ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240792

Estimating the Robustness Radius for Randomized Smoothing with 100× Sample Efficiency

Emmanouil Seferis[a,b], Stefanos Kollias[b] and Chih-Hong Cheng[c]

[a]Fraunhofer IKS, Germany
[b]National Technical University of Athens, Greece
[c]Chalmers & University of Gothenburg, Sweden

Abstract. Randomized smoothing (RS) has successfully been used to improve the robustness of predictions for deep neural networks (DNNs) by adding random noise to create multiple variations of an input, followed by deciding the consensus. To understand if an RS-enabled DNN is effective in the sampled input domains, it is mandatory to sample data points within the operational design domain, acquire the point-wise certificate regarding robustness radius, and compare it with pre-defined acceptance criteria. Consequently, ensuring that a point-wise robustness certificate for any given data point is obtained relatively cost-effectively is crucial. This work demonstrates that reducing the number of samples by one or two orders of magnitude can still enable the computation of a slightly smaller robustness radius (commonly $\approx$ 20% radius reduction) with the same confidence. We provide the mathematical foundation for explaining the phenomenon while experimentally showing promising results on the standard CIFAR-10 and ImageNet datasets.

1 Introduction

Deep Neural Networks (DNNs) have achieved impressive results in tasks ranging from image and speech recognition [21, 17], language understanding [4], or game playing [28]. They have also been the key enabling technology in realizing perception modules for autonomous driving (see [18, 23, 6, 24] for recent survey). Nevertheless, the brittleness of the DNN [31] has been one of the safety concerns [1] contributing to the potential hazards.

Among many of the hardening techniques, *Randomized Smoothing (RS)* [10, 25, 35] is one of the promising model-agnostic methods to ensure the robustness of the DNN in the operational design domain (ODD), where the underlying idea is to make predictions based on the aggregation of outputs from multiple variations of the input data perturbed with random noise. For certification, however, it is imperative to understand how much an RS-enabled DNN can withstand noise, often characterized using the concept of *robustness radius*, i.e., the degree of perturbation where the majority-vote prediction remains consistent. This activity must be applied to the set of representative data points sampled within the ODD. The immediate consequence is its high computational cost: with m data points sampled from the ODD and each data point being perturbed n times, the number of DNN executions needed equals mn, which can be tremendous. It is also not a single-step process, as the introduction of continuous assurance [26, 2], safety certificates should be produced whenever the DNN module is updated. Thus, it is crucial to have highly efficient algorithms in computing the point-wise robustness certificate.

In this paper, we investigate the efficient computation of robustness certificates for each data point. Currently, the estimation of point-wise robustness certificates for RS-enabled DNNs requires creating *hundreds of thousands* of samples (i.e., $n \approx 100000$), as demonstrated by prior results [10, 25, 5]. We counteract with existing know-how and demonstrate that computing such a point-wise robustness certificate for RS-enabled DNNs can be done with substantially fewer samples, with a small reduction in the size of the derived robustness radius while having the same confidence level. We provide the theoretical explanation of such a phenomenon based on the Central Limit Theorem (CLT) serving as an approximation to the binomial distribution (and thereby to the Clopper-Pearson interval calculation), integrating Shore's numerical approximation [27] on the inverse cumulative function of the normal distribution. The dual form of the theoretical result can be used as an *early stopping criterion* to know if adding additional samples can lead to a significantly increased radius matching the acceptance criterion. Finally, the theoretical results are confirmed by extensive experiments on the standard CIFAR-10 and ImageNet datasets, where reducing the number of samples by two to three orders of magnitude still allows for providing a point-wise robustness certificate of adequate radius.

In summary, the main contribution of this paper includes the following:

- An empirical evaluation of understanding the amount of perturbed inputs and its effect on the derived robustness radius for RS-enabled DNNs.
- A theoretical interpretation of the observed phenomenon and the mathematical formulation between the reduction/addition of samples and the decrease/increase of robustness radius.

2 Related Work

Robustness certification is an important aspect of the safe commissioning of learning-enabled systems. For DNNs, knowing the robustness bound can be done by viewing the DNN as a mathematical object (i.e., a program without loops). This enables formal verification techniques such as SMT [19], mixed-integer programming [8, 32], abstract interpretation [16, 15, 29], or convex relaxations [34]. While formal methods are computationally expensive, by viewing the DNN as a composition of functions, one can also apply the Lipchitz analysis [22, 30] to understand the impact of perturbation, thereby forming an estimate on the robustness bound. While these methods are directly applied to the DNN, they can also serve as a non-probabilistic lower bound for RS-enabled DNN, as the robustness bound computed by these methods ensures that all variations will share the same result, thereby creating an uncontested (i.e., predictions under all noises are the consistent) majority.

Randomized smoothing [10] currently represents the state-of-the-art sampling-based robustness prediction methods, as it scales to large DNNs used in practice and is agnostic to their architectural details. Moreover, RS has been extended to additional threat models beyond standard L_2 balls, such as general L_p norms [35], geometric transformations [12], segmentation [13] and more. Nevertheless, the amount of sample variations remains a practical concern. To this end, the closest approach to our work is in [7], where the authors studied the effect of using small sampling numbers (e.g., $n = 1000$); if the certified radius with the small sample size is larger than the accepted radius with high confidence, then there is no need to use more samples. They also qualitatively studied the effect of radius change concerning the sample size decrease from a constant. On the contrary, we start by considering the ideal case of infinite sampling (i.e., $n = \infty$) and formally derive the performance decrease when the sample size is small (e.g., $n = 1000$). Our proved theorems allow us to analytically estimate the potential increase in the point-wise robustness radius (Thm. 4.3) as well as the average radius increase in the whole domain (Thm. 4.5; subject to a certain shape commonly observed in the experiment) due to the increase of samples. The duality of this result enables an early stopping scheme when the increase of sample size is deemed ineffective in achieving the radius; their method will continue to adaptively increase the sample size without an end.

Finally, randomized smoothing is one of the driving forces for ensuring the robustness of prediction, where apart from majority votes, uncertainty can also be estimated. Other sampling-based methods such as MC-dropout [14] also sampling-based methods by randomly disabling some parameters. It is also interesting to know the robustness radius of these methods, which has not been explored to the best of our knowledge.

3 Preliminaries: Randomized Smoothing

Let $f : \mathbb{R}^d \to \{1, \ldots, K\}$ be a classifier mapping inputs $\mathbf{x} \in \mathbb{R}^d$ into K classes. In RS, f is replaced with the following classifier:

$$g_\sigma(\mathbf{x}) = \arg\max_y \mathbb{P}[f(\mathbf{x} + \mathbf{z}) = y], \mathbf{z} \sim N(0, \sigma^2 I) \qquad (1)$$

That is, g_σ perturbs the input $\mathbf{x}$ with noise $\mathbf{z}$ that follows an isotropic Gaussian distribution $N(0, \sigma^2 I)$, and returns the class $A \in \{1, \ldots, K\}$ that gets the majority vote, i.e., the one that f is most likely to return on the perturbed inputs.

