

# PROVEN: Certifying Robustness of Neural Networks with a Probabilistic Approach

Tsui-Wei Weng<sup>1</sup>, Pin-Yu Chen<sup>2</sup>, Lam M. Nguyen<sup>2</sup>, Mark S. Squillante<sup>2</sup>, Ivan Oseledets<sup>3</sup>, Luca Daniel<sup>1</sup>

<sup>1</sup>MIT, <sup>2</sup>IBM Research, <sup>3</sup>Skoltech

## Abstract

With deep neural networks providing state-of-the-art machine learning models for numerous machine learning tasks, quantifying the robustness of these models has become an important area of research. However, most of the research literature merely focuses on the *worst-case* setting where the input of the neural network is perturbed with noises that are constrained within an  $\ell_p$  ball; and several algorithms have been proposed to compute certified lower bounds of minimum adversarial distortion based on such worst-case analysis. In this paper, we address these limitations and extend the approach to a *probabilistic* setting where the additive noises can follow a given distributional characterization. We propose a novel probabilistic framework PROVEN to **PRO**babilitically **VE**rify Neural networks with statistical guarantees – i.e., PROVEN certifies the probability that the classifier’s top-1 prediction cannot be altered under any constrained  $\ell_p$  norm perturbation to a given input. Importantly, we show that it is possible to derive closed-form probabilistic certificates based on current state-of-the-art neural network robustness verification frameworks. Hence, the probabilistic certificates provided by PROVEN come naturally and with almost no overhead when obtaining the worst-case certified lower bounds from existing methods such as Fast-Lin, CROWN and CNN-Cert. Experiments on small and large MNIST and CIFAR neural network models demonstrate our probabilistic approach can achieve up to around 75% improvement in the robustness certification with at least a 99.99% confidence compared with the worst-case robustness certificate delivered by CROWN.

## 1 Introduction

Despite the recent advances and successes of deep neural networks in many machine learning tasks, it has been shown that adversarial examples exist and can be easily crafted, spanning from image classification [1] to speech recognition [2] to malware detection [3] and sparse regression [4], just to name a few. Although deep neural networks have achieved unprecedented performance in these applications, their lack of robustness against adversarial perturbations [5, 6] has raised serious concerns and has drawn a great deal of attention by the machine learning communities, as many safety-critical tasks cannot afford the potential risks incurred by adversarial examples.

While there is a growing interest in crafting adversarial examples with stronger attacks under various settings (e.g., white-box/grey-box/black-box attacks) and in developing effective defense strategies against adversarial attacks, the topic of assessing and verifying robustness properties of neural networks is equally important and challenging. Given a well-trained neural network model, we are interested in measuring its robustness on an arbitrary natural example  $\mathbf{x}_0$  by examining if the neighborhood of  $\mathbf{x}_0$  has the same prediction results; this serves as a robustness proxy for evaluating the ease with which one can turn  $\mathbf{x}_0$  into adversarial examples via adversarial manipulations. Conventionally, the concept of neighborhood is characterized by an  $\ell_p$  ball centered at  $\mathbf{x}_0$  with radius  $\epsilon$  for any  $p \geq 1$ , where larger  $\epsilon$  indicates greater robustness. Ideally, for robustness evaluation, we would like to find the smallest adversarial distortion imposed on  $\mathbf{x}_0$  that will change the model prediction, which is known as the *minimum adversarial distortion*. Unfortunately, it has been shown that computing the minimum adversarial distortion on neural networks with ReLU activations is an NP-complete problem [7, 8], and hence formal verification methods such as Reluplex [7] and Planet [9] are computationally demanding and cannot scale to large realistic networks.

As an alternative to minimum adversarial distortion, the concept of solving for a (non-trivial) *lower bound* on minimum distortion as a certified robustness metric has been recently proposed in [10, 11, 12, 13, 14, 15, 16, 17] and ap-

Table 1: Table of Notation

Notation	Definition	Notation	Definition
$K$	number of output classes	CDF	cumulative distribution function
$f : \mathbb{R}^{n_0} \rightarrow \mathbb{R}^K$	neural network classifier	pdf	probability density function
$\mathbf{x}_0 \in \mathbb{R}^{n_0}$	original input vector	$F_X$	CDF of a random variable $X$
$c = \operatorname{argmin}_i f_i(\mathbf{x}_0)$	predicted class of input $\mathbf{x}_0$	$f_X$	pdf of a random variable $X$
$g_t(\mathbf{x}) = f_c(\mathbf{x}) - f_t(\mathbf{x})$	margin function at $\mathbf{x}$ for class $t$	$\mathbb{P}[g_t(X) > a]$	probability that $g_t(X)$ is greater than $a$
$g_t^L(\mathbf{x}) : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$	linear lower bound of $g_t(\mathbf{x})$	$\gamma_L$	theoretical lower bound of $\mathbb{P}[g_t(X) > a]$
$g_t^U(\mathbf{x}) : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$	linear upper bound of $g_t(\mathbf{x})$	$\gamma_U$	theoretical upper bound of $\mathbb{P}[g_t(X) > a]$

appears to be a promising approach. Theoretical lower bounds have been derived in terms of local Lipschitz constants for continuously-differentiable classifiers [10] and neural networks with ReLU activations [14]. In addition, there have been some recent works on developing algorithms that are able to deliver *certified* lower bounds for fully-connected networks with ReLU activations [12, 14, 16] and general activations [15], and for general convolutional neural networks with commonly-used convolutional layers, pooling layers and residual blocks [17]. Here the term *certified* means that numerical values generated by these approaches are indeed deterministic lower bounds. In other words, all such approaches consider the setting where an input example can be perturbed by any perturbation bounded in an  $\ell_p$  ball, and thus their analyses all belong to the category of *worst-case* analysis.

On the other hand, additional alternative questions of interest include:

- (a) What are the corresponding guarantees under the situation where the input data point is perturbed with some random noises?
- (b) Can we provide confidence levels on the possibility that a given model will never be fooled under this probabilistic setting?

One way to address questions such as (a) and (b) is to relate them to the sensitivity of a target model  $f$  when the noise in the input is known to follow a given distribution. With prior knowledge on the noise distribution, this approach is expected to provide a more informative robustness certification in comparison to the prevailing worst-case analysis. More importantly, this statistical viewpoint of robustness indeed goes beyond the worst-case analysis considered in the adversarial attack setting. The probabilistic robustness certification is readily applicable to understanding the sensitivity and reliability of a target model subject to additive random noises under mild assumptions. For example, such random noises can be caused by data quantization, input preprocessing, or environmental background noises.

Unfortunately, to date there have been relatively little research efforts along these lines. Existing works [18, 19]

require some unverifiable or unrealistic assumptions on the classifier models and decision boundaries, rendering their results less useful in practice, especially for neural network models. This is indeed at the core of the motivation for our work – we seek to develop a probabilistic framework that can address questions such as (a) and (b), without imposing unverifiable assumptions on the models or decision boundaries.

