Regularized Training and Tight Certification for Randomized Smoothed Classifier with Provable Robustness

Huijie Feng[1], **Chunpeng Wu**[2], **Guoyang Chen**[2], **Weifeng Zhang**[2], **Yang Ning**[1]
[1]Cornell University, [2]Alibaba Group Inc.

Abstract

Recently smoothing deep neural network based classifiers via isotropic Gaussian perturbation is shown to be an effective and scalable way to provide state-of-the-art probabilistic robustness guarantee against ℓ_2 norm bounded adversarial perturbations. However, how to train a good base classifier that is accurate and robust when smoothed has not been fully investigated. In this work, we derive a new regularized risk, in which the regularizer can adaptively encourage the accuracy and robustness of the smoothed counterpart when training the base classifier. It is computationally efficient and can be implemented in parallel with other empirical defense methods. We discuss how to implement it under both standard (non-adversarial) and adversarial training scheme. At the same time, we also design a new certification algorithm, which can leverage the regularization effect to provide tighter robustness lower bound that holds with high probability. Our extensive experimentation demonstrates the effectiveness of the proposed training and certification approaches on CIFAR-10 and ImageNet datasets.

Introduction

Modern machine learning models such as deep neural networks have achieved a great success in a wide range of tasks, but are shown to be brittle against *adversarial attacks*. For instance, in image classification small perturbations imperceptible to human eyes may largely deteriorate the performance (Szegedy et al. 2013). Various heuristic approaches are proposed to either *attack* the classifier or *defend* adversarial attacks by making the classifier robust. However, defenses that are empirically observed to be robust to specific types of attacks are later found vulnerable to stronger or adaptive attacks (Carlini and Wagner 2017; Athalye, Carlini, and Wagner 2018; Uesato et al. 2018). Therefore, achieving provable/certifiable robustness starts to draw attention, in which the goal is to guarantee, deterministically or probabilistically, that no attacks within a certain region will alter the prediction of a classifier.

Recently, *randomized smoothing* is shown to be able to provide instance-specific ℓ_2 robustness guarantees (Lécuyer et al. 2018; Li et al. 2018; Cohen, Rosenfeld, and Kolter 2019). Specifically, given a base classifier, the prediction of the smoothed classifier, defined as the most probable prediction over random isotropic Gaussian perturbations, will not change within an ℓ_2 ball whose radius may vary among different inputs. This guarantee does not require assumptions on the base classifier, and is shown to be one of few methods to provide non-trivial robustness guarantee for large scale classification task like ImageNet.

Despite recent advances on the theoretical properties of randomized smoothed classifier, how to train a good base classifier that can achieve both good accuracy and robustness when smoothed under this framework has not been fully investigated. The training procedures employed in most previous works did not fully take into account the ultimate goal of achieving high accuracy and robustness when the trained classifier is smoothed. On the other hand, since smoothed classifiers based on neural networks cannot be evaluated exactly (we will discuss the technical details later), in order to provide robustness guarantee under this framework, a certification algorithm is required to give a lower bound of the certified radius for each instance that will hold with high probability. Nevertheless, how to certify the robustness of smoothed classifiers is under-explored as well.

In this paper, we fill the aforementioned gaps and study how to train and provide robustness certification for randomized smoothed classifier. For training, we derive a regularized risk and discuss how to implement it for training a good base classifier. Specifically, we propose ADRE, an ADaptive Radius Enhancing regularizer, which penalizes examples misclassified by the *smoothed* classifier while encourages the certified radius of correctly classified examples. This regularizer can be implemented efficiently and applied in parallel with other adversarial defense methods. In particular, we discuss how ADRE regularization can be extended to adversarial training scheme that has been widely employed to improve adversarial robustness (Kurakin, Goodfellow, and Bengio 2016; Madry et al. 2017; Salman et al. 2019). At the same time, we introduce T-CERTIFY, a new certification algorithm to provide a tighter

This work was done when Huijie Feng and Chunpeng Wu were interns at Alibaba lower bound of the certified radius that holds with high probability. This algorithm builds upon and extends previous certification approaches and can further improve the robustness guarantee. We assess the effectiveness of ADRE and T-CERTIFY on CIFAR-10 and ImageNet datasets, and demonstrate that both approaches can improve the ℓ_2 robustness of randomized smoothed classifier.

