Certified Adversarial Robustness via Randomized Smoothing

Jeremy Cohen [1] **Elan Rosenfeld** [1] **J. Zico Kolter** [1 2]

Abstract

We show how to turn any classifier that classifies well under Gaussian noise into a new classifier that is certifiably robust to adversarial perturbations under the ℓ_2 norm. While this "randomized smoothing" technique has been proposed before in the literature, we are the first to provide a tight analysis, which establishes a close connection between ℓ_2 robustness and Gaussian noise. We use the technique to train an ImageNet classifier with e.g. a certified top-1 accuracy of 49% under adversarial perturbations with ℓ_2 norm less than 0.5 (=127/255). Smoothing is the only approach to certifiably robust classification which has been shown feasible on full-resolution ImageNet. On smaller-scale datasets where competing approaches to certified ℓ_2 robustness are viable, smoothing delivers higher certified accuracies. The empirical success of the approach suggests that provable methods based on randomization at prediction time are a promising direction for future research into adversarially robust classification. Code and models are available at `http://github.com/locuslab/smoothing`.

1. Introduction

Modern image classifiers achieve high accuracy on i.i.d. test sets but are not robust to small, adversarially-chosen perturbations of their inputs (Szegedy et al., 2014; Biggio et al., 2013). Given an image x correctly classified by, say, a neural network, an adversary can usually engineer an adversarial perturbation δ so small that $x + \delta$ looks just like x to the human eye, yet the network classifies $x + \delta$ as a different, incorrect class. Many works have proposed heuristic methods for training classifiers intended to be robust to adversarial perturbations. However, most of these heuristics have been subsequently shown to fail against suitably powerful adversaries (Carlini & Wagner, 2017; Athalye et al., 2018; Uesato et al., 2018). In response, a line of work on *certifiable robustness* studies classifiers whose prediction at any point x is verifiably constant within some set around x (e.g. Wong & Kolter, 2018; Raghunathan et al., 2018a). In most of these works, the robust classifier takes the form of a neural network. Unfortunately, all existing approaches for certifying the robustness of neural networks have trouble scaling to networks that are large and expressive enough to solve problems like ImageNet.

One workaround is to look for robust classifiers that are not neural networks. In this paper, we analyze an operation we call *randomized smoothing*[1] which transforms any arbitrary base classifier f into a new "smoothed classifier" g that is certifiably robust in ℓ_2 norm. Let f be an arbitrary classifier which maps inputs $\mathbb{R}^d$ to classes $\mathcal{Y}$. For any input x, the smoothed classifier's prediction $g(x)$ is defined to be the class which f is most likely to classify the random variable $\mathcal{N}(x, \sigma^2 I)$ as. That is, $g(x)$ returns the most probable prediction by f of random Gaussian corruptions of x.

If the base classifier f is most likely to classify $\mathcal{N}(x, \sigma^2 I)$ as x's correct class, then the smoothed classifier g will be

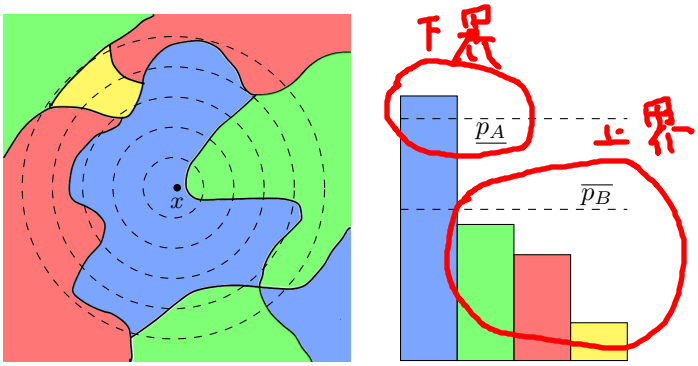


Figure 1. Evaluating the smoothed classifier at an input x. **Left**: the decision regions of the base classifier f are drawn in different colors. The dotted lines are the level sets of the distribution $\mathcal{N}(x, \sigma^2 I)$. **Right**: the distribution $f(\mathcal{N}(x, \sigma^2 I))$. As discussed below, $\underline{p_A}$ is a lower bound on the probability of the top class and $\overline{p_B}$ is an upper bound on the probability of each other class. Here, $g(x)$ is "blue."

pA 是top类概率的下限，pB 是其他类概率的上限。

[1]We adopt this term because it has been used to describe as similar technique in a different context (Duchi et al., 2012).

平滑分类器的预测 g(x) 被定义为 f 最有可能将随机变量 N(x; σ2I) 分类为的类

[1]Carnegie Mellon University [2]Bosch Center for AI. Correspondence to: Jeremy Cohen <jeremycohen@cmu.edu>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

只需估计f将N(x; σ2I)分类为每个类的概率，就可以验证g的预测在任何输入x周围的I2球内是恒定的。

f 将 N(x; σ2I) 分类为最可能类的概率越高，g 可证明返回该类的 x 周围的 '2 半径越大。

correct at x. But the smoothed classifier g will also possess a desirable property that the base classifier may lack: one can verify that g's prediction is constant within an ℓ_2 ball around any input x, simply by estimating the probabilities with which f classifies $\mathcal{N}(x, \sigma^2 I)$ as each class. The higher the probability with which f classifies $\mathcal{N}(x, \sigma^2 I)$ as the most probable class, the larger the ℓ_2 radius around x in which g provably returns that class.

这两个保证都是松散的，因为平滑分类器 g 比保证表明的更稳健。

Lecuyer et al. (2019) proposed randomized smoothing as a provable adversarial defense, and used it to train the first certifiably robust classifier for ImageNet. Subsequently, Li et al. (2018) proved a stronger robustness guarantee. However, both of these guarantees are loose, in the sense that the smoothed classifier g is *provably always* more robust than the guarantee indicates. In this paper, we prove the first tight robustness guarantee for randomized smoothing. Our analysis reveals that smoothing with Gaussian noise naturally induces certifiable robustness under the ℓ_2 norm. We suspect that other, as-yet-unknown noise distributions might induce robustness to other perturbation sets such as general ℓ_p norm balls.

相反，我们为这两个任务提供了蒙特卡罗算法，这些算法保证以任意高的概率成功。

Randomized smoothing has one major drawback. If f is a neural network, it is not possible to *exactly* compute the probabilities with which f classifies $\mathcal{N}(x, \sigma^2 I)$ as each class. Therefore, it is not possible to exactly evaluate the smoothed classifier g or to exactly compute the radius in which g is robust. Instead, we present Monte Carlo algorithms for both tasks that are guaranteed to succeed with arbitrarily high probability.