Let $p_A(\mathbf{x}, \sigma) \overset{\text{def}}{:=} \mathbb{P}[f(\mathbf{x} + \mathbf{z}) = A], \mathbf{z} \sim N(0, \sigma^2 I)$ be the probability of the majority class being A, where we use the term p_A when the context is clear. If $p_A \geq 0.5$, then g_σ is guaranteed to be robust around $\mathbf{x}$, where we define the (guaranteed) *robustness radius* in such a situation as follows[1]:

$$R_\sigma \overset{\text{def}}{:=} \begin{cases} \sigma\Phi^{-1}(p_A) & \text{if } p_A \geq 0.5 \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

where Φ^{-1} is the inverse of the normal cumulative distribution function (CDF). The robustness radius R_σ ensures consistency of prediction, namely $\forall \mathbf{z} \in \mathbb{R}^d : ||\mathbf{z}||_2 \leq R_\sigma \to g_\sigma(\mathbf{x}) = g_\sigma(\mathbf{x} + \mathbf{z})$. The intuition is that a slight perturbation on $\mathbf{x}$ can change the output of f arbitrarily, but not the one of g_σ: since g_σ relies on the consensus over points distributed around $\mathbf{x}$, a small shift cannot change a distribution much. This is the crucial fact where RS resides.

Algorithm 1 Finding the robustness radius (adapted from [10])

```
1: **Input:** point $\mathbf{x}$, classifier $f$, $\sigma$, $n$, $\alpha$
2: **Output:** class $A$ and robustness radius $R_\sigma$ of $\mathbf{x}$
3: sample $n$ noisy samples $\mathbf{x}'_1, ..., \mathbf{x}'_n \sim N(\mathbf{x}, \sigma^2 I)$
4: get majority class $A \leftarrow \arg\max_y \sum_{i=1}^n \mathbf{1}[f(\mathbf{x}'_i) = y]$
5: counts($A$) $\leftarrow \sum_{i=1}^n \mathbf{1}[f(\mathbf{x}'_i) = A]$
6: $\bar{p_A} \leftarrow$ LowerConfBound(counts($A$), $n$, $\alpha$) {compute probability lower bound}
7: **if** $\bar{p_A} \geq \frac{1}{2}$ **then**
8:     return $A, \sigma\Phi^{-1}(\bar{p_A})$
9: **else**
10:    return ABSTAIN
11: **end if**
```

Finally, notice that finding the precise value of p_A is not possible; however, a lower bound $\bar{p_A}$ can be estimated by Monte Carlo sampling with a high degree of confidence $1 - \alpha$, as demonstrated in line 6 of Algo. 1 (technical details will be expanded in later sections). Yet, following existing results [10, 25, 5], the required number of samples n are typically around 10000 to 100000.

4 Reducing the Sample Size

In this section, we perform the theoretical analysis[2] concerning the effect of reducing the sample size in Algo. 1 on the robustness radius and average certified accuracy. The critical point to explore is the dependence of the lower bound $\bar{p_A}$ and the number of samples n (as well as the admissible error rate α).

4.1 The General Approach

Let p_A be probability of input $\mathbf{x}$ being class A, when being fed into g_σ (Eq. (1)). As p_A is unknown to us, the aim is to estimate it by drawing samples, with the goal of obtaining a lower bound $\bar{p_A}$ where the true probability is larger than $\bar{p_A}$ with confidence at least $1 - \alpha$. This interprets the meaning of Line 6 in Algo. 1.

More specifically, let $\mathbf{x}'_i \sim N(\mathbf{x}, \sigma^2 I)$ be noisy versions of $\mathbf{x}$ ($i = 1, ..., n$), and set $Y_i = \mathbf{1}[f(\mathbf{x}'_i) = A]$; Y_i is an indicator Random Variable (RV), taking the value 1 if $f(\mathbf{x}'_i)$ predicts the correct class A, and 0 otherwise. Notice that Y_i's are binomial RVs, with success probability p_A. Further, let $\hat{p} = \frac{Y_1 + \ldots + Y_n}{n}$ be the sample mean, i.e., the empirical estimate of p_A.

Given $\hat{p}$, n and α, there are many ways to estimate the lower bound as used line 6 of Algo. 1. One standard approach is to apply the Clopper-Pearson test [9] to obtain a lower bound, a term we call $\bar{p_A}^{CP}$. Unfortunately, while the Clopper-Pearson test gives us an exact lower bound for binomials, there is no analytical closed-form solution. Therefore, to obtain an algebraically tractable approximation, we apply the Central Limit Theorem (CLT) as our main

[1] In paving the theoretical result, starting Eq. (2), we follow prior RS works (e.g., [10]) that focus on the binary classification setting for simplicity, as it allows us to convey the key theoretical results more easily.

[2] For important results, the created theorems use ≈ (approximately equal) to omit the error terms introduced by numerical approximation; it is an easy exercise to precise all terms, but the resulting formula would be too complex for the reader to grasp the big picture.

tool [33], which states that the distribution of sample means approximates a normal distribution as the sample size gets larger ($n \geq 30$), with mean $\mathbf{E}[\hat{p}] = p_A$ and variance $\mathbf{Var}[\hat{p}] = \frac{p_A(1-p_A)}{n}$:

$$\hat{p} \sim N\left(p_A, \frac{p_A(1-p_A)}{n}\right) \tag{3}$$

Before proving the lower-bound, we now establish a lemma characterizing the approximation of expectations over functions.

Lemma 4.1. *Let X be an RV with finite mean and variance, and let f be a twice continuously differentiable function, with $|f''(x)| \leq 2M$ for all $x \in \mathbb{R}$. Then the following condition holds.*

$$f(\mathbf{E}[X]) - M \cdot \mathbf{Var}[X] \leq \mathbf{E}[f(X)] \leq f(\mathbf{E}[X]) + M \cdot \mathbf{Var}[X] \tag{4}$$

Proof. Since f is twice continuously differentiable on the open interval with f'' continuous on the closed interval between x_0 and x, using Taylor's theorem with the Lagrange form of remainder, one derives:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(\xi)(x - x_0)^2 \tag{5}$$

with $\xi \in (x_0, x)$. Since $|f''(x)| \leq 2M$ for all $x \in \mathbb{R}$, the above gives the following inequality:

$$\begin{aligned} f(x_0) + f'(x_0)(x - x_0) - M(x - x_0)^2 \leq f(x) \\ \leq f(x_0) + f'(x_0)(x - x_0) + M(x - x_0)^2 \end{aligned} \tag{6}$$

Now, replace x by X and x_0 by $\mathbf{E}[X]$ in Eq. (6), and take expectations on both sides, we get the following:

$$\begin{aligned} \mathbf{E}[f(\mathbf{E}[X])] + \mathbf{E}[f'(\mathbf{E}[X])(X - \mathbf{E}[X])] - \mathbf{E}[M(X - \mathbf{E}[X])^2] \\ \leq \mathbf{E}(f(X)) \\ \leq \mathbf{E}[f(\mathbf{E}[X])] + \mathbf{E}[f'(\mathbf{E}[X])(X - \mathbf{E}[X])] + \mathbf{E}[M(X - \mathbf{E}[X])^2] \end{aligned} \tag{7}$$

Eq. (7) can be simplified to Eq. (4), due to the following facts.

