Hidden Cost of Randomized Smoothing

Jeet Mohapatra*
MIT

Ching-Yun Ko*
MIT

Tsui-Wei Weng
MIT-IBM Watson AI Lab

Pin-Yu Chen
IBM Research

Sijia Liu
MIT-IBM Watson AI Lab

Luca Daniel
MIT

Abstract

The fragility of modern machine learning models has drawn a considerable amount of attention from both academia and the public. While immense interests were in either crafting adversarial attacks as a way to measure the robustness of neural networks or devising worst-case analytical robustness verification with guarantees, few methods could enjoy both scalability and robustness guarantees at the same time. As an alternative to these attempts, randomized smoothing adopts a different prediction rule that enables statistical robustness arguments which easily scale to large networks. However, in this paper, we point out the side effects of current randomized smoothing workflows. Specifically, we articulate and prove two major points: 1) the decision boundaries of smoothed classifiers will shrink, resulting in disparity in class-wise accuracy; 2) applying noise augmentation in the training process does not necessarily resolve the shrinking issue due to the inconsistent learning objectives.

1 INTRODUCTION

Current mainstream methods to evaluate robustness of DNNs against adversarial examples [Szegedy et al. (2014); Biggio et al. (2013)] employ robustness verification. Such techniques can guarantee that no adversarial examples can exist within a specified distance r from a given input. As computing the largest possible r has been proven to be NP-complete [Katz et al. (2017)], one popular approach is to derive a certified lower bound of r through convex/linear relaxation [Hein and Andriushchenko (2017); Weng et al. (2018a); Singh et al. (2018); Zhang et al. (2018)], which can be computed efficiently. Nevertheless, these techniques can hardly scale to state-of-the-art DNNs on ImageNet, motivating the idea of applying *randomized smoothing* [Cohen et al. (2019); Lecuyer et al. (2019); Li et al. (2019); Jia et al. (2019); Lee et al. (2019)] (*i.e.* a spatial low-pass filter) to transform the original classifier into a "smoothed" counterpart. This new smoothed classifier now returns the class with the highest probability by querying input data that has been purposely corrupted by isotropic Gaussian noise $N(0, \sigma^2\mathcal{I})$.

Although randomized smoothing allows non-trivial robustness verification for the smoothed classifier on ImageNet, the side-effects of randomized smoothing have not yet been rigorously studied, except for a case-study of one specific binary classifier in [Głuch and Urbanke (2019), p.2] and some impossibility results on accuracy-certification trade-off [Yang et al. (2020); Blum et al. (2020); Kumar et al. (2020)]. The main motivation of this paper is to take a deep dive into the hidden cost of randomized smoothing for general multi-class classifiers.

The development of this paper is as follows: in Section 2 we review basic preliminaries for adversarial robustness certification with randomized smoothing; in Section 3 we fully expose a major hidden cost of randomized smoothing – biased predictions, by providing evidences from both real-life and synthetic datasets; in Section 4 we provide a comprehensive theory exposing the root of the biased prediction – referred to as the *shrinking phenomenon* in the remainder of the paper; in Section 5 we hold a discussion on the effects of data augmentation on the shrinking phenomenon and implications given by our theoretical analysis. Table 1 summarizes our contributions.

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

Table 1: A look-up table of theoretical (**T**) and numerical (**N**) contributions in Section 4.

region geometry	shrinking	vanishing rate σ_{van}	shrinking rate	certified radius
bounded	**T** (Thm. 3)	**T** - lower bnd. (Thm. 4)	**N** - lower bnd. (Fig. 2)	**N** - case study
semi-bounded	**T** (Thm. 5)	not applicable	**T** - lower bnd. (Thm.7)	**N** - case study

2 BACKGROUND

2.1 Randomized Smoothing and Adversarial Robustness

Generally, the prediction of a model for input x_0 is given by taking the highest output of the score function (a neural network) $g(x_0)$. Let e_i denote the i^{th} basis vector with all components 0 and the i^{th} component be 1. Then the base classifier can be given as

$$f(x_0) = e_{\xi_A}; \quad \xi_A = \arg\max_j \ g_j(x_0). \tag{1}$$

Correspondingly, under randomized smoothing the prediction for a model g is given as the "most likely" standard prediction output by the model when noise is added to the input. Conventionally, the resulting classifier is referred to as the *smoothed classifier* and the type of noise added to the input is denoted as the *smoothing measure*. When isotropic Gaussian distribution $\mathcal{N}(0, \sigma^2\mathcal{I})$ is used as the smoothing measure, the smoothed function f_σ is given as

$$f_\sigma(x_0) = e_{\xi_A};$$
$$\xi_A = \arg\max_j \ \mathbb{P}[j = \arg\max_i g_i(x)], \ x \sim \mathcal{N}(x_0, \sigma^2\mathcal{I}).$$

