Scaling provable adversarial defenses

Eric Wong
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA 15213
`ericwong@cs.cmu.edu`

Frank Schmidt
Bosch Center for Artificial Intelligence
Renningen, Germany
`Frank.R.Schmidt@de.bosch.com`

Jan Hendrik Metzen
Bosch Center for Artificial Intelligence
Renningen, Germany
`JanHendrik.Metzen@de.bosch.com`

J. Zico Kolter
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA 15213
`zkolter@cs.cmu.edu`

Abstract

Recent work has developed methods for learning deep network classifiers that are *provably* robust to norm-bounded adversarial perturbation; however, these methods are currently only possible for relatively small feedforward networks. In this paper, in an effort to scale these approaches to substantially larger models, we extend previous work in three main directions. First, we present a technique for extending these training procedures to much more general networks, with skip connections (such as ResNets) and general nonlinearities; the approach is fully modular, and can be implemented automatically (analogous to automatic differentiation). Second, in the specific case of ℓ_∞ adversarial perturbations and networks with ReLU nonlinearities, we adopt a nonlinear random projection for training, which scales *linearly* in the number of hidden units (previous approaches scaled quadratically). Third, we show how to further improve robust error through cascade models. On both MNIST and CIFAR data sets, we train classifiers that improve substantially on the state of the art in provable robust adversarial error bounds: from 5.8% to 3.1% on MNIST (with ℓ_∞ perturbations of $\epsilon = 0.1$), and from 80% to 36.4% on CIFAR (with ℓ_∞ perturbations of $\epsilon = 2/255$). Code for all experiments in the paper is available at `https://github.com/locuslab/convex_adversarial/`.

1 Introduction

A body of recent work in adversarial machine learning has shown that it is possible to learn *provably robust* deep classifiers [Wong and Kolter, 2017, Raghunathan et al., 2018, Dvijotham et al., 2018]. These are deep networks that are verifiably *guaranteed* to be robust to adversarial perturbations under some specified attack model; for example, a certain robustness certificate may guarantee that for a given example x, no perturbation Δ with ℓ_∞ norm less than some specified ϵ could change the class label that the network predicts for the perturbed example $x + \Delta$. However, up until this point, such provable guarantees have only been possible for reasonably small-sized networks. It has remained unclear whether these methods could extend to larger, more representionally complex networks.

In this paper, we make substantial progress towards the goal of scaling these provably robust networks to realistic sizes. Specifically, we extend the techniques of [Wong and Kolter, 2017] in three key ways. First, while past work has only applied to pure feedforward networks, we extend the framework to deal with arbitrary residual/skip connections (a hallmark of modern deep network architectures), and arbitrary activation functions (Dvijotham et al. [2018] also worked with arbitrary activation

Preprint. Work in progress.

functions, but only for feedforward networks, and just discusses network verification rather than robust training). Second, and possibly most importantly, computing the upper bound on the robust loss in [Wong and Kolter, 2017] in the worst case scales *quadratically* in the number of hidden units in the network, making the approach impractical for larger networks. In this work, we use a nonlinear random projection technique to estimate the bound in manner that scales only linearly in the size of the hidden units (i.e., only a constant multiple times the cost of traditional training), and which empirically can be used to train the networks with no degradation in performance from the previous work. Third, we show how to further improve robust performance of these methods, though at the expense of worse non-robust error, using multi-stage cascade models. Through these extensions, we are able to improve substantially upon the verified robust errors obtained by past work.

2 Background and related work

Work in adversarial defenses typically falls in one of three primary categories. First, there is ongoing work in developing heuristic defenses against adversarial examples: [Goodfellow et al., 2015, Papernot et al., 2016, Kurakin et al., 2017, Metzen et al., 2017] to name a few. While this work is largely empirical at this point, substantial progress has been made towards developing networks that seem much more robust than previous approaches. Although a distressingly large number of these defenses are quickly "broken" by more advanced attacks [Athalye et al., 2018], there have also been some methods that have proven empirically resistant to the current suite of attacks; the recent NIPS 2017 adversarial example challenge [Kurakin et al., 2018], for example, highlights some of the progress made on developing classifiers that appear much stronger in practice than many of the ad-hoc techniques developed in previous years. Many of the approaches, though not formally verified in the strict sense during training, nonetheless have substantial theoretical justification for why they may perform well: Sinha et al. [2018] uses properties of statistical robustness to develop an approach that is not much more difficult to train and which empirically does achieve some measure of resistance to attacks; Madry et al. [2017] considers robustness to a first-order adversary, and shows that a randomized projected gradient descent procedure is optimal in this setting. Indeed, in some cases the classifiers trained via these methods can be verified to be adversarially robust using the verification techniques discussed below (though only for very small networks). Despite this progress, we believe it is also crucially important to consider defenses that *are* provably robust, to avoid any possible attack.

Second, our work in this paper relates closely to techniques for the formal verification of neural networks systems (indeed, our approach can be viewed as a convex procedure for verification, coupled with a method for training networks via the verified bounds). In this area, most past work focuses on using exact (combinatorial) solvers to verify the robustness properties of networks, either via Satisfiability Modulo Theories (SMT) solvers [Huang et al., 2017, Ehlers, 2017, Carlini and Wagner, 2017] or integer programming approaches [Lomuscio and Maganti, 2017, Tjeng and Tedrake, 2017, Cheng et al., 2017]. These methods have the benefit of being able to reason exactly about robustness, but at the cost of being combinatorial in complexity. This drawback has so far prevented these methods from effectively scaling to large models or being used within a training setting. There have also been a number of recent attempts to verify networks using non-combinatorial methods (and this current work fits broadly in this general area). For example, Gehr et al. [2018] develop a suite of verification methods based upon abstract interpretations (these can be broadly construed as relaxations of combinations of activations that are maintained as they pass through the network). Dvijotham et al. [2018] use an approach based upon analytically solving an optimization problem resulting from dual functions of the activations (which extends to activations beyond the ReLU). However, these methods apply to simple feedforward architectures without skip connections, and focus only on verification of existing networks.

Third, and most relevant to our current work, there are several approaches that go beyond provable verification, and also integrate the verification procedure into the training of the network itself. For example, Hein and Andriushchenko [2017] develop a formal bound for robustness to ℓ_2 perturbations in two-layer networks, and train a surrogate of their bounds. Raghunathan et al. [2018] develop a semidefinite programming (SDP) relaxation of exact verification methods, and train a network by minimizing this bound via the dual SDP. And Wong and Kolter [2017] present a linear-programming (LP) based upper bound on the robust error or loss that can be suffered under norm-bounded perturbation, then minimize this upper bound during training; the method is particularly efficient since they do not solve the LP directly, but instead show that it is possible to bound the LP optimal value and compute elementwise bounds on the activation functions based on a backward pass through

the network. However, it is still the case that none of these approaches scale to realistically-sized networks; even the approach of [Wong and Kolter, 2017], which empirically has been scaled to the largest settings of all the above approaches, in the worst case scales *quadratically* in the number of hidden units in the network and dimensions in the input. Thus, all the approaches so far have been limited to relatively small networks and problems such as MNIST.

Contributions This paper fits into this third category of integrating verification into training, and makes substantial progress towards scaling these methods to realistic settings. While we cannot yet reach e.g. ImageNet scales, even in this current work, we show that it *is* possible to overcome the main hurdles to scalability of past approaches. Specifically, we develop a provably robust training procedure, based upon the approach in [Wong and Kolter, 2017], but extending it in three key ways. The resulting method: 1) extends to general networks with skip connections, residual layers, and activations besides the ReLU; we do so by using a general formulation based on the Fenchel conjugate function of activations; 2) scales *linearly* in the dimensionality of the input and number of hidden units in the network, using techniques from nonlinear random projections, all while suffering minimal degradation in accuracy; and 3) further improves the quality of the bound with model cascades. We describe each of these contributions in the next section.

