IMPROVED DETERMINISTIC l_2 ROBUSTNESS ON CIFAR-10 AND CIFAR-100

Anonymous authors
Paper under double-blind review

ABSTRACT

Training convolutional neural networks (CNNs) with a strict Lipschitz constraint under the l_2 norm is useful for provable adversarial robustness, interpretable gradients and stable training. While 1-Lipschitz CNNs can be designed by enforcing a 1-Lipschitz constraint on each layer, training such networks requires each layer to have an orthogonal Jacobian matrix (for all inputs) to prevent the gradients from vanishing during backpropagation. A layer with this property is said to be Gradient Norm Preserving (GNP). In this work, we introduce a procedure to certify the robustness of 1-Lipschitz CNNs by relaxing the orthogonalization of the last linear layer of the network that significantly advances the state of the art for both standard and provable robust accuracies on CIFAR-100 (gains of 4.80% and 4.71%, respectively). We further boost their robustness by introducing (i) a novel Gradient Norm preserving activation function called the Householder activation function (that includes every $\mathrm{GroupSort}$ activation) and (ii) a certificate regularization. On CIFAR-10, we achieve significant improvements over prior works in provable robust accuracy (5.81%) with only a minor drop in standard accuracy (-0.29%).

1 INTRODUCTION

Given a neural network $f : \mathbb{R}^d \to \mathbb{R}^k$, the Lipschitz constant[1] $\mathrm{Lip}(f)$ enforces an upper bound on how much the output is allowed to change in proportion to a change in the input. Previous work has demonstrated that a small Lipschitz constant is useful for improved adversarial robustness (Szegedy et al., 2014; Cissé et al., 2017), generalization bounds (Bartlett et al., 2017; Long & Sedghi, 2020), interpretable gradients (Tsipras et al., 2018) and Wasserstein distance estimation (Villani, 2008). $\mathrm{Lip}(f)$ also upper bounds the increase in the norm of gradient during backpropagation and can thus prevent gradient explosion during training, enabling us to train very deep networks (Xiao et al., 2018). While heuristic methods to enforce Lipschitz constraints (Miyato et al., 2018; Gulrajani et al., 2017) have achieved much practical success, they do not provably enforce a bound on $\mathrm{Lip}(f)$ globally and it remains challenging to achieve similar results when $\mathrm{Lip}(f)$ is provably bounded.

Using the property: $\mathrm{Lip}(g \circ h) \leq \mathrm{Lip}(g)\,\mathrm{Lip}(h)$, the Lipschitz constant of the neural network can be bounded by the product of the Lipschitz constant of all layers. While this allows us to construct 1-Lipschitz neural networks by constraining each layer to be 1-Lipschitz, Anil et al. (2018) identified a key difficulty with this approach. Because a 1-Lipschitz layer can only reduce the norm of gradient during backpropagation, backprop through each layer reduces the gradient norm, resulting in small gradient values for layers closer to the input, making training slow and difficult. To address this problem, they introduce Gradient Norm Preserving (GNP) architectures where each layer preserves the gradient norm during backpropagation. This involves constraining the Jacobian of each linear layer to be an orthogonal matrix and using a GNP activation function called $\mathrm{GroupSort}$. $\mathrm{GroupSort}$ activation function (Anil et al., 2018) first separates the vector of preactivations $\mathbf{z} \in \mathbb{R}^m$ into groups of pre-specified sizes, sorts each group in the descending order and then concatenates these sorted groups. When the group size is 2, the resulting activation function is called MaxMin.

In this work, we characterize the MaxMin activation function as a special case of the more general *Householder* (HH) activations. Recall that given $\mathbf{z} \in \mathbb{R}^m$, the HH transformation is a linear function reflecting $\mathbf{z}$ about the hyperplane $\mathbf{v}^T\mathbf{x} = 0$ ($\|\mathbf{v}\|_2 = 1$), given by $(\mathbf{I} - 2\mathbf{v}\mathbf{v}^T)\mathbf{z}$ where $\mathbf{I} - 2\mathbf{v}\mathbf{v}^T$ is

[1] Unless specified, we assume the Lipschitz constant under the l_2 norm in this work.

orthogonal because $\|\mathbf{v}\|_2 = 1$. The *Householder* activation function $\sigma_{\mathbf{v}}$ is defined below:

$$\sigma_{\mathbf{v}}(\mathbf{z}) = \begin{cases} \mathbf{z}, & \mathbf{v}^T\mathbf{z} > 0, \\ (\mathbf{I} - 2\mathbf{v}\mathbf{v}^T)\mathbf{z}, & \mathbf{v}^T\mathbf{z} \leq 0. \end{cases} \tag{1}$$

First, note that since $\mathbf{z} = (\mathbf{I} - 2\mathbf{v}\mathbf{v}^T)\mathbf{z}$ along $\mathbf{v}^T\mathbf{z} = 0$, $\sigma_{\mathbf{v}}$ is continuous. Moreover, the Jacobian $\nabla_{\mathbf{z}}\, \sigma_{\mathbf{v}}$ is either $\mathbf{I}$ or $\mathbf{I} - 2\mathbf{v}\mathbf{v}^T$ (both orthogonal) implying $\sigma_{\mathbf{v}}$ is GNP. Since these properties hold $\forall\, \mathbf{v} : \|\mathbf{v}\|_2 = 1$, $\mathbf{v}$ can be learned during the training. In fact, we prove that any GNP piecewise linear function that changes from $\mathbf{Q}_1\mathbf{z}$ to $\mathbf{Q}_2\mathbf{z}$ ($\mathbf{Q}_1, \mathbf{Q}_2$ are square orthogonal matrices) along $\mathbf{v}^T\mathbf{z} = 0$ must satisfy $\mathbf{Q}_2 = \mathbf{Q}_1\left(\mathbf{I} - 2\mathbf{v}\mathbf{v}^T\right)$ to be continuous (Theorem 3). Thus, this characterization proves that every $\mathrm{GroupSort}$ activation is a special case of the more general Householder activation function (example in Figure 1, discussion in Section 6).