In summary, we propose in this work a novel probabilistic framework **PROVEN** to **PRO**abilistically **VER**ify Neural network robustness. We show that it is possible to extend the conventional worst-case setting to a probabilistic setting based on existing worst-case certification frameworks with very little computational overhead, meaning that the probabilistic certificate comes naturally with nearly no additional overhead beyond worst-case robustness computations by methods such as Fast-Lin [14], CROWN [15] and CNN-Cert [17].

**Contributions.** We highlight the contributions of this paper as follows.

- A probabilistic framework **PROVEN** is proposed for certifying the robustness of neural networks under  $\ell_p$  norm-ball bounded threat models, when the input noise follows a given distributional characterization (zero-mean Gaussian or independent bounded random noises). The established theoretical results are based on an  $\ell_\infty$  constraint on the perturbation, but can be easily extended to other norms such as  $\ell_1$  and  $\ell_2$ .
- Experimental results on large neural networks trained on MNIST and CIFAR datasets show that the robustness certification metric (i.e., the certified lower bound) can be greatly improved under the proposed probabilistic framework in comparison with the worst-case analysis results, even when the statistical risk is small. For example, with a confidence level of 99.99%, which means the robustness metric is almost 100% guaranteed to be certified, the improvement provided by our probabilistic framework over the worst-case analysis can be as high as 78.9% for small networks and 32.8% for large networks.

- In addition to the noticeable improvement in the robustness metric, our probabilistic framework is a general tool that can be readily applied to neural networks with different activation functions, including tanh, sigmoid and arctan, as will be demonstrated in our experiments. Moreover, our proposed method is as computationally efficient as the worst-case analysis, since our probabilistic certificate has a closed-form and its parameters are by-products of worst-case certification frameworks (e.g., Fast-Lin [14], CROWN [15], and CNN-Cert [17]).

## 2 Background and related works

Given an input data example under a specified threat model, typically  $\ell_p$  norm-ball bounded perturbation attacks, the goal of formal verification for adversarial robustness aims to certify a perturbation level  $\epsilon$  such that the top-1 prediction will not be altered by any means. In other words, formal verification guarantees that no attack under the threat model can alter the top-1 prediction of the model if its attack perturbation is smaller than  $\epsilon$ . However, certifying the largest possible  $\epsilon$ , which is equivalent to finding the minimum perturbation required for a successful adversarial attack, has been shown to be an NP-complete problem [7] and thus it is computationally infeasible for large realistic networks. Alternatively, recent works have shown that solving for a lower bound of the minimum perturbation for formal verification can be made more scalable and computationally efficient [12, 14, 16, 20]. Some analytical lower bounds, based solely on model weights, have been derived [1, 21, 10, 13] but they can be loose, even becoming trivial lower bounds (close to 0), or they only apply to 1 or 2 hidden layers. It is worth noting that current robustness verification approaches mainly focus on a “worst-case” analysis, whereas our approach takes a probabilistic viewpoint for robustness certification. As will be evident in the following sections, our probabilistic framework approach PROVEN is able to certify a significantly larger  $\epsilon$  value than the corresponding worst-case analysis result with 99.99% certification guarantees. This indicates that while conventional worst-case robustness certification framework may be too conservative when we have some prior knowledge about the input perturbations (e.g. its distribution), our probabilistic framework will be more applicable in this situation.

In fact, deep neural networks are not only vulnerable to crafted adversarial noises but also to random noises: [22] shows that they can fool LeNet and AlexNet with additive Gaussian noises and [23] shows random perturbations can indeed fool VGG networks; [24] shows they fool Google Cloud Vision API by random Gaussian noises, suggesting random perturbation can be into a serious adversarial attack. Meanwhile, the robustness of classifiers to various kinds of random noises, such as uniform noise in the  $\ell_p$  unit ball and Gaussian noise with an arbitrary covariance matrix, has

been studied in [18]. This can apply to linear classifiers as well as non-linear classifiers with locally approximately flat decision boundaries. The bounds in the uniform  $\ell_p$  case depend on some universal constants, which may be arbitrarily large or small and can impact the quality of these bounds. Recently, the robustness of classifiers to perturbations under the assumption of Gaussian distributed latent input vectors has been studied in [19]. Moreover, all the results in [19] depend on the modulus of continuity constant, which can be arbitrarily large since one cannot control it. Due to these limitations, the bounds in these recent papers cannot be directly used to deliver certified robustness metrics. We note that our probabilistic framework is not limited to supervised neural network models - while this work is under review, a very recent workshop paper [25] takes a similar approach to verify some properties (e.g., monotonicity and convexity) of deep probabilistic models such as variational autoencoders (VAEs) [26] and conditional VAEs. The key differences are that in their setting the uncertainty source is the latent variable sampled from a distribution generated by the encoder (they consider only Gaussian distribution), whereas our uncertainty comes from input perturbations (we consider Gaussian as well as general bounded distributions) and our focus is to verify neural network classifiers in supervised learning instead of generative models. This also indicates a connection between probabilistic robustness certification of neural network classifiers and property verification of deep generative models.

## 3 PROVEN: a probabilistic framework to certify neural network robustness

In this section, we present a general probabilistic framework PROVEN together with related theoretical results to compute the certified bounds in probability that a classifier can never be fooled when the inputs of the classifier are perturbed with some given distributions. We first introduce a worst-case setting, where an input example can be perturbed by any perturbation bounded within an  $\ell_p$  ball, and present corresponding worst-case analysis results. We then show that it is possible to extend these worst-case analysis results to a probabilistic setting where the input perturbations follow some given distributions, and present our probabilistic framework and main theorem. Lastly, we provide *closed-form* probabilistic bounds for various probabilistic distributions that the input perturbations can follow.

### 3.1 Worst-case setting

Let  $f(\mathbf{x}) : \mathbb{R}^{n_0} \rightarrow \mathbb{R}^K$  denote a  $K$ -class neural network classifier of interest, which takes an input  $\mathbf{x}$  (e.g., an image) and outputs the corresponding logit scores over all classes. The ultimate goal is to efficiently find the largest  $\epsilon^*$  such that the original predicted class  $c$  always has a larger score  $f_c(\mathbf{x})$  than the score  $f_t(\mathbf{x})$  of targeted attack class  $t$  when

the input is perturbed within the  $\ell_p$  ball having radius  $\epsilon^*$ . Let  $g_t(\mathbf{x}) = f_c(\mathbf{x}) - f_t(\mathbf{x})$ , it means that we want to find the largest  $\epsilon^*$  such that  $g_t(\mathbf{x}) > 0$  for all  $\mathbf{x}$  satisfying  $\|\mathbf{x} - \mathbf{x}_0\|_p \leq \epsilon^*$  and  $c = \operatorname{argmax}_i f_i(\mathbf{x}_0), t \neq c$ . This  $\epsilon^*$  is a *certified lower bound* of the minimum adversarial distortion as first introduced in Section 1.