3 Scaling provably robust networks

3.1 Robust bounds for general networks via modular dual functions

This section presents an architecture for constructing provably robust bounds for general deep network architectures, using Fenchel duality. Importantly, we derive the dual of each network operation in a fully modular fashion, simplifying the problem of deriving robust bounds of a network to bounding the dual of individual functions. By building up a toolkit of dual operations, we can automatically construct the dual of any network architecture by iterating through the layers of the original network.

The adversarial problem for general networks We consider a generalized k "layer" neural network $f_\theta : \mathbb{R}^{|x|} \to \mathbb{R}^{|y|}$ given by the equations

$$z_i = \sum_{j=1}^{i-1} f_{ij}(z_j), \;\; \text{for } i = 2, \ldots, k \tag{1}$$

where $z_1 = x$, $f_\theta(x) \equiv z_k$ (i.e., the output of the network) and $f_{ij} : \mathbb{R}^{|z_j|} \to \mathbb{R}^{|z_i|}$ is some function from layer j to layer i. Importantly, this differs from prior work in two key ways. First, unlike the conjugate forms found in Wong and Kolter [2017], Dvijotham et al. [2018], we no longer assume that the network consists of linear operations followed by activation functions, and instead opt to work with an arbitrary sequence of k functions. This simplifies the analysis of sequential non-linear activations commonly found in modern architectures, e.g. max pooling or a normalization strategy followed by a ReLU,[1] by analyzing each activation independently, whereas previous work would need to analyze the entire sequence as a single, joint activation. Second, we allow layers to depend not just on the previous layer, but also on all layers before it. This generalization applies to networks with any kind of skip connections, e.g. residual networks and dense networks, and greatly expands the set of possible architectures.

Let $\mathcal{B}(x) \subset \mathbb{R}^{|x|}$, represent some input constraint for the adversary. For this section we will focus on an arbitrary norm ball $\mathcal{B}(x) = \{x + \Delta : \|\Delta\| \le \epsilon\}$. This is the constraint set considered for norm-bounded adversarial perturbations, however other constraint sets can certainly be considered. Then, given an input example x, a known label y^*, and a target label y^{targ}, the problem of finding the most adversarial example within $\mathcal{B}$ (i.e., a so-called *targeted* adversarial attack) can be written as

$$\underset{z_k}{\text{minimize}} \;\; c^T z_k, \;\; \text{subject to} \;\; z_i = \sum_{j=1}^{i-1} f_{ij}(z_j), \;\; \text{for } i = 2, \ldots, k, \;\; z_1 \in \mathcal{B}(x) \tag{2}$$

where $c = e_{y^\star} - e_{y^{\text{targ}}}$.

---

[1]Batch normalization, since it depends on entire minibatches, is formally not covered by the approach, but it can be approximated by considering the scaling and shifting to be generic parameters, as is done at test time.

Dual networks via compositions of modular dual functions   To bound the adversarial problem, we look to its dual optimization problem using the machinery of Fenchel conjugate functions [Fenchel, 1949], described in Definition 1.

Definition 1. *The conjugate of a function f is another function f^* defined by*

$$f^*(y) = \sup_x x^T y - f(x) \tag{3}$$

Specifically, we can lift the constraint $z_{i+1} = \sum_{j=1}^{i} f_{ij}(z_j)$ from Equation 2 into the objective with an indicator function, and use conjugate functions to obtain a lower bound. For brevity, we will use the subscript notation $(\cdot)_{1:i} = ((\cdot)_1, \ldots, (\cdot)_i)$, e.g. $z_{1:i} = (z_1, \ldots, z_i)$. Due to the skip connections, the indicator functions are not independent, so we cannot directly conjugate each individual indicator function. We can, however, still form its dual using the conjugate of a different indicator function corresponding to the backwards direction, as shown in Lemma 1.

Lemma 1. *Let the indicator function for the ith constraint be*

$$\chi_i(z_{1:i}) = \begin{cases} 0 & \text{if } z_i = \sum_{j=1}^{i-1} f_{ij}(z_j) \\ \infty & \text{otherwise,} \end{cases} \tag{4}$$

for $i = 2, \ldots, k$, and consider the joint indicator function $\sum_{i=2}^{k} \chi_i(z_{1:i})$. Then, the joint indicator is lower bounded by $\max_{\nu_{1:k}} \nu_k^T z_k - \nu_1^T z_1 - \sum_{i=1}^{k-1} \chi_i^*(\nu_{i:k})$, where

$$\chi_i^*(\nu_{i:k}) = \max_{z_i} -\nu_i^T z_i + \sum_{j=i+1}^{k} \nu_j^T f_{ji}(z_i) \tag{5}$$

for $i = 1, \ldots, k-1$. Note that $\chi_i^*(\nu_{i:k})$ is the exact conjugate of the indicator for the set $\{x_{i:k} : x_j = f_{ji}(x_i) \ \forall j > i\}$, which is different from the set indicated by χ_i. However, when there are no skip connections (i.e. z_i only depends on z_{i-1}), χ_i^* is exactly the conjugate of χ_i.

We defer the proof of Lemma 1 to Appendix A.1. With structured upper bounds on these conjugate functions, we can bound the original adversarial problem using the dual network described in Theorem 1. We can then optimize the bound using any standard deep learning toolkit using the same robust optimization procedure as in [Wong and Kolter, 2017] but using our bound instead. For the adversarial setting, note that the ℓ_∞ perturbation results in a dual norm of ℓ_1.

Theorem 1. *Let g_{ij} and h_i be any functions such that*

$$\chi_i^*(\nu_{i:k}) \leq h_i(\nu_{i:k}) \quad \text{subject to} \quad \nu_i = \sum_{j=i+1}^{k} g_{ij}(\nu_j) \tag{6}$$

for $i = 1, \ldots, k-1$. Then, the adversarial problem from Equation 2 is lower bounded by

$$J(x, \nu_{1:k}) = -\nu_1^T x - \epsilon \|\nu_1\|_* - \sum_{i=2}^{k} h_i(\nu_{i:k}) \tag{7}$$

where $\|\cdot\|_*$ is the dual norm, and $\nu_{1:k} = g(c)$ is the output of a k layer neural network g on input c, given by the equations

$$\nu_k = -c, \quad \nu_i = \sum_{j=i}^{k-1} g_{ij}(\nu_{j+1}), \quad \text{for } i = 1, \ldots, k-1. \tag{8}$$

We denote the upper bound on the conjugate function from Equation 6 a *dual layer*, and defer the proof to Appendix A.2. To give a concrete example, we present two possible dual layers for linear operators and ReLU activations in Corollaries 1 and 2, and we also depict an example dual residual block in Figure 1.

Corollary 1. *The dual layer for a linear operator $\hat{z}_{i+1} = W_i z_i + b_i$ is*

$$\chi_i^*(\nu_{i:k}) = \nu_{i+1}^T b_i \quad \text{subject to} \quad \nu_i = W_i^T \nu_{i+1}. \tag{9}$$
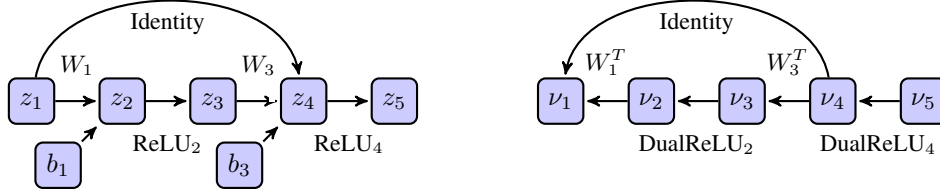
4

Figure 1: An example of the layers forming a typical residual block (left) and its dual (right), using the dual layers described in Corollaries 1 and 2. Note that the bias terms of the residual network go into the dual objective and are not part of the structure of the dual network, and the skip connections remain in the dual network but go in the opposite direction.