For 1-Lipschitz CNNs, the robustness certificate for a sample $\mathbf{x}$ from class l is computed as $\mathcal{M}_f(\mathbf{x})/\sqrt{2}$ where $\mathcal{M}_f(\mathbf{x}) = f_l(\mathbf{x}) - \max_{i \neq l} f_i(\mathbf{x})$. Naturally, larger values of $f_l(\mathbf{x})$ and smaller values of $\max_{j \neq l} f_j(\mathbf{x})$ will lead to larger certificates. This requires the last weight matrix in the network, denoted by $\mathbf{W} \in \mathbb{R}^{k \times m}$ (k is the number of classes, m is the dimension of the penultimate layer, $m > k$), to enforce the following constraints throughout training:

$$\forall j, \;\; \|\mathbf{W}_{j,:}\|_2 = 1, \qquad i \neq j, \;\; \mathbf{W}_{j,:} \perp \mathbf{W}_{i,:}$$

where $\mathbf{W}_{i,:}$ denotes the ith row of $\mathbf{W}$. Now suppose that for some input image with label l, we want to update $\mathbf{W}$ to increase the logit for the lth class. Since $\|\mathbf{W}_{l,:}\|_2$ is constrained to be 1, the logit can only be increased by changing the direction of the vector $\mathbf{W}_{l,:}$. Because the other rows $\{\mathbf{W}_{i,:},\ i \neq l\}$ are constrained to be orthogonal to $\mathbf{W}_{l,:}$, this further requires an update for all the rows of $\mathbf{W}$. Thus during training, any update made to learn some class must necessitate the forgetting of information relevant for the other classes. This can be particularly problematic when the number of classes k and thus the number of orthogonality constraints per row (i.e., $k - 1$) is large (such as in CIFAR-100).

To address this limitation, we propose to keep the last weight layer of the network *unchanged.* But then the resulting function is no longer 1-Lipschitz and the certificate $\mathcal{M}_f(\mathbf{x})/\sqrt{2}$ is not valid. Thus, we introduce a new certification procedure that does not require the last weight layer of the network $\mathbf{W}$ to be orthogonal. Our certificate is then given by the following equation:

$$\min_{i \neq l} \frac{f_l(\mathbf{x}) - f_i(\mathbf{x})}{\|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2}$$

However, a limitation of using the above certificate is that because the weight layers are completely unconstrained, larger norms of rows (i.e., $\|\mathbf{W}_{i,:}\|$) can result in larger values of $\|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2$ and thus smaller certificate values. To address this limitation, we normalize all rows to be of unit norm before computing the logits. While this still requires all the rows to be of unit norm, their directions are now allowed to change freely thus preventing the need to update other rows and forgetting of learned information. We show that this provides significant improvements when the number of classes is large. We call this procedure *Last Layer Normalization* (abbreviated as *LLN*). On CIFAR-100, this significantly improves both the standard ($> 3\%$) and provable robust accuracy ($> 4\%$ at $\rho = 36/255$) across multiple 1-Lipschitz CNN architectures (Table 1).

Another limitation of existing 1-Lipschitz CNNs (Li et al., 2019b; Trockman & Kolter, 2021; Singla & Feizi, 2021) is that their robustness guarantees do not scale properly with the l_2 radius ρ. For example, the provable robust accuracy of (Singla & Feizi, 2021) drops $\sim 30\%$ at $\rho = 108/255$ compared to $36/255$ on CIFAR-10 (Table 2). To address this limitation, we introduce a certificate regularization denoted by CR (Section 5) that when used along with the Householder activation results in significantly improved provable robust accuracy at larger radius values with minimal loss in standard accuracy. On the CIFAR-10 dataset, we achieve significant improvements in the provable robust accuracy for large $\rho = 108/255$ (min gain of $+4.96\%$) across different architectures with minimal loss in the standard accuracy (max drop of -0.56%). Results are in Table 2.

In summary, in this paper, we make the following contributions:

- We introduce a certification procedure *without* orthogonalizing the last linear layer called *Last Layer Normalization*. This procedure significantly enhances the standard and provable robust accuracy when the number of classes is large. Using the LipConvnet-15 network on CIFAR-100, our modification achieves a gain of $+4.71\%$ in provable robust accuracy (at $\rho = 36/255$) with a gain of $+4.80\%$ in standard accuracy (Table 1).

- We introduce a *Certificate Regularizer* that significantly advances the provable robust accuracy with a small reduction in standard accuracy. Using LipConvnet-15 network on CIFAR-10, we achieve $+5.81\%$ improvement in provable robust accuracy (at $\rho = 108/255$) with only a -0.29% drop in standard accuracy over the existing methods (Table 2).
- We introduce a class of piecewise linear GNP activation functions called *Householder* or HH activations. We show that the MaxMin activation is a special case of the HH activation for certain settings. We prove that Householder transformations are *necessary* for any GNP piecewise linear function to be continuous (Theorem 3).

2 Related work

Provably Lipschitz convolutional neural networks: The class of fully connected neural networks (FCNs) which are Gradient Norm Preserving (GNP) and 1-Lipschitz were first introduced by Anil et al. (2018). They orthogonalize weight matrices and use $\mathrm{GroupSort}$ as the activation function to design each layer to be GNP. While there have been numerous works on enforcing Lipschitz constraints on convolution layers (Cissé et al., 2017; Tsuzuku et al., 2018; Qian & Wegman, 2019; Gouk et al., 2020; Sedghi et al., 2019), they either enforce loose Lipschitz bounds or are not scalable to large networks. To ensure that the Lipschitz constraint on convolutional layers is tight, multiple recent works try to construct convolution layers with an orthogonal Jacobian matrix (Li et al., 2019b; Trockman & Kolter, 2021; Singla & Feizi, 2021). These approaches avoid the aforementioned issues and allow the training of large, provably 1-Lipschitz CNNs while achieving impressive results. While 1-Lipschitz CNNs have achieved much practical success, they still use the original MaxMin activation and there is little work so far in designing novel activation functions that are GNP.

Provable defenses against adversarial examples: A provably robust classifier is one for which we can guarantee that the classifier's prediction remains constant within some region around the input. Most of the existing methods for provable robustness either bound the Lipschitz constant of the neural network or the individual layers (Weng et al., 2018; Zhang et al., 2019; 2018; Wong et al., 2018; Wong & Kolter, 2018; Raghunathan et al., 2018; Croce et al., 2019; Singh et al., 2018; Singla & Feizi, 2020; Zhang et al., 2021). However, these methods do not scale to large and practical networks on the ImageNet dataset (Deng et al., 2009). To scale to such large networks, randomized smoothing (Liu et al., 2018; Cao & Gong, 2017; Lécuyer et al., 2018; Li et al., 2019a; Cohen et al., 2019; Salman et al., 2019; Levine et al., 2019; Kumar et al., 2020a;b) has been proposed as a *probabilistically certified defense*. However, certifying robustness with high probability requires generating a large number of noisy samples leading to high inference-time computation. In contrast, the defense we propose is deterministic and hence not comparable to randomized smoothing. While Levine & Feizi (2021) provide deterministic robustness certificates using randomized smoothing, their certificates are in the l_1 norm and not directly applicable for the l_2 threat model studied in this work.

