
Provable Lipschitz Certification for Generative Models

Matt Jordan¹ Alexandros G. Dimakis¹

Abstract

We present a scalable technique for upper bounding the Lipschitz constant of generative models. We relate this quantity to the maximal norm over the set of attainable vector-Jacobian products of a given generative model. We approximate this set by layerwise convex approximations using zonotopes. Our approach generalizes and improves upon prior work using zonotope transformers and we extend to Lipschitz estimation of neural networks with large output dimension. This provides efficient and tight bounds on small networks and can scale to generative models on VAE and DCGAN architectures.

1. Introduction

We study the problem of bounding the Lipschitz constant of generative models. The central technical difficulty is that these are vector-valued functions with high-dimensional outputs, so the techniques for Lipschitz estimation of scalar-valued neural networks do not directly translate. Our approach is to frame the Lipschitz constant estimation problem as an optimization over the range of attainable vector-Jacobian products. We overapproximate this set via layerwise convex approximations, and then solve a relaxed version of the optimization problem.

The primary challenge is computing the feasible set of vector-Jacobian products over a range of inputs and vectors. We generate an overapproximation of this set by representing it as a zonotope. While prior work has performed reachability analysis of neural networks using zonotopes, these approaches only consider the forward pass. Our technique generalizes this prior work and is able to yield tighter bounds. Additionally, we are able to apply our approach to backpropagation, where we pass zonotopes backwards in place of vectors.

We present a general algorithm for mapping zonotopes

¹University of Texas at Austin. Correspondence to: Matt Jordan <mjordan@cs.utexas.edu>.

through a variety of nonlinear operators and relate this to a 2-dimensional geometric problem that we solve optimally. The final norm-maximization that arises turns out to be equivalent to a mathematical question called the Grothendieck problem. We demonstrate that the linear-programming relaxation of this problem can be solved in linear time with our machinery. To fairly compare our work against existing Lipschitz estimation techniques, we first compare the estimated Lipschitz value and runtime of our algorithm on networks of increasing size trained on a toy dataset. We observe that our approach favorably trades-off accuracy for efficiency and yields tighter bounds compared to previous techniques. Further, it can significantly improve the runtime of exact Lipschitz computations. We scale our technique to generative models on MNIST and CIFAR-10, using well-known architectures like DCGAN and VAEs with both fully-connected and convolutional layers (Kingma & Welling, 2013; Radford et al., 2015). Our approach yields much tighter bounds compared to any other technique that can handle vector-valued networks and can scale to such architecture sizes.

2. Related Work

Robustness Certification: Lipschitz estimation is closely related to robustness certification. Here, the goal is to provide a certificate of robustness against adversarial attacks for a specified network and input region. The techniques of interest in this domain are Lipschitz approximation and reachability analysis. It has been noted multiple times that an upper bound to the Lipschitz constant can provide a guarantee of robustness against adversarial examples (Hein & Andriushchenko, 2017; Weng et al., 2018b;a). Reachability analysis is frequently couched in the language of abstract interpretations where the goal is to develop sound transformations to map sets through the forward pass of a given classifier. The classes of sets considered include hyperboxes and zonotopes, as well as polytopes, imageStars, and linear bounds (Singh et al., 2019a;b; Mirman et al., 2018; Zico Kolter & Wong, 2017; Weng et al., 2018a; Xu et al., 2020; Zhang et al., 2018; Singh et al., 2018; Tran et al., 2020).

Lipschitz Approximation: A number of recent works provide either heuristic estimates or provable upper bounds

of the Lipschitz constant of neural networks. For ReLU networks, this problem is known to be NP-hard and has strong inapproximability guarantees (Virmaux & Scaman, 2018; Jordan & Dimakis, 2020). However for small networks it has been shown that this quantity may be estimated to a reasonable degree of accuracy. The majority of these works are either unable to handle vector-valued neural networks or are unable to scale to larger networks. Provable upper bounds of the Lipschitz constant may be attained using interval analysis (Weng et al., 2018a), semidefinite programming (Fazlyab et al., 2019), linear or mixed-integer programming (Jordan et al., 2019), or polynomial optimization (Latorre et al., 2019; Chen et al., 2020). Heuristic estimates of the Lipschitz constant may be attained by either randomness and extreme-value theory (Weng et al., 2018b), or by a greedy algorithm over the activation patterns (Virmaux & Scaman, 2018). Our work can be viewed as a spiritual successor to the interval analysis approach of Lipschitz estimation and the zonotope abstract interpretation approaches for robustness certification.

3. Problem Statement and Relaxation Strategy

We formally define the problem of estimating the Lipschitz constants of vector-valued functions and provide a broad overview of the strategy for our convex relaxation. We start with the notation we will employ.

Notation: We will denote vectors using lowercase letters, matrices using capital letters and sets using calligraphic letters. We denote the unit-norm ball with respect to the α -norm as B_α . We refer to the dual norm of the α -norm as the norm $\|\cdot\|_{\alpha^*}$, defined as $\|v\|_{\alpha^*} := \sup_{u \in B_\alpha} |u^T v|$. For a function $f(\cdot) : \mathbb{R}^k \rightarrow \mathbb{R}^n$, we denote the Jacobian with respect to its argument as $\nabla_x f(x)$, which is a matrix in $\mathbb{R}^{n \times k}$. We will frequently abuse notation and for sets $\mathcal{X}$ and functions f , we write $f(\mathcal{X})$ to denote the set $\{f(x) \mid x \in \mathcal{X}\}$. We use $\odot$ to denote the Hadamard product of two vectors, and $\oplus$ to denote the Minkowski sum of two sets.

Lipschitz constants of vector-valued functions: Given a neural network $f : \mathbb{R}^k \rightarrow \mathbb{R}^n$, an input domain $\mathcal{X} \subseteq \mathbb{R}^k$ and norms $\|\cdot\|_\alpha, \|\cdot\|_\beta$ over $\mathbb{R}^k, \mathbb{R}^n$ respectively, the local Lipschitz constant is formally defined as:

$$L^{(\alpha, \beta)}(f, \mathcal{X}) := \sup_{x, y \in \mathcal{X}} \frac{\|f(x) - f(y)\|_\beta}{\|x - y\|_\alpha}. \quad (1)$$

When f is continuously differentiable and $\mathcal{X}$ is an open set, this quantity may be written as an optimization over matrix norms of the Jacobian. Letting $\nabla_x f(x)$ be the Jacobian of f evaluated at x , we have that

$$L^{(\alpha, \beta)}(f, \mathcal{X}) = \sup_{x \in \mathcal{X}} \|\nabla_x f(x)\|_{\alpha \rightarrow \beta}. \quad (2)$$

The matrix norm $\|M\|_{\alpha \rightarrow \beta}$ is defined as

$$\|M\|_{\alpha \rightarrow \beta} := \sup_{\|v\|_\alpha \leq 1} \|Mv\|_\beta = \sup_{\|u\|_{\beta^*} \leq 1} \|M^T u\|_{\alpha^*}. \quad (3)$$

where the second equality follows from the definition of the dual norm. Combining Equations 2 and 3, we can formulate the problem of computing the Lipschitz constant as an optimization over vector-Jacobian products:

$$L^{(\alpha, \beta)}(f, \mathcal{X}) = \sup_{x \in \mathcal{X}} \sup_{u \in B_{\beta^*}} \|\nabla_x f(x)^T u\|_{\alpha^*}. \quad (4)$$

When f is nonsmooth, as in the case of neural networks with ReLU nonlinearities, the optimization over the Jacobian in Equation 4 is replaced with an optimization over Clarke generalized subgradients (Jordan & Dimakis, 2020). For our purposes, as we seek only to upper bound this quantity, this distinction is of minimal importance.

The strategy we employ to upper bound Equation 4 will rely on two clear steps. The first step is to generate a sound approximation of the set of vector-Jacobian products of f , and the second step is to bound the maximal $\|\cdot\|_{\alpha^*}$ norm of this set. Concisely, we first develop a set $\mathcal{Y}$ satisfying the containment

$$\{\nabla_x f(x)^T u \mid x \in \mathcal{X}, u \in B_{\beta^*}\} \subseteq \mathcal{Y}, \quad (5)$$

and then we upper bound the maximal α^* norm of $\mathcal{Y}$.

We will focus in particular on the $L^{(\infty, 1)}(f, \mathcal{X})$ Lipschitz constant. The choice of $\|\cdot\|_\alpha$ to be the ℓ_∞ norm is standard in the robustness certification literature. The choice of $\|\cdot\|_\beta$ to be the ℓ_1 norm will yield an upper bound for $L^{(\alpha, p)}(f, \mathcal{X})$ for $p \geq 1$ as the dual ball, B_{β^*} is the ℓ_∞ ball and contains all other ℓ_p balls. This approach relies on mapping ℓ_∞ balls through both the forward and backward pass of a network, which zonotopes are well-suited for.

Vector-Jacobian Products of Neural Networks: To handle the step of overapproximating the set of vector-Jacobian products, we turn our attention to the structure of the functions we consider. We consider feedforward neural networks with either fully-connected or convolutional layers and elementwise nonlinearities, σ , such as the ReLU, tanh, or sigmoid operators. A neural network f with L hidden layers may be evaluated according to the following recursion,

$$f(x) := W_L Z_L(x) + b_L \quad (6)$$

$$Z_i(x) = \sigma(\hat{Z}_i(x)) \quad (7)$$

$$\hat{Z}_i(x) = W_i Z_{i-1}(x) + b_i \quad (8)$$

$$Z_0(x) = x, \quad (9)$$

for $i \in \{1, \dots, L\}$. When f has convolutional layers, the affine operator in the definition of $\hat{Z}_i(x)$ may be instantiated as a convolution operator.

While the full Jacobian may be expensive to compute, vector-Jacobian products are a standard operation performed on neural nets via the backpropagation algorithm. This may be evaluated according to the following recursion,

$$\nabla_x f(x)^T u := W_1^T J_1(x)^T Y_1(x, u) \quad (10)$$

$$Y_i(x, u) = J_{i+1}^T(x) \hat{Y}_i(x, u) \quad (11)$$

$$\hat{Y}_i(x, u) = W_{i+1}^T Y_{i+1}(x, u) \quad (12)$$

$$Y_L(x, u) = u, \quad (13)$$

where $J_i(x)$ is the Jacobian of the i^{th} nonlinearity with respect to its input, $\nabla_{\hat{Z}_i} Z_i(x)$, and i ranges from 1 to $L - 1$. For the standard elementwise nonlinearities, $J_i(x)$ is diagonal and may be written as the Hadamard product with a vector. For example, when σ is the ReLU operator, this is a Hadamard product with a vector taking entries in $\{0, 1\}$, depending on the sign of the input to each neuron. When convolutional layers are used in place of fully-connected layers, the transpose convolution operator with no bias terms may be used in place of W_{i+1}^T in the definition of $\hat{Y}_i(x, u)$. For generative models that yield images, the outputs are constrained to the hyperbox $[0, 1]^n$, usually by applying a sigmoid or tanh layer to the output of f . In this case, $\hat{Y}_{L-1}(x, u)$ is $W_L^T J_L^T(x)u$, for $J_L(x)$ denoting the Jacobian with respect to this final nonlinear layer.

As our goal is to generate a set $\mathcal{Y}$ satisfying the containment in equation 5, we can unroll the recursions and iteratively construct sets which serve as sound approximations of each element in the recursion. For an input range of $\mathcal{X}$, our algorithm will yield a collection of sets $\mathcal{Z}_i, \hat{\mathcal{Z}}_i, \mathcal{J}_i, \mathcal{Y}_i, \hat{\mathcal{Y}}_i$ satisfying the containments

$$\mathcal{X} \subseteq \mathcal{Z}_0 \quad B_{\beta^*} \subseteq \hat{\mathcal{Y}}_L \quad (14)$$

$$W_i \mathcal{Z}_{i-1} + b_i \subseteq \hat{\mathcal{Z}}_i \quad W_i^T \hat{\mathcal{Y}}_i \subseteq \mathcal{Y}_{i-1} \quad (15)$$

$$\sigma(\hat{\mathcal{Z}}_i) \subseteq \mathcal{Z}_i \quad \mathcal{J}_i \odot \mathcal{Y}_{i+1} \subseteq \hat{\mathcal{Y}}_i \quad (16)$$

$$\nabla_z \sigma(\mathcal{Z}_i) \subseteq \mathcal{J}_i \quad (17)$$

That is, we first overapproximate the range of attainable values of each layer of the neural net, $\mathcal{Z}_i(\mathcal{X})$ and $\hat{\mathcal{Z}}_i(\mathcal{X})$. This allows us to create sets that contain the true range of gradients for each nonlinearity, $\nabla_{\hat{\mathcal{Z}}_i} \hat{\mathcal{Z}}_i(\mathcal{X})$ as per the left column. Then a similar procedure is used to obtain sets which contain the true range of partial vector-Jacobian products as the backpropagation algorithm is performed, i.e. $Y_i(\mathcal{X}, B_{\beta^*})$. Ultimately this will yield a set that contains the set of attainable vector-Jacobian products $\nabla_x f(\mathcal{X})^T B_{\beta^*}$. Soundness is encapsulated in the following theorem.

