Rethinking Lipschitz Neural Networks for Certified L-infinity Robustness

Anonymous Author(s)
Affiliation
Address
email

Abstract

Designing neural networks with bounded Lipschitz constant is a promising way to obtain certifiably robust classifiers against adversarial examples. However, the relevant progress is rather limited for the important ℓ_∞ perturbation setting, and a principled understanding of how to design better ℓ_∞ Lipschitz networks is still lacking. In this paper, we bridge the gap by studying certified ℓ_∞ robustness from a novel perspective of representing Boolean functions. We derive two fundamental impossibility results that hold for any standard Lipschitz network: one for robust classification on discrete datasets, and the other for (continuous) Lipschitz function approximation. These results identify that norm-bounded affine transformation intrinsically loses expressive power even in the two-dimensional case, and provide insights into how recently proposed Lipschitz networks bypass these impossibilities. Finally, we propose a unified Lipschitz neural network that generalizes prior works, and provide a practical version that can be efficiently trained (making certified robust training free). Extensive experiments show that our approach is scalable, efficient, and consistently yields better certified robustness across multiple datasets and perturbation radii than prior Lipschitz networks.

1 Introduction

Modern neural networks, despite their great success in various applications [21, 13], typically suffer from a severe drawback of lacking adversarial robustness. In classification tasks, given an image $\boldsymbol{x}$ correctly classified by a neural network, there often exists a small adversarial perturbation $\boldsymbol{\delta}$, making the perturbed image $\boldsymbol{x} + \boldsymbol{\delta}$ look indistinguishable to humans but fool the network to predict a wrong class with high confidence [53, 4].

It is well-known that the adversarial robustness of a neural network is closely related to its Lipschitz continuity [8, 57] (see Section 3.1). Accordingly, training neural networks with bounded Lipschitz constant has been considered a promising way to address the problem. A variety of works studied Lipschitz architectures for the ordinary Euclidean norm [57, 56, 34, 50], and recent works even achieved state-of-the-art deterministic certified ℓ_2 robustness [51, 38]. However, when it comes to the more critical (realistic) ℓ_∞ perturbation setting, the progress seems to be rather limited. For example, standard Lipschitz ReLU networks have been shown to perform poorly in terms of ℓ_∞ robustness [57, 23, 2]. While several works proposed more advanced Lipschitz networks [2, 10], the achieved results are still far from satisfactory even on the MNIST dataset. Until recently, Zhang et al. [69, 70] designed a quite unusual 1-Lipschitz network based on a heuristic choice of ℓ_∞-distance function, which surprisingly established state-of-the-art certified ℓ_∞ robustness on multiple datasets over prior works. Yet, it remains unclear why previous Lipschitz networks typically failed, and what is the essential reason behind the success of this particular ℓ_∞-distance network structure.

In this work, we provide an answer to the above question by studying Lipschitz neural networks (w.r.t. ℓ_∞-norm) from a novel perspective of representing discrete Boolean functions. We figured out

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

a fundamental limitation of standard Lipschitz networks in representing a class of logical operations called *symmetric Boolean functions*, which comprises the basic logical AND/OR as special cases. We prove that for any non-constant symmetric Boolean functions of d variables, there exists a finite dataset of size $\mathcal{O}(d)$ such that the certified ℓ_∞ robust radius must vanish as $\mathcal{O}(1/d)$ for any classifier induced by a standard Lipschitz network. Remarkably, since logical AND/OR operations correspond to perhaps the most basic classifiers, our result indicates an intrinsic difficulty of standard Lipschitz networks in fitting real-world datasets with guaranteed certified ℓ_∞ robustness.

Our analysis can be readily extended into the Lipschitz function approximation setting. Noticing the relationship between symmetric Boolean functions and sorting, we then prove that any d-dimensional *order statistic* function (including the max/min) on a compact domain cannot be approximated by any standard Lipschitz network with error $\mathcal{O}(1 - 1/d)$, regardless of the network size. Our impossibility result is significant in that (i) it applies to all Lipschitz activation functions (thus extending prior works [2, 23]), and (ii) a quantitative lower bound of *approximation error* is established.

Equipped by the above impossibility results, we proceed to examine two advanced Lipschitz architectures: the GroupSort network [2] and the recently proposed ℓ_∞-distance net [69, 70]. We find that besides the linear operation, both networks incorporate other Lipschitz aggregation operations into the neuron design, especially the order statistic operations, thus shedding light on how they work. However, for the MaxMin network [2] — a computationally efficient version of the GroupSort network implemented in practice, approximating order statistic functions is still impossible if the network depth is no greater than $\lceil \log_2 d \rceil$. In contrast, we prove that a *two-layer* ℓ_∞-distance net can *exactly* represent any order statistic function on a compact domain, or even *all Boolean functions*. This strongly justifies the empirical success of ℓ_∞-distance net over GroupSort (MaxMin) networks.

Our theoretical insights can also guide in designing better Lipschitz network architectures. Inspired by the importance of order statistics, we propose a general form of Lipschitz network, called SortNet, that extends both GroupSort and ℓ_∞ distance networks and incorporates them into a unified framework. Yet, the full-sort operation is computationally expensive and leads to optimization difficulties (as with the GroupSort network). We further propose a specialized SortNet that can be efficiently trained, by assigning each weight vector $\boldsymbol{w}$ using *geometric series*, i.e. w_i proportional to ρ^i for some $0 \leq \rho < 1$. This leads to a restricted version of SortNet but still covers ℓ_∞-distance net as a special case. For this particular SortNet, we skillfully derive a stochastic estimation that gives an unbiased approximation of the neuron output without performing sorting operations explicitly. This eventually yields an efficient training strategy with similar cost as training standard networks, thus making certified robust training free. Extensive experiments demonstrate that the proposed SortNet is scalable, efficient, and consistently achieves better certified robustness than prior Lipschitz networks across multiple datasets and perturbation radii. In particular, our approach even scales on a variant of ImageNet, and surpasses the best-known result [65] with a 22-fold decrease in training time thanks to our "free" certified training approach.

The main contributions of this paper can be summarized as follows:

- We develop a systematic study for the representation power of Lipschitz neural networks using the tools of Boolean algebra. We prove the impossibility results of standard Lipschitz networks in two settings: a) certified ℓ_∞ robustness on discrete datasets (Section 3.2); b) Lipschitz function approximation (Section 3.3).
- We provide insights into how recently proposed networks can bypass the impossibility results. In particular, we show that a *two-layer* ℓ_∞-distance net can precisely represent any Boolean functions on a compact domain, while *shallow* GroupSort networks cannot (Section 3.4).
- We propose SortNet, a Lipschitz network that generalizes GroupSort and ℓ_∞-distance net. For a special type of SortNet, we derive a stochastic training approach that bypasses the difficulties in calculating sorting operations explicitly and makes certified training free (Section 4).
- Extensive experiments demonstrate that SortNet exhibits better certified robustness on several benchmark datasets over baseline methods with high training efficiency (Section 5).