3 Problem setup and Notation

For a vector $\mathbf{v}$, $\mathbf{v}_j$ denotes its jth element. For a matrix $\mathbf{A}$, $\mathbf{A}_{j,:}$ and $\mathbf{A}_{:,k}$ denote the jth row and k^{th} column respectively. Both $\mathbf{A}_{j,:}$ and $\mathbf{A}_{:,k}$ are assumed to be column vectors (thus $\mathbf{A}_{j,:}$ is the transpose of jth row of $\mathbf{A}$). $\mathbf{A}_{j,k}$ denotes the element in j^{th} row and k^{th} column of $\mathbf{A}$. $\mathbf{A}_{:j,:k}$ denotes the matrix containing the first j rows and k columns of $\mathbf{A}$. The same rules are directly extended to higher order tensors. $\mathbf{I}$ denotes the identity matrix, $\mathbb{R}$ to denote the field of real numbers. For $\theta \in \mathbb{R}$, $\mathbf{J}^{+}(\theta)$ and $\mathbf{J}^{-}(\theta)$ denote the orthogonal matrices with determinants $+1$ and -1 defined as follows:

$$\mathbf{J}^{+}(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \qquad \mathbf{J}^{-}(\theta) = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \tag{2}$$

We construct a 1-Lipschitz neural network, $f : \mathbb{R}^d \to \mathbb{R}^k$ (d is the input dimension, k is the number of classes) by composing 1-Lipschitz convolution layers and GNP activation functions. To certify robustness for some input $\mathbf{x}$ with prediction l, we first define the margin of prediction: $\mathcal{M}_f(\mathbf{x}) = \max\left(0, f_l(\mathbf{x}) - \max_{i \neq l} f_i(\mathbf{x})\right)$ where $f_i(\mathbf{x})$ is the logit for class i and l is the correct label. Using Theorem 7 in Li et al. (2019b), we can derive the robustness certificate (in the l_2 norm)

as $\mathcal{M}_f(\mathbf{x})/\sqrt{2}$. Thus, the l_2 distance of $\mathbf{x}$ to the decision boundary is lower bounded by $\mathcal{M}_f(\mathbf{x})/\sqrt{2}$:

$$\min_{i \neq l} \min_{f_i(\mathbf{x}^*) = f_l(\mathbf{x}^*)} \|\mathbf{x}^* - \mathbf{x}\|_2 \geq \frac{\mathcal{M}_f(\mathbf{x})}{\sqrt{2}} \tag{3}$$

We often use the abbreviation $f_i - f_j : \mathbb{R}^D \to \mathbb{R}$ to denote the function so that:

$$(f_i - f_j)(\mathbf{x}) = f_i(\mathbf{x}) - f_j(\mathbf{x}), \qquad \forall\, \mathbf{x} \in \mathbb{R}^D$$

Our goal is to train the neural network f to achieve the maximum possible provably robust accuracy while also simultaneously improving (or maintaining) standard accuracy.

4 Last Layer Normalization

Existing 1-Lipschitz neural networks require the weight matrices of all the linear layers of the network to be orthogonal to ensure that the network is 1-Lipschitz and the certificate in equation (3) is valid. For the weight matrix in the last layer of the network (that maps the penultimate layer neurons to the logits), $\mathbf{W} \in \mathbb{R}^{k \times m}$ (k is the number of classes, m is the dimension of the penultimate layer, $m > k$), this enforces the following constraints on each row $\mathbf{W}_{i,:} \in \mathbb{R}^m$:

$$\forall j, \;\; \|\mathbf{W}_{j,:}\|_2 = 1, \qquad \forall i \neq j, \;\; \mathbf{W}_{j,:} \perp \mathbf{W}_{i,:} \tag{4}$$

Now suppose that for some input $\mathbf{x}$ with label l, we want to update $\mathbf{W}$ to increase the logit for the lth class. Since $\|\mathbf{W}_{l,:}\|_2$ is constrained to be 1, the gradient update can only change the direction of the vector $\mathbf{W}_{l,:}$. But now, because the other rows $\{\mathbf{W}_{i,:},\ i \neq l\}$ are constrained to be orthogonal to $\mathbf{W}_{l,:}$, this further requires an update for all the rows of $\mathbf{W}$. This has the negative effect that during training, any update made to learn some class may necessitate the forgetting of information relevant for the other classes. This can be particularly problematic when the number of classes k is large (such as in CIFAR-100) and thus the number of orthogonality constraints per row i.e. $k - 1$ is large.

To address this limitation, first observe that the neural network from the input layer to the penultimate layer (i.e excluding the last linear layer) is 1-Lipschitz. Let $g : \mathbb{R}^d \to \mathbb{R}^m$ be this function so that $f(\mathbf{x}) = \mathbf{W}g(\mathbf{x}) + \mathbf{b}$. This equation suggests that even if $\mathbf{W}$ is not orthogonal, the Lipschitz constant of the function $f_l - f_i$, can be computed by multiplying the Lipschitz constant of g (which is 1) and that of $(\mathbf{W}_{l,:} - \mathbf{W}_{i,:})$ (which is $\|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2$). The robustness certificate can then be computed as $\mathcal{M}_f(\mathbf{x})/\|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2$. This procedure leads to the following theorem:

Theorem 1. *Given 1-Lipschitz continuous function $g : \mathbb{R}^d \to \mathbb{R}^m$ and $\mathbf{W} \in \mathbb{R}^{k \times m},\ \mathbf{b} \in \mathbb{R}^k$ (k is the number of classes), construct a new function $f : \mathbb{R}^d \to \mathbb{R}^k$ defined as: $f(\mathbf{x}) = \mathbf{W}g(\mathbf{x}) + \mathbf{b}$. Let $f_l(\mathbf{x}) > \max_{i \neq l} f_i(\mathbf{x})$. The robustness certificate (under the l_2 norm) is given by:*

$$\min_{i \neq l} \min_{f_l(\mathbf{x}^*) = f_i(\mathbf{x}^*)} \|\mathbf{x}^* - \mathbf{x}\|_2 \geq \min_{i \neq l} \frac{f_l(\mathbf{x}) - f_i(\mathbf{x})}{\|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2}$$

