Journal of Machine Learning Research 21 (2020) 1-21 Submitted 3/20; Published 9/20

Random Smoothing Might be Unable to Certify ℓ_∞ Robustness for High-Dimensional Images

Avrim Blum AVRIM@TTIC.EDU
Toyota Technological Institute at Chicago

Travis Dick TBD@SEAS.UPENN.EDU
University of Pennsylvania

Naren Manoj NSM@TTIC.EDU
Toyota Technological Institute at Chicago

Hongyang Zhang* HONGYANZ@TTIC.EDU
Toyota Technological Institute at Chicago

Editor: Pushmeet Kohli

Abstract

We show a hardness result for random smoothing to achieve certified adversarial robustness against attacks in the ℓ_p ball of radius ϵ when $p > 2$. Although random smoothing has been well understood for the ℓ_2 case using the Gaussian distribution, much remains unknown concerning the existence of a noise distribution that works for the case of $p > 2$. This has been posed as an open problem by Cohen et al. (2019) and includes many significant paradigms such as the ℓ_∞ threat model. In this work, we show that *any* noise distribution $\mathcal{D}$ over $\mathbb{R}^d$ that provides ℓ_p robustness for all base classifiers with $p > 2$ must satisfy $\mathbb{E}\,\eta_i^2 = \Omega(d^{1-2/p}\epsilon^2(1-\delta)/\delta^2)$ for 99% of the features (pixels) of vector $\eta \sim \mathcal{D}$, where ϵ is the robust radius and δ is the score gap between the highest-scored class and the runner-up. Therefore, for high-dimensional images with pixel values bounded in $[0, 255]$, the required noise will eventually dominate the useful information in the images, leading to trivial smoothed classifiers.

Keywords: random smoothing, certified adversarial robustness, hardness results, high-dimensional data, ℓ_p adversarial examples

1. Introduction

Adversarial robustness has been a critical object of study in various fields, including machine learning (Zhang et al., 2019; Madry et al., 2018), computer vision (Szegedy et al., 2013; Yang et al., 2020b), and many other domains (Lecuyer et al., 2019). In machine learning and computer vision, the study of adversarial robustness has led to significant advances in defending against attacks in the form of perturbed input images, where the data is high dimensional but each feature is bounded in $[0, 255]$. The problem can be stated as that of learning a non-trivial classifier with high test accuracy on the adversarial images. The adversarial perturbation is either restricted to be in an ℓ_p ball of radius ϵ centered at 0 (Yang et al., 2020c), or is measured under other threat models such as Wasserstein distance and adversarial rotation (Wong et al., 2019; Brown et al., 2018). The focus of this work is the former setting.

*. Corresponding author.

Despite a large amount of work on adversarial robustness, many fundamental problems remain open. One of the challenges is to end the long-standing arms race between adversarial defenders and attackers: defenders design empirically robust algorithms which are later exploited by new attacks designed to undermine those defenses (Athalye et al., 2018). This motivates the study of *certified robustness* (Raghunathan et al., 2018; Wong et al., 2018)—algorithms that are provably robust to the worst-case attacks—among which random smoothing (Cohen et al., 2019; Li et al., 2019; Lecuyer et al., 2019) has received significant attention in recent years. Algorithmically, random smoothing takes a base classifier f as an input, and outputs a smooth classifier g that, given an input example x, outputs the most probable class predicted by f on a distribution of perturbed versions of x.

Random smoothing has many appealing properties that one could exploit: it is agnostic to network architecture, is scalable to deep networks, and perhaps most importantly, achieves state-of-the-art certified ℓ_2 robustness for deep learning based classifiers (Cohen et al., 2019; Li et al., 2019; Lecuyer et al., 2019).

Open problems in random smoothing. Given the rotation invariance of Gaussian distribution, most positive results for random smoothing have focused on the ℓ_2 robustness achieved by smoothing with the Gaussian distribution (see Theorem 5). However, the existence of a noise distribution for general ℓ_p robustness has been posed as an open question by Cohen et al. (2019):

We suspect that smoothing with other noise distributions may lead to similarly natural robustness guarantees for other perturbation sets such as general ℓ_p norm balls.

Several special cases of the conjecture have been proven for $p < 2$. Li et al. (2019) show that ℓ_1 robustness can be achieved with the Laplacian distribution, and Lee et al. (2019) show that ℓ_0 robustness can be achieved with a discrete distribution. Much remains unknown concerning the case when $p > 2$. On the other hand, the most standard threat model for adversarial examples is ℓ_∞ robustness, among which 8-pixel and 16-pixel attacks have received significant attention in the computer vision community (i.e., the adversary can change every pixel by 8 or 16 intensity values, respectively). In this paper, we derive lower bounds on the magnitude of noise required for certifying ℓ_p robustness that highlights a phase transition at $p = 2$. In particular, for $p > 2$, the noise that must be added to each feature of the input examples grows with the dimension d in expectation, while it can be constant for $p \leq 2$.

Preliminaries. Given a base classifier $f : \mathbb{R}^d \to \mathcal{Y}$ and smoothing distribution $\mathcal{D}$, the randomly smoothed classifier is defined as follows: for each class $y \in \mathcal{Y}$, define the score of class y at point x to be $G_y(x; \mathcal{D}, f) = \Pr_{\eta \sim \mathcal{D}}(f(x + \eta) = y)$. Then the smoothed classifier outputs the class with the highest score: $g(x; \mathcal{D}, f) = \operatorname{argmax}_y G_y(x; \mathcal{D}, f)$.

The key property of smoothed classifiers is that the scores $G_y(x; \mathcal{D}, f)$ change slowly as a function of the input point x (the rate of change depends on $\mathcal{D}$). It follows that if there is a gap between the highest and second highest class scores at a point x, the smoothed classifier $g(\cdot; \mathcal{D}, f)$ must be constant in a neighborhood of x. We denote the score gap by $\Delta(x; \mathcal{D}, f) = G_a(x; \mathcal{D}, f) - G_b(x; \mathcal{D}, f)$, where $a = \operatorname{argmax}_y G_y(x; \mathcal{D}, f)$ and $b = \operatorname{argmax}_{y \neq a} G_y(x; \mathcal{D}, f)$.

Definition 1 ($(\mathcal{A}, \delta)$- and (ϵ, δ)-robustness) *For any set $\mathcal{A} \subseteq \mathbb{R}^d$ and $\delta \in [0, 1]$, we say that the smoothed classifier g is* $(\mathcal{A}, \delta)$-robust *if for all $x \in \mathbb{R}^d$ with $\Delta(x; \mathcal{D}, f) > \delta$, we have that $g(x + v; \mathcal{D}, f) = g(x; \mathcal{D}, f)$ for all $v \in \mathcal{A}$. For a given norm $\|\cdot\|$, we also say that g is (ϵ, δ)-robust with respect to $\|\cdot\|$ if it is $(\mathcal{A}, \delta)$-robust with $\mathcal{A} = \{v \in \mathbb{R}^d : \|v\| \leq \epsilon\}$.*

When the base classifier f and the smoothing distribution $\mathcal{D}$ are clear from context, we will simply write $G_y(x)$, $g(x)$, and $\Delta(x)$. We often refer to a sample from the distribution $\mathcal{D}$ as *noise*, and use *noise magnitude* to refer to squared ℓ_2 norm of a noise sample. Finally, we use $\mathcal{D} + v$ to denote the distribution of $\eta + v$, where $\eta \sim \mathcal{D}$.

1.1 Our Results

Our main results derive lower bounds on the magnitude of noise sampled from any distribution $\mathcal{D}$ that leads to (ϵ, δ)-robustness with respect to $\|\cdot\|_p$ for all possible base classifiers $f : \mathbb{R}^d \to \mathcal{Y}$. A major strength of random smoothing is that it provides certifiable robustness guarantees without making any assumption on the base classifier $f : \mathbb{R}^d \to \mathcal{Y}$. For example, the results of Cohen et al. (2019) imply that using a Gaussian smoothing distribution with standard deviation $\sigma = \frac{2\epsilon}{\delta}$ guarantees that $g(\cdot; \mathcal{D}, f)$ is (ϵ, δ)-robust with respect to $\|\cdot\|_2$ for every possible base classifier $f : \mathbb{R}^d \to \mathcal{Y}$. We show that there is a phase transition at $p = 2$, and that ensuring (ϵ, δ)-robustness for all base classifiers f with respect to ℓ_p norms with $p > 2$ requires that the noise magnitude grows non-trivially with the dimension d of the input space. In particular, for image classification tasks where the data is high dimensional and each feature is bounded in the range $[0, 255]$, this implies that for sufficiently large dimensions, the necessary noise will dominate the signal in each example.

The following result, proved in Appendix A, shows that any distribution $\mathcal{D}$ that provides $(\mathcal{A}, \delta)$-robustness for every possible base classifier $f : \mathbb{R}^d \to \mathcal{Y}$ must be approximately translation-invariant to all translations $v \in \mathcal{A}$. More formally, for every $v \in \mathcal{A}$, we must have that the total variation distance between $\mathcal{D}$ and $\mathcal{D} + v$, denoted by $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) := \sup_{\mathcal{S}} |\Pr_{Z \sim \mathcal{D}}(Z \in \mathcal{S}) - \Pr_{Z' \sim \mathcal{D}+v}(Z' \in \mathcal{S})|$, is bounded by δ. The rest of our results are consequences of this approximate translation-invariance property.

Lemma 2 *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$ such that for every (randomized) classifier $f : \mathbb{R}^d \to \mathcal{Y}$, the smoothed classifier $g(\cdot; \mathcal{D}, f)$ is $(\mathcal{A}, \delta)$-robust. Then for all $v \in \mathcal{A}$, we have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \leq \delta$.*

Lower bound on noise magnitude. Our first result is a lower bound on the expected squared ℓ_2-magnitude of a sample $\eta \sim \mathcal{D}$ for any distribution $\mathcal{D}$ that is approximately invariant to ℓ_p-translations of size ϵ.

Theorem 3 *Fix any $p \geq 2$ and let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$ such that there exists a radius ϵ and total variation bound δ satisfying that, for all $v \in \mathbb{R}^d$ with $\|v\|_p \leq \epsilon$, we have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \leq \delta$. Then*

$$\mathop{\mathbb{E}}_{\eta \sim \mathcal{D}} \|\eta\|_2^2 \geq \frac{\epsilon^2 d^{2-2/p}}{800} \cdot \frac{1 - \delta}{\delta^2}.$$

As a consequence of Theorem 3 and Lemma 2, it follows that any distribution that ensures (ϵ, δ)-robustness with respect to $\|\cdot\|_p$ for any base classifier f must also satisfy the same lower bound.