Corollary 2. *Suppose we have lower and upper bounds ℓ_{ij}, u_{ij} on the pre-activations. The dual layer for a ReLU activation $\hat{z}_{i+1} = \max(z_i, 0)$ is*

$$\chi_i^*(\nu_{i:k}) \leq -\sum_{j \in \mathcal{I}_i} \ell_{i,j}[\nu_{ij}]_+ \quad \text{subject to} \quad \nu_i = D_i \nu_{i+1}. \tag{10}$$

where $\mathcal{I}_i^-, \mathcal{I}_i^+, \mathcal{I}$ denote the index sets where the bounds are negative, positive or spanning the origin respectively, and where D_i is a diagonal matrix with entries

$$(D_i)_{jj} = \begin{cases} 0 & j \in \mathcal{I}_i^- \\ 1 & j \in \mathcal{I}_i^+ \\ \frac{u_{i,j}}{u_{i,j} - \ell_{i,j}} & j \in \mathcal{I}_i \end{cases}. \tag{11}$$

We briefly note that these dual layers recover the original dual network described in Wong and Kolter [2017]. Furthermore, the dual linear operation is the exact conjugate and introduces no looseness to the bound, while the dual ReLU uses the same relaxation used in Ehlers [2017], Wong and Kolter [2017]. More generally, the strength of the bound from Theorem 1 relies entirely on the tightness of the individual dual layers to their respective conjugate functions in Equation 6. While any g_{ij}, h_i can be chosen to upper bound the conjugate function, a tighter bound on the conjugate results in a tighter bound on the adversarial problem.

If the dual layers for all operations are linear, the bounds for all layers can be computed with a single forward pass through the dual network using a direct generalization of the form used in Wong and Kolter [2017] (due to their similarity, we defer the exact algorithm to Appendix F). By trading off tightness of the bound with computational efficiency by using linear dual layers, we can efficiently compute all bounds and construct the dual network one layer at a time. The end result is that we can automatically construct dual networks from dual layers in a fully modular fashion, completely independent of the overall network architecture (similar to how auto-differentiation tools proceed one function at a time to compute all parameter gradients using only the local gradient of each function). With a sufficiently comprehensive toolkit of dual layers, we can compute provable bounds on the adversarial problem for any network architecture.

For other dual layers, we point the reader to two resources. For the explicit form of dual layers for batch normalization, residual connections, and the derivations for the linear and ReLU dual layers, we direct the reader to Appendix B. For analytical forms of conjugate functions of other activation functions such as tanh, sigmoid, and max pooling, we refer the reader to Dvijotham et al. [2018].

3.2 Efficient bound computation via random projections

A limiting factor of the proposed algorithm and the work of Wong and Kolter [2017] is its computational complexity: for instance, to compute the bounds exactly for ℓ_∞ norm bounded perturbations in ReLU networks, it is computationally expensive to calculate $\|\nu_1\|_1$ and $\sum_{j \in \mathcal{I}_i} \ell_{ij}[\nu_{ij}]_+$. In contrast to other terms like $\nu_{i+1}^T b_i$ which require only sending a single bias vector through the dual network, the matrices ν_1 and $\nu_{i, \mathcal{I}_i}$ must be explicitly formed by sending an example through the dual network for each input dimension and for each $j \in \mathcal{I}_i$, which renders the entire computation *quadratic* in the number of hidden units. To scale the method for larger, ReLU networks with ℓ_∞ perturbations, we look to random Cauchy projections. Note that for an ℓ_2 norm bounded adversarial perturbation, the dual norm is also an ℓ_2 norm, so we can use traditional random projections [Vempala, 2005]. However, for the remainder of this section we focus on the ℓ_1 case arising from ℓ_∞ perturbations.

Algorithm 1 Estimating $\|\nu_1\|_1$ and $\sum_{j\in\mathcal{I}} \ell_{ij}[\nu_{ij}]_+$

---

input: Linear dual network operations g_{ij}, projection dimension r, lower bounds ℓ_{ij}, d_{ij} from Equation 13, layer-wise sizes $|z_i|$

$R_1^{(1)} := \mathrm{Cauchy}(r, |z_1|)$ // *initialize random matrix for ℓ_1 term*
for $i = 2, \ldots, k$ **do**
  // *pass each term forward through the network*
  for $j = 1, \ldots, i-1$ **do**
    $R_j^{(i)}, S_j^{(i)} := \sum_{k=1}^{i-1} g_{ki}^T(R_i^{(k)}), \sum_{k=1}^{i-1} g_{ki}^T(S_i^{(k)})$
  end for
  $R_i^{(1)}, S_i^{(1)} := \mathrm{diag}(d_i)\mathrm{Cauchy}(|z_i|, r), d_i$ // *initialize terms for layer i*
end for
output: $\mathrm{median}(|R_1|), 0.5\left(-\mathrm{median}(|R_2|) + S_2\right), \ldots, 0.5\left(-\mathrm{median}(|R_{k-1}|) + S_{k-1}\right)$

Estimating with Cauchy random projections From the work of Li et al. [2007], we can use the sample median estimator with Cauchy random projections to directly estimate $\|\nu_1\|_1$ for linear dual networks, and use a variation to estimate $\sum_{j\in\mathcal{I}} \ell_{ij}[\nu_{ij}]_+$, as shown in Theorem 2 (the proof is in Appendix D.1).

Theorem 2. . *Let $\nu_{1:k}$ be the dual network from Equation 1 with linear dual layers and let $r > 0$ be the projection dimension. Then, we can estimate*

$$\|\nu_1\|_1 \approx \mathrm{median}(|\nu_1^T R|) \tag{12}$$

where R is a $|z_1| \times r$ standard Cauchy random matrix and the median is taken over the second axis. Furthermore, we can estimate

$$\sum_{j\in\mathcal{I}} \ell_{ij}[\nu_{ij}]_+ \approx \frac{1}{2}\left(-\mathrm{median}(|\nu_i^T \mathrm{diag}(d_i)R|) + \nu_i^T d_i\right), \quad d_{i,j} = \left\{ \begin{array}{ll} \frac{u_{i,j}}{u_{i,j}-\ell_{i,j}} & j \notin \mathcal{I}_i \\ 0 & j \in \mathcal{I}_i \end{array} \right. \tag{13}$$

where R is a $|z_i| \times r$ standard Cauchy random matrix, and the median is taken over the second axis.

This estimate has two main advantages: first, it is simple to compute, as evaluating $\nu_1^T R$ involves passing the random matrix forward through the dual network (similarly, the other term requires passing a modified random matrix through the dual network; the exact algorithm is detailed in 1). Second, it is memory efficient in the backward pass, as the gradient need only propagate through the median entries.

These random projections reduce the computational complexity of computing these terms to piping r random Cauchy vectors (and an additional vector) through the network. Crucially, the complexity is no longer a quadratic function of the network size: if we fix the projection dimension to some constant r, then the computational complexity is now linear with the input dimension and $\mathcal{I}_i$. Since previous work was either quadratic or combinatorially expensive to compute, estimating the bound with random projections is the fastest and most scalable approach towards training robust networks that we are aware of. At test time, the bound can be computed exactly, as the gradients no longer need to be stored. However, if desired, it is possible to use a different estimator (specifically, the geometric estimator) for the ℓ_∞ norm to calculate high probability bounds on the adversarial problem, which is discussed in Appendix E.1.

3.3 Bias reduction with cascading ensembles

A final major challenge of training models to minimize a robust bound on the adversarial loss, is that the robustness penalty acts as a regularization. For example, in a two-layer ReLU network, the robust loss penalizes $\epsilon\|\nu_1\|_1 = \epsilon\|W_1 D_1 W_2\|_1$, which effectively acts as a regularizer on the network with weight ϵ. Because of this, the resulting networks (even those with large representational capacity), are typically overregularized to the point that many filters/weights become identically zero (i.e., the network capacity is not used).

To address this point, we advocate for using a robust *cascade* of networks: that is, we train a sequence of robust classifiers, where later elements of the cascade are trained (and evaluated) *only on those examples that the previous elements of the cascade cannot certify* (i.e., those examples that lie within ϵ of the decision boundary). This procedure is formally described in the Appendix in Algorithm 2.