There has been a lot of research in developing robustness verification techniques for the *base classifier* in Equation (1) [Hein and Andriushchenko (2017); Weng et al. (2018a); Gehr et al. (2018); Raghunathan et al. (2018); Weng et al. (2018b); Wong and Kolter (2018); Wang et al. (2018); Li et al. (2020)], *i.e.* given g, x_0, ξ_A and p, find the maximum value of r such that $\arg\max_j g_j(x_0 + \delta) = \xi_A, \ \forall \|\delta\|_p \leq r$. However, due to the intrinsic hardness of the problem [Katz et al. (2017); Weng et al. (2018a); Tjeng et al. (2018)], the above approaches can hardly scale to state-of-the-art deep neural networks such as ResNet-50 and VGG-19 nets. On the other hand, it is also possible to perform robustness verification on the *smoothed classifier*. To solve the problem of certification, Lecuyer et al. (2019) first applied differential privacy techniques to derive a non-trivial lower bound of r for $p = 1, 2$. The bound was later improved by Li et al. (2019) via the tools in information theory for $p = 2$. Recently, Cohen et al. (2019) proved a tighter bound of r for $p = 2$ below:

$$r = \frac{\sigma}{2}\left[\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})\right], \tag{2}$$

where σ is the smoothing factor in the Gaussian noise, Φ^{-1} is the inverse of standard Gaussian CDF, and $\underline{p_A}$ and $\overline{p_B}$ are the lower/upper bound on the probability with class ξ_A and ξ_B (ξ_A is the top-1 class of the smoothed classifier and ξ_B is the "runner-up" class), respectively. In practice, Cohen et al. (2019) sets $\overline{p_B} = 1 - \underline{p_A}$ and abstains when $\underline{p_A} < 0.5$, implying that no radius can be certified in this case.

2.2 Data Augmentation

In the seminal work of randomized smoothing , Cohen et al. (2019) and Lecuyer et al. (2019) suggest to apply randomized smoothing during training (noise augmentation) for better classification accuracy. We first recall that a standard learning problem takes the form of

$$\mathcal{R} = \mathbb{E}_{x\in\mathcal{X}}[l(f(x), h(x))],$$

where $\mathcal{X}$, $\mathcal{Y}$, l, f, and h are the input space, the output space, the loss function, a neural network, and the ground-truth classifier, respectively. Given some probability distribution $\mathfrak{D}_p$ the noise smoothing risk takes the form of

$$\begin{aligned}\mathcal{R}_{\text{RS}} &= \mathbb{E}_{x\in\mathcal{X}}[l(f_\sigma(x), h(x))] \\ &= \mathbb{E}_{x\in\mathcal{X}}[l(\mathbb{E}_{z\sim\mathfrak{D}_p}[f(x+z)], h(x))].\end{aligned}$$

Cohen et al. (2019) motivate the use of corrupted samples during training by arguing that, when l is chosen to be the cross entropy and $\mathfrak{D}_p = \mathcal{N}(0, \sigma^2 I)$, the noise augmentation risk

$$\mathcal{R}_{\text{RS-train}} = \mathbb{E}_{x\in\mathcal{X}}[\mathbb{E}_{z\sim\mathfrak{D}_p}[l(f_{\text{train},\sigma}(x+z), h(x))]]$$

constitutes a lower bound of $\mathcal{R}_{\text{RS}}$. We distinguish $f_{\text{train},\sigma}$ from f since they are learned from different objectives. Throughout this paper, we abbreviate Gaussian noise augmentation (*i.e.* $\mathfrak{D}_p$ be the Gaussian centered at the origin) as data augmentation.

3 TWO MOTIVATING EXAMPLES

The major highlight of randomized smoothing techniques in the scope of adversarial robustness is its ability to provide non-trivial robustness guarantees (certified radii) for large networks. With this in mind, as pointed out in [Cohen et al. (2019), Sec. 3.2.2 last para.], for randomized smoothing with parameter

Table 2: The mean certified radii (with ± std.) of CIFAR10 classifiers learned with data augmentation and inferred by the randomized smoothing prediction rule. "certified radius (c)" denotes the correct certified radius.

training σ	0.12	0.25	0.50	1.00	1.50	2.00	3.00
min & max class-wise acc.(%)	$(67.8 \pm 1.9,$ $93.4 \pm 1.3)$	$(55.4 \pm 4.8,$ $89.2 \pm 1.3)$	$(42.4 \pm 4.8,$ $81.9 \pm 2.2)$	$(20.8 \pm 1.3,$ $72.8 \pm 1.5)$	$(9.8 \pm 1.3,$ $61.2 \pm 3.1)$	$(5.4 \pm 0.9,$ $53.2 \pm 3.9)$	$(1.2 \pm 0.8,$ $41.0 \pm 1.0)$
certified radius	0.28 ± 0.01	0.42 ± 0.02	0.51 ± 0.03	0.50 ± 0.01	0.44 ± 0.01	0.38 ± 0.01	0.32 ± 0.01
certified radius (c)	0.34 ± 0.01	0.56 ± 0.01	0.80 ± 0.02	1.07 ± 0.01	1.25 ± 0.03	1.40 ± 0.03	1.80 ± 0.07

σ, the maximum achievable certified radius is around 4σ, implying larger smoothing factor σ is needed for a larger maximum achievable certified radius[1]. This need is further justified in Cohen et al. (2019) by pointing out the trade-off between the sample complexity and certified radii with a fixed smoothing factor. Therefore, one has to use large σ to achieve the state-of-the-art robustness guarantees while avoiding impractical sample complexity.

In Table 2, we validate this point by calculating the certified radii of CIFAR10 smoothed classifiers with base classifier trained with data augmentation[2]. In this experiment, we vary the smoothing factor σ from 0.12 to 3.00, which is used simultaneously in data augmentation and randomized smoothing. When reporting their certified radius, we consider two metrics: 1) certified radius - the mean of all certified radii in the testing set, with the radius assigned to zero for wrongly-classified samples; and 2) correct certified radius - the mean of certified radii of correctly-classified samples in the testing set. We then see that with the increasing smoothing factor σ, the average certified radius of correctly-classified samples keeps rising from only 0.34 to 1.80, obtaining indeed non-trivial robustness guarantees.

On the other hand, the average certified radius of all samples climbs to around 0.5 and then decreases to 0.32. This is because the classification accuracy also drops as one uses larger σ, pushing more samples to have zero certified radius. In order to better understand the drop in accuracy and the affected examples, we provide a case study over a synthetic dataset.