It has been shown in [14, 15, 17] that the output  $f_i(\mathbf{x})$  and the margin function  $g_t(\mathbf{x})$  of a general (convolutional) neural network classifier with general activation functions (including but not limited to ReLU, tanh, arctan, sigmoid) can be bounded by two linear functions. In other words, the authors show that

$$g_t^L(\mathbf{x}) \leq g_t(\mathbf{x}) \leq g_t^U(\mathbf{x}), \quad (1)$$

where  $g_t^L(\mathbf{x}) : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$  and  $g_t^U(\mathbf{x}) : \mathbb{R}^{n_0} \rightarrow \mathbb{R}$  are two linear functions

$$g_t^L(\mathbf{x}) = \mathbf{A}_{t,:}^L \mathbf{x} + d^L \quad \text{and} \quad g_t^U(\mathbf{x}) = \mathbf{A}_{t,:}^U \mathbf{x} + d^U \quad (2)$$

with  $\mathbf{A}_{t,:}^L, \mathbf{A}_{t,:}^U \in \mathbb{R}^{1 \times n_0}$  being two constant row vectors and  $d^L, d^U \in \mathbb{R}$  being two constants related to the network weights  $\mathbf{W}^{(k)}$  and biases  $\mathbf{b}^{(k)}$  as well as the parameters bounding the activation functions in each neuron. The superscripts  $L$  and  $U$  denote the parameters corresponding to the lower bound and the upper bound of  $g_t(\mathbf{x})$ .

As the network output is bounded, the positiveness of the lower bound of  $g_t(\mathbf{x})$  implies that  $g_t(\mathbf{x})$  is positive, i.e.,

$$g_t^L(\mathbf{x}) > 0 \implies g_t(\mathbf{x}) > 0. \quad (3)$$

Here, a *worst-case* analysis can be performed by minimizing the linear function  $g_t^L(\mathbf{x})$  over all possible inputs in the set  $\{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\|_p \leq \epsilon\}$ , which yields a closed-form solution as presented in [14, 15, 17]. Therefore, the condition of whether  $g_t^L(\mathbf{x}) > 0$  can be conveniently checked given some  $\epsilon$  using the closed-form solutions; the largest  $\epsilon$  such that  $g_t^L(\mathbf{x}) > 0$  is called the *certified lower bound*, which can be computed by bisection with respect to  $\epsilon$ .

### 3.2 Our proposed probabilistic framework: PROVEN

In addition to considering the worst-case condition for  $g_t(\mathbf{x}) > 0$  over the norm ball constrained on the input  $\{\mathbf{x} : \|\mathbf{x} - \mathbf{x}_0\|_p \leq \epsilon\}$ , we now show that it is possible to formulate a probabilistic setting and derive bounds with guarantees by building upon the results that the neural network output can be bounded by two linear functions [14, 15, 17]. We start by presenting the problem formulation of the probabilistic setting and then present our main theoretical results in Theorem 3.1.

**Problem formulation.** Consider a neural network classifier  $f(\mathbf{x})$  and an input example  $\mathbf{x}_0$ . Let the predicted class of  $\mathbf{x}_0$  be  $c$ , the targeted attack class  $t$ , and the margin function  $g_t(\mathbf{x}) = f_c(\mathbf{x}) - f_t(\mathbf{x})$ . Suppose the perturbed input

random vector  $X$  follows some given distribution  $\mathcal{D}$ , i.e.,  $X \sim \mathcal{D}$ . We are interested in the probability of the margin function  $g_t(\mathbf{x})$  being greater than some value  $a \in \mathbb{R}$ , i.e.,  $\mathbb{P}[g_t(X) > a]$ .

Given that the neural network  $f(\mathbf{x})$  is highly non-linear and non-convex in  $\mathbf{x}$ , it is hard to directly compute the distribution of  $g_t(X)$  given the input  $X \sim \mathcal{D}$ . Fortunately, we can still derive *analytic lower and upper bounds* for  $\mathbb{P}[g_t(X) > 0]$  with guarantees based on the result in the worst-case analysis that the margin function  $g_t(x)$  can be bounded by two linear functions as shown in (1). The following theorem provides such theoretical guarantees on  $\mathbb{P}[g_t(X) > 0]$ .

#### Theorem 3.1 (Probabilistic bounds of network output)

Let  $f(\mathbf{x})$  be a  $K$ -class neural network classifier function,  $\mathbf{x}_0$  an input example, and  $\epsilon$  such that  $\|\mathbf{x} - \mathbf{x}_0\|_p \leq \epsilon, p \geq 1$ . Let  $c = \operatorname{argmax}_i f_i(\mathbf{x}_0), t (\neq c)$ , be some targeted class and define the margin function  $g_t(\mathbf{x}) = f_c(\mathbf{x}) - f_t(\mathbf{x})$ . Suppose the input random vector  $X \in \mathbb{R}^{n_0}$  follows some given distribution  $\mathcal{D}$  with mean  $\mathbf{x}_0$  and let  $a \in \mathbb{R}$  be some real number. There exists an explicit lower bound  $\gamma_L$  and an explicit upper bound  $\gamma_U$  on the probability  $\mathbb{P}[g_t(X) > a]$  such that

$$\gamma_L \leq \mathbb{P}[g_t(X) > a] \leq \gamma_U, \quad (4)$$

where

$$\gamma_L = 1 - F_{g_t^L(X)}(a), \quad \gamma_U = 1 - F_{g_t^U(X)}(a), \quad (5)$$

$F_Z(z)$  is the cumulative distribution function (CDF) of the random variable  $Z$ , and  $g_t^L(\mathbf{x}), g_t^U(\mathbf{x})$  satisfy Equation (1).

*Proof.* Let  $h_1 : \mathbb{R}^d \rightarrow \mathbb{R}, h_2 : \mathbb{R}^d \rightarrow \mathbb{R}$ , and  $h_1(\mathbf{x}) \geq h_2(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}^d$ . Let  $X \in \mathbb{R}^d$  be a random vector and let  $Y_1 \in \mathbb{R}, Y_2 \in \mathbb{R}^+$  be two random variables. Since  $h_1(\mathbf{x}) \geq h_2(\mathbf{x})$  always hold, we can let  $h_1(X) = Y_1 + Y_2$  and  $h_2(X) = Y_1$  for some random value  $Y_2 \geq 0$ . We therefore have

$$\begin{aligned} \mathbb{P}[h_1(X) > a] &= \mathbb{P}[Y_1 + Y_2 > a] = \mathbb{P}[Y_1 > a - Y_2], \\ \mathbb{P}[h_2(X) > a] &= \mathbb{P}[Y_1 > a]. \end{aligned}$$

Since  $a - Y_2 \leq a$  for any random value  $Y_2 \geq 0$ , we obtain  $\mathbb{P}[Y_1 > a - Y_2] \geq \mathbb{P}[Y_1 > a]$  based on the fact that the cumulative distribution function is nondecreasing [27], which is equivalent to

$$\mathbb{P}[h_1(X) > a] \geq \mathbb{P}[h_2(X) > a]. \quad (6)$$

From the results in [14], we know that the relationships in Equation (1), i.e.,

$$g_t^L(\mathbf{x}) \leq g_t(\mathbf{x}) \leq g_t^U(\mathbf{x}),$$

satisfy  $\|\mathbf{x} - \mathbf{x}_0\|_p \leq \epsilon$  for all  $\mathbf{x}$ . Hence, upon applying Equation (6) to Equation (1) and using the fact that  $\mathbb{P}[Z > a] = 1 - F_Z(a)$ , we obtain

$$\gamma_L \leq \mathbb{P}[g_t(X) > a] \leq \gamma_U,$$

with  $\gamma_L = 1 - F_{g_t^L(X)}(a)$ ,  $\gamma_U = 1 - F_{g_t^U(X)}(a)$ .  $\square$