Theorem 1. *For feedforward neural networks f , an input set $\mathcal{X}$ and sets $\mathcal{Z}_i, \hat{\mathcal{Z}}_i, \mathcal{J}_i, \mathcal{Y}_i, \hat{\mathcal{Y}}_i$ satisfying the containments in Equations 14-17, the set of vector-Jacobian products satisfies*

$$\{\nabla_x f(x)^T u \mid x \in \mathcal{X} \quad u \in B_{\beta^*}\} \subseteq \mathcal{Y}_0. \quad (18)$$

For such a $\mathcal{Y}_0$, the Lipschitz constant of f may be upper-bounded by maximizing the $\|\cdot\|_{\alpha^*}$ norm over the set $\mathcal{Y}_0$.

Abstracting this slightly, we notice that each of the recursive containments follow one of four forms. The first is the mapping of sets through affine operators as in equations 15. Next we require the mapping of sets through the nonlinearity σ or an elementwise multiplication as in equation 16. Third we have the Jacobian operator of σ as in equation 17. In the sequel we will describe a family of sets that trades off expressiveness with efficiency of representation, and then we develop a technique to perform each of these four operations in a way that satisfies the required containments.

4. Hyperboxes and Zonotopes

The key idea to handle the sound approximations required by Theorem 1 is to introduce a family of sets and develop transformations that are closed under this family of sets and preserve the necessary containment for the four classes of operators. The families of sets we will consider here are hyperboxes and zonotopes. Throughout our Lipschitz estimation procedure, we will frequently make use of the fact that linear programs and Minkowski sums of these sets are efficiently computable.

Hyperboxes: An axis-aligned hyperbox in $\mathbb{R}^d$ may be defined by a center point $c \in \mathbb{R}^d$ and a radius vector $r \in \mathbb{R}^d$ with $r \geq 0$, such that the set $H(c, r)$ may be defined as

$$H(c, r) := \{c + r \odot y \mid \|y\|_{\infty} \leq 1\}.$$

Hyperboxes have very efficient representations, and enjoy many nice computational properties. Namely, linear programs over $H(c, r)$ admit a closed-form solution computable in time $O(d)$ as do Minkowski sums:

$$\max_{x \in H(c, r)} a^T x = a^T c + \|a \odot r\|_1$$

$$H(c_1, r_1) \oplus H(c_2, r_2) = H(c_1 + c_2, r_1 + r_2).$$

Zonotopes: Zonotopes are a family of sets that can be much more expressive than hyperboxes but also enjoy many of the same efficient subroutines. A zonotope may be defined as the image of an affine operator applied to a hyperbox, or equivalently, a Minkowski sum of line segments. Typically a zonotope in $\mathbb{R}^d$ is represented in the G-representation, where a center $c \in \mathbb{R}^d$ and a generator matrix $G \in \mathbb{R}^{d \times m}$ are supplied. The number of columns, m , of E is referred to as the degrees of freedom of a zonotope. Formally, these sets are described as

$$Z(c, E) := \{c + E y \mid \|y\|_{\infty} \leq 1\}.$$

Linear programs over zonotopes also admit a closed form solution as

$$\max_{x \in Z(c, E)} a^T x = a^T c + \|E^T a\|_1$$

which follows from the definitions of dual norms. An important and frequently used application of this fact is that coordinate-wise lower and upper bounds may be efficiently computed. Indeed, the smallest hyperbox containing a zonotope may be written as

$$Z(c, E) \subseteq H(c, |E|\vec{1}).$$

where $|E|$ denotes the absolute value operator applied elementwise to E . Similarly, the Minkowski sum of two zonotopes is efficiently computable as

$$Z(c_1, E_1) \oplus Z(c_2, E_2) = Z(c_1 + c_2, E_1 || E_2)$$

where $E_1 || E_2$ is the concatenation of the columns of E_1 with those of E_2 .

5. Zonotopes and Sound Pushforward Operators

Now we describe how to construct sound transformations as required by the operations outlined in Equations 14-17.

Affine operators: Both hyperboxes and zonotopes have efficient sound transformations when mapping through affine operators. Consider an affine operator $x \rightarrow Ax + b$. In general, hyperboxes are not closed under affine operation, but the tightest sound operator maps a hyperbox $H(c, r)$ to the hyperbox $H(Ac + b, |A|r)$. Zonotopes, on the other hand, are closed under affine operation and $Z(c, r)$ maps to $Z(Ac + b, AE)$.

Elementwise nonlinearities: In general, zonotopes may not be closed under elementwise nonlinearities. Here we will demonstrate a strategy for these mappings that are optimal in a sense and improve upon the mappings of zonotopes through elementwise nonlinearities from prior works.

In general, the problem we consider is to map a zonotope $Z \subseteq \mathbb{R}^d$ through an operator $\Phi : \mathbb{R}^d \rightarrow \mathbb{R}^d$ where $\Phi(x) = (\phi(x_1), \dots, \phi(x_d))$. That is, we wish to develop a zonotope Z' satisfying $\Phi(Z) \subseteq Z'$.

The strategy we employ to construct Z' is to retain the structure of Z and incorporate new degrees of freedom. We will scale Z along each coordinate and cover the errors by taking the Minkowski sum with a new zonotope:

$$Z' = (\Lambda \odot Z) \oplus \hat{Z}$$

where Λ is a vector and $\hat{Z}$ represents the a zonotope containing the new degrees of freedom. The following sufficient condition states that this transformation satisfies the desired containment property:

Lemma 1. *For any zonotope $Z \subset \mathbb{R}^d$ and any operator Φ operating over $\mathbb{R}^d$, if $\hat{Z}$ is a zonotope satisfying the containment*

$$\{\Phi(z) - \Lambda \odot z \mid z \in Z\} \subseteq \hat{Z} \quad (19)$$

then

$$\Phi(Z) \subseteq (\Lambda \odot Z) \oplus \hat{Z}.$$

We refer to the set $\{\Phi(z) - \Lambda \odot z \mid z \in Z\}$ as the *residuals*, and reduce the problem to finding a vector Λ and set $\hat{Z}$ containing these residuals. Our strategy is to consider sets $\hat{Z}$ that are axis-aligned hyperboxes, i.e. $\hat{Z} = H(b^*, \mu^*)$. While there are many such residual hyperboxes, a reasonable heuristic would be to choose Λ to yield the smallest hyperbox satisfying Lemma 1.

By considering only transformations that scale each coordinate of Z independently and accounting for the residuals with a hyperbox, it suffices to consider each coordinate individually. In this case, the soundness criterion of Equation 19 reduces to the condition:

$$\{\phi(z_i) - \Lambda_i z_i \mid z \in Z\} \subseteq [b_i - \mu, b_i + \mu].$$

By our heuristic, we would like to minimize the size of the residual interval, 2μ in the above equation. This may be written as a min-max optimization:

$$\min_{\Lambda_i, b_i} \max_{z \in Z} |\phi(z_i) - \Lambda_i z_i - b_i|. \quad (20)$$

Indeed, this may be equivalently be viewed as fitting an affine function $L(z_i) := \Lambda_i z_i + b_i$ to the function $\phi(z_i)$ such that the maximum absolute value deviation between $L(z_i)$ and $\phi(z_i)$ is minimized across all Z . Now assume that the optimal objective value of the above min-max is $2\mu^*$, and the argmin and argmax are (Λ_i^*, b_i^*) and (z_i^*) respectively. Then by definition, we have that

$$\{\phi(z_i) - \Lambda_i^* z_i \mid z \in Z\} \subseteq [b_i^* - \mu^*, b_i^* + \mu^*].$$

By computing the optimal solution to Equation 20 for each coordinate i , we can compute Λ^* and a residual hyperbox $H(b^*, \mu^*)$ satisfying the sufficient condition required by Lemma 1.

It remains to be seen how to efficiently solve the min-max of Equation 20. We consider this problem graphically and notice that any 2-dimensional set of points described as the points with vertical deviation of no more than μ from an affine function $L(z_i)$ is a parallelogram with vertical sides. Indeed, we refer to sets of the form

$$\{(z_i, y_i) \mid z_i \in [l_i, u_i] \mid y_i - L(z_i) \leq \mu\}$$

as *vertical parallelograms*, parameterized by the line $L(\cdot)$ and vertical range μ and denoted as $P(L, \mu)$. As we have shown, the min-max can be reduced to an instance of the vertical-parallelogram fitting problem.

Definition 1. *The vertical parallelogram fitting problem asks the following question. Given a 2-dimensional set S , we seek to find the vertical parallelogram with minimal area that contains the set S .*

Since the horizontal range of the provided set S is fixed, the area of a vertical parallelogram hinges only upon the length of its vertical side. We discuss how to solve this problem in the next section.

Applying this problem to the particular form of the min-max in Equation 20, we arrive at the following theorem:

Theorem 2. *When S is the set $\{(z_i, \phi(z_i)) \mid z_i \in [l_i, u_i]\}$, the solution to the vertical parallelogram fitting problem yields the optimal solution to Equation 20. Repeated calls to this subroutine yields the tightest hyperbox fitting the residuals as in equation 19.*

Jacobians of Elementwise Operators: Now we consider sound operators for the Jacobians of these elementwise nonlinearities, which will ultimately yield the sets $\mathcal{J}_i$ as in equation 17. Specifically the goal is to develop a set containing

$$\{\nabla_z \Phi(z) \mid z \in Z\}$$

for any zonotope Z . While the strategy presented above certainly applies to this case, we choose to develop a hyperbox approximation for this containment. As hyperboxes allow for independence of coordinates and the Φ operator is an elementwise operator, this reduces to computing, for each coordinate i , the values:

$$j_i^{(l)} := \min_{z \in Z} \phi'(z_i) \quad j_i^{(u)} := \max_{z \in Z} \phi'(z_i).$$

When the i^{th} coordinate of Z is bounded in $[l_i, u_i]$, the above minimum and maximum may be solved efficiently for common nonlinearities. Indeed for ReLU, we have that $j_i^{(l)} = \text{sign}(l_i)$ and $j_i^{(u)} = \text{sign}(u_i)$. For the sigmoid and tanh operators, this is $j_i^{(l)} = \phi'(\max(\{|l_i|, |u_i|\}))$ and $j_i^{(u)} = \phi'(\min(\{|l_i|, |u_i|\}))$. This may be computed for every coordinate i and yields the hyperbox with center $\frac{j^{(u)} + j^{(l)}}{2}$ and radius $\frac{j^{(u)} - j^{(l)}}{2}$.

Elementwise multiplication: Using the above machinery, we can handle the elementwise multiplication operator in a sound fashion. Given a zonotope Z and a hyperbox H , we wish to develop a zonotope which contains the set

$$\{x \odot z \mid x \in H, \quad z \in Z\}$$

Parallel to Lemma 1, we employ a strategy where we seek to find the hyperbox that most tightly fits the residual set. This soundness criterion is proved in the following lemma:

Lemma 2. *For any zonotope $Z \subseteq \mathbb{R}^d$ and hyperbox $H \subseteq \mathbb{R}^d$, if $\hat{Z}$ is a zonotope satisfying the containment*

$$\{x \odot z - \Lambda \odot z \mid x \in H \quad z \in Z\} \subseteq \hat{Z} \quad (21)$$

then

$$\{x \odot z \mid x \in H \quad z \in Z\} \subseteq (\Lambda \odot Z) \oplus \hat{Z}.$$

lemma

Since $\odot$ acts elementwise, H is a hyperbox, and we only seek to fit a hyperbox residual, we may again consider each coordinate independently. This reduces to another min-max problem where the maximum is now taken over both Z and H .

$$\min_{\Lambda_i, b_i} \max_{z \in Z, x \in H} |x_i \cdot z_i - \Lambda_i z_i - b_i|. \quad (22)$$

We may again solve this via a reduction vertical-parallelogram fitting problem. We can suppose that z_i is contained in the interval $[l_i^{(z)}, u_i^{(z)}]$ and x_i is contained in the interval $[l_i^{(x)}, u_i^{(x)}]$. Then the following theorem relates the vertical parallelogram fitting problem to the elementwise multiplication operator.