Phase transition at $p = 2$. The lower bound given by Theorem 3 implies a phase transition in the nature of distributions $\mathcal{D}$ that are able to ensure (ϵ, δ)-robustness with respect to $\|\cdot\|_p$ that occurs at $p = 2$. For $p \leq 2$, the necessary expected squared ℓ_2-magnitude of a sample from $\mathcal{D}$ grows only like $\sqrt{d}$, which is consistent with adding a constant level of noise to every feature in the input example (e.g., as would happen when using a Gaussian distribution with standard deviation $\sigma = \frac{2\epsilon}{\delta}$). On the other hand, for $p > 2$, the expected ℓ_2 magnitude of a sample from $\mathcal{D}$ grows strictly faster than $\sqrt{d}$,

which, intuitively, requires that the noise added to each component of the input example must scale with the input dimension d, rather than remaining constant as in the $p \leq 2$ regime. More formally, we prove the following:

Theorem 4 (hardness of random smoothing) *Fix any $p > 2$ and let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$ such that for every (randomized) classifier $f : \mathbb{R}^d \to \mathcal{Y}$, the smoothed classifier $g(\cdot; \mathcal{D}, f)$ is $(\mathcal{A}, \delta)$-robust. Let η be a sample from $\mathcal{D}$. Then at least 99% of the components of η satisfy $\mathbb{E}\,\eta_i^2 = \Omega(\frac{\epsilon^2 d^{1-2/p}(1-\delta)}{\delta^2})$. Moreover, if $\mathcal{D}$ is a product measure of i.i.d. noise (i.e., $\mathcal{D} = (\mathcal{D}')^d$), then the tail of $\mathcal{D}'$ satisfies $\Pr_{\zeta\sim\mathcal{D}'}(|\zeta| > s) \geq \left(\frac{c\epsilon(1-\delta)}{s\delta}\right)^{2p/(p-2)}$ for some $s > c\epsilon(1-\delta)/\delta$, where c is an absolute constant. In other words, $\mathcal{D}'$ is a heavy-tailed distribution.*[1]

The phase transition at $p = 2$ is more clearly evident from Theorem 4. In particular, the variance of most components of the noise must grow with $d^{1-2/p}$, which is an increasing function of d when $p > 2$. Theorem 4 shows that *any* distribution that provides (ϵ, δ)-robustness with respect to $\|\cdot\|_p$ for $p > 2$ must have very high variance in most of its component distributions when the dimension d is large. In particular, for $p = \infty$ the variance grows linearly with the dimension. Similarly, if we use a product distribution to achieve (ϵ, δ)-robustness with respect to $\|\cdot\|_p$ with $p > 2$, then each component of the noise distribution must be heavy-tailed and is likely to generate very large perturbations.

1.2 Technical Overview

Total-variation bound of noise magnitude. Our results demonstrate a strong connection between the required noise magnitude $\mathbb{E}\,\|\eta\|_2^2$ in random smoothing and the total variation distance between $\mathcal{D}$ and its shifted distribution $\mathcal{D} + v$ in the worst-case direction v. The total variation distance has a natural interpretation from the hardness of testing $\mathcal{D}$ v.s. $\mathcal{D} + v$: no classifier can distinguish $\mathcal{D}$ from $\mathcal{D} + v$ with good probability relative to $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v)$. Our analysis applies the following techniques.

Warm-up: one-dimensional case. We begin our analysis of Theorem 3 with the one-dimension case, by studying the projection of the noise $\eta \in \mathbb{R}^d$ on a direction $v \in \mathbb{R}^d$. Chebyshev's inequality implies $\mathbb{E}_{\eta\sim\mathcal{D}}\,|v^\top \eta|^2 \geq \|v\|_2^4\,(1-\delta)/8$. To see this, let η be a sample from $\mathcal{D}$ and let $\eta' = \eta + v$ so that η' is a sample from $\mathcal{D} + v$. Define $Z = v^\top \eta$ and $Z' = v^\top \eta' = Z + \|v\|_2^2$. Define $r = \|v\|_2^2\,/2$ so that the intervals $\mathcal{A} = (-r, r)$ and $\mathcal{B} = [\|v\|_2^2 - r, \|v\|_2^2 + r]$ are disjoint. From Chebyshev's inequality, we have $\Pr(Z \in \mathcal{A}) \geq 1 - \mathbb{E}\,|Z|^2/r^2$. Similarly, $\Pr(Z' \in \mathcal{B}) \geq 1 - \mathbb{E}\,|Z|^2/r^2$ and, since $\mathcal{A}$ and $\mathcal{B}$ are disjoint, this implies $\Pr(Z' \in \mathcal{A}) < \mathbb{E}\,|Z|^2/r^2$. Therefore, $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \geq \Pr(Z \in \mathcal{A}) - \Pr(Z' \in \mathcal{A}) \geq 1 - 2\,\mathbb{E}\,|Z|^2/r^2$. The claim follows from rearranging this inequality and the fact $\delta \geq \mathrm{TV}(\mathcal{D}, \mathcal{D} + v)$.

The remainder of the one-dimensional case is to show $\mathbb{E}_{\eta\sim\mathcal{D}}\,|v^\top \eta| \geq \|v\|_2^2\,\frac{(1-\delta)^2}{8\delta}$. To this end, we exploit a nice property of total variation distance in $\mathbb{R}$: every ϵ-interval $I = [a, a + \epsilon)$ satisfies $\mathcal{D}(I) \leq \mathrm{TV}(\mathcal{D}, \mathcal{D} + \epsilon)$. We note that for any $\tau \geq 0$, rearranging Markov's inequality gives $\mathbb{E}\,|v^\top \eta| \geq \tau \Pr(|v^\top \eta| > \tau) = \tau(1 - \Pr(|v^\top \eta| \leq \tau))$. We can cover the set $\{x \in \mathbb{R} : |x| \leq \tau\}$ using $\lceil \frac{2\tau}{\epsilon} \rceil$ intervals of width $\epsilon = \|v\|_2^2$ and, by this property, each of those intervals has probability

1. A distribution is heavy-tailed, if its tail is not an exponential function of x for all $x > 0$ (i.e., not an sub-exponential or sub-Gaussian distribution) (Vershynin, 2018).

mass at most δ. It follows that $\Pr(|v^\top \eta| \le \tau) \le \lceil \frac{2\tau}{\epsilon} \rceil \delta$, implying $\mathbb{E}\,|v^\top \eta| \ge \tau(1 - \lceil \frac{2\tau}{\epsilon} \rceil \delta)$. Finally, we optimize τ to obtain the bound $\mathbb{E}_{\eta \sim \mathcal{D}}\,|v^\top \eta| \ge \|v\|_2^2 \frac{(1-\delta)^2}{8\delta}$, as desired.

Extension to the d-dimensional case. We use the Pythagorean theorem to bridge from the one-dimensional case to the d-dimensional case. If there exists a set of orthogonal directions v_i's such that $\mathbb{E}_{\eta \sim \mathcal{D}}\,|v_i^\top \eta|^2 \ge \frac{\|v_i\|_2^4}{200} \frac{1-\delta}{\delta^2}$ and $\|v_i\|_2 = \epsilon d^{1/2-1/p}$ (the furthest distance to x in the ℓ_p ball $\mathcal{B}_p(x, \epsilon)$), then the Pythagorean theorem implies the result for the d-dimensional case straightforwardly. The existence of a set of orthogonal directions that satisfy these requirements is easy to find for the ℓ_2 case, because the ℓ_2 ball is isotropic and any set of orthogonal bases of $\mathbb{R}^d$ satisfies the conditions. However, the problem is challenging for the ℓ_p case, since the ℓ_p ball is not isotropic in general. In Corollary 11, we show that there exist at least $d/2$ vectors v_i that satisfy the requirements. Using the Pythagorean theorem in the subspace spanned by such v_i's gives Theorem 3.

Peeling argument and tail probability. We now summarize our main techniques to prove Theorem 4. Since $\|\eta\|_2 \le \sqrt{d}\|\eta\|_\infty$, Theorem 3 implies $\mathbb{E}\,\eta_i^2 \ge \frac{\epsilon d^{1/2-1/p}}{800} \cdot \frac{1-\delta}{\delta^2}$ for at least one index i, which shows that at least one component of η is large. However, this guarantee only highlights the largest pixel of $|\eta|$. Rather than working with the ℓ_∞-norm of η, we apply a similar argument to show that the variance of at least one component of η must be large. Next, we consider the $(d-1)$-dimensional distribution obtained by removing the highest-variance feature. Applying an identical argument, the highest-variance remaining feature must also be large. Each time we repeat this procedure, the strength of the variance lower bound decreases since the dimensionality of the distribution is decreasing. However, we can apply this peeling strategy for any constant fraction of the components of η to obtain lower bounds. The tail-probability guarantee in Theorem 4 follows a standard moment analysis in (Vershynin, 2018).

Summary of our techniques. Our proofs—in particular, the use of the Pythagorean theorem—show that defending against adversarial attacks in the ℓ_p ball of radius ϵ by random smoothing is almost as hard as defending against attacks in the ℓ_2 ball of radius $\epsilon d^{1/2-1/p}$. Therefore, the ℓ_∞ certification procedure—firstly using Gaussian smoothing to certify ℓ_2 robustness and then dividing the ℓ_2 certified radius by $\sqrt{d}$ as in (Salman et al., 2019)—is almost an optimal random smoothing approach for certifying ℓ_∞ robustness. The principle might hold generally for other threat models beyond ℓ_p robustness, and sheds light on the design of new random smoothing and proofs of hardness in the other threat models broadly.

2. Related Works

ℓ_2 robustness. Probably one of the most well-understood results for random smoothing is the ℓ_2 robustness. With Gaussian random noises, Lecuyer et al. (2019) and Li et al. (2019) provided the first guarantee of random smoothing and was later improved by Cohen et al. (2019) with the following theorem.

Theorem 5 (Theorem 1 of Cohen et al. (2019)) *Let $f : \mathbb{R}^d \to \mathcal{Y}$ by any deterministic or random classifier, and let $\eta \sim \mathcal{N}(0, \sigma^2 I)$. Let $g(x) = \operatorname{argmax}_{c \in \mathcal{Y}} \Pr(f(x+\eta) = c)$. Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy: $\Pr(f(x+\eta) = c_A) \ge \underline{p_A} \ge \overline{p_B} \ge \max_{c \ne c_A} \Pr(f(x+\eta) = c)$. Then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < \epsilon$, where $\epsilon = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$, and $\Phi(\cdot)$ is the cumulative distribution function of standard Gaussian distribution.*

Note that Theorem 5 holds for an *arbitrary* classifier. Thus a hardness result of random smoothing—the one in an opposite direction of Theorem 5—requires finding a hard instance of classifier f such that a similar conclusion of Theorem 5 does not hold, i.e., the resulting smoothed classifier g is trivial as the noise variance is too large. Our results in Theorems 3 and 4 are of this type. Beyond the top-1 predictions used in Theorem 5, Jia et al. (2020) studied certified robustness for top-k predictions via random smoothing under Gaussian noise and derive a tight robustness bound in ℓ_2 norm. In this paper, however, we study the standard setting of top-1 predictions.

