2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 979-8-3503-0249-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVPRW59228.2023.00225

Certified Adversarial Robustness Within Multiple Perturbation Bounds

Soumalya Nandi Sravanti Addepalli * Harsh Rangwani * R. Venkatesh Babu
Vision and AI Lab, Indian Institute of Science, Bengaluru

Abstract

Randomized smoothing (RS) is a well known certified defense against adversarial attacks, which creates a smoothed classifier by predicting the most likely class under random noise perturbations of inputs during inference. While initial work focused on robustness to ℓ_2 norm perturbations using noise sampled from a Gaussian distribution, subsequent works have shown that different noise distributions can result in robustness to other ℓ_p norm bounds as well. In general, a specific noise distribution is optimal for defending against a given ℓ_p norm based attack. In this work, we aim to improve the certified adversarial robustness against multiple perturbation bounds simultaneously. Towards this, we firstly present a novel certification scheme, that effectively combines the certificates obtained using different noise distributions to obtain optimal results against multiple perturbation bounds. We further propose a novel training noise distribution along with a regularized training scheme to improve the certification within both ℓ_1 and ℓ_2 perturbation norms simultaneously. Contrary to prior works, we compare the certified robustness of different training algorithms across the same natural (clean) accuracy, rather than across fixed noise levels used for training and certification. We also empirically invalidate the argument that training and certifying the classifier with the same amount of noise gives the best results. The proposed approach achieves improvements on the ACR (Average Certified Radius) metric across both ℓ_1 and ℓ_2 perturbation bounds. Code available at https://github.com/val-iisc/NU-Certified-Robustness

1. Introduction

Deep neural networks are vulnerable to carefully crafted input perturbations known as adversarial attacks [3, 30]. These perturbations are imperceptible to human eyes, but are sufficient to fool the classifier into making wrong predictions. In order to improve the robustness of models against such attacks, one line of research is on *empirical defenses*, which augments the training data with adversarial attacks during training [13, 23, 34, 35, 40]. Another line of research is on *certified defenses*, which gives mathematically proven probabilistic guarantees within guarded areas around an input x, such that any perturbation within that region fails to deceive the network [7, 25, 33]. *Randomized smoothing* (RS) [7] is one such certified defense that makes a network certifiably robust within an ℓ_2 norm ball by creating a smoothed version of a base classifier. This is achieved by using zero mean additive Gaussian noise augmentations of the inputs during both training and inference.

In this work, we firstly highlight the following shortcomings observed in the current literature of randomized smoothing, and further propose solutions to address the same: 1) Existing defenses focus mostly on achieving certified robustness within a given threat model [7, 37]. However, for real-life applications, it is necessary to improve the robustness of models against more than one ℓ_p norm threat models simultaneously [9, 31]. 2) Prior works [7] recommend using the same noise magnitude during both training and certification for optimal performance, but we empirically demonstrate that this does not hold true in practice (Table 3). 3) It has been a standard practice to compare various defenses against a common magnitude of noise that is used for training/ certification [7, 18]. However, we show that this may not be the best way of comparing defenses, since the use of different training/ certification schemes may result in different robustness-accuracy trade-offs, making it hard to compare defenses from an end user perspective. For example, an improved robustness at a lower clean accuracy may not necessarily mean improved performance overall. Following are our contributions to address the above issues:

- We propose a novel certification scheme, that combines the benefits of *Gaussian* and *Uniform* noise based smoothed classifiers to create a stronger hybrid smoothed classifier in terms of both ℓ_1 and ℓ_2 norm robustness guarantees.
- Contrary to existing works, we show that using the same noise during both training and inference does not always lead to optimal robustness. Further, we propose to compare defenses by targeting a fixed value of clean accuracy, rather than by fixing the noise used for training/ inference which, can be misleading.

*Equal Contribution

- We propose the use of *Normal-Uniform* distribution as the training noise distribution, alongside a regularizer that enforces similarity between the training and inference noise distributions. Using the proposed approach, we demonstrate improved certified robustness against both ℓ_2 and ℓ_1 attacks simultaneously.