Theorem 3. *When S is the set $\{(z, x \cdot z) \mid l^{(z)} \leq z \leq u^{(z)} \quad l^{(x)} \leq x \leq u^{(x)}\}$, the solution to the vertical-parallelogram fitting problem yields the optimal solution to Equation 22. Repeated calls to this subroutine yields the tightest hyperbox fitting the residuals as in Equation 21.*

theorem In this sense, we may once again compute the vertical parallelogram fit for each coordinate i to generate the scaling factor Λ and residual hyperbox Z' .

6. Vertical-Parallelogram fitting problem

In the cases of elementwise operators or elementwise multiplication by a hyperbox, we have reduced the problem of tightly fitting the residuals to the vertical parallelogram fitting problem. Here we describe our algorithm to solve this problem and illustrate its use on two examples. Prior work has optimally solved this problem for the ReLU nonlinearity, but is unnecessarily loose for differentiable nonlinearities. We provide full derivations for the ReLU, sigmoid, tanh, and absolute value operators in the appendix.

Algorithm Consider some general 2-dimensional set S . We will walk through the major components of our algorithm and describe the necessary subroutines to fit a vertical parallelogram to S . Recall that vertical-parallelograms are parameterized by a scalar vertical side-length, 2μ , and an affine function $L(\cdot)$ which is parameterized by a slope λ and intercept b .

First we compute the vertical-side length μ . Since vertical parallelograms are convex, any vertical parallelogram containing the target set S must contain its convex hull. Assuming that the set S is bounded in the horizontal dimension by $[l_x, u_x]$, the convex hull of S may be decomposed into a convex upper and lower hull, $h^+(x)$, $h^-(x)$ such that

$$\text{conv}(S) = \{(x, y) \mid x \in [l_x, u_x], \quad h^-(x) \leq y \leq h^+(x)\},$$

where the upper hull is concave and the lower hull is convex. The vertical deviation between these hulls, which we refer

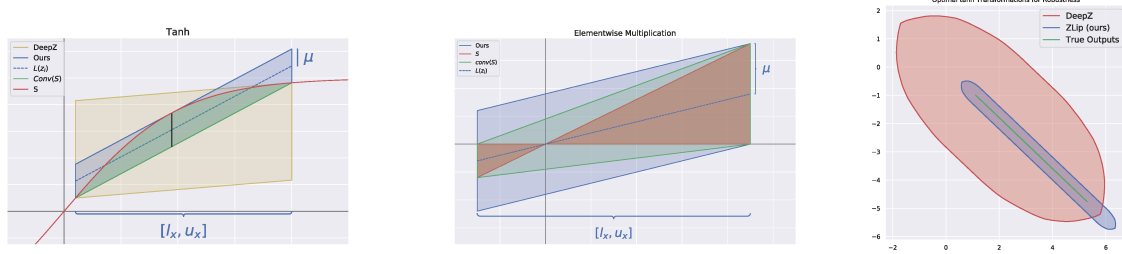


Figure 1. Examples of the vertical parallelogram fitting algorithm for the tanh (left) and elementwise multiplication (middle) operators. The sets S are drawn in red, with their convex hulls in green. The optimal vertical parallelogram for each is drawn in blue. The bound yielded by DeepZ is in yellow for the tanh operator. For classifiers trained on the toy dataset, our improved tanh bounds yield tighter reachability analysis compared to prior work (right). The set of attainable outputs of a network for a specified region is plotted in green, DeepZ yields the set shaded in red, whereas our reachable set is shaded in blue.

to as the altitude at x , is denoted by their difference $h^+(x) - h^-(x)$. The vertical side-length 2μ is lower bounded by the altitude at x for every $x \in [l_x, u_x]$. If the maximum of the altitude across $[l_x, u_x]$ is $2\mu^*$, attained at x^* , the vertical side-length of P must satisfy $2\mu^* \leq 2\mu$. It turns out this is always attainable with equality, i.e. there always exists an affine function L such that $P(L, \mu^*)$ contains S .

For maximum altitude $2\mu^*$ attained at x^* , we compute an affine function L such that the shifted affine function $L(x) + \mu^*$ is greater than $h^+(x)$ for all $x \in [l_x, u_x]$ and vice versa for the lower convex hull, $L(x) - \mu^* \leq h^-(x)$. We first solve for the slope of L . The subgradient of a convex function f , denoted $\delta f(x)$, is the set of slopes of affine functions passing through the point $(x, f(x))$ that lower bound f . Thus, the slope λ must lie in both the subgradient of $h^-(x)$ and $-h^+(x)$. The intersection of these subgradients is guaranteed to be nonempty as x^* minimizes the convex function $h^-(x) - h^+(x)$, implying $0 \in \delta(h^-(x^*) - h^+(x^*))$. Any element in the intersection of these subgradients suffices as a choice for the slope of the fitting parallelogram. We arbitrarily choose such an element and denote it as λ^* .

Finally we compute the intercept of the line L . For the bounds to be tight, the constraint that $L(x) + \mu^* \geq h^+(x)$ must be tight at x^* and likewise for h^- . Hence the line L must pass through the point $(x^*, \frac{h^+(x^*) + h^-(x^*)}{2})$. Provided with a slope λ^* , it is trivial to compute the intercept of a line passing through that point. This concludes the algorithm and yields the desired parallelogram attaining the tightest possible vertical deviation.

The running time of this algorithm depends crucially on the structure of the set S . Computing the upper and lower convex hulls of S may be challenging and algorithms to compute these functions need to be designed by hand. However, the sets we consider admit efficient algorithms to compute these functions. Multiple prior works have correctly computed the convex hull for the set defined by the ReLU

operator, but have provided suboptimal bounds for differentiable nonlinearities. To illustrate its use, we will apply our algorithm to the tanh operator and the elementwise multiplication operator.

Tanh example: Consider the tanh function and the horizontal coordinate bounds $[l_x, u_x]$ where $l_x \geq 0$. In this case, we desire to fit a parallelogram to the set $\{(x, \tanh(x)) \mid x \in [l_x, u_x]\}$. The tanh function is concave for $x \geq 0$, so the convex hull of S has an upper hull of $h^+(x) = \tanh(x)$. The lower hull, $h^-(x)$, is the secant line passing through the points $(l_x, \tanh(l_x))$ and $(u_x, \tanh(u_x))$, which has slope $\lambda = \frac{\tanh(u_x) - \tanh(l_x)}{u_x - l_x}$. The altitude is maximized when the slope of $\tanh(x)$ is equal to λ , which is attained at $x^* = \tanh^{-1}(\sqrt{1 - \lambda})$. This yields the maximal vertical deviation $2\mu^* = \tanh(x^*) - h^-(x^*)$. Since the lower hull is a secant line, the subgradients everywhere are λ , so the slope of $L(\cdot)$ must be λ . The intercept of the parallelogram's line $L(\cdot)$ must pass through the point $(x^*, \tanh(x^*) - \mu^*)$. This is illustrated graphically in figure 1 (left), where the set S is drawn in red and its convex hull is green. The parallelogram yielded by our algorithm is in blue, whereas the parallelogram yielded by prior work is yellow (Singh et al., 2018).

Elementwise multiplication example: Now consider the case of elementwise multiplication. Consider the x -interval with $l_x < 0 < u_x$ multiplied by a value y in the range $[\alpha, \beta]$ for $\alpha \geq 0$. The set $S = \{(x, x \cdot y) \mid l_x \leq x \leq u_x, \alpha \leq y \leq \beta\}$ is the union of the two triangles shown in red in Figure 1 (middle). The convex hull of S is the trapezoid, in green in Figure 1, and the upper hull is the line connecting the points $(l_x, \alpha l_x)$ and $(u_x, \alpha u_x)$. The lower hull is the line connecting the points $(l_x, \beta l_x)$ and $(u_x, \alpha u_x)$. The maximum altitude is attained at one of the endpoints: in the example in Figure 1 this is u_x because $u_x > |l_x|$. The altitude is then $(\alpha - \beta) \cdot u_x$. The slope λ must lie between the slopes of the directional derivatives

of the upper and lower hulls, yielding an admissible range of $[\frac{\alpha u_x - \beta l_x}{u_x - l_x}, \frac{\beta u_x - \alpha l_x}{u_x - l_x}]$. We choose λ to be the midpoint of this interval. The intercept of the center line may then be chosen such that it passes through the point $(u_x, \frac{\alpha + \beta}{2} u_x)$. We plot the resulting parallelogram in blue.

7. Maximizing Norms over Zonotopes

The final step to upper bounding the Lipschitz constant of a network is to compute a maximization of the $\|\cdot\|_{\alpha^*}$ norm over a zonotope Z , which contains the set of all attainable vector-Jacobian products. While maximizing a convex norm over a convex set may be hard in general, it suffices to upper bound this value. We may always upper bound this norm efficiently by transforming Z into the tightest containing hyperbox and computing the norm over this hyperbox. Any maximal- ℓ_p norm of a hyperbox is efficiently computable, so this technique is quite efficient. However as we typically consider the α to be the ℓ_∞ norm, we focus on techniques to maximize the dual ℓ_1 norm specifically. First we show that this problem is equivalent to computing the Grothendieck problem, i.e. to compute the matrix norm $\|\cdot\|_{\infty \rightarrow 1}$.

Theorem 4. *The problem of computing the maximal ℓ_1 norm of a zonotope is equivalent to the $\|\cdot\|_{\infty \rightarrow 1}$ matrix norm: both problems are NP-hard in general. Additionally, any approximation algorithm with approximation ratio α for the Grothendieck problem will yield an approximation algorithm with ratio α for the zonotope ℓ_1 maximization problem and vice versa.*

theorem The Grothendieck problem is well-studied and it has been shown that the semidefinite relaxation yields an approximation ratio of < 1.783 (Braverman et al., 2013). However, this relaxation may be quite slow. We present a novel result that states that the linear-programming relaxation for the ℓ_1 zonotope norm maximization problem, and equivalently the Grothendieck problem, may be computed in linear time.

Theorem 5. *For a zonotope, $Z(c, E)$, the linear programming relaxation of $\max_{z \in Z(c, E)} \|z\|_1$ is computable in time $O(|E|)$ where $|E|$ denotes the number of elements in E .*

theorem The proof follows from applying the vertical-parallelogram fitting algorithm to the absolute value operator and then solving a linear program over the resulting zonotope.

8. Experiments

We highlight that our algorithm, which we refer to as ZLip, is specifically designed to provide Lipschitz estimates of networks with large output dimension. However, the approaches outlined above are applicable to scalar functions as well. As much of the literature focuses on classifiers, we

first compare our approach on a binary classification task against other Lipschitz estimation techniques. Then we apply ZLip to generative models for MNIST and CIFAR-10. A full description of the experimental details and additional experiments on MNIST and CIFAR-10 classifiers are present in the supplementary.

Toy Network Benchmarks: To fairly compare against existing Lipschitz estimation techniques, we present results on the 2-dimensional Circle dataset from (Aziznejad et al., 2020), where the binary classification is resolved by taking the sign of the output. We consider the $L^{(\infty, 1)}(f, \mathcal{X})$ Lipschitz constant for fully-connected networks f with input dimension 2 and a varying amount of layers of width 100 and the ReLU nonlinearity. We report the average Lipschitz estimate and compute time for the following techniques: Fast-Lip, LipSDP, SeqLip, CLEVER, and LipMIP. Fast-Lip and LipSDP provide provable upper bounds (Weng et al., 2018a; Fazlyab et al., 2019). SeqLip and CLEVER provide heuristic estimates and LipMIP computes this quantity exactly (Virmaux & Scaman, 2018; Weng et al., 2018b; Jordan & Dimakis, 2020). LipMIP leverages interval analysis as a first step, so we also consider a modified version that instead uses the layerwise approximations yielded by ZLip.