Related Work and Preliminary

Certified adversarial defenses Certified defenses aim to provide robustness guarantee for classifiers. Specifically, for a certain type of attack, we say a classifier is provable/certifiable robust within some region that may depend on the input, if the outputs of the classifier is constant over this region. For the well studied ℓ_p norm bounded attacks, a variety of methods based on techniques such as mixed integer linear programing (Lomuscio and Maganti 2017; Fischetti and Jo 2017), satisfiability modulo theories (Katz et al. 2017; Ehlers 2017; Huang et al. 2017), bounding local or global Lipschitz constant of the neural network (Hein and Andriushchenko 2017; Cisse et al. 2017; Tsuzuku, Sato, and Sugiyama 2018; Anil, Lucas, and Grosse 2018), convex relaxation (Wong and Kolter 2017; Raghunathan, Steinhardt, and Liang 2018) and many others have been proposed. However, these methods are generally unable to certify large networks, and thus cannot provide meaningful guarantees for tasks like ImageNet classification, mainly due to the intrinsic computational burden or loose relaxation. Compared to these methods, a salient advantage of randomized smoothed classifier is that it circumvents additional assumptions on the base classifier, and thus can fully leverage large expressive neural network to generate a powerful smoothed classifier.

Notations and Randomized Smoothed Classifier Let $\mathcal{D}$ denote the distribution of $(\boldsymbol{x}, y) \in \mathbb{R}^d \times [C]$ where $[C] = \{1, \ldots, C\}$. A soft classification function parameterized by $\boldsymbol{\theta}$, $F(\boldsymbol{x}; \boldsymbol{\theta}) : \mathbb{R}^d \to [0,1]^C$, maps the input to the probability score for each class $c \in [C]$, and the corresponding (hard) classifier $f(\boldsymbol{x}; \boldsymbol{\theta}) : \boldsymbol{x} \to [C]$ outputs the class label with the highest score. We use $F^c(\boldsymbol{x})$ to denote the probability score with respect to class c. For neural network classifiers, the probability scores are typically generated by the softmax function.

Given a (base) classifier f, the smoothed classifier g based on f under isotropic Gaussian perturbation with variance σ^2 is defined as

$$g(\boldsymbol{x}; f, \sigma) = \underset{c \in [C]}{\arg\max}\, G^c(\boldsymbol{x}; f, \sigma), \tag{1}$$

where $G^c(\boldsymbol{x}; f, \sigma) = \mathbb{P}(f(\boldsymbol{x} + \boldsymbol{\delta}) = c)$ is the smoothed probability score and $\boldsymbol{\delta} \sim N(0, \sigma^2 \boldsymbol{I})$. Throughout the paper we simplify the notation by omitting the parameter $\boldsymbol{\theta}$ and/or σ, and use f, g to denote the base and smoothed classifier, respectively. A nice property of g is that, for any given $\boldsymbol{x}$, $g(\boldsymbol{x} + \gamma)$ will yield the same prediction for all $\|\gamma\|_2 \leq R$, where the certified radius R depends on the top probability score $p_A = \max_c G^c(\boldsymbol{x})$ and the "runner up" score $p_B = \max_{c \neq g(\boldsymbol{x})} G^c(\boldsymbol{x})$ (Lécuyer et al. 2018; Li et al. 2018; Cohen, Rosenfeld, and Kolter 2019). Without further assumptions on f, the tight radius is

$$R = \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)), \tag{2}$$

where Φ^{-1} is the quantile function of standard Gaussian distribution (Cohen, Rosenfeld, and Kolter 2019).

