Supplementary Material: 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

1. Background

1.1. Randomized Smoothing

Let $f : \mathbb{R}^d \rightarrow \{1, 2, 3, ..., K\}$ be a neural network that maps a d dimensional image to one of the K classes. Using Randomized Smoothing, the base classifier $f(x)$ can be transformed to a smoothed classifier $g(x)$ that has inherent probabilistic certified guarantees. Given an input x, the smoothed classifier $g(x)$ outputs the most likely class as predicted by the base classifier, across different augmentations of the input image, as shown below:

$$g(x) = \underset{c}{\text{argmax}}\, P[f(x+\varepsilon) = c] \tag{1}$$

Here, ε is generated from a smoothing measure μ. Considering μ to be isotropic Gaussian, Cohen *et al.* [1] show that $g(x)$ inherits certified robustness in ℓ_2 norm through the following theorem.

Theorem 1. *(Restating theorem* 1 *by Cohen* et al. *[1]): Let $\varepsilon \sim N(0, \sigma^2 I)$. Suppose $c_A \in \{1, 2, 3, ..., K\}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ satisfy: $P(f(x+\varepsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq max_{c \neq c_A} P(f(x+\varepsilon) = c)$. Then $g(x+\delta) = c_A$ for all $||\delta||_2 < R$, where $R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$, Φ^{-1} being the inverse of standard Gaussian CDF.*

Yang *et al.* [3] show using the following theorem that by considering μ as a Uniform distribution, $g(x)$ has provable robustness guarantee against ℓ_1 norm constrained attacks.

Theorem 2. *(Restating theorem $I.8$ by Yang* et al. *[3]): Suppose H is a smoothed classifier smoothed by the uniform distribution on the cube $[-\lambda, \lambda]^d$, such that $H(x) = (H(x)_1, ..., H(x)_C)$ is a vector of probabilities that H assigns to each class $1, ..., C$. If H correctly predicts the class y on input x, and the probability of the correct class is $\rho \overset{def}{=} H(x)_y > 1/2$, then H continues to predict the correct class when x is perturbed by any η with $||\eta||_1 < 2\lambda(\rho - 0.5)$.*

1.2. Consistency Regularization

Jeong and Shin [2] attempts to achieve better generalization performance of the base classifier over noise augmentation. Since, during certification the model is evaluated on noise augmented inputs, it is logical to use noise augmented inputs for training also, but the variance of the noise distribution hampers the stability of the training process. This work introduces a regularizer on top of the standard cross entropy loss that controls the prediction consistency over noisy samples. The overall loss function is,

$$L := \frac{1}{m} \sum_i (\mathcal{L}_{CE}(F(x+\delta_i), y) + \lambda \cdot KL(\hat{F}(x)||F(x+\delta_i)) + \eta \cdot H(\hat{F}(x)) \tag{2}$$

where $KL(\cdot||\cdot)$ is the KL-divergence term and $H(\cdot)$ is the entropy term. $F(x)$ is the differentiable function on which the classifier f is built, δ is Gaussian noise and $\hat{F}(x) = \frac{1}{m}\sum_i F(x+\delta_i)$. λ and η are hyperparameters. $\mathcal{L}_{CE}$ is the cross entropy function. The KL term reduces the variance of the predictions while the entropy term prevents the variance to become 0. In the paper, m is fixed as 2 and η is fixed at 0.5. The certification process is exactly same as that of *Gaussian Smoothing*.

1.3. Kurtosis of a distribution

Let X be a real valued random variable, then the kurtosis, $K(X)$ of the probability distribution of X is defined as,

$$K(X) = \frac{\mu_4}{\mu_2^2} - 3 \tag{3}$$

where μ_i is the i^{th} ordered central moment of X, i.e.,

$$\mu_i = E[X - E[X]]^i \tag{4}$$

Kurtosis measures the shape of a distribution in terms of its tailedness. If a probability distribution has fat tails, that is the random variable X has a good amount of area under the curve on its tails then for points in that region, $X - E[X]$ would be large in magnitude. So, $[X - E[X]]^4$ would produce even larger positive values. Therefore a high value of

