Randomized Smoothing (almost) in Real Time?

Emmanouil Seferis [1] **Simon Burton** [1] **Stefanos Kollias** [2]

Abstract

Certifying the robustness of Deep Neural Networks (DNNs) is very important in safety-critical domains. Randomized Smoothing (RS) has been recently proposed as a scalable, model-agnostic method for robustness verification, which has achieved excellent results and has been extended for a large variety of adversarial perturbation scenarios. However, a hidden cost in RS is during interference, since it requires passing *tens-of-thousands* perturbed samples through the DNN in order to perform the verification. In this work, we try to address this challenge, and explore what it would take to perform RS much faster, perhaps even in real-time, and what happens as we decrease the number of samples by orders of magnitude. Surprisingly, we find that *the performance reduction in terms of average certified radius is not too large, even if we decrease the number of samples by two orders of magnitude, or more*. This could possibly pave the way even for real-time robustness certification, under suitable settings. We perform a detailed analysis, both theoretically and experimentally, and show promising results on the standard CIFAR-10 and ImageNet datasets.

1. Introduction & Related Work

Deep Neural Networks (DNNs) have achieved impressive results in many tasks, such as image and speech recognition (Krizhevsky et al., 2017; Graves et al., 2013), language (Brown et al., 2020), or game playing (Silver et al., 2018). Despite that, applying DNNs in safety-critical domains remains challenging.

One part of the challenge is the lack of robustness: namely, it's well known that slight, imperceptible perturbations on DNN inputs can drastically change the prediction outcome - these are the so-called adversarial examples (Szegedy et al., 2013). After empirical adversarial defences turned out to be broken by stronger attacks (Athalye et al., 2018), the researchers' focus shifted on methods for robustness certification: namely to prove that no adversarial examples exist within a certain region around the input, typically relying on Formal Methods (Wong & Kolter, 2018; Gehr et al., 2018).

Recently, Randomized Smoothing (RS) has emerged as a scalable approach for robustness certification (Cohen et al., 2019). RS has been afterwards extended in many ways (Salman et al., 2019; Yang et al., 2020), and applied to many different perturbation scenarios, such as geometric transformations (Fischer et al., 2020) and more. While much more efficient than other certification approaches, in order to certify robustness with RS, it's required to pass multiple perturbed versions of the input through the DNN (noisy samples), typically in the tens or hundreds of thousands.

In this work, we want to investigate what happens if we reduce this number of samples. Counter-intuitively, we find that the effect of this reduction on the average certified radius that RS achieves is much more minimal that expected; for example, reducing the number of samples by $100\times$ decreases the average certified radius just by 50%. This opens up interesting possibilities, perhaps even performing robustness certification in real-time, which we explore. We apply our approach on the standard CIFAR-10 and ImageNet datasets, and we additionally perform a detailed theoretical analysis.

A related work we identified in the literature is (Chen et al., 2022), where the authors determine the minimum number of samples such that the RS robustness radius at a point doesn't drop more than an allowed value. However, this is different from our case, where the number of samples is constrained from the beginning. Moreover, the authors of (Chen et al., 2022) determine the sample number in an algorithmic way, and give no closed-form formula or analysis of the result.

2. Preliminaries: Randomized Smoothing

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

[1]Fraunhofer Institute for Cognitive Systems (IKS), Munich, Germany [2]National Technical University of Athens (NTUA), Athens, Greece. Correspondence to: Emmanouil Seferis <emmanouil.seferis@iks.fraunhofer.de>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Algorithm 1 RS Certification

Input: point $\mathbf{x}$, classifier f, σ, n, a
Output: class c_A and certified radius R of $\mathbf{x}$
sample n noisy samples $\mathbf{x}'_1, ..., \mathbf{x}'_n \sim N(\mathbf{x}, \sigma^2 I)$
get majority class $c_A = \arg\max_y \sum_{i=1}^{n} \mathbf{1}[f(\mathbf{x}'_i) = y]$
counts$(c_A) \leftarrow \sum_{i=1}^{n} \mathbf{1}[f(\mathbf{x}'_i) = c_A]$
$\bar{p_A} \leftarrow$ LowerConfBound(counts$(c_A), n, a)$ {compute probability lower bound}
if $\bar{p_A} \geq \frac{1}{2}$ **then**
  return $c_A, \sigma\Phi^{-1}(\bar{p_A})$
else
  return ABSTAIN
end if

classifier:

$$g(\mathbf{x}) = argmax_y P[f(\mathbf{x} + \mathbf{z}) = y], \mathbf{z} \sim N(0, \sigma I) \quad (1)$$

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

Surprisingly, if $p_A \geq 0.5$ is the probability of the majority class A, then g is robust around $\mathbf{x}$, with a robustness radius of:

$$R = \sigma\Phi^{-1}(p_A) \quad (2)$$

where Φ^{-1} is the inverse of the normal cumulative distribution function (CDF). The intuition is that a slight perturbation on $\mathbf{x}$ can change the output of f arbitrarily, but not the one of g - since g relies on a distribution of points around $\mathbf{x}$, and a small shift cannot change a distribution much. This is the crucial fact where RS resides.

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 high degree of confidence, as shown in algorithm 1 (Cohen et al., 2019). Yet, the samples required to do so are typically around $10.000 - 100.000$, which makes real-time robustness verification impossible.

3. Methodology & Experiments

To inspect the influence of sample number on the average certified radius, we run experiments on CIFAR10 and ImageNet, where we vary the sample numbers. We work with the code-base of (Cohen et al., 2019), using their pre-trained models. The results for CIFAR-10 can be seen in fig 1, and for ImageNet in fig 2.

