Improved techniques for deterministic $l_2$ robustness

Sahil Singla
Department of Computer Science
University of Maryland
ssingla@umd.edu

Soheil Feizi
Department of Computer Science
University of Maryland
sfeizi@umd.edu

Abstract

Training convolutional neural networks (CNNs) with a strict 1-Lipschitz constraint under the $l_2$ norm is useful for adversarial robustness, interpretable gradients and stable training. 1-Lipschitz CNNs are usually designed by enforcing each layer to have an orthogonal Jacobian matrix (for all inputs) to prevent the gradients from vanishing during backpropagation. However, their performance often significantly lags behind that of heuristic methods to enforce Lipschitz constraints where the resulting CNN is not provably 1-Lipschitz. In this work, we reduce this gap by introducing (a) a procedure to certify robustness of 1-Lipschitz CNNs by replacing the last linear layer with a 1-hidden layer MLP that significantly improves their performance for both standard and provably robust accuracy, (b) a method to significantly reduce the training time per epoch for Skew Orthogonal Convolution (SOC) layers (> 30% reduction for deeper networks) and (c) a class of pooling layers using the mathematical property that the $l_2$ distance of an input to a manifold is 1-Lipschitz. Using these methods, we significantly advance the state-of-the-art for standard and provable robust accuracies on CIFAR-10 (gains of +1.79% and +3.82%) and similarly on CIFAR-100 (+3.78% and +4.75%) across all networks. Code is available at https://github.com/singlasahil14/improved_l2_robustness.

1 Introduction

The Lipschitz constant $\text{Lip}(f) : \mathbb{R}^d \rightarrow \mathbb{R}^c$, denoted by Lip($f$), controls the change in the output divided by change in the input (both changes measured in the $l_2$ norm). Previous work provides evidence that a small Lipschitz constant is useful for interpretable saliency maps [Tsipras et al., 2018], generalization bounds [Long and Sedghi, 2020], Wasserstein distance estimation [Villani, 2008], adversarial robustness [Szegedy et al., 2014] and preventing gradient explosion during backpropagation [Xiao et al., 2018]. Several prior works [Miyato et al., 2018, Gulrajani et al., 2017] use heuristic methods to enforce Lipschitz constraints to successfully address problems such as stabilizing GAN training. However, these methods do not enforce a guaranteed lipschitz bound and it remains challenging to achieve strong results with provable guarantees.

Using the composition property i.e. Lip($g \circ h$) $\leq$ Lip($g$) Lip($h$), we can construct a 1-Lipschitz neural network by constraining each layer to be 1-Lipschitz. However, a key difficulty with this approach is that because a 1-Lipschitz layer can only reduce the gradient norm during backpropagation, for deeper networks, this results in small gradient values for layers closer to the input making training slow and difficult. To address this problem, [Anil et al., 2018] introduce Gradient Norm Preserving (GNP) architectures where each layer preserves the gradient norm during backpropagation. For 1-Lipschitz Convolutional Neural Networks (CNNs), this involves using orthogonal convolutions (convolution layers with an orthogonal Jacobian matrix) [Li et al., 2019b, Trockman and Kolter].

In this work, we assume the Lipschitz constant under the $l_2$ norm.

While several works have attempted to construct novel and more expressive orthogonal convolution layers and GNP activation functions called Householder activations [Singla et al., 2022], and using a class of GNP activation functions called Householder activations [Singla et al., 2022].

Among orthogonal convolutions, SOC [Singla and Feizi, 2021] achieves state-of-the-art results on CIFAR-10, 100 [Singla et al., 2022]. SOC uses the following two mathematical properties to construct an orthogonal convolution: (a) the exponential of a skew symmetric matrix is orthogonal and (b) the matrix exponential can be computed using the exponential series. A drawback of using the exponential series is that it requires us to apply the convolution operation multiple times to achieve a reasonable approximation of an orthogonal matrix. Consequently, during backpropagation, computing the gradient with respect to the weights for a single SOC layer requires us to compute the weight gradient for each of these convolutions resulting in significant training overhead. In this work, we show that because the matrix is skew symmetric, using some approximations, the gradients with respect to weights can be computed in significantly reduced time with no loss in performance. This enables us to reduce the training time per epoch using SOC by > 30% for networks with large number of SOC layers (Results in Table 1).

Another limitation of 1-Lipschitz CNNs is that their performance is often significantly below compared to that of standard CNNs. Recently, [Singla et al., 2022] introduced a procedure for certifying robustness by relaxing the orthogonality requirements of the last linear layer (i.e. the linear layer mapping penultimate neurons to class logits) achieving state-of-the-art results on CIFAR-10, 100 [Krizhevsky, 2009]. Since MLPs are more expressive than linear layers, one would expect improved standard accuracy by replacing the last linear layer with them. However, the resulting networks are not 1-Lipschitz and achieving high robustness guarantees (provable robust accuracy) is difficult.

To certify robustness for these networks, note that since the mapping from input to penultimate layer is 1-Lipschitz, the robustness certificate for the penultimate output also provides a certificate for the input. Thus, we first replace the linear layer mapping penultimate output to logits with a 1-hidden layer MLP (Multi layer Perceptron) because such MLPs are easier to certify compared to deep MLPs. To certify robustness for the MLP, we use the Curvature-based Robustness Certificate or CRC [Singla and Feizi, 2020]. To train MLP to have high robustness guarantees, we use a variant of adversarial training that only applies adversarial perturbations to the MLP (not the whole network). We call our certification procedure CRC-Lip. This results in improved results for both the standard and also the provable robust accuracy across all network architectures on CIFAR-10 (≥ 1.63%, ≥ 3.14%) and CIFAR-100 (≥ 2.49%, ≥ 2.27% respectively). Results are in Tables 3 and 4.