2 Related Work

Extensive studies have been devoted to developing neural networks with certified robustness guarantees. Existing approaches can be mainly divided into the following three categories.

Certified defenses for standard networks. A variety of works focus on establishing certified robustness for standard neural networks. However, exactly calculating the certified radius of a standard ReLU network is known to be NP-hard [27]. Researchers thus developed a class of relaxation-based approaches that provide a tight lower bound estimate of the certified robustness efficiently. These approaches typically use convex relaxation to calculate a bound of the neuron outputs under input perturbations layer by layer [61, 62, 60, 49, 40, 18, 58, 72]. See also [3, 15, 16, 44, 45, 64, 11, 33, 59, 22] for more advanced approaches. However, most of these works suffer from high computational costs and are hard to scale up to large datasets. Currently, the only scalable convex relaxation approach is based on *interval bound propagation* (IBP) [40, 20, 71, 65, 48], but the produced bound is known to be loose [47], and a recent study showed that IBP cannot achieve enough certified robustness on simple datasets for any standard ReLU network [39].

Certified defenses using Lipschitz networks. On the other hand, Lipschitz networks inherently imply certified robustness, resulting in a *much simpler* certification process based on the output margin (see Proposition 3.1). Yet, most prior works can only handle the ℓ_2-norm Lipschitz situation by leveraging specific mathematical properties such as the spectral norm [8, 67, 19, 57, 17, 43, 2, 34, 38] or orthogonality of weight matrices [37, 56, 50, 51]. For the ℓ_∞-norm, standard Lipschitz networks were shown to give only a vanishingly small certified radius [57]. Huster et al. [23] found that standard Lipschitz ReLU networks cannot represent certain simple functions such as the absolute value, which inspired the first expressive Lipschitz architecture called the GroupSort network [2]. Since then, GroupSort has been extensively investigated [10, 54], but its performance is still much worse than the above relaxation-based approaches even on MNIST. Recently, Zhang et al. [69, 70] first proposed a practical 1-Lipschitz architecture w.r.t. ℓ_∞-norm based on a special neuron called the ℓ_∞-distance neuron, which can scale to TinyImageNet with state-of-the-art certified robustness over relaxation-based approaches. However, despite its practical success, it is rather puzzling how such a simple architecture can work while prior approaches all failed. Answering this question may require an in-depth re-examination of Lipschitz networks (w.r.t. ℓ_∞-norm), which is the focus of this paper.

Certified defenses via randomized smoothing. As a rather different and parallel research line, randomized smoothing typically provides *probabilistic* certified ℓ_2 robustness guarantees. Due to the wide applicability, randomized smoothing has been scaled up to ImageNet and achieves state-of-the-art certified accuracy for ℓ_2 perturbations [32, 36, 9, 46, 68, 26]. However, certifying robustness with high probability requires sampling a large number of noisy inputs (e.g., 10^5) for a single image, leading to a high computational cost at inference. Moreover, theoretical results pointed out that it cannot achieve non-trivial certified ℓ_∞ robustness if the perturbation radius is larger than $\Omega(d^{-1/2})$ where d is the input dimension [66, 5, 29, 63].

3 The Representation Power of Lipschitz Neural Networks

3.1 Preliminaries

Notations. We use boldface letters to denote vectors (e.g., $\boldsymbol{x}$) or vector functions (e.g., $\boldsymbol{f}$), and use x_i (or f_i) to denote its i-th element. For a unary function σ, $\sigma(\boldsymbol{x})$ applies $\sigma(\cdot)$ element-wise on vector $\boldsymbol{x}$. The ℓ_p-norm ($p \geq 1$) and ℓ_∞-norm of a vector $\boldsymbol{x}$ are defined as $\|\boldsymbol{x}\|_p = (\sum_i |x_i|^p)^{1/p}$ and $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$, respectively. The matrix ∞-norm is defined as $\|\mathbf{W}\|_\infty = \max_i \|\mathbf{W}_{i,:}\|_1$ where $\mathbf{W}_{i,:}$ is the i-th row of the matrix $\mathbf{W}$. The k-th largest element of a vector $\boldsymbol{x}$ is denoted as $x_{(k)}$. We use $[n]$ to denote the set $\{1, \cdots, n\}$, and use $\boldsymbol{e}_i$ to denote the unit vector with the i-th element being one. We adopt the big O notations by using $\mathcal{O}(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$ to hide universal constants.

Lipschitzness. A mapping $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is said to be L-Lipschitz continuous w.r.t. norm $\|\cdot\|$ if for any pair of inputs $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$,

$$\|\boldsymbol{f}(\boldsymbol{x}_1) - \boldsymbol{f}(\boldsymbol{x}_2)\| \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|. \tag{1}$$

If the mapping $\boldsymbol{f}$ represented by a neural network has a small Lipschitz constant L, then (1) implies that the change of network output can be strictly controlled under input perturbations, resulting in *certified* robustness guarantees as shown in the following proposition.

Proposition 3.1. (Certified robustness of Lipschitz networks) *For a neural network $\boldsymbol{f}$ with Lipschitz constant L under ℓ_p-norm $\|\cdot\|_p$, define the resulting classifier g as $g(\boldsymbol{x}) := \arg\max_k f_k(\boldsymbol{x})$ for an input $\boldsymbol{x}$. Then g is provably robust under perturbations $\|\boldsymbol{\delta}\|_p < \frac{c}{L}\operatorname{margin}(\boldsymbol{f}(\boldsymbol{x}))$, i.e.*

$$g(\boldsymbol{x} + \boldsymbol{\delta}) = g(\boldsymbol{x}) \quad \textit{for all } \boldsymbol{\delta} \textit{ with } \|\boldsymbol{\delta}\|_p < \frac{c}{L}\operatorname{margin}(\boldsymbol{f}(\boldsymbol{x})). \tag{2}$$

Here $c = \sqrt[p]{2}/2$ is a constant depending only on the norm $\|\cdot\|_p$, which is $1/2$ for the ℓ_∞-norm, and $\mathrm{margin}(\boldsymbol{f}(\boldsymbol{x}))$ is the margin between the largest and second largest output logits.

Li et al. [37] proved the case of $p = 2$, and we give a proof for general p in Appendix B.1. It can be seen that the robust radius is inversely proportional to the Lipschitz constant L. Note that the bound on the certified radius (2) is *tight* and cannot be improved without additional assumptions on $\boldsymbol{f}$.