We observe that the reduced sample sizes do not decrease

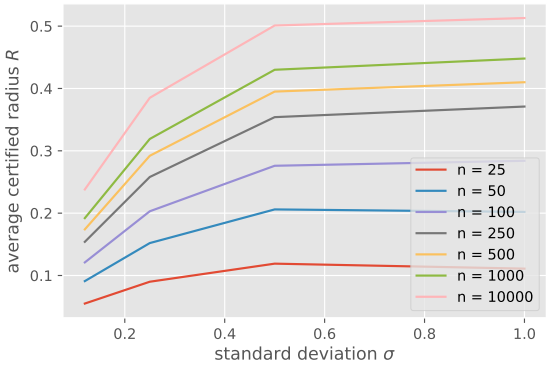

Figure 1. Average certified radius for each noise level σ and sample number n on CIFAR-10, for the models of (Cohen et al., 2019)

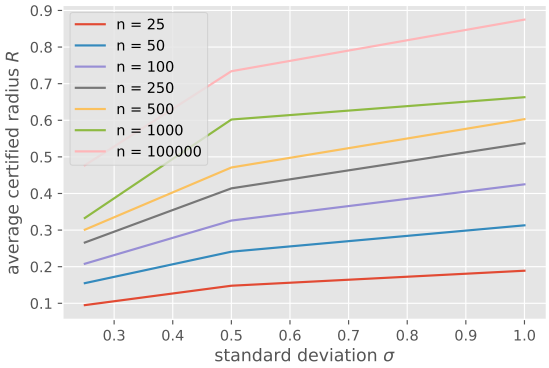

Figure 2. Average certified radius for each noise level σ and sample number n on ImageNet, for the models of (Cohen et al., 2019)

the average certified radius $\bar{R}$ as much as expected: for example, in the case of CIFAR-10, a $10\times$ decrement (from 10.000 to 1000) reduces $\bar{R}$ by only around 20% across noise levels σ. Moreover, a $100\times$ decrement reduces $\bar{R}$ by only 50%. Similarly for the case of ImageNet, a reduction of n from 100.000 to 100 reduces $\bar{R}$ by merely 50%!

Moreover, the decrement of $\bar{R}$ due to the reduced sample size could even be remedied via improvements in the training process of RS. To showcase this, we also run tests on the improved RS models of (Salman et al., 2019), where the authors come up with several ideas on how to train models such that they can obtain better certified radii. The results are shown in fig 3 for CIFAR-10, and in fig 4 in the Appendix for ImageNet.

As we can see, the overall dependency of $\bar{R}$ in terms of n is similar as before. However, for any given n and σ,

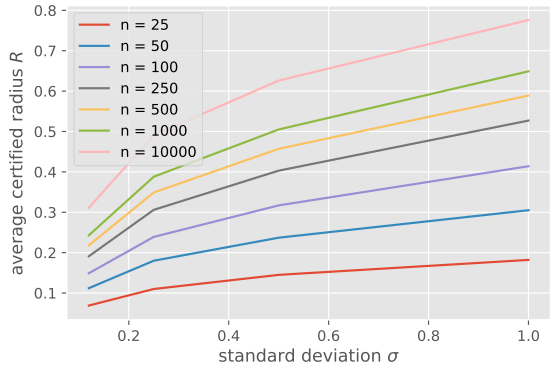

Figure 3. Average certified radius for each noise level σ and sample number n on CIFAR-10, for the best models of (Salman et al., 2019)

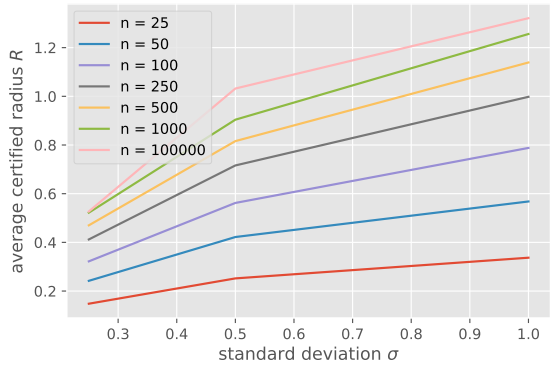

Figure 4. Average certified radius for each noise level σ and sample number n on ImageNet, for the best models of (Salman et al., 2019)

the approach of (Salman et al., 2019) demonstrates a larger certified radius. This shows that the reduction due to n could be compensated by improving RS training. Indeed, we see that $\bar{R}$ at $n = 100$ for (Salman et al., 2019) in the case of CIFAR-10 is roughly equal to the one of (Cohen et al., 2019) at $n = 10.000$! In the case of ImageNet, we can also observe an improvement, but not that large.

The previous results could perhaps even open-up the possibility of performing RS robustness certification in real-time. Consider for example an application such as autonomous driving (AD): there, each frame has to be processed as soon as it's captured, since results such as object detection need to be obtained immediately. Hence, such applications operate essentially with a batch size $b = 1$. However, maybe a larger batch B could be processed by the GPU within the real-time constraints imposed by the use-case. This additional, "unused" batch size could perhaps be used to perform RS in real-time! As we show before, this comes at a cost of a reduced $\bar{R}$, yet this reduction is mild.

To further inspect this possibility, we benchmark a 3080 NVIDIA GPU in order to determine the maximum batch size B a model can process before violating the real-time constraint. For that, we set a maximum processing time of $t = 1/25s = 40ms$, which is the time between two frames in standard video. The results for CIFAR-10 and ImageNet models are shown in fig. 5. As we see, CIFAR-10 allows B's of up to 200 and more, which would make real-time RS possible. However, the situation for a ResNet50 model on ImageNet is different, where B remains at around 15 for full-resolution images, or around 80 for half-resolution. Based on the results of fig. 2, this would lead to a larger reduction of $\bar{R}$, of around 80%.