While several works have attempted to construct novel and more expressive orthogonal convolution layers and GNP activation functions, current state-of-the-art 1-Lipschitz CNNs still use pooling layers by taking the max of different elements. In this work, we introduce a class of 1-Lipschitz pooling layers called projection pooling using the following mathematical property: Given a manifold \( \mathcal{M} \subset \mathbb{R}^n \) and input \( x \in \mathbb{R}^n \), the function \( d_{\mathcal{M}} : \mathbb{R}^n \rightarrow \mathbb{R} \) defined as the \( l_2 \) distance of \( x \) to \( \mathcal{M} \) is 1-Lipschitz. Thus, to construct a pooling layer, we can first define a learnable manifold with parameters \( \Theta \) (Example in Section 5). During training, for input \( x \in \mathbb{R}^n \), the pooling layer simply outputs \( d_{\mathcal{M}}(x) \in \mathbb{R} \) as the output, resulting in decrease in input dimension by a factor of \( n \). Moreover, since the output \( d_{\mathcal{M}}(x) \) is also a function of \( \Theta \), \( \Theta \) can be learned during training. However, solving for the distance \( d_{\mathcal{M}}(x) \) can be difficult especially when \( \mathcal{M} \) is a high-dimensional manifold.

### Table 1: CIFAR-10 results using faster gradients, projection pooling, CRC-Lip

<table>
<thead>
<tr>
<th>SOC layers</th>
<th>Time/epoch (secs)</th>
<th>Reduction</th>
<th>Standard accuracy</th>
<th>Provable robust accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ours</td>
<td>Previous</td>
<td>Ours</td>
<td>Previous</td>
</tr>
<tr>
<td>6</td>
<td>25.34</td>
<td>30.63</td>
<td>-17.27%</td>
<td>79.36%</td>
</tr>
<tr>
<td>11</td>
<td>37.11</td>
<td>48.44</td>
<td>-23.39%</td>
<td>79.57%</td>
</tr>
<tr>
<td>16</td>
<td>49.27</td>
<td>66.74</td>
<td>-25.34%</td>
<td>79.44%</td>
</tr>
<tr>
<td>21</td>
<td>61.59</td>
<td>83.18</td>
<td>-29.74%</td>
<td>79.13%</td>
</tr>
<tr>
<td>26</td>
<td>71.51</td>
<td>100.70</td>
<td>-31.17%</td>
<td>79.19%</td>
</tr>
<tr>
<td>31</td>
<td>84.00</td>
<td>119.55</td>
<td>-31.17%</td>
<td>79.64%</td>
</tr>
<tr>
<td>36</td>
<td>95.06</td>
<td>137.10</td>
<td>-30.66%</td>
<td>78.57%</td>
</tr>
<tr>
<td>41</td>
<td>106.01</td>
<td>156.25</td>
<td>-32.16%</td>
<td>78.41%</td>
</tr>
</tbody>
</table>
To address this limitation, (a) we use 2D projection pooling layers that reduce the dimension by factor of 2 and (b) we construct these layers using piecewise linear curves for which the distance can be computed efficiently by computing the minimum distance to all line segments and the connecting points (Example in Appendix Figure [2]). If the curve is closed and without self-intersections, we can also define a signed projection pooling for which the signs of the output for points inside and outside the region enclosed by the curve are different (x and y in Figure [2]). This allows the subsequent layers to distinguish between the inputs inside and outside the region. In this work, we show some preliminary results using a simple 2D projection pooling layer (Section 6). We leave the problem of constructing high performance 2D projection pooling layers open for future research.

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

- We introduce a method for faster computation of the weight gradient for SOC layers. For deeper networks, we observe reduction in training time per epoch by ≥ 30% (Table1).

- We introduce a certification procedure called CRC-Lip that replaces the last linear layer with a 1-hidden layer MLP and results in significantly improved standard and provable robust accuracy. For deeper networks (≥ 35 layers), we observe improvements of ≥ 5.84%, ≥ 8.00% in standard and ≥ 8.76%, ≥ 8.65% in provable robust accuracy (l2 radius 36/255) on CIFAR-10,100 respectively.

- We introduce a large class of 1-Lipschitz pooling layers called projection pooling using the mathematical property that the l2 distance $d_M(x)$ of an input x to the manifold M is 1-Lipschitz.

- On CIFAR-10, across all architectures, we achieve the best standard and provable robust accuracy (at 36/255) of 79.57, 67.13% respectively (gain of +1.79%, +3.82% from prior works). Similarly, on CIFAR-100, we achieve 51.84, 39.27% (+3.78%, +4.75% from prior works). These results establish new state-of-the-art results in the standard and provable robust accuracy on both datasets.