- $\mathbf{E}(f(\mathbf{E}[X])) = f(\mathbf{E}[X])$,
- $\mathbf{E}[f'(\mathbf{E}[X])(X - \mathbf{E}[X])] = f'(\mathbf{E}[X])\mathbf{E}[X - \mathbf{E}[X]] = f'(\mathbf{E}[X])(\mathbf{E}[X] - \mathbf{E}[X]) = 0$, and
- $\mathbf{E}[M(X - \mathbf{E}[X])^2] = M\mathbf{Var}[X]$

□

By applying the CLT via Eq. (3), we derive a lower-bound for p_A as follows:

Lemma 4.2. *Let $Y_1, ..., Y_n$ be Bernoulli RVs, with success probability p_A where $0 < p_l \leq p_A \leq p_h < 1$ with p_l, p_h constants, and $\hat{p} = \frac{Y_1 + ... + Y_n}{n}$. Assume $n \geq 30$ such that CLT holds. Then the following two conditions hold:*

1. *$\bar{p_A}^{CP} \approx \hat{p} - z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where $z_\alpha = \Phi^{-1}(1 - \frac{\alpha}{2})$ is the $1 - \frac{\alpha}{2}$ quantile of the normal distribution $N(0, 1)$.*
2. *$\mathbf{E}[\bar{p_A}^{CP}]$, i.e., the expected value of $\bar{p_A}^{CP}$, is equal to $p_A - z_\alpha\sqrt{\frac{p_A(1-p_A)}{n}} + \delta$, where $\delta \in [-c\mathbf{Var}[\hat{p}], c\mathbf{Var}[\hat{p}]]$ with c being a constant.*

Proof. The first item is the standard normal interval approximation for the binomial, under the CLT approximation [3]. For the second item, consider the function $f(p) = p - z_a\sqrt{\frac{p(1-p)}{n}}$. For $0 < p_l \leq p_A \leq p_h < 1$, $|f''(p)| = \frac{z_a}{4\sqrt{n}[p(1-p)]^{3/2}}$ is bounded by some constant c.

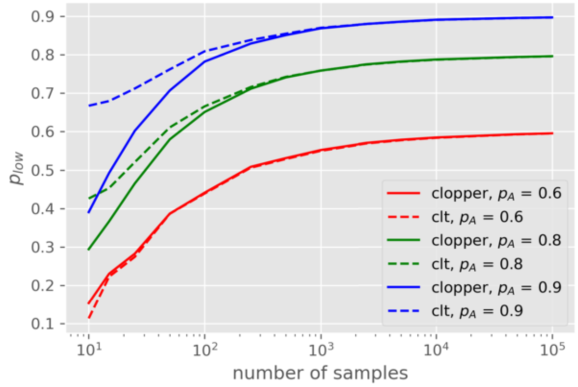


Figure 1: Plot of the average lower bounds for the Clopper-Pearson method and Lemma 4.2, obtained over 100 trials, for multiple values of p_A and n.

By taking Lemma 1 where X is assigned with $\hat{p}$ and M with c, we obtain

$$f(\mathbf{E}[\hat{p}]) - c\mathbf{Var}[\hat{p}] \leq \mathbf{E}[f(\hat{p})] \leq f(\mathbf{E}[\hat{p}]) + c\mathbf{Var}[\hat{p}] \tag{8}$$

By applying condition 1, using the definition of f, and applying Eq. (8), we obtain the following.

$$\begin{aligned} \mathbf{E}[\bar{p_A}^{CP}] \approx \mathbf{E}[\hat{p} - z_\alpha\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}] = \mathbf{E}[f(\hat{p})] \Rightarrow \\ \mathbf{E}_{\hat{p}}[f(\hat{p})] \in [f(\mathbf{E}[\hat{p}]) - c\mathbf{Var}[\hat{p}], f(\mathbf{E}[\hat{p}]) + c\mathbf{Var}[\hat{p}]] \end{aligned}$$

Finally, as $\mathbf{E}_{\hat{p}}[\hat{p}]$ equals p_A (following the definition in Eq. (3)), we can derive $\mathbf{E}_{\hat{p}}[\bar{p_A}^{CP}] \approx p_A - z_\alpha\sqrt{\frac{p_A(1-p_A)}{n}} + \delta$ where $\delta \in [-c\mathbf{Var}[\hat{p}], c\mathbf{Var}[\hat{p}]]$, establishing the validity of the second condition.

□

(Remark) In Lemma 4.2, the assumption on δ being negligible is reasonable in practice, e.g., $\delta \in [-0.0006, 0.0006]$ even for $p_A = 0.95$, with $n = 1000$.

Fig. 1 illustrates the lower-bound computed using Clopper-Pearson method and Lemma 4.2. For each point in the line, it is created by fixing the n value (number of samples), repeating the trial for 100 times, followed by taking the average. We observe that the distance between them is small, and decreases rapidly as n exceeds 100.

4.2 The Impact of Sample Size on Robustness Radius

We now study how the sample size influences the robustness radius. Let $R_\sigma^{\alpha,n}(p_A)$ be the expected robustness radius estimated with n samples and error rate α subject to noise level σ, provided that the original probability is p_A but being approximated by $\bar{p_A}^{CP}$. That is, $R_\sigma^{\alpha,n}(p_A) \overset{\text{def}}{:=} \mathbf{E}[\sigma\Phi^{-1}(\bar{p_A}^{CP})]$. Ideally, if we had an infinite number of samples, $\bar{p_A}^{CP}$ will be equal to p_A, and the robustness radius would be, by Eq. (2): $R_\sigma = R_\sigma^{0,\infty}(p_A) = \sigma\Phi^{-1}(p_A)$. However, for a finite number of samples, we do not have access to p_A, but to a lower bound of it, as we saw before.

To study how much the robustness radius, $R_\sigma^{\alpha,n}(p_A)$, drops due to the reduction of finite samples, we use the following approximation stated in Eq. (9) for $\Phi^{-1}(p)$, which is valid for $p \geq \frac{1}{2}$ [27]. In our case, the validity assumption for $p \geq \frac{1}{2}$ is not a problem, as otherwise, if $p_A < \frac{1}{2}$, the robustness radius is 0, implying that g_σ failed to predict the correct class (which is by common sense, undesired).

$$\Phi^{-1}(p) \approx \frac{1}{0.1975}[p^{0.135} - (1-p)^{0.135}] \tag{9}$$

By applying Lemma 4.2 followed by integrating Eq. (9), we can examine the influence n has on $R_\sigma^{\alpha,n}(p_A)$, which leads to the following theorem.