Proof. We proceed by computing the Lipschitz constant of the function $f_l - f_i$.
The gradient of the function: $f_l - f_i$ at $\mathbf{x}$ can be computed using the chain rule:

$$\nabla_{\mathbf{x}} (f_l - f_i) = (\mathbf{W}_{l,:} - \mathbf{W}_{i,:})^T \nabla_{\mathbf{x}} g$$

Since g is given to be 1-Lipschitz, the Lipschitz constant of $f_l - f_i$ can be computed using the above equation as follows:

$$\|\nabla_{\mathbf{x}} (f_l - f_i)\|_2 \leq \|(\mathbf{W}_{l,:} - \mathbf{W}_{i,:})^T (\nabla_{\mathbf{x}} g)\|_2$$
$$\|\nabla_{\mathbf{x}} (f_l - f_i)\|_2 \leq \|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2 \|\nabla_{\mathbf{x}} g\|_2 \leq \|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2$$

Thus, the distance of $\mathbf{x}$ to the decision boundary $f_l - f_i = 0$, is lower bounded by:

$$\min_{f_l(\mathbf{x}^*) = f_i(\mathbf{x}^*)} \|\mathbf{x}^* - \mathbf{x}\|_2 \geq \frac{f_l(\mathbf{x}) - f_i(\mathbf{x})}{\|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2}$$

Thus, the distance to decision boundary across all classes $i \neq l$ is lower bounded by:

$$\min_{i \neq l} \min_{f_l(\mathbf{x}^*) = f_i(\mathbf{x}^*)} \|\mathbf{x}^* - \mathbf{x}\|_2 \geq \min_{i \neq l} \frac{f_l(\mathbf{x}) - f_i(\mathbf{x})}{\|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2}$$

□

However, in our experiments, we found that using this procedure directly i.e without any constraint on the weight matrix $\mathbf{W}$ often results in large norms of row vectors $\|\mathbf{W}_{i,:}\|_2$, and thus large values of $\|\mathbf{W}_{l,:} - \mathbf{W}_{i,:}\|_2$ and smaller certificates (Theorem 1). To address this problem, we normalize all rows of the matrix to be of unit norm before computing the logits so that for the input $\mathbf{x}$, the logit $g_i(\mathbf{x})$ can be computed as follows:

$$g_i(\mathbf{x}) = \frac{(\mathbf{W}_{i,:})^T f(\mathbf{x})}{\|\mathbf{W}_{i,:}\|_2} + \mathbf{b}_i$$

The robustness certificate can then be computed as follows:

$$\min_{i \neq l} \frac{g_l(\mathbf{x}) - g_i(\mathbf{x})}{\|\mathbf{W}^{(n)}_{l,:} - \mathbf{W}^{(n)}_{i,:}\|_2}, \qquad \text{where } \forall j, \;\; \mathbf{W}^{(n)}_{j,:} = \frac{\mathbf{W}_{j,:}}{\|\mathbf{W}_{j,:}\|_2}, \; g_j(\mathbf{x}) = \left(\mathbf{W}^{(n)}_{j,:}\right)^T f(\mathbf{x}) + \mathbf{b}_j$$

While each row $\mathbf{W}_{i,:}$ is still constrained to be of unit norm, unlike with orthogonality constraints (equation (4)), their directions are now allowed to change freely. This provides significant improvements when the number of classes is large. We call this procedure *Last Layer Normalization* (abbreviated as *LLN*).

5 Certificate Regularization

A limitation of using cross entropy loss for training 1-Lipschitz CNNs is that it is not explicitly designed to maximize the margin $\mathcal{M}_f(\mathbf{x})$ and thus, the robustness certificate. That is, once the cross entropy loss becomes small, the gradients will no longer try to further increase the margin $\mathcal{M}_f(\mathbf{x})$ even though the network may have the capacity to learn bigger margins.

To address this limitation, we can simply *subtract* the certificate i.e $-\mathcal{M}_f(\mathbf{x})/\sqrt{2}$ from the usual cross entropy loss function during training. Observe that we subtract the certificate because we want to *maximize* the certificate values while *minimizing* the cross entropy loss. However, in our experiments we found that this regularization term excessively penalizes the network for the misclassified examples and as a result, the certificate values for the correctly classified inputs are not large. Thus, we propose to use the following regularized loss function during training:

$$\min_{\Omega} \quad \mathbb{E}_{(\mathbf{x},l)\sim D}\left[\ell\left(f_\Omega(\mathbf{x}), l\right) - \gamma \,\text{relu}\left(\frac{\mathcal{M}_f(\mathbf{x})}{\sqrt{2}}\right)\right] \tag{5}$$

In the above equation, f_Ω denotes the 1-Lipschitz neural network parametrized by Ω, $f_\Omega(\mathbf{x})$ denotes the logits for the input $\mathbf{x}$, $\ell\left(f_\Omega(\mathbf{x}), l\right)$ is the cross entropy loss for input $\mathbf{x}$ with label l and $\gamma > 0$ is the regularization coefficient for maximizing the certificate. We have the minus sign in front of the regularization term $\gamma\,\text{relu}(\mathcal{M}_f(\mathbf{x})/\sqrt{2})$ because we want to *maximize* the certificate while *minimizing* the cross entropy loss. For wrongly classified inputs, $\mathcal{M}_f(\mathbf{x})/\sqrt{2} < 0 \implies \text{relu}(\mathcal{M}_f(\mathbf{x})/\sqrt{2}) = 0$. This ensures that the optimization tries to increase the certificates *only* for the correctly classified inputs. We call the above mentioned procedure *Certificate Regularization* (abbreviated as *CR*).