ℓ_p robustness. Beyond the ℓ_2 robustness, random smoothing also achieves the state-of-the-art certified ℓ_p robustness for $p < 2$. Lee et al. (2019) provided adversarial robustness guarantees and associated random-smoothing algorithms for the discrete case where the adversary is ℓ_0 bounded. Li et al. (2019) suggested replacing Gaussian with Laplacian noise for the ℓ_1 robustness. Dvijotham et al. (2020) introduced a general framework for proving robustness properties of smoothed classifiers in the black-box setting using f-divergence. However, much remains unknown concerning the effectiveness of random smoothing for ℓ_p robustness with $p > 2$. Salman et al. (2019) proposed an algorithm for certifying ℓ_∞ robustness, by firstly certifying ℓ_2 robustness via the algorithm of Cohen et al. (2019) and then dividing the certified ℓ_2 radius by $\sqrt{d}$. However, the certified ℓ_∞ radius by this procedure is as small as $\mathcal{O}(1/\sqrt{d})$, in contrast to the constant certified radius as discussed in this paper.

Training algorithms. While random smoothing certifies inference-time robustness for any given base classifier f, the certified robust radius might vary a lot for different training methods. This motivates researchers to design new training algorithms for the base classifier f that are particularly well suited for use with random smoothing. Zhai et al. (2020) trained a robust smoothed classifier via maximizing the certified radius. In contrast to using naturally trained classifier in (Cohen et al., 2019), Salman et al. (2019) combined adversarial training of Madry et al. (2018) with random smoothing in the training procedure of f. In our experiments, we introduce a new baseline that combines TRADES (Zhang et al., 2019) with random smoothing to train a robust smoothed classifier.

Hardness of random smoothing. At the time of submission of this paper, we are aware of two contemporaneous independent works by Kumar et al. (2020) and Yang et al. (2020a) which study the similar problem of hardness of random smoothing. Kumar et al. (2020) showed that the noise magnitude must have a polynomial dependency on d for $p > 2$ if the noise distribution $\mathcal{D}$ is: (1) a product measure of i.i.d. noise from symmetric distribution with a continuous support; (2) generalized Gaussian distribution; (3) uniform distribution within a finite region. In contrast, our statements hold for arbitrary distributions. Yang et al. (2020a) showed that with only label statistics under random input perturbations, random smoothing cannot achieve nontrivial certified accuracy against perturbations of ℓ_p norm $\Omega(\min(1, d^{\frac{1}{p}-\frac{1}{2}}))$, while the dependency on other factors beyond dimension d was unclear. In this work, we provide more precise characterization on the ℓ_p norm of perturbations when random smoothing fails, which is $\Omega((\mathbb{E}\,\eta_i^2)^{\frac{1}{2}}\frac{\delta}{\sqrt{1-\delta}}d^{\frac{1}{p}-\frac{1}{2}})$ for at least 99% of the components η_i of noise η, where δ is the score gap between the highest-scored class and the runner-up.

3. Analysis of Main Results

In this section we prove Theorem 3 and Theorem 4.

3.1 Analysis of Theorem 3

In this section we prove Theorem 3. Our proof has two main steps: first, we study the one-dimensional version of the problem and prove two complementary lower bounds on the magnitude of a sample η drawn from a distribution $\mathcal{D}$ over $\mathbb{R}$ with the property that for all $v \in \mathbb{R}$ with $|v| \leq \epsilon$ we have $\mathrm{TV}(\mathcal{D}, \mathcal{D}+v) \leq \delta$. Next, we show how to apply this argument to $\Omega(d)$ orthogonal 1-dimensional subspaces in $\mathbb{R}^d$ to lower bound the expected magnitude of a sample drawn from a distribution $\mathcal{D}$ over $\mathbb{R}^d$, with the property that for all $v \in \mathbb{R}^d$ with $\|v\|_p \leq \epsilon$, we have $\mathrm{TV}(\mathcal{D}, \mathcal{D}+v) \leq \delta$.

One-dimensional results. Our first result lower bounds the magnitude of a sample from any distribution $\mathcal{D}$ in terms of the total variation distance between $\mathcal{D}$ and $\mathcal{D}+\epsilon$ for any $\epsilon \geq 0$.

Lemma 6 *Let $\mathcal{D}$ be any distribution on $\mathbb{R}$, η be a sample from $\mathcal{D}$, $\epsilon \geq 0$, and let $\delta = \mathrm{TV}(\mathcal{D}, \mathcal{D}+\epsilon)$. Then we have*[2]

$$\mathbb{E}\,|\eta|^2 \geq \frac{\epsilon^2}{200} \cdot \frac{1-\delta}{\delta^2}.$$

We prove Lemma 6 using two complementary lower bounds. The first lower bound is tighter for large δ, while the second lower bound is tighter when δ is close to zero. Taking the maximum of the two bounds proves Lemma 6.

Lemma 7 *Let $\mathcal{D}$ be any distribution on $\mathbb{R}$, η be a sample from $\mathcal{D}$, $\epsilon \geq 0$, and let $\delta = \mathrm{TV}(\mathcal{D}, \mathcal{D}+\epsilon)$. Then we have*

$$\mathbb{E}\,|\eta|^2 \geq \frac{\epsilon^2}{8} \cdot (1-\delta).$$

Proof Let $\eta' = \eta + \epsilon$ so that η' is a sample from $\mathcal{D}+\epsilon$ and define $r = \epsilon/2$ so that the sets $\mathcal{A} = (-r, r)$ and $\mathcal{B} = [\epsilon - r, \epsilon + r]$ are disjoint. From Chebyshev's inequality, we have that $\Pr(\eta \in \mathcal{A}) = 1 - \Pr(|\eta| \geq r) \geq 1 - \frac{\mathbb{E}\,|\eta|^2}{r^2}$. Further, since $\eta' \in \mathcal{B}$ if and only if $\eta \in \mathcal{A}$, we have $\Pr(\eta' \in \mathcal{B}) \geq 1 - \frac{\mathbb{E}\,|\eta|^2}{r^2}$. Next, since $\mathcal{A}$ and $\mathcal{B}$ are disjoint, it follows that $\Pr(\eta' \in \mathcal{A}) \leq 1 - \Pr(\eta' \in \mathcal{B}) \leq 1 - 1 + \frac{\mathbb{E}\,|\eta|^2}{r^2} = \frac{\mathbb{E}\,|\eta|^2}{r^2}$. Finally, we have $\delta \geq \Pr(\eta \in \mathcal{A}) - \Pr(\eta' \in \mathcal{A}) \geq 1 - \frac{2\,\mathbb{E}\,|\eta|^2}{r^2} = 1 - \frac{8\,\mathbb{E}\,|\eta|^2}{\epsilon^2}$. Rearranging this inequality proves the claim. ■

Next, we prove a tighter bound when δ is close to zero. The key insight is that no interval $I \subseteq \mathbb{R}$ of width ϵ can have probability mass larger than $\mathrm{TV}(\mathcal{D}, \mathcal{D}+\epsilon)$. This implies that the mass of $\mathcal{D}$ cannot concentrate too close to the origin, leading to lower bounds on the expected magnitude of a sample from $\mathcal{D}$.

Lemma 8 *Let $\mathcal{D}$ be any distribution on $\mathbb{R}$, η be a sample from $\mathcal{D}$, $\epsilon \geq 0$, and let $\delta = \mathrm{TV}(\mathcal{D}, \mathcal{D}+\epsilon)$. Then we have*

$$\mathbb{E}\,|\eta| \geq \frac{\epsilon}{8} \cdot \frac{(1-\delta)^2}{\delta},$$

which implies $\mathbb{E}\,|\eta|^2 \geq \frac{\epsilon^2}{64} \cdot \frac{(1-\delta)^4}{\delta^2}$.

2. We do not try to optimize constants throughout the paper.

Proof The key step in the proof is to show that every interval $\mathcal{I} = [a, a+\epsilon)$ of length ϵ has probability mass at most δ under the distribution $\mathcal{D}$. Once we have established this fact, then the proof is as follows: for any $\tau \geq 0$, rearranging Markov's inequality gives $\mathbb{E}\,|\eta| \geq \tau \Pr(|\eta| > \tau) = \tau(1 - \Pr(|\eta| \leq \tau))$. We can cover the set $\{x \in \mathbb{R} : |x| \leq \tau\}$ using $\lceil \frac{2\tau}{\epsilon} \rceil$ intervals of width ϵ and each of those intervals has probability mass at most δ. It follows that $\Pr(|\eta| \leq \tau) \leq \lceil \frac{2\tau}{\epsilon} \rceil \delta$, implying $\mathbb{E}\,|\eta| \geq \tau(1 - \lceil \frac{2\tau}{\epsilon} \rceil \delta)$. Since $\lceil \frac{2\tau}{\epsilon} \rceil \leq \frac{2\tau}{\epsilon} + 1$, we have $\mathbb{E}\,|\eta| \geq (1-\delta)\tau - \frac{2\delta}{\epsilon}\tau^2$. Finally, we optimize τ to get the strongest bound. The strongest bound is obtained at $\tau = \frac{\epsilon(1-\delta)}{4\delta}$, which gives $\mathbb{E}\,|\eta| \geq \frac{\epsilon(1-\delta)^2}{8\delta}$.

It remains to prove the claim that all intervals of length ϵ have probability mass at most δ. Let $\mathcal{I} = [a, a+\epsilon)$ be any such interval. The proof has two steps: first, we partition $\mathbb{R}$ using a collection of translated copies of the interval $\mathcal{I}$, and show that the difference in probability mass between any pair of intervals in the partition is at most δ. Then, given that there must be intervals with probability mass arbitrarily close to zero, this implies that the probability mass of any interval (and in particular, the probability mass of $\mathcal{I}$) is upper bounded by δ.