Theorem 4.3. *For an RS classifier g_σ, given a point $\mathbf{x}$, assume that we estimate p_A by drawing n samples and compute the lower bound from the empirical $\hat{p}$ with confidence $1-\alpha$. Assume that the conditions in Lemma 4.2 hold, while $|\frac{d^2\Phi^{-1}(p)}{dp^2}|\mathbf{Var}[\hat{p}]$ and δ in Lemma 4.2 are negligible. Then*

$$R_\sigma^{\alpha,n}(p_A) \approx \sigma\Phi^{-1}(p_A - t_{\alpha,n}) \tag{10}$$

where $t_{\alpha,n} = z_\alpha\sqrt{\frac{p_A(1-p_A)}{n}}$. In addition, $R_\sigma^{\alpha,n}(p_A)$ is approximately equal to:

$$\begin{aligned} R_\sigma^{\alpha,n}(p_A) \approx 5.063\sigma[p_A^{0.135} - (1-p_A)^{0.135} - \\ 0.135\frac{z_\alpha}{\sqrt{n}}(p_A^{-0.365}(1-p_A)^{1/2} + p_A^{1/2}(1-p_A)^{-0.365})] \end{aligned} \tag{11}$$

The practical usage of Thm. 4.3 occurs when we have already used n samples to compute the radius. With σ being a constant, under a given specified α, one can reverse-engineer to uncover the appropriate p_A value making Eq. (11) holds. This allows us to predict the robustness radius increase when one changes from n to substantially larger values such as $100n$. The maximum increase occurs when $n \to \infty$, where the robustness radius is approximately equal to the term in Eq. (12).

$$R_\sigma^{0,\infty}(p_A) \approx 5.063\,\sigma[p_A^{0.135} - (1-p_A)^{0.135}] \tag{12}$$

Proof. As the condition of Lemma 4.2 holds, $\bar{p_A}^{CP} \approx \hat{p} - t_{\alpha,n}$. Using Eq. (4), we get

$$\begin{aligned} &\sigma\Phi^{-1}(\mathbf{E}[\bar{p_A}^{CP}]) - M\mathbf{Var}[\hat{p}] \\ &\leq R_\sigma^{\alpha,n}(p_A) = \mathbf{E}[\sigma\Phi^{-1}(\bar{p_A}^{CP})] \\ &\sigma\Phi^{-1}(\mathbf{E}[\bar{p_A}^{CP}]) + M\mathbf{Var}[\hat{p}] \end{aligned} \tag{13}$$

where M is the upper bound of $|\frac{d^2\Phi^{-1}(p)}{dp^2}|$ in the interval $[p_l, p_h)$. As $|\frac{d^2\Phi^{-1}(p)}{dp^2}|\mathbf{Var}[\hat{p}]$ is negligible, we have:

$$R_\sigma^{\alpha,n}(p_A) = \mathbf{E}[\sigma\Phi^{-1}(\bar{p_A}^{CP})] \approx \sigma\Phi^{-1}(\mathbf{E}[\bar{p_A}^{CP}]) \tag{14}$$

By applying the second condition of Thm 4.2 and with δ being negligible, we have the following.

$$R_\sigma^{\alpha,n}(p_A) \approx \sigma\Phi^{-1}(\mathbf{E}_{\hat{p}}[\bar{p_A}^{CP}]) \approx \sigma\Phi^{-1}(p_A - t_{\alpha,n}) \tag{15}$$

Next, we replace Φ^{-1} by the approximation of Eq. (9), and we get:

$$R_\sigma^{\alpha,n}(p_A) \approx \sigma\frac{1}{0.1975}[(p_A - t_{\alpha,n})^{0.135} - (1 - p_A + t_{\alpha,n})^{0.135}] \tag{16}$$

For further simplification, we use binomial theorem, $(1+x)^a = 1 + ax + \frac{a(a-1)}{2!}x^2 + ...$ valid for $|x| < 1$ on both terms of Eq. (16), and keep only the 1st order terms. Doing that gives:

$$\begin{aligned} A &\overset{\text{def}}{:=} \left(p_0 - z_\alpha\sqrt{\frac{p_A(1-p_A)}{n}}\right)^{0.135} \\ &= p_A^{0.135}\left(1 - \frac{z_\alpha}{\sqrt{n}}p_A^{-1/2}(1-p_A)^{1/2}\right)^{0.135} \Rightarrow \\ A &\approx p_A^{0.135}(1 - 0.135\frac{z_\alpha}{\sqrt{n}}p_A^{-1/2}(1-p_A)^{1/2}) = p_A^{0.135} \\ &- 0.135\frac{z_\alpha}{\sqrt{n}}p_A^{-0.365}(1-p_A)^{1/2} \\ B &\overset{\text{def}}{:=} \left(1 - p_A + z_\alpha\sqrt{\frac{p_A(1-p_A)}{n}}\right)^{0.135} = \\ &(1-p_A)^{0.135}\left(1 + \frac{z_\alpha}{\sqrt{n}}p_A^{1/2}(1-p_A)^{-1/2}\right)^{0.135} \Rightarrow \\ B &\approx (1-p_A)^{0.135}(1 + 0.135\frac{z_\alpha}{\sqrt{n}}p_A^{1/2}(1-p_A)^{-1/2}) \\ &= (1-p_A)^{0.135} + 0.135\frac{z_\alpha}{\sqrt{n}}p_A^{1/2}(1-p_A)^{-0.365} \end{aligned} \tag{17}$$

Substituting in Eq. (16) and combining terms gets Eq. (11).

□

(Remark) In Thm. 4.3, the assumption on $|\frac{d^2\Phi^{-1}(p)}{dp^2}|\mathbf{Var}[\hat{p}]$ being negligible is reasonable, as $\mathbf{Var}[\hat{p}] \overset{\text{def}}{:=} \frac{p_A(1-p_A)}{n}$, and when n is around 1000, the value can at most be 0.00025. The second derivative of inverse normal CDF $|\frac{d^2\Phi^{-1}(p)}{dp^2}|$, when p is not too close to 1, is reasonably sized. For example, when $p = 0.9$, $|\frac{d^2\Phi^{-1}(p)}{dp^2}| = 27.77$, making the product term $|\frac{d^2\Phi^{-1}(p)}{dp^2}|\mathbf{Var}[\hat{p}] = 0.0069$ still small. We observe in the experiments that even when n is not very big (cf. Sec. 5), the approximation and the observed behavior remain similar.

4.3 Average Robustness Radius Drop

In the previous section, we examine the effect of n on the robustness radius for one specific point. But ultimately, what really interests us is the effect on the average robustness radius over the entire dataset. This is the average robustness radius that we expect to lose by reducing the number of samples.

For any particular point, the robustness radius depends on p_A, the probability that g_σ outputs the correct class A (Eq. (2)). To answer questions about the average behavior of the robustness radius, we need to consider the probability distribution of p_A: we devote the probability density function (pdf) of p_A to be $\Pr(p_A)$. We can roughly imagine $\Pr(p_A)$ as a histogram over the p_A values obtained from our dataset.

Formally, the average robustness radius is then given by Eq. (18). The integration can start at 0.5, as the robustness radius $R_\sigma^{\alpha,n}(p_A)$ equals 0 for $p_A < 0.5$ (cf. Eq. (2)).

$$\begin{aligned} \bar{R}_\sigma(\alpha, n) \overset{\text{def}}{:=} \mathbf{E}_{\Pr(p_A)}[R_\sigma^{\alpha,n}(p_A)] \\ = \int_0^1 R_\sigma^{\alpha,n}(p_A)\Pr(p_A)dp_A = \int_{0.5}^1 R_\sigma^{\alpha,n}(p_A)\Pr(p_A)dp_A \end{aligned} \tag{18}$$

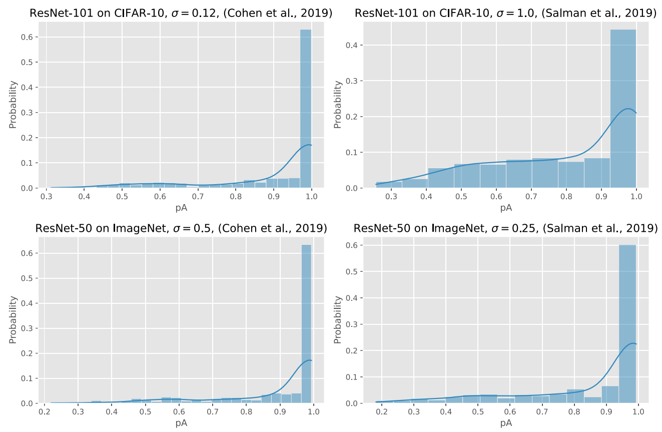


Figure 2: Plots of histograms and density plots of p_0 obtained for different models and datasets, as shown in the figure titles. The values of p_0 were estimated empirically using $n = 100000$ samples.