Table 1: Number of hidden units, parameters, and time per epoch for various architectures.

Model	Dataset	# hidden units	# parameters	Time (s) / epoch
Small	MNIST	4804	166406	74
	CIFAR	6244	214918	48
Large	MNIST	28064	1974762	667
	CIFAR	62464	2466858	466
Resnet	MNIST	82536	3254562	2174
	CIFAR	107496	4214850	1685

Table 2: Results on MNIST, and CIFAR10 with small networks, large networks, residual networks, and cascaded variants.

Dataset	Model	Epsilon	Single model error		Cascade error	
			Robust	Standard	Robust	Standard
MNIST	Small, Exact	0.1	4.48%	1.26%	-	-
MNIST	Small	0.1	4.99%	1.37%	**3.13%**	**3.13%**
MNIST	Large	0.1	**3.67%**	**1.08%**	3.42%	3.18%
MNIST	Small	0.3	**43.10%**	14.87%	**33.64%**	**33.64%**
MNIST	Large	0.3	45.66%	**12.61%**	41.62%	35.24%
CIFAR10	Small	2/255	52.75%	38.91%	39.35%	39.35%
CIFAR10	Large	2/255	46.59%	**31.28%**	38.84%	36.08%
CIFAR10	Resnet	2/255	**46.11%**	31.72%	**36.41%**	**35.93%**
CIFAR10	Small	8/255	79.25%	72.24%	71.71%	71.71%
CIFAR10	Large	8/255	83.43%	80.56	79.24%	79.14%
CIFAR10	Resnet	8/255	**78.22%**	**71.33%**	**70.95%**	**70.77%**

4 Experiments

Dataset and Architectures We evaluate the techniques in this paper on two main datasets: MNIST digit classification [LeCun et al., 1998] and CIFAR10 image classification [Krizhevsky, 2009].[2] We test on a variety of deep and wide convolutional architectures, with and without residual connections. All code for these experiments is available at `https://github.com/locuslab/convex_adversarial/`. The small network is the same as that used in [Wong and Kolter, 2017], with two convolutional layers of 16 and 32 filters and a fully connected layer of 100 units. The large network is a scaled up version of it, with four convolutional layers with 32, 32, 64, and 64 filters, and two fully connected layers of 512 units. The residual networks use the same structure used by [Zagoruyko and Komodakis, 2016] with 4 residual blocks with 16, 16, 32, and 64 filters. We highlight a subset of the results in Table 2, and briefly describe a few key observations below. We leave more extensive experiments and details regarding the experimental setup in Appendix G. All results except where otherwise noted use random projection of 50 dimensions.

Summary of results For the different data sets and models, the final robust and nominal test errors are given in Table 2. We emphasize that in all cases we report the *robust test error*, that is, our *upper bound* on the possible test set error that the classifier can suffer under *any* norm-bounded attack (thus, considering different empirical attacks is orthogonal to our main presentation and not something that we include, as we are focused on verified performance). As we are focusing on the particular random projections discussed above, all experiments consider attacks with bounded ℓ_∞ norm, plus the ReLU networks highlighted above. On MNIST, the (non-cascaded) large model reaches a final robust error of 3.7% for $\epsilon = 0.1$, and the best cascade reaches 3.1% error. This contrasts with the best previous bound of 5.8% robust error for this epsilon, from [Wong and Kolter, 2017]. On CIFAR10, the ResNet model achieves 46.1% robust error for $\epsilon = 2/255$, and the cascade lowers this to 36.4% error. In contrast, the previous best *verified* robust error for this ϵ, from [Dvijotham et al., 2018], was 80%. While the robust error is naturally substantially higher for $\epsilon = 8/255$ (the amount typically

---

[2]We fully realize the irony of a paper with "scaling" in the title that currently maxes out on CIFAR10 experiments. But we emphasize that when it comes to certifiably robust networks, the networks we consider here, as we illustrate below in Table 1, are more than an order of magnitude larger than any that have been considered previously in the literature. Thus, our emphasis is really on the potential scaling properties of these approaches rather than large-scale experiments on e.g. ImageNet sized data sets.
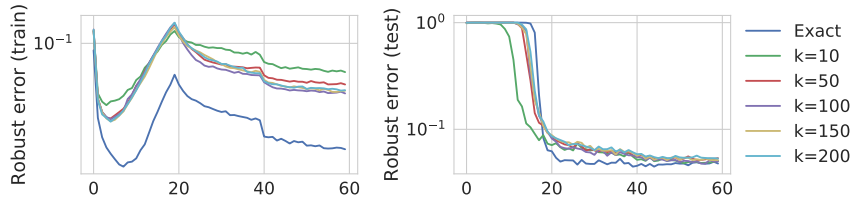
Figure 2: Training and testing robust error curves over epochs on the MNIST dataset using k projection dimensions. The ϵ value for training is scheduled from 0.01 to 0.1 over the first 20 epochs.
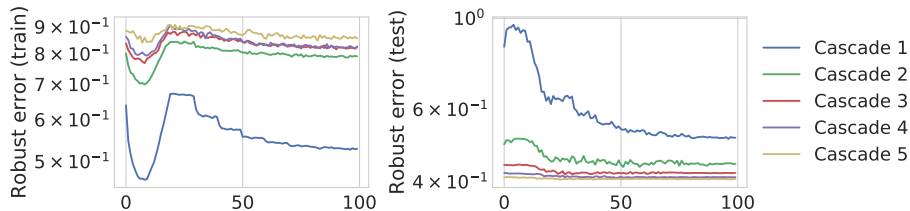

Figure 3: Robust error curves as we add models to the cascade for the CIFAR10 dataset on a small model. The ϵ value for training is scheduled to reach 2/255 after 20 epochs. The training curves are for each individual model, and the testing curves are for the whole cascade up to the stage.

considered in empirical works), we are still able to achieve 71% provable robust error; for comparison, the best *empirical* robust performance against current attacks is 53% error at $\epsilon = 8/255$ Madry et al. [2017], and most heuristic defenses have been broken to beyond this error Athalye et al. [2018].

Number of random projections In the MNIST dataset (the only data set where it is trivial to run exact training without projection), we have evaluated our approach using different projection dimensions as well as exact training (i.e., without random projections). We note that using substantially lower projection dimension does not have a significant impact on the test error. This fact is highlighted in Figure 2. Using the same convolutional architecture used by Wong and Kolter [2017], which previously required gigabytes of memory and took hours to train, it is sufficient to use only 10 random projections to achieve comparable test error performance to training with the exact bound. Each training epoch with 10 random projections takes less than a minute on a single GeForce GTX 1080 Ti graphics card, while using less than 700MB of memory, achieving significant speedup and memory reduction over Wong and Kolter [2017].

Cascades Finally, we consider the performance of the cascaded versus non-cascaded models. In all cases, cascading the models is able to improve the robust error performance, sometimes substantially, for instance decreasing the robust error on CIFAR10 from 46.1% to 36.4% for $\epsilon = 2/255$. However, this comes at a cost as well: the *nominal* error *increases* throughout the cascade (this is to be expected, since the cascade essentially tries to force the robust and nominal errors to match). Thus, there is substantial value to both improving the single-model networks *and* integrating cascades into the prediction.

5 Conclusion

In this paper, we have presented a general methodology for deriving dual networks from compositions of dual layers based on the methodology of conjugate functions to train classifiers that are provably robust to adversarial attacks. Importantly, the methodology is linearly scalable for ReLU based networks against ℓ_∞ norm bounded attacks, making it possible to train large scale, provably robust networks that were previously out of reach, and the obtained bounds can be improved further with model cascades. While this marks a significant step forward in scalable defenses for deep networks, there are several directions for improvement. One particularly important direction is better architecture development: a wide range of functions and activations not found in traditional deep residual networks may have better robustness properties or more efficient dual layers that also allow for scalable training. But perhaps even more importantly, we also need to consider the nature

of adversarial perturbations beyond just norm-bounded attacks. Better characterizing the space of perturbations that a network "should" be resilient to represents one of the major challenges going forward for adversarial machine learning.