Standard Lipschitz networks. Throughout this paper, we refer to standard neural networks as neural networks formed by affine layers (e.g., fully-connected or convolutional layers) and element-wise activation functions. Based on the Lipschitz property of composite functions, most prior works enforce the 1-Lipschitzness of a multi-layer neural network by constraining each layer to be a 1-Lipschitz mapping. For the ℓ_∞-norm, it is further equivalent to constraining the weight matrices to have bounded ∞-norm, plus using Lipschitz activation functions [2], which can be formalized as

$$\boldsymbol{x}^{(l)} = \sigma^{(l)}(\mathbf{W}^{(l)}\boldsymbol{x}^{(l-1)} + \boldsymbol{b}^{(l)}) \quad \text{s.t. } \|\mathbf{W}^{(l)}\|_\infty \le 1 \text{ and } \sigma^{(l)} \text{ being 1-Lipschitz, } l \in [L]. \tag{3}$$

Usually $\sigma^{(L)}(x) = x$ is the identity function. The network takes $\boldsymbol{x}^{(0)} := \boldsymbol{x}$ as the input and outputs $\boldsymbol{x}^{(L)}$. For K-class classification problems, $\boldsymbol{x}^{(L)} \in \mathbb{R}^K$ and the network predicts the class $g(\boldsymbol{x}) := \arg\max_{k\in[K]} x_k^{(L)}$. We refer to the resulting network as a standard Lipschitz network.

3.2 Certified robustness on discrete Boolean datasets

In this section, we will construct a class of counterexamples for which certified ℓ_∞ robustness can be arbitrarily poor using standard Lipschitz networks. We focus on the *Boolean dataset*, a discrete dataset where both inputs and labels are Boolean-valued and the relationship between inputs and their labels $(\boldsymbol{x}^{(i)}, y^{(i)}) \in \{0,1\}^d \times \{0,1\}$ can be described using a Boolean function $g^{\mathrm{B}} : \{0,1\}^d \to \{0,1\}$. The key motivation lies in the finding that Boolean vectors correspond to the vertices of a d-dimensional hypercube, and thus are geometrically related to the ℓ_∞-distance metric. Moreover, the ℓ_∞-distance between any two different data points in a Boolean dataset is always 1, which means that the dataset is *well-separated*. This yields the following proposition, stating that it is always possible to achieve *optimal* certified ℓ_∞ robustness on Boolean datasets by using Lipschitz classifiers.

Proposition 3.2. *For any Boolean dataset $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^n$, there exists a classifier $\hat{g} : \mathbb{R}^d \to \{0,1\}$ induced by a 1-Lipschitz mapping $\hat{\boldsymbol{f}} : \mathbb{R}^d \to \mathbb{R}^2$, such that $\hat{g}$ can fit the whole dataset with $\mathrm{margin}(\hat{\boldsymbol{f}}(\boldsymbol{x}^{(i)})) = 1\ \forall i \in [n]$, thus achieving an optimal certified ℓ_∞ robust radius of $1/2$.*

The key observation in the proof (Appendix B.2) is that one can construct a so-called *nearest neighbor classifier* that achieves a large margin on the whole dataset and is *Lipschitz*. Based on Proposition 3.2, it is natural to ask whether standard Lipschitz networks of the form (3) can perform well on Boolean datasets. Unfortunately, we show it is not the case, even on a class of simple datasets constructed using *symmetric* Boolean functions.

Definition 3.3. A Boolean function $g^{\mathrm{B}} : \{0,1\}^d \to \{0,1\}$ is symmetric if it is invariant under input permutation, i.e. $g^{\mathrm{B}}(x_1, \cdots, x_d) = g^{\mathrm{B}}(x_{\pi(1)}, \cdots, x_{\pi(d)})$ for any $\boldsymbol{x} \in \{0,1\}^d$ and $\pi \in S_d$.

Remark 3.4. Two of the most basic operations in Boolean algebra are the logical AND/OR, both of which belong to the class of symmetric Boolean functions. Other important examples include the exclusive-or (XOR, also called the parity function), NAND, NOR, and the majority function (or generally, the threshold functions). See Appendix A.2 for a detailed description of these examples.

Theorem 3.5. *For any non-constant symmetric Boolean function $g^B : \{0,1\}^d \to \{0,1\}$, there exists a Boolean dataset with labels $y^{(i)} = g^B(\boldsymbol{x}^{(i)})$, such that no standard Lipschitz network can achieve a certified ℓ_∞ robust radius larger than $1/2d$ on the dataset.*

Implications. Theorem 3.5 shows that the certified radius of standard Lipschitz networks must vanish as dimension d grows, which is in stark contrast to the constant radius given by Proposition 3.2. Also, note that the logical AND/OR functions are perhaps the most basic classifiers (which simply make predictions based on the existence of binary features). It is thus not surprising to see standard Lipschitz networks perform poorly even on MNIST (a dataset with almost binary inputs) [57].

The proof of Theorem 3.5 is deferred to Appendix B.3, where we also prove that the bound of $1/2d$ is *tight* (Theorem B.8). Moreover, we discuss the sample complexity by proving that a dataset of size $\mathcal{O}(d)$ already suffices to give $\mathcal{O}(1/d)$ certified radius (Corollary B.10). To the best of our knowledge, Theorem 3.5 is the first impossibility result directly targeting the certified robustness of standard Lipschitz networks with an *explicit* upper bound on the certified radius. Next, we will extend our analysis to the function approximation setting and make discussions with literature results [23, 2].

3.3 Lipschitz function approximation

Classic approximation theory has shown that standard neural networks are universal function approximators [12, 35], in that they can approximate any continuous function on a compact domain arbitrarily well. For 1-Lipschitz neural networks, an analogous question is whether they can approximate all Lipschitz functions accordingly. Unfortunately, the results in Section 3.2 already imply a negative answer. Indeed, by combining Proposition 3.2 and Theorem 3.5, $\hat{\boldsymbol{f}}$ is clearly a 1-Lipschitz function that cannot be approximated by any standard Lipschitz networks.