3.1 Synthetic Datasets

Consider the binary-classification problem on the dataset ($\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$) given as mixture of Gaussians:

$$\mathcal{X}_1 = (\frac{1}{2} - \epsilon) \cdot \mathcal{N}(-a, \sigma_o^2) + \epsilon \cdot \mathcal{N}(ka, \sigma_o^2);$$
$$\mathcal{X}_2 = \frac{1}{2} \cdot \mathcal{N}(0, \sigma_o^2);$$

where $a, k, \sigma_o \in \mathbb{R}^+/\{0\}$. Then we have

Theorem 1. *Consider a classifier f_{train,σ_t} given as the naive-Bayes classifier obtained by training on the dataset $\mathcal{X}$ with data augmentation of variance σ_t. Let the class-wise accuracy of the two classes with f_{train,σ_t} using the randomized smoothing prediction rule be given as $Acc_1(\sigma_t)$ and $Acc_2(\sigma_t)$. Then we define the bias $(\Delta(\sigma_t))$ to be the gap between class-wise accuracies $(\Delta(\sigma_t) = |Acc_1(\sigma_t) - Acc_2(\sigma_t)|)$. For $k > \frac{1}{2\epsilon} - 1$, class I decision region grows in size at a rate of $\Theta(\sigma_t^2)$ and thus the bias is large for large σ_t.*

It is quite well-known that using higher σ leads to lowering of accuracy. In general, previous works have stated the existence of a robustness-accuracy trade-off. Here, we notice another interesting and quite important problem that is created by randomized smoothing: randomized smoothing based models for high values of σ_t are biased in their predictions. Some classes are favored a lot more than others, resulting in huge difference in class-wise accuracies.

In order to better understand the extent of the bias possible, we also study the limiting case of $\sigma_o \to 0$. This allows us to effectively study large bias without having $\sigma_t \to \infty$. In particular, we consider the dataset($\mathcal{X}'$) with probability mass function :

$$\rho(0, 1) = \frac{1}{2}; \quad \rho(-a, 2) = \frac{1}{2} - \epsilon; \quad \rho(ka, 2) = \epsilon,$$

with a, k defined as before. For this new dataset, we see that

Theorem 2. *Consider a classifier f_{train,σ_t} given as the naive-Bayes classifier obtained by training on the dataset $\mathcal{X}'$ with data augmentation of variance σ_t. The bias of the classifier f_{train,σ_t} using the randomized smoothing prediction rule is $1 - \epsilon$, if $k > \frac{e^2}{\epsilon} - 1$ and* $\sigma_t \geq a\sqrt{\frac{k(k+1)}{2ln(2\epsilon(k+1)) - \frac{2k}{k+2}}}$.

[1] One can also gain insights from that the certified radius r is proportional to the smoothing factor σ (*cf. Equation 2*).

[2] Throughout the paper, all the classification results and certified radii are obtained with the open-source code provided by Salman et al. (2019).

Table 3: The class-wise accuracy (%) in percentile of classifiers and smoothing factors used in Cohen et al. (2019).

	CIFAR10					ImageNet				
percentile	1st	25th	50th	75th	100th	1st	25th	50th	75th	100th
$\sigma = 0.00$	78	88	91	93	96	14	66	78	88	100
0.12	0	8	15	24	100	0	36	52	66	96
0.25	0	0	0	0	72	0	2	10	20	82
0.50	0	0	0	0	98	0	0	0	0	56

To give intuitive understanding of the critical smoothing factor in Theorem 2, we fix the scale of the dataset $a(k+1)$ to be $[0,1]$ as is common-practice in the literature [Cohen et al. (2019); Salman et al. (2019)]. Then, we observe the shrinking effects happen at $\sigma \approx 0.7$ which is well within the realm of smoothing factors used in practice (Cohen et al. (2019); Salman et al. (2019) use smoothing factors upto 1.0 for data augmentation and randomized smoothing). This idea can be extended to several more general and interesting cases: a multi-class case giving accuracy $\frac{1}{c} + \epsilon$ by having class 1 with the same distribution and the rest of the classes with distributions similar to that of class 2's; and a binary-class case where adopting data augmentation does not change the optimal solution but the subsequent randomized smoothing inference still gets low accuracy for a high enough smoothing factor σ. The proofs of Theorem 1 and Theorem 2 are included in the supplementary materials for interested readers.