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.

Chih-Hong Cheng, Georg Nührenberg, and Harald Ruess. Maximum resilience of artificial neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, pages 251–268. Springer, 2017.

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.

Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, 2017.

Werner Fenchel. On conjugate convex functions. *Canad. J. Math*, 1(73-77), 1949.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. AI^2: Safety and robustness certification of neural networks with abstract interpretation. In *IEEE Conference on Security and Privacy*, 2018.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL `http://arxiv.org/abs/1412.6572`.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*. 2017.

Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks. In *International Conference on Computer Aided Verification*, pages 3–29. Springer, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, et al. Adversarial attacks and defences competition. *arXiv preprint arXiv:1804.00097*, 2018.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Ping Li, Trevor J Hastie, and Kenneth W Church. Nonlinear estimators and tail bounds for dimension reduction in l1 using cauchy random projections. *Journal of Machine Learning Research*, 8(Oct): 2497–2532, 2007.

Alessio Lomuscio and Lalit Maganti. An approach to reachability analysis for feed-forward relu neural networks. *arXiv preprint arXiv:1706.07351*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On detecting adversarial perturbations. In *International Conference on Learning Representations*, 2017.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Security and Privacy (SP), 2016 IEEE Symposium on*, pages 582–597. IEEE, 2016.

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

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018.

Vincent Tjeng and Russ Tedrake. Verifying neural networks with mixed integer programming. *CoRR*, abs/1711.07356, 2017. URL http://arxiv.org/abs/1711.07356.

Santosh S Vempala. *The random projection method*, volume 65. American Mathematical Soc., 2005.

Eric Wong and J Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

A Conjugates and lower bounds with duality

A.1 Conjugates of the joint indicator function

Here, we derive the conjugate of $\sum_{i=2}^{k} \chi_i(z_{1:i})$. It is mathematically convenient to introduce addition variables $\hat{z}_{1:k}$ such that $\hat{z}_i = z_i$ for all $i = 1, \ldots, k$, and rephrase it as the equivalent constrained optimization problem.

$$\max_{z_{1:k}, \hat{z}_{1:k}} 0$$

$$\text{subject to } \hat{z}_i = \sum_{j=1}^{i-1} f_{ij}(z_j) \quad \text{for} \quad i = 2, \ldots, k \tag{14}$$

$$z_i = \hat{z}_i \quad \text{for } i = 1, \ldots, k$$

We introduce Lagrangian variables $\nu_{1:k}, \hat{\nu}_{2:k}$ to get the following Lagrangian:

$$L(z_{1:k}, \hat{z}_{1:k}, \nu_{1:k}, \hat{\nu}_{2:k}) = \sum_{i=2}^{k} \hat{\nu}_i^T \left(\hat{z}_i - \sum_{j=1}^{i-1} f_{ij}(z_j) \right) + \sum_{i=1}^{k} \nu_i^T (z_i - \hat{z}_i) \tag{15}$$

Grouping up terms by $z_i, \hat{z}_i$ and rearranging the double sum results in the following expression:

$$L(z_{1:k}, \hat{z}_{1:k}, \nu_{1:k}, \hat{\nu}_{2:k}) = -\nu_1^T \hat{z}_1 + \sum_{i=2}^{k} (\hat{\nu}_i - \nu_i)^T \hat{z}_i + \sum_{i=1}^{k} \left(\nu_i^T z_i - \sum_{j=i+1}^{k} \hat{\nu}_j^T f_{ji}(z_i) \right) \tag{16}$$

From the KKT stationarity conditions for the derivative with respect to $\hat{z}_i$, we know that $\hat{\nu}_i = \nu_i$. Also note that in the summand, the last term for $i = k$ has no double summand, so we move it out for clarity.

$$L(z_{1:k}, \nu_{1:k}) = -\nu_1^T \hat{z}_1 + \nu_k^T z_k + \sum_{i=1}^{k-1} \left(\nu_i^T z_i - \sum_{j=i+1}^{k} \nu_j^T f_{ji}(z_i) \right) \tag{17}$$

Finally, we minimize over z_i for $i = 2, \ldots, k - 1$ to get the conjugate form

$$L(z_1, z_k, \nu_{1:k}) = -\nu_1^T \hat{z}_1 + \nu_k^T z_k - \sum_{i=1}^{k-1} \max_{z_i} \left(-\nu_i^T z_i + \sum_{j=i+1}^{k} \nu_j^T f_{ji}(z_i) \right)$$

$$= -\nu_1^T z_1 + \nu_k^T z_k - \sum_{i=1}^{k-1} \chi_i^*(\nu_{i:k}) \tag{18}$$

A.2 Proof of Theorem 1

First, we rewrite the primal problem by bringing the function and input constraints into the objective with indicator functions I. We can then apply Lemma 1 to get the following lower bound on the adversarial problem:

$$\underset{\nu_{1:k}}{\text{maximize}} \, \underset{z_1, z_k}{\text{minimize}} \, (c^T + \nu_k)^T z_k + I(z_1 \in \mathcal{B}(x)) - \nu_1^T z_1 - \sum_{i=1}^{k-1} \chi_i^*(\nu_{i:k}) \tag{19}$$

Minimizing over z_1 and z_k, note that

$$\min_{\hat{z}_k} (c + \nu_k)^T \hat{z}_k = I(\nu_k = -c)$$

$$\min_{\hat{z}_1} I(z_1 \in \mathcal{B}(x)) - \nu_1^T z_1 = -I_{\hat{z}_1 \in \mathcal{B}(x)}^*(\nu_1) \tag{20}$$

Note that if $\mathcal{B}(x) = \{x + \Delta : \|\Delta\| \le \epsilon\}$ for some norm, then $I_{\hat{z}_1 \in \mathcal{B}(x)}^*(\nu_1) = \nu_1^T x + \epsilon \|\nu_1\|_*$ where $\| \cdot \|$ is the dual norm, but any sort of input constraint can be used so long as its conjugate can be bounded. Finally, consider the last term can be bounded with the dual layer:

$$\min_{z_i} \nu_i^T z_i - \sum_{j=i+1}^{k} \nu_j^T f_{ji}(z_i) = -\chi_i^*(\nu_{i:k}) \ge -h_i(\nu_{i:k}) \quad \text{subject to} \quad \nu_i = \sum_{j=i+1}^{k} g_{ij}(\nu_j) \tag{21}$$

11

Combining these all together, we get that the adversarial problem from Equation 2 is lower bounded by

$$\underset{\nu}{\text{maximize}} \quad -\nu_1^T x - \epsilon \|\nu_1\|_* - \sum_{i=1}^{k-1} h_i(\nu_{i:k})$$

$$\text{subject to} \quad \nu_k = -c \tag{22}$$

$$\nu_i = \sum_{j=i+1}^{k} g_{ij}(\nu_j)$$

B Dual layers

In this section, we derive the dual layers for standard building blocks of deep learning.

