# 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 $f : \mathbb{R}^d \to \{1, 2, 3, ..., K\}$ 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:

$$g(x) = \arg\max_c P[f(x + \varepsilon) = c]$$

(1)

Here, $\varepsilon$ is generated from a smoothing measure $\mu$. Considering $\mu$ to be isotropic Gaussian, Cohen et al. [1] show that $g(x)$ inherits certified robustness in $\ell_2$ 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 $p_A, p_B \in [0, 1]$ satisfy: $P(f(x + \varepsilon) = c_A) \geq p_A \geq p_B \geq \max_{c \neq c_A} P(f(x + \varepsilon) = c)$. Then $g(x + \delta) = c_A$ for all $||\delta||_2 < R$, where $R = \frac{\sigma}{2} (\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$, $\Phi^{-1}$ being the inverse of standard Gaussian CDF.

Yang et al. [3] show using the following theorem that by considering $\mu$ as a uniform distribution, $g(x)$ has provable robustness guarantee against $\ell_1$ norm constrained attacks.

**Theorem 2.** (Restating theorem 1.8 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 $\eta$ with $||\eta||_1 < 2\lambda(\rho - 0.5)$.

*Equal Contribution

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,

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

(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, $\delta$ is Gaussian noise and $\hat{F}(x) = \frac{1}{m} \sum_i F(x + \delta_i)$. $\lambda$ and $\eta$ 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 $\eta$ 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,

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

(3)

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

$$\mu_i = E[X - E[X]]^i$$

(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
$E[X - E[X]]^4$ or $\mu_4$ denotes a fat tailed distribution. The very similar argument follows to conclude that a random variable with low value of $\mu_4$ has a thin tailed probability distribution. In general a standardized metric is used hence, $\frac{\mu_4}{\sigma^4}$. The above metric $K(X)$ is used to measure the tailedness of a distribution relative to Gaussian distribution. For a normal distribution, $\frac{\mu_4}{\sigma^4} = 3$, 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 $X \sim N(0, \sigma^2)$ and $Y \sim U(-\lambda, \lambda)$ independently, then $Z = X + Y$ will follow a Normal-Uniform distribution with parameters $(\sigma_N, \lambda)$. The $\sigma_N$ 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 $\sigma_U$ instead of $\lambda$ in all our experiments. Hence the Normal-Uniform distribution is defined with the parameters $(\sigma_N, \sigma_U)$. 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 $X \sim N(0, \sigma^2)$ and $Y \sim U(-\lambda, \lambda)$ with $X$ and $Y$ being independent, then $Z = X + Y$ has the pdf as $f_Z(z) = \frac{1}{\sigma \sqrt{2\pi}} \left[ \Phi \left( \frac{z - \lambda}{\sigma} \right) - \Phi \left( \frac{-z - \lambda}{\sigma} \right) \right]$ with $z \in (-\infty, \infty)$, where $\Phi(.)$ is the CDF of standard normal distribution.

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

$= \int_{t+y \leq t} f_x(x) f_y(y) \, dx \, dy$

$= \int_{-\infty}^{\infty} f_Y(y) \left[ \int_{t-y}^{t} f_X(x) \, dx \right] dy$

$= \int_{-\infty}^{\infty} f_Y(y) \Phi \left( \frac{t-y}{\sigma} \right) \, dy$

So, the pdf is 

$f_Z(z) = \frac{1}{\sigma} \int_{-\infty}^{\infty} f_Y(y) \phi \left( \frac{z-y}{\sigma} \right) \, dy$

$= \frac{1}{\sigma} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} \phi \left( \frac{z-y}{\sigma} \right) \, dy$

$= \frac{1}{\sigma} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi} \sigma} \phi \left( \frac{z-\lambda}{\sigma} \right) \, dy$

$= \frac{1}{\sigma} \Phi \left( \frac{z-\lambda}{\sigma} \right) - \frac{1}{\sigma} \Phi \left( \frac{-z-\lambda}{\sigma} \right)$

\[\square\]

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 $F_Z(t) = \frac{1}{\sqrt{2\pi} \sigma} \left[ \Phi \left( \frac{t+\lambda}{\sigma} \right) + \phi \left( \frac{t+\lambda}{\sigma} \right) - \Phi \left( \frac{-t-\lambda}{\sigma} \right) - \phi \left( \frac{-t-\lambda}{\sigma} \right) \right]$ with $t \in (-\infty, \infty)$, where $\Phi(.)$ and $\phi(.)$ are the CDF and pdf of standard normal distribution respectively.

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

Let $I = \int_{-\infty}^{t} \Phi \left( \frac{z-\lambda}{\sigma} \right) \, dx = \sigma \int_{-\infty}^{\frac{t+\lambda}{\sigma}} \Phi(z) \, dz$

$= \sigma \left[ \Phi \left( \frac{t+\lambda}{\sigma} \right) - \Phi \left( \frac{-t-\lambda}{\sigma} \right) \right]$

Similarly calculating for the second term and replacing in the original equation, we get 

$F_Z(t) = \frac{1}{\sqrt{2\pi} \sigma} \left[ \Phi \left( \frac{t+\lambda}{\sigma} \right) + \phi \left( \frac{t+\lambda}{\sigma} \right) - \Phi \left( \frac{-t-\lambda}{\sigma} \right) - \phi \left( \frac{-t-\lambda}{\sigma} \right) \right]$

\[\square\]

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, $\mu_4 = E(Z^4 - E(Z))^4 = E(Z^4) = E[X^4 + 4X^3Y + 6X^2Y^2 + 4XY^3 + Y^4]$

$= E(X^4) + 6E(X^2)E(Y^2) + E(Y^4) = 3\sigma^4 + 6\sigma^2 \lambda^2 + \frac{\lambda^4}{3}$

We have $E(Y^4) = \int_{-\lambda}^{\lambda} y^4 \, dy = \frac{2\lambda^5}{15}$

So, 

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

\[\square\]
Table 1. Effect of $\beta$ 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($\sigma = 0.65$).