In Figure 2, we plot the reported Lipschitz estimate and runtime of these other techniques applied on input regions that are random hyperboxes of ℓ_∞ radius 0.1 centered at elements in the test set. These plots can demonstrate where each technique lies with respect to the efficiency-accuracy tradeoff. In varying the architecture size, we observe that ZLip yields the tightest provable upper bounds for small networks, and only begins to provide looser bounds than LipSDP at 9 hidden layers, at which point LipSDP is three orders of magnitude slower than ZLip. Additionally, using ZLip in place of interval analysis in LipMIP can provide speedups of up to 100x while preserving the exactness of Lipschitz computation.

Generative Models: We now scale our approach to generative models for the MNIST and CIFAR-10 datasets. We train multiple VAEs and GANs using fully-connected and convolutional layers and the ReLU and tanh nonlinearities (Kingma & Welling, 2013; Radford et al., 2015). To evaluate over VAEs, we consider input sets $\mathcal{X}$ that are ℓ_∞ balls surrounding the encodings of images taken from the test set. For GAN evaluation, we consider ℓ_∞ balls surrounding random inputs from the training distribution and evaluate $L^{(\infty, 1)}(f, \mathcal{X})$ of the generator. The only other nontrivial Lipschitz estimation approach that tolerates vector-valued networks and is able to scale to networks of this size is Fast-Lip. Full experimental details are presented in the Appendix, as well as experiments with input sets of different radii.

Table 1 displays results for random inputs with ℓ_∞ radius of

Table 1. Evaluation of ZLip and Fast-Lip on generative models trained on the MNIST and CIFAR-10 datasets, evaluated over inputs of radius 0.05. Times are reported in seconds. The column F/Z denotes the ratio of the estimate returned by Fast-Lip to our estimate. For larger networks, ZLip can yield upper bounds that are several orders of magnitude tighter than Fast-Lip.

Network	MNIST					CIFAR-10				
	ZLip		Fast-Lip		F/Z	ZLip		Fast-Lip		F/Z
	Value	Time	Value	Time		Value	Time	Value	Time	
VAESmall	6.81×10^2	0.831	6.30×10^3	0.0017	10.29	4.07×10^3	4.97	1.13×10^4	0.0029	2.79
VAEMed	2.39×10^3	1.22	1.06×10^6	0.0029	1024	8.28×10^3	5.99	1.06×10^6	0.0042	133.6
VAEBig	6.56×10^3	1.38	5.71×10^7	0.0042	46746	8.39×10^e	5.41	1.74×10^7	0.0054	2227.8
VAECNN	8.75×10^2	1.37	3.58×10^4	0.0028	47.75	5.97×10^3	10.3	6.16×10^4	0.0031	10.38
FFGAN	1.55×10^7	0.455	2.95×10^8	0.0044	52.99	1.24×10^7	1.14	1.21×10^8	0.081	10.01
DCGAN	4.34×10^5	3.46	2.24×10^7	0.0056	55.33	2.11×10^6	8.15	4.63×10^7	0.0127	31.19
VAETanh	3.55×10^3	2.17	7.96×10^4	0.0032	24.19	2.26×10^3	3.90	2.97×10^5	0.0044	132.8
VC-Tanh	2.40×10^3	2.95	7.71×10^4	0.0031	37.25	5.76×10^3	1.97	1.86×10^5	0.0033	32.47

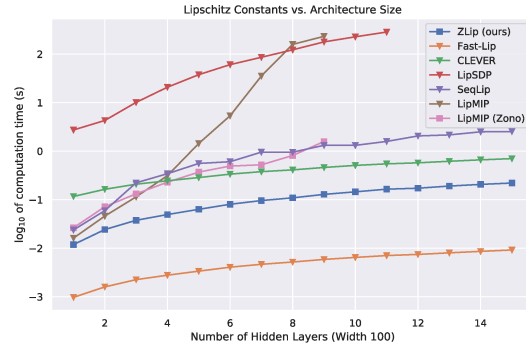
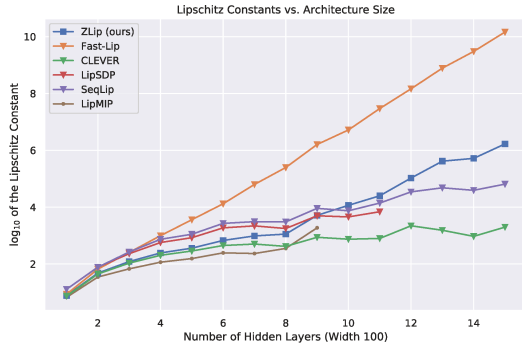


Figure 2. Comparisons of Lipschitz estimate (left) and running time (right) for various Lipschitz estimation techniques on networks trained on the Circle dataset. The horizontal axis is the number of hidden layers of the networks considered and the vertical axis is in log-scale. Our approach is scalable to larger networks while still providing reasonably tight bounds. Also note that using our approach within LipMIP dramatically improves the efficiency.

0.05. We report the average Lipschitz bound and runtime, as well as the ratio of the Fast-Lip value to the ZLip value, denoted F/Z. As the network size scales, we notice that the values reported by both ZLip and Fast-Lip increase. This is likely due to both the true Lipschitz constant of the network increasing as well as the bounds becoming looser. However the bounds provided by ZLip are comparatively much tighter than Fast-Lip for larger networks. We attribute this to the fact that the interval analysis of Fast-Lip introduces error in both the affine and nonlinear layers, whereas our approach only introduces error in the nonlinear layers. While an increase in network size accompanies an increase in runtime for both techniques and ZLip is indeed slower than Fast-Lip, ZLip remains tractable for even the largest networks we consider.

9. Conclusion

We have presented a principled way to efficiently upper bound the Lipschitz constant of neural networks with large output dimension. Our technique improves upon prior works for zonotope approximations of neural networks and is applicable to the operators required by the backpropagation algorithm. This approach yields tighter provable Lipschitz bounds for classifiers and is able to scale effectively to familiar generative models for the MNIST and CIFAR10 datasets, yielding improvements of up to three orders of magnitude for Lipschitz estimation of these networks.

Acknowledgements: This research has been supported by NSF Grants CCF 1763702, 1934932, AF 1901292, 2008710, 2019844 research gifts by Western Digital, WNCG IAP, computing resources from TACC and the Archie Straiton Fellowship.

References

- Aziznejad, S., Gupta, H., Campos, J., and Unser, M. Deep neural networks with trainable activations and controlled lipschitz constant. January 2020.
- Braverman, M., Makarychev, K., Makarychev, Y., and Naor, A. THE GROTHENDIECK CONSTANT IS STRICTLY SMALLER THAN KRIVINE’S BOUND. *Forum of Mathematics, Pi*, 1, 2013.
- Chen, T., Lasserre, J.-B., Magron, V., and Pauwels, E. Semi-algebraic optimization for lipschitz constants of ReLU networks. February 2020.
- Fazlyab, M., Robey, A., Hassani, H., Morari, M., and Pappas, G. Efficient and accurate estimation of lipschitz constants for deep neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 11423–11434. Curran Associates, Inc., 2019.
- Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. May 2017.
- Jordan, M. and Dimakis, A. G. Exactly computing the local lipschitz constant of ReLU networks. March 2020.
- Jordan, M., Lewis, J., and Dimakis, A. G. Provable certificates for adversarial examples: Fitting a ball in the union of polytopes. March 2019.
- Kingma, D. P. and Welling, M. Auto-Encoding variational bayes. December 2013.
- Latorre, F., Rolland, P., and Cevher, V. Lipschitz constant estimation of neural networks via sparse polynomial optimization. September 2019.
- Mirman, M., Gehr, T., and Vechev, M. Differentiable abstract interpretation for provably robust neural networks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3578–3586, Stockholmsmässan, Stockholm Sweden, 2018. PMLR.
- Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. November 2015.
- Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. T. Fast and effective robustness certification. *NeurIPS*, 1(4):6, 2018.
- Singh, G., Gehr, T., Püschel, M., and Vechev, M. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL):41:1–41:30, January 2019a.
- Singh, G., Gehr, T., Püschel, M., and Vechev, M. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, January 2019b.
- Tran, H.-D., Bak, S., Xiang, W., and Johnson, T. T. Verification of deep convolutional neural networks using ImageStars. In *Computer Aided Verification*, pp. 18–42. Springer International Publishing, 2020.
- Virmaux, A. and Scaman, K. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 3835–3844. Curran Associates, Inc., 2018.
- Weng, T.-W., Zhang, H., Chen, H., Song, Z., Hsieh, C.-J., Boning, D., Dhillon, I. S., and Daniel, L. Towards fast computation of certified robustness for ReLU networks. April 2018a.
- Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., and Daniel, L. Evaluating the robustness of neural networks: An extreme value theory approach. January 2018b.
- Xu, K., Zhang, H., Wang, S., Wang, Y., Jana, S., Lin, X., and Hsieh, C.-J. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. November 2020.
- Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. Efficient neural network robustness certification with general activation functions. November 2018.
- Zico Kolter, J. and Wong, E. Provable defenses against adversarial examples via the convex outer adversarial polytope. November 2017.

A. Proofs of Lemmas and Theorems

Theorem 1. For feedforward neural networks f , an input set $\mathcal{X}$ and sets $\mathcal{Z}_i, \hat{\mathcal{Z}}_i, \mathcal{J}_i, \mathcal{Y}_i, \hat{\mathcal{Y}}_i$ satisfying the containments in Equations 14-17, the set of vector-Jacobian products satisfies

$$\{\nabla_x f(x)^T u \mid x \in \mathcal{X} \quad u \in B_{\beta^*}\} \subseteq \mathcal{Y}_0. \quad (18)$$

For such a $\mathcal{Y}_0$, the Lipschitz constant of f may be upper-bounded by maximizing the $\|\cdot\|_{\alpha^*}$ norm over the set $\mathcal{Y}_0$.

Proof. Suppose $\mathcal{Z}_i, \hat{\mathcal{Z}}_i, \mathcal{J}_i, \mathcal{Y}_i, \hat{\mathcal{Y}}_i$ satisfy the containments in equation 14-17. Now consider any $x \in \mathcal{X}$ and $u \in B_{\beta^*}$. The proof follows from repeated applications of the following statement: for any function g , if $\mathcal{A} \subseteq \mathcal{B}$, then $g(\mathcal{A}) \subseteq g(\mathcal{B})$. We iteratively apply this statement to the forward recursion to see that $\hat{Z}_i(x) \in \hat{\mathcal{Z}}_i$ for all i , and similarly for $Z_i(x) \in \mathcal{Z}_i$. From equation 17, $J_{i+1}^T(x) \in \mathcal{J}_{i+1}$ for all i . We may now perform the backward recursion to see that $Y_i(x, u) \in \mathcal{Y}_i$ and similarly for $\hat{Y}_i(x, u), \hat{\mathcal{Y}}_i$. Repeating this for all i yields the desired result. $\square$

Lemma 1. For any zonotope $Z \subset \mathbb{R}^d$ and any operator Φ operating over $\mathbb{R}^d$, if $\hat{Z}$ is a zonotope satisfying the containment

$$\{\Phi(z) - \Lambda \odot z \mid z \in Z\} \subseteq \hat{Z} \quad (19)$$

then

$$\Phi(Z) \subseteq (\Lambda \odot Z) \oplus \hat{Z}.$$

Proof. By assumption, for every $z \in Z$, there exists a $\hat{z} \in \hat{Z}$ such that $\Phi(z) - \Lambda \odot z = \hat{z}$. This implies that, $\Phi(z) = (\Lambda \odot z) + \hat{z}$. By definition,

$$(\Lambda \odot Z) \oplus \hat{Z} := \{\Lambda \odot z + \hat{z} \mid z \in Z \quad \hat{z} \in \hat{Z}\},$$

so $\Phi(z) \in (\Lambda \odot Z) \oplus \hat{Z}$ for every $z \in Z$. $\square$

Theorem 2. When S is the set $\{(z_i, \phi(z_i)) \mid z_i \in [l_i, u_i]\}$, the solution to the vertical parallelogram fitting problem yields the optimal solution to Equation 20. Repeated calls to this subroutine yields the tightest hyperbox fitting the residuals as in equation 19.