B.1 Linear operators

Suppose $f_i(z_i) = W_i z_i + b_i$ for some linear operator W_i and bias terms b_i. Then,

$$\begin{aligned}
\chi_i^*(\nu_i, \nu_{i+1}) &= \max_{z_i} -z_i^T \nu_i + (W_i z_i + b_i)^T \nu_{i+1} \\
&= \max_{z_i} z_i^T (W_i^T \nu_{i+1} - \nu_i) + b_i^T \nu_{i+1} \\
&= \max_{z_i} I(\nu_i = W_i^T \nu_{i+1}) + b_i^T \nu_{i+1} \\
&= b_i^T \nu_{i+1} \quad \text{subject to} \quad \nu_i = W_i^T \nu_{i+1}
\end{aligned} \tag{23}$$

B.2 Residual linear connections

Suppose $f_i(z_i, z_j) = W_i z_i + z_j + b_i$ and $z_{j+1} = W_j z_j + b_j$ for some $j < i - 1$ for linear operators W_i, W_j and bias term b_i, b_j. Then,

$$\begin{aligned}
\chi_i^*(\nu_i) &= \max_{z_i} -z_i^T \nu_i + (W_i z_i + b_i)^T \nu_{i+1} \\
&= b_i^T \nu_{i+1} \quad \text{subject to} \quad \nu_i = W_i^T \nu_{i+1}
\end{aligned} \tag{24}$$

and

$$\begin{aligned}
\chi_i^*(\nu_j) &= \max_{z_j} -z_j^T \nu_j + z_j^T \nu_i + (W_j z_j + b_j)^T \nu_{j+1} \\
&= b_j^T \nu_j \quad \text{subject to} \quad \nu_j = W_j^T \nu_{j+1} + \nu_i
\end{aligned} \tag{25}$$

B.3 ReLU activations

The proof here is the same as that presented in Appendix A3 of Wong and Kolter [2017], however we reproduce a simplified version here for the reader. The conjugate function for the ReLU activation is the following:

$$\chi^*(\nu_i, \nu_{i+1}) = \max_{z_i} -z_i^T \nu_i + \max(z_i, 0)\nu_{i+1} \tag{26}$$

Suppose we have lower and upper bounds ℓ_i, u_i on the input z_i. If $u_i \leq 0$, then $\max(z_i, 0) = 0$, and so

$$\chi^*(\nu_i, \nu_{i+1}) = \max_{z_i} -z_i^T \nu_i = 0 \quad \text{subject to} \quad \nu_i = 0 \tag{27}$$

Otherwise, if $\ell_i \geq 0$, then $\max(z_i, 0) = z_i$ and we have

$$\chi^*(\nu_i, \nu_{i+1}) = \max_{z_i} -z_i^T \nu_i + z_i^T \nu_{i+1} = 0 \quad \text{subject to} \quad \nu_i = \nu_{i+1} \tag{28}$$

Lastly, suppose $\ell_i < 0 < u_i$. Then, we can upper bound conjugate by taking the maximum over a convex outer bound of the ReLU, namely $\mathcal{S}_i = \{(z_i, z_{i+1}) : z_{i+1} \geq 0, z_{i+1} \geq z_i, -u_i \odot z_i + (u_i - \ell_i) \odot z_{i+1} \leq -u_i \odot \ell_i\}$, where $\odot$ denotes element-wise multiplication:

$$\chi^*(\nu_i, \nu_{i+1}) \leq \max_{\mathcal{S}_i} -z_i^T \nu_i + z_{i+1}^T \nu_{i+1} \tag{29}$$

The maximum must occur either at the origin $(0,0)$ or along the line $-u_{ij}z_{ij} + (u_{ij} - \ell_{ij})z_{i+1,j} = -u_{ij}\ell_{ij}$, so we can upper bound it again with

$$
\begin{aligned}
\chi^*(\nu_{ij}, \nu_{i+1,j}) &\leq \max_{z_{ij}} \left[-z_{ij}\nu_{ij} + \left(\frac{u_{ij}}{u_{ij} - \ell_{ij}} z_{ij} - \frac{u_{ij}\ell_{ij}}{u_{ij} - \ell_{ij}} \right) \nu_{i+1,j} \right]_+ \\
&= \max_{z_{ij}} \left[\left(\frac{u_{ij}}{u_{ij} - \ell_{ij}} \nu_{i+1,j} - \nu_{ij} \right) z_{ij} - \frac{u_{ij}\ell_{ij}}{u_{ij} - \ell_{ij}} \nu_{i+1,j} \right]_+ \\
&= \left[-\frac{u_{ij}\ell_{ij}}{u_{ij} - \ell_{ij}} \nu_{i+1,j} \right]_+ \quad \text{subject to} \quad \nu_{ij} = \frac{u_{ij}}{u_{ij} - \ell_{ij}} \nu_{i+1,j} \\
&= -\ell_{ij}\left[\nu_{ij} \right]_+ \quad \text{subject to} \quad \nu_{ij} = \frac{u_{ij}}{u_{ij} - \ell_{ij}} \nu_{i+1,j}
\end{aligned}
\tag{30}
$$

Let $\mathcal{I}_i^-, \mathcal{I}_i^+, \mathcal{I}$ and D_i be as defined in the corollary. Combining these three cases together, we get the final upper bound:

$$
\chi_i^*(\nu_{i:k}) \leq - \sum_{j \in \mathcal{I}_i} \ell_{i,j}[\nu_{i+1,j}]_+ \quad \text{subject to} \quad \nu_i = D_i\nu_{i+1}
\tag{31}
$$

B.4 Batch normalization

As mentioned before, we only consider the case of batch normalization with a fixed mean and variance. This is true during test time, and at training time we can use the batch statistics as a heuristic. Let μ_i, σ_i be the fixed mean and variance statistics, so batch normalization has the following form:

$$
BN(z_i) = \gamma \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta
\tag{32}
$$

where γ, β are the batch normalization parameters. Then,

$$
z_i = \gamma \frac{\hat{z}_i - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta = D_i z_i + d_i
\tag{33}
$$

where $D_{i+1} = \text{diag}\left(\frac{\gamma}{\sqrt{\sigma^2 + \epsilon}} \right)$ and $d_{i+1} = \beta - \frac{\mu}{\sqrt{\sigma^2 + \epsilon}}$. and so we can simply plug this into the linear case to get

$$
\chi_i^*(\nu_{i:k}) = d_i^T \nu_{i+1} \quad \text{subject to} \quad \nu_i = D_i\nu_{i+1}
\tag{34}
$$

Note however, that batch normalization has the effect of shifting the activations to be centered more around the origin, which is exactly the case in which the robust bound becomes looser. In practice, we find that while including batch normalization may improve convergence, it reduces the quality of the bound.

C Cascade construction

The full algorithm for constructing cascades as we describe in the main text is shown in Algorithm 2. To illustrate the use of the cascade, Figure 4 shows a two stage cascade on a few data points in two dimensional space. The boxes denote the adversarial ball around each example, and if the decision boundary is outside of the box, the example is certified.

D Estimation using Cauchy random projections

D.1 Proof of Theorem 2

Estimating $\|\hat{\nu}_1\|_{1,:}$ Recall the form of $\hat{\nu}_1$,

$$
\hat{\nu}_1 = IW_1^T D_2 W_2^T \dots D_n W_n^T = g(I)
$$

where we include the identity term to make explicit the fact that we compute this by passing an identity matrix through the network g. Estimating this term is straightforward: we simply pass in a Cauchy random matrix R, and take the median absolute value:

$$
\|\hat{\nu}_1\|_{1,:} \approx \text{median}(|RW_1^T D_2 W_2^T \dots D_n W_n^T|) = \text{median}(|g(R)|)
$$

where the median is taken over the minibatch axis.

Algorithm 2 Training robust cascade of k networks and making predictions

---

input: Initialized networks $f_1, \ldots, f_k$, training examples X, y, robust training procedure denoted RobustTrain, test example x^*

for $i = 1, \ldots, k$ **do**
   $f_i := \text{RobustTrain}(f_i, X, y)$ *// Train network*
   // remove certified examples from dataset
   $X, y := \{x_i, y_i : J(x, g(e_{f(x_i)} - e_{y^{targ}})) > 0, \ \forall y^{targ} \neq f(x_i)\}$
end for
for $i = 1, \ldots, k$ **do**
   if $J(x, g(e_{f_i(x^*)} - e_{y^{targ}})) < 0 \ \forall y^{targ} \neq f_i(x^*)$ **then**
      output: $f_i(x^*)$ *// return label if certified*
   end if
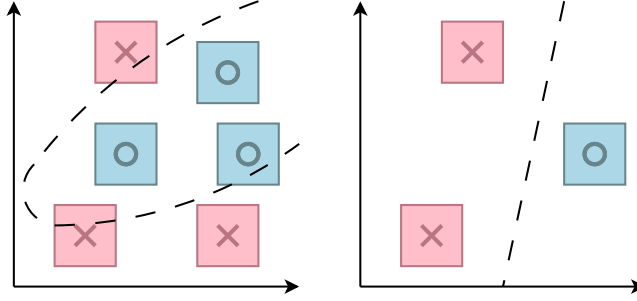**end for**
output: no certificate

Figure 4: An example of a two stage cascade. The first model on the left can only robustly classify three of the datapoints. After removing the certified examples, the remaining examples can now easily be robustly classified by a second stage classifier.