As discussed in Section 3.1, the neural network output and the margin function can be bounded by two linear functions [14, 15, 17]. Here, we take an additional step to investigate the relationship between the margin function and its linear bounds in the probabilistic setting. Specifically, Theorem 3.1 shows that the probability of the neural network margin function being greater than some value  $a$  can also be bounded by the CDFs of its linear bounds. Note that in the worst-case analysis of Section 3.1, we usually concern ourselves with the margin function  $g_t(x) > 0$ , i.e.,  $a = 0$ . Analogously, in the probabilistic setting, we concern ourselves with the probability of the margin function  $g_t(x) > 0$ . This is indeed the guarantee provided by Theorem 3.1: when the input  $X \sim \mathcal{D}$ , the result guarantees that the probability of  $g_t(X) > a$  is at least  $\gamma_L$  and at most  $\gamma_U$ .

### 3.3 Evaluating the probabilistic bounds

Theorem 3.1 provides us with a theoretical lower bound  $\gamma_L$  and upper bound  $\gamma_U$  for  $\mathbb{P}[g_t(X) > a]$ . In practice, we would like to numerically compute such bounds. Below we show it is possible to obtain explicit forms for  $\gamma_L$  and  $\gamma_U$  in terms of  $\mathbf{A}_{t,:}^L, \mathbf{A}_{t,:}^U, d^L, d^U$ , as well as the parameters of the probability distributions of input perturbations. By Theorem 3.1,  $\gamma_L$  and  $\gamma_U$  only depend on the CDFs  $F_{g_t^L(X)}$  and  $F_{g_t^U(X)}$ , and we observe that  $g_t^L(X)$  and  $g_t^U(X)$  are both linear functions of  $X$  as follows:

$$\begin{aligned} g_t^L(X) &= \sum_{i=1}^{n_0} \mathbf{A}_{t,i}^L X_i + d^L, \\ g_t^U(X) &= \sum_{i=1}^{n_0} \mathbf{A}_{t,i}^U X_i + d^U. \end{aligned}$$

Hence, the problem of computing the CDFs  $F_{g_t^L(X)}$  and  $F_{g_t^U(X)}$  becomes a problem of computing the CDFs of a weighted sum of  $X_i$  given  $X \sim \mathcal{D}$ . We primarily consider the following two cases:

- (i) When  $X_i$  are independent random variables with probability density function (pdf)  $f_{X_i}$ ;
- (ii) When  $X$  follows a multivariate normal distribution with mean  $\mathbf{x}_0$  and covariance  $\Sigma$ .

It also appears that these results may be extended to address some forms of negative correlation [28, 29].

#### 3.3.1 Case (i)

When  $X_i$  are independent random variables with probability density function  $f_{X_i}$ , there are two approaches for computing the CDFs of the weighted sum.

**Approach 1: Direct convolutions.** The pdf of the weighted sum is simply the convolution of the pdfs for each of the weighted random variables  $\mathbf{A}_{t,i}^L X_i$ . Specifically, we have

$$f_{\mathbf{A}_{t,:}^L X} = \bigotimes_{i=1}^{n_0} f_{\mathbf{A}_{t,i}^L X_i},$$

where  $\bigotimes_{i=1}^N h_i$  denotes convolution over the  $N$  functions  $h_1$  to  $h_N$ . The CDF of  $\mathbf{A}_{t,:}^L X$  can therefore be obtained from the pdf  $f_{\mathbf{A}_{t,:}^L X}$  and we obtain  $F_{g_t^L(X)}(z) = F_{\mathbf{A}_{t,:}^L X}(z - d^L)$ ; similarly,  $F_{g_t^U(X)}(z) = F_{\mathbf{A}_{t,:}^U X}(z - d^U)$ . Hence, we have

$$\gamma_L = 1 - F_{\mathbf{A}_{t,:}^L X}(a - d^L), \quad \gamma_U = 1 - F_{\mathbf{A}_{t,:}^U X}(a - d^U).$$

**Approach 2: Probabilistic inequalities.** Approach 1 is useful in cases where  $n_0$  is not large. However, for large  $n_0$ , it might not be easy to directly compute the CDF through convolutions. For such cases, an alternative approach can be applying the probabilistic inequalities on the CDFs. Since we want to provide *guarantees* on the probability in (4), we need to find a lower bound on  $\gamma_L$  and an upper bound on  $\gamma_U$  via the probabilistic inequalities. These results are given in the following corollary.

**Corollary 3.2** *Let  $X_i$  be bounded independent random variables with  $X_i \in [\mathbf{x}_{0i} - \epsilon, \mathbf{x}_{0i} + \epsilon], \forall i \in [n_0]$ , and symmetric around the mean  $\mathbf{x}_{0i}$ . Define*

$$\mu_L = \mathbf{A}_{t,:}^L \mathbf{x}_0 + d^L, \quad \mu_U = \mathbf{A}_{t,:}^U \mathbf{x}_0 + d^U.$$

*Then, we have*

$$\begin{aligned} \gamma_L &\geq \begin{cases} 1 - \exp\left(-\frac{(\mu_L - a)^2}{2\epsilon^2 \|\mathbf{A}_{t,:}^L\|_2^2}\right), & \text{if } \mu_L - a \geq 0 \\ 0, & \text{otherwise;} \end{cases} \\ \gamma_U &\leq \begin{cases} \exp\left(-\frac{(\mu_U - a)^2}{2\epsilon^2 \|\mathbf{A}_{t,:}^U\|_2^2}\right), & \text{if } -\mu_U + a \geq 0 \\ 1, & \text{otherwise.} \end{cases} \end{aligned}$$

*Proof.* Let  $W_i = \mathbf{A}_{t,i}^L (X_i - \mathbf{x}_{0i})$  and  $\mu_L = \mathbf{A}_{t,:}^L \mathbf{x}_0 + d^L$ . We then have  $-|\mathbf{A}_{t,i}^L| \epsilon \leq W_i \leq |\mathbf{A}_{t,i}^L| \epsilon$  where  $W_i$  is symmetric with respect to zero since  $X_i$  is symmetric. By using the fact that the sum of independent symmetric random variables is still a symmetric random variable [30], we derive

$$\begin{aligned} \gamma_L &= \mathbb{P}[g_t^L(X) > a] \\ &= \mathbb{P}\left[\sum_{i=1}^{n_0} W_i > a - \mu_L\right] \\ &= \mathbb{P}\left[\sum_{i=1}^{n_0} W_i < -a + \mu_L\right] \\ &= 1 - \mathbb{P}\left[\sum_{i=1}^{n_0} W_i \geq -a + \mu_L\right]. \end{aligned}$$

