Certified Robustness via Randomized Smoothing over Multiplicative Parameters of Input Transformations

Nikita Muravev[1,2] and **Aleksandr Petiushko**[2,3*]
[1]Huawei Moscow Research Center
[2]Lomonosov Moscow State University
[3]Nuro, Inc.
nikita.muravev@huawei.com, apetiushko@gmail.com

Abstract

Currently the most popular method of providing robustness certificates is randomized smoothing where an input is smoothed via some probability distribution. We propose a novel approach to randomized smoothing over multiplicative parameters. Using this method we construct certifiably robust classifiers with respect to a gamma correction perturbation and compare the result with classifiers obtained via other smoothing distributions (Gaussian, Laplace, uniform). The experiments show that asymmetrical Rayleigh distribution allows to obtain better certificates for some values of perturbation parameters. To the best of our knowledge it is the first work concerning certified robustness against the multiplicative gamma correction transformation and the first to study effects of asymmetrical distributions in randomized smoothing.

1 Introduction

It is well known [Szegedy *et al.*, 2014] that modern classification models are vulnerable to adversarial attacks. There were many attempts [Goodfellow *et al.*, 2015; Kurakin and Ian Goodfellow, 2017; Liu *et al.*, 2018] to construct empirically robust classifiers, but all of them are proven to be vulnerable to some new more powerful attacks [Carlini and Wagner, 2017; Athalye *et al.*, 2018] . As a result, a lot of certification methods have been introduced [Wong and Kolter, 2018; Li *et al.*, 2021; Alfarra *et al.*, 2021] that provide provably correct robustness certificates. Among this variety of certification methods the randomized smoothing remains one of the most effective and feasible approaches which is easy to implement and to parallel. Randomized smoothing [Cohen *et al.*, 2019; Li *et al.*, 2019; Lecuyer *et al.*, 2019] can be applied to data and classifiers of any nature since it does not use any information about the internal structure of the classifier. It is also easily scalable to large networks unlike other methods. All these benefits lead to the fact that the randomized smoothing approach is currently one of the most popular and powerful solutions to robustness certification tasks.

Consequently, there are a lot of works devoted to the randomized smoothing. While most of them concern a setting of robustness within l_p- or l_∞-balls, some generalize the approach to cover parameterized transformations [Fischer *et al.*, 2020; Li *et al.*, 2021]. These parameters usually add up to each other when composing transformations, which allows to smooth classifiers over them. Yet there are some other types of perturbations such that their parameters do not add up but multiply to each other (e.g. a volume change of audio signals or a gamma correction of images). Though it is possible to reduce this case to the additive one by a logarithm (like it is done in [Fischer *et al.*, 2020]), we find that alternative distributions (like Rayleigh) may result in better certificates. Thus we propose a new approach of smoothing directly over multiplicative parameters. We prove the theoretical guarantees for classifiers smoothed with Rayleigh distribution and compare them with those obtained via other smoothing distributions for the gamma correction perturbation. Our experiments show that the Rayleigh smoothing cannot be consistently outperformed by the others and thus may be more suitable for certain values of the perturbation parameter.

The main contributions of this work are:

- Direct generalization of the randomized smoothing method to the case of multiplicative parameters;
- Practical comparison of the proposed method and other smoothing techniques for the gamma correction perturbation, which shows superiority of our method for some factors;
- Construction of the first certifiably robust classifiers for the gamma correction perturbation.

2 Background

The randomized smoothing is a method of replacement of an original classifier f with its "smoothed" version $g(x) = \arg\max_c \mathbb{P}_\beta(f \circ \psi_\beta(x) = c)$ that returns the class the base classifier f is most likely to return when an instance is perturbed by some randomly distributed composable transformation ψ. A right choice of the distribution β allows to predict a robustness certificate for a given input x for this smoothed classifier g using its confidence in its own answer. Here a robustness certificate is a range of transformations such that they do not change the predicted class $g(x)$ when

*Work done at Huawei Moscow Research Center

The extended version of the paper with an appendix https://arxiv.org/abs/2106.14432

applied to the given input x. One expects the new (smoothed) classifier to be reasonably good if the base one is sufficiently robust.