To gain further insights into the structure of unrepresentable 1-Lipschitz functions, let us consider the *continuousization* of discrete Boolean functions. For the symmetric case, one needs to find a class of 1-Lipschitz continuous functions that are also invariant under permutations. It can be found that a simple class of candidates is the *order statistics*, i.e. the k-th largest elements of a vector. The k-th order statistic $x_{(k)}$ is indeed 1-Lipschitz and is precisely the continuousization of the k-threshold Boolean function defined as $g^{\mathrm{B},k}(\boldsymbol{x}) = \mathbb{I}(\sum_i x_i \geq k)$. In particular, $x_{(1)} = \max_i x_i$ and $x_{(d)} = \min_i x_i$ corresponds to the continuousization of logical OR and logical AND functions, respectively. We thus arrived at the following theorem (proven in Appendix B.4):

Theorem 3.6. *Any standard Lipschitz network $f : \mathbb{R}^d \to \mathbb{R}$ cannot approximate the simple 1-Lipschitz functions $\boldsymbol{x} \to x_{(k)}$ for some $k \in [d]$ on a bounded domain $\mathcal{K} \in [0,1]^d$ if $d \geq 2$. Moreover, there must exist a point $\widehat{\boldsymbol{x}} \in \mathcal{K}$, such that*

$$|f(\widehat{\boldsymbol{x}}) - \widehat{x}_{(k)}| \geq \frac{1}{2} - \frac{1}{2d}. \tag{4}$$

Remark 3.7. The above theorem indicates that order statistics cannot be *uniformly* approximated using any standard Lipschitz network regardless of the network size. Note that the trivial constant function $\tilde{f}(\boldsymbol{x}) = 1/2$ already achieves an approximation error of $1/2$ uniformly on $\mathcal{K}$, implying that standard Lipschitz networks can hardly improve upon trivial solutions. Our result can also be easily generalized to weaker forms of non-uniform approximation, e.g., using the ℓ_p-norm as distance metrics [42], by proving that there exists a hypercube $\mathcal{B}_\infty(\widehat{\boldsymbol{x}})$ centered at $\widehat{\boldsymbol{x}}$ with length $\Theta(1)$, such that $|f(\tilde{\boldsymbol{x}}) - \tilde{x}_{(k)}| \geq \Theta(1)$ holds for all $\tilde{x} \in \mathcal{B}_\infty(\widehat{\boldsymbol{x}})$ when $d \geq 2$. See Corollary B.13 for details.

Discussion with prior works [2, 23]. The work of Anil et al. also gave negative results on the expressive power of standard Lipschitz networks[1] (Theorem 1 in [2]). They proved a different (weaker) version of Theorem 3.6, showing that if the activation function σ is *monotonic*, the network cannot *precisely* represent non-linear 1-Lipschitz functions whose gradient norm is 1 almost everywhere (e.g., the absolute value function proven before by [23]). The intuition is that any monotonic non-linear 1-Lipschitz activation (e.g., ReLU) must have regions with slopes less than 1, leading to *gradient attenuation* during backpropagation. The authors thus attributed the reason to the activation function which is not *gradient-norm preserving* (GNP). However, such an explanation is still not satisfactory, as GNP can be easily achieved using a non-monotonic activation function (e.g., $\sigma(x) = |x|$), but the resulting network still lacks expressive power due to Theorem 3.6, which cannot be explained. Instead, we draw a rather different conclusion, arguing that the lack of expressiveness is due to the inability of *weight-constrained* affine transformations to perform basic Boolean operations (even for $d = 2$). A further justification is given in Section 3.4. Compared with [2], the form of Theorem 3.6 is more fundamental, in the sense that it does not make assumptions on the activation function, and it gives an *explicit error bound* on the approximation that is arbitrarily close to the plain fit $f(\boldsymbol{x}) = 1/2$.

3.4 Investigating more advanced Lipschitz networks

Seeing the above impossibility results, we then examine two representative works of (non-standard) Lipschitz networks in literature: the GroupSort network [2] and the recently proposed ℓ_∞-distance net [69, 70]. Notably, both networks are Lipschitz-universal function approximators and thus fully expressive. The GroupSort network makes minimum changes to standard Lipschitz networks (3), by replacing element-wise activation σ with GroupSort layers. GroupSort partitions the input vector into groups, sorts the sub-vector of each group in descending order, and finally concatenates the resulting sub-vectors. Since sorting is computationally expensive, the authors considered a practical version of GroupSort with a group size of 2, called MaxMin, which simply calculates the maximum and minimum pair by pair [6]. ℓ_∞-distance net, on the other hand, is fundamentally different from

[1] The result is described w.r.t. ℓ_2-norm, but with some effort, it can be extended to the ℓ_∞-norm case.

standard Lipschitz networks. Each neuron in an ℓ_∞-distance net is designed based on the ℓ_∞-distance function $y = \|\boldsymbol{x} - \boldsymbol{w}\|_\infty + b$ (with parameters $\{\boldsymbol{w}, b\}$). Despite the somewhat unusual structure, ℓ_∞-distance net has been shown to substantially outperform GroupSort (MaxMin) in terms of certified ℓ_∞ robustness according to [69], a puzzling thing to be understood.

We provide a possible explanation for this. We find that both networks incorporate order statistics into the neuron design, either explicitly (GroupSort) or implicitly (ℓ_∞-distance net), thus bypassing the impossibility result in Theorem 3.6. Indeed, an ℓ_∞-distance neuron can exactly represent the max function on a bounded domain, by assigning the weight $\boldsymbol{w} = -c\mathbf{1}$ and the bias $b = -c$ with a sufficiently large constant c:

$$y = \|\boldsymbol{x} - \boldsymbol{w}\|_\infty + b = \max_i |x_i - (-c)| - c = \max_i x_i \quad \text{if } c \geq \max_i -x_i, \tag{5}$$

and thus can represent the logical OR operation. In general, we have the following theorem:

Theorem 3.8. *A two-layer ℓ_∞-distance net can exactly represent the following functions:* (i) *any discrete Boolean function;* (ii) *any continuous order-statistic function on a compact domain.*

We give a proof in Appendix B.5. Our proof uses the fundamental result in Boolean algebra that any Boolean function can be written in its disjunctive normal form (DNF, see Appendix A.1), which can be further reduced to using only the logical OR operation of literals and thus be realized by a two-layer ℓ_∞-distance net. To represent order statistics, we formulate them as nested max-min functions, which can also be realized by a two-layer ℓ_∞-distance net. Therefore, the construction in our proof provides strong justification for the mechanism behind the success of ℓ_∞-distance nets, since each ℓ_∞-distance neuron can be regarded as a basic "logical gate".

For GroupSort networks with a group size $G \geq d$, a similar result holds. However, it is not the case for practically used MaxMin networks ($G = 2$), where we have the following impossibility result:

Theorem 3.9. *An M-layer MaxMin network $f : \mathbb{R}^d \to \mathbb{R}$ cannot approximate any k-th order statistic function on a bounded domain $\mathcal{K} \in [0, 1]^d$ if $M \leq \lceil \log_2 d \rceil$ (no matter how wide the network is). Moreover, there exists a point $\widehat{\boldsymbol{x}} \in \mathcal{K}$, such that*

$$|f(\widehat{\boldsymbol{x}}) - x_{(k)}| \geq \frac{1}{2} - \frac{2^{M-2}}{d} \geq \frac{1}{4} \quad \textit{if } M \leq \lfloor \log_2 d \rfloor. \tag{6}$$

In Appendix B.6, we also discuss the tightness of Theorem 3.9 by proving that a depth of $\mathcal{O}(\log_2 d)$ is sufficient to represent any order statistic function (Theorem B.19). However, for general discrete Boolean functions, we can only prove that an upper bound on the minimum depth is $\mathcal{O}(d)$. Based on our knowledge of Boolean circuit theory, we conjecture that the lower bound of depth is also $\Omega(d)$ and we will leave the study as future work. Note that both results are in stark contrast to Theorem 3.8, which says that a constant depth is sufficient for ℓ_∞-distance nets.