2. Background

2.1. Randomized Smoothing

A base classifier f is first trained using cross-entropy loss on images with Gaussian noise based augmentations, where noise sampled from $\mathcal{N}(0, \sigma^2 I)$ is added to every image. During certification, f is transformed to a smoothed classifier g such that for a given test image x, g outputs the most probable class across different noise augmentations of x, using noise sampled from the same distribution as that used for training. Since it is not feasible to compute this probability exactly for Neural Networks, it is estimated using a random sample of n noise vectors generated from the smoothing distribution for each test image x. The augmented images are passed through the trained base classifier to obtain n predictions for x. Let $\hat{c}_A$ be the class predicted the most number of times and n_A be the number of times $\hat{c}_A$ has been predicted, then the estimated probability is $p_A = n_A/n$. Next a one-sided $(1-\alpha)$ lower confidence bound of p_A is calculated as $\underline{p_A}$. If $\underline{p_A} > 0.5$, then the ℓ_2 norm robust radius is returned as $\sigma\Phi^{-1}(\underline{p_A})$, otherwise the sample is said to be *ABSTAINED*. Cohen *et al*. [7] use $n = 100,000$ and $\alpha = 0.001$, resulting in a 0.1% chance of returning a falsely certified data point. More details on the computation of certified radius for ℓ_1 and ℓ_2 threat models are presented in Section-1.1 of the Supplementary.

2.2. Evaluation metrics for Certified Robustness

We use the following metrics for evaluation: *Clean accuracy* and *robustness*. Clean accuracy refers to the accuracy on natural unperturbed images. For robustness, we use a metric called *Average Certified Radius* or ACR, introduced by Zhai *et al*. [38]. ACR is the average value of certified radii of all the considered test data points as shown below:

$$ACR := \frac{1}{|D_{test}|} \sum_{(x,y)\in D_{test}} CR(f,x).\mathbf{1}_{g(x)=y} \quad (1)$$

where D_{test} is the considered test dataset, $CR(f,x)$ is the certified radius of x for base classifier f, and $\mathbf{1}_{g(x)=y}$ is the indicator function for correct classification. As the noise level σ is increased, some data points will have higher certified radii, but the number of wrongly classified data points would also increase. For misclassified data points we would have, $CR(f,x).\mathbf{1}_{g(x)=y} = 0$. So increasing the noise magnitude σ, does not necessarily increase the ACR. Hence in addition to robustness, ACR also captures the robustness-accuracy trade-off of a model.

3. Related Works

3.1. Empirical Defenses

Empirical Defenses are heuristic based approaches that achieve adversarial robustness without specific mathematical guarantees. Their adversarial robustness is thus evaluated against strong empirical attacks such as AutoAttack [10] and GAMA [28]. Adversarial training [13, 23, 34] is one such defense, which maximizes a classification loss to generate adversarial attacks and minimizes the loss on the attacked images for training. Although empirical defenses achieve the best possible robustness today [35, 40], they do not provide robustness guarantees, which may be crucial for security critical applications such as autonomous driving. In the history of empirical defenses, several early methods which used gradient obfuscation as a defense strategy [4, 15, 27, 36] have later been broken later by stronger attacks [2], highlighting the need for robustness guarantees.

3.2. Certified Defenses

Certified defenses provide provable probabilistic or exact guarantees [7, 25, 26, 33, 37] to ensure that for any input x, the classifier's output is consistent within a defined neighbourhood around x, i.e., any attack within this radius is unsuccessful to deceive the network. Certified defenses can be categorized as, i) Exact methods [5,6,12,17,19,22], ii) Convex Optimization based methods [8,14,24,32,33,39], and iii) Randomized Smoothing [7, 37] based methods. The former two categories are highly computationally expensive and architecture dependent, thus not scalable to large networks. Being architecture independent, randomized smoothing has an unparalleled advantage over the other two categories in this regard. Cohen *et al*. [7] provide a tight ℓ_2 certified robustness guarantee and Yang *et al*. [37] provide the same for an ℓ_1 adversary along with the theoretical foundation.

3.3. Robustness to Multiple Threat Models

Most existing defenses aim to achieve robustness against only a single type of adversary, such as attacks constrained within a given ℓ_p norm bound. Tramer *et al*. [31] propose an empirical defense against more than one ℓ_p norm attacks simultaneously, by performing adversarial training on all or the worst attack amongst all considered threat models for the given sample. Croce and Hein [11] propose a defense that utilizes the geometrical relation between different ℓ_p norm balls to obtain robustness against a union of ℓ_p norm threat models. The authors also introduce a similar framework [9] for provable robustness against multiple ℓ_p norm perturbations. However, to the best of our knowledge, certi-

fied robustness guarantee for more than one ℓ_p norm bound using randomized smoothing is unexplored.