Proof. Suppose that $\{\Lambda_i, b_i, \mu_i\}_{i=1}^d$ is the set of solutions to the vertical parallelogram fitting problem for each set $S_i = \{(z_i, \phi(z_i)) \mid z_i \in [l_i, u_i]\}$. Since the coordinate-wise bounds l_i, u_i are chosen such that $l \leq z \leq u$ for all $z \in Z$, the containment holds:

$$\{\Phi(z) - \Lambda \odot z - b \mid z \in Z\} \subseteq \{\Phi(z) - \Lambda \odot z - b \mid l \leq z \leq u\}.$$

By definition of solutions to the vertical parallelogram fitting problem, the set of vectors $\{\Phi(z) - \Lambda \odot z - b \mid l \leq z \leq u\}$ is contained in the hyperbox $H(0, \mu)$. Adding b to each element of each set, we see that $\{\Phi(z) - \Lambda \odot z \mid l \leq z \leq u\}$ is contained in the hyperbox $H(b, \mu)$, thus satisfying the assumptions of Lemma 1. $\square$

Lemma 2. For any zonotope $Z \subseteq \mathbb{R}^d$ and hyperbox $H \subseteq \mathbb{R}^d$, if $\hat{Z}$ is a zonotope satisfying the containment

$$\{x \odot z - \Lambda \odot z \mid x \in H \quad z \in Z\} \subseteq \hat{Z} \quad (21)$$

then

$$\{x \odot z \mid x \in H \quad z \in Z\} \subseteq (\Lambda \odot Z) \oplus \hat{Z}.$$

Proof. By assumption, for every $z \in Z$ and every $x \in H$, there exists a $\hat{z} \in \hat{Z}$ such that $x \odot z - \Lambda \odot z = \hat{z}$. This implies that, $x \odot z = (\Lambda \odot z) + \hat{z}$. And by definition

$$(\Lambda \odot Z) \oplus \hat{Z} := \{\Lambda \odot z + \hat{z} \mid z \in Z \quad \hat{z} \in \hat{Z}\},$$

so $x \odot z \in (\Lambda \odot Z) \oplus \hat{Z}$ for every $z \in Z$. $\square$

Theorem 3. When S is the set $\{(z, x \cdot z) \mid l^{(z)} \leq z \leq u^{(z)} \quad l^{(x)} \leq x \leq u^{(x)}\}$, the solution to the vertical-parallelogram fitting problem yields the optimal solution to Equation 22. Repeated calls to this subroutine yields the tightest hyperbox fitting the residuals as in Equation 21.

Proof. Suppose that $\{(\Lambda_i, b_i, \mu_i)\}_{i=1}^d$ is the set of solutions to the vertical parallelogram fitting problem for each set $S_i = \{(z_i, x_i \odot z_i) \mid l_i^{(z)} \leq z_i \leq u_i^{(z)}, \quad l_i^{(x)} \leq x_i \leq u_i^{(x)}\}$. Let H be the hyperbox with lower and upper-bounds denoted by $l^{(x)}, u^{(x)}$. Since the coordinate-wise bounds l_i, u_i are chosen such that $l^{(z)} \leq z \leq u^{(z)}$ for all $z \in Z$, the containment holds:

$$\{x \odot z - \Lambda \odot z - b \mid z \in Z \quad x \in H\} \subseteq \{\Phi(z) - \Lambda \odot z - b \mid l^{(z)} \leq z \leq u^{(z)} \quad x \in \mathcal{H}\}.$$

By definition of solutions to the vertical parallelogram fitting problem, the set of vectors $\{x \odot (z) - \Lambda \odot z - b \mid l^{(z)} \leq z \leq u^{(z)} \quad x \in H\}$ is contained in the hyperbox $H(0, \mu)$. Adding b to each element of each set, we see that $\{x \odot z - \Lambda \odot z \mid l \leq z \leq u \quad x \in H\}$ is contained in the hyperbox $H(b, \mu)$, thus satisfying the assumptions of Lemma 2. $\square$

Theorem 4. *The problem of computing the maximal ℓ_1 norm of a zonotope is equivalent to the $\|\cdot\|_{\infty \rightarrow 1}$ matrix norm: both problems are NP-hard in general. Additionally, any approximation algorithm with approximation ratio α for the Grothendieck problem will yield an approximation algorithm with ratio α for the zonotope ℓ_1 maximization problem and vice versa.*

Proof. We prove this via a strict reduction in showing that any instance of one problem may be converted into an instance of the other and will keep the same optimal value. To do this, we first note that for any matrix M , with $(0||M)$ denoting the zero-column prepended to the columns of M , that $\|M\|_{\infty \rightarrow 1} = \|(0||M)\|_{\infty \rightarrow 1}$. This follows since

$$\|(0||M)\|_{\infty \rightarrow 1} = \max_{v \in B_\infty} \|(0||M)v\|_1 = \max_{u \in B_\infty} \|Mu\|_1 = \|M\|_{\infty \rightarrow 1}.$$

Next, it suffices to show that

$$\max_{z \in Z(c, E)} \|z\|_1 = \|(c||E)\|_{\infty \rightarrow 1}, \quad (23)$$

for if this were true, certainly any zonotope could be reduced to a matrix-norm maximization problem, and any matrix norm problem could first prepend the zero column to the matrix and be reduced to a zonotope norm-maximization problem. Any α -approximation algorithm for one problem could provide an α -approximation for any instance of the other via this reduction.

First we show that $\max_{z \in Z(c, E)} \|z\|_1 \leq \|(c||E)\|_{\infty \rightarrow 1}$. As the right-hand-side may be written

$$\|(c||E)\|_{\infty \rightarrow 1} = \max_{|v_0| \leq 1} \max_{v \in B_\infty} \|v_0 \cdot c + Ev\|_1$$

and whereas the left-hand side of Equation 23 is the same optimization with v_0 restricted to 1. Therefore the $(\leq)$ direction of Equation 23 holds. For the other direction, consider any integral solution to the RHS,

$$(v_0^*, v^*) \in \arg \max_{|v_0| \leq 1, v \in B_\infty} \|v_0 \cdot c + Ev\|_1.$$

Without loss of generality, v_0^* may be chosen to be 1, and the point $(c + Ev^*)$ is in $Z(c, E)$. Hence there's a point in $Z(c, E)$ with ℓ_1 norm at least that of $\|(c||E)\|_{\infty \rightarrow 1}$, thus proving the $(\geq)$ direction of equality 23. $\square$

Theorem 5. *For a zonotope, $Z(c, E)$, the linear programming relaxation of $\max_{z \in Z(c, E)} \|z\|_1$ is computable in time $O(|E|)$ where $|E|$ denotes the number of elements in E .*

Proof. First we write down the Linear-programming relaxation of the zonotope-norm maximization problem and then relate this to the mapping of the zonotope through the absolute value operator, by our vertical-parallelogram fitting procedure. The final result follows from linear programs being efficiently solvable over zonotopes.

Consider some zonotope $Z(c, E) \subseteq \mathbb{R}^d$ which has coordinate-wise upper and lower bounds $[l_i, u_i]$ for every $i \in [d]$. We partition the coordinates into three sets of indices: S^-, S^+, S such that $S^- := \{i \mid u_i \leq 0\}$, $S^+ := \{i \mid l_i > 0\}$ and S is the set of indices not in either S^- or S^+ . We may write down the familiar mixed-integer programming relaxation for the

absolute value operator by introducing $|S|$ continuous variables, $\{t_i\}_{i \in S}$, and d integer variables $\{a_i\}_{i \in S}$, where $t_i \in \mathbb{R}$ and $a_i \in \{0, 1\}$:

$$\max \sum_{i \in S} t_i - \sum_{i \in S^-} z_i + \sum_{i \in S^+} z_i \quad (24)$$

$$t_i \geq z_i \quad (25)$$

$$t_i \geq -z_i \quad (26)$$

$$t_i \leq -z_i + 2 \cdot u_i \cdot a_i \quad (27)$$

$$t_i \leq z_i - 2 \cdot l_i \cdot (1 - a_i) \quad (28)$$

$$a_i \in \{0, 1\} \quad (29)$$

$$z \in Z(c, E) \quad (30)$$

Where the constraints enforce that $t_i = |z_i|$. The first two constraints require that $t_i \geq |z_i|$. To show $t_i \leq |z_i|$, we proceed by cases. When $z_i > 0$, then 27 implies that $a_i = 1$, for otherwise $t_i < 0$ contradicting the first constraint. This causes 28 to imply $t_i \leq z_i$. When $z_i < 0$, 28, $a_i = 0$, for otherwise 27 again implies that $t_i < 0$. This causes 27 to imply $t_i \leq -z_i$. When $z_i = 0$, either case can hold and $t_i = 0$. The linear programming relaxation lets a_i be in the range $[0, 1]$ instead of $\{0, 1\}$.

For any fixed z_i , we can compute the maximum value of t_i under this relaxation, which is a function of the now-continuous variable, a_i . By setting the upper bounds to equality, the optimal value of a_i is $a_i = \frac{z_i - l_i}{u_i - l_i}$ and t is then upper bounded by

$$t_i \leq \frac{-z_i + 2u_i \cdot (z_i - l_i)}{u_i - l_i}$$

. We observe that this is an equivalent relaxation to the upper-hull provided by the absolute value operator and our vertical-parallelogram fitting procedure (next section). This allows us to rewrite the optimization above as

$$\max_{z \in Z(c, E)} \sum_{i \in S} \frac{-z_i + 2u_i \cdot (z_i - l_i)}{u_i - l_i} - \sum_{i \in S^-} z_i + \sum_{i \in S^+} z_i$$

which we notice is a linear program over a zonotope. The objective vector may be developed in $O(d)$ time, and linear programs may be solvable over zonotopes in $O(|E|)$ time. $\square$

B. Pseudocode

Algorithm 1 ZLip

Require: L -layer feedforward neural network f , input set $\mathcal{X}$, norm β

Returns: Zonotope $\mathcal{Y}_0 \supseteq \{\nabla_x f(x)^T v \mid x \in \mathcal{X}, v \in B_{\beta^*}\}$

```

function ZLIP( $f, \mathcal{X}, \beta$ )
     $\mathcal{Z}_0 \leftarrow \text{Zonotope}(\mathcal{X})$                                 ▷ Cast input set to zonotope
    for  $i \leftarrow 1$  to  $L$  do                                    ▷ Forward pass (e.g., DeepZ)
         $\hat{\mathcal{Z}}_i \leftarrow \text{map\_affine}(W_i, b_i, \mathcal{Z}_{i-1})$ 
         $\mathcal{Z}_i \leftarrow \text{map\_nonlin}(\sigma, \hat{\mathcal{Z}}_i)$ 
         $\mathcal{J}_i \leftarrow \text{elementwise\_jacobian}(\sigma, \mathcal{Z}_i)$         ▷ Gradient range for  $\nabla_z \sigma(\mathcal{Z}_i)$ 
    end for
     $\hat{\mathcal{Y}}_L \leftarrow \text{Zonotope}(\mathcal{B}_{\beta^*})$                         ▷ Cast dual ball to zonotope
    for  $i \leftarrow L$  to 1 do                                    ▷ Backward pass
         $\mathcal{Y}_{i-1} \leftarrow \text{map\_affine}(W_i^T, 0, \hat{\mathcal{Y}}_i)$ 
         $\hat{\mathcal{Y}}_{i-1} \leftarrow \text{elementwise\_mul}(\mathcal{J}_i, \mathcal{Y}_{i-1})$ 
    end for
     $\hat{\mathcal{Y}}_0 \leftarrow \text{map\_affine}(W_0^T, 0, \hat{\mathcal{Y}}_0)$ 
    return  $\hat{\mathcal{Y}}_0$ 
end function

```

Algorithm 2 Vertical Parallelogram Fitting

Require: Function $\sigma : \mathbb{R} \rightarrow \mathcal{P}(\mathbb{R})$, and interval $[c - |E|, c + |E|]$

Returns: Slope Λ^* , Altitude μ^* , center b^*

```

function VP_FIT( $\sigma, c, E$ )
     $\mathcal{I} \leftarrow c \pm |E|$ 
     $S \leftarrow \{(x, \sigma(x)) \mid x \in \mathcal{I}\}$ 
     $h^-, h^+ \leftarrow \text{conv\_hull}(S)$                                 ▷ Possibly hard, depends on  $\sigma$ 
     $x^* \leftarrow \text{argmax}_{x \in \mathcal{I}} h^+(x) - h^-(x)$ 
     $\mu^* \leftarrow h^+(x^*) - h^-(x^*)$                                 ▷ Altitude
     $\Lambda^* \leftarrow \delta(h^-(x^*)) \cap \delta(-h^+(x^*))$                 ▷ Slope of parallelogram's non-vert side
     $b^* \leftarrow \frac{1}{2} \cdot (h^+(x^*) + h^-(x^*))$                     ▷ Intercept
    return  $\Lambda^*, \mu^*, b^*$ 
end function