This technique has only two major drawbacks. The first one is that the transformation we want out classifier to be robust against has to be composable. It is a strong constraint since even theoretically composable transformations lose this property in practice due to rounding and interpolation errors. The other one is that generally we are unable to calculate the smoothed classifier g directly and instead have to use approximations. Since we cannot calculate the actual prediction of g, we use the Monte-Carlo approximation algorithm with n samples to obtain the result with an arbitrary high level of confidence. The introduced parameter n controls the trade-off between the approximation accuracy and the evaluation time on inference. Though it is possible to smooth a default undefended base classifier f, experiments show that the best results can be achieved by smoothing of already (empirically) robust base classifier f. Thus during training of the base classifier f we use the same type of augmentation we expect the smoothed classifier g to be robust against. For a more detailed description of the randomized smoothing method see [Cohen *et al.*, 2019; Li *et al.*, 2019; Lecuyer *et al.*, 2019].

3 Generalization of Smoothing for Multiplicative Parameters

Let us extend the notion of composable transformations in case of multiplicative parameters.

Definition 1. *A parameterized map $\psi_\delta : X \to X, \quad \delta \in \mathcal{B} \subset \mathbb{R}^n$ is called multiplicatively composable if*

$$(\psi_\delta \circ \psi_\theta)(x) = \psi_{(\delta \cdot \theta)}(x),\ \forall x \in X,\ \forall \delta, \theta \in \mathcal{B}, \quad (1)$$

where $\delta \cdot \theta \in \mathcal{B}$ means the element-wise multiplication of vectors.

Usually multiplicative parameters are positive, thus one needs probability distributions with positive supports for randomized smoothing. Yet unlike the additive case an identity transformation corresponds to the parameter value equal to 1 rather than 0. So a conventional exponential distribution [Li *et al.*, 2021] is unsuitable here (we want to concentrate probability mass at 1). Thus we propose a Rayleigh distribution for these needs.

Definition 2. *A random variable ζ has a Rayleigh distribution with the scale parameter $\sigma > 0$ ($\zeta \sim Rayleigh(\sigma)$) if its probability density function (PDF) has a form*

$$p_\zeta(z) = \sigma^{-2} z e^{-z^2/(2\sigma^2)}, z \geq 0. \quad (2)$$

For classifiers smoothed with the Rayleigh distribution the following robustness guarantee can be obtained:

Theorem 3. *Let $x \in \mathbb{R}^m$, $f : \mathbb{R}^m \to Y$ be a classifier, $\psi_\beta : \mathbb{R}^m \to \mathbb{R}^m$ be a multiplicatively composable transformation for $\beta \sim Rayleigh(\sigma)$ and $g(x) = \arg\max_c \mathbb{P}_\beta(f \circ \psi_\beta(x) = c)$. Denote*

$$p_A = \mathbb{P}_\beta(f \circ \psi_\beta(x) = c_A),$$

$$p_B = \max_{c_B \neq c_A} \mathbb{P}_\beta(f \circ \psi_\beta(x) = c_B).$$

If

$$p_A \geq \underline{p_A} > \overline{p_B} \geq p_B,$$

then $g \circ \psi_\gamma(x) = c_A$ for all γ satisfying $\gamma_1 < \gamma < \gamma_2$, where γ_1, γ_2 are the only solutions of the following equations

$$F(\gamma_1^{-1} F^{-1}(\overline{p_B})) + F(\gamma_1^{-1} F^{-1}(1 - \underline{p_A})) = 1, \quad (3)$$

$$F(\gamma_2^{-1} F^{-1}(\underline{p_A})) + F(\gamma_2^{-1} F^{-1}(1 - \overline{p_B})) = 1, \quad (4)$$

and $F(z) = 1 - e^{-z^2/(2\sigma^2)}$ is the CDF of β.