Training the Base Classifier To train the base classifier, the most common approach was applying canonical empirical risk minimization with a single draw of Gaussian noise added on the training samples as a data augmentation procedure (Lécuyer et al. 2018; Cohen, Rosenfeld, and Kolter 2019). Stability training that penalizes the difference between the logits from original and Gaussian augmented example was also proposed (Li et al. 2018). Very recently, adversarial training was applied to significantly improve the certified ℓ_2 robustness of randomized smoothed classifier (Salman et al. 2019), where adding multiple Gaussian perturbation for a single training example was also employed. In this paper, we formalize the idea of single and multiple Gaussian augmentation as approximately minimizing a perturbed risk, based on which we derive the proposed ADRE regularized risk. We further adapt adversarial training to our regularized procedure and demonstrate through experiments that ADRE regularizer is also effective in this case.

Robustness Certification The robustness radius for a given example under the framework of randomized smoothing requires identifying and evaluating p_A and p_B. Unfortunately, for neural network based smoothed classifier, exact evaluation is intractable. In practice, we can only give a lower bound of the certified radius by estimating a lower and upper bound for p_A and p_B, denoted by $\underline{p_A}$ and $\overline{p_B}$, respectively. Simultaneous confidence interval for multinomial distribution (Sison and Glaz 1995) was applied in (Li et al. 2018). However, from statistical perspective, without prior knowledge about the true top and "runner-up" class, constructing confidence intervals for class probabilities is not sufficient to provide rigorous robustness certification. Another approach named CERTIFY firstly estimates $\underline{p_A}$, and then chooses $\overline{p_B} = 1 - \underline{p_A}$, which can be loose in some cases (Cohen, Rosenfeld, and Kolter 2019). In particular, the proposed ADRE regularizer encourages robustness by penalizing the "runner-up" probability for correctly classified examples, and thus this approach may not fully express the improved robustness. In contrast, the proposed T-CERTIFY estimate $\underline{p_A}$ and $\overline{p_B}$ separately, and is shown to provide tighter lower bound for the true certified radius.

While the radius in (2) holds for arbitrary base classifier, under the framework of randomized smoothing we wish to train a base classifier that can consistently make correct predictions under isotropic Gaussian perturbation to achieve high accuracy and large certified radius. Consequently, standard empirical risk minimization may not yield a desired base classifier, since the original and perturbed samples can be very different in high dimension, especially when σ is large. Instead, consider the following perturbed risk

$$\begin{aligned} R_{per}(\boldsymbol{\theta}, \mathcal{D}, \mathcal{P}) &= \mathbb{E}_{\mathcal{D}\times\mathcal{P}}\Big[L(F(\boldsymbol{x}+\boldsymbol{\delta};\boldsymbol{\theta}), y)\Big] \\ &= \mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}_{\mathcal{P}}\big[L(F(\boldsymbol{x}+\boldsymbol{\delta};\boldsymbol{\theta}), y)\big|\boldsymbol{x}.y\big]\Big], \quad (3) \end{aligned}$$

where $\boldsymbol{\delta} \sim \mathcal{P}$ is the perturbation distribution and L is some loss function. Although $\mathcal{P}$ and L can be arbitrary, in this paper we focus on $\mathcal{P} \overset{d}{=} N(0, \sigma^2 \boldsymbol{I})$ independent of $\mathcal{D}$ and cross entropy loss $l_{CE}(F(\boldsymbol{x}), y) = -\log(F^y(\boldsymbol{x}))$. We write for simplicity $R_{per}(\boldsymbol{\theta}, \mathcal{D}, \mathcal{P}) = R_{per}(\boldsymbol{\theta})$ without confusion. Intuitively, minimizing (3) yields a classifier that has low risk, and thus high accuracy under Gaussian perturbation.