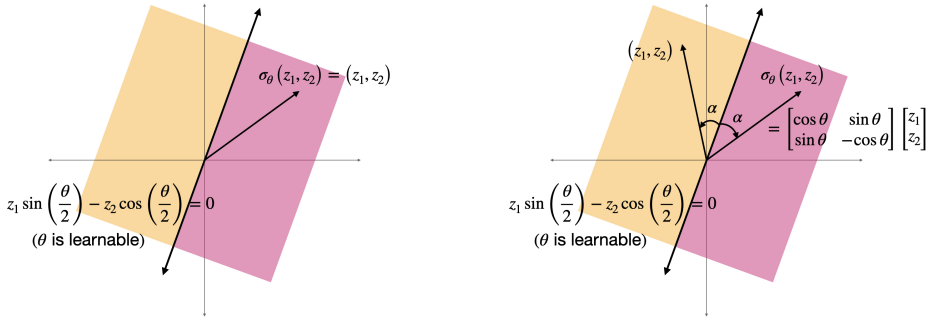
6 Householder Activation Functions

Recall that for the MaxMin activation function, $\mathrm{MaxMin}(z_1, z_2) = (z_1, z_2)$ if $z_1 > z_2$ and (z_2, z_1) otherwise. Thus, the Jacobian of MaxMin for $z_1 > z_2$ case is $\mathbf{I} = \mathbf{J}^+(0)$ while for $z_1 \leq z_2$ is $\mathbf{J}^-(\pi/2)$. We first show that MaxMin is a special case of the more general Householder activation functions where the Jacobian $\mathbf{J}^-(\pi/2)$ is replaced with $\mathbf{J}^-(\theta)$ and the conditions $z_1 > z_2$ are replaced with $z_1 \sin(\theta/2) > z_2 \cos(\theta/2)$ (similarly for $\leq$). The construction of Householder activations in 2 dimensions, denoted by σ_θ, is given in the following Theorem:

Theorem 2. *The function $\sigma_\theta : \mathbb{R}^2 \to \mathbb{R}^2$ defined as*

$$\sigma_\theta\left(z_1, z_2\right) = \begin{cases} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} & \text{if } \; z_1 \sin\left(\theta/2\right) - z_2 \cos\left(\theta/2\right) > 0 \\ \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} & \text{if } \; z_1 \sin\left(\theta/2\right) - z_2 \cos\left(\theta/2\right) \leq 0 \end{cases} \tag{6}$$

is continuous and is called 2D *Householder Activation of Order* 1.



(a) Case 1: $z_1 \sin(\theta/2) - z_2 \cos(\theta/2) > 0$ (b) Case 2: $z_1 \sin(\theta/2) - z_2 \cos(\theta/2) \leq 0$

Figure 1: Illustration of the Householder activation, σ_θ. In each colored region, σ_θ is linear. The Jacobian is $\mathbf{I}$ when (z_1, z_2) lies in the pink region (Case 1) and $\mathbf{I} - 2\mathbf{v}\mathbf{v}^T$ in the other region (Case 2) where $\mathbf{v} = [\sin(\theta/2) \quad -\cos(\theta/2)]^T$. Both of these matrices are orthogonal implying σ_θ is GNP.

The two cases are demonstrated in Figure 1a and Figure 1b, respectively. Since σ_θ is continuous, GNP and 1-Lipschitz $\forall\, \theta \in \mathbb{R}$, θ is a learnable parameter. For $\theta = \pi/2$ in equation (6), σ_θ is equivalent to MaxMin. Thus, σ_θ is at least as expressive as MaxMin.

Recall that given $\mathbf{z} \in \mathbb{R}^m$, the Householder (HH) transformation reflects $\mathbf{z}$ about the hyperplane $\mathbf{v}^T\mathbf{x} = 0$ where $\|\mathbf{v}\|_2 = 1$. The linear transformation is given by the equation $(\mathbf{I} - 2\mathbf{v}\mathbf{v}^T)\mathbf{z}$ where $\mathbf{I} - 2\mathbf{v}\mathbf{v}^T$ is orthogonal because $\|\mathbf{v}\|_2 = 1$. Now, consider the nonlinear function $\sigma_\mathbf{v}$ defined below:

Definition 1. ***(Householder Activation of Order*** 1***)*** *The activation function* $\sigma_\mathbf{v} : \mathbb{R}^m \to \mathbb{R}^m$*, applied on* $\mathbf{z} \in \mathbb{R}^m$*, is called the* m*-dimensional Householder Activation of Order* 1*:*

$$\sigma_\mathbf{v}(\mathbf{z}) = \begin{cases} \mathbf{z}, & \mathbf{v}^T\mathbf{z} > 0, \\ (\mathbf{I} - 2\mathbf{v}\mathbf{v}^T)\mathbf{z}, & \mathbf{v}^T\mathbf{z} \leq 0. \end{cases} \tag{7}$$

Since $\sigma_\mathbf{v}$ is linear when $\mathbf{v}^T\mathbf{z} > 0$ or $\mathbf{v}^T\mathbf{z} < 0$, it is also continuous in both cases. At the hyperplane separating the two cases (i.e., $\mathbf{v}^T\mathbf{z} = 0$) we have: $(\mathbf{I} - 2\mathbf{v}\mathbf{v}^T)\mathbf{z} = \mathbf{z} - 2(\mathbf{v}^T\mathbf{z})\mathbf{v} = \mathbf{z}$ (both linear functions are equal). Thus, $\sigma_\mathbf{v}$ is continuous $\forall\, \mathbf{z} \in \mathbb{R}^m$. Moreover, the Jacobian is either $\mathbf{I}$ or $\mathbf{I} - 2\mathbf{v}\mathbf{v}^T$ which are both square orthogonal matrices. Thus, $\sigma_\mathbf{v}$ is also GNP and 1-Lipschitz. Since these properties hold for all $\mathbf{v}$ satisfying $\|\mathbf{v}\|_2 = 1$, $\mathbf{v}$ can be made a learnable parameter.