A proof is similar to the one provided in [Fischer *et al.*, 2020], though it is also possible to obtain the same result using the constrained adversarial certification framework [Zhang *et al.*, 2020] or TSS framework [Li *et al.*, 2021]. All variants are given in Appendix A, B. Notice that the certificate bounds γ_1, γ_2 can be easily found numerically (e.g. by a simple bisection method). But it is also possible to use a trivial estimation $\overline{p_B} = 1 - \underline{p_A}$ for the probability of the top two class. In that case the equations can be solved analytically:

$$\gamma_1 = \frac{F^{-1}(1 - \underline{p_A})}{F^{-1}(\frac{1}{2})} = \sqrt{\frac{\log \underline{p_A}}{\log \frac{1}{2}}}, \quad (5)$$

$$\gamma_2 = \frac{F^{-1}(\underline{p_A})}{F^{-1}(\frac{1}{2})} = \sqrt{\frac{\log(1 - \underline{p_A})}{\log \frac{1}{2}}}. \quad (6)$$

The only parameter to be chosen for the Rayleigh distribution is the scale σ. It seems reasonable to choose it in such a way that either the median or the mean of the random value equals one (value multiplied by 1 stays the same). We try both variants and find that the median equal to 1 results in better certificates (See Appendix C for comparison details). Thus hereafter the scale is equal to $\frac{1}{\sqrt{2 \ln(2)}}$.

The above theorem can be generalized in case of transformations with multiple multiplicative parameters.

Theorem 4. *Let $x \in \mathbb{R}^m$, $f : \mathbb{R}^m \to Y$ be a classifier, $\psi_\beta : \mathbb{R}^m \to \mathbb{R}^m$, $\beta = (\beta_1, ..., \beta_n)^T$ be a multiplicatively composable transformation for independent and identically distributed random variables $\beta_i \sim Rayleigh(\sigma)$ and $g(x) = \arg\max_c \mathbb{P}_\beta(f \circ \psi_\beta(x) = c)$. Denote*

$$p_A = \mathbb{P}_\beta(f \circ \psi_\beta(x) = c_A),$$

$$p_B = \max_{c_B \neq c_A} \mathbb{P}_\beta(f \circ \psi_\beta(x) = c_B).$$

If

$$p_A \geq \underline{p_A} > \overline{p_B} \geq p_B,$$

then $g \circ \psi_\gamma(x) = c_A$ for all $\gamma \in \Omega$, where Ω is a region defined by the following inequality

$$\mathbb{P}((\gamma_1^2 - 1)\beta_1^2 + ... + (\gamma_n^2 - 1)\beta_n^2 \leq r) > \mathbb{P}((\gamma_1^2 - 1)\beta_1^2 + ... + (\gamma_n^2 - 1)\beta_n^2 \geq \theta), \quad (7)$$

where r and θ are the only solutions of the following equations

$$\mathbb{P}((1 - \gamma_1^{-2})\beta_1^2 + ... + (1 - \gamma_n^{-2})\beta_n^2 \leq r) = \underline{p_A}, \quad (8)$$

$$\mathbb{P}((1 - \gamma_1^{-2})\beta_1^2 + ... + (1 - \gamma_n^{-2})\beta_n^2 \geq \theta) = \overline{p_B}. \quad (9)$$

$\underline{p_A}$	$\overline{p_B}$	γ_1	γ_2
0.600	0.400	0.86	1.15
	0.200	0.71	1.33
0.700	0.300	0.72	1.32
	0.100	0.54	1.56
0.800	0.200	0.57	1.52
0.900	0.100	0.39	1.82
0.990	0.010	0.12	2.58
0.999	0.001	0.04	3.16

Table 1: The calculated robustness certificates (γ_1, γ_2) for the top two class probabilities $\underline{p_A}, \overline{p_B}$.

A proof is analogous to the one for a single parameter case and can be found in Appendix A.

Usually we cannot evaluate the probabilities p_A, p_B. Thus we use the Clopper-Pearson [Clopper and Pearson, 1934] bounds $p_B \leq \overline{p_B} < \underline{p_A} \leq p_A$ that can be calculated with an arbitrary high confidence probability $1 - \alpha$. In case $\underline{p_A} \leq \frac{1}{2}$ the classifier g abstains from answering (i.e. returns the "abstain" answer).

We present γ_1, γ_2 values for some $\underline{p_A}, \overline{p_B}$ in Table 1 to show the certificates that can possibly be achieved.

4 Experiments

There are a lot of multiplicatively composable transformations. We choose the gamma correction of images for our experiments due to its importance and simplicity.

Gamma correction is a very popular operation in image processing. There are a lot of works on how it is used for image enhancement [Veluchamy and Subramani, 2019] and medical image processing [Somasundaram and Kalavathi, 2011; Agarwal and Mahajan, 2017]. Yet to the best of our knowledge no works on certified robustness to gamma correction attacks have been released. Thus we are eager to both fill this gap and try out our new smoothing technique on it.