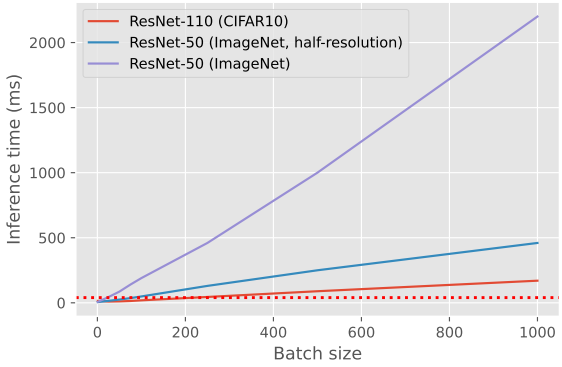

Figure 5. Benchmark of an NVIDIA 3080 GPU on the CIFAR-10 ResNet-110 and ImageNet ResNet-50 models of (Cohen et al., 2019). For ImageNet, we also measure performance on half-resolution images. The horizontal red line indicates the $40ms$ assumed real-time threshold.

Finally, note that the presented use-case above is a bit unrealistic: even if we operate at batch size $b = 1$ in a use-case like AD, the GPU capabilities will not be left un-utilized: for example, in AD they will be used to process multiple frames from different viewpoints at each time step. Hence, our assumption that b can be increased up to B is a bit naive; yet, our results show that a mild level of parallelization could still lead to similar conclusions.

3.1. Theoretical Analysis

To better understand the results found in the experimental section, we attempt to theoretically investigate the effect of sample size on the average certified radius $\bar{R}$. By making a series of approximations, in an effort to reach a closed-form formula, we obtain the following results:

Proposition 3.1. *Suppose we perform Monte-Carlo sampling in order to estimate a lower bound for $\bar{p_A}$ with confidence at least $1-a$ as described in (Cohen et al., 2019). Then, if the true probability is p_0 and we use n samples, $\bar{p_A}$ is approximately:*

$$\bar{p_A} \approx p_0 - z_a \sqrt{\frac{p_0(1-p_0)}{n}} \tag{3}$$

where p_0 is the true probability of A, and z_a is the $1-a/2$ quantile of the normal CDF. Similarly, the certified radius at that point is approximately equal to:

$$R(p_0, n, a, \sigma) \approx \sigma \Phi^{-1}\left(p_0 - z_a \sqrt{\frac{p_0(1-p_0)}{n}}\right) \tag{4}$$

Proposition 3.2. *The certified radius $R(p_0, n, a, \sigma)$ of Proposition 3.1 satisfies the following approximate formula:*

$$R(p_0, n, a, \sigma) \approx 5.063\sigma[p_0^{0.135} - (1-p_0)^{0.135} - 0.135\frac{z_a}{\sqrt{n}}(p_0^{-0.365}(1-p_0)^{1/2} + p_0^{1/2}(1-p_0)^{-0.365})] \tag{5}$$

Proposition 3.3. *Assume that the true probability p_0 of the majority class A follows a uniform distribution in the interval $[0.5, 1)$ across input points $\mathbf{x}$. Then, the drop of the average certified radius $\bar{R}$ using n samples from the ideal case of $n = \infty$ is approximately equal to:*

$$\frac{\bar{R}(n, a, \sigma)}{\bar{R}(\infty, a, \sigma)} \approx 1 - 2\frac{z_a}{\sqrt{n}} \tag{6}$$

The proofs are given in the Appendix. In fig. 6 we plot the approximation of eq. 6 for $a = 0.001$.

We see that the obtained curve roughly captures the dependency of $\frac{\bar{R}(n,a,\sigma)}{\bar{R}(\infty,a,\sigma)}$ we observed in the experiments. First, the radius drop is independent of the noise level σ; indeed, in the experiments we found approximately the same radius reduction across sigmas for each dataset. Second, we observe that the reduction of $\bar{R}(n, a, \sigma)$ from $n = 10.000$ to $n = 1.000$ is around $\approx 85\%$, which is what we see in the experiments. Similarly, the formula shows that there's little difference for $n = 10.000$ and $n = 100.000$ also in agreement with the observations. On the other hand, the predicted reduction as we shift n from 10.000 to 100 is around 37%, which is a bit smaller to the one we saw in experiments.

However, these deviations are to be expected, since Proposition 3.3 relies on the simplifying assumption that the probability distribution of p_0 (the majority class probability) is uniform in $[0.5, 1]$. This is a strong simplification: while the PDF of p_0 is indeed skewed towards 1, it's by no means uniform; it varies strongly for different datasets and models, and we couldn't identify a well-known family of distributions (for example Gaussians) that captures its behavior. Moreover, an additional detail is that in the cases where the model fails to predict correctly, the certified radius is 0; this again depends on the specific model and dataset. Thus, computing the value of $\bar{R}(n, a, \sigma)$ a-priori is not possible. However, eq. 6 seems to capture the general behavior. In fig. 13 in the Appendix, we plot the histograms of p_0 for various models and datasets.

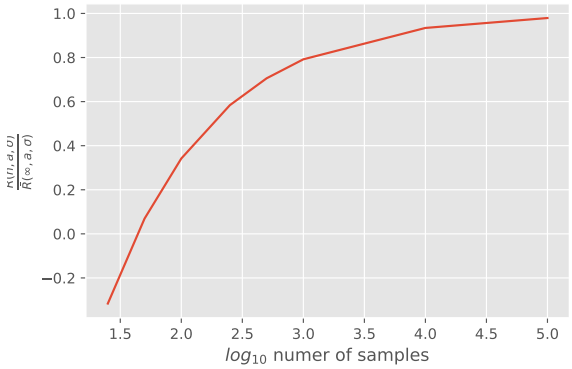

Figure 6. Plots of eq. 6 as a function of n, for $a = 0.001$

3.2. Potential Applications

We think that being able to do robustness estimation with less samples can have multiple applications in AI Safety and beyond. For example, applications such as Automated Driving or Robotics, it might be beneficial to be able to perform robustness certification at run-time. Moreover, during DNN training or in scenarios such as learning from human feedback (Christiano et al., 2017), one could also consider the robustness of the different preferences: being able to estimate the robustness radius via less samples could improve the training process.