Despite this drawback, randomized smoothing enjoys several compelling advantages over other certifiably robust classifiers proposed in the literature: it makes no assumptions about the base classifier's architecture, it is simple to implement and understand, and, most importantly, it permits the use of arbitrarily large neural networks as the base classifier. In contrast, other certified defenses do not currently scale to large networks. Indeed, smoothing is the only certified adversarial defense which has been shown feasible on the full-resolution ImageNet classification task.

Table 1. Approximate certified accuracy on ImageNet. Each row shows a radius r, the best hyperparameter σ for that radius, the approximate certified accuracy at radius r of the corresponding smoothed classifier, and the standard accuracy of the corresponding smoothed classifier. To give a sense of scale, a perturbation with ℓ_2 radius 1.0 could change one pixel by 255, ten pixels by 80, 100 pixels by 25, or 1000 pixels by 8. Random guessing on ImageNet would attain 0.1% accuracy.

ℓ_2 RADIUS	BEST σ	CERT. ACC (%)	STD. ACC(%)
0.5	0.25	49	67
1.0	0.50	37	57
2.0	0.50	19	57
3.0	1.00	12	44

Figure 2. The smoothed classifier's prediction at an input x (left) is defined as the most likely prediction by the base classifier on random Gaussian corruptions of x (right; $\sigma = 0.5$). Note that this Gaussian noise is much larger in magnitude than the adversarial perturbations to which g is provably robust. One interpretation of randomized smoothing in high dimension is that these large random perturbations "drown out" small adversarial perturbations.

对高维随机平滑的一种解释是，这些大的随机扰动“淹没”了小的对抗性扰动。

We use randomized smoothing to train state-of-the-art certifiably ℓ_2-robust ImageNet classifiers; for example, one of them achieves 49% provable top-1 accuracy under adversarial perturbations with ℓ_2 norm less than 127/255 (Table 1). We also demonstrate that on smaller-scale datasets like CIFAR-10 and SHVN, where competing approaches to certified ℓ_2 robustness are feasible, randomized smoothing can deliver better certified accuracies, both because it enables the use of larger networks and because it does not constrain the expressivity of the base classifier.

2. Related Work

Many works have proposed classifiers intended to be robust to adversarial perturbations. These approaches can be broadly divided into *empirical* defenses, which empirically seem robust to known adversarial attacks, and *certified* defenses, which are *provably* robust to certain kinds of adversarial perturbations.

Empirical defenses The most successful empirical defense to date is *adversarial training* (Goodfellow et al., 2015; Kurakin et al., 2017; Madry et al., 2018), in which adversarial examples are found during training (often using projected gradient descent) and added to the training set. Unfortunately, it is typically impossible to tell whether a prediction by an empirically robust classifier is truly robust to adversarial perturbations; the most that can be said is that a specific attack was unable to find any. In fact, many heuristic defenses proposed in the literature were later "broken" by stronger adversaries (Carlini & Wagner, 2017; Athalye et al., 2018; Uesato et al., 2018; Athalye & Carlini, 2018).

为了给出尺度感，半径为 2 的扰动 1.0 可以将 1 个像素更改为 255、10 个像素乘以 80、100 像素乘以 25 或 1000 像素乘以 8。ImageNet 上的随机猜测将达到 0.1% 的准确度

一些工作提出了用于证明经过通用训练的网络的稳健性的算法，而其他作品（提出了一种稳健的训练方法和一种互补的认证机制。

Aiming to escape this cat-and-mouse game, a growing body of work has focused on defenses with formal guarantees.

Certified defenses A classifier is said to be *certifiably robust* if for any input x, one can easily obtain a guarantee that the classifier's prediction is constant within some set around x, often an ℓ_2 or ℓ_∞ ball. In most work in this area, the certifiably robust classifier is a neural network. Some works propose algorithms for certifying the robustness of generically trained networks, while others (Wong & Kolter, 2018; Raghunathan et al., 2018a) propose both a robust training method and a complementary certification mechanism.

Certification methods are either *exact* (a.k.a "complete") or *conservative* (a.k.a "sound but incomplete"). In the context of ℓ_p norm-bounded perturbations, exact methods take a classifier g, input x, and radius r, and report whether or not there exists a perturbation δ within $\|\delta\| \leq r$ for which $g(x) \neq g(x+\delta)$. In contrast, conservative methods either certify that no such perturbation exists or decline to make a certification; they may decline even when it is true that no such perturbation exists. Exact methods are usually based on Satisfiability Modulo Theories (Katz et al., 2017; Carlini et al., 2017; Ehlers, 2017; Huang et al., 2017) or mixed integer linear programming (Cheng et al., 2017; Lomuscio & Maganti, 2017; Dutta et al., 2017; Fischetti & Jo, 2018; Bunel et al., 2018). Unfortunately, no exact methods have been shown to scale beyond moderate-sized (100,000 activations) networks (Tjeng et al., 2019), and networks of that size can only be verified when they are trained in a manner that impairs their expressivity.

Conservative certification is more scalable. Some conservative methods bound the *global* Lipschitz constant of the neural network (Gouk et al., 2018; Tsuzuku et al., 2018; Anil et al., 2019; Cisse et al., 2017), but these approaches tend to be very loose on expressive networks. Others measure the *local* smoothness of the network in the vicinity of a particular input x. In theory, one could obtain a robustness guarantee via an upper bound on the local Lipschitz constant of the network (Hein & Andriushchenko, 2017), but computing this quantity is intractable for general neural networks. Instead, a panoply of practical solutions have been proposed in the literature (Wong & Kolter, 2018; Wang et al., 2018a;b; Raghunathan et al., 2018a;b; Wong et al., 2018; Dvijotham et al., 2018b;a; Croce et al., 2019; Gehr et al., 2018; Mirman et al., 2018; Singh et al., 2018; Gowal et al., 2018; Weng et al., 2018a; Zhang et al., 2018). Two themes stand out. Some approaches cast verification as an optimization problem and import tools such as relaxation and duality from the optimization literature to provide conservative guarantees (Wong & Kolter, 2018; Wong et al., 2018; Raghunathan et al., 2018a;b; Dvijotham et al., 2018b;a). Others step through the network layer by layer, maintaining at each layer an outer approximation of the set of activations reachable by a perturbed input (Mirman et al., 2018; Singh et al., 2018; Gowal et al., 2018; Weng et al., 2018a; Zhang et al., 2018). None of these local certification methods have been shown to be feasible on networks that are large and expressive enough to solve modern machine learning problems like the ImageNet classification task. Also, all method either assume specific network architectures (e.g. ReLU activations or a layered feedforward structure) or require extensive customization for new network architectures.