If we consider an input image in the RGB-format as a tensor x with entries in $[0, 1]$, then the gamma correction G_γ with the gamma factor γ simply raises x to the power γ in the element-wise manner: $G_\gamma(x) = x^\gamma$. For a human eye it looks like change of brightness (Fig.1).

Obviously, this transformation is multiplicatively composable:

$$G_\beta \circ G_\gamma(x) = (x^\gamma)^\beta = x^{\gamma \cdot \beta} = G_{\gamma \cdot \beta}(x). \tag{10}$$

But it should be noticed that in reality a gamma corrected image is likely to be converted to some image format with colour channel of limited width afterwards. In that case some information is lost and the resulting transformation is no longer composable. Thus we have two settings: 1) "idealized" — when the colour channel is so wide that the conversion error is negligible in comparison with the rounding error of a machine; 2) "realistic" — when the conversion error is significant and cannot be ignored.

Other works usually consider idealized setting where the only type of error we encounter is an interpolation error. Following their lead we conduct most of our experiments in this setting though we also show how to deal with conversion errors.

4.1 Idealized Setting

As it is mentioned before the multiplicative parameters β, γ can be converted into the additive ones $a = \log_c \gamma, b = \log_c \beta$ via logarithm:

$$\gamma \cdot \beta = c^{\log_c \gamma} \cdot c^{\log_c \beta} = c^a \cdot c^b = c^{a+b}, \tag{11}$$

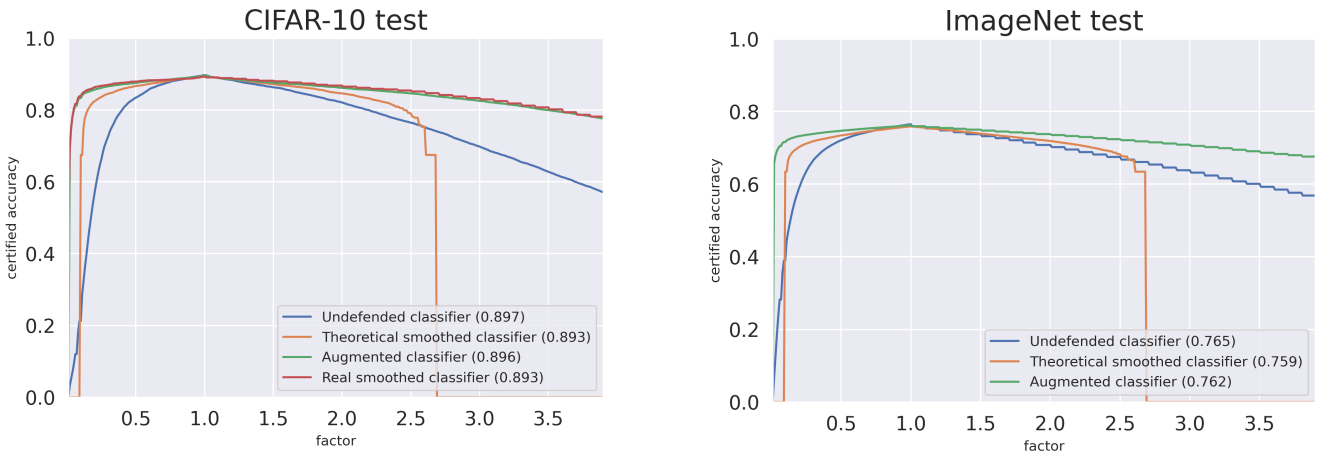
where c is some fixed base. Thus one can smooth over these new parameters with the standard Gaussian distribution to obtain certificates on the original ones. The certified accuracy for a factor γ is the proportion of correctly classified test images whose certificates contain the value γ. All certificates are obtained with the mistake probability $\alpha = 0.001$.

We compare the theoretical certificates for the smoothed classifier with empirical ones for: a) the same smoothed classifier (real smoothed classifier), b) the base architecture without augmentations on training or smoothing (undefended classifier), and c) the base architecture with augmentations on training but without smoothing (augmented classifier) (Fig.2a). Numbers in brackets show values at 1, that is benign or clean accuracy: the portion of unperturbed images that were correctly classified.