From the Hoeffding inequality [31], we obtain the following upper bound on the term  $\mathbb{P}[\sum_{i=1}^{n_0} W_i \geq -a + \mu_L]$  when  $-a + \mu_L > 0$ :

$$\mathbb{P}\left[\sum_{i=1}^{n_0} W_i \geq -a + \mu_L\right] \leq \exp\left(-\frac{(\mu_L - a)^2}{2\epsilon^2 \|\mathbf{A}_{t,:}^L\|_2^2}\right),$$

and thus  $\gamma_L \geq 1 - \exp\left(-\frac{(\mu_L - a)^2}{2\epsilon^2 \|\mathbf{A}_{t,:}^L\|_2^2}\right)$ . When  $-a + \mu_L \leq 0$ , we use the trivial bound of  $\gamma_L = 0$ . Similarly, for  $\gamma_U$ , we can define  $\mu_U$  correspondingly and directly apply the Hoeffding inequality to obtain  $\gamma_U \leq \exp\left(-\frac{(\mu_U - a)^2}{2\epsilon^2 \|\mathbf{A}_{t,:}^U\|_2^2}\right)$ , or use the trivial bound of  $\gamma_U = 1$ .  $\square$

### 3.3.2 Case (ii)

When  $X$  follows a multivariate normal distribution with mean  $\mathbf{x}_0$  and covariance  $\Sigma$ , we are able to obtain an explicit form for the CDFs  $F_{g_t^L(X)}$  and  $F_{g_t^U(X)}$  based on the fact that the sum of the normally distributed random variables still follows the normal distribution [30]. Note that we include here both cases where (a)  $X_i$  are independent Gaussian random variables ( $\Sigma$  is a diagonal matrix) and (b)  $X_i$  are correlated random variables ( $\Sigma$  is a general covariance matrix and positive semidefinite). The result is stated in the following corollary.

**Corollary 3.3** *Let  $X$  follow a multivariate normal distribution with mean  $\mathbf{x}_0$  and covariance  $\Sigma$ . Define*

$$\begin{aligned} \mu_L &= \mathbf{A}_{t,:}^L \mathbf{x}_0 + d^L, \quad \sigma_L^2 = \mathbf{A}_{t,:}^L \Sigma (\mathbf{A}_{t,:}^L)^\top, \\ \mu_U &= \mathbf{A}_{t,:}^U \mathbf{x}_0 + d^U, \quad \sigma_U^2 = \mathbf{A}_{t,:}^U \Sigma (\mathbf{A}_{t,:}^U)^\top, \end{aligned}$$

where  $\top$  denotes the transpose operator. We then have

$$\gamma_L \approx \frac{1}{2} - \frac{1}{2} \operatorname{erf}\left(\frac{a - \mu_L}{\sigma_L \sqrt{2}}\right), \quad \gamma_U \approx \frac{1}{2} - \frac{1}{2} \operatorname{erf}\left(\frac{a - \mu_U}{\sigma_U \sqrt{2}}\right)$$

with  $\operatorname{erf}(\cdot)$  as the error function.

*Proof.* The result is obtained in a straightforward manner from the fact [30] that if  $X \sim \mathcal{N}(\mu, \Sigma)$ , then its linear combination  $Z = wX + v$  also follows the normal distribution:  $Z \sim \mathcal{N}(w\mu + v, w\Sigma w^\top)$ . The CDF of  $Z$  is then given by  $\frac{1}{2}(1 + \operatorname{erf}(\frac{z - \mu_Z}{\sigma_Z \sqrt{2}}))$ , leading to the stated approximations.  $\square$

**Remark 3.4** *Note that in our framework, all possible inputs have to lie in the  $\ell_p$  ball with given radius  $\epsilon$ . Thus, in order to apply the Gaussian perturbation in our setting, we need to set an upper limit on the variance of the input such that 99.7% of the density is within the  $\ell_p$  ball, i.e., the 3- $\sigma$  rule. See Section 4 Methods for more details.*

**Connection to  $\ell_1$  and  $\ell_2$  norms.** Our foregoing probabilistic analysis is established under the  $\ell_\infty$  norm constraint. We note that this presented analysis can be easily extended to  $\ell_1$  and  $\ell_2$  norms by using the norm inequalities:  $\|\mathbf{x}\|_1 \leq \sqrt{n_0} \|\mathbf{x}\|_2 \leq n_0 \|\mathbf{x}\|_\infty$ .

## 4 Experiments

**Methods.** We apply Corollaries 3.2 and 3.3 to compute the largest  $\epsilon$  (denoted as  $\epsilon_{\text{PROVEN}}$ ) that PROVEN can certify with confidence of at least  $\gamma_L$  when the input follows the two cases discussed in Section 3.3. The certified lower bound computed by the worst-case analysis in [14] and [15] is denoted as  $\epsilon_{\text{worst-case}}$ . Below is the setting of the input distributions in our simulations:

- **Case (i).**  $X_i$  are independent random variables bounded in  $[\mathbf{x}_{0i} - \epsilon_{\text{worst-case}}, \mathbf{x}_{0i} + \epsilon_{\text{worst-case}}]$  with mean  $\mathbf{x}_{0i}$ . The results are presented in Table 2.
- **Case (ii).**  $X$  follows a multivariate normal distribution with mean  $\mathbf{x}_0$  and covariance  $\Sigma$ . We consider both situations where  $\Sigma$  is a positive diagonal matrix or a positive semidefinite matrix with diagonals whose square roots are less than or equal to  $\epsilon_{\text{worst-case}}/3$ . The results are presented in Figure 1.

Note that in all the Tables, we express  $\gamma_L$  as a percentage. We report  $\epsilon_{\text{PROVEN}}$  for the following values:  $\{(100 - \eta), 75, 50, 25, 5, 0\}\%$  where  $\eta = 10^{-2}$  and calculate the improvement of  $\epsilon_{\text{PROVEN}}$  over  $\epsilon_{\text{worst-case}}$  obtained by  $(100 - \eta)\%$  in the last column in Table 2 for Case (i). The results in Table 2 are averaged over 10 randomly selected images in the test sets. On the other hand, we also investigate how robust it is for the results in Table 2 by computing the average  $\epsilon_{\text{PROVEN}}$  over randomly chosen  $\{10, 50, 100\}$  images in 100 random trials. We report the mean and standard deviation in Table 4 and show that (a) the variation of using 10 sample average in Table 2 is less  $\sim 10\%$  and (b) the average  $\epsilon_{\text{PROVEN}}$  and improvement has less deviations when we use 50 or 100 samples.

**Model and Dataset.** We use the publicly available pre-trained models provided in [14] and [15] as classifier models, which are fully-connected feed-forward neural networks with ReLU activation as well as general activations including tanh, sigmoid and arctan on the MNIST [32] and CIFAR-10 [33] datasets. We denote a network with  $m$  layers and  $n$  neurons per layer as  $m \times [n]$  in the Tables.

**Implementation and Setup.** We implement PROVEN<sup>1</sup> in Python and perform experiments on a laptop with 8 Intel Cores i7-4700 HQ CPU at 2.40 GHz.

**Result on small and large ReLU networks.** We perform simulations on both small 2-3 layer MNIST networks with 20 neurons per layer and large 2-7 layer MNIST and CIFAR networks with 1024 or 2048 neurons per layer; the results are summarized in Tables 2a and 2b. These results show that on the small networks, PROVEN can certify up to 78.9% more with respect to the certified lower bound at the expense of decreasing the confidence by only  $\eta = 10^{-2}$ .