Unfortunately, training deep MinMax networks is known to be challenging due to severe optimization difficulties [10]. Consequently, prior works only use a shallow MaxMin network with 4 layers [2, 10], which severely lacks expressive power. One possible solution is to increase the group size, and several works explored this aspect using toy examples and observed significant benefits empirically [10, 54]. However, a large group size involves computationally expensive sorting operations and makes the network very hard to train, limiting its value in practice.

4 A Unified Framework of Lipschitz Neural Networks

The above theoretical results have justified order statistics as a crucial component in representing a class of Boolean functions, shedding light on how GroupSort and ℓ_∞-distance net work. Based on these insights, in this section, we will propose a unified framework of Lipschitz networks that take the respective advantage of prior Lipschitz architectures, and then give a practical (specialized) version that enables efficient training.

Consider a general Lipschitz network constructed using the following three basic types of 1-Lipschitz operations: (i) norm-bounded affine transformations, e.g. $y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}$ ($\|\boldsymbol{w}\|_1 \leq 1$) and $y = x + b$; (ii) 1-Lipschitz unary activation functions σ; (iii) the order statistics. The first two types are extensively used in standard Lipschitz networks, while the last type is motivated by Section 3 and is of crucial importance. We propose the following network which naturally combines the above components:

Definition 4.1. (SortNet) Define an M-layer fully-connected SortNet $\boldsymbol{f}$ as follows. The network takes $\boldsymbol{x} = \boldsymbol{x}^{(0)}$ as input, and the k-th unit in the l-th hidden layer $x_k^{(l)}$ is computed by

$$x_k^{(l)} = (\boldsymbol{w}^{(l,k)})^{\mathrm{T}} \operatorname{sort}(\sigma(\boldsymbol{x}^{(l-1)} + \boldsymbol{b}^{(l,k)})), \quad \text{s.t. } \|\boldsymbol{w}^{(l,k)}\|_1 \leq 1, \quad l \in [M], k \in [d_l] \tag{7}$$

where d_l is the hidden size of the l-th layer, and $\mathrm{sort}(\boldsymbol{x}) := (x_{(1)}, \cdots, x_{(d)})^{\mathrm{T}}$ calculates all order statistics of vector $\boldsymbol{x}$. The network outputs $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{x}^{(M)} + \boldsymbol{b}^{\text{out}}$. Here $\{\boldsymbol{w}^{(l,k)}\}$, $\{\boldsymbol{b}^{(l,k)}\}$ and $\boldsymbol{b}^{\text{out}}$ are network parameters.

It is easy to see that SortNet is 1-Lipschitz w.r.t. ℓ_∞-norm. We now show that SortNet is a general architecture that extends both GroupSort and ℓ_∞-distance networks into a unified framework.

Proposition 4.2. *Any GroupSort network with an arbitrary group size on a compact input domain can be represented by a SortNet with the same topological structure using activation $\sigma(x) = x$.*

Proposition 4.3. *Any ℓ_∞-distance net can be represented by a SortNet with the same topological structure by fixing the weights $\boldsymbol{w}^{(l,k)} = \boldsymbol{e}_1$ and using the absolute-value activation $\sigma(x) = |x|$.*

See Appendix C for a proof. As a result, SortNet can exploit the respective advantage of these Lipschitz networks. Compared with GroupSort, SortNet can freely use activation functions such as the absolute value, thus easily addressing the problem claimed in [23, 2]. Moreover, unlike GroupSort, the bias vector $\boldsymbol{b}^{(l,k)}$ in SortNet (7) can be assigned diversely for different neurons in the same layer. In this way, one can control the sorting behavior of each neuron individually by varying the bias value without disturbing the output of other neurons, which is very beneficial (see Appendix C.1 for details). Compared with ℓ_∞-distance net, SortNet adds linear transformation and incorporates all order statistics (rather than only the maximum), thus can represent certain functions more effectively.

A practical version of SortNet. As with GroupSort networks, we also design a practical (specialized) version of SortNet which enjoys efficient training. But different from the MaxMin network that reduces the group size, we keep the full-dimensional order statistics as they are crucial for the expressive power (Section 3.4). The key observation is that in (7), the only required computation is the linear combination of order statistics (i.e. $\boldsymbol{w}^{\mathrm{T}}\,\mathrm{sort}(\cdot)$), rather than the entire sorting results (i.e. $\mathrm{sort}(\cdot)$). We find that for certain carefully designed choices of the weight vector $\boldsymbol{w}$, there exist efficient approximation algorithms that can give a good estimation of $\boldsymbol{w}^{\mathrm{T}}\,\mathrm{sort}(\cdot)$. In particular, we propose an assignment of the weight vector that follows geometric series, i.e. w_i proportional to ρ^i, in which case we have the following result:

Proposition 4.4. *Let $\boldsymbol{w} \in \mathbb{R}^d$ be a vector satisfying $w_i = (1-\rho)\rho^{i-1}, i \in [d]$ for some $0 \le \rho < 1$. Then for any vector $\boldsymbol{x} \in \mathbb{R}^d_+$ with non-negative elements,*

$$\boldsymbol{w}^{\mathrm{T}}\,\mathrm{sort}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{s}\sim\mathrm{Ber}(1\text{-}\rho)}[\max_i s_i x_i]. \tag{8}$$

Here $\boldsymbol{s}$ is a random vector following independent Bernoulli distribution with probability $1-\rho$ being 1 and ρ being 0.