The empirical certificates are obtained via evaluations of classifiers' predictions on perturbed images. For every image x starting from gamma factor $\gamma_0 = 1$ we evaluate classifier's predictions on $G_{\gamma_k}(x)$ for every $\gamma_k = \gamma_{k-1} + \epsilon$ with a step $\epsilon = 0.01$. If the classifier predicts a wrong label on the input

(a) $\gamma = 0.5$ (b) $\gamma = 1$ (c) $\gamma = 2$

Figure 1: Demonstration of the gamma correction transformation on a frog image.



(a) ResNet-20 on CIFAR-10

(b) ResNet-50 on ImageNet

Figure 2: Comparison of theoretically predicted certificates with empirically obtained ones for the gamma correction perturbation.

$G_{\gamma_k}(x)$ for some γ_k, we assume that it is robust only in the interval $[1, \gamma_{k-1}]$ and stop the certification procedure. The same procedure is used to find the left end of a certification interval. We use $\epsilon = 0.1$ searching for the right ends of certificate intervals for some classifiers (you can spot them by the less smooth right parts of corresponding graphs) due to computational complexity of this process. The resulting certificate can be over optimistic but still it provides a good approximation of the actual robustness. One can see that the real accuracy is bounded from below by the theoretically guaranteed one, which is expected.

We also observe that the smoothed classifier has almost the same actual robustness as the just augmented one. Thus it can be concluded that smoothing provides theoretical certificates at a cost of extra time on inference but does not significantly impact the actual robustness of the augmented base classifier.

The randomized smoothing method is easily scalable to deep models and large datasets. We repeat certification experiments for ResNet-50 on ImageNet [Deng *et al.*, 2009] (Fig.2b) but do not approximate actual robustness due to computational complexity of this process.

In order to compare Rayleigh smoothing with other smoothing distributions we train several smoothed ResNet-110 classifiers on CIFAR-10 dataset with uniform, Gaussian and Laplace distributions following TSS protocol (including consistency regularization [Jeong and Shin, 2020]). We also use the TSS certification protocol with $\alpha = 0.001, n = 100000$ and 500 images randomly picked from the test set for certification. The results are presented at Fig.3. To ease comparison the factor axes have been re-scaled for factors smaller than 1 to represent uniform multiplicative scale and a dotted line has been used for Rayleigh smoothing.

One can see that the Rayleigh smoothing outperforms all competitors for small gamma factors. Apparently asymmetric nature of the Rayleigh distribution allows to obtain much better certificates for small perturbation values than it would be possible with any conventional symmetric distribution. See Appendix E for more detailed analysis of this phenomenon.

4.2 Realistic Setting

Our study shows that if we convert gamma corrected images to some format with 8 bits per colour channel (RGB true colour), then the conversion error is usually too big to be ignored. Thus we need to make the base classifier f robust in

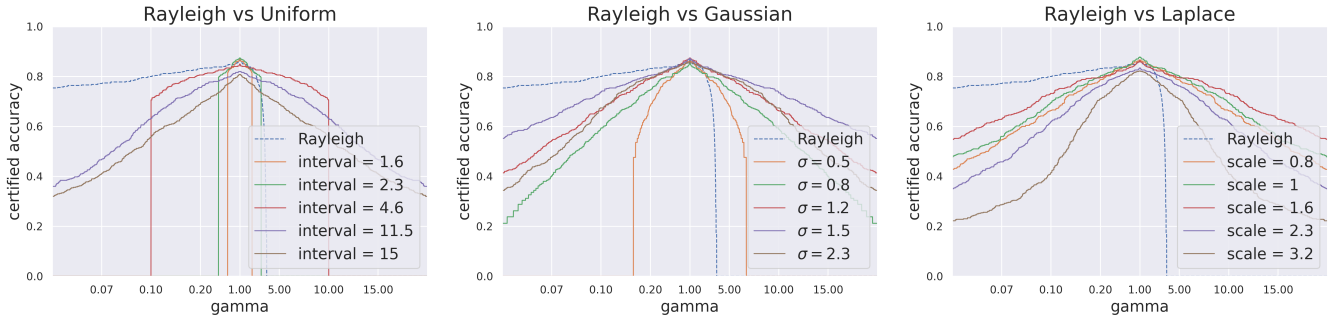


Figure 3: Comparison of Rayleigh smoothing with uniform, Gaussian and Laplace smoothing for ResNet-110 on CIFAR-10 for the gamma correction perturbation. The base is equal to e everywhere, the mean equals 0 for all alternate distributions and we only change the length of support (uniform), standard deviation (Gaussian) and scale (Laplace) parameters.