<sup>1</sup><https://github.com/lilyweng/PROVEN>

In other words, PROVEN guarantees that at least 99.99% of the  $\epsilon$  computed (e.g., 0.04394 in MNIST  $2 \times [20]$ , Table 2a) is a certified lower bound as compared to 0.02722 for the  $\epsilon_{\text{worst-case}}$  delivered by Fast-Lin[14], where the improvement we obtained for this model is 61.4%. Tables 2a and 2b are both ReLU activations and the only difference is the bounding techniques applied on the ReLU activations, where the bounding technique in Table 2b is adaptive and thus can certify more [15]. For large networks, PROVEN can certify up to 76%, which is significant. Interestingly, when the bounding technique is better, it also helps our probabilistic bounds – the improvement is significant, and even for the large CIFAR network with around 10,000 neurons, we can still obtain 10 – 15% improvement. For the cases where the input perturbations are Gaussians, the results are presented in Figure 1.

**Results on large networks with general activations.** We also ran experiments on various MNIST and CIFAR networks with non-ReLU activations, e.g., tanh, sigmoid and arctan. The results are summarized in Tables 2c to 2e. In comparison to the same architecture but with ReLU activations, the improvement of these activations are better than the non-adaptive bounding technique in general, and can achieve up to 32.8% on large networks. Note that the computational overhead of our approach compared to the worst-case analysis [14, 15] is very little, as we only need to perform a few binary searches on the  $\epsilon$  that will satisfy Corollary 3.2.

## 5 Conclusions and future works

We proposed a novel probabilistic framework PROVEN to certify the robustness of neural networks and derived theoretical bounds on the robustness certification with statistical guarantees. PROVEN is a general tool that can build on top of existing state-of-the-art neural network robustness certification algorithms (Fast-Lin, CROWN and CNN-Cert) and hence can be readily applied to certify fully-connected and convolutional neural networks with different activation functions. Experimental results on large neural networks demonstrated significant benefits of PROVEN over the standard worst-case analysis results.

Table 2: The largest  $\epsilon$  that PROVEN can certify with confidence of at least  $\gamma_L = \{99.99, 75, 50, 25, 5\}\%$  when  $X_i$  are independent random variables in Case (i). We compare the largest  $\epsilon$  that PROVEN can certify with 99.99% with the largest  $\epsilon$  from state-of-the-art worst-case robustness certification algorithms [14, 15] and show in the last column that PROVEN can certify more than the worst-case analysis by giving up 0.01% confidence.

(a) Relu activation

Certification Method Guarantees $\gamma_L$	Worst-case [14] 100% <sup>†</sup>	Our probabilistic approach: PROVEN					Certification improvement <sup>†</sup>
		99.99% <sup>†</sup>	75%	50%	25%	5%	
MNIST 2×[20]	0.02722	<b>0.04394</b>	0.04782	0.04824	0.04859	0.04897	<b>61.4%</b>
MNIST 3×[20]	0.02127	<b>0.02694</b>	0.02831	0.02847	0.02860	0.02874	<b>26.7%</b>
MNIST 2×[1024]	0.02904	<b>0.03572</b>	0.03758	0.03778	0.03796	0.03814	<b>23.0%</b>
MNIST 3×[1024]	0.02082	<b>0.02253</b>	0.02303	0.02309	0.02313	0.02318	<b>8.2 %</b>
MNIST 4×[1024]	0.00796	<b>0.00813</b>	0.00817	0.00818	0.00818	0.00818	<b>2.1 %</b>
CIFAR 5×[2048]	0.00183	<b>0.00186</b>	0.00186	0.00186	0.00186	0.00186	<b>1.6 %</b>
CIFAR 7×[1024]	0.00189	<b>0.00192</b>	0.00192	0.00193	0.00193	0.00193	<b>1.6 %</b>

(b) Relu activation with adaptive bounds

Certification Method Guarantees	Worst-case [15] 100% <sup>†</sup>	Our probabilistic approach: PROVEN					Certification improvement <sup>†</sup>
		99.99% <sup>†</sup>	75%	50%	25%	5%	
MNIST 2×[20]	0.02746	<b>0.04912</b>	0.05212	0.05246	0.05276	0.05307	<b>78.9 %</b>
MNIST 3×[20]	0.02236	<b>0.03828</b>	0.03966	0.03981	0.03995	0.04009	<b>71.2 %</b>
MNIST 2×[1024]	0.03158	<b>0.05560</b>	0.05756	0.05779	0.05798	0.05818	<b>76.1 %</b>
MNIST 3×[1024]	0.02397	<b>0.03524</b>	0.03583	0.03589	0.03595	0.03601	<b>47.1 %</b>
MNIST 4×[1024]	0.00962	<b>0.01288</b>	0.01293	0.01294	0.01295	0.01295	<b>33.9 %</b>
CIFAR 5×[2048]	0.00228	<b>0.00264</b>	0.00265	0.00265	0.00265	0.00265	<b>15.8 %</b>
CIFAR 7×[1024]	0.00189	<b>0.00209</b>	0.00210	0.00210	0.00210	0.00210	<b>10.6 %</b>

(c) Tanh activation

Certification Method Guarantees	Worst-case [15] 100% <sup>†</sup>	Our probabilistic approach: PROVEN					Certification improvement <sup>†</sup>
		99.99% <sup>†</sup>	75%	50%	25%	5%	
MNIST 2×[1024]	0.02232	<b>0.02915</b>	0.03005	0.03013	0.03022	0.03033	<b>30.6%</b>
MNIST 3×[1024]	0.01121	<b>0.01360</b>	0.01376	0.01378	0.01380	0.01381	<b>21.3 %</b>
MNIST 4×[1024]	0.00682	<b>0.00745</b>	0.00750	0.00750	0.00751	0.00751	<b>9.2 %</b>
CIFAR 5×[2048]	0.00081	<b>0.00085</b>	0.00085	0.00085	0.00085	0.00085	<b>4.9 %</b>

(d) Sigmoid activation

Certification Method Guarantees	Worst-case [15] 100% <sup>†</sup>	Our probabilistic approach: PROVEN					Certification improvement <sup>†</sup>
		99.99% <sup>†</sup>	75%	50%	25%	5%	
MNIST 2×[1024]	0.02785	<b>0.03285</b>	0.03404	0.03419	0.03426	0.03441	<b>18.0%</b>
MNIST 3×[1024]	0.01856	<b>0.02296</b>	0.02342	0.02348	0.02353	0.02358	<b>23.7 %</b>
MNIST 4×[1024]	0.01778	<b>0.02170</b>	0.02224	0.02229	0.02232	0.02237	<b>22.1 %</b>