2 Provable defenses against adversarial examples

Provable defenses against adversarial examples: For a provably robust classifier, we can guarantee that its predictions remain constant within some region around the input. Most of the existing methods for provable robustness either bound the Lipschitz constant or use convex relaxations [Singh et al., 2017, 2018a, 2019a, b, c, Weng et al., 2018, Salman et al., 2019a, b, Zhang et al., 2019, 2018, Wong et al., 2018, Wong and Kolter, 2018, Singh et al., 2018b, Raghunathan et al., 2018, Dvijotham* et al., 2018, Croce et al., 2019, Gowal et al., 2019, Dvijotham* et al., 2020, Lu and Kumar, 2020, Singla and Feizi, 2020, Bunel et al., 2020, Leino et al., 2021, Leino and Fredrikson, 2021, Zhang et al., 2021, 2022, Wang et al., 2021, Huang et al., 2021, Müller et al., 2021, Singh et al., 2021, Palma et al., 2021]. However, these methods are often not scalable to large neural networks while achieving high performance. In contrast, randomized smoothing [Liu et al., 2018, Cao and Gong, 2017, Lécuyer et al., 2018, Lu et al., 2019a, Cohen et al., 2019, Salman et al., 2019a, Levine et al., 2019, Kumar et al., 2020a, b, Salman et al., 2020, 2021] scales to large neural networks but is a probabilistically certified defense: certifying robustness with high probability requires generating a large number of noisy samples leading to high inference-time. The defense we propose in this work is deterministic and not comparable to randomized smoothing.

Provably Lipschitz neural networks: The class of Gradient Norm Preserving (GNP) and 1-Lipschitz fully connected neural networks was first introduced by Anil et al., 2018. To design each layer to be GNP, they orthogonalize weight matrices and use a class of piecewise linear GNP activations called GroupSort. Later, Singla et al., 2022 proved that for any piecewise linear GNP function to be continuous, different Jacobian matrices in a neighborhood must change via householder transformations, implying that GroupSort is a special case of more general HouseHolder activations. Several previous works enforce Lipschitz constraints on convolution layers using spectral normalization, clipping or regularization [Cissé et al., 2017, Tsuzuku et al., 2018, Qian and Wegman, 2019, Gouk et al., 2020, Sedghi et al., 2019]. However, these methods either enforce loose lipschitz bounds or do not scale to large networks. To ensure that the Lipschitz constraint on convolutional layers is tight, recent works construct convolution layers with an orthogonal Jacobian [Lu et al., 2019b, Trockman and Kolter, 2021, Singla and Feizi, 2021, Yu et al., 2022, Su et al., 2022, Kiani et al., 2022]. These approaches avoid the aforementioned issues and allow training of large, provably 1-Lipschitz CNNs achieving state-of-the-art results for deterministic l2 robustness.
3 Problem setup and Notation

For a vector \( \mathbf{v}, \mathbf{v}_j \) denotes its \( j^{th} \) element. For a matrix \( \mathbf{A}, \mathbf{A}_{jk} \) and \( \mathbf{A}(:,k) \) denote the \( j^{th} \) row and \( k^{th} \) column respectively. Both \( \mathbf{A}_{jk} \) and \( \mathbf{A}(:,k) \) are assumed to be column vectors (thus \( \mathbf{A}_{jk} \) is the transpose of \( j^{th} \) row of \( \mathbf{A} \)). \( \mathbf{A}_{jk} \) denotes the element in \( j^{th} \) row and \( k^{th} \) column of \( \mathbf{A} \). \( \mathbf{A}_{jk}, \mathbf{A}_{j} \) denotes the matrix containing the first \( j \) rows and \( k \) columns of \( \mathbf{A} \). We define \( \mathbf{A}_{j} = \mathbf{A}_{j,:} \) and \( \mathbf{A}_{j} = \mathbf{A}_{:,j} \). Similar notation applies to higher order tensors. \( \mathbf{I} \) denotes the identity matrix, \( \mathbb{R} \) to denote the field of real numbers. We construct a 1-Lipschitz neural network, \( f: \mathbb{R}^d \rightarrow \mathbb{R}^c \) (\( d \) is the input dimension, \( c \) is the number of classes) by composing 1-Lipschitz convolution layers and GNP activation functions. We often use the abbreviation \( f^{\exp} \).

We approximate the exponential series:

\[
\exp(w.r.t. \text{layer output } z)
\]

Among the existing orthogonal convolution layers in the literature, Skew Orthogonal Convolutions (or SOC) by Singla and Feizi \[2021\] achieves state-of-the-art results on CIFAR-10,100 \[Singla et al. 2022\]. SOC first constructs a convolution filter \( \mathbf{L} \) whose Jacobian is skew-symmetric. This is followed by a convolution exponential \[Hoogeboom et al., 2020\] operation. Since the exponential of a skew-symmetric matrix is orthogonal, the Jacobian of the resulting layer is an orthogonal matrix.

However, a drawback of this procedure is that convolution exponential requires multiple convolution operations per SOC layer to achieve a reasonable approximation of the orthogonal Jacobian. Consequently, if we use \( k \) convolution operations in the SOC layer during forward pass, we need to compute the gradient with respect to the weights \( \mathbf{L} \) (called convolution weight gradient) per convolution operation \( k \) times which can lead to slower training time especially when the number of SOC layers is large. To address this limitation, we show that even if we use \( k \) convolutions in the forward pass of an SOC layer, the weight gradient for the layer can be computed using a single convolution weight gradient during backpropagation leading to significant reduction in training time.

For simplicity, let us first consider the case of an orthogonal fully connected layer (i.e. not convolution) with the same input and output size \( (n) \). Later, we will see that our analysis leads to improvements for orthogonal convolutional layers. Further, assume that the weights i.e. \( \mathbf{A} \in \mathbb{R}^{n \times n} \) are skew-symmetric i.e. \( \mathbf{A} = -\mathbf{A}^T \) and given the input \( x \in \mathbb{R}^n \), the output \( z \in \mathbb{R}^n \) is computed as follows:

\[
\mathbf{A} = -\mathbf{A}^T
\]

\[
\mathbf{z} = \sum_{i=0}^{k} \frac{\mathbf{A}^i}{i!} \mathbf{x} + \mathbf{b}, \quad \text{where } \mathbf{A} = -\mathbf{A}^T
\]

We approximate the exponential series: \( \exp(\mathbf{A}) = \sum_{i=0}^{\infty} \frac{\mathbf{A}^i}{i!} \) using a finite number of terms \( (k) \).