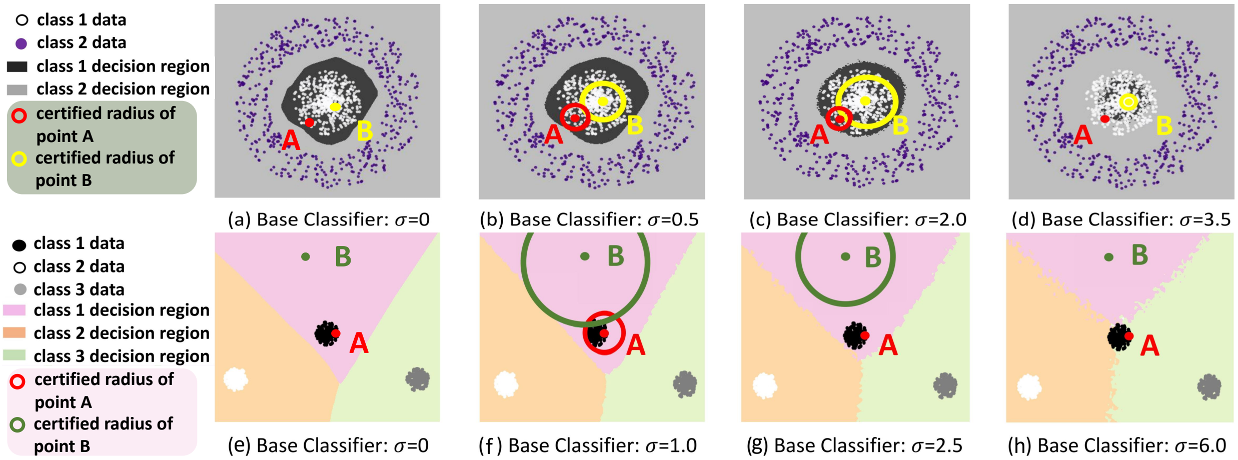
3.2 Real-Life Datasets

In the existing literature, randomized smoothing remains a legitimate way of providing adversarial robustness. However, the results on the synthetic datasets suggest randomized smoothing is biased towards some classes. In order to see if the bias is present in real-life datasets we consider a new metric, namely the min and max class-wise accuracy, where we calculate separately for each class their classification accuracy and report the minimum and the maximum. In Table 2 we give the performance of randomized smoothing based classifiers under the new metric. With this metric, one can then readily see that despite the increasing trend in certified radii, the class-wise accuracies becomes more imbalanced at higher smoothing factor σ. Specifically, when the smoothing factor $\sigma = 0.12$, the smoothed network with base classifier being trained by data augmentation with the same magnitude of Gaussian noise classifies "cat" samples with 67% accuracy and "automobile" samples with 92% accuracy. However, when $\sigma = 1.00$, this gap evolves to 22% accuracy ("cat") versus 68% accuracy ("ship"). This comes as an unpleasant surprise since it essentially means despite the current success of randomized smoothing in adversarial robustness, the method can lead to biased predictions, causing fairness issues.

As remarked earlier, a randomized smoothing model differs from other models in two phases, data augmentation during training and smoothing during inference. As the statistical guarantees given by randomized smoothing depend on the smoothing during inference, we focus on its role in producing the bias. Before proceeding, we verify that the bias problem still persists in the absence of augmentation during training. We conduct the smoothing experiments on the pretrained models provided by Cohen et al. (2019). In Table 3, we report the smoothing factors σ and corresponding class-wise accuracies (sorted ascendingly) in percentile of [1st,25th,50th,75th,100th]. That is, the 1st and 100th in the percentile correspond to the lowest (min) and highest (max) class-wise accuracy, respectively For CIFAR10, the [25th, 75th] percentile corresponds to the [3rd, 8th] lowest per-class accuracy. One can then see that originally more than 3/4 of the classes in datasets have reasonable accuracy, which decreases as σ goes bigger. Eventually, when $\sigma = 0.5$, more than 3/4 of the classes have 0 accuracy. Notably, $\sigma = 0.5$ is a reasonable number under the current randomized smoothing regime since the largest sigma used by Cohen et al. (2019) and Salman et al. (2019) is 1.0. Thus, we see that randomized smoothing produces biased results even in the absence of data augmentation during training. In the next section, we analyze how biased predictions are caused by randomized smoothing depending on the geometry of the underlying data distribution.

4 THEORETICAL CHARACTERIZATION OF THE SHRINKING PHENOMENON

Before we start our theoretical characterization, we first give a visual inspection of how randomized smoothing can change the decision regions. Specially, Figure 1 illustrates two toy examples, in which the decision regions of class 1 data (the dark green region in the first row and the pink region in the second row) shrink with larger smoothing factors σ. As consequences of the shrinkage, the class-wise accuracy for class 1 data



(a) Base Classifier: σ=0 (b) Base Classifier: σ=0.5 (c) Base Classifier: σ=2.0 (d) Base Classifier: σ=3.5

(e) Base Classifier: σ=0 (f) Base Classifier: σ=1.0 (g) Base Classifier: σ=2.5 (h) Base Classifier: σ=6.0

Figure 1: The 1st row shows examples of **bounded** decision regions for smoothed classifiers. The 2nd row shows examples of **semi-bounded** decision regions. The class 1 decision regions shrink as the smoothing factor σ increases from left to right. In case (h) with larges σ, the decision region has shrunk so much that class 1 data are completely misclassified. We also plot the certified radius (Equation 2) of point A and B and show that it may decrease as σ increases.

drops drastically, leading to the biased prediction.

Indeed, in this section, we aim to take a close look at this *shrinking phenomenon* of randomized smoothing, uncovering the fundamental problem of the technique. Moreover, we conduct a rigorous study providing also the bounds of extreme values, beyond which the shrinking phenomenon will happen. Our results are tight and prove the prevalence of such phenomena. In order to facilitate this analysis we perform the following reductions.

Problem Reductions. By the definition of randomized smoothing, the smoothed function depends on the base classifier only through the indicator function f. As the smoothed function f_σ only depends on the partitioning of the input space created by the base classifier g, we shift our focus from the output of g to how it partitions the input space, *i.e.*, we are interested in characterizing all possible partitions of the input space that can lead to biased prediction as one applies randomized smoothing with a high σ. As it is hard to measure a decrease in accuracy directly from the geometry of the classifier, we approximate the decrease in accuracy using the mismatch in partitions of input space provided by f and by f_σ.

However, the problem of characterizing the partitions of the space into multiple classes is intractable. So we instead focus on tracking the behaviour of the decision boundary of a single class with respect to randomized smoothing. Without loss of generality, we set the concerned class as class 1. In this case, we analyze the misclassification rate for class 1 by the region size of the input space that is partitioned as class 1 under f but not under f_σ. Considering that for any $x \in \mathbb{R}^d$, the necessary condition for it to be classified as class 1 is to have $f_\sigma(x)_1 \geq \frac{1}{c}$, so we do a worst-case analysis by assuming the reformed class 1 partition is defined by exactly $f_\sigma(x)_1 \geq \frac{1}{c}$. If this overestimated reformed class 1 partition is still smaller than the original, then for sure the actual misclassification rate will be higher than the analysis herein.