3.4. Randomized Smoothing (RS) based methods

Cohen *et al*. [7] recommend training and certifying with the same noise augmentations to achieve optimal performance. Yang *et al*. [37] also use the same strategy of training and certifying with the same noise magnitude. Few recent works [18, 38] propose the use of regularized robust training methods to improve the certified radius of the smoothed classifier. Zhai *et al*. [38] propose MACER, which uses a robustness loss in addition to the standard training framework provided by Cohen *et al*. [7], to maximize the robust radius. Jeong and Shin [18] propose a regularizer that controls the prediction consistency over noise to increase the certified robustness of smoothed classifiers, while also controlling the accuracy-robustness trade-off. Motivated by this, we propose a regularizer that is better suited to the proposed defense of using Normal-Uniform noise during training, and a combination of Normal and Uniform noise during certification. Although these works differ in their training methods, they also perform training and certification with the same noise augmentations and compare results across fixed levels of noise magnitudes. We highlight the limitations of the same (Table 3) and propose a novel comparison perspective across different methods.

4. Proposed Approach

We now discuss the various contributions of this work during training, certification and inference in greater detail.

4.1. Proposed Certification Method for robustness within multiple threat models

As noted in prior works [7, 37], the optimal smoothing distribution for ℓ_1 norm robustness is Uniform distribution and that for ℓ_2 norm robustness is Gaussian distribution. In order to achieve robustness within both ℓ_1 and ℓ_2 norm bounds, we perform certification twice using noise sampled from Uniform and Gaussian distributions respectively. While the certification process for a single threat model is straightforward, combining the certificates obtained for different smoothing distributions effectively is non-trivial. Towards this, we propose a novel strategy to effectively combine the predictions and certifications obtained from the individual smoothing distributions. A Uniform (Gaussian) smoothed classifier gives 4 outcomes for every test image: confidence in the predicted label (Abstained or not), predicted label y_U (y_G), ℓ_1 norm certified radius $r_U^{l_1}$ ($r_G^{l_1}$) and ℓ_2 norm certified radius $r_U^{l_2}$ ($r_G^{l_2}$). We build a hybrid smoothed classifier from these 8 outputs using the rules presented in Table 1 depending on the different possibilities arising from the predictions and their confidence. Table 2 shows that the proposed certification method improves the ℓ_1 ACR when compared to the use of Gaussian smoothing alone for certification, without diminishing the ℓ_2 ACR. The clean accuracy also improves using this approach.

Table 1. Proposed certification method that combines the certificates of the two smoothed classifiers to build a stronger hybrid classifier with improved robustness in both ℓ_1 and ℓ_2 norm bounds.

CATEGORY	ℓ_1 RADIUS	ℓ_2 RADIUS	PREDICTED LABEL
$y_U = y_G$	$Max(r_U^{l_1}, r_G^{l_1})$	$Max(r_U^{l_2}, r_G^{l_2})$	COMMON LABEL
ONLY y_U IS ABSTAINED	$r_G^{l_1}$	$r_G^{l_2}$	y_G
ONLY y_G IS ABSTAINED	$r_U^{l_1}$	$r_U^{l_2}$	y_U
BOTH ARE ABSTAINED	0	0	ABSTAINED
$y_U \neq y_G$	0	0	ABSTAINED

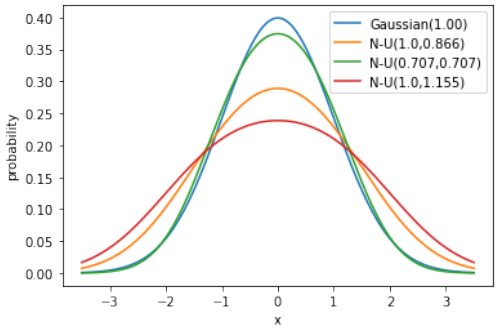

Figure 1. Shape of the proposed Normal-Uniform distribution for different kurtosis values.

4.2. Noise Magnitude for training and inference

We empirically show that the use of same noise magnitude (σ) for both training and inference does not always result in the best overall performance. As shown in Table 3, both clean accuracy and ACR improve when a lower σ is used during certification. When a consistency regularizer is used, a higher σ at test time improves robust accuracy with a marginal drop in clean accuracy. We thus select the best combination of train and test σ, rather than using the same value for both.