Unfortunately, $\Pr(p_A)$ depends heavily on the particular model and dataset used, and does not seem to follow any well-known family of distributions such as Gaussian. This can be seen in Fig. 2, where we estimate the histogram of p_A for different models of [10] and [25]. Nevertheless, we notice that $\Pr(p_A)$ is skewed towards 1 in all cases we tested: namely, most of the mass of $\Pr(p_A)$ is concentrated in a small interval $(\beta, 1)$ on the right, while the mass outside it - and especially in the interval $[0, 0.5]$ is close to zero. Intuitively, this is the behavior we expect, where most of the data points in the sample are far from the decision boundary; otherwise, the robustness radius of a point would be very small.

Inspired by the shape of the distribution as illustrated in Fig. 2, we form a distribution as characterized in Lemma 4.4, where κ_1 and κ_2 shall be small, similar to that of Fig. 2.

Lemma 4.4. *Assume that* $\Pr(p_A)$ *follows a piecewise uniform distribution across input points* $\mathbf{x}$, *where*

$$\Pr(p_A) \overset{\text{def}}{:=} \begin{cases} \kappa_1 & \text{if } p_A \in [0, 0.5) \\ \kappa_2 & \text{if } p_A \in [0.5, \beta) \\ \kappa_3 \overset{\text{def}}{:=} \frac{1-\kappa_1(0.5)-\kappa_2(\beta-0.5)}{1-\beta} & \text{if } p_A \in [\beta, 1) \end{cases} \quad (19)$$

Then $\bar{R}_\sigma(\alpha, n) = \kappa_2 \int_{0.5}^{\beta} R_\sigma^{\alpha,n}(p_A)dp_A + \kappa_3 \int_\beta^1 R_\sigma^{\alpha,n}(p_A)dp_A$

Proof. The proof immediately follows Eq. (18) and (19). □

By integrating Eq. (16) into Lemma 4.4, with κ_1, κ_2 and β being constants, the integrals can be calculated. In the following, we exemplify the computation by computing the result of robustness radius drop when $\kappa_1 = 0$ and $\kappa_2 = 0$, a case where $\Pr(p_A)$ is uniform in the interval $[\beta, 1)$ with cases $\beta \geq 0.8$ and $\beta = 0.5$.

Theorem 4.5. *Assume that* $\Pr(p_A)$ *follows a uniform distribution in the interval* $[\beta, 1)$ *across data points* $\mathbf{x}$ *in the input domain The decrease of the average certified radius* $\bar{R}_\sigma(\alpha, n)$ *using* n *samples from the ideal case of* $n = \infty$ *is approximately equal to:*

$$r_\sigma(\alpha, n) := \frac{\bar{R}_\sigma(\alpha, n)}{\bar{R}_\sigma(0, \infty)} \approx 1 - \Theta \frac{z_\alpha}{\sqrt{n}} \quad (20)$$

where

$$\Theta \overset{\text{def}}{:=} \begin{cases} 1.64 & \text{if } \beta \in [0.8, 1) \\ 2 & \text{if } \beta = 0.5 \end{cases} \quad (21)$$

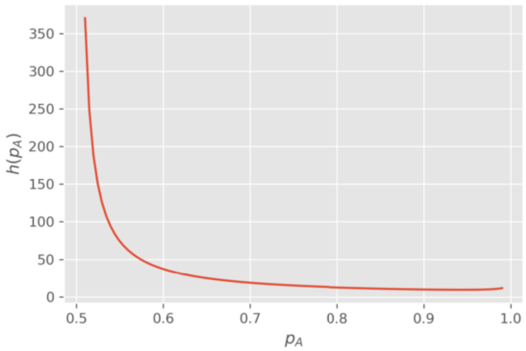


Figure 3: Plot of $h(p_A)$ in the interval $[0.5, 1]$

Proof. We proceed with the proof by separating cases.
[Case 1 - $\beta \in [0.8, 1)$] Recall that Eq. (11) gives us $R_\sigma^{\alpha,n}(p_A)$ for a particular point with class probability p_A, while Eq. (12) provides the result of $R_\sigma^{0,\infty}(p_A)$. Consider the ratio characterized below.

$$\frac{R_\sigma^{\alpha,n}(p_A)}{R_\sigma^{0,\infty}(p_A)} = 1 - 0.135 \frac{z_\alpha}{\sqrt{n}} h(p_A) \quad (22)$$

where

$$h(p_A) = \frac{p_A^{-0.365}(1-p_A)^{1/2} + p_A^{1/2}(1-p_A)^{-0.365}}{p_A^{0.135} - (1-p_A)^{0.135}} \quad (23)$$

Crucially, $h(p_A)$ is almost constant within an interval close to 1, as illustrated in Fig. 3. For instance, in the interval $(\beta, 1)$ with $\beta \geq 0.8$, we find $h(p_A) \approx 12.14$. Substituting this value inside Eq. (22), we obtain:

$$\frac{R_\sigma^{\alpha,n}(p_A)}{R_\sigma^{0,\infty}(p_A)} \approx 1 - 1.64 \frac{z_\alpha}{\sqrt{n}} \quad (24)$$

Therefore:
$$\begin{aligned} \bar{R}_\sigma(\alpha, n) &= \int_0^1 R_\sigma^{\alpha,n}(p_A) \Pr(p_A) dp_A \\ &\approx (1 - 1.64 \frac{z_\alpha}{\sqrt{n}}) \int_\beta^1 R_\sigma^{0,\infty}(p_A) \Pr(p_A) dp_A \\ &= (1 - 1.64 \frac{z_\alpha}{\sqrt{n}}) \int_0^1 R_\sigma^{0,\infty}(p_A) \Pr(p_A) dp_A \\ &= (1 - 1.64 \frac{z_\alpha}{\sqrt{n}}) \bar{R}_\sigma(0, \infty) \end{aligned} \quad (25)$$

In Eq. (25), the equality of expanding the integral from $\int_\beta^1$ to $\int_0^1$ comes from the fact that $\Pr(p_A) = 0$ when $p_A \in [0, \beta)$. As $\int_\beta^1 R_\sigma^{0,\infty}(p_A) \Pr(p_A) dp_A$ is exactly the definition of $\bar{R}_\sigma(0, \infty)$, we obtain the required formula. Interestingly, the derivation in this case holds for density functions $Pr[p_A]$ in $[\beta, 1)$ of any form.