Problem Formulation. We formulate our problem as to characterize the "decision regions" that will shrink or drift after applying randomized smoothing. Formally, the decision region $\mathcal{D}$ of class 1 data is determined by the classifier f via $\mathcal{D} = \{x \mid f(x)_1 = 1\}$. By adopting randomized smoothing, we obtain $f_\sigma(x) = \int_{x' \in \mathbb{R}^d} f(x')p(x')dx'$ with the decision region denoted by $\mathcal{D}_\sigma = \{x \mid (f_\sigma(x))_1 \geq \frac{1}{c}\}$. The scope of this section is to investigate under what conditions (*w.r.t.* the classifier and smoothing factor σ) will the shrinking occur. On the whole, the shrinking effect depends highly on the geometry of the data distribution. However, considering the intractable numbers of possible decision region geometry, we will only discuss here two major classes of the geometries (*bounded* in Section 4.1 and *semi-bounded* in Section 4.2) for multidimensional data (*i.e.* $d > 1$). We supplement $d = 1$ discussions in the supplementary materials for readers' references. All the proofs are also deferred to the supplementary materials due to page limit.

4.1 Bounded Decision Region

In this section, we aim at proving the shrinking side-effects incurred by the smoothing filter when the decision region is bounded. Formally, we say a decision region is bounded and shrinks according to the following definition:

Definition 1 (Bounded Decision Regions)**.** *If the decision region (disconnected or connected) of class 1 data is a bounded set in the Euclidean space (can be bounded by a ball of finite radius), then we call these decision regions bounded decision regions.*

We denote the smallest ball that contains the original decision region of f by $S_{\mathcal{D}}$ ($\mathcal{D} \subseteq S_{\mathcal{D}}$). Similarly, we let the smallest ball that contains the smoothed decision region (the decision region of smoothed classifier) be $S_{\mathcal{D}_\sigma}$ ($\mathcal{D}_\sigma \subseteq S_{\mathcal{D}_\sigma}$).

Definition 2 (Shrinking of Bounded Decision Regions)**.** *A bounded decision region is considered to have shrinked after applying smoothing filters if the radius R_σ of $S_{\mathcal{D}_\sigma}$ is strictly smaller than the radius R of $S_{\mathcal{D}}$, i.e. $R_\sigma < R$, where $S_{\mathcal{D}}$ and $S_{\mathcal{D}_\sigma}$ are the smallest balls containing the original decision region and the smoothed decision region, respectively.*

For randomized smoothing, we observe that

Corollary 1. *The smallest ball $S_{\mathcal{D}_\sigma}$ containing the smoothed decision region is contained within the smoothed version of $S_{\mathcal{D}}$, i.e. $S_{\mathcal{D}_\sigma} \subseteq (S_{\mathcal{D}})_\sigma$*[3]*.*

Theorem 3. *A bounded decision region shrinks after applying randomized smoothing filters with large σ. Specifically, if $\sigma > \frac{R\sqrt{c}}{\sqrt{2(d-1)}}$, then $R_\sigma < R$ (cf. Definition 2).*

Analysis of bounded decision regions with randomized smoothing. As we have proven that any bounded decision region shrinks after applying randomized smoothing filters, we will investigate in this part of the paper *how fast* the decision region (quantified by R_σ) shrinks/vanishes. From Corollary 1, we have that the smallest ball $S_{\mathcal{D}_\sigma}$ containing the smoothed decision region is contained within the smoothed version of $S_{\mathcal{D}}$. Therefore we only consider the worst case when we have a ball-like decision region. Without loss of generality, we consider a case when the decision region of class 1 data characterized by the network function is exactly $\{x \in \mathbb{R}^d \mid \|x\|_2 \leq R\}$.

Theorem 4 (Vanishing Rate in the Ball-like Decision Region Case)**.** *The decision region of class 1 data vanishes when smoothing factor $\sigma_{van} > \frac{R\sqrt{c}}{\sqrt{d}}$.*

We validate Theorem 4 for binary classification ($c = 2$) by substituting R by $R = 1$ and plot the shrinking rate (the derivatives of R_σ with respect to σ) of the decision region as a function of the smoothing factor σ for different input data dimensions $d = \{3, 8, 20, 30, 40, 50\}$ in Figure 2. Notably, the x-axis in Figure 2 is the varying smoothing factor σ and the y-axis is the rate of the shrinkage concerning class 1 decision region. We then see that overall the region vanishes at smaller smoothing factor σ_{van} with the growing dimension. For example, the shrinking rate curve stops at smoothing factor $\sigma_{\text{van}} = 0.651$ when $d = 3$ but at smoothing factor $\sigma_{\text{van}} = 0.141$ when $d = 50$. We collect these vanishing smoothing factors with different data dimensions and compare with the theoretical lower bounds found in Theorem 4 in the appendix to demonstrate the tightness of our theoretical lower bound. In a multi-class case, the certifiability and prediction do not follow the same setting as in Cohen et al. (2019). For the certifiability, the effective number of classes is 2 as Cohen et al. (2019) treats it as a one vs all setting. Therefore one would be unable to certify any radius with some smoothing factor $\sigma < \sigma_{\text{van}}$ in the multi-class case. We further elaborate on this point about certifiability in Section 4.3.