[*] Equal Contribution

$E[X - E[X]]^4$ or μ_4 denotes a fat tailed distribution. The very similar argument follows to conclude that a random variable with low value of μ_4 has a thin tailed probability distribution. In general a standardized metric is used hence, $\frac{\mu_4}{\mu_2^2}$. The above metric $K(X)$ is used to measure the tailedness of a distribution relative to Gaussian distribution. For a normal distribution, $\frac{\mu_4}{\mu_2^2} = 3$, so $K(X)$ becomes 0 and the distribution is called a *mesokurtic* distribution. A distribution with a negative $K(X)$ is called a *platykurtic* distribution and has thinner tails than that of a Gaussian distribution, whereas a distribution with a positive $K(X)$ is called a *leptokurtic* distribution which has fatter tails than a Gaussian distribution. A platykurtic distribution is less prone to generate outliers than a Gaussian distribution while a leptokurtic distribution produces outliers with a higher probability than that of a Gaussian distribution. For example, Uniform distribution is a platykurtic distribution with $K(X) = -1.2$. It does not generate outliers at all, whereas a standard Laplace distribution has $K(X) = 3$ which generates much more outliers than Gaussian distribution.

2. Theoretical properties of Normal-Uniform distribution

Let $X \sim N(0, \sigma_N^2)$ and $Y \sim U(-\lambda, \lambda)$ independently, then $Z = X + Y$ will follow a *Normal-Uniform* distribution with parameters (σ_N, λ). The σ_N denotes the standard deviation of Normal distribution. For a $U(-\lambda, \lambda)$ distribution, the standard deviation is $\sigma_U = \frac{\lambda}{\sqrt{3}}$. Therefore to define both the distributions with the same parameter, we have used σ_U instead of λ in all our experiments. Hence the *Normal-Uniform* distribution is defined with the parameters (σ_N, σ_U). By controlling these parameters, we can effectively control the shape of the distribution (Figure 1) and can make it to behave more like a Gaussian distribution or a Uniform distribution accordingly (Figure 2).

Lemma 1. *If $X \sim \mathcal{N}(0, \sigma^2)$ and $Y \sim \mathcal{U}(-\lambda, \lambda)$ with X and Y being independent, then $Z = X + Y$ has the pdf as $f_Z(z) = \frac{1}{2\lambda}[\Phi(\frac{z+\lambda}{\sigma}) - \Phi(\frac{z-\lambda}{\sigma})]$ with $z \in (-\infty, \infty)$, where $\Phi(.)$ is the CDF of standard normal distribution.*

Proof. pdf of X is $f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$ with $x \in \mathbb{R}$. Pdf of Y is $f_Y(y) = \frac{1}{2\lambda}$ with $y \in (-\lambda, \lambda)$. Then, $F_Z(t) = P[Z \leq t]$

$$= \iint_{x+y \leq t} f_X(x) f_Y(y)\, dx\, dy$$
$$= \int_{-\infty}^{\infty} f_Y(y) \{\int_{-\infty}^{t-y} f_X(x) dx\} dy$$
$$= \int_{-\infty}^{\infty} f_Y(y) \Phi(\tfrac{t-y}{\sigma}) dy$$

So, the pdf is

$$f_Z(z) = \frac{1}{\sigma} \int_{-\infty}^{\infty} f_Y(y) \phi(\tfrac{t-y}{\sigma}) dy$$
$$= \frac{1}{\sigma} \int_{-\lambda}^{\lambda} \frac{1}{2\lambda} \phi(\tfrac{t-y}{\sigma}) dy = \frac{1}{2\lambda}[\Phi(\tfrac{z+\lambda}{\sigma}) - \Phi(\tfrac{z-\lambda}{\sigma})]$$

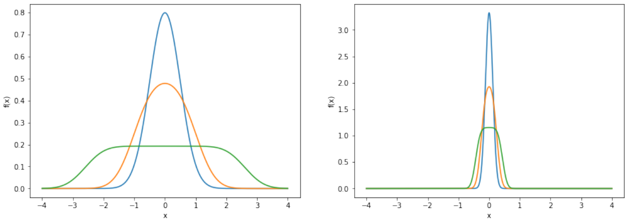

Figure 1. By controlling σ_N and σ_U, the shape of Normal-Uniform probability distribution function (pdf) can be adjusted from bell shaped to flat surfaced. **Left**: Comparison of pdf of Normal($\sigma_N = 0.5$)(Blue), Normal-Uniform($\sigma_N = 0.5, \sigma_U = 0.577$)(Orange), Normal-Uniform($\sigma_N = 0.5, \sigma_U = 1.500$)(Green) **Right**: Comparison of pdf of Normal($\sigma_N = 0.12$)(Blue), Normal-Uniform($0, \sigma_N = 0.12, \sigma_U = 0.144$)(Orange), Normal-Uniform($\sigma_N = 0.12, \sigma_U = 0.25$)(Green)