Related work involving noise Prior works have proposed using a network's robustness to Gaussian noise as a proxy for its robustness to adversarial perturbations (Weng et al., 2018b; Ford et al., 2019), and have suggested that Gaussian data augmentation could supplement or replace adversarial training (Zantedeschi et al., 2017; Kannan et al., 2018). Smilkov et al. (2017) observed that averaging a classifier's input gradients over Gaussian corruptions of an image yields very interpretable saliency maps. The robustness of neural networks to random noise has been analyzed both theoretically (Fawzi et al., 2016; Franceschi et al., 2018) and empirically (Dodge & Karam, 2017). Finally, Webb et al. (2019) proposed a statistical technique for estimating the noise robustness of a classifier more efficiently than naive Monte Carlo simulation; we did not use this technique since it appears to lack formal high-probability guarantees. While these works hypothesized relationships between a neural network's robustness to random noise and *the same network's* robustness to adversarial perturbations, randomized smoothing instead uses a classifier's robustness to random noise *to create a new classifier* robust to adversarial perturbations.

Randomized smoothing Randomized smoothing has been studied previously for adversarial robustness. Several works (Liu et al., 2018; Cao & Gong, 2017) proposed similar techniques as heuristic defenses, but did not prove any guarantees. Lecuyer et al. (2019) used inequalities from the differential privacy literature to prove an ℓ_2 and ℓ_1 robustness guarantee for smoothing with Gaussian and Laplace noise, respectively. Subsequently, Li et al. (2018) used tools from information theory to prove a stronger ℓ_2 robustness guarantee for Gaussian noise. However, all of these robustness guarantees are loose. In contrast, we prove a tight robustness guarantee in ℓ_2 norm for randomized smoothing with Gaussian noise.

3. Randomized smoothing 本文方案

Consider a classification problem from $\mathbb{R}^d$ to classes $\mathcal{Y}$. As discussed above, randomized smoothing is a method for constructing a new, "smoothed" classifier g from an arbitrary base classifier f. When queried at x, the smoothed classifier g returns whichever class the base classifier f is most likely

.

认证方法要么是精确的（又名"完整"），要么是保守的（又名"健全但不完整"）

可满足性模理论

混合整数线性规划

限制了神经网络的全局 Lipschitz 常数

to return when x is perturbed by isotropic Gaussian noise:

$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(x+\varepsilon) = c) \quad (1)$$
$$\text{where } \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

An equivalent definition is that $g(x)$ returns the class c whose pre-image $\{x' \in \mathbb{R}^d : f(x') = c\}$ has the largest probability measure under the distribution $\mathcal{N}(x, \sigma^2 I)$. The noise level σ is a hyperparameter of the smoothed classifier g which controls a robustness/accuracy tradeoff; it does not change with the input x. We leave undefined the behavior of g when the argmax is not unique.

We will first present our robustness guarantee for the smoothed classifier g. Then, since it is not possible to exactly evaluate the prediction of g at x or to certify the robustness of g around x, we will give Monte Carlo algorithms for both tasks that succeed with arbitrarily high probability.

3.1. Robustness guarantee

Suppose that when the base classifier f classifies $\mathcal{N}(x, \sigma^2 I)$, the most probable class c_A is returned with probability p_A, and the "runner-up" class is returned with probability p_B. Our main result is that smoothed classifier g is robust around x within the ℓ_2 radius $R = \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$, where Φ^{-1} is the inverse of the standard Gaussian CDF. This result also holds if we replace p_A with a lower bound $\underline{p_A}$ and we replace p_B with an upper bound $\overline{p_B}$.

"亚军"类以概率 pB 返回

定理1

Theorem 1. *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function, and let $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. Let g be defined as in (1). Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ satisfy:*

$$\mathbb{P}(f(x+\varepsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c) \quad (2)$$

Then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < R$, where

$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})) \quad (3)$$

We now make several observations about Theorem 1:

- Theorem 1 assumes nothing about f. This is crucial since it is unclear which well-behavedness assumptions, if any, are satisfied by modern deep architectures.
- The certified radius R is large when: (1) the noise level σ is high, (2) the probability of the top class c_A is high, and (3) the probability of each other class is low.
- The certified radius R goes to ∞ as $\underline{p_A} \to 1$ and $\overline{p_B} \to 0$. This should sound reasonable: the Gaussian distribution is supported on all of $\mathbb{R}^d$, so the only way that $f(x+\varepsilon) = c_A$ with probability 1 is if $f = c_A$ almost everywhere.

Both Lecuyer et al. (2019) and Li et al. (2018) proved ℓ_2 robustness guarantees for the same setting as Theorem 1, but with different, smaller expressions for the certified radius. However, our ℓ_2 robustness guarantee is *tight*: if (2) is all that is known about f, then it is impossible to certify an ℓ_2 ball with radius larger than R. In fact, it is impossible to certify any superset of the ℓ_2 ball with radius R:

Theorem 2. *Assume $\underline{p_A} + \overline{p_B} \leq 1$. For any perturbation δ with $\|\delta\|_2 > R$, there exists a base classifier f consistent with the class probabilities (2) for which $g(x+\delta) \neq c_A$.*

Theorem 2 shows that Gaussian smoothing naturally induces ℓ_2 robustness: if we make no assumptions on the base classifier beyond the class probabilities (2), then the set of perturbations to which a Gaussian-smoothed classifier is provably robust is *exactly* an ℓ_2 ball.

The complete proofs of Theorems 1 and 2 are in Appendix A. We now sketch the proofs in the special case when there are only two classes.