some l_2-ball with a radius larger then the conversion error for the most images. For these needs Fischer et al. [Fischer *et al.*, 2020] propose to initially smooth the base classifier with the additive Gaussian noise to make it robust in l_2-norm. This new classifier can then be used as a base one for smoothing over transformation parameters. Thus instead of n samples for the Monte-Carlo simulation algorithm we use n_ε samples to simulate the smoothing in l_2-norm and n_γ samples to simulate the smoothing over the gamma correction. Since we need n_ε samples for every gamma correction sample, the resulting number of samples is equal to $n_\varepsilon \cdot n_\gamma$.

There are two types of guarantees that can be obtained by this method: 1) a distributional guarantee, when the conversion error is pre-calculated on the training dataset; 2) an individual guarantee, when we calculate the conversion error for each input at inference time.

For a given input image x and the gamma correction transformation G we define

$$\varepsilon(\beta, \gamma, x) := G_\beta \circ G_\gamma(x) - G_{\beta \cdot \gamma}(x) \tag{12}$$

as the conversion error for gamma factors β, γ. In the idealized setting we assume $\varepsilon(\beta, \gamma, x) = 0$ for all inputs. But now we need to estimate this value with some upper bound E.

In case of distributional guarantees the conversion error can be estimated with E such that

$$q_E = \mathbb{P}_{x \sim D, \beta \sim Rayleigh} \left(\max_{\gamma \in \Gamma} ||\varepsilon(\beta, \gamma, x)||_2 \leq E \right), \tag{13}$$

where D is the data distribution, Γ is a fixed interval of possible attacks and $1 - q_E$ is a desirable error rate.

The conversion errors E are estimated for intervals Γ of gamma factors $\Gamma_1 = [0.86, 1.15]$ and $\Gamma_2 = [0.71, 1.33]$ with the same error rate $q_E = 0.9$ as $E_1 = 0.18$ and $E_2 = 0.22$ respectively. These intervals are chosen randomly but in such a way that they could be certificate intervals obtained with the Rayleigh smoothing for some input images, i.e. there exist such $\underline{p_A}, \overline{p_B}$ that the corresponding certificates obtained via Theorem 3 equal to Γ_1 or Γ_2. Then we smooth the base classifier with the Gaussian noise with a deviation $\sigma = 0.25$, that is found experimentally to deliver the best results for our base models and datasets.

The resulting mistake probability ρ of that smoothed classifier on a ball $B_E(x)$ for a given x can be estimated as the sum of mistake probabilities on all steps:

$$\rho \leq \alpha + 1 - q_E + \alpha_E, \tag{14}$$

where $1 - \alpha$ is the confidence of the base classifier, q_E is the probabilistic guarantee for E and $1 - \alpha_E$ is the confidence with which E is obtained. In all our experiments we set $\alpha_E = 0.01$.

This smoothed classifier is then smoothed with the Rayleigh distributed gamma correction. For the certification procedure the probabilities $\underline{p_A}, \overline{p_B}$ are adjusted by ρ:

$$\underline{p_A}' = \underline{p_A} - \rho, \ \overline{p_B}' = \overline{p_B} + \rho \tag{15}$$

such that we take into consideration the mistake probability of the base classifier we have estimated previously. In order to preserve the correctness of conversion error estimations the obtained certificates are clipped to the selected attack intervals Γ.

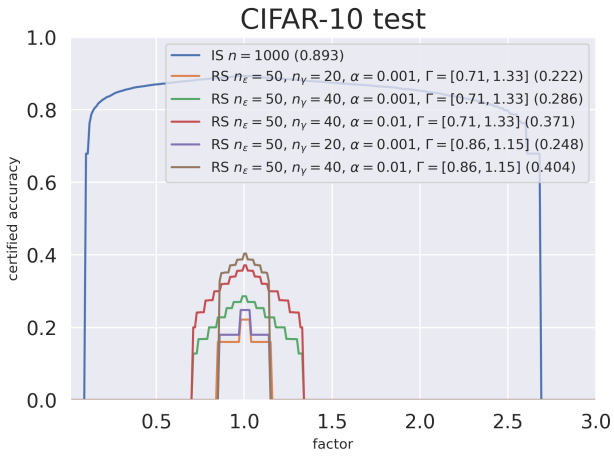


Figure 4: Certified accuracy of smoothed ResNet-20 classifiers in the idealized (IS) and the realistic (RS) settings on CIFAR-10 for the gamma correction perturbation.