with $z \in (-\infty, \infty)$. □

Lemma 2. *If $X \sim N(0, \sigma^2)$ and $Y \sim U(-\lambda, \lambda)$ with X and Y being independent, then $Z = X + Y$ has the cdf as $F_Z(t) = \frac{\sigma}{2\lambda}[\frac{t+\lambda}{\sigma}\Phi(\frac{t+\lambda}{\sigma}) + \phi(\frac{t+\lambda}{\sigma}) - \frac{t-\lambda}{\sigma}\Phi(\frac{t-\lambda}{\sigma}) - \phi(\frac{t-\lambda}{\sigma})]$ with $t \in (-\infty, \infty)$, where $\Phi(.)$ and $\phi(.)$ are the CDF and pdf of standard normal distribution respectively.*

Proof. $F_Z(t) = P[X \leq t] = \int_{-\infty}^{t} \frac{1}{2\lambda}[\Phi(\frac{z+\lambda}{\sigma}) - \Phi(\frac{z-\lambda}{\sigma})]dx$

Let $I = \int_{-\infty}^{t} \Phi(\frac{x+\lambda}{\sigma})dx = \sigma \int_{-\infty}^{\frac{t+\lambda}{\sigma}} \Phi(z)dz$
$= \sigma[\Phi(z)z - \int \phi(z)zdz]_{-\infty}^{\frac{t+\lambda}{\sigma}} = \sigma[\Phi(z)z + \phi(z)]_{-\infty}^{\frac{t+\lambda}{\sigma}}$

$= \sigma[\Phi(\frac{t+\lambda}{\sigma})(\frac{t+\lambda}{\sigma}) + \phi(\frac{t+\lambda}{\sigma})]$
Similarly calculating for the second term and replacing in the original equation, we get
$F_Z(t) = \frac{\sigma}{2\lambda}[\frac{t+\lambda}{\sigma}\Phi(\frac{t+\lambda}{\sigma}) + \phi(\frac{t+\lambda}{\sigma}) - \frac{t-\lambda}{\sigma}\Phi(\frac{t-\lambda}{\sigma}) - \phi(\frac{t-\lambda}{\sigma})]$ □

Lemma 3. *If $X \sim N(0, \sigma^2)$ and $Y \sim U(-\lambda, \lambda)$ with X and Y being independent, then $Z = X + Y$ has $K(Z) = \frac{3\sigma^4 + 2\sigma^2\lambda^2 + (\lambda^4/5)}{(\sigma^2 + (\lambda^2/2))^2} - 3$*

Proof. We have $E(Z) = 0$ and $Var(Z) = Var(X) + Var(Y) = \sigma^2 + \frac{\lambda^2}{3}$
Now, $\mu_4 = E(Z - E(Z))^4 = E(Z)^4 = E[X^4 + 4X^3Y + 6X^2Y^2 + 4XY^3 + Y^4]$
$= E(X^4) + 6E(X^2)E(Y^2) + E(Y^4) = 3\sigma^4 + 6\sigma^2\frac{\lambda^2}{3} + E(Y^4)$

We have $E(Y^4) = \int_{-\lambda}^{\lambda} \frac{y^4}{2\lambda} dy = \frac{y^5}{10\lambda}|_{-\lambda}^{\lambda} = \frac{\lambda^4}{5}$

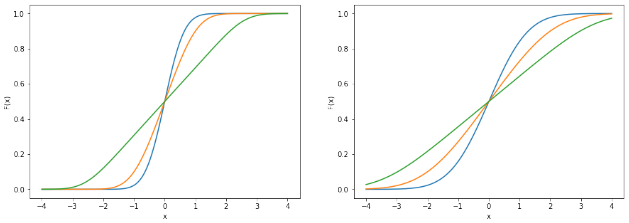

Figure 2. Similar to figure 1, by controlling the σ_N and σ_U, the shape of Normal-Uniform cdf can be adjusted from S shaped to a straightline. **Left**: Comparison of cdfs of Normal($\sigma_N = 0.5$)(Blue), Normal-Uniform($\sigma_N = 0.5, \sigma_U = 0.577$)(Orange), Normal-Uniform($\sigma_N = 0.5, \sigma_U = 1.500$)(Green) **Right**: Comparison of cdfs of Normal($\sigma_N = 1.0$)(Blue), Normal-Uniform($\sigma_N = 1.0, \sigma_U = 1.155$)(Orange), Normal-Uniform($\sigma_N = 1.0, \sigma_U = 2.00$)(Green)

so $K(Z) = \frac{\mu_4}{\mu_2^2} - 3 = \frac{3\sigma^4 + 2\sigma^2\lambda^2 + (\lambda^4/5)}{(\sigma^2 + (\lambda^2/2))^2} - 3$ □

We can vary the shape of the distribution from a bell shaped curve to flat surfaced curve by controlling the σ and λ appropriately as shown in figure 1.