4.3. Normal-Uniform Noise Distribution

In this work, we aim to achieve improved adversarial robustness to ℓ_2 norm perturbations using Gaussian smoothing and to ℓ_l norm perturbations using Uniform smoothing simultaneously. Towards this, we propose to use a combination of both distributions during training, which we refer to as the *Normal-Uniform* distribution as shown in Figure 1. We consider two independent random variables, $X \sim N(0, \sigma_N^2)$, and $Y \sim U(-\sqrt{3}\sigma_U, \sqrt{3}\sigma_U)$. Then, $Z = X + Y$ is defined to follow a *Normal-Uniform* distribution with parameters (σ_N, σ_U).

The use of Normal-Uniform noise for training allows the model to be trained on noise sampled from both Gaussian and Uniform distributions simultaneously, while also allowing the choice of parameters such that Z has a negative

Table 2. The proposed certification strategy can significantly increase the ℓ_1 ACR without compromising on the ℓ_2 ACR.

TRAINING	CERTIFICATION	CLEAN ACC	ℓ_2 ACR	ℓ_1 ACR
$NU(\sigma_N = 0.50, \sigma_U = 0.433)$+$R_S(\beta = 3)$	$Gaussian(\sigma = 0.60)$	59.30	0.742	0.742
$NU(\sigma_N = 0.50, \sigma_U = 0.433)$+$R_S(\beta = 3)$	$Gaussian(\sigma = 0.60) + Uniform(\sigma = 0.65)$	60.26	0.742	**0.783**
$NU(\sigma_N = 1.00, \sigma_U = 0.866)$+$R_S(\beta = 2)$	$Gaussian(\sigma = 1.00)$	44.49	0.769	0.769
$NU(\sigma_N = 1.00, \sigma_U = 0.866)$+$R_S(\beta = 2)$	$Gaussian(\sigma = 1.00) + Uniform(\sigma = 1.16)$	45.13	0.769	**0.823**

Table 3. The use of different noise magnitudes (σ) during training and testing may result in better overall performance as shown on full testset of CIFAR-10 for Gaussian smoothing [7] with and without consistency regularization [18].

METHOD	TRAIN σ	TEST σ	CLEAN-ACC	ℓ_2 ACR
GAUSSIAN	1.00	1.00	45.90	0.492
GAUSSIAN	1.00	0.80	47.00	**0.531**
GAUSSIAN + CONSISTENCY REG.	0.25	0.25	74.41	0.545
GAUSSIAN + CONSISTENCY REG.	0.25	0.30	72.81	**0.584**

Table 4. Robustness to perturbations on CIFAR-10: Negative kurtosis K(X) coupled with a higher training σ can have significantly better ACR. The last two rows in each block use the same noise magnitude, showing that negative kurtosis improves performance.

TRAINING DISTRIBUTION(X)	TRAINING σ	K(X)	TEST σ	CLEAN-ACC	ℓ_2 ACR
GAUSSIAN ($\sigma = 0.25$) [7]	0.250	0.00		77.64	0.429
GAUSSIAN ($\sigma = 0.331$)	0.331	0.00	0.25	67.95	0.391
NU ($\sigma_N = 0.25, \sigma_U = 0.217$)	0.331	-0.22		75.20	**0.450**
GAUSSIAN ($\sigma = 0.50$) [7]	0.500	0.00		64.19	0.508
GAUSSIAN ($\sigma = 0.661$)	0.661	0.00	0.50	56.89	0.526
NU ($\sigma_N = 0.50, \sigma_U = 0.433$)	0.661	-0.22		61.26	**0.560**
GAUSSIAN ($\sigma = 1.00$) [7]	1.000	0.00		45.90	0.492
GAUSSIAN($\sigma = 1.323$)	1.323	0.00	1.00	37.71	0.477
NU ($\sigma_N = 1.00, \sigma_U = 0.866$)	1.323	-0.22		45.70	**0.592**

kurtosis. The **kurtosis** of a probability distribution is the measure of its shape in terms of its tailedness. A distribution with negative kurtosis (called *platykurtic distribution*) has thinner tails than that of a Gaussian distribution, i.e., the distribution is lesser prone to generating extreme values when compared to the Gaussian distribution. The use of a platykurtic distribution during training reduces the probability of generating outliers, thereby improving the training stability as shown in Table 4, where ACR is higher using the proposed distribution. The last two rows in each block use the same noise magnitude, showing that the negative kurtosis improves performance. More details on **kurtosis** are provided in Section-1.3 of the Supplementary.