```

Algorithm 3 Auxiliary Functions

```

function MAP_NONLIN( $\sigma, Z(c, E)$ )  $\triangleright Z(c, E) \subseteq \mathbb{R}^d, \sigma$  is elementwise
    for  $i \leftarrow 1$  to  $d$  do
         $\Lambda_i, b_i^*, \mu_i^* \leftarrow \text{VP\_Fit}(\sigma_i, c_i, E_i^T)$ 
    end for
    return  $Z(b^*, \text{diag}(\mu^*)) \oplus (\Lambda \odot Z(c, E))$ 
end function

function ELEMENTWISE_MUL( $H(l, u), Z(c, E)$ )
    for  $i \leftarrow 1$  to  $d$  do
         $\Lambda_i, b_i^*, \mu_i^* \leftarrow \text{VP\_Fit}([l_i, u_i], c_i, E_i^T)$   $\triangleright$  Overloading VP_Fit signature
    end for
    return  $Z(b^*, \text{diag}(\mu^*)) \oplus (\Lambda \odot Z(c, E))$ 
end function
    
```

C. Detailed Derivations for Vertical Parallelogram Fitting

We recall the algorithm from Section 6 for fitting a vertical parallelogram to a 2-dimensional set S . The first step was to compute the upper and lower convex hulls of S . For sets of the form $\{(x, f(x)) \mid x \in [l, u]\}$ for some differentiable function, this is equivalent to the biconjugate and the biconjugate of the negation of f . In this case, there exists a simple algorithm to yield the upper-convex hull. The lower-convex hull may be found the same way, but for the set $\{(x, -f(x))\}$. Observe that if f is convex over $[l, u]$, then the upper convex hull is the secant line between the endpoints $(l, f(l))$ and $(u, f(u))$ and the lower-convex hull is $f(x)$; vice versa for concave functions. For functions that are neither convex or concave over $[l, u]$, the upper convex hull may be piecewise continuous, alternating between secant-line segments and f . For function f , let the secant line of f between x_1 and x_2 be denoted as $\text{Sec}^f(x_1, x_2)$.

We take inspiration from the gift-wrapping procedure for finding convex hulls of finite 2-dimensional point sets. In gift-wrapping, the idea is to find the left-most point in the set and sweep a ray clockwise until an intersection with another point in the set is found. The sweep continues, with ray now starting at the newly intersected point until the left-most point is intersected again. Our procedure sweeps performs a sweep over rays of decreasing slope, noting that any intersecting point must lie on the set S and thus the ray is a secant line. Hence, the key subroutine to find the upper hull is to solve a maximization over slopes of secant lines. For a fixed x_0 , the slope of the secant line $\text{Sec}^f(x_0, x)$ is $\frac{f(x) - f(x_0)}{x - x_0}$, and the maximization we seek to solve is a constrained variant of this,

$$\max_{x \in [x_0, u]} \frac{f(x) - f(x_0)}{x - x_0}. \quad (31)$$

When f is differentiable, then we can differentiate the above objective and set to zero and solve for x in the equality

$$f'(x) = \frac{f(x) - f(x_i)}{x - x_i} \quad (32)$$

This procedure may be repeated until the max is attained at an $x \geq u_i$, for which the final secant line spans between $(x_i, f(x_i))$ and $(u, f(u))$.

Once piecewise forms for the convex upper and lower hulls are formed, the maximal altitude can be computed by maximizing the piecewise function $h^+(x) - h^-(x)$. The proper slope for the tightest fitting vertical parallelogram is attained by considering an element in the intersection of subgradients of h^- and $-h^+$ at their maximal altitude.

C.1. Sigmoid/Tanh

We demonstrate the above procedure for finding convex upper and lower hulls of sets $\{(x, f(x)) \mid x \in [l, u]\}$ where f is S -shaped like sigmoid and tanh. We say a function f is S -shaped if it is monotonically increasing and there exists an x' such that f is concave for all $x \geq x'$. We break this into cases, based on the values of $[l, u]$. If f is either convex or concave over the entire interval $[l, u]$, then the upper hull is either the secant line or f respectively, and vice versa for the lower hull. In this case, the maximum altitude is attained at the x where $f'(x)$ is equal to the slope of this secant line.

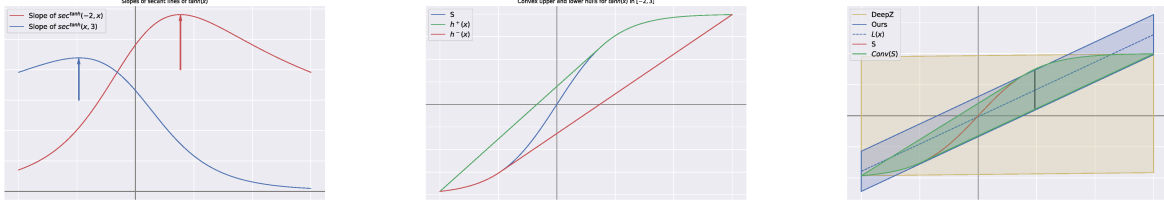


Figure 3. Procedure for computing the tightest fitting vertical-parallelgram to the set $S = \{(x, \tanh(x)) \mid x \in [-2, 3]\}$. (Left) We plot the slope of the secant-line from -2 to x in red and the secant line from x to 3 in blue, where the maximum is marked with a vertical hash. (Middle) We plot the upper and lower convex hulls for S . (Right) We plot the convex hull of S and the parallelgram we produce versus that of prior work.

For cases where f is both convex and concave for portions of $[l, u]$, then we proceed with the iterative giftwrap procedure, starting at $x_0 = l$. For monotonically increasing $\tanh$ and sigmoid functions, this means that $l < 0 < u$ and we seek to find a point in the concave ($x \geq 0$) portion of f such that Equation 32 is satisfied. Since f is concave for $x \geq 0$, the function

$$f'(x) - \frac{f(x) - f(l)}{x - l}$$

only has one zero for $x \geq 0$, which may be found numerically¹. Letting x^* be such an x , we have that the secant-line is given by $y = \frac{f(x^*) - f(l)}{x^* - l}(x - l) + f(l)$ and the upper convex hull for is

$$h^+(x) = \begin{cases} \frac{f(x^*) - f(l)}{x^* - l}(x - l) + f(l) & \text{if } x \in [l, x^*] \\ f(x) & \text{if } x \in [x^*, u] \end{cases}$$

A similar procedure may be considered for the lower convex hull, also yielding a convex hull like

$$h^-(x) = \begin{cases} f(x) & \text{if } x \in [l, x^\dagger] \\ \frac{f(x^\dagger) - f(l)}{x^\dagger - l}(x - l) + f(l) & \text{if } x \in [x^\dagger, u] \end{cases}$$

where $x^\dagger$ is the minimum amongst zeros for the function $f'(x) - \frac{f(x) - f(u)}{x - u}$. Then the altitude, $h^+(x) - h^-(x)$, is a concave piecewise function with three pieces, segmented into the intervals $[l, x^\dagger]$, $[x^\dagger, x^*]$, $[x^*, u]$. If the slope of the linear component of the upper hull is λ^+ and the slope of the linear component of the lower hull is λ^- , then the maximum is attained at one of four points: i) the point in $[l, x^\dagger]$ where $f'(x) = \lambda^-$; ii) the point in $[x^*, u]$ where $f'(x) = \lambda^+$; iii) x^* , or iv) $x^\dagger$. It is trivial to check all these points where f is the sigmoid or $\tanh$ function.

Regardless of which case we are in, once the upper and lower hulls have been computed and their difference has been maximized, the remaining steps are simple. It suffices to compute the subgradients of the lower hull and negative upper hull at that point, pick an element of their intersection and find the line that passes through the proper midpoint. This yields the right altitude and affine function for a vertical parallelgram.

C.2. Elementwise Multiplication

Now we consider the full set of cases for elementwise multiplication. Letting S be the 2-dimensional set $\{(x, y \cdot x) \mid x \in [l, u], y \in [\alpha, \beta]\}$. When $\alpha = \beta$, then the set S is a line segment, and has vertical parallelgram with height 0 and slope α . When $\alpha < \beta$, and $l > 0$ the convex upper hull is $h^+(x) = \beta x$ and the convex lower hull is $h^-(x) = \alpha x$; vice versa for when $u < 0$. The only remaining case is when $\alpha < \beta$, and $l < 0 < u$, which is the case considered in the main paper, where $h^+(x)$ is the line passing through $(l, \alpha l)$ and $(u, \beta u)$; and the lower hull passes through $(l, \beta l)$ and $(u, \alpha u)$.

In all of the above cases, the upper and lower hulls are affine functions and their maximum over $[l, u]$ is attained at either l or u . The admissible slopes are exactly the range $[\alpha, \beta]$.

C.3. Absolute Value

¹We keep track of the numerical error and ensure that the vertical height of the parallelgram accounts for this.

Now we consider the absolute value function, as used in the proof of Theorem 5. Consider the set $\{(x, |x|) \mid x \in [l, u]\}$. When $l \cdot u \geq 0$, then the set S lies on either the $y = x$ or $y = -x$ line, and the vertical parallelogram fit is equivalent to mapping through an affine function. On the other hand, when $l < 0 < u$, the set S has a lower convex hull of $h^-(x) = |x|$, because $|x|$ is a convex function. The upper hull is the secant line connecting $(l, |l|)$ and $(u, |u|)$, which may be written as $h^+(x) = \frac{u+l}{u-l}(x) - \frac{2ul}{u-l}$. The maximum of the altitude, $h^+(x) - h^-(x)$, is attained at $x = 0$, for which the maximum altitude is $-\frac{2ul}{u-l}$ and the height of the vertical parallelogram is $\mu = \frac{-ul}{u-l}$. The slope must be the slope of the (linear) upper hull, $\frac{u+l}{u-l}$ and the central line of the parallelogram must pass through the point $(0, \frac{-ul}{u-l})$.

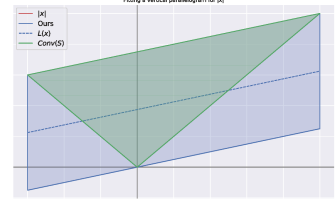


Figure 4. Convex hull and vertical parallelogram for the set $S = \{(x, |x|) \mid x \in [-3, 5]\}$.

D. Experiments

D.1. Model Architectures

Here we will describe the structure of each of the architectures considered. For networks with only fully connected and elementwise nonlinearities, we denote the architectures by $[n_{in}, n_1, \dots, n_L \dots n_{out}]$, where n_i denotes the number of neurons in the i^{th} hidden layer, and ReLU nonlinearities are implied between each layer. We will use the notation “FC X” to denote fully connected layers with an output of X neurons, and $\text{Conv}_s(C \times W \times H)$ to denote convolutional layers with a stride of s and output dimension of C channels, and kernels of size $W \times H$. Transpose convolutional layers are denoted $\text{Conv}_s^T(C \times W \times H)$. For layers of the same size repeated k times, we’ll denote this as $[\text{layers}]^{\times k}$.

Toy Networks: For the networks trained on the toy dataset, the input and output dimension are each 2, and the scalar-valued output is attained by taking the dot product with the vector $[+1, -1]$. These have varying depth, but all have architectures like

$$x \rightarrow \left[FC100 \rightarrow ReLU \right]^{\times k} \rightarrow FC2 \rightarrow z$$

where the number of hidden layers denotes the number of ReLU layers in the network.

Generators for MNIST and CIFAR: For MNIST and CIFAR, we trained 6 VAEs and 2 GANs. These each have the same architecture with the exception of varying input/output shapes. The VAEs each have a latent dimension of 20 and the GANs have an input dimension of 100. We use the notation D and C to denote the output dimension and channel: (784, 1) for MNIST and (3072, 3) for CIFAR-10.