[Case 2 - $\beta = 0.5$] When $\beta = 0.5$, $\Pr(p_A) = 2$ when $p_A \in [0, 1)$. For the average robustness radius, we get:

$$\bar{R}_\sigma(\alpha, n) = \int_0^1 R_\sigma^{\alpha,n}(p_A) \Pr(p_A) dp_A = 2 \int_{0.5}^1 R_\sigma^{\alpha,n}(p_A) dp_A \quad (26)$$

Substituting Eq. (11), we can perform the integration and obtain:

$$\bar{R}_\sigma(\alpha, n) = 2 \int_{0.5}^1 5.063\sigma[p_A^{0.135} - (1-p_A)^{0.135} \\ -0.135 \frac{z_\alpha}{\sqrt{n}} (p_A^{-0.365}(1-p_A)^{1/2} + p_A^{1/2}(1-p_A)^{-0.365})] dp_A \quad (27)$$

The integrals of the form p_A^a and $(1 - p_A)^a$ can be computed easily, while the integrals of the terms $p_A^a(1 - p_A)^b$ are integrals of the Beta function, and can be evaluated numerically. Based on the calculations, we get:

$$\bar{R}_\sigma(\alpha, n) = \sigma \left(0.796 - 1.603 \frac{z_\alpha}{\sqrt{n}}\right) \tag{28}$$

Finally, diving the terms, we see that the ratio of robustness radius drop is independent of σ, and is approximately equal to:

$$\frac{\bar{R}_\sigma(\alpha, n)}{\bar{R}_\sigma(0, \infty)} \approx 1 - 2\frac{z_\alpha}{\sqrt{n}} \tag{29}$$

which concludes the proof. □

(Observation) (i) Note that even when one changes the distribution and falls back to the form in Eq. (19), the computation in Eq. (27) will be changed, but in the final computation of $\frac{\bar{R}_\sigma(\alpha,n)}{\bar{R}_\sigma^\infty}$, σ appearing on both the denominator and the numerator will be canceled out, implying that the average robustness radius drop is *independent of the noise level* σ. (ii) The two cases we have demonstrated hint at a tendency where Θ shall fall inside the interval $[1.64, 2]$, when the distribution adjusts from uniform to a small peak.

4.4 Certified Accuracy Drop

Apart from the average robustness radius, another important quantity in evaluating robust classifiers is the average certified accuracy: We denote it by acc_R, which is the fraction of points that g_σ classifies correctly, and with robustness radius at least larger than a threshold R.

To understand the situation, consider again the distribution of $\Pr(p_A)$, and assume that we are evaluating acc_{R_0} for some radius of interest R_0. By Eq. (2), this corresponds to a probability p_0:

$$R_0 = \sigma\Phi^{-1}(p_0) \Leftrightarrow p_0 = \Phi(R_0/\sigma) \tag{30}$$

By applying similar simplifying assumptions on $\Pr(p_A)$ as in Thm. 4.5, we can obtain an upper bound on the certified accuracy drop. One can easily extend the result to accommodate the distribution characterized in Eq. (19).

Theorem 4.6. *Let $acc_{R_0}(\alpha, n)$ be the certified accuracy g_σ obtains using n samples and error rate α, and let acc_{R_0} be the ideal case where $n = \infty$. Assume that $\Pr(p_A)$ follows a uniform distribution in the interval $[0.5, 1)$ across input points $\mathbf{x}$. The certified accuracy drop, $\Delta acc_{R_0}(\alpha, n) = acc_{R_0} - acc_{R_0}(\alpha, n)$ satisfies:*

$$\Delta acc_{R_0}(\alpha, n) \leq \frac{z_\alpha}{\sqrt{n}} \tag{31}$$

Proof. Let $p_0 = \Phi(R_0/\sigma)$; then, for acc_{R_0} we have that:

$$acc_{R_0} = \int_{p_0}^{1} \Pr(p_A) dp_A \tag{32}$$

Nevertheless, when we use n samples, we can measure only the $(1-\alpha)$-lower bound of p_A, which, by Theorem 4.2, is approximately equal to: $\bar{p_A}^{CP} = p_A - t_{\alpha,n}$.

So, now a point will be included in the integration if we have $\bar{p_A}^{CP} \geq p_0$. Via syntactic rewriting, we have

$$\bar{p_A}^{CP} \geq p_0 \Rightarrow p_A - t_{\alpha,n} \geq p_0 \Rightarrow p_A \geq p_0 + t_{\alpha,n} \tag{33}$$

Table 1: Average robustness radius for each noise level σ and sample size n on CIFAR-10, for the models of [10] (with $\alpha = 0.001$)

σ/n	25	50	100	250	500	1000	10000
0.12	0.055	0.091	0.121	0.154	0.174	0.192	0.238
0.25	0.09	0.152	0.203	0.258	0.292	0.319	0.385
0.50	0.119	0.206	0.276	0.354	0.395	0.43	0.501
1.00	0.111	0.202	0.284	0.371	0.41	0.448	0.513

Table 2: Average robustness radius for each noise level σ and sample size n on ImageNet, for the models of [10] (with $\alpha = 0.001$)

σ/n	25	50	100	250	500	1000	10000
0.25	0.095	0.155	0.208	0.266	0.301	0.333	0.477
0.50	0.148	0.241	0.326	0.414	0.471	0.602	0.734
1.00	0.189	0.313	0.425	0.537	0.603	0.663	0.875

For $t_{\alpha,n}$ we notice that:

$$t_{\alpha,n} = z_\alpha \sqrt{\frac{p_A(1 - p_A)}{n}} \Rightarrow t_{\alpha,n} \leq \frac{z_\alpha}{2\sqrt{n}} \tag{34}$$

since the quantity $p_A(1 - p_A)$ with $p_A \in [0, 1]$ is maximized for $p_A = 0.5$, and has value $1/4$.

Hence, all points satisfying $p_A \geq p_0 + \frac{z_\alpha}{2\sqrt{n}}$ will be included in the integration, and the interval that will be excluded will be at most $[p_0, p_0 + \frac{z_\alpha}{2\sqrt{n}}]$. So, we finally obtain:

$$\Delta acc_{R_0}(\alpha, n) \leq \int_{p_0}^{1} \Pr(p_A) dp_A - \int_{p_0 + \frac{z_\alpha}{2\sqrt{n}}}^{1} \Pr(p_A) dp_A \Rightarrow$$

$$\Delta acc_{R_0}(\alpha, n) \leq \int_{p_0}^{p_0 + \frac{z_\alpha}{2\sqrt{n}}} \Pr(p_A) dp_A \tag{35}$$

Under the assumption that $\Pr(p_A)$ is uniform in $[0.5, 1)$, we have $\Pr(p_A) = 2$, and the last integral is simply $\frac{z_\alpha}{2\sqrt{n}} \times 2$, which yields the required result. □