We provide distributional guarantees for smoothed classifiers on CIFAR-10 for different numbers of samples n_ε, n_γ, mistake probabilities α and attack intervals Γ (Fig. 4). A significant drop in accuracy can be seen when we switch from the idealized setting to the realistic one. It happens because we are not able to preserve the same number of samples for smoothing over the gamma parameter (due to the computational complexity) as well as because our base classifier has to be robust in l_2-norm.

One can see that the smaller attack interval allows the lower estimation E (which is expected) and thus the higher accuracy can be obtained. It should also be noticed that a greater mistake probability α increases the accuracy due to fewer "abstain" answers by both smoothed classifiers, but at the same time a greater α decreases the resulting confidence with which certificates are obtained.

The achieved certification results can be found in Appendix D.

4.3 Experiments With Scaling Interpolation Error

The other type of distortion we experiment with is a scaling interpolation error. By that we mean a transformation $S_r(x)$ which scales the image x by factor r followed by re-scaling to the original size. For a human eye it looks like a loss of clarity (Fig. 5). Note that this transformation is not the same as scaling (e.g. like here [Li *et al.*, 2021]) where there is no re-scaling to the original size and black padding is used instead. The resulting interpolation error is a perturbation we want our classifier to be robust against. It is clear that in the idealized setting scale factors multiply to each other when we scale an image several times and in that case the interpolation error does not exist. Thus this transformation is multiplicatively composable. But in reality we face non-zero interpolation errors that cannot be ignored. Though it is possible to handle this problem via double smoothing like it is done with the gamma correction, we find that the average interpo-

(a) $\gamma = 0.5$ (b) $\gamma = 1$ (c) $\gamma = 2$

Figure 5: Demonstration of the scaling interpolation error on a frog image.

lation error on CIFAR-10 images is too great to be handled via Gaussian smoothing with satisfactory accuracy. Fischer et al. [Fischer *et al.*, 2020] encounter the same problem dealing with rotations. They propose several techniques to reduce interpolation errors (preprocessing with vignetting and low-pass filter, noise partitioning). Some of these techniques can be applied to our case but we find that the problem persists even for high-resolution images (e.g. ImageNet) that is not the case for rotations.

It should be noticed that down-scaling presents a greater challenge than up-scaling since the latter is theoretically invertible and results in smaller interpolation errors. Therefore it seems reasonable to use non-symmetric (with respect to multiplication) distributions for smoothing like the Rayleigh one, which has a distinct bias towards smaller values rather than bigger ones. One can expect such distributions to provide better certification results than the standard Gaussian and other symmetric distributions for perturbations whose symmetric parameter values do not cause the same level of distortion.

Of course one could use the TSS framework [Li *et al.*, 2021] to deal with the interpolation errors but this approach does not require any smoothing distribution and thus it does not allow to exploit the asymmetrical nature of scaling distortions via asymmetrical probability distributions.

5 Further Research

There are some recent works revealing the limitations of the standard Gaussian smoothing [Hayes, 2020; Zhang *et al.*, 2020], especially in case of multidimensional perturbations with l_p-ball certificates for large p. It seems reasonable to search for alternative smoothing distributions not only among those suitable for additive parameters but also among distributions of multiplicative parameters. Indeed, it is interesting to find a good family of distributions for smoothing over multiplicative parameters, so that one could vary the variance like in additive case with uniform, Gaussian and Laplace distributions. Moreover, an example of the scaling perturbation shows that we may be more interested in asymmetric distributions which are able to exploit asymmetrical nature of certain transformations. There is a chance that it will allow to avoid the pitfalls and limitations of conventional smoothing distributions and achieve better results in the robustness certification tasks.

Also this technique can be applied to other types of perturbations and even to other types of data (audio signals, video, etc.). Therefore further experiments can include construction of certifiably robust audio and video classifiers based on the Rayleigh smoothing and comparison of them with those obtained via other methods.