4. Conclusion

In this work, we try to address the large number of samples required for RS-based robustness certification, and investigate what happens as the number of samples gets reduced by orders of magnitude. Unexpectedly, we find that the resulting reduction in the average certified radius is much milder than expected. This could perhaps even open-up the possibility of performing RS certification in real-time, under specific settings. We also analyze the phenomenon theoretically, and our findings align with the experiments.

Moreover, another interesting finding is that the loss of certified radius from reducing the sample numbers can be partially compensated by improving the training of RS models, as we found in the experiments. This opens an interesting path for future work: namely, is it possible to train RS models in such a way that the required number of samples for certification becomes lower? And how efficient would that be? We plan to investigate these questions next.

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901, 2020.

Chen, R., Li, J., Yan, J., Li, P., and Sheng, B. Input-specific robustness certification for randomized smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6295–6303, 2022.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Fischer, M., Baader, M., and Vechev, M. Certified defense to image transformations via randomized smoothing. *Advances in Neural information processing systems*, 33: 8404–8417, 2020.

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., and Vechev, M. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2018.

Graves, A., Mohamed, A.-r., and Hinton, G. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pp. 6645–6649. Ieee, 2013.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32, 2019.

Shore, H. 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.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Thulin, M. The cost of using exact confidence intervals for a binomial proportion. 2014.

Wong, E. and Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pp. 5286–5295. PMLR, 2018.

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

A. Additional results

Here, we plot the results of section 3, but now we plot the average certified radius $\bar{R}$ as a function of the number of samples n used for certification. This can show the dependency of $\bar{R}$ with n more clearly.

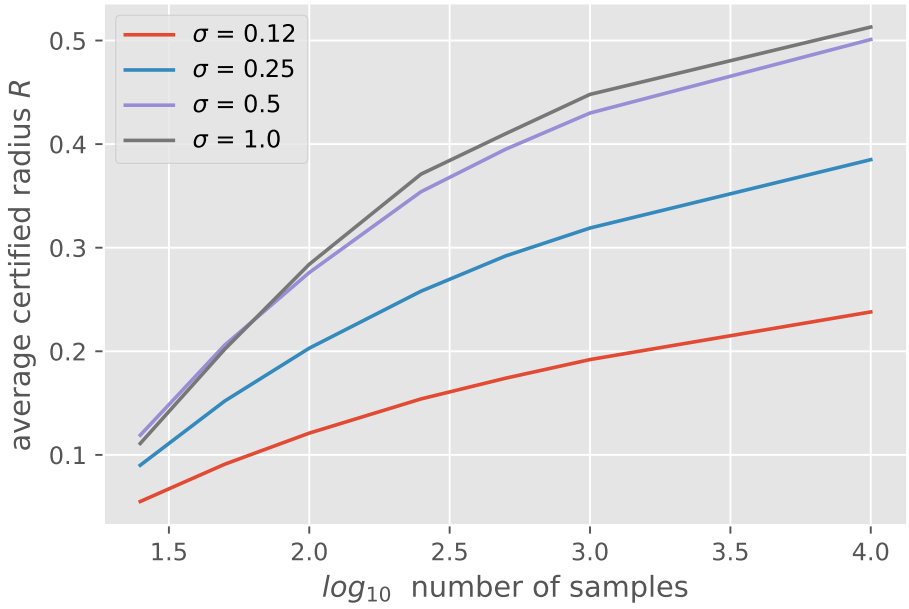

Figure 7. Average certified radius for each noise level σ and sample number n on CIFAR-10, for the models of (Cohen et al., 2019)

B. Deferred Proofs

Proof. (Proposition 3.1) Consider a binomial Random Variable (RV) $X \sim Bin(p_0)$, with probability p_0. Suppose we draw n i.i.d. samples $X_1, X_2, ..., X_n$ from $Bin(p_0)$, and consider the percentage of successes, $\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$. $\bar{X}$ will have a mean of $\mu_{\bar{x}} = E[X] = p_0$, and standard deviation:

$$
\begin{aligned}
&Var[\bar{X}] = Var[\frac{1}{n}(X_1 + ... + X_n)] = \frac{1}{n^2}(Var[X_1] + ... + Var[X_n]) \Leftrightarrow \\
&Var[\bar{X}] = \frac{1}{n^2} \cdot nVar[X] = \frac{Var[X]}{n} = \frac{p_0(1-p_0)}{n} \Leftrightarrow \\
&\sigma_{\bar{X}} = \sqrt{Var[\bar{X}]} = \sqrt{\frac{p_0(1-p_0)}{n}},
\end{aligned}
$$

since $Var[\bar{X}] = p_0(1-p_0)$ for a binomial RV $X \sim Bin(p_0)$, and $X_1, ..., X_n$ are i.i.d.

Now, due to the Central Limit Theorem (CLT), $\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$ will approximately follow a Normal distribution with parameters $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$; for for $n \geq 30$, this approximation will be very accurate. Therefore, the measured success probability, $\bar{X}$, will lie with probability $1 - a$ in the following interval:

$$
\begin{aligned}
&\bar{X} \in [\mu_{\bar{x}} - z_a\sigma_{\bar{x}}, \mu_{\bar{x}} + z_a\sigma_{\bar{x}}] \Leftrightarrow \\
&\bar{X} \in [p_0 - z_a\sqrt{\frac{p_0(1-p_0)}{n}}, p_0 + z_a\sqrt{\frac{p_0(1-p_0)}{n}}]
\end{aligned}
\tag{7}
$$