定理1的证明过程

Theorem 1 (binary case). *Suppose $\underline{p_A} \in (\frac{1}{2}, 1]$ satisfies $\mathbb{P}(f(x+\varepsilon) = c_A) \geq \underline{p_A}$. Then $g(x+\delta) = c_A$ for all $\|\delta\|_2 < \sigma\Phi^{-1}(\underline{p_A})$.*

Proof sketch. Fix a perturbation $\delta \in \mathbb{R}^d$. To guarantee that $g(x+\delta) = c_A$, we need to show that f classifies the translated Gaussian $\mathcal{N}(x+\delta, \sigma^2 I)$ as c_A with probability $> \frac{1}{2}$. However, all we know about f is that f classifies $\mathcal{N}(x, \sigma^2 I)$ as c_A with probability $\geq \underline{p_A}$. This raises the question: out of all possible base classifiers f which classify $\mathcal{N}(x, \sigma^2 I)$ as c_A with probability $\geq \underline{p_A}$, which one f^* classifies $\mathcal{N}(x+\delta, \sigma^2 I)$ as c_A with the smallest probability? One can show using an argument similar to the Neyman-Pearson lemma (Neyman & Pearson, 1933) that this "worst-case" f^* is a linear classifier whose decision boundary is normal to the perturbation δ (Figure 3):

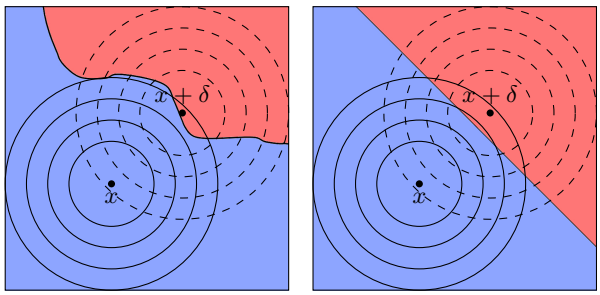
$$f^*(x') = \begin{cases} c_A & \text{if } \delta^T(x' - x) \leq \sigma\|\delta\|_2 \Phi^{-1}(\underline{p_A}) \\ c_B & \text{otherwise} \end{cases} \quad (4)$$

This "worst-case" f^* classifies $\mathcal{N}(x+\delta, \sigma^2 I)$ as c_A with probability $\Phi\left(\Phi^{-1}(\underline{p_A}) - \frac{\|\delta\|_2}{\sigma}\right)$. Therefore, to ensure that even the "worst-case" f^* classifies $\mathcal{N}(x+\delta, \sigma^2 I)$ as c_A with probability $> \frac{1}{2}$, we solve for those δ for which

$$\Phi\left(\Phi^{-1}(\underline{p_A}) - \frac{\|\delta\|_2}{\sigma}\right) > \frac{1}{2}$$

which is equivalent to the condition $\|\delta\|_2 < \sigma\Phi^{-1}(\underline{p_A})$. □

Theorem 2 is a simple consequence: for any δ with $\|\delta\|_2 > R$, the base classifier f^* defined in (4) is consistent with (2); yet if f^* is the base classifier, then $g(x+\delta) = c_B$.



同心圆是密度等值线

Figure 3. Illustration of f^* in two dimensions. The concentric circles are the density contours of $\mathcal{N}(x, \sigma^2 I)$ and $\mathcal{N}(x + \delta, \sigma^2 I)$. Out of all base classifiers f which classify $\mathcal{N}(x, \sigma^2 I)$ as c_A (blue) with probability $\geq \underline{p_A}$, such as both classifiers depicted above, the "worst-case" f^*, which classifies $\mathcal{N}(x + \delta, \sigma^2 I)$ as c_A with minimal probability, is the classifier depicted on the right: a linear classifier with decision boundary normal to the perturbation δ.

Figure 5 (left) plots our ℓ_2 robustness guarantee against the guarantees derived in prior work. Observe that our R is much larger than that of Lecuyer et al. (2019) and moderately larger than that of Li et al. (2018). Appendix I derives the other two guarantees using this paper's notation.

如果 f 是线性的，则平滑分类器 g 与基分类器 f 相同

Linear base classifier A two-class linear classifier $f(x) = \text{sign}(w^T x + b)$ is already certifiable: the distance from any input x to the decision boundary is $|w^T x + b|/\|w\|$, and no perturbation δ with ℓ_2 norm less than this distance can possibly change f's prediction. In Appendix B we show that if f is linear, then the smoothed classifier g is identical to the base classifier f. Moreover, we show that our bound (3) will certify the true robust radius $|w^T x + b|/\|w\|$, rather than a smaller, overconservative radius. Therefore, when f is linear, there always exists a perturbation δ just beyond the certified radius which changes g's prediction.

更高分辨率的图像可以在其类别区分内容被破坏之前容忍更高水平的各向同性高斯噪声。

Noise level can scale with image resolution Since our expression (3) for the certified radius does not depend explicitly on the data dimension d, one might worry that randomized smoothing is less effective for images of higher resolution — certifying a fixed ℓ_2 radius is "less impressive" for, say, a 224×224 image than for a 56×56 image. However, as illustrated by Figure 4, images in higher resolution can tolerate higher levels σ of isotropic Gaussian noise before their class-distinguishing content gets destroyed. As a consequence, in high resolution, smoothing can be performed with a larger σ, leading to larger certified radii. See Appendix G for a more rigorous version of this argument.

3.2. Practical algorithms

We now present practical Monte Carlo algorithms for evaluating $g(x)$ and certifying the robustness of g around x. More details can be found in Appendix C.

3.2.1. Prediction

Evaluating the smoothed classifier's prediction $g(x)$ requires identifying the class c_A with maximal weight in the categorical distribution $f(x + \varepsilon)$. The procedure described in pseudocode as PREDICT draws n samples of $f(x + \varepsilon)$ by running n noise-corrupted copies of x through the base classifier. Let $\hat{c}_A$ be the class which appeared the largest number of times. If $\hat{c}_A$ appeared much more often than any other class, then PREDICT returns $\hat{c}_A$. Otherwise, it abstains from making a prediction. We use the hypothesis test from Hung & Fithian (2019) to calibrate the abstention threshold so as to bound by α the probability of returning an incorrect answer. PREDICT satisfies the following guarantee:

Proposition 1. *With probability at least $1 - \alpha$ over the randomness in* PREDICT, PREDICT *will either abstain or return $g(x)$. (Equivalently: the probability that* PREDICT *returns a class other than $g(x)$ is at most α.)*

The function SAMPLEUNDERNOISE(f, x, num, σ) in the pseudocode draws num samples of noise, $\varepsilon_1 \dots \varepsilon_{\text{num}} \sim \mathcal{N}(0, \sigma^2 I)$, runs each $x + \varepsilon_i$ through the base classifier f, and returns a vector of class counts. BINOMPVALUE(n_A, $n_A + n_B$, p) returns the p-value of the two-sided hypothesis test that $n_A \sim \text{Binomial}(n_A + n_B, p)$.