(e) Arctan activation

Certification Method Guarantees	Worst-case [15] 100% <sup>†</sup>	Our probabilistic approach: PROVEN					Certification improvement <sup>†</sup>
		99.99% <sup>†</sup>	75%	50%	25%	5%	
MNIST 2×[1024]	0.02105	<b>0.02796</b>	0.02907	0.02915	0.02924	0.02936	<b>32.8%</b>
MNIST 3×[1024]	0.01250	<b>0.01462</b>	0.01486	0.01488	0.01490	0.01493	<b>17.0 %</b>
MNIST 4×[1024]	0.00726	<b>0.00829</b>	0.00836	0.00837	0.00838	0.00838	<b>14.2 %</b>
CIFAR 5×[2048]	0.00078	<b>0.00089</b>	0.00089	0.00089	0.00089	0.00089	<b>14.1 %</b>



Figure 1: We plot the improvement of the largest  $\epsilon$  certified by PROVEN with various confidence ( $\gamma_L = \{99.99, 75, 50, 25, 5\}\%$ ) over the largest  $\epsilon$  certified by worst-case robustness certification algorithms [14, 15]. We consider both input perturbations being independent/correlated Gaussian random variables as in Case (ii) and independent random variables as in Case (i). The  $x$ -axis label in the figure:  $\gamma_L$ ;  $y$ -axis label: Certification improvement of PROVEN over  $\epsilon_{\text{worst-case}}$ . The models are 2-4 layers MNIST networks with 1024 nodes per layer and ReLU activations.

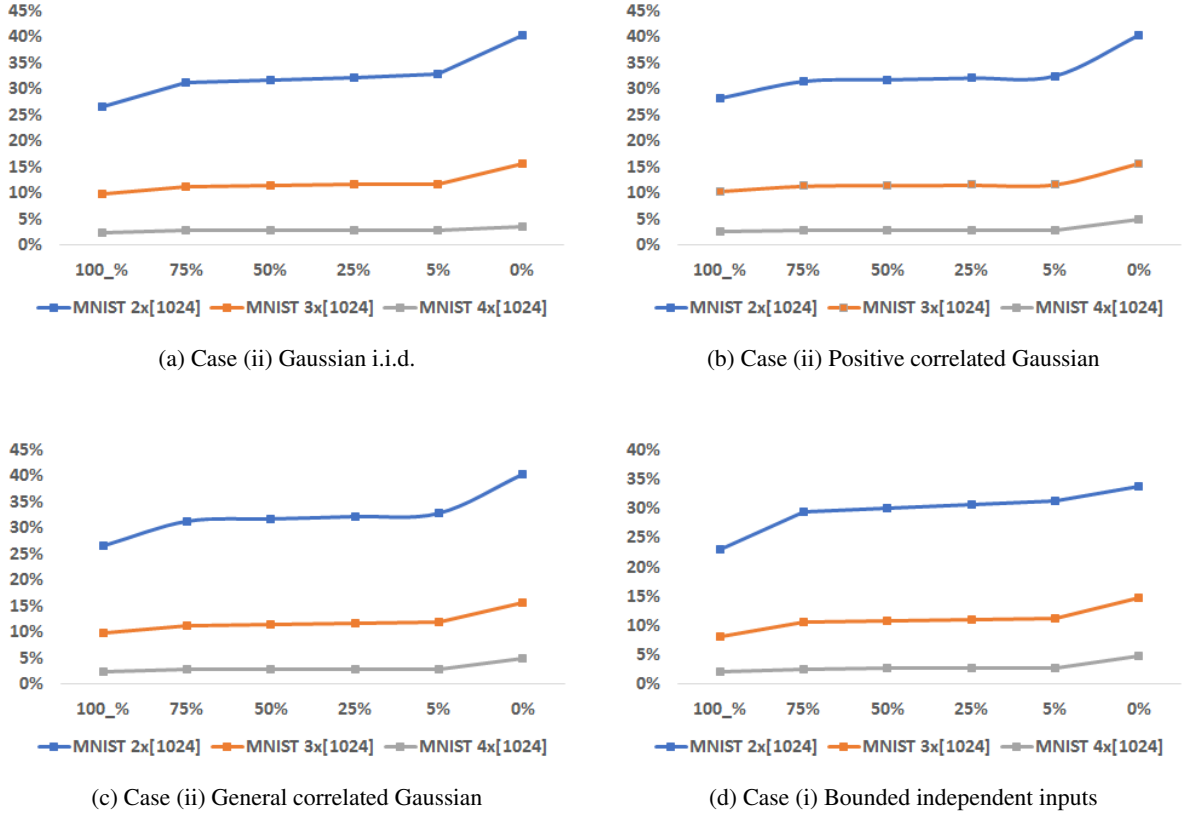


Table 3: Summary of the improvement of our approach (we certify the bound with at least 99.99% confidence) compared to  $\epsilon_{\text{worst-case}}$  [15].

model	Relu	Relu-ada	tanh	sigmoid	arctan
MNIST 2 $\times$ [1024]	23.0%	76.1%	30.6%	18.0%	32.8%
MNIST 3 $\times$ [1024]	8.2%	47.1%	21.3%	23.7%	17.0%
MNIST 4 $\times$ [1024]	2.1%	33.9%	9.2%	22.1%	14.2%

Table 4: With input perturbations are independent random variables in case (i), we randomly choose  $\{10, 50, 100\}$  input samples (images) in each trial and compute average of the largest  $\epsilon$  that can be certified by worst-case analysis [15] (denoted as  $\epsilon_{\text{worst-case}}$ ) and PROVEN with 99.99% confidence (denoted as  $\epsilon_{\text{PROVEN}}$ ) and the improved certification of  $\epsilon_{\text{PROVEN}}$  over  $\epsilon_{\text{worst-case}}$  (denoted as Improv.). We present the mean and std of the average  $\epsilon$  and improvements for  $\{10, 50, 100\}$  samples in total 100 random trials and it shows that the mean and std converge as number of sample increases.

(a) MNIST  $3 \times [1024]$ , ReLU activation with adaptive bounds

100 rand trials	10 samples			50 samples			100 samples		
	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.
Mean	0.02559	0.03703	44.75%	0.02581	0.03734	44.70%	0.02579	0.03733	44.74%
std	0.00165	0.00222	1.12%	0.00076	0.00102	0.57%	0.00054	0.00071	0.43%

(b) MNIST  $3 \times [1024]$ , tanh activation

100 rand trials	10 samples			50 samples			100 samples		
	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.
Mean	0.01195	0.01375	15.17%	0.01193	0.01374	15.22%	0.01192	0.01374	15.25%
std	0.00065	0.00068	2.66%	0.00030	0.00030	1.27%	0.00020	0.00021	0.77%

(c) MNIST  $4 \times [1024]$ , ReLU activation with adaptive bounds

100 rand trials	10 samples			50 samples			100 samples		
	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.
Mean	0.00998	0.01329	33.18%	0.00994	0.01325	33.24%	0.00997	0.01328	33.21%
std	0.00051	0.00066	0.57%	0.00021	0.00027	0.27%	0.00014	0.00018	0.15%

(d) MNIST  $3 \times [1024]$ , tanh activation

100 rand trials	10 samples			50 samples			100 samples		
	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.
Mean	0.01195	0.01375	15.17%	0.01193	0.01374	15.22%	0.01192	0.01374	15.25%
std	0.00065	0.00068	2.66%	0.00030	0.00030	1.27%	0.00020	0.00021	0.77%