Proof. Without loss of generality, assume $x_1, \cdots, x_d$ are different from each other. Denote $j_1, \cdots, j_d$ as the sorting indices such that $\mathrm{sort}(\boldsymbol{x}) = (x_{j_1}, \cdots, x_{j_d})$. Then

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{s}\sim\mathrm{Ber}(\rho)}[\max_i s_i x_i] &= \textstyle\sum_{k\in[d]} \Pr_{\boldsymbol{s}\sim\mathrm{Ber}(\rho)}\left[\max_i s_i x_i = x_{j_k}\right] x_{j_k} \\
&= \textstyle\sum_{k\in[d]} \Pr_{\boldsymbol{s}\sim\mathrm{Ber}(\rho)}\left[s_{j_k} = 1 \text{ and } s_{j_i} = 0\ \forall 1 \le i < k\right] x_{j_k} \\
&= \textstyle\sum_{k\in[d]} (1-\rho)\rho^{k-1} \cdot x_{(k)} = \boldsymbol{w}^{\mathrm{T}}\,\mathrm{sort}(\boldsymbol{x}).
\end{aligned}$$

□

It is easy to check that the weight $\boldsymbol{w}$ in the above proposition satisfies $\|\boldsymbol{w}\|_1 \le 1$, which guarantees the Lipschitzness. The non-negative condition on $\boldsymbol{x}$ in Proposition 4.4 holds when using a suitable activation function in neuron (7), such as the absolute value function.

Proposition 4.4 suggests that one can use $\max_i s_i x_i$ to give an *unbiased* estimation of $\boldsymbol{w}^{\mathrm{T}}\,\mathrm{sort}(\boldsymbol{x})$. In this way, the expensive sorting operation is avoided and replaced by a max operation, thus significantly reducing the computational cost in training. Note that $\boldsymbol{s}$ is a random Bernoulli vector, so the above calculation is similar to applying a mask on the input of each neuron, like dropout [52]. It means that the introduced stochasticity may further prevent overfitting and benefit generalization performance.

Regarding the value of ρ. When $\rho = 0$, only the maximum value is taken into the computation and the resulting network can recover the ℓ_∞-distance net by choosing the activation function $\sigma(x) = |x|$. This means the specialized SortNet still extends ℓ_∞-distance net and thus has sufficient expressive power. When $\rho > 0$, all order statistics become utilized. A simple way of selecting ρ is to regard it as a hyper-parameter and set its value by cross-validation, which is adopted in our experiments. One can also consider treating ρ as learnable parameters for each neuron that participate in the optimization process, but this involves calculating the gradient of ρ which may be complicated due to the stochastic sampling procedure (8). We will leave the study as future work.

5 Experiments

In this section, we perform extensive empirical evaluations of the proposed SortNet architecture as well as various prior works in the certified ℓ_∞ robustness area. To show the scalability of different approaches, we consider a variety of benchmark datasets, including MNIST [31], CIFAR-10 [28], TinyImageNet [30], and ImageNet (64×64) [7]. Due to space limitations, a complete training recipe is given in Appendix E. Our code and trained models will be released.

SortNet model configuration. Since SortNet generalizes the ℓ_∞-distance net, we simply follow the same model configurations as [69] and consider two types of models. The first one is a simple SortNet consisting of M fully-connected layers with a hidden size of 5120, which is used in MNIST and CIFAR-10. Like [69], we choose $M = 5$ for MNIST and $M = 6$ for CIFAR-10. Since SortNet is Lipschitz, we directly apply the margin-based certification method to calculate the certified accuracy (Proposition 3.1). To achieve the best results on ImageNet-like datasets, in our second type of model we consider using a composite architecture consisting of a base SortNet backbone and a prediction head (denoted as SortNet+MLP). Following [69], the SortNet backbone has 5 layers with a width of 5120 neurons, which serves as a robust feature extractor. The top prediction head is a lightweight 2-layer perceptron with 512 hidden neurons (or 2048 for ImageNet), which takes the robust features as input to give classification results. We also try a larger SortNet backbone, denoted as SortNet+MLP (2x), that has roughly four times the training cost (see Appendix E.2 for architectural details). We use the same approach as [69] to train and certify these models, i.e. by combining margin-based certification and interval bound propagation [20].

Baseline methods and metrics. We compare SortNet with representative literature approaches including relaxation-based certification (for standard networks), margin-based certification (using Lipschitz networks), and mixed-integer linear programming (MILP) [55]. In Appendix G, we also discuss randomized smoothing approaches [9, 46], which provide probabilistic guarantees rather than deterministic ones. For each method in these tables, we report five metrics: (i) training efficiency, measured by the wall-clock time per training epoch; (ii) certification efficiency, measured by the time needed to calculate the certified accuracy on the test dataset; (iii) the clean test accuracy without perturbation (denoted as Clean); (iv) the robust test accuracy under 100-step PGD attack (denoted as PGD); (v) the certified robust test accuracy (denoted as Certified). For a fair comparison, we reproduce most of baseline methods using the official codes and report the wall-clock time under the same NVIDIA-RTX 3090 GPU. These results are presented in Tables 1, 2 and 3.

Performance on MNIST. The results are presented in Table 1. Following the common practice, we consider both a small perturbation radius $\epsilon = 0.1$ and a larger one $\epsilon = 0.3$. It can be seen that the SortNet models can achieve **98.14%** ($\epsilon = 0.1$) and **93.40%** ($\epsilon = 0.3$) certified accuracy, respectively, both of which outperform all previous baseline methods. In contrast, the GroupSort network can only achieve a trivial certified accuracy for $\epsilon = 0.3$. This matches our theory in Section 3.4, indicating that the expressive power of shallow MaxMin networks is insufficient in real-world applications.

Performance on CIFAR-10. The results are presented in Table 2. Following the common practice, we consider two perturbation radii: $\epsilon = 2/255$ and $\epsilon = 8/255$. Our models can achieve **56.94%** ($\epsilon = 2/255$) and **40.39%** ($\epsilon = 8/255$) certified accuracy, respectively. Moreover, the training approach proposed in Section 4 is very efficient, e.g., with a training time of **13**$\sim$**14** seconds per epoch. For both radii, our models perform the best among all existing approaches that can be certified in a reasonable time. Compared with relaxation-based methods, the certified accuracy of SortNet models is much higher (typically 3$\sim$6 point improvement for both radii), despite our training speed being several times faster. Such results may indicate that certified ℓ_∞ robustness can be better achieved by designing suitable Lipschitz models (w.r.t. ℓ_∞ norm) than by devising sophisticated relaxation procedures for non-Lipschitz models.

Performance on TinyImageNet and ImageNet. To demonstrate the scalability of SortNet models, we finally run experiments on two large-scale datasets: Tiny-ImageNet and ImageNet (64×64). Notably, the ImageNet dataset has 1000 classes and contains 1.28 million images for training and 50,000 images for testing. Due to both the large size and the huge number of classes, achieving certified ℓ_∞ robustness on the ImageNet level has long been a challenging task.

Table 3 presents our results along with existing baselines. Among them, we achieve **18.18%** certified accuracy on TinyImageNet and achieve **9.54%** certified accuracy on ImageNet, both of which establish state-of-the-art results. The gap is most prominent on ImageNet, where our small SortNet+MLP model already outperforms the largest model of [65] while being **22** times faster to train. Even for the largest model (SortNet+MLP 2x), the training is still 7 times faster, resulting in a

Table 1: Comparison of our results with existing methods on MNIST dataset.