<table>
<thead>
<tr>
<th>$\beta$</th>
<th>Clean ACC</th>
<th>$\ell_1$ ACR</th>
<th>$\ell_2$ ACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>62.00</td>
<td>0.601</td>
<td>0.570</td>
</tr>
<tr>
<td>2</td>
<td>60.40</td>
<td>0.743</td>
<td>0.714</td>
</tr>
<tr>
<td>3</td>
<td>59.00</td>
<td>0.776</td>
<td>0.734</td>
</tr>
<tr>
<td>4</td>
<td>58.20</td>
<td><strong>0.779</strong></td>
<td>0.749</td>
</tr>
<tr>
<td>6</td>
<td>55.20</td>
<td>0.773</td>
<td><strong>0.751</strong></td>
</tr>
<tr>
<td>8</td>
<td>52.60</td>
<td>0.771</td>
<td>0.749</td>
</tr>
<tr>
<td>10</td>
<td>55.20</td>
<td>0.777</td>
<td>0.741</td>
</tr>
<tr>
<td>12</td>
<td>55.40</td>
<td>0.766</td>
<td>0.750</td>
</tr>
<tr>
<td>14</td>
<td>54.60</td>
<td>0.770</td>
<td>0.749</td>
</tr>
<tr>
<td>16</td>
<td>52.00</td>
<td>0.758</td>
<td>0.737</td>
</tr>
<tr>
<td>18</td>
<td>50.80</td>
<td>0.769</td>
<td>0.753</td>
</tr>
<tr>
<td>20</td>
<td>51.60</td>
<td>0.765</td>
<td>0.751</td>
</tr>
<tr>
<td>24</td>
<td>9.400</td>
<td>0.215</td>
<td>0.215</td>
</tr>
</tbody>
</table>

Figure 2. Similar to figure 1, by controlling the $\sigma_N$ and $\sigma_U$, 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($\sigma_N = 1.0, \sigma_U = 2.00.0$)(Green)

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

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

3. Ablation study

3.1. Effect of tuning parameter $\beta$

We investigate the effect of the regularizer tuning parameter $\beta$. As usual, when we increase $\beta$, initially the ACRs increase and clean accuracy decreases. A prominent robustness-accuracy trade-off is visible in table 1. However, as we further increase $\beta$, the clean accuracy oscillates between 55% and 52%, while the $\ell_1$ ACR gets stagnant around 0.770 and $\ell_2$ 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.

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

$$R_U = KL(F(x + NU)||F(x + U))$$ (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 $\ell_1$ ACR. The $\ell_2$ ACR drops though as can be seen in 3rd and 4th rows of table 2. On the other hand, regularizer $R_N$ gives a better $\ell_2$ 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 $\ell_2$ ACR than that of $\ell_1$ 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.

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.

<table>
<thead>
<tr>
<th>TRAINING</th>
<th>CERTIFICATION</th>
<th>CLEAN ACC</th>
<th>$\ell_1$ ACR</th>
<th>$\ell_2$ ACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{NU}(\sigma_N = 0.50, \sigma_U = 0.433) + R_N(\beta = 6)$</td>
<td>Gaussian($\sigma = 0.60$) + $\text{Unif}(\sigma = 0.650)$</td>
<td>60.00</td>
<td>0.770</td>
<td>0.746</td>
</tr>
<tr>
<td>$\text{NU}(\sigma_N = 0.50, \sigma_U = 0.433) + R_U(\beta = 6)$</td>
<td>$\text{Unif}(\sigma = 1.00)$ + $\text{Unif}(\sigma = 1.160)$</td>
<td>58.40</td>
<td>0.778</td>
<td>0.710</td>
</tr>
<tr>
<td>$\text{NU}(\sigma_N = 0.50, \sigma_U = 0.433) + R_S(\beta = 3)$</td>
<td>$\text{Unif}(\sigma = 1.00)$</td>
<td>59.00</td>
<td>0.776</td>
<td>0.734</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>TRAINING</th>
<th>$K(X)$</th>
<th>CLEAN ACC</th>
<th>$\ell_1$ ACR</th>
<th>$\ell_2$ ACR</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{NU}(\sigma_N = 1.00, \sigma_U = 0.866)$</td>
<td>0.22</td>
<td>45.00</td>
<td>0.671</td>
<td>0.625</td>
</tr>
<tr>
<td>$\text{NU}(\sigma_N = 1.00, \sigma_U = 1.155)$</td>
<td>-0.39</td>
<td>38.40</td>
<td>0.682</td>
<td>0.606</td>
</tr>
</tbody>
</table>

![Figure 3. Shape of Normal-Uniform distribution for different kurtosis values.](image)

The choice of $\sigma_N$ and $\sigma_U$ may seem arbitrary, but a specific relation selects them. For first row, $\sigma_U = \frac{1}{\sqrt{3}} \sigma_N$, 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 $\sigma$ 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 $\ell_1$ ACR under Gaussian noise training when the model is certified with our proposed hybrid 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.
References