Even if the true smoothed classifier g is robust at radius R, PREDICT will be vulnerable in a certain sense to adversarial perturbations with ℓ_2 norm slightly less than R. By engineering a perturbation δ for which $f(x + \delta + \varepsilon)$ puts mass just over $\frac{1}{2}$ on class c_A and mass just under $\frac{1}{2}$ on class c_B, an adversary can force PREDICT to abstain at a high rate. If this scenario is of concern, a variant of Theorem 1 could be proved to certify a radius in which $\mathbb{P}(f(x + \delta + \varepsilon) = c_A)$ is larger by some margin than $\max_{c \neq c_A} \mathbb{P}(f(x + \delta + \varepsilon) = c)$.

3.2.2. Certification

Evaluating *and* certifying the robustness of g around an input x requires not only identifying the class c_A with maximal weight in $f(x + \varepsilon)$, but also estimating a lower bound $\underline{p_A}$ on the probability that $f(x + \varepsilon) = c_A$ and an upper bound $\overline{p_B}$ on the probability that $f(x + \varepsilon)$ equals any other class. Doing all three of these at the same time in a statistically correct manner requires some care. One simple

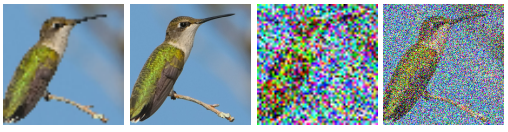


Figure 4. Left to right: clean 56 x 56 image, clean 224 x 224 image, noisy 56 x 56 image ($\sigma = 0.5$), noisy 224 x 224 image ($\sigma = 0.5$).

Pseudocode for certification and prediction

 # evaluate g at x
function PREDICT(f, σ, x, n, α)
  counts $\leftarrow$ SAMPLEUNDERNOISE(f, x, n, σ)
  $\hat{c}_A, \hat{c}_B \leftarrow$ top two indices in counts
  $n_A, n_B \leftarrow$ counts[$\hat{c}_A$], counts[$\hat{c}_B$]
  if BINOMPVALUE($n_A, n_A + n_B, 0.5$) $\leq \alpha$ **return** $\hat{c}_A$
  else return ABSTAIN

 # certify the robustness of g around x
function CERTIFY($f, \sigma, x, n_0, n, \alpha$)
  counts0 $\leftarrow$ SAMPLEUNDERNOISE(f, x, n_0, σ)
  $\hat{c}_A \leftarrow$ top index in counts0
  counts $\leftarrow$ SAMPLEUNDERNOISE(f, x, n, σ^2)
  $\underline{p_A} \leftarrow$ LOWERCONFBOUND(counts[$\hat{c}_A$], n, $1-\alpha$)
  if $\underline{p_A} > \frac{1}{2}$ **return** prediction $\hat{c}_A$ and radius $\sigma\,\Phi^{-1}(\underline{p_A})$
  else return ABSTAIN

solution is presented in pseudocode as CERTIFY: first, use a small number of samples from $f(x+\varepsilon)$ to take a guess at c_A; then use a larger number of samples to estimate $\underline{p_A}$; then simply take $\overline{p_B} = 1 - \underline{p_A}$.

Proposition 2. *With probability at least $1-\alpha$ over the randomness in* CERTIFY*, if* CERTIFY *returns a class $\hat{c}_A$ and a radius R (i.e. does not abstain), then g predicts $\hat{c}_A$ within radius R around x: $g(x+\delta) = \hat{c}_A \;\; \forall \; \|\delta\|_2 < R$.*

The function LOWERCONFBOUND($k, n, 1-\alpha$) in the pseudocode returns a one-sided $(1-\alpha)$ lower confidence interval for the Binomial parameter p given a sample $k \sim \text{Binomial}(n, p)$.

Certifying large radii requires many samples Recall from Theorem 1 that R approaches ∞ as $\underline{p_A}$ approaches 1. Unfortunately, it turns out that $\underline{p_A}$ approaches 1 so slowly with n that R also approaches ∞ very slowly with n. Consider the most favorable situation: $f(x) = c_A$ everywhere. This means that g is robust at radius ∞. But after observing n samples of $f(x+\varepsilon)$ which all equal c_A, the tightest (to our knowledge) lower bound would say that with probability least $1-\alpha$, $\underline{p_A} \geq \alpha^{(1/n)}$. Plugging $\underline{p_A} = \alpha^{(1/n)}$ and $\overline{p_B} = 1 - \underline{p_A}$ into (3) yields an expression for the certified radius as a function of n: $R = \sigma\,\Phi^{-1}(\alpha^{1/n})$. Figure 5 (right) plots this function for $\alpha = 0.001, \sigma = 1$. Observe that certifying a radius of 4σ with 99.9% confidence would require $\approx 10^5$ samples.

3.3. Training the base classifier

Theorem 1 holds regardless of how the base classifier f is trained. However, in order for g to classify the labeled example (x, c) correctly and robustly, f needs to consistently classify $\mathcal{N}(x, \sigma^2 I)$ as c. In high dimension, the Gaussian distribution $\mathcal{N}(x, \sigma^2 I)$ places almost no mass near its mode x. As a consequence, when σ is moderately high, the distribution of natural images has virtually disjoint support from the distribution of natural images corrupted by $\mathcal{N}(0, \sigma^2 I)$; see Figure 2 for a visual demonstration. Therefore, if the base classifier f is trained via standard supervised learning on the data distribution, it will see no noisy images during training, and hence will not necessarily learn to classify $\mathcal{N}(x, \sigma^2 I)$ with x's true label. Indeed, we observed empirically that when neural network base classifiers are trained on noiseless data, they cannot recognize noisy images.

Therefore, in this paper we follow Lecuyer et al. (2019) and train the base classifier with Gaussian data augmentation at variance σ^2. A justification for this procedure is provided in Appendix F. However, we suspect that there may be room to improve upon this training scheme, perhaps by training the base classifier so as to maximize the smoothed classifier's certified accuracy at some tunable radius r.

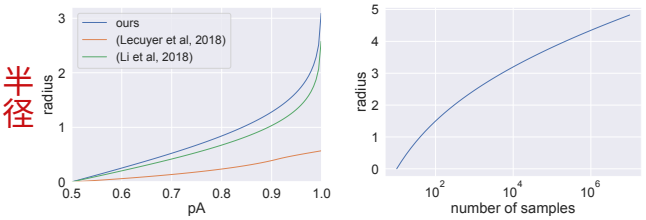


Figure 5. **Left**: Certified radius R as a function of $\underline{p_A}$ (with $\overline{p_B} = 1 - \underline{p_A}$ and $\sigma = 1$) under all three randomized smoothing bounds. **Right**: A plot of $R = \sigma\,\Phi^{-1}(\alpha^{1/n})$ for $\alpha = 0.001$ and $\sigma = 1$. The radius we can certify with high probability grows slowly with the number of samples, even in the *best* case where $f(x) = c_A$ everywhere.