While the above arguments suggest that HH transformations are *sufficient* to ensure such functions are continuous, we also prove that they are *necessary*. That is, we prove that if a GNP piecewise linear function $g : \mathbb{R}^m \to \mathbb{R}^m$ transitions between different linear functions $\mathbf{Q}_1\mathbf{z}$ and $\mathbf{Q}_2\mathbf{z}$ (in an open set $S \subset \mathbb{R}^m$) along a hyperplane $\mathbf{v}^T\mathbf{z} = 0$ (where $\|\mathbf{v}\|_2 = 1$), then g is continuous in S *if and only if* $\mathbf{Q}_2 = \mathbf{Q}_1(\mathbf{I} - 2\mathbf{v}\mathbf{v}^T)$. This theoretical result provides a general principle for designing piecewise linear GNP activation functions. The formal result is stated in the following Theorem:

Theorem 3. *Given an open set* $S \subset \mathbb{R}^m$*, orthogonal square matrices* $\mathbf{Q}_1 \neq \mathbf{Q}_2$*, and vector* $\mathbf{v} \in \mathbb{R}^m$ $(\|\mathbf{v}\|_2 = 1)$ *such that* $S \cap \{\mathbf{z} : \mathbf{v}^T\mathbf{z} = 0\} \neq \emptyset$*, the function* g *defined as follows:*

$$g(\mathbf{z}) = \begin{cases} \mathbf{Q}_1\mathbf{z}, & \mathbf{z} \in S, \mathbf{v}^T\mathbf{z} > 0, \\ \mathbf{Q}_2\mathbf{z}, & \mathbf{z} \in S, \mathbf{v}^T\mathbf{z} \leq 0 \end{cases} \tag{8}$$

is continuous in S *if and only if* $\mathbf{Q}_2 = \mathbf{Q}_1(\mathbf{I} - 2\mathbf{v}\mathbf{v}^T)$.

Proof of Theorem 3 is in Appendix A.1 and that of Theorem 2 is discussed in A.2 as a specific case of Theorem 3. Note that since the matrix $\mathbf{I} - 2\mathbf{v}\mathbf{v}^T$ has determinant -1, the above theorem necessitates that $\det(\mathbf{Q}_1) = -\det(\mathbf{Q}_2)$ i.e the determinant of the Jacobian must change sign whenever the Jacobian of a piecewise linear GNP activation function changes.

A major limitation of using $\sigma_\mathbf{v}$ (equation (7)) directly is that it has only 2 linear regions and is thus limited in its expressive power. In contrast, MaxMin first divides the preactivation $\mathbf{z} \in \mathbb{R}^m$ (assuming m is divisible by 2) into $m/2$ groups of size 2 each. Since each group has 2 linear regions, we get

$2^{m/2}$ linear regions from the $m/2$ groups. Thus, to increase the expressive power, we similarly divide $\mathbf{z}$ into $m/2$ groups of size 2 each and apply the 2-dimensional Householder activation function of Order 1 (σ_θ) to each group resulting in $2^{m/2}$ linear regions (same as MaxMin).

To apply σ_θ to the output of a convolution layer $\mathbf{z} \in \mathbb{R}^{m \times n \times n}$ (m is the number of channels and $n \times n$ is the spatial size), we first split $\mathbf{z}$ into 2 tensors along the channel dimension giving the tensors: $\mathbf{z}_{:m/2,:,:}$ and $\mathbf{z}_{m/2:,:,:}$. Each of these tensors is of size $m/2 \times n \times n$ giving $n^2 m/2$ groups. We use the same θ for each pair of channels (irrespective of spatial location) resulting in $m/2$ learnable parameters. We initialize each $\theta = \pi/2$ so that σ_θ is equivalent to MaxMin at initialization.

7 EXPERIMENTS

Our goal is to evaluate the effectiveness of the three changes proposed in this work: (a) Last Layer Normalization, (b) Certificate regularization and (c) Householder activation functions. We perform experiments under the setting of provably robust image classification on CIFAR-10 and CIFAR-100 datasets using the same 1-Lipschitz CNN architectures used by Singla & Feizi (2021) (LipConvnet-5, 10, 15, . . . , 40) due to their superior performance over the prior works. We compare with the three orthogonal convolution layers in the literature: SOC (Singla & Feizi, 2021), BCOP (Li et al., 2019b) and Cayley (Trockman & Kolter, 2021) using MaxMin as the activation function.

We use SOC with MaxMin as the primary baseline for comparison in the maintext due to their superior performance over prior works (BCOP, Cayley). Results using BCOP and Cayley convolutions are given in Appendix Sections E and F for completeness. We use the same implementations for these convolution layers as given in their respective github repositories. We compare the provable robust accuracy using 3 different values of the l_2 perturbation radius $\rho = 36/255,\ 72/255,\ 108/255$. In both Tables 2 and 1, for all networks, we use SOC as the convolution layer. The symbol HH (in Tables 2, 1) is for the 2D Householder Activation of order 1 or σ_θ (defined in equation (6)). We observe negligible inference time overhead due to any of our changes (Appendix Tables 3, 9).In fact, LLN leads to reduced training time because it relaxes the orthogonality constraint.

All experiments were performed using 1 NVIDIA GeForce RTX 2080 Ti GPU. All networks were trained for 200 epochs with initial learning rate of 0.1, dropped by a factor of 0.1 after 100 and 150 epochs. For Certificate Regularization (or CR), we set the parameter $\gamma = 0.5$.

7.1 RESULTS ON CIFAR-100

In Table 1, for each architecture, the row "SOC + MaxMin" uses the MaxMin activation, "+ LLN" adds Last Layer Normalization (uses MaxMin), "+ HH" replaces MaxMin with σ_θ (also uses LLN) , "+ CR" also adds Certificate Regularization with $\gamma = 0.1$ (uses both σ_θ and LLN). The column, "Increase (Standard)" denotes the increase in standard accuracy relative to "SOC + MaxMin".

By adding LLN (the row "+ LLN"), we observe gains in standard (min gain of 1.10%) and provable robust accuracy (min gain of 1.71% at $\rho = 36/255$) across all the LipConvnet architectures (gains relative to "SOC + MaxMin"). These gains are smallest for the LipConvnet-40 network with the maximum depth. However, replacing MaxMin with the σ_θ activation further improves the standard (min gain of 3.65%) and provable robust accuracy (min gain of 4.46% at $\rho = 36/255$) across all networks (again relative to "SOC + MaxMin"). We observe that replacing MaxMin with σ_θ significantly improves the performance of the deeper LipConvnet-$35, 40$ networks.

Adding CR further improves the provable robust accuracy while only slightly reducing the standard accuracy. Because LLN significantly improves the standard accuracy, we compare the standard accuracy numbers between rows "+ CR" and "+ LLN" to evaluate the drop due to CR. We observe a small drop in standard accuracy ($-0.04\%, -0.11\%$) only for LipConvnet-5 and LipConvnet-15 networks. For the other networks, the standard accuracy actually *increases*.