Motivating Adaptive Radius Enhancing Regularization The perturbed risk (3) tends to yield a randomized smoothed classifier with high accuracy. However, the tradeoff between robustness and accuracy has been widely observed, both empirically and theoretically (Fawzi, Fawzi, and Frossard 2018; Tsipras et al. 2019; Zhang et al. 2019). Meanwhile, although Gaussian augmentation has also been observed to yield a (base) classifier with improved robustness (Kannan, Kurakin, and Goodfellow 2018), it is not clear whether it will generate a smoothed classifier with large certified robustness. In fact, without additional assumptions on the curvature or complexity of the base classifier, it is difficult to build a direct connection between the base and smoothed classifier . Thus, the resulting base classifier from (3) may still be suboptimal regarding robustness when smoothed.

It is clear that for any given input $\boldsymbol{x}$, the certified radius directly depends on the top and "runner-up" probability score of the smoothed classifier. Notice that for any fixed input $\boldsymbol{x}$ a certified radius exists no matter g makes a correct prediction or not. However, while a large radius when $\boldsymbol{x}$ is correctly predicted is desired, a misclassified $\boldsymbol{x}$ with large radius is detrimental. This motivates the following measure

$$R_{adre}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}\Bigg[\underbrace{L'(G(\boldsymbol{x};\boldsymbol{\theta}), \arg\max_{c\neq y} G^c(\boldsymbol{x};\boldsymbol{\theta}))}_{(\spadesuit)}\Bigg], \quad (4)$$

where L' is some loss function. To interpret this, we consider two cases

- when g makes a correct prediction, $\arg\max_{c\neq y} G^c(\boldsymbol{x};\boldsymbol{\theta})$ is the "runner-up" class, and in this case ($\spadesuit$) serves as a measure of robustness for the smoothed classifier, where a larger value suggests a higher robustness.
- when g makes a wrong prediction, $\arg\max_{c\neq y} G^c(\boldsymbol{x};\boldsymbol{\theta})$ is top class, and in this case ($\spadesuit$) corresponds to the radius of a misclassified example, where a larger value indicates a smaller radius.

Therefore, we can think of R_{adre} as a balanced measure between accuracy and robustness for the smoothed classifier g. For concreteness, in this paper we also choose L' as the cross entropy loss. Following this, we propose ADRE, an ADaptive Radius Enhancing regularized risk

$$R_{reg}(\boldsymbol{\theta}) = R_{per}(\boldsymbol{\theta}) - \lambda R_{adre}(\boldsymbol{\theta}), \quad (5)$$

where λ is a hyper-parameter. Here the first component R_{per} corresponds to the classification accuracy of the base classifier under perturbation. For the second component, we use R_{adre} as a regularization term that adaptively encourages the certified radius and accuracy for the smoothed counterpart of the trained base classifier. We call the training procedure based on (5) as ADRE$_{\text{REG}}$.

Connection to Large Margin Training The goal of achieving large certified radius for correctly classified example is closely related to the objective of obtaining large margin classifier. Notice that $R = \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)) \geq \phi \cdot \frac{\sigma}{2}(p_A - p_B)$, where $\phi > 0$ is the lower bound of the derivative of Φ^{-1}. From (4) we can see that R_{adre} acts a similar role as promoting $G^y(\boldsymbol{x}) - \max_{c\neq y} G^c(\boldsymbol{x})$, which is equivalent to $p_A - p_B$ when the smoothed classifier correctly classify $\boldsymbol{x}$. Therefore, the proposed ADRE regularizer can be treated as a large margin regularizer under the framework of randomized smoothing. Different from directly maximizing the margin of the trained classifier such as in (Ding et al. 2018; Elsayed et al. 2018), we exploit R_{adre} that is tailored to randomized smoothed classifiers to guide the trained base classifier in the direction of higher robustness when smoothed.