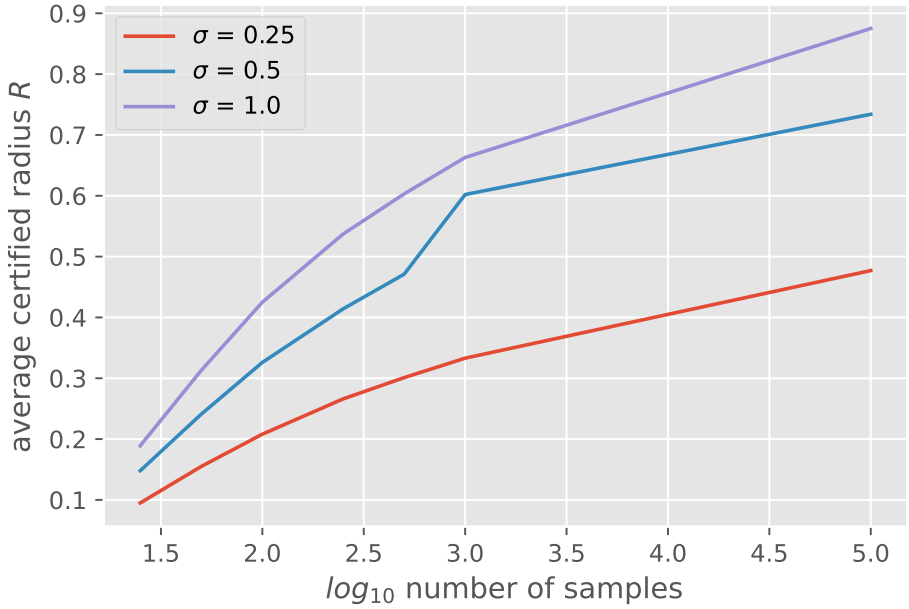

Figure 8. Average certified radius for each noise level σ and sample number n on ImageNet, for the models of (Cohen et al., 2019)

where $z_a = \Phi^{-1}(1 - a/2)$ is the $1 - a/2$ quantile of the Normal distribution $N(0, 1)$. For example, for the typical value of $a = 0.001$ we have $z_a = \Phi^{-1}(0.9995) = 3.2905$.

On the other hand, the Clopper-Pearson interval method used in RS will also return an $1 - a$ confidence interval $[p_{low}, p_{high}]$ for the true success probability p_0, given $\bar{X}$ and n. The difference is that the Clopper-Pearson interval relies on the true, Beta distribution and is exact, while the Gaussian interval of eq. 7 is approximate and doesn't necessarily satisfy the $1 - a$ confidence level. Nevertheless, for $n \geq 30$, the approximation is very close, and we can approximate the interval $[p_{low}, p_{high}]$ with the one of eq. 7. Hence, the lower bound for the success probability p_0 will be, in expectation (Thulin, 2014):

$$p_{low} \approx p_0 - z_a \sqrt{\frac{p_0(1-p_0)}{n}} := \bar{p}(n, a) \tag{8}$$

So, as the number of samples is reduced from n to $n' < n$, the drop on the lower bound success probability will be:

$$\bar{p}(n, a) - \bar{p}(n', a) = z_a \sqrt{p_0(1-p_0)} \cdot \left(\frac{1}{\sqrt{n'}} - \frac{1}{\sqrt{n}} \right) \tag{9}$$

Finally, according to (Cohen et al., 2019), the certified radius at a point $\mathbf{x}$ satisfies $R \geq \sigma \Phi^{-1}(\bar{p})$, where $\bar{p}$ is a lower bound for the success probability of the correct class, that holds with confidence $1 - a$. Therefore, the certified radii we get using n samples will be:

$$R(p_0, n, a, \sigma) \approx \sigma \Phi^{-1} \left(p_0 - z_a \sqrt{\frac{p_0(1-p_0)}{n}} \right) \tag{10}$$

if $p_0 - z_a \sqrt{\frac{p_0(1-p_0)}{n}} \geq \frac{1}{2}$, and 0 otherwise (since in that case the class doesn't have the majority). As the average certified

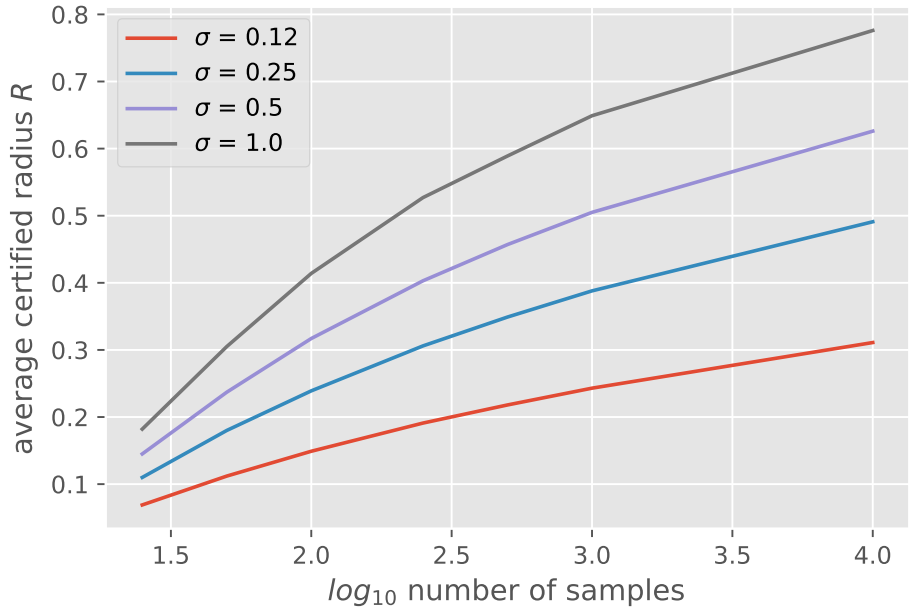

Figure 9. Average certified radius for each noise level σ and sample number n on CIFAR-10, for the best models of (Salman et al., 2019)

radius $\bar{R}$ is computed as an average over many samples, we expect the approximations above to be quite precise for this purpose.