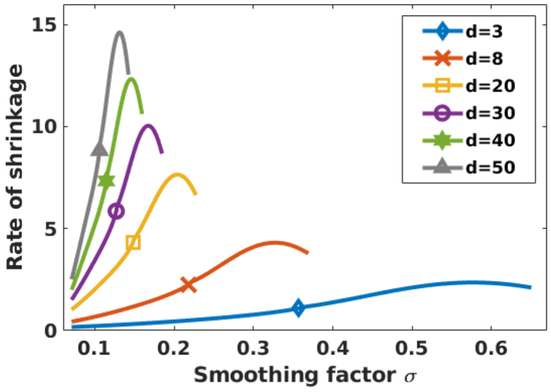


Figure 2: The shrinking rate of the decision region quantified by R_σ for different input data dimension d.

4.2 Semi-bounded Decision Region

In this section, we discuss the case when the decision region is semi-bounded and is not a half-space. Formally, we say a decision region is semi-bounded and shrinks according to the following definitions:

Definition 3 (Semi-bounded Decision Regions)**.** *If a decision region is not bounded and there exists a half-space $\mathcal{H}$ (decided by a hyperplane) that contains the unbounded decision region, then we call it semi-bounded decision region. We say a semi-bounded decision region is bounded in v-direction if there $\exists k \in \mathbb{R}/\infty$ such that for $\forall x \in \mathcal{D}$, $v^T x < k$.*

An illustrative example of semi-bounded decision regions is shown as Figure 1, where we have 3 clusters of

[3] $(S_{\mathcal{D}})_\sigma := \{x \mid (u_\sigma(x))_1 \geq \frac{1}{c}\}$, where $u_\sigma(x) = \int_{x' \in \mathbb{R}^d} \mathbf{1}_{S_{\mathcal{D}}} p(x') dx'$ and $p(x')$ is the pdf of Gaussian centered at x.

data points denoting three different classes' data and their decision regions. Observing the change in the decision region of class 1, we define "shrinking" as

Definition 4 (Shrinking of Semi-bounded Decision Regions)**.** *A semi-bounded decision region bounded in v-direction is distinguished as shrinked along the direction after applying smoothing filters if the upper bound of projections of the decision region onto direction v shrinks, i.e.* $\Upsilon^v_{\mathcal{D}_\sigma} < \Upsilon^v_{\mathcal{D}}$, *where* $\Upsilon^v_{\mathcal{D}} = \max_{x\in\mathcal{D}} v^T x, \Upsilon^v_{\mathcal{D}_\sigma} = \max_{x\in\mathcal{D}_\sigma} v^T x$.

With this definition of shrinking of semi-bounded decision regions, we demonstrate in the following that any "narrow" semi-bounded decision region bounded in v-dimension will shrink along the direction (*cf.* Figure 1(e-h)). We quantify the size of a decision region as follows:

Definition 5 (θ, v-Bounding Cone for a Decision Region)**.** *A* θ, v *cone is defined as a right circular cone* $\mathcal{C}$ *with axis along* $-v$ *and aperture* 2θ*. Then we define the* θ, v*-bounding cone* $\mathcal{C}^{\mathcal{D}}_{\theta,v}$ *for* $\mathcal{D}$ *as the* θ, v *cone that has the smallest projection on* v *and contains* $\mathcal{D}$*, i.e.,* $\mathcal{C}^{\mathcal{D}}_{\theta,v} = \arg\min_{\mathcal{D}\subseteq\mathcal{C}_{\theta,v}} \Upsilon^v_{\mathcal{C}_{\theta,v}}$.

Theorem 5. *A semi-bounded decision region that has a narrow bounding cone shrinks along v-direction after applying randomized smoothing filters with high* σ*, i.e. if the region admits a bounding cone* $\mathcal{C}^{\mathcal{D}}_{\theta,v}$ *with* $\tan(\theta) < \sqrt{\frac{(d-1)}{2c\log(c-1)}}$*, then for* $\sigma > (\Upsilon^v_{\mathcal{C}^{\mathcal{D}}_{\theta,v}} - \Upsilon^v_{\mathcal{D}})\tan(\theta)\sqrt{\frac{c}{d-1}} \cdot \frac{2(d-1)}{(d-1)-2\tan^2(\theta)c\log(c-1)}$, $\Upsilon^v_{\mathcal{D}_\sigma} < \Upsilon^v_{\mathcal{D}}$ *(cf. Definition 4).*

Concretely, the narrowness condition (the larger the easier to fulfil) of the cone for MNIST dataset [LeCun (1998)] relaxes to $0.43\pi = 76.7°$, meaning that if any single class's decision region can be bounded by a θ, v cone with θ being less than 76.7°, then shrinking effect happens. Correspondingly, this narrowness condition for CIFAR10 dataset [LeCun (1998)] is $0.46\pi = 83.2°$ and $0.42\pi = 75.2°$ for ImageNet dataset [Russakovsky et al. (2015)]. Notably, for binary classification tasks ($c = 2$), according to Theorem 5, the condition for shrinking reduces to $\tan(\theta) < \infty$ that implies $\theta < \pi/2$. In other words, when there are only two classes, as long as the semi-decision region is not a half-space, it **will** shrink.