3. Ablation study

3.1. Effect of tuning parameter β

We investigate the effect of the regularizer tuning parameter β. As usual, when we increase β, initially the ACRs increase and clean accuracy decreases. A prominent robustness-accuracy trade-off is visible in table 1. However, as we further increase β, the clean accuracy oscillates between 55% and 52%, while the ℓ_1 ACR gets stagnant around 0.770 and ℓ_2 ACR around 0.750 before both dropping drastically.

3.2. Effect of choice of KL term

Our proposed regularizer has two KL terms, each associated with one smoothed classifier. This results in 3 forward passes for each batch of images during training as we need 3 different outputs $F(x + NU), F(x + U), F(x + N)$ to calculate the regularizer. In this section we try out the following regularizers having only one KL term.

$$R_N = KL(F(x + NU)||F(x + N)) \quad (5)$$

$$R_U = KL(F(x + NU)||F(x + U)) \quad (6)$$

Table 2 shows the results under few setups. In most cases, proposed *similarity* regularizer dominates the robustness for comparable clean accuracy. The models trained with R_U regularizer provide better clean accuracy than R_S and also comparable ℓ_1 ACR. The ℓ_2 ACR drops though as can be seen in 3^{rd} and 4^{th} rows of table 2. On the other hand, regularizer R_N gives a better ℓ_2 ACR as compared to R_U at the cost of slight decrease in clean accuracy. The use of R_U, forces the training noise to behave more like a Uniform distribution, whereas R_N makes it more similar with Gaussian distribution. Table 4 shows that training with Gaussian noise and certifying with Uniform noise performs better than doing the opposite. That is why, using R_U results in higher decrease in ℓ_2 ACR than that of ℓ_1 ACR while using R_N. Our proposed *similarity* regularizer performs better in terms of both the ACRs at comparable clean accuracy as compared to both R_N and R_U. However, if one has to choose between R_N and R_U only, R_N is more preferable.

Table 1. Effect of β when trained on Normal-Uniform($\sigma_N = 0.50, \sigma_U = 0.433$) and certified on our proposed method on a subset of 500 test images of CIFAR10 with Gaussian smoothing($\sigma = 0.60$) + Uniform smoothing($\sigma = 0.65$).

β	Clean Acc	ℓ_1 ACR	ℓ_2 ACR
0	62.00	0.601	0.570
2	60.40	0.743	0.714
3	59.00	0.776	0.734
4	58.20	**0.779**	0.749
6	55.20	0.773	**0.751**
8	52.60	0.771	0.749
10	55.20	0.777	0.741
12	55.40	0.766	0.750
14	54.60	0.770	0.749
16	52.00	0.758	0.737
18	50.80	0.769	0.753
20	51.60	0.765	0.751
24	9.400	0.215	0.215

3.3. Effect of kurtosis

So far we have used a kurtosis value of -0.22 for all our proposed training experiments. As we know a Gaussian distribution has kurtosis value of 0 and a Uniform distribution has kurtosis value of -1.2, so when we introduce too much negative kurtosis in our training distribution, then it deviates from a Gaussian distribution and behaves more like a Uniform distribution (Figure 3). In the next subsection, we have evidence that training with Gaussian noise and then certifying with Uniform noise works better than training with Uniform noise and certifying with Gaussian noise. So keeping a mild negative kurtosis results in better robustness guarantees. Table 3 shows that decreasing the kurtosis further does not help in terms of the robustness-accuracy trade-off. The

Table 2. Effect of using a single KL term instead of two KL terms in the regularizer on a subset of 500 test images of CIFAR10.

Training	Certification	Clean Acc	ℓ_1 ACR	ℓ_2 ACR
$NU(\sigma_N = 0.50, \sigma_U = 0.433)$+$R_N(\beta = 6)$	$Gaussian(\sigma = 0.60) + Unif(\sigma = 0.650)$	60.00	0.770	0.746
$NU(\sigma_N = 0.50, \sigma_U = 0.433)$+$R_U(\beta = 6)$		58.40	0.778	0.710
$NU(\sigma_N = 0.50, \sigma_U = 0.433)$+$\mathbf{R_S}(\beta = 3)$		59.00	0.776	0.734
$NU(\sigma_N = 1.00, \sigma_U = 0.866)$+$R_N(\beta = 6)$	$Gaussian(\sigma = 1.00) + Unif(\sigma = 1.160)$	42.80	0.806	0.763
$NU(\sigma_N = 1.00, \sigma_U = 0.866)$+$R_U(\beta = 6)$		45.00	0.828	0.742
$NU(\sigma_N = 1.00, \sigma_U = 0.866)$+$\mathbf{R_S}(\beta = 4)$		44.00	0.858	0.789

Table 3. Effect of amount of negative kurtosis $K(X)$ in the training distribution when certified on our proposed method with Gaussian smoothing($\sigma = 1.00$) + Uniform smoothing($\sigma = 1.160$) when tested on a subset of 500 images of CIFAR10.