4.4. Similarity Regularizer

Consistency regularizers are commonly used in the literature of adversarial defenses [1, 28, 29, 40]. In empirical defenses, Zhang *et al*. [41] introduce TRADES regularizer that enforces similarity between the outputs of a clean image and the corresponding perturbed image. Li *et al*. [21] introduce *stability training* for certified robustness in the same spirit as TRADES, where a Gaussian noise augmented image is used instead of the adversarially perturbed image.

Table 5. Effect of the proposed *similarity regularizer* R_S when Normal-Uniform noise is used during training, along with Gaussian noise during certification on CIFAR-10.

TRAINING	CERTIFICATION	CLEAN ACC	ℓ_2 ACR
$NU(\sigma_N = 0.50, \sigma_U = 0.433)$	$\mathcal{N}(0.60^2)$	61.77	0.570
$NU(\sigma_N = 0.50, \sigma_U = 0.433)$+$R_S(\beta = 3)$		59.30	**0.742**
$NU(\sigma_N = 1.00, \sigma_U = 0.866)$	$\mathcal{N}(1.00^2)$	45.70	0.592
$NU(\sigma_N = 1.00, \sigma_U = 0.866)$+$R_S(\beta = 2)$		44.49	**0.769**

Zhai *et al*. [38] introduce MACER, that uses a regularizer to maximize the certified radius for ℓ_2 norm perturbations. The current state-of-the-art method [18] also uses a consistency regularizer to enforce consistency in predictions over noisy samples. In this work, we introduce the following *similarity regularizer* which is based on KL-divergence,

$$R_S = \mathrm{KL}(f(x + \mathbf{NU}) || f(x + \mathbf{N})) + \mathrm{KL}(f(x + \mathbf{NU}) || f(x + \mathbf{U})) \quad (2)$$

where, f is the base classifier that outputs the K dimensional probability vector, $\mathbf{NU}$, $\mathbf{N}$ and $\mathbf{U}$ are noise vectors sampled from the Normal-Uniform(σ_N, σ_U), Normal$(0, \sigma_N^2)$ and Uniform$(-\sqrt{3}\sigma_U, \sqrt{3}\sigma_U)$ distributions respectively. σ_N and σ_U are the standard deviations of Gaussian and Uniform sub-parts of the Normal-Uniform distribution respectively.

Although we propose to use the Normal-Uniform noise distribution during training, we obtain certification using noise from the individual Normal and Uniform distributions as recommended by Cohen *et al*. [7] and Yang *et al*. [37] respectively. The certificates obtained are further combined using the process discussed in Section-4.1. Since the noise distributions used during training and certification are different, the proposed regularizer (Eq.2) helps in aligning predictions of the Normal-Uniform corrupted images with that of each distribution individually, thereby aligning the training and certification stages. The overall loss function used in the proposed method is shown below, where $\mathcal{L}_{\mathrm{CE}}(.)$ is the cross entropy loss and β is a hyperparamater.

$$L := \mathcal{L}_{\mathrm{CE}}(f(x + \mathbf{NU}), y) + \beta \cdot R_S \quad (3)$$

As shown in Table 5, on using the proposed regularizer R_S, we observe a significant boost in the ACR against ℓ_2 norm perturbations at slightly lower clean accuracy, when Gaussian noise based smoothing is used for certification.

4.5. Comparison Strategy across Methods

A standard practice of comparing robust classifiers across different training strategies is to use a common noise magnitude (σ) during inference. However, the use of different training strategies can result in vastly different clean and robust performance metrics for the same noise magnitude used during inference, making it very difficult to compare a method with higher clean accuracy and lower ACR against a method with lower clean accuracy and better ACR. Given that variation of training and inference noise magnitude does give the flexibility of trading off robust accuracy with clean accuracy, from an end user perspective, it is important to know which method gives the best ACR for a specified value of clean accuracy. We therefore propose to firstly tune the training and certification σ to achieve the required level of clean accuracy with optimal ACR, and further compare the ACR across different methods.

5. Experiments and Results

5.1. Experimental Setup

We present an empirical evaluation of the proposed approach on the full test set of CIFAR-10 [20] unless specified otherwise. As discussed in Section-2.2, we evaluate the performance of models on their clean accuracy and Average Certified radius or ACR. To reproduce results from the baselines, we use the codes officially released by the authors. We use the model architecture as ResNet-110 [16], as used in the baselines. Our proposed model is trained for 300 epochs with a batch size of 400. We use cosine learning rate schedule with SGD optimizer. In this work, we refer to the Gaussian Smoothing baseline proposed by Cohen *et al*. [7] as *Gaussian* and the Consistency Regularization baseline by Jeong *et al*. [18] as *Gaussian+Consistency* or *Gaussian+cons*.