4. Experiments

In adversarially robust classification, one metric of interest is the *certified test set accuracy* at radius r, defined as the fraction of the test set which g classifies correctly with a prediction that is certifiably robust within an ℓ_2 ball of radius r. However, if g is a randomized smoothing classifier, computing this quantity exactly is not possible, so we instead report the *approximate certified test set accuracy*, defined as the fraction of the test set which CERTIFY classifies correctly (without abstaining) and certifies robust with a radius $R \geq r$. Appendix D shows how to convert the approximate certified accuracy into a lower bound on the true certified accuracy that holds with high probability over the randomness in CERTIFY. However Appendix H.2 demonstrates that when α is small, the difference between these two quantities is negligible. Therefore, in our experiments we omit the step for simplicity and report approximate certified accuracies.

无论基分类器 f 如何训练，定理 1 都成立。 然而，为了让 g 正确且稳健地对标记示例 (x; c) 进行分类，f 需要一致地将 N(x; σ2I) 分类为 c。

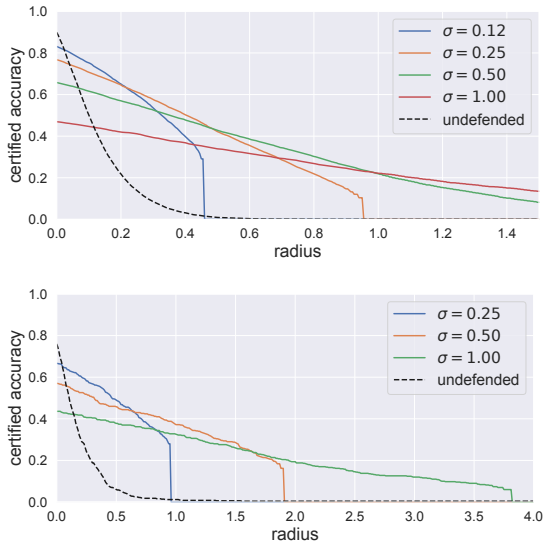


Figure 6. Approximate certified accuracy attained by randomized smoothing on CIFAR-10 (**top**) and ImageNet (**bottom**). The hyperparameter σ controls a robustness/accuracy tradeoff. The dashed black line is an upper bound on the empirical robust accuracy of an undefended classifier with the base classifier's architecture.

In all experiments, unless otherwise stated, we ran CERTIFY with $\alpha = 0.001$, so there was at most a 0.1% chance that CERTIFY returned a radius in which g was not truly robust. Unless otherwise stated, when running CERTIFY we used $n_0 = 100$ Monte Carlo samples for selection and $n = 100{,}000$ samples for estimation.

In the figures above that plot certified accuracy as a function of radius r, the certified accuracy always decreases gradually with r until reaching some point where it plummets to zero. This drop occurs because for each noise level σ and number of samples n, there is a hard upper limit to the radius we can certify with high probability, achieved when all n samples are classified by f as the same class.

ImageNet and CIFAR-10 results We applied randomized smoothing to CIFAR-10 (Krizhevsky, 2009) and ImageNet (Deng et al., 2009). On each dataset we trained several smoothed classifiers, each with a different σ. On CIFAR-10 our base classifier was a 110-layer residual network; certifying each example took 15 seconds on an NVIDIA RTX 2080 Ti. On ImageNet our base classifier was a ResNet-50; certifying each example took 110 seconds. We also trained a neural network with the base classifier's architecture on clean data, and subjected it to a DeepFool ℓ_2 adversarial attack (Moosavi-Dezfooli et al., 2016), in order to obtain an empirical upper bound on its robust accuracy. We certified the full CIFAR-10 test set and a subsample of 500 examples from the ImageNet test set.

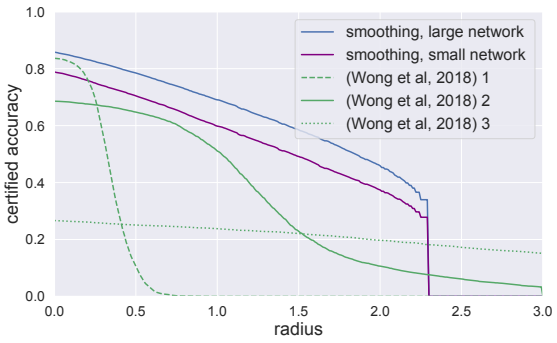


Figure 7. Comparison betwen randomized smoothing and Wong et al. (2018). Each green line is a small resnet classifier trained and certified using the method of Wong et al. (2018) with a different setting of its hyperparameter ϵ. The purple line is our method using the same small resnet architecture as the base classifier; the blue line is our method with a larger neural network as the base classifier. Wong et al. (2018) gives deterministic robustness guarantees, whereas smoothing gives high-probabaility guaranatees; therefore, we plot here the certified accuracy of Wong et al. (2018) against the "approximate" certified accuracy of smoothing.

Figure 6 plots the certified accuracy attained by smoothing with each σ. The dashed black line is the empirical upper bound on the robust accuracy of the base classifier architecture; observe that smoothing improves substantially upon the robustness of the undefended base classifier architecture. We see that σ controls a robustness/accuracy tradeoff. When σ is low, small radii can be certified with high accuracy, but large radii cannot be certified. When σ is high, larger radii can be certified, but smaller radii are certified at a lower accuracy. This observation echoes the finding in Tsipras et al. (2019) that adversarially trained networks with higher robust accuracy tend to have lower standard accuracy. Tables of these results are in Appendix E.

Figure 8 (**left**) plots the certified accuracy obtained using our Theorem 1 guarantee alongside the certified accuracy obtained using the analogous bounds of Lecuyer et al. (2019) and Li et al. (2018). Since our expression for the certified radius R is greater (and, in fact, tight), our bound delivers higher certified accuracies. Figure 8 (**middle**) projects how the certified accuracy would have changed had CERTIFY used more or fewer samples n (under the assumption that the relative class proportions in `counts` would have remained constant). Finally, Figure 8 (**right**) plots the certified accuracy as the confidence parameter α is varied. Observe that the certified accuracy is not very sensitive to α.