□

In fig. 11 we plot the difference of p_{low} as given by the Clopper-Pearson method (drawing n samples with probability p_0) and the approximation of Proposition 3.1 for various values of n and p_0, with $a = 0.001$. As we observe, the deviation is low even for small n's.

Next, we can also approximate Φ^{-1} with a closed-form function, in order to get an approximate closed-form formula for the radius $R(p_0, n, a, \sigma)$ in eq. 10:

Proof. (Proposition 3.2) The goal here is to make a series of approximations, in order to obtain a closed-form formula for the certified radius $R(p_0, n, a, \sigma)$. The first step is to replace the inverse normal CDF $\Phi^{-1}(.)$ with a closed-form approximation. From (Shore, 1982) we have the following approximate formula:

$$\begin{aligned} x_p = \Phi^{-1}(p) \Leftrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_p} e^{-t^2/2} dt = p \Rightarrow \\ x_p \approx \frac{1}{0.1975}(p^{0.135} - (1-p)^{0.135}) \end{aligned} \tag{11}$$

where the last line in eq. 11 is from (Shore, 1982), where the approximation is valid for $\frac{1}{2} \leq p \leq 1$, which is what we need, since $R(p_0, n, a, \sigma) = 0$ for $p_0 < \frac{1}{2}$.

In fig. 12 we plot the exact values of $\Phi^{-1}(p)$ and our approximation for $p \in [0.5, 1)$. As we see, the approximation formula very close to the true values.

Using the approximation of eq. 11 and substituting in eq. 10, we get the following approximation for the certified radius $R(p_0, n, a, \sigma)$:

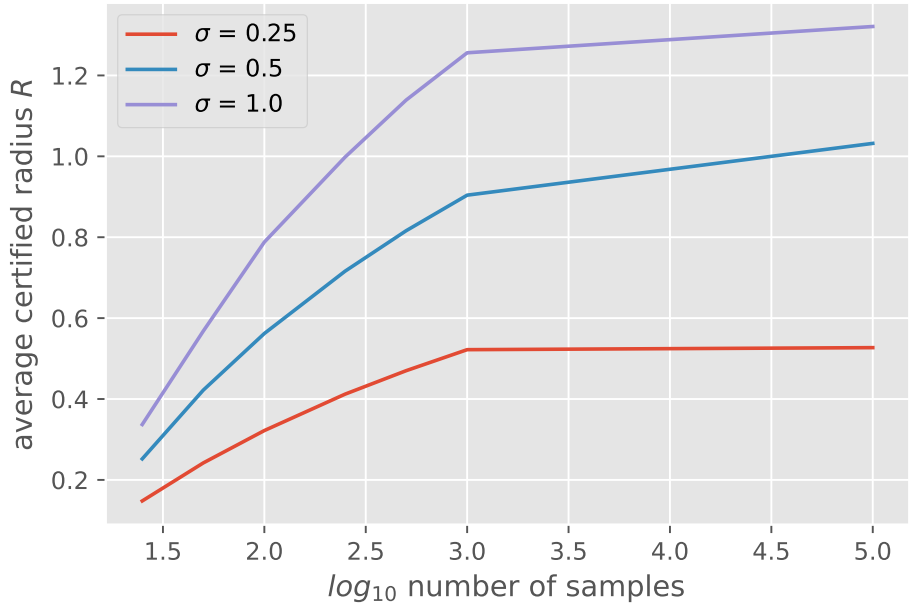

Figure 10. Average certified radius for each noise level σ and sample number n on ImageNet, for the best models of (Salman et al., 2019)

$$
\begin{aligned}
&R(p_0, n, a, \sigma) \approx \sigma \Phi^{-1}\left(p_0 - z_a\sqrt{\frac{p_0(1-p_0)}{n}}\right) \Rightarrow \\
&R(p_0, n, a, \sigma) \approx 5.063\sigma\left[\left(p_0 - z_a\sqrt{\frac{p_0(1-p_0)}{n}}\right)^{0.135} - \left(1 - p_0 + z_a\sqrt{\frac{p_0(1-p_0)}{n}}\right)^{0.135}\right]
\end{aligned}
\tag{12}
$$

To further simplify this equation, we'll apply the binomial theorem, $(1+x)^a = 1 + ax + \frac{a(a-1)}{2!}x^2 + ...$ valid for $|x| < 1$ on both terms of eq. 12, and keep only the 1st order terms. Doing that, we get:

$$
\begin{aligned}
&A = \left(p_0 - z_a\sqrt{\frac{p_0(1-p_0)}{n}}\right)^{0.135} = p_0^{0.135}\left(1 - \frac{z_a}{\sqrt{n}}p_0^{-1/2}(1-p_0)^{1/2}\right)^{0.135} \Rightarrow \\
&A \approx p_0^{0.135}(1 - 0.135\frac{z_a}{\sqrt{n}}p_0^{-1/2}(1-p_0)^{1/2}) = p_0^{0.135} - 0.135\frac{z_a}{\sqrt{n}}p_0^{-0.365}(1-p_0)^{1/2} \\
&B = \left(1 - p_0 + z_a\sqrt{\frac{p_0(1-p_0)}{n}}\right)^{0.135} = (1-p_0)^{0.135}\left(1 + \frac{z_a}{\sqrt{n}}p_0^{1/2}(1-p_0)^{-1/2}\right)^{0.135} \Rightarrow \\
&B \approx (1-p_0)^{0.135}(1 + 0.135\frac{z_a}{\sqrt{n}}p_0^{1/2}(1-p_0)^{-1/2}) = (1-p_0)^{0.135} + 0.135\frac{z_a}{\sqrt{n}}p_0^{1/2}(1-p_0)^{-0.365}
\end{aligned}
\tag{13}
$$