(e) CIFAR  $5 \times [2048]$ , ReLU activation with adaptive bounds

100 rand trials	10 samples			50 samples			100 samples		
	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.
Mean	0.00224	0.00264	18.07%	0.00222	0.00262	17.93%	0.00222	0.00263	18.06%
std	0.00020	0.00025	2.39%	0.00009	0.00011	1.12%	0.00005	0.00006	0.55%

(f) CIFAR  $5 \times [2048]$ , arctan activation

100 rand trials	10 samples			50 samples			100 samples		
	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.
Mean	0.00091	0.00100	9.28%	0.00091	0.00100	9.32%	0.00092	0.00100	9.32%
std	0.00008	0.00009	3.17%	0.00003	0.00003	1.15%	0.00001	0.00002	0.56%

(g) CIFAR  $7 \times [1024]$ , ReLU activation with adaptive bound

100 rand trials	10 samples			50 samples			100 samples		
	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.	$\epsilon_{\text{worst-case}}$	$\epsilon_{\text{PROVEN}}$	Improv.
Mean	0.00176	0.00195	10.68%	0.00174	0.00192	10.73%	0.00174	0.00193	10.70%
std	0.00018	0.00020	1.87%	0.00007	0.00008	0.75%	0.00003	0.00004	0.37%

## References

- [1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.
- [2] M. M. Cisse, Y. Adi, N. Neverova, and J. Keshet, “Houdini: Fooling deep structured visual and speech recognition models with adversarial examples,” in *NIPS*, 2017.
- [3] Q. Wang, W. Guo, K. Zhang, A. G. Ororbia II, X. Xing, X. Liu, and C. L. Giles, “Adversary resistant deep neural networks with an application to malware detection,” in *SIGKDD*. ACM, 2017.
- [4] P.-Y. Chen, B. Vinzamuri, and S. Liu, “Is ordered weighted  $\ell_1$  regularized regression robust to adversarial perturbation? a case study on oscar,” *arXiv preprint arXiv:1809.08706*, 2018.
- [5] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” *ICLR, arXiv preprint arXiv:1412.6572*, 2015.
- [6] B. Biggio and F. Roli, “Wild patterns: Ten years after the rise of adversarial machine learning,” *arXiv preprint arXiv:1712.03141*, 2017.
- [7] G. Katz, C. Barrett, D. L. Dill, K. Julian, and M. J. Kochenderfer, “Reluplex: An efficient smt solver for verifying deep neural networks,” in *International Conference on Computer Aided Verification*. Springer, 2017, pp. 97–117.
- [8] A. Sinha, H. Namkoong, and J. Duchi, “Certifiable distributional robustness with principled adversarial training,” *ICLR, arXiv preprint arXiv:1710.10571*, 2018.
- [9] R. Ehlers, “Formal verification of piece-wise linear feed-forward neural networks,” in *International Symposium on Automated Technology for Verification and Analysis*. Springer, 2017, pp. 269–286.
- [10] M. Hein and M. Andriushchenko, “Formal guarantees on the robustness of a classifier against adversarial manipulation,” in *Advances in Neural Information Processing Systems*, 2017, pp. 2263–2273.
- [11] T.-W. Weng, H. Zhang, P.-Y. Chen, J. Yi, D. Su, Y. Gao, C.-J. Hsieh, and L. Daniel, “Evaluating the robustness of neural networks: An extreme value theory approach,” *ICLR, arXiv preprint arXiv:1801.10578*, 2018.
- [12] J. Z. Kolter and E. Wong, “Provable defenses against adversarial examples via the convex outer adversarial polytope,” *arXiv preprint arXiv:1711.00851*, 2017.
- [13] A. Raghunathan, J. Steinhardt, and P. Liang, “Certified defenses against adversarial examples,” *ICLR, arXiv preprint arXiv:1801.09344*, 2018.
- [14] T.-W. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, D. Boning, I. S. Dhillon, and L. Daniel, “Towards fast computation of certified robustness for relu networks,” *ICML, arXiv preprint arXiv:1804.09699*, 2018.
- [15] H. Zhang, T.-W. Weng, P.-Y. Chen, C.-J. Hsieh, and L. Daniel, “Efficient neural network robustness certification with general activation functions,” in *Advances in Neural Information Processing Systems (NIPS)*, 2018.
- [16] T. Gehr, M. Mirman, D. Drachler-Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev, “Ai2: Safety and robustness certification of neural networks with abstract interpretation,” in *IEEE Symposium on Security and Privacy (SP)*, vol. 00, 2018, pp. 948–963.
- [17] A. Boopathy, T.-W. Weng, P.-Y. Chen, S. Liu, and L. Daniel, “Cnn-cert: An efficient framework for certifying robustness of convolutional neural networks,” in *AAAI*, Jan 2019.
- [18] J.-Y. Franceschi, A. Fawzi, and O. Fawzi, “Robustness of classifiers to uniform  $\ell_p$  and gaussian noise,” *arXiv preprint arXiv:1802.07971*, 2018.
- [19] A. Fawzi, H. Fawzi, and O. Fawzi, “Adversarial vulnerability for any classifier,” *arXiv preprint arXiv:1802.08686*, 2018.
- [20] K. Dvijotham, R. Stanforth, S. Goyal, T. Mann, and P. Kohli, “A dual approach to scalable verification of deep networks,” *arXiv preprint arXiv:1803.06567*, 2018.
- [21] J. Peck, J. Roels, B. Goossens, and Y. Saeys, “Lower bounds on the robustness to adversarial perturbations,” in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 804–813.
- [22] A. Bibi, M. Alfadly, and B. Ghanem, “Analytic expressions for probabilistic moments of pl-dnn with gaussian input,” in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
- [23] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, “Robustness of classifiers: from adversarial to random noise,” in *Advances in Neural Information Processing Systems*, 2016, pp. 1632–1640.
- [24] H. Hosseini, B. Xiao, and R. Poovendran, “Google’s cloud vision api is not robust to noise,” in *Machine Learning and Applications (ICMLA), 2017 16th IEEE International Conference on*. IEEE, 2017, pp. 101–105.

- [25] K. Dvijotham, A. F. Marta Garnelo, and P. Kohli, “Robustness of classifiers: from adversarial to random noise,” in *arXiv preprint arXiv:1812.02795*, 2018.
- [26] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [27] M. Shaked and G. Shanthikumar, *Stochastic Orders*. Springer, 2007.
- [28] A. Panconesi and A. Srinivasan, “Randomized distributed edge coloring via an extension of the chernoff-hoeffding bounds,” *SIAM Journal on Computing*, vol. 26, no. 2, pp. 350–368, 1997.
- [29] D. P. Dubhashi and D. Ranjan, “Balls and bins: A study in negative dependence,” *Random Structures and Algorithms*, vol. 13, no. 2, pp. 99–124, 1998.
- [30] Y. S. Chow and H. Teicher, *Probability Theory: Independence, Interchangeability, Martingales*, 3rd ed. Springer, 2003.
- [31] S. I. Resnick, *A Probability Path*. Birkhäuser, 2014.
- [32] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [33] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” Citeseer, Tech. Rep., 2009.