Comparison to baselines We compared randomized smoothing to three baseline approaches for certified ℓ_2 robustness: the duality approach from Wong et al. (2018), the Lipschitz approach from Tsuzuku et al. (2018), and the

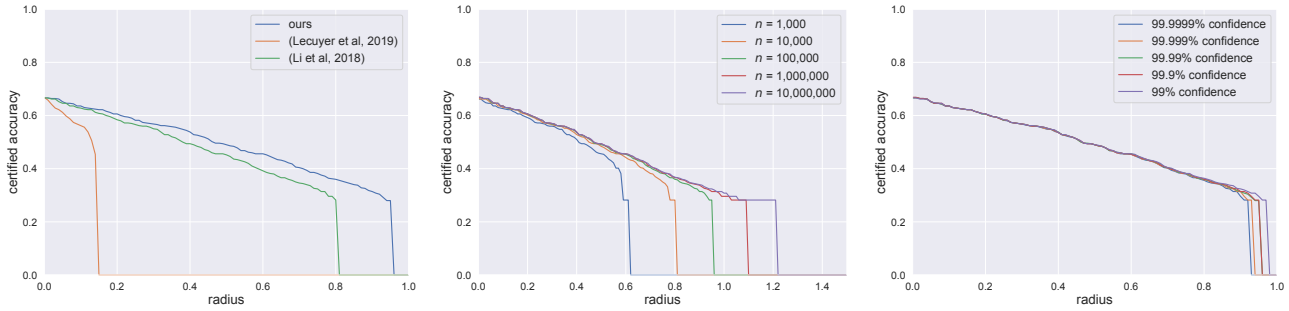


Figure 8. Experiments with randomized smoothing on ImageNet with $\sigma = 0.25$. **Left**: certified accuracies obtained using our Theorem 1 versus those obtained using the robustness guarantees derived in prior work. **Middle**: projections for the certified accuracy if the number of samples n used by CERTIFY had been larger or smaller. **Right**: certified accuracy as the failure probability α of CERTIFY is varied.

approach from Weng et al. (2018a); Zhang et al. (2018). The strongest baseline was Wong et al. (2018); we defer the comparison to the other two baselines to Appendix H.

In Figure 7, we compare the largest publicly released model from Wong et al. (2018), a small resnet, to two randomized smoothing classifiers: one which used the same small resnet architecture for its base classifier, and one which used a larger 110-layer resnet for its base classifier. First, observe that smoothing with the large 110-layer resnet substantially outperforms the baseline (across all hyperparameter settings) at all radii. Second, observe that smoothing with the small resnet also outperformed the method of Wong et al. (2018) at all but the smallest radii. We attribute this latter result to the fact that neural networks trained using the method of Wong et al. (2018) are "typically overregularized to the point that many filters/weights become identically zero," per that paper. In contrast, the base classifier in randomized smoothing is a fully expressive neural network.

Prediction It is computationally expensive to certify the robustness of g around a point x, since the value of n in CERTIFY must be very large. However, it is far cheaper to evaluate g at x using PREDICT, since n can be small. For example, when we ran PREDICT on ImageNet ($\sigma = 0.25$) using $n = 100$, making each prediction only took 0.15 seconds, and we attained a top-1 test accuracy of 65% (Appendix E).

As discussed earlier, an adversary can potentially force PREDICT to abstain with high probability. However, it is relatively rare for PREDICT to abstain on the actual data distribution. On ImageNet ($\sigma = 0.25$), PREDICT with failure probability $\alpha = 0.001$ abstained 12% of the time when $n = 100$, 4% when $n = 1000$, and 1% when $n = 10{,}000$.

Empirical tightness of bound When f is linear, the bound in Theorem 1 is tight, in that there always exists a class-changing perturbation just beyond the certified radius. Since deep neural networks are not linear, we empirically assessed the tightness of our bound by subjecting an ImageNet randomized smoothing classifier ($\sigma = 0.25$) to a projected gradient descent-style adversarial attack. For each example, we ran CERTIFY with $\alpha = 0.01$, and, if the example was correctly classified and certified robust at radius R, we tried finding an adversarial example for g within radius $1.5R$ and within radius $2R$. We succeeded 17% of the time at radius $1.5R$ and 53% of the time at radius $2R$. See Appendix J.3 for more details on the attack.

5. Conclusion

Theorem 2 establishes that smoothing with Gaussian noise naturally confers adversarial robustness in ℓ_2 norm: if we have no knowledge about the base classifier beyond the distribution of $f(x + \varepsilon)$, then the set of perturbations to which the smoothed classifier is provably robust is precisely an ℓ_2 ball. We suspect that smoothing with other noise distributions may lead to similarly natural robustness guarantees for other perturbation sets such as general ℓ_p norm balls.

Our strong empirical results suggest that randomized smoothing is a promising direction for future research into adversarially robust classification. Most empirical approaches (except PGD adversarial training) have been "broken," and provable approaches based on certifying neural network classifiers have not been shown to scale to networks of modern size. It seems to be computationally infeasible to reason in any sophisticated way about the decision boundaries of a large, expressive neural network. Randomized smoothing circumvents this problem: the smoothed classifier is not itself a neural network, though it leverages the discriminative ability of a neural network base classifier. To make the smoothed classifier robust, one need simply make the base classifier classify well under noise. In this way, randomized smoothing reduces the unsolved problem of adversarially robust classification to the comparably solved domain of supervised learning.

6. Acknowledgements

We thank Mateusz Kwaśnicki for help with Lemma 4 in the appendix, Aaditya Ramdas for pointing us toward the work of Hung & Fithian (2019), and Siva Balakrishnan for helpful discussions regarding the confidence interval in Appendix D. We thank Tolani Olarinre, Adarsh Prasad, Ben Cousins, and Ramon Van Handel for useful conversations. Finally, we are very grateful to Vaishnavh Nagarajan, Arun Sai Suggala, Shaojie Bai, Mikhail Khodak, Han Zhao, and Zachary Lipton for reviewing drafts of this work. Jeremy Cohen is supported by a grant from the Bosch Center for AI.

References

Anil, C., Lucas, J., and Grosse, R. B. Sorting out lipschitz function approximation. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Athalye, A. and Carlini, N. On the robustness of the cvpr 2018 white-box adversarial example defenses. *The Bright and Dark Sides of Computer Vision: Challenges and Opportunities for Privacy and Security*, 2018.