**Forward pass \((\mathbf{z})\):** To compute \( \mathbf{z} \) during the forward pass, we use the following iterations:

\[
\mathbf{u}^{(i)} = \begin{cases} 
\mathbf{x} & i = k - 1 \\
\mathbf{x} + (\mathbf{A} \mathbf{u}^{(i+1)})/(i+1) & i \leq k - 2
\end{cases}
\]

(2)

It can be shown that \( \mathbf{z} = \mathbf{u}^{(0)} \) (Appendix C). During backpropagation, given the gradient of loss \( \ell \) w.r.t. layer output \( \mathbf{z} \) i.e. \( \nabla_\mathbf{z} \ell \), we want to compute \( \nabla_\mathbf{z} \ell \) (input gradient) and \( \nabla_\mathbf{A} \ell \) (weight gradient).

**Input gradient \((\nabla_\mathbf{z} \ell)\):** To compute \( \nabla_\mathbf{z} \ell \), observe that \( \mathbf{z} \) is a linear function of \( \mathbf{x} \). Thus, using the chain rule, skew-symmetry \((\mathbf{A}^T = -\mathbf{A})\) and the property \((\mathbf{A}^i)^T = (\mathbf{A}^T)^i\), we have:

\[
\nabla_\mathbf{z} \ell = \left( \sum_{i=0}^{k} \frac{\mathbf{A}^i}{i!} \right)^T \nabla_\mathbf{z} \ell = \left( \sum_{i=0}^{\infty} \frac{(-\mathbf{A})^i}{i!} \right)^T \nabla_\mathbf{z} \ell
\]
We can again approximate the exponential series using the same number of terms as in the forward pass i.e. $k$. To compute the finite term approximation, we use the following iterations:

$$
\mathbf{v}^{(i)} = \begin{cases} 
\nabla \mathbf{z} \ell & i = k - 1 \\
\nabla \mathbf{z} \ell - (\mathbf{A} \mathbf{v}^{(i+1)})/(i + 1) & i \leq k - 2
\end{cases}
$$

(3)

Similar to forward pass, it can be shown that $\nabla_x \ell = \mathbf{v}^{(0)}$ (Appendix C).

**Weight gradient** ($\nabla_{\mathbf{A}} \ell$): We first derive the exact expression for $\nabla_{\mathbf{A}} \ell$ in the Theorem below:

**Theorem 1** The gradient of the loss function $\ell$ w.r.t $\mathbf{A}$ i.e. $\nabla_{\mathbf{A}} \ell$ is given by:

$$
\nabla_{\mathbf{A}} \ell = - \sum_{i=1}^{k} \left( (\mathbf{A}^{i-1}\mathbf{x}) \left( \mathbf{v}^{(i)} \right)^T - \mathbf{v}^{(i)} (\mathbf{A}^{i-1}\mathbf{x})^T \right)
$$

(4)

where $\mathbf{v}^{(i)}$ is defined as in equation (3).

Note that the first outer product i.e. $(\mathbf{A}^{i-1}\mathbf{x}) \left( \mathbf{v}^{(i)} \right)^T$ and the second i.e. $(\mathbf{v}^{(i)}) (\mathbf{A}^{i-1}\mathbf{x})^T$ are transpose of each other implying that each term in the summation is skew-symmetric. Although these outer products can be computed in a straightforward way for orthogonal fully connected layers, this is not the case for orthogonal convolution layers (SOC in this case). This is because, for a convolution filter $\mathbf{L} \in \mathbb{R}^{p \times q \times r \times s}$, the term analogous to $\mathbf{A}^{i-1}\mathbf{x}$ is a patch of size $q \times r \times s$ and that analogous to $\mathbf{v}^{(i)}$ is another patch of size $p \times 1 \times 1$ resulting in an outer product of the desired size $p \times q \times r \times s$ per patch. Thus, for SOC layers, each term inside the summation is computed by summing over the outer products for all such patches. For large input sizes, the number of such patches can often be large, making this computation expensive. To address this limitation, we use the following approximation:

$$
\nabla_{\mathbf{A}} \ell \approx - \left( \mathbf{u}^{(1)} \left( \mathbf{v}^{(1)} \right)^T - \mathbf{v}^{(1)} \left( \mathbf{u}^{(1)} \right)^T \right)
$$

(5)

In Appendix B we show that the above approximation is principled because after subtracting the exact and approximation gradients (equations (4) and (5)) and simplifying, each term in the resulting series is divided by a large value in its denominator ($\approx 0$). This approximation is useful because in equation (5), the outer product need not be computed once whereas in equation (4), the outer products need to be computed $k$ times. Also, both $\mathbf{u}^{(1)}$ and $\mathbf{v}^{(1)}$ are computed while computing $\mathbf{u}^{(0)} = \mathbf{z}$ during the forward pass and $\mathbf{v}^{(0)} = \nabla_x \mathbf{z}$ during the backward pass using the recurrences in equations (2) and (3). Thus, we can simply store $\mathbf{u}^{(1)}$, $\mathbf{v}^{(1)}$ during the forward/backward pass respectively so that $\nabla_{\mathbf{A}} \ell$ can be computed directly using equation (5). In our experiments, we observe that this leads to significant reduction in training time with almost no drop in performance (Tables 1, 3, 4).