6 Related Work

In this section we want to discuss similar works in certification theory and point out some key differences between them and the present paper. Usually randomized smoothing is applied to classifiers having R^d as their domain, resulting in the threat model when the signal is directly fed into the model without interpolation, rounding or clipping (like in [Li *et al.*, 2021]). In real life it is unlikely that we will be able to perturb the original signal itself. Usually we have only a digital record of the signal (e.g. an image) and cannot access the original. Thus this setting seems unrealistic and we call it "idealized" in the paper. We go further and propose to consider "realistic" attacks that take into account limitations of digital image processing (limited channel width and bounds on minimal and maximal pixel values). Notice that gamma correction preserves pixel values within $[0, 1]^3$ RGB-area (unlike brightness or contrast changes) and thus can be easily certified against in "realistic" setting via above mentioned double smoothing technique.

The realistic setting results in a big certified accuracy drop, but we believe that it is a much more adequate way to simulate real-world attacks.

7 Conclusion

This work proposes a novel approach of randomized smoothing directly over multiplicative parameters of input transformations. We prove certification theorems and provide experimental comparison of the proposed method with conventional smoothing distributions for gamma correction transformations and show that it cannot be consistently outperformed by the latter. We also propose to consider realistic attacks which include additional conversion errors caused by the limited color channel width. As a result, the first certifiably robust image classifiers in the idealized and realistic settings are constructed for the gamma correction perturbation.

References

[Agarwal and Mahajan, 2017] Monika Agarwal and Rashima Mahajan. Medical images contrast enhancement using quad weighted histogram equalization with adaptive gama correction and homomorphic filtering. *Procedia Computer Science*, 115:509–517, 2017. 7th International Conference on Advances in Computing & Communications, ICACC-2017, 22-24 August 2017, Cochin, India.

[Alfarra *et al.*, 2021] Motasem Alfarra, Adel Bibi, Naeemullah Khan, Philip H. S. Torr, and Bernard Ghanem. Deformrs: Certifying input deformations with randomized smoothing. *ArXiv*, 2021.

[Athalye *et al.*, 2018] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, AISec '17, page 3–14, New York, NY, USA, 2017. Association for Computing Machinery.

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

[Cohen *et al.*, 2019] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. *Proceedings of Machine Learning Research*, 97:1310–1320, 2019.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: a large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 06 2009.

[Fischer *et al.*, 2020] Marc Fischer, Maximilian Baader, and Martin Vechev. Certified defense to image transformations via randomized smoothing. *Advances in Neural Information Processing Systems*, 33:8404–8417, 2020.

[Goodfellow *et al.*, 2015] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[Hayes, 2020] J. Hayes. Extensions and limitations of randomized smoothing for robustness guarantees. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3413–3421, 2020.

[Jeong and Shin, 2020] Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10558–10570. Curran Associates, Inc., 2020.

[Kurakin and Ian Goodfellow, 2017] Alexey Kurakin and Samy Bengio Ian Goodfellow. Adversarial machine learning at scale. *ArXiv*, 2017.

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

[Li *et al.*, 2019] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *Advances in Neural Information Processing Systems*, 32, 2019.

[Li *et al.*, 2021] Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. Tss: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, page 535–557, New York, NY, USA, 2021. Association for Computing Machinery.

[Liu *et al.*, 2018] Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[Somasundaram and Kalavathi, 2011] Karuppanagounder Somasundaram and Palanisamy Kalavathi. Medical image contrast enhancement based on gamma correction. *International Journal of Knowledge Management and e-Learning*, 3:15–18, 02 2011.

[Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, J. Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *international conference on learning representations*, 2014.

[Veluchamy and Subramani, 2019] Magudeeswaran Veluchamy and Bharath Subramani. Image contrast and color enhancement using adaptive gamma correction and histogram equalization. *Optik*, 183:329–337, 2019.

[Wong and Kolter, 2018] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5286–5295, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[Zhang *et al.*, 2020] Dinghuai Zhang, Mao Ye, Chengyue Gong, Zhanxing Zhu, and Qiang Liu. Black-box certification with randomized smoothing: A functional optimization based framework. *Advances in Neural Information Processing Systems*, 33:2316–2326, 2020.