5.2. Comparing Different Smoothing Measures

As discussed in Section-4.5, comparison across different approaches was done against a fixed value of smoothing σ in prior works [18, 38]. In Section-4.5, we motivate the need for comparing against a fixed value of clean accuracy instead. We now present another scenario where comparing across a fixed noise σ can give misleading results.

We consider a comparison between methods that use different smoothing distributions during inference. For example, Yang *et al*. [37] theoretically show that Uniform distribution is optimal for robustness within ℓ_1 norm bound. However, it is not straightforward to show this in practice by comparing two methods that use different noise distributions for certification, as shown in Table 6. Merely comparing two methods against the same inference noise magnitude using the ℓ_1 ACR metric gives an impression that Gaussian smoothing is a better approach. However, one cannot conclude which smoothing is better overall because they create two different types of robustness-accuracy trade-offs as shown in the left image of Figure 2. There are approximately 40% data-points which have an ℓ_1 certified radius greater than or equal to 0.5 under Gaussian smoothing, but with uniform smoothing all the data-points have certified radii less than 0.5. On the other hand, the latter gives more correctly classified data-points than the former, making it hard to interpret which approach is better. Therefore, comparing two differently smoothed classifiers by fixing the smoothing noise level apriori fails to capture the true robustness-accuracy trade-off. The plot on the right in Figure 2, and results in Table 7 clearly show that smoothing with Uniform noise gives better ℓ_1 robustness, which rightly reflects the theoretical results by Yang *et al*. [37]. The proposed comparison method also works effectively using the proposed approach where we combine certifications from two different smoothing strategies during inference, and hence the noise level of the hybrid classifier cannot be determined easily.

Table 6. Comparing ℓ_1 certified robustness using noise from Gaussian and Uniform distributions for training and certification: While in theory, Uniform noise is known to be better, merely comparing ACR using a fixed noise magnitude for certification gives a false impression that Gaussian smoothing is optimal.

σ	MODEL	CLEAN-ACC	ℓ_1 ACR
0.25	GAUSSIAN SMOOTHING	76.40	0.430
0.25	UNIFORM SMOOTHING	86.40	0.331

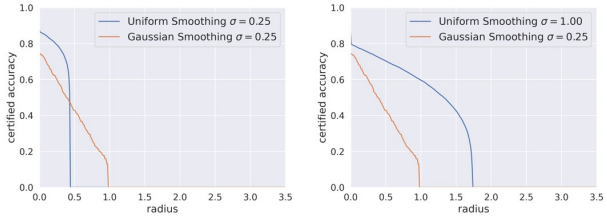

Figure 2. Comparison of ℓ_1 norm robustness-accuracy trade-off using Gaussian and Uniform smoothing for (**Left**) a fixed noise magnitude (σ) of 0.25. (**Right**) a fixed clean accuracy of 75%.

Table 7. Comparison of Gaussian and Uniform Smoothing against fixed clean accuracy clearly shows that Uniform distribution is better for ℓ_1 certified robustness as shown by Yang *et al*. [37].

σ	MODEL	CLEAN-ACC	ℓ_1 ACR
0.25	GAUSSIAN SMOOTHING	76.40	0.430
1.00	UNIFORM SMOOTHING	75.20	**0.967**

5.3. SOTA Comparison Results

The results presented so far have focused on highlighting the individual contributions of our work. In this section, we present the overall results using all the rec-

Table 8. Performance of the proposed approach when compared to baselines across different fixed levels of clean accuracy (in each block), in terms of ℓ_1, ℓ_2 and average ACR on the full test set of CIFAR-10.