Analysis of the semi-bounded case with randomized smoothing. As in Section 4.1, we conduct the analysis using the worst-case ball-like bounded decision region, here we correspondingly consider a solid right circular cone along the v direction. The shrinkage in this case serves as a non-trivial lower bound. Without loss of generality, we consider a θ, v solid right circular cone $\{x \in \mathbb{R}^d \mid v^T x - \|v\|\|x\|cos(\theta) \leq 0\}$ as the decision region $\mathcal{D}$ of class 1 data, where $-v = [0, \ldots, 0, 1]^T \in \mathbb{R}^d$. Since the semi-bounded decision region is unbounded and will shrink but will not vanish, we emphasize in this section only on giving the shrinking rate with respect to the smoothing factor σ, the number of classes c, the angle θ, and the data dimension d with randomized smoothing. Two major theorems regarding the shrinking rate in the solid cone-like decision region are:

Theorem 6. *The shrinkage of class 1 decision region is proportional to the smoothing factor, i.e.* $\Upsilon^v_{\mathcal{D}} - \Upsilon^v_{\mathcal{D}_\sigma} \propto \sigma$.

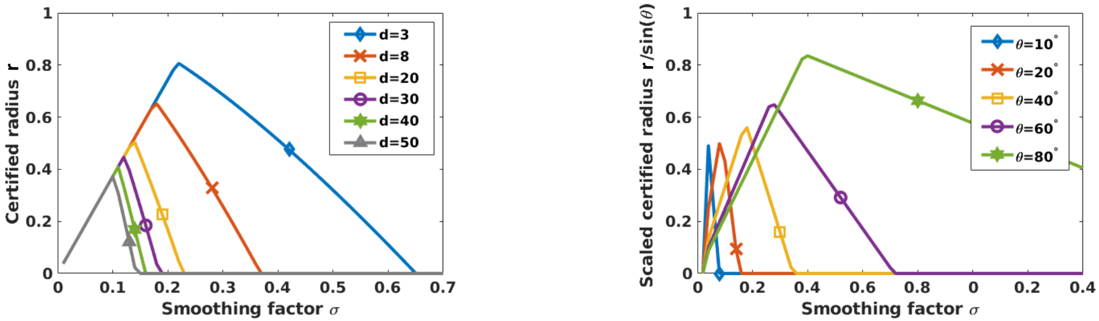
With the above Theorem 6, we can fix the smoothing factor to $\sigma = 1$ and further obtain a lower bound of the shrinking rate *w.r.t* c, θ, and d:

Theorem 7. *The shrinking rate of class 1 decision region is at least* $\sqrt{\frac{d-1}{c\tan^2(\theta)}} \cdot \frac{(d-1)-2\tan^2(\theta)c\log(c-1)}{2(d-1)}$*, i.e.* $\frac{\Upsilon^v_{\mathcal{D}_\sigma} - \Upsilon^v_{\mathcal{D}_{\sigma+\delta}}}{\delta} > \sqrt{\frac{d-1}{c\tan^2(\theta)}} \cdot \frac{(d-1)-2\tan^2(\theta)c\log(c-1)}{2(d-1)}$.

4.3 Remarks on Certified Radii

In the case of bounded decision region, the point at the origin has the highest probability to be classified as class 1 (see supplementary materials for the proof). Therefore when it has less than 0.5 probability to be classified as class 1, the decision region vanishes and no point can be certified (certified radius $r = 0$). Specifically, Figure 3(a) describes the certified radius r of the point at the origin using Equations (2) as a function of the smoothing factor σ and shows that the maximum certified radius (the peak) decreases with the increasing dimension. We include complete certified radius behavioral plots for different dimensions in the supplementary materials. As training samples are normally scaled in practice, they lie within a ball of radius $R \leq \sqrt{d}/2$. According to Theorem 4, for this ball, the upper bound of σ_{van} is $1/\sqrt{2} \approx 0.707$. So in practice, if the decision region of any class lies within the volume spanned by the training samples, its certifiable region vanishes for $\sigma \geq 0.708$, regardless of the input-space dimension d.

In the case of semi-bounded decision region, the point on the axis has the highest probability to be classified as class 1, thus we study the certified radius of a point $x_0 = [0, \ldots, 0, 1]$ as a function of cone narrowness θ and smoothing factor σ. Acknowledging that the minimum distance from x_0 to θ, v cones is $\sin(\theta)$, we show in Figure 3(b) the scaled certified radius $r/\sin(\theta)$ when $d = 25$. One can then readily verify that overall the peak scaled certified radius decreases with θ, *e.g.* the scaled certified radius at x_0 can be as large as 0.84 when $\theta = 80°$, while it is at most 0.49 when $\theta = 10°$. Moreover, we point out that certified radii drop to zero when we keep increasing the smoothing factor σ - the "narrower" (smaller θ) the decision region is,



(a)

(b)

Figure 3: (a) The certified radius r of the point at the origin for different input data dimension d; (b) The scaled certified radius $\frac{r}{\sin(\theta)}$ of a point on the axis v for cones with different apertures (2θ).

the faster they drop to zero. We discuss the effect of input data dimension d on the certified radius in the supplementary materials.

Interestingly, the certified radii increase with the growing smoothing factor σ but begin to decrease at certain point - larger certified radius can normally be obtained by larger smoothing factor σ according to Equation (2) but the dominance is taken over by the vanishing decision region when the σ is enough-close to σ_{van}. This also explains the eventual decrease in the average certified radius seen in Table 2. For small values of σ the average certified radius keeps increasing to a point ($\sigma_{thres} \in [0.50, 1.00]$) after which the effect of the vanishing decision region reduces the average certified radius.

5 EFFICACY OF DATA AUGMENTATION

As Section 4 proves that the biased prediction comes from the shrinking phenomenon of randomized smoothing, we want to hold a discussion herein investigating whether the state-of-the-art workflow for boosting randomized smoothing accuracy can solve this issue.