Method		Train Time (s)	MNIST ($\epsilon = 0.1$)			MNIST ($\epsilon = 0.3$)		
			Clean	PGD	Certified	Clean	PGD	Certified
Relaxation	IBP [20]	17.5	98.92	97.98	97.25	97.88	93.22	91.79
	IBP [48]	34.7	98.84	–	97.95	97.67	–	93.10
	CROWN-IBP [71]	60.3	98.83	98.19	97.76	98.18	93.95	92.98
Lipschitz	GroupSort (MaxMin) [2]	–	97.0	84.0	79.0	97.0	34.0	2.0
	ℓ_∞-dist Net [69]	17.2	98.66	97.85	97.73	98.54	94.62	92.64
	ℓ_∞-dist Net+MLP [69]	17.2	98.86	97.77	97.60	98.56	95.05	93.09
	ℓ_∞-dist Net [70]	17.0	98.93	98.03	97.95	98.56	94.73	93.20
	SortNet	**10.6**	99.01	98.21	**98.14**	98.46	94.64	**93.40**
MILP	COLT [3]	–	99.2	–	97.1	97.3	–	85.7

Table 2: Comparison of our results with existing methods on CIFAR-10 dataset.

Method		Time (s)		$\epsilon = 2/255$			$\epsilon = 8/255$		
		Train	Certify	Clean	PGD	Certified	Clean	PGD	Certified
Relaxation	CAP [62]	659.0	7,570	68.28	–	53.89	28.67	–	21.78
	IBP [20]	19.0	2.74	61.46	50.28	44.79	50.99	31.27	29.19
	IBP [48]	70.4	4.02	66.84	–	52.85	48.94	–	34.97
	CROWN-IBP [71]	87.2	7.01	71.52	59.72	53.97	45.98	34.58	33.06
	CROWN-IBP [65]	45.0	4.02	–	–	–	46.29	35.69	33.38
Lipschitz	ℓ_∞-dist Net [69]	19.7	1.73	60.33	51.55	50.94	56.80	36.19	33.30
	ℓ_∞-dist Net+MLP [69]	19.7	1.74	65.62	51.47	51.05	50.80	36.51	35.42
	ℓ_∞-dist Net [70]	18.9	1.73	60.61	54.28	54.12	54.30	41.84	40.06
	SortNet	14.0	8.00	65.96	57.03	56.67	54.84	41.50	**40.39**
	SortNet+MLP	**13.4**	8.01	67.72	57.83	**56.94**	54.13	41.58	39.99
MILP	COLT [3]	252.0	$\sim 10^5$	78.4	–	60.5	51.7	–	27.5

Table 3: Comparison of our results with existing methods on TinyImageNet and ImageNet datasets.

Method	TinyImageNet ($\epsilon = 1/255$)				ImageNet 64×64 ($\epsilon = 1/255$)			
	Time (s)	Clean	PGD	Certified	Time (s)	Clean	PGD	Certified
IBP [20]	735	26.46	20.60	14.85	11,026	15.96	9.12	6.13
IBP [48]	284	25.71	–	17.64	–	–	–	–
CROWN-IBP [65]	1,256	27.82	20.52	15.86	16,269	16.23	10.26	8.73
ℓ_∞-dist Net+MLP [69]	55	21.82	–	16.31	–	–	–	–
ℓ_∞-dist Net [70]	55	12.57	11.09	11.04	–	–	–	–
SortNet+MLP	**39**	24.17	20.57	17.92	**715**	13.48	10.93	9.02
SortNet+MLP (2x larger)	156	25.69	21.57	**18.18**	2,192	14.79	11.93	**9.54**

training overhead of 4 days using two GPUs. We suspect that continuing to increase the model size will yield better results, given the noticeable improvement of the larger model over the smaller one.

Comparing with ℓ_∞-distance net. As can be seen from these tables, SortNet models consistently achieve better certified accuracy than ℓ_∞-distance nets for all different datasets and perturbation levels. The performance gap is quite prominent compared with the original work [69]. Very recently, a follow-up paper [70] significantly improved the performance of ℓ_∞-distance net using a carefully designed training strategy as well as several training tricks, creating a strong baseline on CIFAR-10. However, the authors did not present the results on large-scale datasets, and we find their approach does not suit the ImageNet-like datasets when the number of classes is large (see Appendix E.6 for details). In contrast, SortNet models enjoy great scalability ranging from MNIST to ImageNet and consistently outperform [70]. The improvement is also remarkable for the $\epsilon = 2/255$ case on CIFAR-10 ($\sim 3\%$ increase in certified accuracy and $\sim 7\%$ increase in clean accuracy).

In Appendix F, we conduct ablation studies on CIFAR-10 by varying the value of ρ and comparing SortNet models ($\rho > 0$) with ℓ_∞-distance net ($\rho = 0$), under *the same training strategy* in this paper without additional tricks. We observe a large gain in certified accuracy when switching from ℓ_∞-distance net to general SortNet. This empirically indicates that incorporating other order statistics has extra benefits in certified ℓ_∞ robustness than using only the maximum (the first order statistic).

Conclusion. In this paper, we study certified ℓ_∞ robustness from the novel perspective of representing Boolean functions. Our analysis points out an inherent problem of standard Lipschitz networks, and provides novel insights on how recently proposed Lipschitz networks resolve the problem. Guided by the theoretical results, we also design a new Lipschitz network with better empirical performance than prior works. We believe both the theoretical framework and empirical improvements may contribute to future research of Lipschitz neural networks in the certified ℓ_∞ robustness area.

References

[1] Miklós Ajtai, János Komlós, and Endre Szemerédi. An 0 (n log n) sorting network. In *Proceedings of the fifteenth annual ACM symposium on Theory of computing*, pages 1–9, 1983.

[2] Cem Anil, James Lucas, and Roger Grosse. Sorting out lipschitz function approximation. In *International Conference on Machine Learning*, pages 291–301, 2019.

[3] Mislav Balunovic and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In *International Conference on Learning Representations*, 2020.

[4] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[5] Avrim Blum, Travis Dick, Naren Manoj, and Hongyang Zhang. Random smoothing might be unable to certify ℓ_∞ robustness for high-dimensional images. *arXiv preprint arXiv:2002.03517*, 2020.

[6] Artem Chernodub and Dimitri Nowicki. Norm-preserving orthogonal permutation linear unit activation functions (oplu). *arXiv preprint arXiv:1604.02313*, 2016.

[7] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017.

[8] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863, 2017.