Training	$K(X)$	Clean Acc	ℓ_1 ACR	ℓ_2 ACR	Avg ACR
$NU(\sigma_N = 1.00, \sigma_U = 0.866)$	-0.22	45.00	0.671	0.625	**0.648**
$NU(\sigma_N = 1.00, \sigma_U = 1.155)$	-0.39	38.40	0.682	0.606	0.644

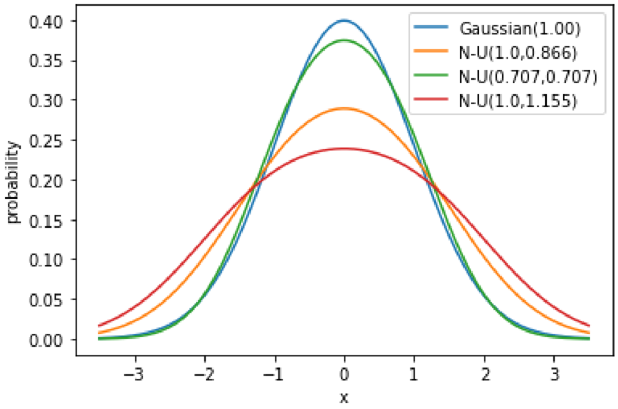

Figure 3. Shape of Normal-Uniform distribution for different kurtosis values.

choice of σ_N and σ_U may seem arbitrary, but a specific relation selects them. For first row, $\sigma_U = \frac{1.5}{\sqrt{3}}\sigma_N$, and for the last row, $\sigma_U = \frac{2}{\sqrt{3}}\sigma_N$.

3.4. Performance of the baselines

In this section we evaluate the performance of the baselines on our proposed certification method. As per the suggestions provided by the considered baselines [1, 2], the noise level σ is fixed at different levels in prior and then the model is trained with proposed methods from the considered baselines at that fixed noise level. For certification we create a Gaussian smoothed classifier and a Uniform smoothed classifier each with the same fixed noise level and combine the certificates as per our proposed method. Table 4 describes the results for $\sigma \in \{0.25, 0.50\}$ and shows that the performance is inferior to our proposed training noise distribution. We get an insignificant improvement in ℓ_1 ACR under Gaussian noise training when the model is certified with our proposed hybrid smoothed classifier over Gaussian smoothed classifier. This shows that the proposed use of Normal-Uniform noise distribution as the training noise plays a key role in order to create highly robust hybrid smoothed classifier. Another notable finding is that performance of the models trained with Gaussian noise augmentation and certified under Uniform smoothing are far better than the performance of the models trained with Uniform noise augmentation and certified under Gaussian Smoothing.

Table 4. Performance of the baselines under proposed certification method when tested on a sample of 500 test images of CIFAR10.

σ	Training	Certification	Clean Acc	ℓ_1 ACR	ℓ_2 ACR
0.25	Gaussian	Gaussian	76.40	0.430	0.430
		Uniform	76.60	0.276	0.005
		Ours	**76.60**	**0.438**	**0.430**
	Gaussian+Consistency	Gaussian	73.40	0.535	0.535
		Uniform	73.20	0.291	0.005
		Ours	**73.60**	**0.538**	**0.535**
	Uniform	Gaussian	44.60	0.180	0.180
		Uniform	86.40	0.331	0.006
		Ours	58.80	0.279	0.178
0.50	Gaussian	Gaussian	64.80	0.523	0.523
		Uniform	65.20	0.406	0.007
		Ours	**65.40**	**0.543**	**0.523**
	Gaussian+Consistency	Gaussian	64.60	0.702	0.702
		Uniform	64.80	0.450	0.008
		Ours	**65.00**	**0.720**	**0.702**
	Uniform	Gaussian	14.80	0.062	0.062
		Uniform	79.40	0.568	0.01
		Ours	36.60	0.271	0.062

References

[1] 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, 4

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

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