For each integer $i \in \mathbb{Z}$, let $\mathcal{I}_i = \mathcal{I} + i\epsilon = \{x + i\epsilon : x \in \mathcal{I}\}$ be a copy of the interval $\mathcal{I}$ translated by $i\epsilon$. By construction the set of intervals $\mathcal{I}_i$ for $i \in \mathbb{Z}$ forms a partition of $\mathbb{R}$. For any indices $i < j$, we can express the difference in probability mass between $\mathcal{I}_i$ and $\mathcal{I}_j$ as a telescoping sum: $\mathcal{D}(\mathcal{I}_j) - \mathcal{D}(\mathcal{I}_i) = \sum_{k=i}^{j-1} [\mathcal{D}(\mathcal{I}_{k+1}) - \mathcal{D}(\mathcal{I}_k)]$. We will show that for any $i < j$, the telescoping sum is contained in $[-\delta, \delta]$. Let $P = \{k \in (i, j] : \mathcal{D}(\mathcal{I}_{k+1}) - \mathcal{D}(\mathcal{I}_k) > 0\}$ be the indices of the positive terms in the sum. Then, since the telescoping sum is upper bounded by the sum of its positive terms and the intervals are disjoint, we have

$$\mathcal{D}(\mathcal{I}_j) - \mathcal{D}(\mathcal{I}_i) \leq \sum_{k \in P} [\mathcal{D}(\mathcal{I}_{k+1}) - \mathcal{D}(\mathcal{I}_k)] = \mathcal{D}\left(\bigcup_{k \in P} \mathcal{I}_{k+1}\right) - \mathcal{D}\left(\bigcup_{k \in P} \mathcal{I}_k\right).$$

For all $k \in P$ we have $\eta \in \mathcal{I}_k$ if and only if $\eta + \epsilon \in \mathcal{I}_{k+1}$, which implies $\Pr(\eta \in \bigcup_{k \in P} \mathcal{I}_k) = \Pr(\eta + \epsilon \in \bigcup_{k \in P} \mathcal{I}_{k+1})$. Combined with the definition of the total variation distance, it follows that

$$\begin{aligned}
\mathcal{D}\left(\bigcup_{k \in P} \mathcal{I}_{k+1}\right) - \mathcal{D}\left(\bigcup_{k \in P} \mathcal{I}_k\right) &= \Pr\left(\eta \in \bigcup_{k \in P} \mathcal{I}_{k+1}\right) - \Pr\left(\eta \in \bigcup_{k \in P} \mathcal{I}_k\right) \\
&= \Pr\left(\eta \in \bigcup_{k \in P} \mathcal{I}_{k+1}\right) - \Pr\left(\eta + \epsilon \in \bigcup_{k \in P} \mathcal{I}_{k+1}\right) \leq \delta,
\end{aligned}$$

and therefore $\mathcal{D}(\mathcal{I}_j) - \mathcal{D}(\mathcal{I}_i) \leq \delta$. A similar argument applied to the negative terms of the telescoping sum guarantees that $\mathcal{D}(\mathcal{I}_j) - \mathcal{D}(\mathcal{I}_i) \geq -\delta$, proving that $|\mathcal{D}(\mathcal{I}_j) - \mathcal{D}(\mathcal{I}_i)| \leq \delta$.

Finally, for any $\alpha > 0$, there must exist an interval $\mathcal{I}_j$ such that $\mathcal{D}(\mathcal{I}_j) < \alpha$ (since otherwise the total probability mass of all the intervals would be infinite). Since no pair of intervals in the partition can have probability masses differing by more than δ, this implies that $\mathcal{D}(\mathcal{I}) \leq \alpha + \delta$ for any α. Taking the limit as $\alpha \to 0$ shows that $\mathcal{D}(\mathcal{I}) \leq \delta$, completing the proof. ■

Finally, Lemma 6 follows from Lemmas 7 and 8, and the fact that for any $\delta \in (0, 1]$, we have $\max\{\frac{1-\delta}{8}, \frac{(1-\delta)^4}{64\delta^2}\} \geq \frac{1}{200} \cdot \frac{1-\delta}{\delta^2}$.

Extension to the d-dimensional case. For the remainder of this section we turn to the analysis of distributions $\mathcal{D}$ defined over $\mathbb{R}^d$. First, we use Lemma 6 to lower bound the magnitude of noise drawn from $\mathcal{D}$ when projected onto any one-dimensional subspace.

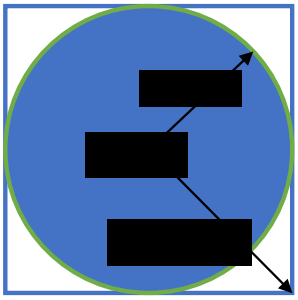

Figure 1: Vectors pointing towards the corner of the cube in $\mathbb{R}^d$ have large ℓ_2 norm but small ℓ_p norm.

Corollary 9 *Let $\mathcal{D}$ be any distribution on $\mathbb{R}^d$, η be a sample from $\mathcal{D}$, $v \in \mathbb{R}^d$, and let $\delta = \mathrm{TV}(\mathcal{D}, \mathcal{D}+v)$. Then we have*

$$\mathop{\mathbb{E}}_{\eta\sim\mathcal{D}} \frac{|v^\top \eta|^2}{\|v\|_2^2} \geq \frac{\|v\|_2^2}{200} \cdot \frac{1-\delta}{\delta^2}.$$

Proof Let η be a sample from $\mathcal{D}$, $\eta' = \eta + v$ be a sample from $\mathcal{D} + v$, and define $Z = v^\top \eta$ and $Z' = v^\top \eta' = Z + \|v\|_2^2$. Then the total variation distance between Z and Z' is bounded by δ, and Z' corresponds to a translation of Z by a distance $\|v\|_2^2$. Therefore, applying Lemma 6 with $\epsilon = \|v\|_2^2$, we have that $\mathbb{E}\,|v^\top \eta|^2 = \mathbb{E}\,|Z|^2 \geq \|v\|_2^4 \cdot \frac{1-\delta}{200\delta^2}$. Rearranging this inequality completes the proof. ■

Intuitively, Corollary 9 shows that for any vector $v \in \mathbb{R}^d$ such that $\mathrm{TV}(\mathcal{D}, \mathcal{D}+v)$ is small, the expected magnitude of a sample $\eta \sim \mathcal{D}$ when projected onto v cannot be much smaller than the length of v. The key idea for proving Theorem 3 is to construct a large number of orthogonal vectors $v_1, \ldots, v_b$ with small ℓ_p norms but large ℓ_2 norms. Then $\mathcal{D}$ will have to be "spread out" in all of these directions, resulting in a large expected ℓ_2 norm. We begin by showing that whenever d is a power of two, we can find an orthogonal basis for $\mathbb{R}^d$ in $\{\pm 1\}^d$.

Lemma 10 *For any $n \geq 0$ there exist $d = 2^n$ orthogonal vectors $v_1, \ldots, v_d \in \{\pm 1\}^d$.*

Proof The proof is by induction on n. For $n = 0$, we have $d = 1$ and the vector $v_1 = (1)$ satisfies the requirements. Now suppose the claim holds for n and let $v_1, \ldots, v_d$ be orthogonal in $\{\pm 1\}^d$ for $d = 2^n$. For each $i \in [d]$, define $a_i = (v_i, v_i) \in \{\pm 1\}^{2d}$ and $b_i = (v_i, -v_i) \in \{\pm 1\}^{2d}$. We will show that these vectors are orthogonal. For any indices i and j, we can compute the inner products between pairs of vectors among a_i, a_j, b_i, and b_j: $a_i^\top a_j = 2v_i^\top v_j$, $b_i^\top b_j = 2v_i^\top v_j$, and $a_i^\top b_j = v_i^\top v_j - v_i^\top v_j = 0$. Therefore, for any $i \neq j$, since $v_i^\top v_j = 0$, we are guaranteed that $a_i^\top a_j = 0$, $b_i^\top b_j = 0$, and $a_i^\top b_j = 0$. It follows that the 2^{d+1} vectors $a_1, \ldots, a_d, b_1, \ldots, b_d$ are orthogonal. ■

From this, it follows that for any dimension d, we can always find a collection of $b \geq d/2$ vectors that are short in the ℓ_p norm, but long in the ℓ_2 norm. Intuitively, these vectors are the vertices of a hypercube in a b-dimensional subspace. Figure 1 depicts the construction.

Corollary 11 *For any $p \geq 2$ and dimension d, there exist $b \geq d/2$ orthogonal vectors $v_1, \ldots, v_b \in \mathbb{R}^d$ such that $\|v_i\|_2 = b^{1/2-1/p} \geq (d/2)^{1/2-1/p}$ and $\|v_i\|_p = 1$ for all $i \in [b]$. This holds even when $p = \infty$.*

Proof Let n be the largest integer such that $2^n \leq d$. We must have $2^n > d/2$, since otherwise $2^{n+1} \leq d$. We now apply Lemma 10 to find $b = 2^n$ orthogonal vectors $u_1, \ldots, u_b \in \{\pm 1\}^b$. For each $i \in [b]$, we have that $\|u_i\|_p = b^{1/p}$. Finally, for $i \in [b]$, define $v_i = (u_i \cdot b^{-1/p}, 0, \ldots, 0) \in \mathbb{R}^d$ to be a normalized copy of u_i padded with $d - b$ zeros. For all $i \in [b]$, we have $\|v_i\|_p = 1$ and $\|v_i\|_2 = (b \cdot b^{-2/p})^{1/2} = b^{1/2-1/p} \geq (d/2)^{1/2-1/p}$. ■

With this, we are ready to prove Theorem 3.

Theorem 3 *Fix any $p \geq 2$ and let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$ such that there exists a radius ϵ and total variation bound δ satisfying that, for all $v \in \mathbb{R}^d$ with $\|v\|_p \leq \epsilon$, we have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \leq \delta$. Then*

$$\mathop{\mathbb{E}}_{\eta \sim \mathcal{D}} \|\eta\|_2^2 \geq \frac{\epsilon^2 d^{2-2/p}}{800} \cdot \frac{1-\delta}{\delta^2}.$$

Proof Let η be a sample from $\mathcal{D}$. By scaling the vectors from Corollary 11 by ϵ, we obtain $b > d/2$ vectors $v_1, \ldots, v_b \in \mathbb{R}^d$ with $\|v_i\|_p = \epsilon$ and $\|v_i\|_2 = \epsilon \cdot b^{1/2-1/p}$. By assumption we must have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v_i) \leq \delta$, since $\|v_i\|_p \leq \epsilon$, and Corollary 9 implies that $\mathbb{E} \frac{|v_i^\top \eta|^2}{\|v_i\|_2^2} \geq \frac{\|v_i\|_2^2}{200} \frac{1-\delta}{\delta^2}$ for each i. We use this fact to bound $\mathbb{E} \|\eta\|_2^2$.