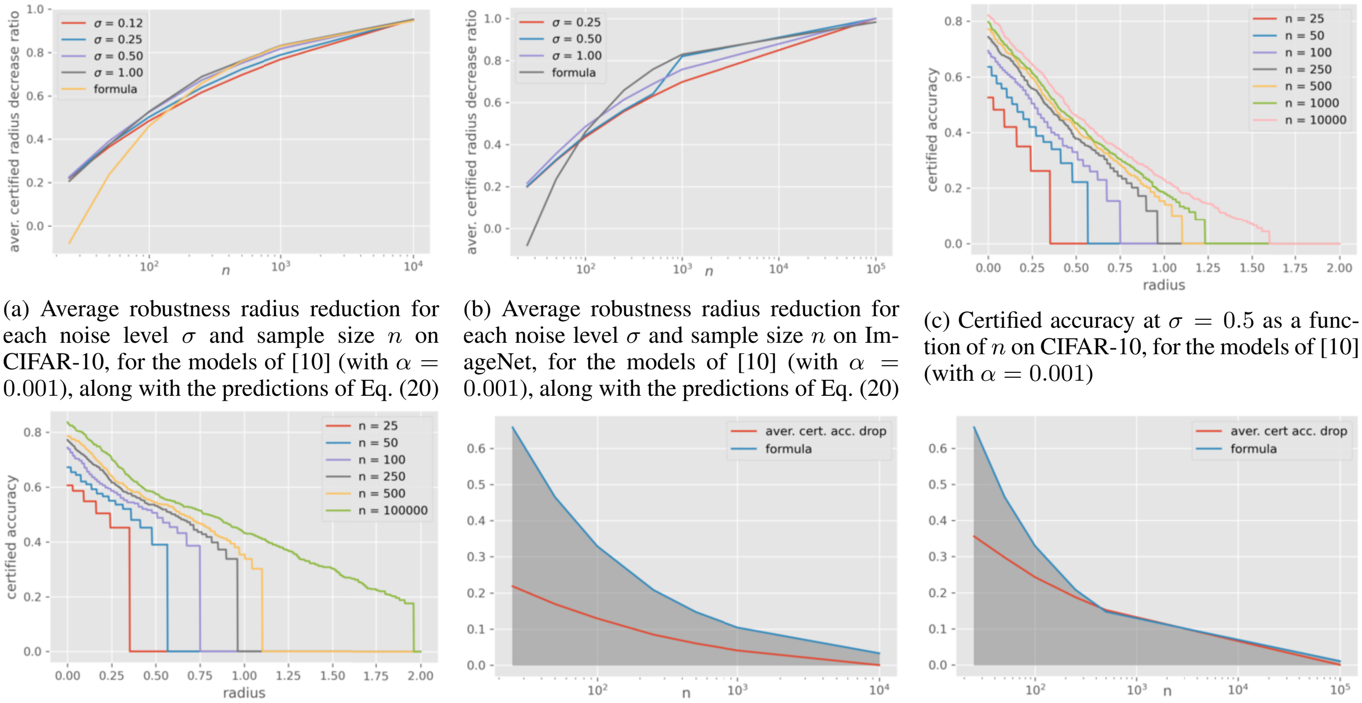
5 Evaluation

In this section, we analyze experimentally the influence of n on the average robustness radius and certified accuracy, and compare the theoretical results of Sec. 4 with the actual measurements by running Algo. 1. We work with the standard CIFAR-10 [20] and ImageNet [11] datasets, as done in several seminal papers on RS [10, 25, 5].

To measure the influence of n (sample size), we take the classifiers from Cohen et al. [10] made available for repeatability evaluation. They have trained different models that work best with the corresponding σ. We fix $\alpha = 0.001$ which is the typical value, and measure the average robustness radius as a function of n, for CIFAR-10 and ImageNet. We subsample every 20-th example in CIFAR-10 and every 100-th in ImageNet, following the protocol of [10]. The results can be seen in Table 1 and 2. To inspect these also visually, we plot the measured dependency of $\bar{R}_\sigma(\alpha, n)$ from n for the different σ values, along with the predictions of Eq. (20). For each experiment, we approximate the ratio $r_\sigma(\alpha, n)$ using as $\bar{R}_\sigma^\infty$ the value we obtain for $n = 100000$ samples. The results are shown in Fig. 4a and Fig. 4b.

From Table 1 and 2, we observe that the reduced sample sizes do not decrease the average robustness radius $\bar{R}_\sigma(\alpha, n)$ as much as expected: for example, in the case of CIFAR-10, a 10× decrement (from 10000 to 1000) reduces $\bar{R}$ by only around 20% across noise levels σ. Moreover, a 100× decrement reduces $\bar{R}$ by only 50%. Similarly, for the case of ImageNet, a reduction of n from 100000 to 100 reduces $\bar{R}$ by merely 50%.

Analyzing the measurements across all datasets and models, the results well-support the theoretical bound characterized by Thm. 4.5.



(a) Average robustness radius reduction for each noise level σ and sample size n on CIFAR-10, for the models of [10] (with $\alpha = 0.001$), along with the predictions of Eq. (20)

(b) Average robustness radius reduction for each noise level σ and sample size n on ImageNet, for the models of [10] (with $\alpha = 0.001$), along with the predictions of Eq. (20)

(c) Certified accuracy at $\sigma = 0.5$ as a function of n on CIFAR-10, for the models of [10] (with $\alpha = 0.001$)

(d) Certified accuracy at $\sigma = 0.5$ as a function of n on ImageNet, for the models of [10] (with $\alpha = 0.001$)

(e) Plot of average certified accuracy drop for the models of [10], at $\sigma = 0.5$, along with the predictions of Eq. (31) (CIFAR-10).

(f) Plot of average certified accuracy drop for the models of [10], at $\sigma = 0.5$, along with the predictions of Eq. (31) (ImageNet).

Figure 4: Evaluation results

First, the radius drop is independent of the noise level σ; indeed, in the experiments we found approximately the same radius reduction across different σ values for each dataset. Second, we observe that the reduction of $\bar{R}_\sigma(\alpha, n)$ from $n = 10000$ to $n = 1000$ is around $\approx 85\%$, which is what we see in the experiments. Similarly, the formula shows that there is little difference for $n = 10000$ and $n = 100000$ also in agreement with the observations. On the other hand, the predicted reduction as we shift n from 10000 to 100 is around 48%, which is slightly larger than the one we saw in experiments. This is to be expected, as Eq. (20) captures the tendency of having a burst in the distribution (cf. Fig. 2), and is "unaware" of the specific model and dataset details; recall that for every dataset and every value of σ, there is a corresponding distinct classifier provided by [10]. Thus, Eq. (20) delivers decent predictions among 2 datasets across 7 different models.

Next, we want to investigate the effect of n on certified accuracy. For that, we measure the certified accuracy of the models of [10] on CIFAR-10 and ImageNet: We choose one model and plot the certified accuracy curve for each value of n. The results are shown in Fig. 4c and Fig. 4f,which show the certified accuracy of a model for each given radius R_0, $acc_{R_0}(\alpha, n)$.

We observe that the distance between the curves (e.g., the robustness radius drop) is roughly constant, until a curve drops to zero, in accordance with Eq. (31). Further, in order to compare the predictions of Eq. (31) with reality, we plot the average certified accuracy drop, averaged over radii, and compare it to the theoretically expected value. This is illustrated in Fig. 4e and 4f. The predictions of Eq. (31) form a "conservative envelope" in Fig. 4e and 4f; they are larger than the certified accuracy drop observed in practice. Notice that there is no strong guarantee that this must be so, as Thm. 4.6 relies on some simplifying assumptions that may not hold universally for all models and datasets; however, what we are mostly interested in is predicting the general trend, which Eq. (31) seems to capture adequately.

6 Conclusion

In this paper, we addressed the challenge of the sample requirements for robustness certification based on randomized smoothing (RS). Our investigation revealed that significant reductions in the number of samples have a less severe impact on the average robustness radius and certified accuracy than previously anticipated. Through detailed empirical and theoretical analysis across multiple model architectures trained on CIFAR-10 and ImageNet, we observed consistent behavior.