Estimating $\sum_i [\nu_{i,:}]_+$ Recall the form of $\nu = \nu_j$ for some layer j,

$$\nu_j = I D_j W_j^T \ldots D_n W_n^T = g_j(I)$$

Note that for a vector x,

$$\sum_i [x]_+ = \frac{\|x\|_1 + 1^T x}{2}$$

So we can reuse the ℓ_1 approximation from before to get

$$\sum_i [\nu_{i,:}]_+ = \frac{\|\nu\|_{1,:} + 1^T \nu}{2} \approx \frac{|\operatorname{median}(g_j(R)) + g_j(1^T)|}{2}$$

which involves using the same median estimator and also passing in a single example of ones through the network.

Estimating $\sum_{i \in \mathcal{I}} \ell_i [\nu_{i,:}]_+$ The previous equation, while simple, is not exactly the term in the objective; there is an addition ℓ_1 factor for each row, and we only add rows in the $\mathcal{I}$ set. However, we can deal with this by simply passing in a modified input to the network, as we will see shortly:

$$
\begin{aligned}
\sum_{i \in \mathcal{I}} \ell_i [\nu_{i,:}]_+ &= \sum_{i \in \mathcal{I}} \ell_i \frac{|\nu_{i,:}| + \nu_{i,:}}{2} \\
&= \frac{1}{2} \left(\sum_{i \in \mathcal{I}} \ell_i |\nu_{i,:}| + \sum_{i \in \mathcal{I}} \ell_i \nu_{i,:} \right) \\
&= \frac{1}{2} \left(\sum_{i \in \mathcal{I}} \ell_i |g_j(I)_i| + \sum_{i \in \mathcal{I}} \ell_i g_j(I)_i \right)
\end{aligned}
\tag{35}
$$

14

Note that since g_j is just a linear function that does a forward pass through the network, for any matrix A, B,

$$Ag_j(B) = ABD_jW_j^T \ldots D_nW_n^T = g_j(AB).$$

So we can take the multiplication by scaling terms ℓ to be an operation on the input to the network (note that we assume $\ell_i < 0$, which is true for all $i \in \mathcal{I}$)

$$\sum_{i \in \mathcal{I}} \ell_i[\nu_{i,:}]_+ = \frac{1}{2}\left(-\sum_{i \in \mathcal{I}}|g_j(\operatorname{diag}(\ell))_i| + \sum_{i \in \mathcal{I}} g_j(\operatorname{diag}(\ell))_i\right) \tag{36}$$

Similarly, we can view the summation over the index set $\mathcal{I}$ as a summation after multiplying by an indicator matrix $1_{\mathcal{I}}$ which zeros out the ignored rows. Since this is also linear, we can move it to be an operation on the input to the network.

$$\sum_{i \in \mathcal{I}} \ell_i[\nu_{i,:}]_+ = \frac{1}{2}\left(-\sum_i|g_j(1_{\mathcal{I}}\operatorname{diag}(\ell))_i| + \sum_i g_j(1_{\mathcal{I}}\operatorname{diag}(\ell))_i\right) \tag{37}$$

Let the linear, preprocessing operation be $h(X) = X1_{\mathcal{I}}\operatorname{diag}(\ell)$ so

$$h(I) = 1_{\mathcal{I}}\operatorname{diag}(\ell).$$

Then, we can observe that the two terms are simply an $\ell_{1,:}$ operation and a summation of the network output after applying g_j to $h(I)$ (where in the latter case, since everything is linear we can take the summation inside both g and h to make it $g_j(h(1^T))$):

$$\sum_{i \in \mathcal{I}} \ell_i[\nu_{i,:}]_+ = \frac{1}{2}\left(-\|g_j(h(I))\|_{1,:} + g_j(h(1^T))\right) \tag{38}$$

The latter term is cheap to compute, since we only pass a single vector. We can approximate the first term using the median estimator on the compound operations $g \circ h$ for a Cauchy random matrix R:

$$\sum_{i \in \mathcal{I}} \ell_i[\nu_{i,:}]_+ \approx \frac{1}{2}\left(-\operatorname{median}(|g_j(h(R))|) + g_j(h(1^T))\right) \tag{39}$$

The end result is that this term can be estimated by generating a Cauchy random matrix, scaling its terms by ℓ and zeroing out columns in $\mathcal{I}$, then passing it through the network and taking the median. $h(R)$ can be computed for each layer lower bounds ℓ, and cached to be computed for the next layer, similar to the non-approximate case.

E   High probability bounds

In this section, we derive high probability certificates for robustness against adversarial examples. Recall that the original certificate is of the form

$$J(g(c, \alpha)) < 0,$$

so if this holds we are guaranteed that the example cannot be adversarial. What we will show is an equivalent high probability statement: for $\delta > 0$, with probability at least $(1 - \delta)$,

$$J(g(c, \alpha)) \leq \tilde{J}(g(c, \alpha))$$

where $\tilde{J}$ is equivalent to the original J but using a high probability ℓ_1 upper bound. Then, if $\tilde{J}(g(c, \alpha)) < 0$ then with high probability we have a certificate.

E.1   High probability bounds using the geometric estimator

While the median estimator is a good heuristic for training, it is still only an estimate of the bound. At test time, it is possible to create a provable bound that holds with high probability, which may be desired if computing the exact bound is computationally impossible.

In this section, we derive high probability certificates for robustness against adversarial examples. Recall that the original certificate is of the form

$$J(g(c, \alpha)) < 0,$$

so if this holds we are guaranteed that the example cannot be adversarial. What we will show is an equivalent high probability statement: for $\delta > 0$, with probability at least $(1 - \delta)$,

$$J(g(c, \alpha)) \leq \tilde{J}(g(c, \alpha))$$

where $\tilde{J}$ is equivalent to the original J but using a high probability upper bound on the ℓ_1 norm. Then, if $\tilde{J}(g(c, \alpha)) < 0$ then with high probability we have a certificate.

E.2   Tail bounds for the geometric estimator

From Li et al. [2007], the authors also provide a geometric mean estimator which comes with high probability tail bounds. The geometric estimator is

$$\|\hat{\nu}_1\|_{1,j} \approx \prod_{i=1}^{k} |g(R)_{i,j}|^{1/k}$$

and the relevant lower tail bound on the ℓ_1 norm is

$$P\left(\frac{1}{1-\epsilon} \prod_{i=1}^{k} |g(R)_{i,j}|^{1/k} \leq \|\hat{\nu}_1\|_{1,j}\right) \leq \exp\left(-k\frac{\epsilon^2}{G_{L,gm}}\right) \tag{40}$$

where

$$G_{L,gm} = \frac{\epsilon^2}{\left(-\frac{1}{2}\log\left(1 + \left(\frac{2}{\pi}\log(1-\epsilon)\right)^2\right) + \frac{2}{\pi}\tan^{-1}\left(\frac{2}{\pi}\log(1-\epsilon)\right)\log(1-\epsilon)\right)}$$

Thus, if $\exp\left(-k\frac{\epsilon^2}{G_{L,gm}}\right) \leq \delta$, then with probability $1 - \delta$ we have that

$$\|\hat{\nu}_1\|_{1,j} \leq \frac{1}{1-\epsilon} \prod_{i=1}^{k} |g(R)_{i,j}|^{1/k} = \text{geo}(R)$$

which is a high probability upper bound on the ℓ_1 norm.