Substituting in eq. 12 and combining terms, we finally get:

$$
R(p_0, n, a, \sigma) \approx 5.063\sigma\left[p_0^{0.135} - (1-p_0)^{0.135} - 0.135\frac{z_a}{\sqrt{n}}(p_0^{-0.365}(1-p_0)^{1/2} + p_0^{1/2}(1-p_0)^{-0.365})\right] \tag{14}
$$

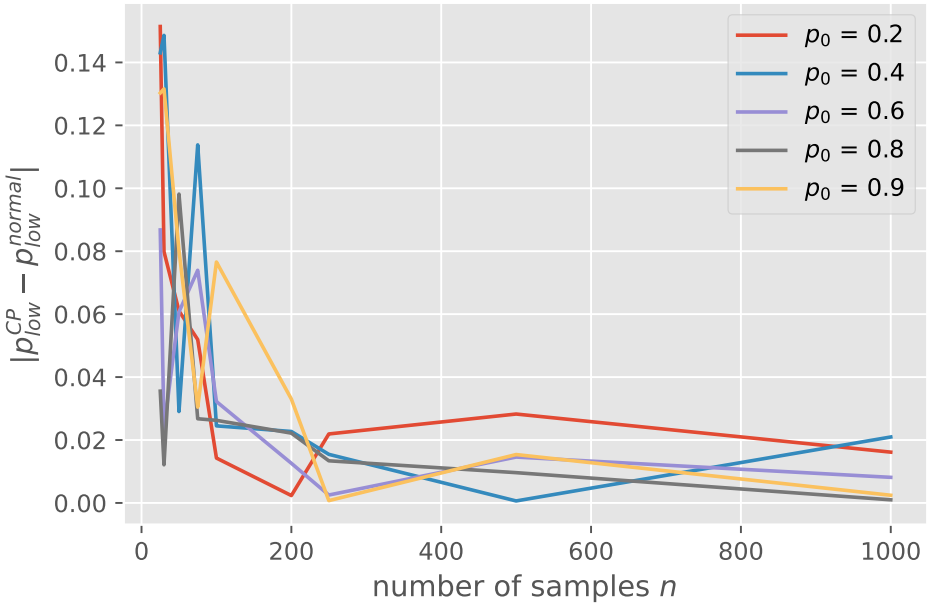

Figure 11. Absolute difference of the lower bounds p_{low} computed by Clopper-Pearson and Proposition 3.1 for various values of n and p_0, and $a = 0.001$.

asd required.

□

Proof. (Proposition 3.3) Here, we make the (simplifying) assumption that the distribution of success probabilities p_0 across input samples $\mathbf{x}$ will be uniform in $[0.5, 1]$. Under this assumption, the average certified radius will be:

$$\bar{R}(n, a, \sigma) = 2 \int_{p_0=0.5}^{1} R(p_0, n, a, \sigma) dp_0 \tag{15}$$

since the PDF of p_0 is $p(p_0) = \frac{1}{1-0.5} = 2$. Substituting eq. 5, we can perform the integration and obtain:

$$\begin{aligned} &\bar{R}(n, a, \sigma) = 2 \int_{p_0=0.5}^{1} R(p_0, n, a, \sigma) dp_0 \Leftrightarrow \\ &\bar{R}(n, a, \sigma) = 10.126\sigma \int_{p_0=0.5}^{1} \left[p_0^{0.135} - (1-p_0)^{0.135} - 0.135 \frac{z_a}{\sqrt{n}} (p_0^{-0.365}(1-p_0)^{1/2} + p_0^{1/2}(1-p_0)^{-0.365}) \right] dp_0 \end{aligned} \tag{16}$$

The integrals of the form p_0^a and $(1-p_0)^a$ can be computed easily, while the integrals of the terms $p_0^a(1-p_0)^b$ are integrals of the Beta function, and can be evaluated numerically. Doing the calculations, we finally get:

$$\bar{R}(n, a, \sigma) = \sigma \left(0.796 - 1.603 \frac{z_a}{\sqrt{n}} \right) \tag{17}$$

Finally, using that we see that the certified radius drop is independent of σ, and is approximately equal to:

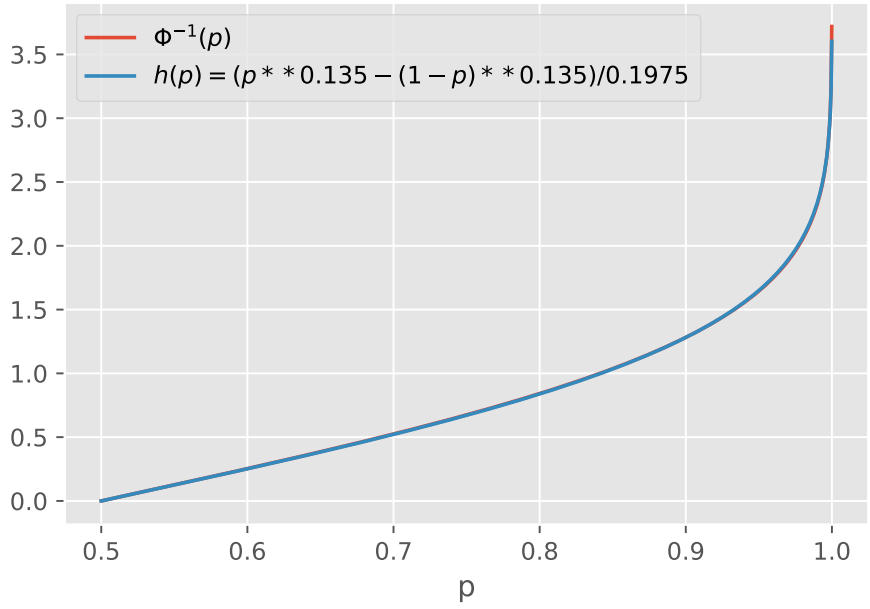

Figure 12. Plots of $\Phi^{-1}(p)$ and our approximation for $p \in [0.5, 1)$

$$\frac{\bar{R}(n, a, \sigma)}{\bar{R}(\infty, a, \sigma)} \approx 1 - 2\frac{z_a}{\sqrt{n}} \tag{18}$$

which is the required formula.