Looking ahead, we see promising opportunities for applying our RS speedup techniques in AI safety. One potential application is in the robustness profiling of classifiers. Typically, such assessments would require testing with 100000 samples per input data point; however, our methodology could significantly lower the cost by enabling robustness assessments with as few as 100 samples, subsequently extrapolating to larger datasets. Moreover, the emergence of foundational models and their availability via API calls presents another area ripe for application. These models, whether Large Language Models (LLMs) or Vision-Language Models (VLMs), are susceptible to adversarial attacks. Our approach could feasibly enable robustness verification for each query without the need for 100000 samples, a previously untenable requirement.

References

[1] Stephanie Abrecht, Alexander Hirsch, Shervin Raafatnia, and Matthias Woehrle, 'Deep learning safety concerns in automated driving perception', *arXiv preprint arXiv:2309.03774*, (2023).

[2] Saddek Bensalem, Panagiotis Katsaros, Dejan Ničković, Brian Hsuan-Cheng Liao, et al., 'Continuous engineering for trustworthy learning-enabled autonomous systems', in *International Conference on Bridging the Gap between AI and Reality (AISoLA)*, pp. 256–278. Springer, (2023).

[3] Lawrence D Brown, T Tony Cai, and Anirban DasGupta, 'Interval estimation for a binomial proportion', *Statistical science*, **16**(2), 101–133, (2001).

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al., 'Language models are few-shot learners', *Advances in Neural Information Processing Systems (NeurIPS)*, **33**, 1877–1901, (2020).

[5] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter, '(certified!!) adversarial robustness for free!', in *International Conference on Learning Representations (ICLR)*, (2022).

[6] Long Chen, Shaobo Lin, Xiankai Lu, Dongpu Cao, Hangbin Wu, Chi Guo, Chun Liu, and Fei-Yue Wang, 'Deep neural network based vehicle and pedestrian detection for autonomous driving: A survey', *IEEE Transactions on Intelligent Transportation Systems*, **22**(6), 3234–3246, (2021).

[7] Ruoxin Chen, Jie Li, Junchi Yan, Ping Li, and Bin Sheng, 'Input-specific robustness certification for randomized smoothing', in *AAAI Conference on Artificial Intelligence*, volume 36, pp. 6295–6303, (2022).

[8] Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess, 'Maximum resilience of artificial neural networks', in *International Symposium on Automated Technology for Verification and Analysis (ATVA)*, pp. 251–268. Springer, (2017).

[9] Charles J Clopper and Egon S Pearson, 'The use of confidence or fiducial limits illustrated in the case of the binomial', *Biometrika*, **26**(4), 404–413, (1934).

[10] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter, 'Certified adversarial robustness via randomized smoothing', in *International Conference on Machine Learning (ICML)*, pp. 1310–1320. PMLR, (2019).

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, 'ImageNet: A large-scale hierarchical image database', in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, (2009).

[12] Marc Fischer, Maximilian Baader, and Martin Vechev, 'Certified defense to image transformations via randomized smoothing', *Advances in Neural Information Processing Systems (NeurIPS)*, **33**, 8404–8417, (2020).

[13] Marc Fischer, Maximilian Baader, and Martin Vechev, 'Scalable certified segmentation via randomized smoothing', in *International Conference on Machine Learning (ICML)*, pp. 3340–3351. PMLR, (2021).

[14] Yarin Gal and Zoubin Ghahramani, 'Dropout as a bayesian approximation: Representing model uncertainty in deep learning', in *International Conference on Machine Learning (ICML)*, pp. 1050–1059. PMLR, (2016).

[15] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev, 'Ai2: Safety and robustness certification of neural networks with abstract interpretation', in *IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, (2018).

[16] 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).

[17] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, 'Speech recognition with deep recurrent neural networks', in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6645–6649. IEEE, (2013).

[18] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu, 'A survey of deep learning techniques for autonomous driving', *Journal of Field Robotics*, **37**(3), 362–386, (2020).

[19] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer, 'Reluplex: An efficient smt solver for verifying deep neural networks', in *International Conference on Computer Aided Verification (CAV)*, pp. 97–117. Springer, (2017).

[20] Alex Krizhevsky, Geoffrey Hinton, et al., 'Learning multiple layers of features from tiny images', (2009).

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, 'ImageNet classification with deep convolutional neural networks', *Communications of the ACM*, **60**(6), 84–90, (2017).

[22] Klas Leino, Zifan Wang, and Matt Fredrikson, 'Globally-robust neural networks', in *International Conference on Machine Learning (ICML)*, pp. 6212–6222. PMLR, (2021).

[23] Sajjad Mozaffari, Omar Y Al-Jarrah, Mehrdad Dianati, Paul Jennings, and Alexandros Mouzakitis, 'Deep learning-based vehicle behavior prediction for autonomous driving applications: A review', *IEEE Transactions on Intelligent Transportation Systems*, **23**(1), 33–47, (2020).

[24] Ashish Pandharipande, Chih-Hong Cheng, Justin Dauwels, Sevgi Z Gurbuz, Javier Ibanex-Guzman, Guofa Li, Andrea Piazzoni, Pu Wang, and Avik Santra, 'Sensing and machine learning for automotive perception: A review', *IEEE Sensors Journal*, (2023).

[25] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang, 'Provably robust deep learning via adversarially trained smoothed classifiers', *Advances in Neural Information Processing Systems (NeurIPS)*, **32**, (2019).

[26] Philipp Schleiss, Francesco Carella, and Iwo Kurzidem, 'Towards continuous safety assurance for autonomous systems', in *International Conference on System Reliability and Safety (ICSRS)*, pp. 457–462. IEEE, (2022).

[27] Haim Shore, 'Simple approximations for the inverse cumulative function, the density function and the loss integral of the normal distribution', *Journal of the Royal Statistical Society Series C: Applied Statistics*, **31**(2), 108–114, (1982).

[28] David Silver, Thomas Hubert, Julian Schrittwieser, et al., 'A general reinforcement learning algorithm that masters chess, shogi, and go through self-play', *Science*, **362**(6419), 1140–1144, (2018).

[29] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev, 'An abstract domain for certifying neural networks', *Proceedings of the ACM on Programming Languages*, **3**(POPL), 1–30, (2019).

[30] Sahil Singla, Surbhi Singla, and Soheil Feizi, 'Improved deterministic l2 robustness on cifar-10 and cifar-100', in *International Conference on Learning Representations (ICLR)*, (2021).

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, 'Intriguing properties of neural networks', *International Conference of Learning Representations (ICLR)*, (2014).

[32] Vincent Tjeng, Kai Xiao, and Russ Tedrake, 'Evaluating robustness of neural networks with mixed integer programming', *International Conference of Learning Representations (ICLR)*, (2019).

[33] Larry Wasserman, *All of statistics: a concise course in statistical inference*, volume 26, Springer, 2004.

[34] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter, 'Scaling provable adversarial defenses', *Advances in Neural Information Processing Systems (NeurIPS)*, **31**, (2018).

[35] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li, 'Randomized smoothing of all shapes and sizes', in *International Conference on Machine Learning (ICML)*, pp. 10693–10705. PMLR, (2020).