E.3   Upper bound on $J(g(c, \alpha))$

In order to upper bound $J(g(c, \alpha))$, we must apply the ℓ_1 upper bound for *every* ℓ_1 term. Let $n_1, \ldots, n_k$ denote the number of units in each layer of a k layer neural network, then we enumerate all estimations as follows:

1. The ℓ_1 norm computed at each intermediary layer when computing iterative bounds. This results in $n_2 + \cdots + n_{k-1}$ estimations.
2. The $\sum_{j \in \mathcal{I}_i} \ell_{i,j}[\nu_{i,j}]_+$ term for each $i = 2, \ldots, k - 1$, computed at each intermediary layer when computing the bounds. This results in $n_3 + 2n_4 + \cdots + (k-3)n_{k-1}$.

In total, this is $n_2 + 2n_3 + \cdots + (k-2)n_{k-1} = N$ total estimations. In order to say that *all* of these estimates hold with probability $1 - \delta$, we can do the following: we bound each estimate in Equation 40 with probability δ/N, and use the union bound over all N estimates. We can then conclude that with probability at most δ, any estimate is not an upper bound, and so with probability $1 - \delta$ we have a proper upper bound.

E.4   Achieving δ/N tail probability

There is a problem here: if δ/N is small, then ϵ becomes large, and the bound gets worse. In fact, since $\epsilon < 1$, when k is fixed, there's actually a lower limit to how small δ/N can be.

To overcome this problem, we take multiple samples to reduce the probability. Specifically, instead of directly using the geometric estimator, we use the maximum over multiple geometric estimators

$$\text{maxgeo}(R_1, \ldots, R_m) = \max(\text{geo}(R_1), \ldots, \text{geo}(R_m)),$$

where R_i are independent Cauchy random matrices. If each one has a tail probability of δ, then the maximum has a tail probability of δ^m, which allows us to get arbitrarily small tail probabilities at a rate exponential in m.

E.5 High probability tail bounds for network certificates

Putting this altogether, let $\delta > 0$, let $N > 0$ be the number of estimates needed to calculate a certificate, and let m be the number of geometric estimators to take a maximum over. Then with probability $(1 - \delta)$, if we bound the tail probability for each geometric estimate with $\hat{\delta} = \left(\frac{\delta}{N}\right)^{1/m}$, then we have an upper bound on the certificate.

E.5.1 MNIST example

As an example, suppose we use the MNIST network from **??**. Then, let $\delta = 0.01$, $m = 10$, and note that $N = 6572$. Then, $\hat{\delta} = 0.26$, which we can achieve by using $k = 200$ and $\epsilon = 0.22$.

F AutoDual

In this section, we describe our generalization of the bounds computation algorithm from [Wong and Kolter, 2017] to general networks using dual layers, which we call AutoDual.

Efficient construction of the dual network via linear dual operators The conjugate form, and consequently the dual layer, for certain activations requires knowing lower and upper bounds for the pre-activations, as was done for ReLU activations in Algorithm 1 of Wong and Kolter [2017]. While the bound in Equation 7 can be immediately used to compute all the bounds on intermediate nodes of the network one layer at a time, this requires performing a backwards pass through the dual network whenever we need to compute the bounds. However, if the operators g_{ij} of the dual layers are all affine operators $g_{ij}(\nu_{i+1}) = A_{ij}^T \nu_{i+1}$ for some affine operator A_{ij}, we can apply a generalization of the lower and upper bound computation found in Wong and Kolter [2017] to compute all lower and upper bounds, and consequently the dual layers, of the entire network with a single forward pass in a layer-by-layer fashion. With the lower and upper bounds, we can also use the same algorithm to automatically construct the dual network. The resulting algorithm, which we call AutoDual, is described in Algorithm 3.

In practice, we can perform several layer-specific enhancements on top of this algorithm. First, many of the A_{ji} operators will not exist simply because most architectures (with a few exceptions) don't have a large number of skip connections, so these become no ops and can be ignored. Second, we can lazily skip the computation of layer-wise bounds until necessary, e.g. for constructing the dual layer of ReLU activations. Third, since many of the functions h_j in the dual layers are functions of $B^T \nu_i$ for some matrix B and some $i \geq j$, we can initialize $\nu_i^{(i)}$ with B instead of the identity matrix, typically passing a much smaller matrix through the dual network (in many cases, B is a single vector).

G Experiments

In this section, we provide more details on the experimental setup, as well as more extensive experiments on the effect of model width and model depth on the performance that were not mentioned above.

We use a parameter k to control the width and depth of the architectures used in the following experiments. The Wide(k) networks have two convolutional layers of $4 \times k$ and $8 \times k$ filters followed by a $128 \times k$ fully connected layer. The Deep(k) networks have k convolutional filters with 8 filters followed by k convolutional filters with 16 filters.

Downsampling Similar to prior work, in all of our models we use strided convolutional layers with 4 by 4 kernels to downsample. When downsampling is not needed, we use 3 by 3 kernels without striding.

G.1 MNIST

Experimental setup For all MNIST experiments, we use the Adam optimizer with a learning rate of 0.001 with a batch size of 50. We schedule ϵ starting from 0.01 to the desired value over the first

---
Algorithm 3 Autodual: computing the bounds and dual of a general network
---

   input: Network operations f_{ij}, data point x, ball size ϵ
   // initialization
   $\nu_1^{(1)} := I$
   $\ell_2 := x - \epsilon$
   $u_2 := x + \epsilon$
   for $i = 2, \ldots, k-1$ **do**
     // initialize new dual layer
     Create dual layer operators A_{ji} and h_i from f_{ji}, ℓ_j and u_j for all $j \leq i$
     $\nu_i^{(i)} := I$.
     // update all dual variables
     for $j = 1, \ldots, i-1$ **do**
       $\nu_j^{(i)} := \sum_{k=1}^{j-1} A_{ki} \nu_j^{(k)}$
     end for
     // compute new bounds
     $\ell_{i+1} := x^T \nu_1^{(i)} - \epsilon \|\nu_1^{(i)}\|_: + \sum_{j=1}^{i} h_j(\nu_j^{(i)}, \ldots, \nu_i^{(i)})$
     $u_{i+1} := x^T \nu_1^{(i)} + \epsilon \|\nu_1^{(i)}\|_: - \sum_{j=1}^{i} h_j(-\nu_j^{(i)}, \ldots, -\nu_i^{(i)})$
     // $\| \cdot \|_:$ for a matrix here denotes the norm of all rows
   end for
   output: bounds $\{\ell_i, u_i\}_{i=2}^{k}$, dual layer operators A_{jk}, h_i

---

Table 3: Results on different widths and depths for MNIST

Dataset	Model	Epsilon	Robust error	Error
MNIST	Wide(1)	0.1	6.51%	2.27%
MNIST	Wide(2)	0.1	5.46%	1.55%
MNIST	Wide(4)	0.1	4.94%	1.33%
MNIST	Wide(8)	0.1	4.79%	1.32%
MNIST	Wide(16)	0.1	5.27%	1.36%
MNIST	Deep(1)	0.1	5.28%	1.78%
MNIST	Deep(2)	0.1	4.37%	1.28%
MNIST	Deep(3)	0.1	4.20%	1.15%

20 epochs, after which we decay the learning rate by a factor of 0.5 every 10 epochs for a total of 60 epochs.

Model width and depth We find that increasing the capacity of the model by simply making the network deeper and wider on MNIST is able boost performance. However, when the model becomes overly wide, the test robust error performance begins to degrade due to overfitting. These results are shown in Table 3.

G.2 CIFAR10

Experimental setup For all CIFAR10 experiments, we use the SGD optimizer with a learning rate of 0.05 with a batch size of 50. We schedule ϵ starting from 0.001 to the desired value over the first 20 epochs, after which we decay the learning rate by a factor of 0.5 every 10 epochs for a total of 60 epochs.