Let $\mathbf{Q} \in \mathbb{R}^{b \times d}$ be the matrix whose i^{th} row is given by $v_i / \|v_i\|_2$ so that $\mathbf{Q}$ is the orthogonal projection matrix onto the subspace spanned by the vectors $v_1, \ldots, v_b$. Then we have $\mathbb{E} \|\eta\|_2^2 \geq \mathbb{E} \|\mathbf{Q}\eta\|_2^2 = \sum_{i=1}^b \mathbb{E} \frac{|v_i^\top \eta|^2}{\|v_i\|_2^2} \geq \sum_{i=1}^b \frac{\|v_i\|_2^2}{200} \frac{1-\delta}{\delta^2}$, where the first inequality follows because orthogonal projections are non-expansive, the equality follows from the Pythagorean theorem, and the last inequality follows from Corollary 9. Using the fact that $\|v_i\|_2 = \epsilon \cdot b^{1/2-1/p}$, we have that $\mathbb{E} \|\eta\|_2^2 \geq \frac{\epsilon^2 b^{2-2/p}}{200} \cdot \frac{1-\delta}{\delta^2}$. Finally, since $b > d/2$ and $(1/2)^{2-2/p} \geq 1/4$ for $p \geq 2$, we have $\mathbb{E} \|\eta\|_2^2 \geq \frac{\epsilon^2 d^{2-2/p}}{800} \cdot \frac{1-\delta}{\delta^2}$, as required. ■

Finally, Lemma 2 provides the main connection between random smoothing guarantees and total variation distance which is used to anchor the above-mentioned arguments.

3.2 Analysis of Theorem 4

In this section we prove the variance and heavy-tailed properties from Theorem 4 separately.

Combining Theorem 3 with a peeling argument, we are able to lower bound the marginal variance in most of the coordinates of η.

Lemma 12 *Fix any $p \geq 2$ and let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$ such that there exists a radius ϵ and total variation bound δ so that for all $v \in \mathbb{R}^d$ with $\|v\|_p \leq \epsilon$ we have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \leq \delta$. Let η be a sample from $\mathcal{D}$ and σ be the permutation of $[d]$ such that $\mathbb{E}[\eta_{\sigma(1)}^2] \geq \cdots \geq \mathbb{E}[\eta_{\sigma(d)}^2]$. Then for any $i \in [d]$, we have $\mathbb{E}[\eta_{\sigma(i)}^2] \geq \frac{\epsilon^2 (d-i+1)^{1-2/p}}{800} \frac{1-\delta}{\delta^2}$.*

Proof For each index i, let $P_i : \mathbb{R}^d \to \mathbb{R}^{d-i+1}$ be the projection $P_i(x) = (x_{\sigma(i)}, x_{\sigma(i+1)}, \ldots, x_{\sigma(d)})$ and $\mathcal{D}_i$ be the distribution of $P_i(\eta)$. First we argue that for each $i \in [d]$ and any $v \in \mathbb{R}^{d-i+1}$ with $\|v\|_p \leq \epsilon$, we must have $\mathrm{TV}(\mathcal{D}_i, \mathcal{D}_i + v) \leq \delta$. To see this, let $z \in \mathbb{R}^d$ be the vector such that $P_i(z) = v$ and $z_{\sigma(1)} = \cdots = z_{\sigma(i-1)} = 0$. Then $\mathrm{TV}(\mathcal{D}_i, \mathcal{D}_i + v) = \sup_{\mathcal{A} \subseteq \mathbb{R}^{d-i+1}} |\Pr(P_i(\eta) \in \mathcal{A}) - \Pr(P_i(\eta) + v \in \mathcal{A})| = \sup_{\mathcal{A} \subseteq \mathbb{R}^{d-i+1}} |\Pr(P_i(\eta) \in \mathcal{A}) - \Pr(P_i(\eta + z) \in \mathcal{A})| \leq \sup_{\mathcal{A} \subseteq \mathbb{R}^d} |\Pr(\eta \in \mathcal{A}) - \Pr(\eta + z \in \mathcal{A})| = \mathrm{TV}(\mathcal{D}, \mathcal{D} + z)$. Next, since $\|z\|_p = \|v\|_p \leq \epsilon$, we must have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \leq \delta$.

Now fix an index $i \in [d]$ and let Z be a sample from $\mathcal{D}_i$. Applying Theorem 3 to Z, we have that $\mathbb{E}\,\|Z\|_2^2 \geq \frac{\epsilon^2 (d-i+1)^{2-2/p}}{800} \cdot \frac{1-\delta}{\delta^2}$. Since there must exist at least one index l such that $\mathbb{E}[Z_l^2] \geq \frac{1}{d-i+1} \sum_{j=1}^{d-i+1} \mathbb{E}[Z_j^2]$, it follows that at least one coordinate l must satisfy $\mathbb{E}[Z_l^2] \geq \frac{\epsilon^2 (d-i+1)^{1-2/p}}{800} \cdot \frac{1-\delta}{\delta^2}$. Finally, since the coordinates of Z are the $(d-i+1)$ coordinates of η with the smallest variance, it follows that $\mathbb{E}[\eta_{\sigma(i)}^2] \geq \frac{\epsilon^2 (d-i+1)^{1-2/p}}{800} \cdot \frac{1-\delta}{\delta^2}$, as required. ■

Lemma 12 implies that any distribution $\mathcal{D}$ over $\mathbb{R}^d$ such that for all $v \in \mathbb{R}^d$ with $\|v\|_p \leq \epsilon$ we have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \leq \delta$ for $p > 2$ must have high marginal variance in most of its coordinates. In particular, for any constant $c \in [0, 1]$, the top c-fraction of coordinates must have marginal variance at least $\Omega(d^{1-2/p} \epsilon^2 \frac{1-\delta}{\delta^2})$. For $p > 2$, this bound grows with the dimension d. Our next lemma shows that when $\mathcal{D}$ is a product measure of d i.i.d. one-dimension distribution $\mathcal{D}'$ in the standard coordinate, the distribution $\mathcal{D}'$ must be heavy-tailed. The lemma is built upon a fact that $\mathbb{E}\|\eta\|_2 \geq \Omega(\epsilon d^{1-1/p} \frac{1-\delta}{\delta})$, with a similar analysis as that of Theorem 3. We defer the proof of this fact to the appendix (see Lemma 24). Note that the fact implies that $\mathbb{E}\|\eta\|_\infty \geq \Omega(\epsilon d^{1/2-1/p} \frac{1-\delta}{\delta})$ by the equivalence between ℓ_2 and ℓ_∞ norms. We then have the following lemma.

Lemma 13 *Let $h(\delta) = \frac{1-\delta}{\delta}$ and $p > 2$. Let $X_1, ..., X_d$ be d random variables in $\mathbb{R}$ sampled i.i.d. from distribution $\mathcal{D}'$. Then "$\mathbb{E} \max_{i \in [d]} |X_i| \geq C d^{1/2-1/p} \epsilon h(\delta)$" implies "$\Pr_{X \sim \mathcal{D}'}[|X| > x] > \left(\frac{c \epsilon h(\delta)}{x}\right)^{2p/(p-2)}$ for some $x > c \epsilon h(\delta)$ with an absolute constant $c > 0$", that is, in sufficiently high dimensions, $\mathcal{D}'$ is a heavy-tailed distribution.*

Proof Denote by $G(x) = \Pr_{X \sim \mathcal{D}'}[|X| > x]$ the complementary Cumulative Distribution Function (CDF) of $\mathcal{D}'$. We only need to show that "$G(x) \leq \left(\frac{\epsilon h(\delta)}{24x}\right)^{2p/(p-2)}$ for all $x > \frac{\epsilon h(\delta)}{24}$" implies "$\mathbb{E} \max_{i \in [d]} |X_i| < C d^{1/2-1/p} \epsilon h(\delta)$ for a constant $C > 0$". We note that

$$
\begin{aligned}
\mathbb{E} \max_{i \in [d]} |X_i| &= \int_0^\infty \Pr_{X_i \sim \mathcal{D}} \left[\max_{i \in [d]} |X_i| > x \right] dx \\
&= \int_0^{\frac{\epsilon h(\delta)}{24}} \Pr_{X_i \sim \mathcal{D}} \left[\max_{i \in [d]} |X_i| > x \right] dx + \int_{\frac{\epsilon h(\delta)}{24}}^\infty [1 - (1 - G(x))^d] dx \\
&\leq \frac{\epsilon h(\delta)}{24} + \frac{\epsilon h(\delta)}{24} \int_1^\infty \left[1 - \left(1 - \frac{1}{t^{2p/(p-2)}} \right)^d \right] dt \\
&= \frac{\Gamma(\frac{p+2}{2p}) \epsilon h(\delta)}{24} \frac{\Gamma(d+1)}{\Gamma(d + \frac{p+2}{2p})} \sim d^{1/2-1/p} \epsilon h(\delta),
\end{aligned}
$$

where the second equality holds because for any i.i.d. $Y_i \sim \mathcal{D}$ with CDF $F(x)$, the CDF of $\max_{i\in[d]} Y_i$ is given by $(1-F(x))^d$, the first inequality holds by the change of variable, and the last $\sim$ relation holds because $\frac{\Gamma(d+1)}{\Gamma(d+\frac{p+2}{2p})} \sim d^{1/2-1/p}$. ■

Combining Lemmas 12 and 13 with Lemma 2 and the fact that $\mathbb{E}\|\eta\|_\infty \geq \Omega(\epsilon d^{1/2-1/p}\frac{1-\delta}{\delta})$ completes the proof of Theorem 4.

4. Experiments

In this section, we evaluate the certified ℓ_∞ robustness and verify the tightness of our lower bounds by numerical experiments. Experiments run with two NVIDIA GeForce RTX 2080 Ti GPUs. We release our code and trained models at `https://github.com/hongyanz/TRADES-smoothing`.

4.1 Certified ℓ_∞ Robustness

Despite the hardness results of random smoothing on certifying ℓ_∞ robustness with large perturbation radius, we evaluate the certified ℓ_∞ robust accuracy of random smoothing on the CIFAR-10 data set when the perturbation radius is as small as 2/255, given that the data dimension $32 \times 32 \times 3$ is not too high relative to the 2-pixel attack. The goal of this experiment is to show that random smoothing based methods might be hard to achieve very promising robust accuracy (e.g., $\geq 70\%$) even when the perturbation radius is as small as 2 pixels.

Experimental setup. Our experiments exactly follow the setups of (Salman et al., 2019). Specifically, we train the models on the CIFAR-10 training set and test it on the CIFAR-10 test sets. We apply the ResNet-110 architecture (He et al., 2016) for the CIFAR-10 classification task. The output size of the last layer is 10. Our training procedure is a modification of (Salman et al., 2019): Salman et al. (2019) used adversarial training of Madry et al. (2018) to train a soft-random-smoothing classifier by injecting Gaussian noise. In our training procedure, we replace the adversarial training with TRADES (Zhang et al., 2019), a state-of-the-art defense model which won the first place in the NeurIPS 2018 Adversarial Vision Challenge (Brendel et al., 2020). In particular, we minimize the empirical risk of the following loss:

$$\min_f \mathbb{E}_{X,Y}\mathbb{E}_{\eta\sim\mathcal{N}(0,\sigma^2 I)}\Big[\mathcal{L}(f(X+\eta),Y) + \beta \max_{X'\in\mathbb{B}_2(X,\epsilon)} \mathcal{L}(f(X+\eta), f(X'+\eta))\Big],$$

where η is the injected Gaussian noise, $\mathcal{L}$ is the cross-entropy loss or KL divergence, (X,Y) is the clean data with label, and f is a neural network classifier which outputs the logits of an instance. For a fixed f, the inner maximization problem is solved by PGD iterations, and we update the parameters in the outer minimization and inner maximization problems alternatively. In our training procedure, we set ℓ_2 perturbation radius $\epsilon = 0.435$, perturbation step size 0.007, number of PGD iterations 10, regularization parameter $\beta = 6.0$, initial learning rate 0.1, standard deviation of injected Gaussian noise 0.12, batch size 256, and run 55 epochs on the training data set. We decay the learning rate by a factor of 0.1 at epoch 50. We use random smoothing of Cohen et al. (2019) to certify ℓ_2 robustness of the base classifier. We obtain the ℓ_∞ certified radius by scaling the ℓ_2 robust radius by a factor of $1/\sqrt{d}$. For fairness, we do not compare with the models using extra unlabeled data, ImageNet pretraining, or ensembling tricks.

