparable to Kumar et al. (2020)'s bound. Besides, the error between our improved upper bound (Eq. 15) and the real radius is surprisingly small. Similar results hold for other random test samples as we show in the next experiment.

Average Over 100 Random Test Samples. As the result from random sample cannot convincingly show which bound is better, in this experiment we select 100 test samples randomly and calculate the mean and the standard deviation of the difference between the upper bounds and the certified radius. Note, the y-axis of the plots show (upper bound − certified radius) instead of the absolute value of the difference, because the divergence bound by Hayes (2020) is smaller than the certified radius (hence the difference is negative).

On Fig. 6 we can see that our improved upper bound (Eq. 15) outperforms the other bounds, and the gap to the certified radius is close to zero for all values of σ. The standard deviation tends to increase with σ for most bounds, and is on average smaller for our bounds. This shows that we cannot hope for a significantly better certificate without making additional assumptions.

9 CONCLUSION

In this work, we show that the limitations of high-dimensional randomized smoothing extend to large family of continuous and origin-symmetric distributions for l_2 adversarial robustness. Without a robust training procedure, the l_2 certified radius of the smoothed classifier could be of order $O(1/\sqrt{d})$. While noise augmentation partially alleviates the issue, it does not overcome it. We purpose two upper bounds of the l_2 certified radius, which indicate the smoothing distribution must not be concentrated at the origin and should assign a large probability mass to samples at distance of order $\Omega(\sqrt{d})$ to avoid the curse of dimensionality. However, this may lead to poor clean accuracy. Our upper bounds outperform the existing bounds for Gaussian smoothing, and the gap between our bounds and the certified radius is small. We also adapt our result to spherical symmetric distribution for l_p robustness. When $p > 2$, the worst case l_p certified radius with spherical symmetric distributions is of order $O(1/d^{1-\frac{1}{p}})$. We complete the picture and show that the curse of dimensionality applies broadly.

Future work. Extending our upper bounds for l_p ($p > 2$) to continuous and origin-symmetric distributions, and other tasks such as regression, is a viable future direction. Upper bounding the certified radius when additional information is available (e.g. second-order constraints) can help us better understand the limitations of randomized smoothing.

Acknowledgements

This research was supported by the German Research Foundation through the Emmy Noether grant GU1409/2-1, and by BMW AG.

References

N. Akhtar and A. Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

A. Blum, T. Dick, N. Manoj, and H. Zhang. Random smoothing might be unable to certify l robustness for high-dimensional images. *Journal of Machine Learning Research*, 21:1–21, 2020.

A. Bojchevski, J. Klicpera, and S. Günnemann. Efficient robustness certificates for discrete data: Sparsity-aware randomized smoothing for graphs, images and more. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Workshop on Artificial Intelligence and Security, AISec*, 2017.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. *ICML*, 2019.

K. D. Dvijotham, J. Hayes, B. Balle, Z. Kolter, C. Qin, A. Gyorgy, K. Xiao, S. Gowal, and P. Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020.

H. X. Y. M. Hao-Chen, L. D. Deb, H. L. J.-L. T. Anil, and K. Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

J. Hayes. Extensions and limitations of randomized smoothing for robustness guarantees. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 786–787, 2020.

J. Jeong and J. Shin. Consistency regularization for certified robustness of smoothed classifiers. In *Adavnces in Neural Information Processing Systems, NeurIPS*, 2020.

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

M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, may 2019. doi: 10.1109/sp.2019.00044.

G.-H. Lee, Y. Yuan, S. Chang, and T. Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems*, 2019.

B. Li, C. Chen, W. Wang, and L. Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, pages 9464–9474, 2019.

S. Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.

H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep

learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 11292–11303, 2019.

H. Salman, M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In *Adavnces in Neural Information Processing Systems, NeurIPS*, 2020.

V. Tjeng, K. Y. Xiao, and R. Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019.

F. Tramer, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

G. Yang, T. Duan, J. E. Hu, H. Salman, I. Razenshteyn, and J. Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020a.

Y.-Y. Yang, C. Rashtchian, H. Zhang, R. Salakhutdinov, and K. Chaudhuri. A closer look at accuracy vs. robustness. In *Adavnces in Neural Information Processing Systems, NeurIPS*, 2020b.

D. Zhang, M. Ye, C. Gong, Z. Zhu, and Q. Liu. Black-box certification with randomized smoothing: A functional optimization based framework. In *Adavnces in Neural Information Processing Systems, NeurIPS*, 2020a.

H. Zhang. Random smoothing might be unable to certify l_∞ robustness for high-dimensional images. 2020.

H. Zhang, P. Zhang, and C.-J. Hsieh. Recurjac: An efficient recursive algorithm for bounding jacobian matrix of neural networks and its applications. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 5757–5764, 2019.

H. Zhang, H. Chen, C. Xiao, S. Gowal, R. Stanforth, B. Li, D. Boning, and C.-J. Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020b.

T. Zheng, D. Wang, B. Li, and J. Xu. Towards assessment of randomized mechanisms for certifying adversarial robustness. *arXiv preprint arXiv:2005.07347*, 2020.