TRAINING SCHEME	CERTIFICATION	CLEAN ACC	ℓ_1 ACR	ℓ_2 ACR	AVG ACR
$Gaussian(\sigma = 1.00)$ [7]	$Gaussian(\sigma = 1.00)$	45.90	0.492	0.492	0.492
$Gaussian(\sigma = 1.00)$+CONS [18]	$Gaussian(\sigma = 1.00)$	45.96	0.762	0.762	0.762
$NU(\sigma_N = 1.00, \sigma_U = 0.866)$+$R_S(\beta = 2)$(OURS)	$Gaussian(\sigma = 1.00) + Unif(\sigma = 1.16)$(OURS)	45.13	**0.823**	**0.769**	**0.796**
$Gaussian(\sigma = 0.84)$ [7]	$Gaussian(\sigma = 0.84)$	50.62	0.513	0.513	0.513
$Gaussian(\sigma = 0.81)$+CONS [18]	$Gaussian(\sigma = 0.81)$	50.39	0.762	0.762	0.762
$NU(\sigma_N = 0.75, \sigma_U = 0.65)$+$R_S(\beta = 3)$(OURS)	$Gaussian(\sigma = 0.75) + Unif(\sigma = 0.90)$(OURS)	51.07	**0.824**	**0.768**	**0.796**
$Gaussian(\sigma = 0.73)$ [7]	$Gaussian(\sigma = 0.73)$	55.98	0.521	0.521	0.521
$Gaussian(\sigma = 0.68)$+CONS [18]	$Gaussian(\sigma = 0.68)$	55.52	0.761	**0.761**	0.761
$NU(\sigma_N = 0.60, \sigma_U = 0.52)$+$R_S(\beta = 4)$(OURS)	$Gaussian(\sigma = 0.60) + Unif(\sigma = 0.70)$(OURS)	55.80	**0.790**	0.762	**0.771**
$Gaussian(\sigma = 0.62)$ [7]	$Gaussian(\sigma = 0.62)$	60.34	0.527	0.527	0.527
$Gaussian(\sigma = 0.57)$+CONS [18]	$Gaussian(\sigma = 0.57)$	60.47	0.734	0.734	0.734
$NU(\sigma_N = 0.50, \sigma_U = 0.433)$+$R_S(\beta = 3)$(OURS)	$Gaussian(\sigma = 0.60) + Unif(\sigma = 0.65)$(OURS)	60.26	**0.783**	**0.742**	**0.762**
$Gaussian(\sigma = 0.50)$ [7]	$Gaussian(\sigma = 0.50)$	64.19	0.508	0.508	0.508
$Gaussian(\sigma = 0.48)$+CONS [18]	$Gaussian(\sigma = 0.48)$	64.91	0.707	0.707	0.707
$NU(\sigma_N = 0.40, \sigma_U = 0.346)$+$R_S(\beta = 3)$(OURS)	$Gaussian(\sigma = 0.50) + Unif(\sigma = 0.50)$(OURS)	65.27	**0.731**	**0.708**	**0.720**
$Gaussian(\sigma = 0.39)$ [7]	$Gaussian(\sigma = 0.39)$	70.37	0.503	0.503	0.503
$Gaussian(\sigma = 0.37)$+CONS [18]	$Gaussian(\sigma = 0.37)$	70.78	0.648	**0.648**	0.648
$NU(\sigma_N = 0.30, \sigma_U = 0.260)$+$R_S(\beta = 2)$(OURS)	$Gaussian(\sigma = 0.40) + Unif(\sigma = 0.35)$(OURS)	70.91	**0.666**	0.637	**0.652**
$Gaussian(\sigma = 0.30)$ [7]	$Gaussian(\sigma = 0.30)$	74.26	0.455	0.455	0.455
$Gaussian(\sigma = 0.25)$+CONS [18]	$Gaussian(\sigma = 0.25)$	74.41	0.545	0.545	0.545
$NU(\sigma_N = 0.20, \sigma_U = 0.17)$+$R_S(\beta = 4)$(OURS)	$Gaussian(\sigma = 0.30) + Unif(\sigma = 0.25)$(OURS)	74.52	**0.577**	**0.556**	**0.566**

ommendations presented in this paper. During training we use the proposed Normal-Uniform (NU) distribution along with the proposed similarity regularizer (Equation 2). During certification we create two smoothed classifiers, one with Uniform smoothing and the other with Gaussian smoothing, and further combine them using the proposed certification method presented in Section 4.1. As discussed in Section 4.5, we compare across different methods against fixed levels of clean accuracy set to the following - $45\%, 50\%, 55\%, 60\%, 65\%, 70\%$ and 75%. As shown in Table 8, the proposed approach has significantly higher ACR in all cases when compared to the baselines. The training parameters σ_N and σ_U are chosen to set the kurtosis of the Normal-Uniform distribution to a negative value (-0.22), which gives the relation between σ_N and σ_U as $\sigma_U = \sqrt{3}\sigma_N/2$. We further tune the hyper-parameter (σ_N) to achieve the required level of clean accuracy. Experimentally, a low value of $\beta \in \{2, 3, 4\}$ for our proposed regularization (Equation 3) gives the best performance for the considered levels of clean accuracy. We further note from Table 8 that for Gaussian smoothing [7], reducing the clean accuracy below 60% by increasing the noise level results in a drop in ACR, which does not happen in the proposed method.