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks against machine learning at test time. *Joint European Conference on Machine Learning and Knowledge Discovery in Database*, 2013.

Blanchard, G. Lecture Notes, 2007. URL http://www.math.uni-potsdam.de/~blanchard/lectures/lect_2.pdf.

Bunel, R. R., Turkaslan, I., Torr, P., Kohli, P., and Mudigonda, P. K. A unified view of piecewise linear neural network verification. In *Advances in Neural Information Processing Systems 31*. 2018.

Cao, X. and Gong, N. Z. Mitigating evasion attacks to deep neural networks via region-based classification. *33rd Annual Computer Security Applications Conference*, 2017.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017.

Carlini, N., Katz, G., Barrett, C., and Dill, D. L. Provably minimally-distorted adversarial examples. *arXiv preprint arXiv: 1709.10207*, 2017.

Cheng, C.-H., Nührenberg, G., and Ruess, H. Maximum resilience of artificial neural networks. *International Symposium on Automated Technology for Verification and Analysis*, 2017.

Cisse, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.

Clopper, C. J. and Pearson, E. S. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):pp. 404–413, 1934. ISSN 00063444.

Croce, F., Andriushchenko, M., and Hein, M. Provable robustness of relu networks via maximization of linear regions. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Dodge, S. and Karam, L. A study and comparison of human and deep learning recognition performance under visual distortions. *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, 2017.

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 2012. URL https://doi.org/10.1137/110831659.

Dutta, S., Jha, S., Sanakaranarayanan, S., and Tiwari, A. Output range analysis for deep neural networks. *arXiv preprint arXiv:1709.09130*, 2017.

Dvijotham, K., Gowal, S., Stanforth, R., Arandjelovic, R., O'Donoghue, B., Uesato, J., and Kohli, P. Training verified learners with learned verifiers. *arXiv preprint arXiv:1805.10265*, 2018a.

Dvijotham, K., Stanforth, R., Gowal, S., Mann, T., and Kohli, P. A dual approach to scalable verification of deep networks. *Proceedings of the Thirty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, 2018b.

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

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems 29*. 2016.

Fischetti, M. and Jo, J. Deep neural networks and mixed integer linear optimization. *Constraints*, 23(3):296–309, July 2018.

Ford, N., Gilmer, J., and Cubuk, E. D. Adversarial examples are a natural consequence of test error in noise. In *Proceedings of the 36th International Conference on Machine Learning*, 2019.

Franceschi, J.-Y., Fawzi, A., and Fawzi, O. Robustness of classifiers to uniform ℓ_p and gaussian noise. In *21st International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2018.

Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., and Vechev, M. T. AI2: safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy, SP 2018, Proceedings, 21-23 May 2018, San Francisco, California, USA*, pp. 3–18, 2018.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Gouk, H., Frank, E., Pfahringer, B., and Cree, M. Regularisation of neural networks by enforcing lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models, 2018.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems 30*. 2017.

Huang, X., Kwiatkowska, M., Wang, S., and Wu, M. Safety verification of deep neural networks. *Computer Aided Verification*, 2017.

Hung, K. and Fithian, W. Rank verification for exponential families. *The Annals of Statistics*, (2):758–782, 04 2019.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Katz, G., Barrett, C., Dill, D. L., Julian, K., and Kochenderfer, M. J. Reluplex: An efficient smt solver for verifying deep neural networks. *Lecture Notes in Computer Science*, pp. 97–117, 2017. ISSN 1611-3349.

Kolter, J. Z. and Madry, A. Adversarial robustness: Theory and practice. https://adversarial-ml-tutorial.org/adversarial_examples/, 2018.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, 2009.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial machine learning at scale. 2017. URL https://arxiv.org/abs/1611.01236.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*, 2019. 首个提出使用randomized smoothing作为可证明对抗防御

Li, B., Chen, C., Wang, W., and Carin, L. Second-order adversarial attack and certifiable robustness. *arXiv preprint arXiv:1809.03113*, 2018. 本文说这是第二篇可证明鲁棒工作

Liu, X., Cheng, M., Zhang, H., and Hsieh, C.-J. Towards robust neural networks via random self-ensemble. In *The European Conference on Computer Vision (ECCV)*, September 2018.

Lomuscio, A. and Maganti, L. An approach to reachability analysis for feed-forward relu neural networks, 2017.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Mirman, M., Gehr, T., and Vechev, M. Differentiable abstract interpretation for provably robust neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., and Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Neyman, J. and Pearson, E. S. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. 证明中使用的基础引理

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018a.

Raghunathan, A., Steinhardt, J., and Liang, P. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems 31*, 2018b.

Singh, G., Gehr, T., Mirman, M., Püschel, M., and Vechev, M. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems 31*. 2018.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Tjeng, V., Xiao, K. Y., and Tedrake, R. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyGIdiRqtm.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SyxAb30cY7.

Tsuzuku, Y., Sato, I., and Sugiyama, M. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems 31*. 2018.

Uesato, J., O'Donoghue, B., Kohli, P., and van den Oord, A. Adversarial risk and the dangers of evaluating against weak attacks. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Wang, S., Chen, Y., Abdou, A., and Jana, S. Mixtrain: Scalable training of formally robust neural networks. *arXiv preprint arXiv:1811.02625*, 2018a.

Wang, S., Pei, K., Whitehouse, J., Yang, J., and Jana, S. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems 31*. 2018b.

Webb, S., Rainforth, T., Teh, Y. W., and Kumar, M. P. Statistical verification of neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1xcx3C5FX.

Weng, L., Zhang, H., Chen, H., Song, Z., Hsieh, C.-J., Daniel, L., Boning, D., and Dhillon, I. Towards fast computation of certified robustness for ReLU networks. In *Proceedings of the 35th International Conference on Machine Learning*, 2018a.

Weng, T.-W., Zhang, H., Chen, P.-Y., Yi, J., Su, D., Gao, Y., Hsieh, C.-J., and Daniel, L. Evaluating the robustness of neural networks: An extreme value theory approach. In *International Conference on Learning Representations*, 2018b.

Wong, E. and Kolter, J. Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Wong, E., Schmidt, F., Metzen, J. H., and Kolter, J. Z. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems 31*, 2018.

Zantedeschi, V., Nicolae, M.-I., and Rawat, A. Efficient defenses against adversarial attacks. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security - AISec '17*, 2017.

Zhang, H., Weng, T.-W., Chen, P.-Y., Hsieh, C.-J., and Daniel, L. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems 31*. 2018.