- **VAESmall:** $[D, 400, 200, 20, 200, 400, D]$. Where the decoder is just the $[20, 200, 400, D]$ subnetwork.
- **VAEMed:** $[D, 400, 200, 100, 50, 100, 200, 400, D]$ where the decoder is just the $[50, 100, 200, 400, D]$ subnetwork.
- **VAEBig:** $[D, 400, 200, 200, 200, 200, 100, 200, 200, 200, 200, 400, D]$, where the decoder is just the $[100, 200, 200, 200, 200, 400, D]$ subnetwork.
- **VAECNN:** $x \rightarrow \text{Conv}_2(16 \times 4 \times 4) \rightarrow ReLU \rightarrow \text{Conv}_2(32 \times 4 \times 4) \rightarrow ReLU \rightarrow FC50 \rightarrow FC800 \rightarrow ReLU \rightarrow \text{Conv}_2^T(16 \times 5 \times 5) \rightarrow ReLU \rightarrow \text{Conv}_2^T(C \times 4 \times 4) \rightarrow Sigmoid$.
- **VAETanh:** Same as VAEMed with $\tanh$ nonlinearities in place of ReLU.
- **VC-Tanh:** Same as VAECNN with $\tanh$ nonlinearities in place of ReLU.
- **FFGAN:** $[100, 256, 512, 1024, D]$.
- **DCGAN:** $x \rightarrow \text{Conv}_1^T(256 \times 4 \times 4) \rightarrow ReLU \rightarrow \text{Conv}_2^T(128 \times 4 \times 4) \rightarrow ReLU \rightarrow \text{Conv}_2^T(64 \times 4 \times 4) \rightarrow ReLU \rightarrow \text{Conv}_2^T(C \times 4 \times 4) \rightarrow \tanh$.

Classifiers for MNIST and CIFAR: For MNIST and CIFAR, we trained three fully connected networks. We refer to these as $\{tiny, small, med, \} - *$. Each of these were trained with both the ReLU and tanh nonlinearities with both standard and PGD adversarial training. We will describe the training techniques in the next section.

- **Tiny*:** $[D, 20, 20, 10]$.
- **Small*:** $[D, 100, 100, 100, 10]$.
- **Med*:** $[D] + [100] \times 6 + [10]$.

D.2. Datasets, Training Methods, and Computing Environment

Computing Environment: All networks were trained using Pytorch on a machine with $2 \times$ GeForce RTX 2070 GPU’s. All Lipschitz evaluations were performed using the CPU only, an Intel i7-9700K. Mixed integer programming evaluations were performed using 4 cores and the Gurobi optimizer.

Datasets: The CIFAR-10 and MNIST datasets are standard. For the dataset taken from (Aziznejad et al., 2020), we generate 1000 points uniformly randomly in $[-1, 1]^2$ and attach the label 1 if the norm of the data point is $\leq \frac{2}{\pi}$.

Training Methods: We outline the methods used to train each set of networks:

- **Toy Dataset Nets:** We trained using the CrossEntropy loss and the Adam optimizer with a learning rate of 0.001, where we trained for 200 epochs.
- **VAEs:** For the VAE networks, we train with the standard VAE loss: a sum of binary cross-entropy between the reconstructed example and original input, and a KL-divergence term. We train for 50 epochs for all VAEs with the Adam optimizer and a learning rate of 0.001
- **GANs:** For the GANs, we train using the binary cross-entropy applied to the discriminator output. We train with Adam ($lr = 0.001$) and 25 epochs.
- **Classifiers:** For the classifiers considered, we always train for 50 epochs using Adam ($lr = 0.001$) and either the standard Cross-Entropy loss or the PGD loss for adversarial training. For MNIST networks, we allow an adversarial budget of ℓ_∞ norm of 0.1, and the adversary takes 10 steps with step-size 0.2. For CIFAR-10, the adversary has a budget of $\frac{8}{255}$ and takes 10 steps of stepsize $\frac{2}{255}$.

D.3. Varying Radius

We consider the effect of varying the radius of the region we evaluate Lipschitz constants over. We expect that as the radius increases, the bounds for the preactivations at each layer will become more loose. For ReLU networks, when zero is strictly contained in the preactivation bounds, a new degree of freedom needs to be introduced to the layerwise zonotope approximations. Hence, as more degrees of freedom are introduced, we expect the bound returned by ZLip to be looser and the runtime will be less efficient since the representation of each zonotope will be larger. However, other methods will likely also yield looser bounds as well. We also note that a more intelligent strategy that prunes the zonotope representation size to a more manageable size may be employed, though we leave such performance improvements for future work.

Toy Dataset: First we examine the effect of changing the evaluation radius on the network trained on the circle dataset with 6 hidden layers of width 100 and the ReLU nonlinearity. We report the results in table 2, where we have evaluated over 64 elements from the test set and ℓ_∞ balls of each radius. We set a timeout of 120s for the mixed-integer programming approaches, at which point the tightest upper bound is returned, explaining the discrepancy between the results on the first two columns. The proportion of timed out examples is presented in Figure 5. We observe that for networks with small input dimension, CLEVER is fairly accurate and can be viewed as a surrogate for the true Lipschitz constant when the LipMIP results time out. In this case we see that for all but the global evaluation returned by SeqLip, all techniques provide looser bounds as the radius increases, however ZLip remains significantly tighter than Fast-Lip and even yields a tighter result than the tightest upper bound provided by LipMIP after this procedure times out.

Generative Models: Here we present results on the generative models when we vary the radius of the input region. We focus primarily on the VAEs described above for both the MNIST and CIFAR-10 datasets. These results for VAEs on MNIST are displayed in Figure 6. As expected, increasing radius size increases the estimated Lipschitz bound for both ZLip and Fast-Lip. The relative gap decreases as the radius increases, indicating that ZLip is comparatively tighter when few neurons are unstable, which we attribute to ZLip being able to perfectly map the affine layers. It is also expected that Fast-Lip does not get slower even as more neurons become unstable, as the representation size of a hyperbox in $\mathbb{R}^d$ will always be $O(d)$, and looser neuron bounds just increases the looseness of the approximation without increasing runtime.

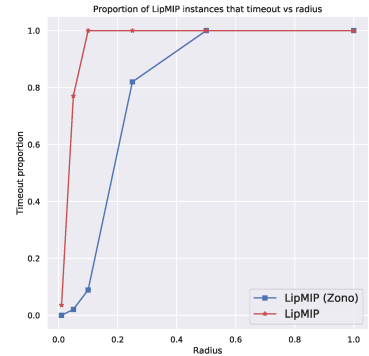


Figure 5. Proportion of timed out examples for LipMIP versus LipMIP using ZLip as a first step across various radii.

Table 2. Evaluation of various Lipschitz computation techniques for the network with 6 hidden layers trained on the Circle dataset.

Lipschitz Values for 6×100 Circle Network						
Radius	LipMIP (Zono)	LipMIP	ZLip	Fast-Lip	SeqLip	CLEVER
0.01	2.62×10^2	2.97×10^2	2.98×10^2	2.82×10^3	2.72×10^3	2.95×10^2
0.05	3.69×10^2	2.79×10^3	4.22×10^2	9.94×10^3	2.72×10^3	3.61×10^2
0.1	4.55×10^2	6.09×10^3	6.17×10^2	1.29×10^4	2.72×10^3	4.39×10^2
0.25	8.91×10^2	1.67×10^4	1.55×10^3	2.00×10^4	2.72×10^3	4.65×10^2
0.5	2.51×10^3	2.20×10^4	3.70×10^3	2.46×10^4	2.72×10^3	6.72×10^2
1.0	6.49×10^3	2.82×10^4	5.81×10^3	3.10×10^4	2.72×10^3	9.04×10^2

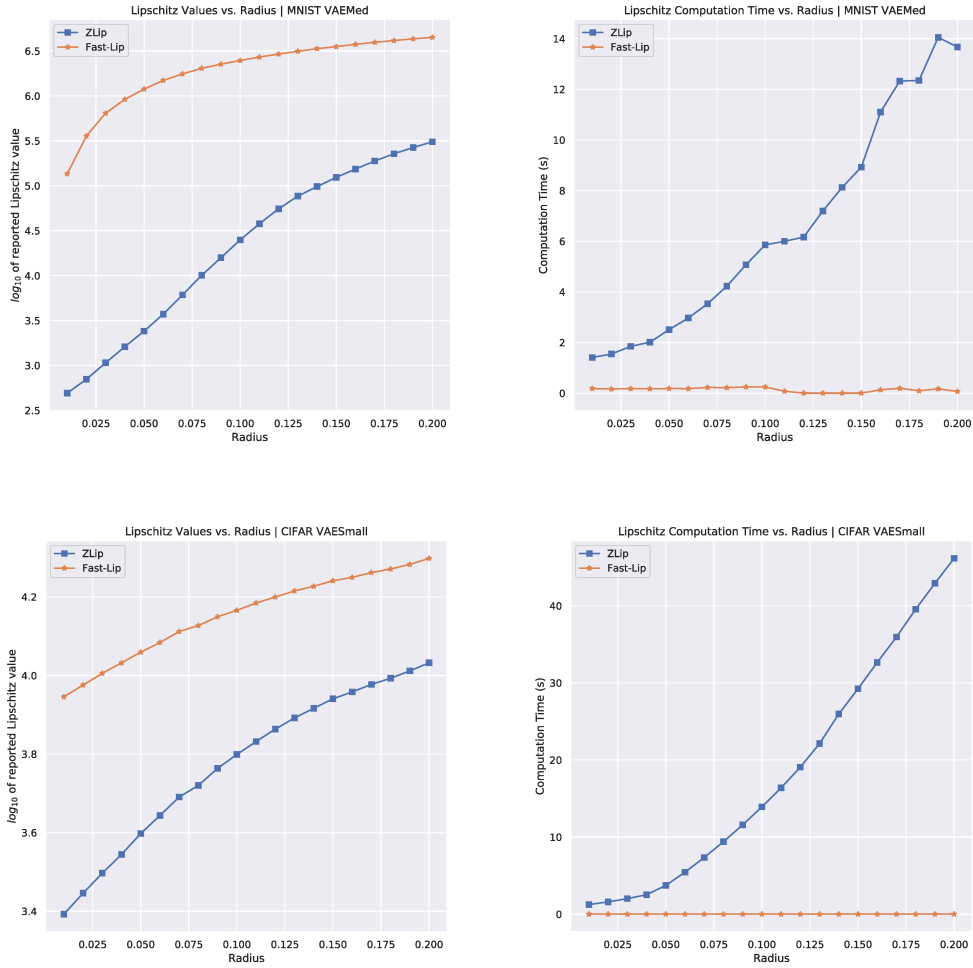


Figure 6. Reported Lipschitz constants and times for the MNIST MedVAE as we increase the radius of the region over which we evaluate Lipschitz constant. (Top Left) reports the values on a log-scale, noticing that both ZLip and Fast-Lip increase their estimate as the radius increases, the relative gap is largest for small radii. (Top Right) we display times in seconds versus the radius. The increase in running time of ZLip is due to the increase in number of unstable neurons, which increases the size of the representations of the zonotopes that must be passed through each layer. (Bottom) Reported Lipschitz constants and times for the CIFAR SmallVAE as we increase radius.

D.4. Ablating the choice of abstract domain:

ZLip operates by iteratively building *zonotopes* to satisfy the containments of equations 14-17, specifically we use zonotopes in both the forward pass (such as DeepZ, DiffAI), but also zonotopes in the backward pass. To examine the importance of zonotopes in both directions, we replace the zonotopes with hyperboxes in one or both of the directions. Note that using hyperboxes in both the forward and backward directions is FastLip. We denote the method that uses hyperboxes in the forward pass, but zonotopes in the backward pass as ‘Hyperbox $\rightarrow$ Zono’, and vice versa.

Toy Dataset: In Figure 7 we compare the performance of the different techniques on networks trained on the toy dataset. The y-axis records the logarithm of the average ratio between each method’s Lipschitz estimate and ZLip’s Lipschitz estimate for each example. We evaluate on 100 random points for each x-value. On the left panel we vary the architecture size while evaluating on inputs that are hyperboxes with radius 0.1. For every network considered, on average, the abstract domain in the forward pass is more important, and this gap becomes more apparent as the size of the network increases. On the right panel, we fix a network and evaluate the performance as we vary the size of the certified region. For small radii, the choice of the forward domain is more important, however as the radius increases, the backward domain becomes more important. We conjecture this is because, for large radii, there is much uncertainty about which ReLU’s are fixed and the performance of zonotopes and hyperboxes in the forward pass becomes equivalent.