[9] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[10] Jeremy EJ Cohen, Todd Huster, and Ra Cohen. Universal lipschitz approximation in bounded depth neural networks. *arXiv preprint arXiv:1904.04861*, 2019.

[11] Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of relu networks via maximization of linear regions. In *the 22nd International Conference on Artificial Intelligence and Statistics*, pages 2057–2066. PMLR, 2019.

[12] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, 2019.

[14] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020.

[15] Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. *arXiv preprint arXiv:1803.06567*, 2018.

[16] Krishnamurthy Dj Dvijotham, Robert Stanforth, Sven Gowal, Chongli Qin, Soham De, and Pushmeet Kohli. Efficient neural network verification with exactness characterization. In *Uncertainty in Artificial Intelligence*, pages 497–507. PMLR, 2020.

[17] Farzan Farnia, Jesse Zhang, and David Tse. Generalizable adversarial training via spectral normalization. In *International Conference on Learning Representations*, 2019.

[18] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pages 3–18. IEEE, 2018.

[19] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael Cree. Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.

[20] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Yujia Huang, Huan Zhang, Yuanyuan Shi, J Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local lipschitz bounds. *Advances in Neural Information Processing Systems*, 34, 2021.

[23] Todd Huster, Cho-Yu Jason Chiang, and Ritu Chadha. Limitations of the lipschitz constant as a defense against adversarial examples. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 16–29. Springer, 2018.

[24] Sergey Ioffe and Christian Szegedy. In *International Conference on Machine Learning*, pages 448–456, 2015.

[25] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[26] Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. *Advances in Neural Information Processing Systems*, 33, 2020.

[27] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.

[28] Alex Krizhevsky et al. Learning multiple layers of features from tiny images. *Technical Report TR-2009*, 2009.

[29] Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. In *International Conference on Machine Learning*, pages 5458–5467. PMLR, 2020.

[30] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[31] Yann LeCun, Corinna Cortes, and Chris Burges. Mnist handwritten digit database, 1998.

[32] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

[33] Sungyoon Lee, Jaewook Lee, and Saerom Park. Lipschitz-certifiable training with a tight outer bound. *Advances in Neural Information Processing Systems*, 33, 2020.

[34] Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

[35] Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

[36] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. *Advances in Neural Information Processing Systems*, 32:9464–9474, 2019.

[37] Qiyang Li, Saminul Haque, Cem Anil, James Lucas, Roger B Grosse, and Joern-Henrik Jacobsen. Preventing gradient attenuation in lipschitz constrained convolutional networks. *Advances in Neural Information Processing Systems*, 32:15390–15402, 2019.

[38] Laurent Meunier, Blaise Delattre, Alexandre Araujo, and Alexandre Allauzen. A dynamical system perspective for lipschitz neural networks. *arxiv:2110.12690*, 2022.

[39] Matthew Mirman, Maximilian Baader, and Martin Vechev. The fundamental limits of interval arithmetic for neural networks. *arXiv preprint arXiv:2112.05235*, 2021.

[40] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pages 3578–3586. PMLR, 2018.

[41] Ryan O'Donnell and Karl Wimmer. Approximation by dnf: examples and counterexamples. In *International Colloquium on Automata, Languages, and Programming*, pages 195–206. Springer, 2007.

[42] Allan Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999.

[43] Haifeng Qian and Mark N. Wegman. L2-nonexpansive neural networks. In *International Conference on Learning Representations*, 2019.

[44] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

[45] Aditi Raghunathan, Jacob Steinhardt, and Percy S. Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10900–10910, 2018.

[46] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, and Sébastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 11292–11303, 2019.

[47] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A convex relaxation barrier to tight robustness verification of neural networks. *Advances in Neural Information Processing Systems*, 32:9835–9846, 2019.

[48] Zhouxing Shi, Yihan Wang, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Fast certified robust training with short warmup. *Advances in Neural Information Processing Systems*, 34, 2021.

[49] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems*, pages 10802–10813, 2018.

[50] Sahil Singla and Soheil Feizi. Skew orthogonal convolutions. In *International Conference on Machine Learning*, volume 139, pages 9756–9766. PMLR, 2021.

[51] Sahil Singla, Surbhi Singla, and Soheil Feizi. Improved deterministic l2 robustness on CIFAR-10 and CIFAR-100. In *International Conference on Learning Representations*, 2022.

[52] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[53] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[54] Ugo Tanielian and Gerard Biau. Approximating lipschitz continuous functions with groupsort neural networks. In *International Conference on Artificial Intelligence and Statistics*, pages 442–450. PMLR, 2021.

[55] Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019.

[56] Asher Trockman and J Zico Kolter. Orthogonalizing convolutional layers with the cayley transform. In *International Conference on Learning Representations*, 2021.

[57] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in Neural Information Processing Systems*, 31:6541–6550, 2018.

[58] Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[59] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. *Advances in Neural Information Processing Systems*, 34, 2021.

[60] 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 *International Conference on Machine Learning*, pages 5276–5285. PMLR, 2018.

[61] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR, 2018.

[62] Eric Wong, Frank R Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8410–8419, 2018.

[63] Yihan Wu, Aleksandar Bojchevski, Aleksei Kuvshinov, and Stephan Günnemann. Completing the picture: Randomized smoothing suffers from the curse of dimensionality for a large family of distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 3763–3771. PMLR, 2021.

[64] Kai Y Xiao, Vincent Tjeng, Nur Muhammad Mahi Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing relu stability. In *International Conference on Learning Representations*, 2019.

[65] Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems*, 33, 2020.

[66] Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pages 10693–10705. PMLR, 2020.

[67] Yuichi Yoshida and Takeru Miyato. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

[68] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020.

[69] Bohang Zhang, Tianle Cai, Zhou Lu, Di He, and Liwei Wang. Towards certifying l-infinity robustness using neural networks with l-inf-dist neurons. In *International Conference on Machine Learning*, pages 12368–12379. PMLR, 2021.

[70] Bohang Zhang, Du Jiang, Di He, and Liwei Wang. Boosting the certified robustness of l-infinity distance nets. In *International Conference on Learning Representations*, 2022.

[71] Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020.

[72] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *Advances in Neural Information Processing Systems*, 31:4939–4948, 2018.

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 the paper, we pointed out some of the problems that we have not yet addressed, both in terms of theoretical results (Section 3.4) and better implementations for training (Section 4). They are left for future work.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] Our work is only for academic research purpose, without any foreseeable negative societal impacts.
   (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] Complete proofs are given either in the main text or in Appendix B.
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] We give complete implementation details in Appendix E. Our code and models will be released once the paper is published. We use public benchmark datasets which can be downloaded freely.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Appendix E.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] See Appendix F.
   (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] See Appendix E.
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? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We use public benchmark datasets which can be downloaded freely.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We use public benchmark datasets which can be downloaded freely.
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]