5 **Curvature-based Robustness Certificate**

A key property of 1-Lipschitz CNNs is that the output of each layer is 1-Lipschitz with respect to the input. Given an input $\mathbf{x} \in \mathbb{R}^d$, consider the penultimate output $g(\mathbf{x}) \in \mathbb{R}^m$ and logits $f(\mathbf{x}) \in \mathbb{R}^c$ for some 1-Lipschitz CNN. Existing robustness certificates [Li et al., 2019b, Singla et al., 2022] rely on the linearity of the function from $g(\mathbf{x}) \rightarrow f(\mathbf{x})$. However, since MLPs have higher expressive power than linear functions [Cybenko, 1989], one way to improve performance could be to replace this mapping with MLPs. However, certifying robustness for deep MLPs is difficult.

To address these challenges, we first replace the mapping from $g(\mathbf{x}) \rightarrow f(\mathbf{x})$ with a 1-hidden layer MLP because they are easier to certify compared to deeper networks. Because computing exact certificates for ReLU networks is known to be NP-complete [Simha et al., 2018], we use the differentiable Softplus activation [Dugas et al., 2020] to certify robustness. To certify robustness, we use the Curvature-based Robustness Certificate or CRC [Singla and Feizi, 2020] because it provides exact certificates for a significant fraction of inputs for shallow MLPs. We provide a brief review of CRC in Appendix Section D. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$ be a 1-Lipschitz continuous function and $h : \mathbb{R}^m \rightarrow \mathbb{R}^c$ be a 1-hidden layer MLP such that $f = h \circ g$. Further, let $l$ be the predicted class for input $\mathbf{x}$ i.e. $f_l(\mathbf{x}) \geq \max_{i \neq l} f_i(\mathbf{x})$. Since $g$ is 1-Lipschitz, it can be shown that if $h$ is provably robust in an $l_2$ radius $R$ around input $g(\mathbf{x})$, then $f$ is also provably robust in the $l_2$ radius $R$ around $\mathbf{x}$. The resulting procedure is called CRC-Lip and is given in the following proposition:
Although max we simply use an identity map (same gradient output as the gradient input) to backpropagate through only an adversarial training procedure that applies to the input of the MLP \( g(\mathbf{x}) \) (i.e. not the input of the neural network \( \mathbf{x} \)) and (b) curvature regularization to reduce the curvature of the MLP. It was shown in [Singla and Feizi 2021] that combining adversarial training with curvature regularization leads to significantly improved provable robust accuracy with small reduction in standard accuracy.

This results in the following loss function for training:

\[
\min_{\Omega, \Psi} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} \left( \max_{\|\mathbf{y}^* - g(\mathbf{x})\|_2 \leq \rho} \ell(h_{\Omega}(\mathbf{y}^*, l)) \right) + \gamma \mathcal{K}_h
\]

In the above equation, \( \Omega \) denote the parameters of the MLP (i.e. \( h \)), \( \Psi \) are the parameters of the 1-Lipschitz function \( g \), \( \mathcal{K}_h \) is the bound on the curvature of the MLP, \( \rho \) denotes the \( l_2 \) perturbation radius and \( \gamma \) denotes the curvature regularization coefficient. The curvature bound \( \mathcal{K}_h \) is the same as in [Singla and Feizi 2020]. Also, since we apply adversarial perturbations in the penultimate layer, we also need to backpropagate through this procedure to enable training of previous layers. To this end, we simply use an identity map (same gradient output as the gradient input) to backpropagate through the adversarial training procedure and find that it works well in practice, achieving significantly better results compared to the state-of-the-art. It is possible that a more principled method of backpropagation may lead to better results and we leave that avenue open for future research.

### 6 Projection pooling layers

In 1-Lipschitz CNNs, pooling is usually carried out as follows: given input \( \mathbf{X} \in \mathbb{R}^{q \times r \times r} \) (\( r \) is even), we first use rearrangement [Jacobsen et al. 2018] illustrated in Figure 1(a) to construct \( \mathbf{X}' \in \mathbb{R}^{4q \times (r/2) \times (r/2)} \). Next, we apply an orthogonal convolution which gives an output of the same size \( \mathbf{Z} \in \mathbb{R}^{4q \times (r/2) \times (r/2)} \) and divide it into two tensors of equal sizes (along the channel dimension) giving \( \mathbf{Z}_{2q} \in \mathbb{R}^{2q \times (r/2) \times (r/2)} \) and \( \mathbf{Z}_{2q} \in \mathbb{R}^{2q \times (r/2) \times (r/2)} \). We then define \( \max(\mathbf{Z}_{2q}, \mathbf{Z}_{2q}) \) (or either one of \( \mathbf{Z}_{2q}, \mathbf{Z}_{2q} \)) as the output of the pooling layer.

Although \( \max \) is 1-Lipschitz, its expressive power is limited. For example, consider the 1-Lipschitz function \( \|x, y\|_2 = \sqrt{x^2 + y^2} \). It is easy to see that if \( x = y \), the error between \( \|x, y\|_2 \) and \( \max(x, y) \) can be arbitrarily large for large values of \( x, y \). To address such limitations, we construct expressive pooling layers using the following mathematical property:

**Theorem 2** Given \( \mathbf{x} \in \mathbb{R}^n \) and manifold \( \mathcal{M} \subset \mathbb{R}^n \), the distance function (in \( l_2 \) norm) is 1-Lipschitz:

\[
d_M(\mathbf{x}) = \min_{\mathbf{x}^* \in \mathcal{M}} \|\mathbf{x}^* - \mathbf{x}\|_2 \implies |d_M(\mathbf{x}) - d_M(\mathbf{y})| \leq \|\mathbf{x} - \mathbf{y}\|_2
\]
We perform experiments under the setting of provably robust image classification on CIFAR-10 and CIFAR-100 datasets. We use the LipConvnet-5, 10, 15, . . . , 40 architectures for comparison. We use SOC as the orthogonal convolution and MaxMin as the activation in all architectures. 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 adversarial training with curvature regularization, we use $\rho = 36/255$ (0.1411), $\gamma = 0.5$ for CIFAR-10 and $\rho = 0.2$, $\gamma = 0.75$ for CIFAR-100. We find that certifying robustness using CRC-Lip is computationally expensive for CIFAR-100 due to large number of classes. To address this issue, we only use classes with top-10 logits (instead of all 100 classes) for CIFAR-100 (Table 4). Since CRC-Lip requires us to solve a convex optimization, we consider a certificate to be valid if the input is correctly classified and gradient at the optimal solution is $\leq 10^{-6}$ (0 otherwise). All results are

<table>
<thead>
<tr>
<th>Output Size</th>
<th>Layer</th>
<th>Repeats</th>
</tr>
</thead>
<tbody>
<tr>
<td>$32 \times 32 \times 32$</td>
<td>Conv + MaxMin</td>
<td>q × r × r</td>
</tr>
<tr>
<td>$64 \times 16 \times 16$</td>
<td>LipBlock-n/5</td>
<td>Input</td>
</tr>
<tr>
<td>$128 \times 8 \times 8$</td>
<td>LipBlock-n/5</td>
<td>$q \times r \times r$</td>
</tr>
<tr>
<td>$256 \times 4 \times 4$</td>
<td>LipBlock-n/5</td>
<td>Conv + MaxMin</td>
</tr>
<tr>
<td>$512 \times 2 \times 2$</td>
<td>LipBlock-n/5</td>
<td>$4q \times (r/2) \times (r/2)$</td>
</tr>
<tr>
<td>$1024 \times 1 \times 1$</td>
<td>LipBlock-n/5</td>
<td>Pooling</td>
</tr>
<tr>
<td># of classes</td>
<td>Linear/MLP-1</td>
<td>$2q \times (r/2) \times (r/2)$</td>
</tr>
</tbody>
</table>

Table 2: LipConvnet-n and LipBlock-n/5 architectures

operation to an input $x \in \mathbb{R}^n$, we simply output its $l_2$ distance to the manifold $M(\phi)$ denoted by $d_{M(\phi)}(x)$. Since $d_{M(\phi)}(x) \in \mathbb{R}$ while $x \in \mathbb{R}^n$, this operation reduces the input size by a factor of $n$.

As an example, let $M_u = \{u\}$. Here $u \in \mathbb{R}^n$ is the learnable parameter and the distance function $d_{M_u}(x) = \|x - u\|_2$. Even in this very simple case, for $u = 0$, $d_{M_u}(x) = \|x\|_2$ and for $n = 2$, this function can learn to represent the function $\|x, y\|_2$ discussed earlier exactly. We also prove that a signed $l_2$ distance function can be defined when $M$ satisfies certain properties:

**Corollary 1** If $M$ is a connected manifold that divides $\mathbb{R}^n$ into two nonempty connected sets $A$ and $B$ such that $A \cap B = \phi$ and every path from $a \in A$ and $b \in B$ intersects a point on $M$, then there exists a signed $l_2$ distance function with different signs in $A$ and $B$.

Proofs of Theorems 2 and Corollary 1 are given in Appendix A.3 and A.4 respectively.

In practice, computing $d_M$ can be difficult for high-dimensional $M$. To tackle this issue, we use 2D pooling layers where $M$ is defined to be a piecewise linear curve in 2D. The $l_2$ distance can then be computed by finding the minimum distance to all the line segments and connecting points. We emphasize that even in this relatively simple case, $M$ can have large number of parameters and $d_M$ can still be efficient to compute because these individual distances can be computed in parallel. We use $M(\phi)$ defined below (illustrated in Figure 1b) lines $L_1$, $L_2$ correspond to $\phi = +\theta, -\theta$:

$$M(\phi) = \{R(\cos \phi, \sin \phi) : R \geq 0, \phi \in \{+\theta, -\theta\}\}$$

In each colored region (Figure 1b), $d_{M(\phi)}$ can be computed using the following:

$$d_{M(\phi)}(R(\cos \alpha, \sin \alpha)) = \begin{cases} R\sin(\alpha - \theta), & 0 \leq \alpha \leq \theta + \pi/2 \\ -R\sin(\alpha + \theta), & 0 < -\alpha \leq \theta + \pi/2 \\ R, & \text{otherwise} \end{cases}$$ (9)

To apply projection pooling on $Z_{2q}$, $Z_{2q}$, discussed previously, we output $d_{M(\phi)}(Z_{2q}, Z_{2q})$.

7 Experiments

We perform experiments under the setting of provably robust image classification on CIFAR-10 and CIFAR-100 datasets. We use the LipConvnet-5, 10, 15, . . . , 40 architectures for comparison. We use SOC as the orthogonal convolution and MaxMin as the activation in all architectures. 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 adversarial training with curvature regularization, we use $\rho = 36/255$ (0.1411), $\gamma = 0.5$ for CIFAR-10 and $\rho = 0.2$, $\gamma = 0.75$ for CIFAR-100. We find that certifying robustness using CRC-Lip is computationally expensive for CIFAR-100 due to large number of classes. To address this issue, we only use classes with top-10 logits (instead of all 100 classes) for CIFAR-100 (Table 4). Since CRC-Lip requires us to solve a convex optimization, we consider a certificate to be valid if the input is correctly classified and gradient at the optimal solution is $\leq 10^{-6}$ (0 otherwise). All results are
Table 3: Provable robustness results on CIFAR-10 (LipConvnet-5, 10 results in Appendix Table 5)