In the supplementary material, we present a detailed ablation study of the proposed method. We present the impact of variation in the hyperparameter β, impact of choice of regularizer and the effect of Kurtosis on the final performance of the certified classifier in terms of its clean accuracy, ℓ_1 ACR and ℓ_2 ACR.

6. Conclusion

In this work, we propose several aspects related to the training, certification and inference strategies for improving certified adversarial robustness within multiple perturbation bounds. Contrary to prior belief, we show that training and inference noise levels need not be the same to achieve optimal results. Further, we propose to compare different methods against a fixed clean accuracy rather that using a fixed noise level during inference. Thirdly, we propose an effective way of combining the certifications obtained using two different noise distributions. We next present the proposed training methodology of using noise sampled from the Normal-Uniform distribution during training, and ensuring similar representations during inference by using a consistency regularizer w.r.t. both Gaussian and Uniform distributions that are used during inference. We motivate the need for each of the individual strategies using several experimental results, and also present the impact of combining all discussed strategies to achieve state-of-the-art results in robustness against a combination of ℓ_1 and ℓ_2 perturbation bounds. We hope our work motivates further research on certified robustness within multiple perturbation bounds, which is very important in a real world setting.

7. Acknowledgments

This work was supported by the research grant CRG/2021/005925 from SERB, DST, Govt. of India. Sravanti is supported by Google PhD Fellowship, and Harsh is supported by Prime Minister's Research Fellowship.

References

[1] Sravanti Addepalli, B S Vivek, Arya Baburaj, Gaurang Sriramanan, and R Venkatesh Babu. Towards Achieving Adversarial Robustness by Enforcing Feature Consistency Across Bit Planes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning (ICML)*, 2018. 2

[3] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. *Lecture Notes in Computer Science*, page 387–402, 2013. 1

[4] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[5] Nicholas Carlini, Guy Katz, Clark Barrett, and David L. Dill. Provably minimally-distorted adversarial examples, 2017. 2

[6] Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess. Maximum resilience of artificial neural networks, 2017. 2

[7] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320. PMLR, 09–15 Jun 2019. 1, 2, 3, 4, 5, 6

[8] Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of relu networks via maximization of linear regions, 2018. 2

[9] Francesco Croce and Matthias Hein. Provable robustness against all adversarial l_p-perturbations for $p \geq 1$, 2020. 1, 2

[10] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020. 2

[11] Francesco Croce and Matthias Hein. Adversarial robustness against multiple l_p-threat models at the price of one and how to quickly fine-tune robust models to another threat model, 2021. 2

[12] Souradeep Dutta, Susmit Jha, Sriram Sanakaranarayanan, and Ashish Tiwari. Output range analysis for deep neural networks, 2017. 2

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015. 1, 2

[14] 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, 2018. 2

[15] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 5

[17] Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks, 2016. 2

[18] Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers, 2020. 1, 3, 4, 5, 6

[19] Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks, 2017. 2

[20] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009. 5

[21] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise, 2019. 4

[22] Alessio Lomuscio and Lalit Maganti. An approach to reachability analysis for feed-forward relu neural networks, 2017. 2

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 2

[24] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. 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 3578–3586. PMLR, 10–15 Jul 2018. 2

[25] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples, 2020. 1, 2

[26] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. 2

[27] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[28] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R Venkatesh Babu. Guided Adversarial Attack for Evaluating and Enhancing Adversarial Defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 4

[29] Gaurang Sriramanan, Sravanti Addepalli, Arya Baburaj, and R Venkatesh Babu. Towards Efficient and Effective Adversarial Training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 4

[30] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks, 2014. 1

[31] Florian Tramer and Dan Boneh. Adversarial training and robustness for multiple perturbations. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[32] Shiqi Wang, Yizheng Chen, Ahmed Abdou, and Suman Jana. Mixtrain: Scalable training of verifiably robust neural networks, 2018. 2

[33] Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope, 2018. 1, 2

[34] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2

[35] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 1, 2

[36] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[37] Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes, 2020. 1, 2, 3, 4, 5

[38] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius, 2020. 2, 3, 4, 5

[39] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions, 2018. 2

[40] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, 2019. 1, 2, 4

[41] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy, 2019. 4