7.2 RESULTS ON CIFAR-10

In Table 2, for each architecture, the row "SOC + MaxMin" uses the MaxMin activation, the row "+ HH" uses σ_θ activation (replacing MaxMin) and the row "+ CR" also adds Certificate Regularization with $\gamma = 0.1$ (again using σ_θ as the activation). Due to the small number of classes in CIFAR-10, we

Architecture	Methods	Standard Accuracy	Provable Robust Acc. ($\rho =$)			Increase
			36/255	72/255	108/255	(Standard)
LipConvnet-5	SOC + MaxMin	42.71%	27.86%	17.45%	9.99%	_
	+ LLN	45.86%	31.93%	21.17%	13.21%	+3.15%
	+ HH	**46.36%**	32.64%	21.19%	13.12%	**+3.65%**
	+ CR	45.82%	**32.99%**	**22.48%**	**14.79%**	+3.11%
LipConvnet-10	SOC + MaxMin	43.72%	29.39%	18.56%	11.16%	_
	+ LLN	46.88%	33.32%	22.08%	13.87%	+3.16%
	+ HH	**47.96%**	34.30%	22.35%	14.48%	**+4.24%**
	+ CR	47.07%	**34.53%**	**23.50%**	**15.66%**	+3.35%
LipConvnet-15	SOC + MaxMin	42.92%	28.81%	17.93%	10.73%	_
	+ LLN	47.72%	33.52%	21.89%	13.76%	+4.80%
	+ HH	**47.72%**	33.97%	22.45%	13.81%	**+4.80%**
	+ CR	47.61%	**34.54%**	**23.16%**	**15.09%**	+4.69%
LipConvnet-20	SOC + MaxMin	43.06%	29.34%	18.66%	11.20%	_
	+ LLN	46.86%	33.48%	22.14%	14.10%	+3.80%
	+ HH	47.71%	34.22%	22.93%	14.57%	+4.65%
	+ CR	**47.84%**	**34.77%**	**23.70%**	**15.84%**	**+4.78%**
LipConvnet-25	SOC + MaxMin	43.37%	28.59%	18.18%	10.85%	_
	+ LLN	46.32%	32.87%	21.53%	13.86%	+2.95%
	+ HH	**47.70%**	34.00%	22.67%	14.57%	**+4.33%**
	+ CR	46.87%	**34.09%**	**23.41%**	**15.61%**	+3.50%
LipConvnet-30	SOC + MaxMin	42.87%	28.74%	18.47%	11.21%	_
	+ LLN	46.18%	32.82%	21.52%	13.52%	+3.31%
	+ HH	46.80%	33.72%	22.70%	14.31%	+3.93%
	+ CR	**46.92%**	**34.17%**	**23.21%**	**15.84%**	**+4.05%**
LipConvnet-35	SOC + MaxMin	42.42%	28.34%	18.10%	10.96%	_
	+ LLN	45.22%	32.10%	21.28%	13.25%	+2.80%
	+ HH	46.21%	32.80%	21.55%	14.13%	+3.79%
	+ CR	**46.88%**	**33.64%**	**23.34%**	**15.73%**	**+4.46%**
LipConvnet-40	SOC + MaxMin	41.84%	28.00%	17.40%	10.28%	_
	+ LLN	42.94%	29.71%	19.30%	11.99%	+1.10%
	+ HH	**45.84%**	**32.79%**	21.98%	14.07%	**+4.00%**
	+ CR	45.03%	32.57%	**22.37%**	**14.76%**	+3.19%

Table 1: Results for provable robustness against adversarial examples on the CIFAR-100 dataset. Results with BCOP and Cayley convolutions are in Appendix Tables 10 and 11.

do not observe significant gains using Last Layer Normalization or LLN (Appendix Table 7). Thus, we do not include LLN for any of the results in Table 2. The column, "Increase $(108/255)$" denotes the increase in provable robust accuracy with $\rho = 108/255$ relative to "SOC + MaxMin".

For LipConvnet-$25, 30, 35, 40$ architectures, we observe gains in both the standard and provable robust accuracy by replacing MaxMin with the HH activation (i.e σ_θ). Similar to what we observe for CIFAR-100 in Table 1, the gains in provable robust accuracy ($\rho = 108/255$) are significantly higher for deeper networks: LipConvnet-35 (3.65%) and LipConvnet-40 (4.35%) with decent gains in standard accuracy (1.71 and 1.61% respectively).

Architecture	Methods	Standard Accuracy	Provable Robust Acc. ($\rho =$) 36/255	72/255	108/255	Increase (108/255)
LipConvnet-5	SOC + MaxMin	75.78%	59.18%	42.01%	27.09%	_
	+ HH	**76.30%**	60.12%	43.20%	27.39%	+0.30%
	+ CR	75.31%	**60.37%**	**45.62%**	**32.38%**	**+5.29%**
LipConvnet-10	SOC + MaxMin	76.45%	60.86%	44.15%	29.15%	_
	+ HH	**76.86%**	61.52%	44.91%	29.90%	+0.75%
	+ CR	76.23%	**62.57%**	**47.70%**	**34.15%**	**+5.00%**
LipConvnet-15	SOC + MaxMin	76.68%	61.36%	44.28%	29.66%	_
	+ HH	**77.41%**	61.92%	45.60%	31.10%	+1.44%
	+ CR	76.39%	**62.96%**	**48.47%**	**35.47%**	**+5.81%**
LipConvnet-20	SOC + MaxMin	76.90%	61.87%	45.79%	31.08%	_
	+ HH	**76.99%**	61.76%	45.59%	30.99%	-0.09%
	+ CR	76.34%	**62.63%**	**48.69%**	**36.04%**	**+4.96%**
LipConvnet-25	SOC + MaxMin	75.24%	60.17%	43.55%	28.60%	_
	+ HH	**76.37%**	61.50%	44.72%	29.83%	+1.23%
	+ CR	75.21%	**61.98%**	**47.93%**	**34.92%**	**+6.32%**
LipConvnet-30	SOC + MaxMin	74.51%	59.06%	42.46%	28.05%	_
	+ HH	**75.25%**	59.90%	43.85%	29.35%	+1.30%
	+ CR	74.23%	**60.64%**	**46.51%**	**34.08%**	**+6.03%**
LipConvnet-35	SOC + MaxMin	73.73%	58.50%	41.75%	27.20%	_
	+ HH	**75.44%**	61.05%	45.38%	30.85%	+3.65%
	+ CR	74.25%	**61.30%**	**47.60%**	**35.21%**	**+8.01%**
LipConvnet-40	SOC + MaxMin	71.63%	54.39%	37.92%	24.13%	_
	+ HH	**73.24%**	58.12%	42.24%	28.48%	+4.35%
	+ CR	72.59%	**59.04%**	**44.92%**	**32.87%**	**+8.74%**

Table 2: Results for provable robustness against adversarial examples on the CIFAR-10 dataset. Results with BCOP and Cayley convolutions are in Appendix Tables 4 and 5.