□

At this point, we have to note that the uniformity assumption of p_0 is very naive. However, from the experiments we weren't able to identify some well-known probability distribution for p_0 (e.g. Gaussian, etc.), although it's apparent that the values are skewed towards 1. Some histograms of p_0 for different models and datasets, estimated using $n = 100.000$, are shown in fig. 13.

C. A remark on the certification algorithm

In this section, we comment on a detail on the RS certification algorithm alg. 1. Namely, alg. 1 uses the same random samples to both estimate the majority class c_A, as well as to estimate the lower bound $\bar{p_A}$. Works such as (Chen et al., 2022) use the same method.

However, in the work of (Cohen et al., 2019), the certification algorithm is slightly different, as shown in alg. 2.

The difference is that alg. 2 first draws a small number n_0 of samples to estimate which is the majority class c_A, and then draws a large number n of additional samples to certify it. On the other hand, alg. 1 uses the same samples to estimate c_A and certify it, in an attempt to further reduce the number of samples.

In order to bridge this discrepancy, we show that both algorithms are equivalent, setting $n_0 = n$.

Lemma C.1. *Algorithms alg. 2 and alg. 1 are equivalent when $n_0 = n$.*

Proof. We will show that both algorithms behave in the same way for all cases. First, consider the estimation phase: in both alg. 2 and alg. 1, $n_0 = n$ random samples are used to determine c_A. Now, we distingusih two cases:

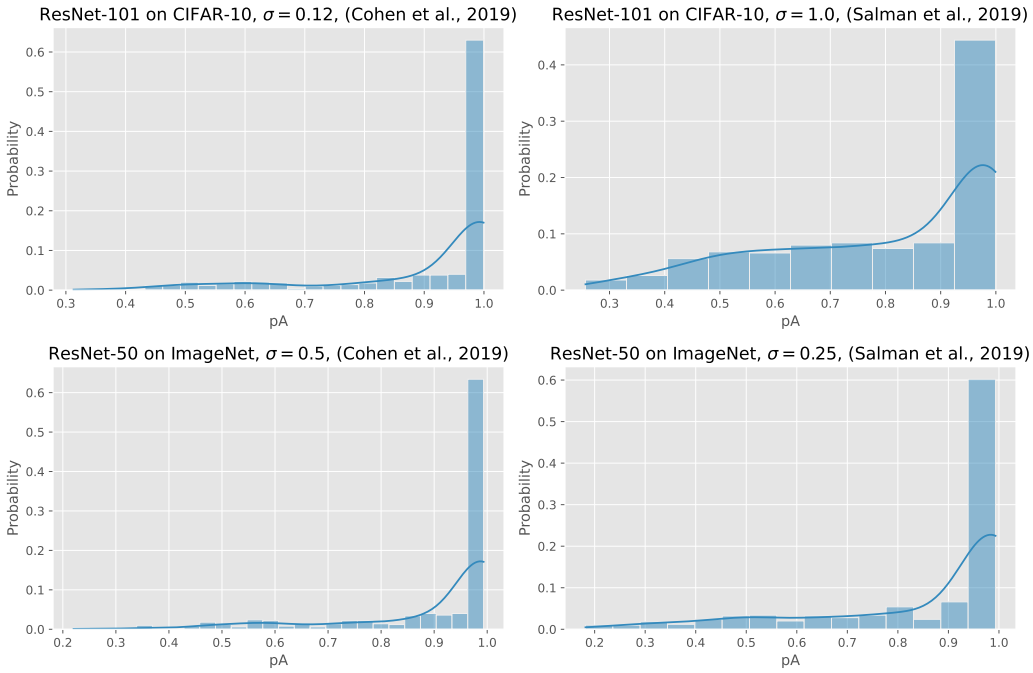

Figure 13. Plots of histograms and density plots of p_0 obtained for different models and datasets, as shown in the figure titles. The values of p_0 we estimated empirically using $n = 100.000$ samples.

.

Suppose the estimated c_A is not equal to the true class y. In that case, the certified radius is 0 in both cases (even if the later stage returns a positive radius, it's invalid).

On the other hand, assume that $c_A = y$. Then, the certification phase of alg. 1 is equivalent to the following: given y, certify it! To do this, alg. 1 uses n random samples. But this is also what alg. 2 does; that is, conditioned that c_A is correct, both algorithms use n random samples to estimate it, and their behavior is equivalent again.

Thus, alg. 2 and alg. 1 are equivalent in all cases. □

To verify Lemma C.1 experimentally, we run the experiments one additional time, using alg. 2 with $n_0 = 100$, and found no noticeable differences.

Algorithm 2 RS Certification (Cohen et al., 2019)

Input: point $\mathbf{x}$, classifier f, σ, n_0, n, a
Output: class c_A and certified radius R of $\mathbf{x}$
sample n_0 noisy samples $\mathbf{x}'_1, ..., \mathbf{x}'_{n_0} \sim N(\mathbf{x}, \sigma^2 I)$
get majority class $c_A = \arg\max_y \sum_{i=1}^{n_0} \mathbf{1}[f(\mathbf{x}'_i) = y]$
sample n noisy samples $\mathbf{x}''_1, ..., \mathbf{x}''_n \sim N(\mathbf{x}, \sigma^2 I)$
counts$(c_A) \leftarrow \sum_{i=1}^{n} \mathbf{1}[f(\mathbf{x}''_i) = c_A]$
$\bar{p_A} \leftarrow$ LowerConfBound(counts$(c_A), n, a)$ {compute probability lower bound}
if $\bar{p_A} \geq \frac{1}{2}$ **then**
  return $c_A, \sigma\Phi^{-1}(\bar{p_A})$
else
  return ABSTAIN
end if