Implementation Given training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, in practice our objective naturally becomes to minimize

$$\frac{1}{n}\sum_{i=1}^{n} L_i - \lambda P_i, \quad (6)$$

where $L_i = \mathbb{E}_{\mathcal{P}}[l_{CE}(F(\boldsymbol{x}_i + \boldsymbol{\delta};\boldsymbol{\theta}), y_i)]$ and $P_i = l_{CE}(G(\boldsymbol{x}_i;\boldsymbol{\theta}), \arg\max_{c\neq y} G^c(\boldsymbol{x}_i;\boldsymbol{\theta}))$.

However, for a neural network base classifier, it is intractable to evaluate both L_i and G exactly, and thus we will approximate both terms during training. Given a training pair $(\boldsymbol{x}', y')$, for the first term we use the unbiased estimator

$$\hat{L}(\boldsymbol{x}', y';\boldsymbol{\theta}) = \frac{1}{k}\sum_{j=1}^{k} l_{CE}(F(\boldsymbol{x}' + \boldsymbol{\delta}_j;\boldsymbol{\theta}), y'). \quad (7)$$

For the second term, we will substitute G by

$$\hat{G}(\boldsymbol{x}';\boldsymbol{\theta}) = \frac{1}{k}\sum_{j=1}^{k} F(\boldsymbol{x}' + \boldsymbol{\delta}_j;\boldsymbol{\theta}). \quad (8)$$

Essentially, for both terms we sample i.i.d Gaussian perturbations and substitute the conditional expected loss and the smoothed probability score by finite sample estimators. Note that for G, we average over a finite sample of base classifier probability scores F under perturbation instead of employing the fraction of counts, defined as

$$\frac{1}{k}\sum_{j=1}^{k}\big(\mathbb{1}\{f(\boldsymbol{x}' + \boldsymbol{\delta}_{ij}) = c\}\big)_{c=1}^{C} \in [0,1]^C, \quad (9)$$

where $\mathbb{1}$ is the indicator function. Although (9) is an unbiased estimator for $G(\boldsymbol{x}')$, due to computational constraint, in practice k cannot be too large, which is problematic both

Algorithm 1 ADRE regularized Training

```
input      : Training sample $\mathcal{D}_N$
parameter: variance $\sigma > 0$; tuning parameter $\lambda \geq 0$; number of perturbations $k > 0$;
           # for adversarial training
           attack steps $M$; step size $\alpha$; radius $\epsilon$;
for each epoch do
    for each minibatch $\{(\boldsymbol{x}_i, y_i)\}_{i \in [B]} \subset \mathcal{D}_N$ do
        $\boldsymbol{\delta}_{ij} \overset{i.i.d}{\sim} N(0, \sigma^2), j \in [k]$
        if adversarial training then
            for $m = 1, \dots, M$ do
                $\hat{G}_i \leftarrow \frac{1}{k} \sum_{j=1}^{k} F(\boldsymbol{x}_i + \boldsymbol{\delta}_{ij}; \boldsymbol{\theta})$
                $\boldsymbol{x}_i \leftarrow$ PGD-step$(\boldsymbol{x}_i, \hat{G}_i, \alpha, \epsilon)$
            end
        end
        $l_{per}^{(i)} \leftarrow \frac{1}{k} \sum_{j=1}^{k} Loss(F(\boldsymbol{x}_i + \boldsymbol{\delta}_{ij}; \boldsymbol{\theta}), y_i)$
        $\hat{G}_i \leftarrow \frac{1}{k} \sum_{j=1}^{k} F(\boldsymbol{x}_i + \boldsymbol{\delta}_{ij}; \boldsymbol{\theta})$
        $\hat{y}_i \leftarrow \arg\max_{c \neq y_i} \hat{G}_i^c$
        $l_{adre}^{(i)} \leftarrow Loss(\hat{G}_i, \hat{y}_i)$
    end
    $\nabla L = \nabla \frac{1}{B} \sum_{i=1}^{B} \{ l_{per}^{(i)} - \lambda l_{adre}^{(i)} \}$
    $\boldsymbol{\theta} \leftarrow$ Step$(\boldsymbol{\theta}, \nabla L)$ #update using proper optimizer
end
```

statistically and numerically, especially when the number of classes is large. Instead, by applying (8) we implicitly conduct smoothing when estimating $G(\boldsymbol{x}')$.