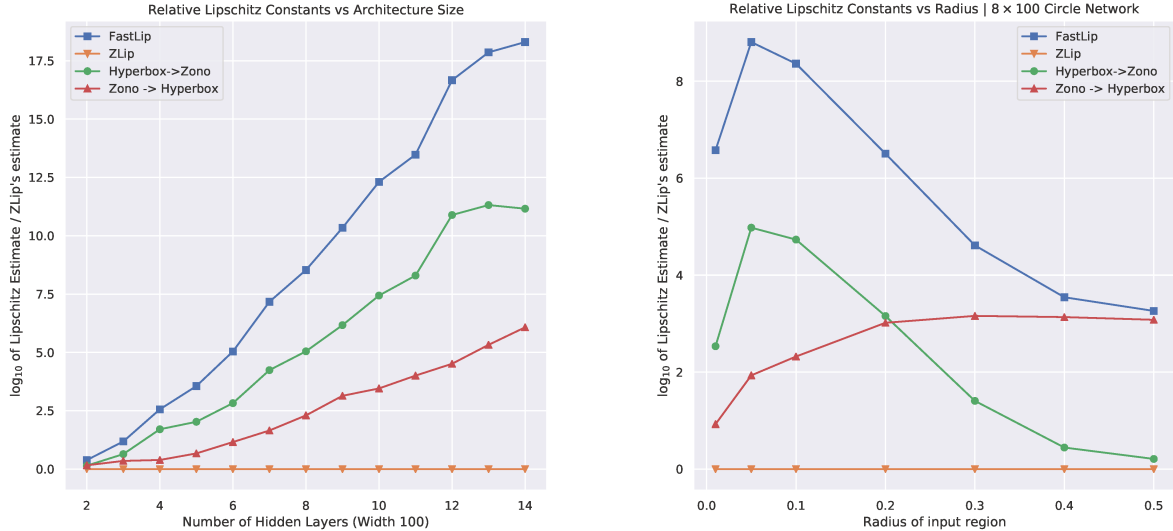


Figure 7. Log-mean-ratio of reported Lipschitz estimates (relative to ZLip) on varying networks trained on the Circle dataset. (Left) reports the values as we vary the network size, demonstrating that the forward domain is more important as the network size increases. (Right) reports the values for a fixed network as we vary the input radius size. Larger input radii have more uncertain ReLU neurons and the choice in forward domain becomes relatively less important.

MNIST Generative models: In Table 3 we evaluate the choice of abstract domains on MNIST generative models. We consider the decoders of the three different VAE’s trained on MNIST and evaluate over two different radii, considering hyperboxes of the denoted radius surrounding the encodings of random test MNIST examples. Here we again see that neither of the single-zonotope approaches is unilaterally better, however larger models tend to benefit more from using zonotopes in the forward pass.

Table 3. Mean ratio of Lipschitz estimates provided by other abstract domains relative to the estimate provided by ZLip, evaluated on MNIST VAE decoders. We see that as the model becomes larger, using zonotopes in the forward pass yields a tighter bound.

Radius	0.01			0.1		
Method	FastLip	$H \rightarrow Z$	$Z \rightarrow H$	FastLip	$H \rightarrow Z$	$Z \rightarrow H$
VAESmall	7.85	1.33	5.89	9.39	3.08	3.11
VAEMed	276.41	12.71	43.78	410.92	51.81	12.25
VAEBig	4238.83	123.21	85.65	26018.02	1357.93	26.64

D.5. Experimental Results on Classifiers

MNIST: For completeness and comparison against other networks on more realistic networks, we present results of our Lipschitz bounding technique versus several recent works for a variety of networks trained both with the standard classification loss as well as those trained adversarially. We evaluate the Lipschitz value returned by SeqLip, LipSDP, Fast-Lip, CLEVER, and ZLip for inputs of radius 0.1 centered at elements taken from the test set. If f is the trained classifier which has outputs in $\mathbb{R}^{10}$, we consider the Lipschitz constant of the network $f_i(\cdot) - f_{i+1}(\cdot)$ for each example where the true label is i . First we present the values for the MNIST networks trained with the standard CrossEntropy Loss in Table 4 and times are presented in Table 5. We remark that the bounds reported by CLEVER for networks with high dimension have been shown to be quite loose, so it is unclear what the correct Lipschitz value is for each of these networks. The most salient points here are that the values returned by ZLip are comparable to those returned by LipSDP at a significantly faster runtime. We also note that LipSDP errors when applied to networks as large as MedReLU or MedTanh.

For the PGD trained MNIST networks, we present the values and times in tables 6, 7. In direct comparison to the tables for the networks trained with CrossEntropy loss, we notice that all methods report lower values for the adversarially trained network. This tracks with prior work that adversarial regularization serves as a form of Lipschitz regularization.

CIFAR-10: The same experiments as above were performed on networks trained to classify the CIFAR-10 dataset. We present these results in Tables 8-11, but note that these results are qualitatively very similar to the results for the MNIST networks.

Table 4. Lipschitz values reported by various networks evaluated on the various Classifiers described above. All numbers report an average over regions of radius 0.1 centered at examples from the test set.

Lipschitz Estimates (MNIST) — CrossEntropy Loss					
Network	SeqLip	SDP	Fast-Lip	CLEVER	ZLip
TinyReLU	3.51×10^3	2.94×10^3	7.36×10^3	1.15×10^1	5.59×10^3
SmallReLU	2.12×10^4	1.52×10^4	1.94×10^5	9.59×10^0	9.34×10^4
MedReLU	1.08×10^6	—	1.45×10^8	1.06×10^1	1.77×10^7
TinyTanh	1.33×10^4	1.14×10^4	2.56×10^4	2.97×10^1	1.98×10^4
SmallTanh	5.80×10^4	4.29×10^4	3.24×10^5	3.26×10^1	1.68×10^5
MedTanh	5.50×10^6	—	4.29×10^8	3.56×10^2	6.94×10^7

Table 5. Times for MNIST classifiers trained with the Cross-Entropy loss

Lipschitz Times (MNIST) — CrossEntropy Loss					
Network	SeqLip	SDP	Fast-Lip	CLEVER	ZLip
TinyReLU	5.08×10^{-1}	1.78×10^1	1.75×10^{-3}	5.98×10^1	4.20×10^{-1}
SmallReLU	2.28×10^0	2.63×10^2	9.73×10^{-2}	1.01×10^2	9.36×10^{-1}
MedReLU	3.35×10^0	—	1.07×10^{-1}	1.75×10^2	2.13×10^0
TinyTanh	5.90×10^1	1.82×10^1	1.96×10^{-3}	5.98×10^1	4.22×10^{-1}
SmallTanh	4.56×10^0	2.28×10^2	1.05×10^{-1}	1.09×10^2	9.87×10^{-1}
MedTanh	9.91×10^0	—	1.08×10^{-1}	1.77×10^2	2.03×10^0

Table 6. Values for MNIST classifiers trained with the PGD loss

Lipschitz Estimates (MNIST) — PGD Loss					
Network	SeqLip	SDP	Fast-Lip	CLEVER	ZLip
TinyReLU	3.97×10^2	2.66×10^2	3.48×10^2	1.07×10^0	1.42×10^2
SmallReLU	1.52×10^3	9.35×10^2	1.98×10^4	1.85×10^0	5.39×10^3
MedReLU	3.91×10^4	1.64×10^4	1.45×10^7	2.26×10^0	9.36×10^5
TinyTanh	1.38×10^3	8.61×10^2	6.86×10^2	3.14×10^0	5.17×10^2
SmallTanh	7.96×10^3	4.81×10^3	5.68×10^4	6.24×10^0	2.42×10^4
MedTanh	2.73×10^5	1.30×10^5	5.98×10^7	1.90×10^1	6.27×10^6

Table 7. Times for MNIST classifiers trained with the PGD loss

Lipschitz Times (MNIST) — PGD Loss					
Network	SeqLip	SDP	Fast-Lip	CLEVER	ZLip
TinyReLU	1.18×10^{-2}	1.14×10^1	1.74×10^{-3}	2.71×10^1	1.10×10^{-1}
SmallReLU	2.71×10^1	1.77×10^2	1.13×10^{-1}	1.13×10^2	1.12×10^0
MedReLU	3.10×10^0	4.42×10^2	1.11×10^{-1}	1.76×10^2	2.17×10^0
TinyTanh	6.11×10^{-1}	1.62×10^1	1.84×10^{-3}	5.33×10^1	3.77×10^{-1}
SmallTanh	4.61×10^0	1.77×10^2	8.66×10^{-2}	7.72×10^1	7.31×10^{-1}
MedTanh	3.29×10^0	5.00×10^2	1.16×10^{-1}	1.86×10^2	2.15×10^0

Table 8. Values for CIFAR-10 classifiers trained with the Cross Entropy loss

Lipschitz Values (CIFAR-10) — CrossEntropy Loss					
Network	SeqLip	SDP	Fast-Lip	CLEVER	ZLip
TinyReLU	1.11×10^3	8.51×10^2	1.13×10^3	6.05×10^{-1}	8.89×10^2
SmallReLU	1.27×10^4	7.23×10^3	1.55×10^5	1.43×10^0	6.32×10^4
MedReLU	4.30×10^5	1.78×10^5	2.42×10^8	2.22×10^0	2.13×10^7
TinyTanh	4.12×10^3	3.22×10^3	5.52×10^3	2.28×10^0	4.32×10^3
SmallTanh	2.61×10^4	1.77×10^4	1.72×10^5	5.73×10^0	7.82×10^4
MedTanh	1.28×10^6	—	1.86×10^8	1.26×10^1	2.22×10^7

Table 9. Values for CIFAR-10 classifiers trained with the PGD loss

Lipschitz Values (CIFAR-10) — PGD Loss					
Network	SeqLip	SDP	Fast-Lip	CLEVER	ZLip
TinyReLU	3.01×10^2	1.92×10^2	4.63×10^1	1.17×10^{-1}	3.95×10^1
SmallReLU	2.03×10^3	1.22×10^3	2.15×10^4	3.62×10^{-1}	7.75×10^3
MedReLU	3.61×10^4	1.26×10^4	1.70×10^7	5.22×10^{-1}	1.26×10^6
TinyTanh	1.81×10^3	1.31×10^3	1.05×10^3	8.94×10^{-1}	8.91×10^2
SmallTanh	1.51×10^4	7.25×10^3	2.62×10^4	2.67×10^0	1.35×10^4
MedTanh	2.57×10^5	9.56×10^4	1.41×10^7	3.14×10^0	1.81×10^6

Table 10. Times for CIFAR-10 classifiers trained with the CrossEntropy loss

Lipschitz Times (CIFAR-10) — CrossEntropy Loss					
Network	SeqLip	SDP	Fast-Lip	CLEVER	ZLip
TinyReLU	2.32×10^{-1}	5.02×10^2	9.59×10^{-2}	5.91×10^1	6.24×10^{-1}
SmallReLU	7.14×10^0	2.88×10^3	1.07×10^{-1}	1.06×10^2	1.16×10^0
MedReLU	7.45×10^1	4.81×10^3	6.38×10^{-2}	9.88×10^1	1.32×10^0
TinyTanh	6.45×10^{-1}	5.05×10^2	1.18×10^{-1}	7.28×10^1	8.66×10^{-1}
SmallTanh	3.11×10^1	3.25×10^3	1.17×10^{-1}	1.22×10^2	1.22×10^0
MedTanh	1.09×10^0	6.78×10^3	6.34×10^{-3}	3.98×10^0	1.89×10^{-1}

Table 11. Times for CIFAR-10 classifiers trained with the PGD loss

Lipschitz Times (CIFAR-10) — PGD Loss					
Network	SeqLip	SDP	Fast-Lip	CLEVER	ZLip
TinyReLU	1.47×10^{-1}	3.94×10^2	1.09×10^{-1}	6.74×10^1	7.93×10^{-1}
SmallReLU	4.11×10^1	2.52×10^3	9.89×10^{-2}	9.12×10^1	1.07×10^0
MedReLU	5.22×10^1	4.01×10^3	1.14×10^{-1}	1.89×10^2	2.39×10^0
TinyTanh	8.12×10^{-2}	4.60×10^2	1.13×10^{-1}	7.02×10^1	8.25×10^{-1}
SmallTanh	4.97×10^0	3.28×10^3	1.13×10^{-1}	1.22×10^2	1.23×10^0
MedTanh	1.13×10^0	3.90×10^3	5.77×10^{-3}	3.98×10^0	1.89×10^{-1}