5.1 Counteracting Shrinking Effect Of Smoothing

Through the above arguments, we see that to countereffect the shrinkage induced by randomized smoothing, one will want to obtain larger decision regions for geometrically compact classes. Assuming a well-balanced distribution of classes, compact classes have a larger number of points near the margin compared to more spread-out classes. As a result, data augmentation expands the compact classes a lot more compared to other classes, partially alleviating the shrinking issue caused by smoothing. As a result, we see that the experiments in Table 3 (without data augmentation) have a much bigger bias in prediction compared to the experiments in Table 4 column “1-standard”, *e.g.* when $\sigma = 0.12$, Table 3 reads 0 versus 100 and Table 4 reads 67 versus 94.

However, it is important to note that the two effects do not exactly cancel each other out. Especially for high values of σ, the expansion caused by data augmentation can cause some of the more compact classes to dominate over all other classes, resulting in a highly biased classifier. Table 4 shows that the bias of the classifier consistently increases with increasing values of σ regardless of the number of augmenting points used. This signals two important observations: the need to limit the use of high values of smoothing factor σ and the need for a data geometry dependent augmentation scheme to properly counteract the shrinking effect caused by smoothing.

5.2 Heavy Data Augmentation

Besides showing the minimum and maximum class-wise accuracies of multiple CIFAR10 classifiers trained with standard data augmentation, we also give in Table 4 the corresponding accuracies for an enhanced version of data augmentation. Essentially, different from the standard data augmentation implementation, where only one point is used to estimate the expectation $\mathbb{E}_{z\sim\mathcal{D}_p}[l(f_{\text{train},\sigma}(x+z), h(x))]$ inside $\mathcal{R}_{\text{RS-train}}$, we evaluate the expectation using {10, 25, 50} points, reducing the estimation bias. We denote this scheme as heavy data augmentation. Using a larger number of augmentation allows us to approximate the augmented distribution more closely and remove any unnecessary bias that is caused by using a bad approximation of the data augmentation. The results in Table 4 show that the bias is slightly reduced by using a larger number {10, 25} of augmentation points but the problem still

Table 4: The minimum and maximum class-wise accuracy (%) of CIFAR10 classifiers learned with data augmentation and inferred by the randomized smoothing prediction rule. The smaller the gap between the maximum and the minimum class-wise accuracies is, the better.

#Augmentation Points	1 (standard)		10		25		50	
class-wise acc.	min	max	min	max	min	max	min	max
$\sigma = 0.12$	67	94	76	96	78	96	68	97
0.25	55	90	68	92	65	93	48	84
0.50	46	84	51	84	52	81	0	87
1.00	22	73	28	74	27	72	3	64

remains. Particularly, we see the relative improvement from increasing augmentation points becomes smaller with a larger smoothing factor σ. It is also worth noting that the gap in accuracies blows when we use up to 50 heavy data augmentation points, performing even worse than using the standard data augmentation. These observations signal it to be a more fundamental problem relating to the way we do data augmentation.

6 CONCLUSION

In this paper, we provide a theoretical characterization showing that randomized smoothing during inference can lead to a drastic gap among class-wise accuracies, even when it is included in the training phase. In addition, we observe that the smoothing during inference is very sensitive to the distribution of the data and can have wildly-different effects on different classes depending on the data geometry. A similar analysis could be extended to other smoothing functions in addition to Gaussian smoothing. Crucially, our results point out the need for limiting the use of large values of σ, as well as the need for data-geometry dependent noise augmentation schemes.

Acknowledgements

J. Mohapatra, C.-Y Ko, and L. Daniel are supported by MIT-IBM Watson AI Lab.

References

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. (2013). Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402.

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

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

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., and Vechev, M. (2018). Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *IEEE Symposium on Security and Privacy (SP)*, volume 00, pages 948–963.

Głuch, G. and Urbanke, R. (2019). Constructing a provably adversarially-robust classifier from a high accuracy one.

Hein, M. and Andriushchenko, M. (2017). Formal guarantees on the robustness of a classifier against adversarial manipulation. In *NeurIPS*.

Jia, J., Cao, X., Wang, B., and Gong, N. Z. (2019). Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *International Conference on Learning Representations*.

Katz, G., Barrett, C., Dill, D. L., et al. (2017). Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer.

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

LeCun, Y. (1998). The mnist database of handwritten digits.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. (2019). Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE.

Lee, G.-H., Yuan, Y., Chang, S., and Jaakkola, T. S. (2019). Tight certificates of adversarial robustness for randomly smoothed classifiers. *arXiv preprint arXiv:1906.04948*.

Li, B., Chen, C., Wang, W., and Carin, L. (2019). Certified adversarial robustness with additive noise. In *NeurIPS*.

Li, L., Qi, X., Xie, T., and Li, B. (2020). Sok: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131*.

Raghunathan, A., Steinhardt, J., and Liang, P. (2018). Certified defenses against adversarial examples. *ICLR*.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252.

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

Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. (2018). Fast and effective robustness certification. In *NeurIPS*.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2014). Intriguing properties of neural networks. *ICLR*.

Tjeng, V., Xiao, K. Y., and Tedrake, R. (2018). Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*.

Wang, S., Pei, K., Whitehouse, J., Yang, J., and Jana, S. (2018). Efficient formal safety analysis of neural networks. In *NeurIPS*.

Weng, T.-W., Zhang, H., Chen, H., Song, Z., Hsieh, C.-J., Boning, D., Dhillon, I. S., and Daniel, L. (2018a). Towards fast computation of certified robustness for relu networks. *ICML*.

Weng, T.-W., Zhang, H., Chen, P.-Y., et al. (2018b). Evaluating the robustness of neural networks: An extreme value theory approach. *ICLR*.

Wong, E. and Kolter, Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope. In *ICML*.

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

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. (2018). Efficient neural network robustness certification with general activation functions. In *NeurIPS*.