<table>
<thead>
<tr>
<th>LipConvnet Method</th>
<th>Standard Accuracy</th>
<th>Provable Robust Accuracy</th>
<th>Increase (Standard)</th>
<th>Increase (36/255)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>77.78%</td>
<td>62.75%</td>
<td>46.34%</td>
<td>31.38%</td>
</tr>
<tr>
<td>+ Fast</td>
<td>77.75%</td>
<td>62.52%</td>
<td>46.23%</td>
<td>31.19%</td>
</tr>
<tr>
<td>+ CRC</td>
<td>79.44%</td>
<td>66.99%</td>
<td>52.56%</td>
<td>38.30%</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>77.50%</td>
<td>63.31%</td>
<td>46.42%</td>
<td>31.53%</td>
</tr>
<tr>
<td>+ Fast</td>
<td>77.13%</td>
<td>62.05%</td>
<td>45.86%</td>
<td>31.13%</td>
</tr>
<tr>
<td>+ CRC</td>
<td>79.13%</td>
<td>66.45%</td>
<td>52.45%</td>
<td>38.12%</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>77.18%</td>
<td>62.46%</td>
<td>45.78%</td>
<td>31.16%</td>
</tr>
<tr>
<td>+ Fast</td>
<td>76.94%</td>
<td>61.91%</td>
<td>45.59%</td>
<td>30.69%</td>
</tr>
<tr>
<td>+ CRC</td>
<td>79.19%</td>
<td>66.28%</td>
<td>51.74%</td>
<td>37.99%</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>74.43%</td>
<td>59.65%</td>
<td>43.76%</td>
<td>29.16%</td>
</tr>
<tr>
<td>+ Fast</td>
<td>74.69%</td>
<td>58.84%</td>
<td>43.33%</td>
<td>28.93%</td>
</tr>
<tr>
<td>+ CRC</td>
<td>78.64%</td>
<td>66.05%</td>
<td>51.31%</td>
<td>37.30%</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>72.73%</td>
<td>57.18%</td>
<td>42.08%</td>
<td>28.09%</td>
</tr>
<tr>
<td>+ Fast</td>
<td>72.91%</td>
<td>57.58%</td>
<td>41.52%</td>
<td>27.37%</td>
</tr>
<tr>
<td>+ CRC</td>
<td>78.57%</td>
<td>65.94%</td>
<td>52.04%</td>
<td>37.63%</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>71.33%</td>
<td>55.74%</td>
<td>39.32%</td>
<td>26.06%</td>
</tr>
<tr>
<td>+ Fast</td>
<td>71.60%</td>
<td>56.15%</td>
<td>39.82%</td>
<td>25.63%</td>
</tr>
<tr>
<td>+ CRC</td>
<td>78.41%</td>
<td>65.51%</td>
<td>51.32%</td>
<td>37.30%</td>
</tr>
</tbody>
</table>

reported using the complete test sets of CIFAR-10 and CIFAR-100. We compare the provable robust accuracy using 3 different \(l_0\) perturbation radii: 36/255, 72/255, 108/255.

**Table details:** For the baseline ("Baseline" in Tables 3 and 4), we use the standard max pooling with the certificate based on LLN [Singla et al., 2022] due to their superior performance over prior works. In Tables 3 and 4 for each architecture, "+ Fast" adds faster gradient computation, "+ CRC" replaces max pooling with projection pooling (equation (9)) and replaces the last linear layer with a 1-hidden layer MLP (CRC-Lip certificate) while also using faster gradients. For each architecture, the columns "Increase (Standard)" and "Increase (36/255)" denote the increase in standard and provable robust accuracy relative to "Baseline" standard and provable robust accuracy (36/255). Results where Projection pooling and CRC-Lip are added separately are given in Appendix Tables 5 and 6.

**LipConvnet Architecture:** We use a 1-Lipschitz CNN architecture called LipConvnet-n where n is a multiple of 5 and \(n+1\) is the total number of convolution layers. It consists of an initial SOC layer that expands the number of channels from 3 to 32. This is followed by 5 blocks that reduce spatial dimensions (height and width) by half while doubling the number of channels. The architecture is summarized in Tables 2a and 2b. The last layer outputs the class logits and is either a linear layer in which case we use LLN to certify robustness or a single-hidden layer MLP where we use CRC-Lip.