Method	Certified Robust Accuracy	Natural Accuracy
TRADES + Random Smoothing	62.6%	78.8%
Salman et al. (2019)	60.8%	82.1%
Zhang et al. (2020)	54.0%	72.0%
Wong et al. (2018)	53.9%	68.3%
Mirman et al. (2018)	52.2%	62.0%
Gowal et al. (2018)	50.0%	70.2%
Xiao et al. (2019)	45.9%	61.1%

Table 1: Certified ℓ_∞ robustness at a radius of 2/255 on the CIFAR-10 data set (without extra unlabelled data or pre-trained model).

Experimental results. We compare TRADES + random smoothing with various baseline methods of certified ℓ_∞ robustness with radius $2/255$. We summarize our results in Table 4.1. All results are reported according to the numbers in their original papers.[3] It shows that TRADES with random smoothing achieves state-of-the-art performance on certifying ℓ_∞ robustness at radius 2/255 and enjoys higher robust accuracy than other methods. However, for all approaches, there are still significant gaps between the robust accuracy and the desired accuracy that is acceptable in the security-critical tasks (e.g., robust accuracy $\geq 70\%$), even when the certified radius is chosen as small as 2 pixels.

4.2 Effectiveness of Lower Bounds

For random smoothing, Theorem 4 suggests that the certified ℓ_∞ robust radius ϵ be (at least) proportional to $\sigma/\sqrt{d}$, where σ is the standard deviation of injected noise. In this section, we verify this dependency by numerical experiments on the CIFAR-10 data set and Gaussian noise.

Experimental setups. We apply the ResNet-110 architecture (He et al., 2016) for classification.[4] The output size of the last layer is 10. We vary the size of the input images with $32 \times 32 \times 3$, $48 \times 48 \times 3$, and $64 \times 64 \times 3$ by calling the *resize* function. We keep the quantity $\sigma/(\sqrt{d}\epsilon)$ as an absolute constant by setting the standard deviation σ as 0.12, 0.18, and 0.24, and the ℓ_2 perturbation radius as 0.435, 0.6525, and 0.87 in the TRADES training procedure for the three input sizes, respectively. Our goal is to show that the accuracy curves of the three input sizes behave similarly. In our training procedure, we set perturbation step size 0.007, number of perturbation iterations 10, regularization parameter $\beta = 6.0$, learning rate 0.1, batch size 256, and run 55 epochs on the training data set. We use random smoothing (Cohen et al., 2019) with varying σ's to certify the ℓ_2 robustness. The ℓ_∞ certified radius is obtained by scaling the ℓ_2 robust radius by a factor of $1/\sqrt{d}$.

3. We report the performance of (Salman et al., 2019) according to the results: `https://github.com/Hadisalman/smoothing-adversarial/blob/master/data/certify/best_models/cifar10/ours/cifar10/DDN_4steps_multiNoiseSamples/4-multitrain/eps_255/cifar10/resnet110/noise_0.12/test/sigma_0.12`, which is the best result in the folder "best models" by Salman et al. (2019). When a method was not tested under the 2/255 threat model in its original paper, we will not compare with it as well in our experiment.
4. The input size of the architecture is adaptive by applying the adaptive pooling layer.

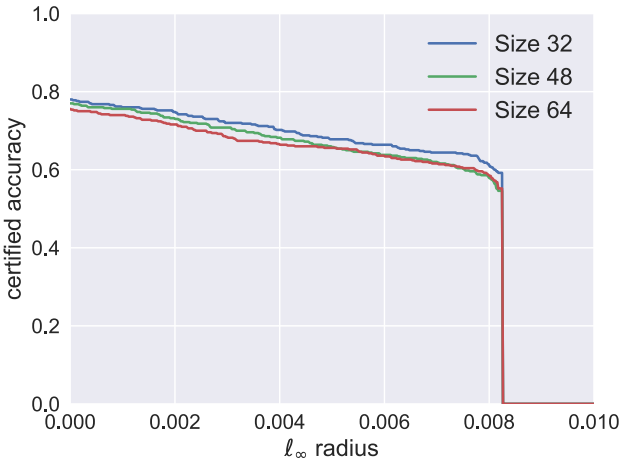

Figure 2: Certified accuracy of ResNet-110 models under varying input sizes by random smoothing.

We summarize our results in Figure 2. We observe that the three curves of varying input sizes behave similarly. This empirically supports our theoretical finding in Theorem 4 that the certified ℓ_∞ robust radius ϵ should be proportional to the quantity $\sigma/\sqrt{d}$. In Figure 2, the certified accuracy is monotonously decreasing until reaching some point where it plummets to zero. The phenomenon has also been observed by Cohen et al. (2019) and was explained by a hard upper limit to the radius we can certify, which is achieved when all samples are classified by f as the same class.

5. Conclusions

In this paper, we show a hardness result of random smoothing on certifying adversarial robustness against attacks in the ℓ_p ball of radius ϵ when $p > 2$. We focus on a lower bound on the necessary noise magnitude: *any* noise distribution $\mathcal{D}$ over $\mathbb{R}^d$ that provides ℓ_p robustness with $p > 2$ for all base classifiers must satisfy $\mathbb{E}\,\eta_i^2 = \Omega(d^{1-2/p}\epsilon^2(1-\delta)/\delta^2)$ for 99% of the features (pixels) of vector η drawn from $\mathcal{D}$, where δ is the score gap between the highest-scored class and the runner up in the framework of random smoothing. For high-dimensional images where the pixels are bounded in $[0, 255]$, the required noise will eventually dominate the useful information in the images, leading to trivial smoothed classifiers.

The proof roadmap of our results shows that defending against adversarial attacks in the ℓ_p ball of radius ϵ is almost as hard as defending against attacks in the ℓ_2 ball of radius $\epsilon d^{1/2-1/p}$, for random smoothing. We thus suggest combining random smoothing with dimensionality reduction techniques, such as principal component analysis or auto-encoder, to circumvent our hardness results, which is left open as future works. Another related open question is whether one can improve our lower bounds, or show that the bounds are tight.

Acknowledgments

This work was supported in part by the National Science Foundation under grant CCF-1815011 and by the Defense Advanced Research Projects Agency under cooperative agreement HR00112020003. The

views expressed in this work do not necessarily reflect the position or the policy of the Government and no official endorsement should be inferred. Approved for public release; distribution is unlimited.

Appendix A. Total-Variation based Robustness

First, we argue that for any points x and x', we must have $g(x) = g(x')$ whenever the total variation distance between $\mathcal{D} + x$ and $\mathcal{D} + x'$ is sufficiently small compared to the gap $\Delta(x)$, where $\mathcal{D} + x$ denotes the distribution of $\eta + x$ with $\eta \sim \mathcal{D}$.

Lemma 14 *For any distribution $\mathcal{D}$ on $\mathbb{R}^d$, base classifier $f : \mathbb{R}^d \to \mathcal{Y}$, and pair of points $x, x' \in \mathbb{R}^d$, if $\Delta(x) > 2\,\mathrm{TV}(\mathcal{D} + x, \mathcal{D} + x')$, then we have $g(x) = g(x')$.*

Proof To simplify notation, let $\delta = \mathrm{TV}(\mathcal{D}+x, \mathcal{D}+x')$ and let η be a sample from $\mathcal{D}$ so that $\eta+x$ is a sample from $\mathcal{D}+x$ and $\eta+x'$ is a sample from $\mathcal{D}+x'$. By the definition of the total variation distance, for any class $y \in \mathcal{Y}$, we have $\delta \geq \big|\Pr\big(f(x+\eta) = y\big) - \Pr\big(f(x'+\eta) = y\big)\big| = |G_y(x) - G_y(x')|$. Now let $y = g(x)$ and $y' \neq y$. Then we have

$$\begin{aligned} G_y(x') &\geq G_y(x) - \delta \\ &\geq G_{y'}(x) - \delta + \Delta(x) \\ &\geq G_{y'}(x') - 2\delta + \Delta(x). \end{aligned}$$

Whenever $\Delta(x) > 2\delta$, we are guaranteed that $G_y(x') > G_{y'}(x')$ for all y', and it follows that $g(x') = y = g(x)$. ■

As a consequence of Lemma 14, we can provide certified robustness guarantees for the smoothed classifier g in terms of balls defined by the total variation distance. In particular, for any $x \in \mathbb{R}^d$ and any $\delta \in (0, 1]$, define

$$\mathcal{B}_{\mathrm{TV}}(x, \delta; \mathcal{D}) = \{x' \in \mathbb{R}^d : \mathrm{TV}(\mathcal{D} + x, \mathcal{D} + x') < \delta\}$$

to be the set of points x' around x such that the distributions $\mathcal{D} + x$ and $\mathcal{D} + x'$ have total variation distance at most δ. When the distribution $\mathcal{D}$ is clear from context, we will simply write $\mathcal{B}_{\mathrm{TV}}(x, \delta)$.

Corollary 15 *For any distribution $\mathcal{D}$, base classifier f, and $x \in \mathbb{R}^d$ we have $g(x') = g(x)$ for all $x' \in \mathcal{B}_{TV}(x, \Delta(x)/2)$.*

Note that the ball $\mathcal{B}_{\mathrm{TV}}(x, \delta)$ is translation invariant (i.e., for any center $x \in \mathbb{R}^d$, we have $\mathcal{B}_{\mathrm{TV}}(x, \delta) = \mathcal{B}_{\mathrm{TV}}(0, \delta) + x$) and the definition of the ball only depends on the distribution $\mathcal{D}$. Therefore, if we can relate the balls for a given distribution $\mathcal{D}$ to those of a norm $\|\cdot\|_p$, then Corollary 15 implies robustness with respect to that norm. Let $\mathcal{B}_p(x, r) = \{x' \in \mathbb{R}^d : \|x' - x\|_p < r\}$ denote the ℓ_p ball of radius r centered at x.