Again similar to CIFAR-100, adding CR further boosts the provable robust accuracy while slightly reducing the standard accuracy. Comparing "+ CR" with "SOC + MaxMin", we observe small drops in standard accuracy for LipConvnet-$5, 10, \ldots, 30$ networks (max. drop of -0.56%), and gains for LipConvnet-35 ($+0.52\%$) and LipConvnet-40 ($+0.96\%$). For provable robust accuracy ($\rho = 108/255$), we observe very significant gains of $> 4.96\%$ for all networks and $> 8\%$ for the deeper LipConvnet-35, 40 networks.

8 Conclusion

In this work, we introduce a procedure to certify robustness of 1-Lipschitz convolutional neural networks without orthogonalizing of the last linear layer of the network. Additionally, we introduce a certificate regularization that significantly improves the provable robust accuracy for these models at higher l_2 radii. Finally, we introduce a class of GNP activation functions called Householder (or HH) activations and prove that the Jacobian of any Gradient Norm Preserving (GNP) piecewise linear function is only allowed to change via Householder transformations for the function to be continuous which provides a general principle for designing piecewise linear GNP functions. These ideas lead to improved deterministic ℓ_2 robustness certificates on CIFAR-10 and CIFAR-100.

REPRODUCIBILITY

The code for the experiments is provided in the supplementary materials.

ETHICS STATEMENT

This paper introduces novel set of techniques for improving the provable robustness of 1-Lipschitz CNNs. We do not see any foreseeable negative consequences associated with our work.

REFERENCES

Cem Anil, James Lucas, and Roger B. Grosse. Sorting out lipschitz function approximation. In *ICML*, 2018.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6241–6250, USA, 2017. Curran Associates Inc. ISBN 978-1-5108-6096-4. URL http://dl.acm.org/citation.cfm?id=3295222.3295372.

Xiaoyu Cao and Neil Zhenqiang Gong. Mitigating evasion attacks to deep neural networks via region-based classification. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC 2017, pp. 278–287, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450353458. doi: 10.1145/3134600.3134606. URL https://doi.org/10.1145/3134600.3134606.

Moustapha Cissé, Piotr Bojanowski, Edouard Grave, Yann N. Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pp. 854–863. PMLR, 2017. URL http://proceedings.mlr.press/v70/cisse17a.html.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, 2019.

Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of relu networks via maximization of linear regions. *AISTATS 2019*, 2019.

J. Deng, Wei Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J. Cree. Regularisation of neural networks by enforcing lipschitz continuity, 2020.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30, pp. 5767–5777. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/892c3b1c6dccd52936e27cbd0ff683d6-Paper.pdf.

Aounon Kumar, Alexander Levine, Soheil Feizi, and Tom Goldstein. Certifying confidence via randomized smoothing. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 5165–5177. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/37aa5dfc44dddd0d19d4311e2c7a0240-Paper.pdf.

Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5458–5467. PMLR, 13–18 Jul 2020b. URL http://proceedings.mlr.press/v119/kumar20b.html.

Mathias Lécuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and S. K. K. Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE S&P 2019*, 2018.

Alexander Levine and Soheil Feizi. Improved, deterministic smoothing for l1 certified robustness. In *ICML*, 2021.

Alexander Levine, Sahil Singla, and Soheil Feizi. Certifiably robust interpretation in deep learning, 2019.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alche-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 9464–9474. Curran Associates, Inc., 2019a. URL https://proceedings.neurips.cc/paper/2019/file/335cd1b90bfa4ee70b39d08a4ae0cf2d-Paper.pdf.

Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger Grosse, and Jörn-Henrik Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. *Conference on Neural Information Processing Systems*, 2019b.

Xuanqing Liu, Minhao Cheng, Huan Zhang, and C. Hsieh. Towards robust neural networks via random self-ensemble. In *ECCV*, 2018.

Philip M. Long and Hanie Sedghi. Generalization bounds for deep convolutional neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=r1e_FpNFDr.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

Haifeng Qian and Mark N. Wegman. L2-nonexpansive neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxGSsR9FQ.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *NeurIPS*, 2018.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 11292–11303. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/3a24b25a7b092a252166a1641ae953e7-Paper.pdf.

Hanie Sedghi, Vineet Gupta, and Philip M. Long. The singular values of convolutional layers. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rJevYoA9Fm.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Puschel, and Martin T. Vechev. Fast and effective robustness certification. In *NeurIPS*, 2018.

Sahil Singla and Soheil Feizi. Second-order provable defenses against adversarial attacks. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8981–8991. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/singla20a.html.

Sahil Singla and Soheil Feizi. Skew orthogonal convolutions. In *ICML*, 2021. URL https://arxiv.org/abs/2105.11417.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL http://arxiv.org/abs/1312.6199.

Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Pbj8H_jEHYv.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *ICLR*, 2018.

Yusuke Tsuzuku, I. Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *NeurIPS*, 2018.

Cedric Villani. Optimal transport, old and new, 2008.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for ReLU networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5276–5285. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/weng18a.html.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5286–5295, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/wong18a.html.

Eric Wong, Frank R. Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses. In *NeurIPS*, 2018.

Lechao Xiao, Yasaman Bahri, Jascha Sohl-Dickstein, Samuel Schoenholz, and Jeffrey Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/xiao18a.html.

Bohang Zhang, Tianle Cai, Zhou Lu, Di He, and Liwei Wang. Towards certifying linfinity robustness using neural networks with linfinity-dist neurons. In *ICML*, 2021.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems (NIPS), arXiv preprint arXiv:1811.00866*, dec 2018.

Huan Zhang, Pengchuan Zhang, and Cho-Jui Hsieh. Recurjac: An efficient recursive algorithm for bounding jacobian matrix of neural networks and its applications. In *AAAI Conference on Artificial Intelligence (AAAI), arXiv preprint arXiv:1810.11783*, dec 2019.