**Correcting the certificates:** Since SOC is an approximation, the Lipschitz constant of each SOC layer can be slightly more than 1 and if we use a large number of SOC layers (e.g. 41 in LipConvnet-40), the Lipschitz constant of the full network \(\text{Lip}(f)\) can be significantly larger than 1. To mitigate this issue, we take the following steps: (a) We use a large number of terms \(k = 15\) during test time to approximate the exponential which results in a small approximation error (using the error bound in Singla and Feizi [2021]) and (b) We compute \(\text{Lip}(f)\) by multiplying the lipschitz constant of all SOC layers (using the power method) and then divide the certificate by \(\text{Lip}(f)\).
<table>
<thead>
<tr>
<th>LipConvnet-</th>
<th>Methods</th>
<th>Standard Accuracy</th>
<th>Provable Robust Accuracy</th>
<th>Increase (standard)</th>
<th>Increase (36/255)</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>Baseline</td>
<td>48.06%</td>
<td>34.52%</td>
<td>23.08%</td>
<td>14.70%</td>
</tr>
<tr>
<td></td>
<td>+ Fast</td>
<td>47.97%</td>
<td>33.84%</td>
<td>22.66%</td>
<td>14.26%</td>
</tr>
<tr>
<td></td>
<td>+ CRC</td>
<td>50.79%</td>
<td>37.50%</td>
<td>26.16%</td>
<td>17.27%</td>
</tr>
<tr>
<td>20</td>
<td>Baseline</td>
<td>47.37%</td>
<td>33.99%</td>
<td>23.40%</td>
<td>14.69%</td>
</tr>
<tr>
<td></td>
<td>+ Fast</td>
<td>46.41%</td>
<td>33.07%</td>
<td>22.06%</td>
<td>14.00%</td>
</tr>
<tr>
<td></td>
<td>+ CRC</td>
<td>51.84%</td>
<td>38.54%</td>
<td>27.32%</td>
<td>18.53%</td>
</tr>
<tr>
<td>25</td>
<td>Baseline</td>
<td>45.77%</td>
<td>32.08%</td>
<td>21.36%</td>
<td>13.64%</td>
</tr>
<tr>
<td></td>
<td>+ Fast</td>
<td>45.28%</td>
<td>31.67%</td>
<td>20.69%</td>
<td>13.26%</td>
</tr>
<tr>
<td></td>
<td>+ CRC</td>
<td>51.59%</td>
<td>39.27%</td>
<td>27.94%</td>
<td>19.06%</td>
</tr>
<tr>
<td>30</td>
<td>Baseline</td>
<td>46.39%</td>
<td>33.08%</td>
<td>22.02%</td>
<td>13.77%</td>
</tr>
<tr>
<td></td>
<td>+ Fast</td>
<td>45.86%</td>
<td>32.54%</td>
<td>21.18%</td>
<td>12.77%</td>
</tr>
<tr>
<td></td>
<td>+ CRC</td>
<td>50.97%</td>
<td>38.77%</td>
<td>27.73%</td>
<td>19.28%</td>
</tr>
<tr>
<td>35</td>
<td>Baseline</td>
<td>43.42%</td>
<td>30.36%</td>
<td>19.71%</td>
<td>12.66%</td>
</tr>
<tr>
<td></td>
<td>+ Fast</td>
<td>42.78%</td>
<td>29.88%</td>
<td>19.73%</td>
<td>12.52%</td>
</tr>
<tr>
<td></td>
<td>+ CRC</td>
<td>51.42%</td>
<td>39.01%</td>
<td>28.94%</td>
<td>20.29%</td>
</tr>
<tr>
<td>40</td>
<td>Baseline</td>
<td>41.72%</td>
<td>28.53%</td>
<td>18.37%</td>
<td>11.49%</td>
</tr>
<tr>
<td></td>
<td>+ Fast</td>
<td>42.07%</td>
<td>28.51%</td>
<td>18.86%</td>
<td>11.89%</td>
</tr>
<tr>
<td></td>
<td>+ CRC</td>
<td>50.11%</td>
<td>38.69%</td>
<td>28.45%</td>
<td>20.05%</td>
</tr>
</tbody>
</table>

### 7.1 Results using Faster gradient computation

We show the reduction in training time per epoch (in seconds) on CIFAR-10 in Table 1 and CIFAR-100 in Appendix Table 8. In both Tables, we observe that for deeper networks (≥ 25 layers), the reduction in time per epoch is ≈ 30%. The corresponding standard and provable robust accuracy numbers are given in Tables 3 (CIFAR-10) and 4 (CIFAR-100) in the row "+ Fast". From the columns "Increase (Standard)" and "Increase (36/255)", we observe that the performance is similar to the baseline across all network architectures. For deeper networks: LipConvnet-35, 40, we observe an increase in performance for CIFAR-10 and small decrease (< 0.64%) for CIFAR-100.

### 7.2 Results using projection pooling and CRC-Lip

We observe that using CRC and projection pooling (row "+ CRC") leads to significant improvements in performance across all LipConvnet architectures. On CIFAR-10, we observe significant improvements in both the standard (≥ 1.63%, column "Increase (Standard)") and provable robust accuracy (≥ 3.14%, column "Increase (36/255)") across all architectures. On CIFAR-100, we also observe significant improvements in the standard (≥ 2.49%) and provable robust accuracy (≥ 2.27%). For deeper networks (LipConvnet-35, 40), on CIFAR-10, we observe even more significant gains in the standard (≥ 5.84%) and provable robust accuracy (≥ 8.76%). Similarly on CIFAR-100, we observe gains of ≥ 8.00% (standard) and ≥ 8.65% (provable robust).

Our results establish a new state-of-the-art for both the standard and provable robust accuracy across all attack radii. In Table 3 (CIFAR-10), the best "Baseline" standard and provable robust accuracy values (at 36/255, 72/255, 108/255) are 77.78%, 63.31%, and 46.42%, respectively. The best "+ CRC" values are 79.57%, 67.13%, 53.17%, and 38.60%. This results in improvements of +1.79% and +3.82%, +6.75%, and +7.07% respectively. Similarly, in Table 4 (CIFAR-100), the best "Baseline" standard and provable robust values are 48.06% and 34.52%, respectively. The best "+ CRC" values are 51.84% (+3.78%) and 39.27% (+4.75%), respectively.
Acknowledgements

This project was supported in part by NSF CAREER AWARD 1942230, a grant from NIST 60NANB20D134, HR000119S0026 (GARD), ONR YIP award N00014-22-1-2271, Army Grant No. W911NF2120076 and the NSF award CCF2212458.

References


Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] in Section 7 we discuss how CRC-Lip might lead to slower certification.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A] because it is expensive to run.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]