Corollary 16 *Fix any $p > 0$, radius $\epsilon \geq 0$, distribution $\mathcal{D}$, and let $\delta \in [0, 1]$ be the smallest total variation bound such that $\mathcal{B}_p(0, r) \subseteq \mathcal{B}_{TV}(0, \delta)$. For any base classifier $f : \mathbb{R}^d \to \mathcal{Y}$ and any point $x \in \mathbb{R}^d$ with $\Delta(x) > 2\delta$, for all $x' \in \mathcal{B}_p(x, r)$ we have $g(x') = g(x)$.*

The following lemma is in an opposite direction as Corollary 16.

Lemma 17 *Let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$ such that for every (randomized) classifier $f : \mathbb{R}^d \to \mathcal{Y}$, the smoothed classifier $g(\cdot; \mathcal{D}, f)$ is $(\mathcal{A}, \delta)$-robust. Then for all $v \in \mathcal{A}$, we have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \leq \delta$.*

Proof Suppose there exists a vector $v \in \mathcal{A}$ such that $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) > \delta$. We show that this implies there is a randomized binary classifier $f : \mathbb{R}^d \to \mathcal{Y}$, such that $g(\cdot; \mathcal{D}, f)$ is not $(\mathcal{A}, \delta)$-robust. It follows that if $g(\cdot; \mathcal{D}, f)$ is $(\mathcal{A}, \delta)$-robust for all randomized classifiers f, then we must have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \leq \delta$ for all $v \in \mathcal{A}$.

Fix any $v \in \mathcal{A}$ such that $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) > \delta$ and let $\mathcal{D}' = \mathcal{D} + v$ be shorthand notation for the translated distribution. Since $\delta < \mathrm{TV}(\mathcal{D}, \mathcal{D}') = \sup_{\mathcal{S} \subseteq \mathbb{R}^d} \mathcal{D}(\mathcal{S}) - \mathcal{D}'(\mathcal{S})$, there exists a set $\mathcal{S} \subseteq \mathbb{R}^d$ so that $\mathcal{D}(\mathcal{S}) - \mathcal{D}'(\mathcal{S}) > \delta$. Let $\mathcal{S}^c = \{x \in \mathbb{R}^d : x \notin \mathcal{S}\}$ denote the complement of $\mathcal{S}$. We assume without loss of generality that $\mathcal{Y} = \{0, 1\}$ and define the randomized classifier f to take value 1 with probability α and 0 with probability $1 - \alpha$ for all $x \in S$ and to take value 1 with probability β and 0 with probability $1 - \beta$ for all $x \in \mathcal{S}^c$, where $\alpha = 1 - \frac{\mathcal{D}'(\mathcal{S})}{2}$ and $\beta = \frac{1}{2} - \frac{\mathcal{D}'(\mathcal{S})}{2}$. That is,

$$f(x) = \begin{cases} 1 \text{ w.p. } \alpha \text{ and } 0 \text{ otherwise}, & \text{if } x \in \mathcal{S}; \\ 1 \text{ w.p. } \beta \text{ and } 0 \text{ otherwise}, & \text{if } x \in \mathcal{S}^c. \end{cases}$$

Note that since $\mathcal{D}'(\mathcal{S}) \in [0, 1]$ we have that $\alpha \in [\frac{1}{2}, 1]$ and $\beta \in [0, \frac{1}{2}]$ are both valid probabilities. For any distribution P over $\mathbb{R}^d$ we have

$$\begin{aligned} \Pr_{Z \sim P}(f(Z) = 1) &= \Pr_{Z \sim P}\big(f(Z) = 1 \mid Z \in \mathcal{S}\big) \cdot P(\mathcal{S}) + \Pr_{Z \sim P}\big(f(Z) = 1 \mid Z \in \mathcal{S}^c\big) \cdot P(\mathcal{S}^c) \\ &= \alpha P(\mathcal{S}) + \beta P(\mathcal{S}^c) \\ &= P(\mathcal{S}) \cdot (\alpha - \beta) + \beta \\ &= \frac{1}{2} + \frac{P(\mathcal{S}) - \mathcal{D}'(\mathcal{S})}{2}. \end{aligned}$$

Therefore, we have $G_1(0; \mathcal{D}, f) = \Pr_{Z \sim \mathcal{D}}\big(f(Z) = 1\big) = \frac{1}{2} + \frac{\mathcal{D}(\mathcal{S}) - \mathcal{D}'(\mathcal{S})}{2} > \frac{1}{2} + \frac{\delta}{2}$. Similarly, we have that $G_1(v; \mathcal{D}, f) = \Pr_{Z \sim \mathcal{D}}\big(f(Z + v) = 1\big) = \Pr_{Z \sim \mathcal{D}'}\big(f(Z) = 1\big) = \frac{1}{2}$. It follows that $g(0; \mathcal{D}, f) = 1$, $\Delta(0; \mathcal{D}, f) = 2G_1(0; \mathcal{D}, f) - 1 > 2(\frac{1}{2} + \frac{\delta}{2}) - 1 = \delta$, and $g(v; \mathcal{D}, f) = 0$ (since $G_1(v; \mathcal{D}, f) = G_0(v; \mathcal{D}, f) = \frac{1}{2}$ and the ties are broken lexicographically). It follows that $g(\cdot; \mathcal{D}, f)$ is not robust to the adversarial translation $v \in \mathcal{A}$, as required. ■

Appendix B. Total Variation Bounds for Specific Distributions

In this section, we provide the total variation bounds for isotropic Gaussian and uniform distributions.

B.1 Isotropic Gaussian

In this section we give bounds for the total variation distance between shifted copies of Gaussian distributions with an isotropic covariance matrix. Our results are derived from the following theorem due to Devroye et al. (2018).

Theorem 18 (Theorem 1.2 of Devroye et al. (2018)) *Suppose $d > 1$, let $\mu_1 \neq \mu_2 \in \mathbb{R}^d$ and let Σ_1, Σ_2 be positive definite $d \times d$ matrices. Let $v = \mu_1 - \mu_2$ and let Π be a $d \times d - 1$ matrix whose*

columns form a basis for the subspace orthogonal to v. Define the function

$$tv(\mu_1, \Sigma_1, \mu_2, \Sigma_2) = \max\left\{ \left\|(\Pi^\top \Sigma_1 \Pi)^{-1} \Pi^\top \Sigma_2 \Pi - I_{d-1}\right\|_F, \frac{|v^\top(\Sigma_1 - \Sigma_2)v|}{v^\top \Sigma_1 v}, \frac{v^\top v}{\sqrt{v^\top \Sigma_1 v}} \right\},$$

where $\|\cdot\|_F$ denotes the Frobenius norm and I_{d-1} is the $(d-1)$-dimensional identity matrix. Then we have

$$\frac{1}{200} \leq \frac{\mathrm{TV}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2))}{\min\{1, tv(\mu_1, \Sigma_1, \mu_2, \Sigma_2)\}} \leq \frac{9}{2}.$$

Theorem 18 takes a simpler form when $\Sigma_1 = \Sigma_2 = \sigma^2 I$ and $\mu_1 = 0$ because then the first and last terms in the max of $tv(\mu_1, \Sigma_1, \mu_2, \Sigma_2)$ are zero, giving the following:

Corollary 19 *Suppose $d > 1$ and let $v \in \mathbb{R}^d$ and $\sigma > 0$. Then*

$$\frac{1}{200} \cdot \min\left\{1, \frac{\|v\|_2}{\sigma}\right\} \leq \mathrm{TV}\big(\mathcal{N}(0, \sigma^2 I), \mathcal{N}(v, \sigma^2 I)\big) \leq \frac{9}{2} \cdot \min\left\{1, \frac{\|v\|_2}{\sigma}\right\}.$$

We can use this result to show that the variance bounds given by Lemma 12 are nearly tight, except for the dependence on the total variation bound, δ.

Corollary 20 *Fix any d, radius $\epsilon > 0$, total variation bound $\delta \in [0, 1]$, and $p \geq 2$. Setting $\sigma = \frac{9}{2}\frac{\epsilon}{\delta} d^{1/2-1/p}$ guarantees that for all $v \in \mathbb{R}^d$ with $\|v\|_p \leq \epsilon$ we have $\mathrm{TV}(\mathcal{N}(0, \sigma^2 I), \mathcal{N}(v, \sigma^2 I)) \leq \delta$. Moreover, if $\eta \sim \mathcal{N}(0, \sigma^2 I)$ then $\mathbb{E}[\eta_i^2] = \sigma^2 = (\frac{9}{2})^2 \cdot \frac{\epsilon^2}{\delta^2} \cdot d^{1-2/p}$ and $\mathbb{E}[\|\eta\|_2] \leq \frac{9}{2}\frac{\epsilon}{\delta} \cdot d^{1-1/p}$.*

Proof Since for every $v \in \mathbb{R}^d$ with $\|v\|_p \leq \epsilon$ we have $\|v\|_2 \leq \epsilon \cdot d^{1/2-1/p}$, it is sufficient to choose σ as in the statement. To bound $\mathbb{E}[\|\eta\|_2]$, we use Jensen's inequality: $\mathbb{E}[\|\eta\|_2] \leq \sqrt{\mathbb{E}[\|\eta\|_2^2]} = \sqrt{d}\sigma = \frac{9}{2}\frac{\epsilon}{\delta} d^{1-1/p}$. ■

B.2 Uniform Distribution on $\mathcal{B}_\infty(0, r)$

In this section, let $\mathcal{U}_r$ denote the uniform distribution on $\mathcal{B}_\infty(0, r)$ with density $p_r(x) = \frac{1}{(2r)^d}\mathbb{I}\{x \in \mathcal{B}_\infty(0, r)\}$.

Lemma 21 *For any dimension d, any vector $v \in \mathbb{R}^d$, and any radius $r \geq 0$, we have*

$$\mathrm{TV}(\mathcal{U}_r, \mathcal{U}_r + v) = 1 - \prod_{i=1}^{d} \max\left\{0, 1 - \frac{|v_i|}{2r}\right\}.$$

Proof To simplify notation, let $\mathcal{A} = \mathcal{B}_\infty(0, r)$ and $\mathcal{B} = \mathcal{B}_\infty(v, r)$. Since $\mathcal{U}_r$ has a density function, we can write the total variation distance as

$$\begin{aligned}
\mathrm{TV}(\mathcal{U}_r, \mathcal{U}_r + v) &= \frac{1}{2}\int_{\mathbb{R}^d} |p_r(x) - p_r(x - v)|\, dx \\
&= \frac{1}{2}(2r)^{-d} \int_{\mathbb{R}^d} |\,\mathbb{I}\{x \in \mathcal{A}\} - \mathbb{I}\{x \in \mathcal{B}\}|\, dx \\
&= \frac{1}{2}(2r)^{-d}\,\mathrm{Vol}(\mathcal{A} \triangle \mathcal{B}),
\end{aligned}$$

where $\mathcal{A} \triangle \mathcal{B}$ denotes the symmetric difference of $\mathcal{A}$ and $\mathcal{B}$. Since $\mathrm{Vol}(\mathcal{A} \triangle \mathcal{B}) = \mathrm{Vol}(\mathcal{A}) + \mathrm{Vol}(\mathcal{B}) - 2\,\mathrm{Vol}(\mathcal{A} \cap \mathcal{B})$, it is sufficient to calculate the volume of $\mathcal{A} \cap \mathcal{B}$. The intersection is a hyper-rectangle with side length $\max\{0, 2r - |v_i|\}$ in dimension i. Therefore, the volume of the intersection is given by $\mathrm{Vol}\big(\mathcal{A} \cap \mathcal{B}\big) = \prod_{i=1}^{d} \max\{0, 2r - |v_i|\}$. Combined with the fact that $\mathrm{Vol}(\mathcal{A}) = \mathrm{Vol}(\mathcal{B}) = (2r)^d$ this gives

$$\begin{aligned}\mathrm{TV}(\mathcal{U}_r, \mathcal{U}_r + v) &= \frac{1}{(2r)^d}\left((2r)^d - \prod_i \max\{0, 2r - |v_i|\}\right) \\ &= 1 - \prod_i \max\left\{0, 1 - \frac{|v_i|}{2r}\right\},\end{aligned}$$

as required. ■

We can also compute the TV-distance for the worst-case shift v with $\|v\|_\infty \le \epsilon$.

Corollary 22 *For any $\epsilon \ge 0$, the vector $v = (\epsilon, \dots, \epsilon) \in \mathbb{R}^d$ satisfies*

$$v \in \operatorname*{argmax}_{v:\|v\|_\infty \le \epsilon} \mathrm{TV}(\mathcal{U}_r, \mathcal{U}_r + v),$$

and $\mathrm{TV}(\mathcal{U}_r, \mathcal{U}_r + v) = \min\{1, 1 - (1 - \frac{\epsilon}{2r})^d\}$. Finally, for $\epsilon \in [0, r]$, we have

$$1 - e^{-\frac{d\epsilon}{2r}} \le \max_{v:\|v\|_\infty \le \epsilon} \mathrm{TV}(\mathcal{U}_r, \mathcal{U}_r + v) \le 1 - 4^{-\frac{d\epsilon}{2r}}.$$

Proof To see that $v = (\epsilon, \dots, \epsilon)$ is a maximizer, observe that the optimization problem decouples over the components v_i and that to maximize the term corresponding to component v_i we want to choose $|v_i|$ as large as possible. It follows that all vectors $v \in \{\pm\epsilon\}^d$ are maximizers.

The bounds for when $\epsilon \in [0, r]$ follow from the fact that for any $z \in [0, \frac{1}{2}]$, we have $4^{-x} \le 1 - z \le e^{-z}$ applied with $z = 1 - \frac{\epsilon}{2r}$. ■

Corollary 23 *Fix any dimension d, radius $\epsilon > 0$, and total variation bound $\delta \in [0, 1]$. Setting $r = \frac{1}{2}\frac{\epsilon}{\delta} d \log(4)$ guarantees that for all $v \in \mathbb{R}^d$ such that $\|v\|_\infty \le \epsilon$ we have $\mathrm{TV}(\mathcal{U}_r, \mathcal{U}_r + v) \le \delta$. Moreover, if $\eta \sim \mathcal{U}_r$ then $\mathbb{E}[\eta_i^2]$ is the variance of a uniform random variable on $[-r, r]$, which is $\frac{\sqrt{2}\log(4)^2}{48}\frac{\epsilon^2}{\delta^2} d^2 \le \frac{\epsilon^2}{\delta^2} d^2$.*

Proof This follows by determining the smallest value of r for which $1 - 4^{-\frac{d\epsilon}{2\delta}} \le \delta$. ■

Appendix C. Lower Bound on ℓ_2-Norm of Noise

Lemma 24 *Fix any $p \ge 2$ and let $\mathcal{D}$ be a distribution on $\mathbb{R}^d$ such that there exists a radius ϵ and total variation bound δ satisfying that for all $v \in \mathbb{R}^d$ with $\|v\|_p \le \epsilon$ we have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v) \le \delta$. Then*

$$\mathop{\mathbb{E}}_{\eta \sim \mathcal{D}} \|\eta\|_2 \ge \frac{\epsilon d^{1-1/p}}{24} \cdot \frac{1 - \delta}{\delta}.$$

Proof We first prove the lemma in the one-dimensional case. Let $\eta' = \eta + \epsilon \in \mathbb{R}$ so that η' is a sample from $\mathcal{D} + \epsilon$ and define $r = \epsilon/2$ so that the sets $\mathcal{A} = (-r, r)$ and $\mathcal{B} = (\epsilon - r, \epsilon + r)$ are disjoint. From Markov's inequality, we have that $\Pr(\eta \in \mathcal{A}) = 1 - \Pr(|\eta| \geq r) \geq 1 - \frac{\mathbb{E}|\eta|}{r}$. Further, since $\eta' \in \mathcal{B}$ if and only if $\eta \in \mathcal{A}$, we have $\Pr(\eta' \in \mathcal{B}) \geq 1 - \frac{\mathbb{E}|\eta|}{r}$. Next, since $\mathcal{A}$ and $\mathcal{B}$ are disjoint, it follows that $\Pr(\eta' \in \mathcal{A}) \leq 1 - \Pr(\eta' \in \mathcal{B}) \leq 1 - 1 + \frac{\mathbb{E}|\eta|}{r} = \frac{\mathbb{E}|\eta|}{r}$. Finally, we have $\delta \geq \Pr(\eta \in \mathcal{A}) - \Pr(\eta' \in \mathcal{A}) \geq 1 - \frac{2\,\mathbb{E}|\eta|}{r} = 1 - \frac{4\,\mathbb{E}|\eta|}{\epsilon}$. Rearranging this inequality proves the claim that $\mathbb{E}|\eta| \geq (1-\delta)\epsilon/4$. Combining with Lemma 8, we obtain that $\mathbb{E}\,|\eta| \geq \frac{\epsilon}{12} \cdot \frac{1-\delta}{\delta}$, due to the fact that for any $\delta \in (0, 1]$ we have $\max\{\frac{1-\delta}{4}, \frac{(1-\delta)^2}{8\delta}\} \geq \frac{1}{12} \cdot \frac{1-\delta}{\delta}$.

We now prove the d-dimensional case. Let η be a sample from $\mathcal{D}$. By scaling the vectors from Corollary 11 by ϵ, we obtain $b > d/2$ vectors $v_1, \ldots, v_b \in \mathbb{R}^d$ with $\|v_i\|_p = \epsilon$ and $\|v_i\|_2 = \epsilon \cdot b^{1/2-1/p}$. By assumption we must have $\mathrm{TV}(\mathcal{D}, \mathcal{D} + v_i) \leq \delta$, since $\|v_i\|_p \leq \epsilon$, and the above-mentioned one-dimensional case implies that $\mathbb{E}\,\frac{|v_i^\top \eta|}{\|v_i\|_2} \geq \frac{\|v_i\|_2}{12}\frac{1-\delta}{\delta}$ for each i. We use this fact to bound $\mathbb{E}\,\|\eta\|_2$.

Let $\mathbf{Q} \in \mathbb{R}^{b \times d}$ be the matrix whose i^{th} row is given by $v_i / \|v_i\|_2$ so that $\mathbf{Q}$ is the orthogonal projection matrix onto the subspace spanned by the vectors $v_1, \ldots, v_b$. Then we have $\mathbb{E}\,\|\eta\|_2 \geq \mathbb{E}\,\|\mathbf{Q}\eta\|_2 \geq \frac{1}{\sqrt{b}}\,\mathbb{E}\,\|\mathbf{Q}\eta\|_1 = \frac{1}{\sqrt{b}}\sum_{i=1}^b \mathbb{E}\,\frac{|v_i^\top \eta|}{\|v_i\|_2} \geq \frac{1}{\sqrt{b}}\sum_{i=1}^b \frac{\|v_i\|_2}{12}\frac{1-\delta}{\delta}$, where the first inequality follows because orthogonal projections are non-expansive, the second inequality follows from the equivalence of ℓ_2 and ℓ_1 norms, and the last inequality follows from $\mathbb{E}\,\frac{|v_i^\top \eta|}{\|v_i\|_2} \geq \frac{\|v_i\|_2}{12}\frac{1-\delta}{\delta}$. Using the fact that $\|v_i\|_2 = \epsilon \cdot b^{1/2-1/p}$, we have that $\mathbb{E}\,\|\eta\|_2 \geq \frac{\epsilon b^{1-1/p}}{12} \cdot \frac{1-\delta}{\delta}$. Finally, since $b > d/2$ and $(1/2)^{1-1/p} \geq 1/2$ for $p \geq 2$, we have $\mathbb{E}\,\|\eta\|_2 \geq \frac{\epsilon d^{1-1/p}}{24} \cdot \frac{1-\delta}{\delta}$, as required. ■

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283, 2018.

Wieland Brendel, Jonas Rauber, Alexey Kurakin, Nicolas Papernot, Behar Veliqi, Sharada P Mohanty, Florian Laurent, Marcel Salathé, Matthias Bethge, Yaodong Yu, Hongyang Zhang, et al. Adversarial vision challenge. In *The NeurIPS'18 Competition*, pages 129–153. Springer, 2020.

Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320, 2019.

Luc Devroye, Abbas Mehrabian, and Tommy Reddad. The total variation distance between high-dimensional Gaussians. *arXiv preprint arXiv:1810.08693*, 2018.

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

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

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

Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *International Conference on Learning Representations*, 2020.

Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *International Conference on Machine Learning*, pages 5567–5576, 2020.

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

Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 4911–4922, 2019.

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

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3578–3586, 2018.

Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10877–10887, 2018.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 11289–11300, 2019.

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

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

Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, pages 8400–8409, 2018.

Eric Wong, Frank R Schmidt, and J Zico Kolter. Wasserstein adversarial examples via projected Sinkhorn iterations. In *International Conference on Machine Learning*, pages 6808–6817, 2019.

Kai Y Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing ReLU stability. In *International Conference on Learning Representations*, 2019.

Greg Yang, Tony Duan, Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 1310–1320, 2020a.

Xiao Yang, Fangyun Wei, Hongyang Zhang, Xiang Ming, and Jun Zhu. Design and interpretation of universal adversarial patches in face detection. In *European Conference on Computer Vision*, 2020b.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Ruslan Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *Advances in Neural Information Processing Systems*, 2020c.

Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. MACER: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482, 2019.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020.