We can also justify (8) following $G(\boldsymbol{x}') = (\mathbb{E}[\mathbb{1}(f(\boldsymbol{x}' + \boldsymbol{\delta}) = c)])_{c=1}^{C} \approx \mathbb{E}[F(\boldsymbol{x}' + \delta)]$.

The detailed training procedure is described in Algorithm 1. Notice that we use the same set of perturbations in both l_{per} and $\hat{G}$. Empirically, we find this saves half of forward pass computation without sacrificing accuracy and robustness compared to the case where two different sets of perturbations are applied. Our implementation of ADRE$_{\text{REG}}$ also unifies and generalizes different Gaussian data augmentation techniques applied in previous works when $\lambda = 0$ (Lécuyer et al. 2018; Cohen, Rosenfeld, and Kolter 2019; Salman et al. 2019). We also note that Algorithm 1 unifies the adversarial training scheme which will be discussed later.

Alternative Formulations One may consider directly balancing off accuracy and robustness based on the following objective

$$\mathbb{E}_{\mathcal{D}}\Big[\mathbb{E}[l_{CE}(F(\boldsymbol{x} + \boldsymbol{\delta}; \boldsymbol{\theta}), y) | \boldsymbol{x}, y] - \lambda' \big(\Phi^{-1}(\max_{c} G^c(\boldsymbol{x}; \boldsymbol{\theta})) - \Phi^{-1}(\max_{c \neq g(\boldsymbol{x};\boldsymbol{\theta})} G^c(\boldsymbol{x}; \boldsymbol{\theta}))\big)\Big], \tag{10}$$

where the first part stays the same, but the second part corresponds to the expected certified radius. Although this looks somewhat natural, empirically we observe that minimizing this objective with plug-in approximation (8) is not stable and may converge to bad local minima, especially when λ is relatively large. This is reasonable since the second part of (10) does not involve the correct label, and a classifier that consistently makes wrong prediction with high confidence can have low risk. Therefore, minimizing this objective can easily converge to bad local minima with such property. We speculate that more careful initialization may be required to yield desired base classifier in this case.

Regularized Smoothed Adversarial Training Adversarial training has been widely used to boost the robustness of classifiers, and is arguably the most effective type of empirical defense method against adversarial attacks (Kurakin, Goodfellow, and Bengio 2016; Madry et al. 2017). Generally speaking, the objective can be formulated as minimizing the worst case risk over an adversarial region with strength ϵ, denoted by S_ϵ

$$\mathbb{E}_{\mathcal{D}}\Big[\max_{\boldsymbol{x}' \in S_\epsilon(\boldsymbol{x})} L(F(\boldsymbol{x}'; \boldsymbol{\theta}), y)\Big]. \tag{11}$$

While adversarial training is typically used to improve empirical robustness of classifiers, it is also recently found helpful to improve provable robustness for smoothed classifier (Salman et al. 2019).

In this section, we describe an ℓ_2 attack scheme based on ADRE regularization, which can be incorporated into training for obtaining robust smoothed classifier. Formally, given $(\boldsymbol{x}^0, y)$ we seek for an adversarial example

$$\tilde{\boldsymbol{x}} = \max_{\|\boldsymbol{x}^0 - \boldsymbol{x}'\|_2 \leq \epsilon} \Big\{ l_{CE}(G(\boldsymbol{x}'; \boldsymbol{\theta}), y) - \lambda l_{CE}(G(\boldsymbol{x}'; \boldsymbol{\theta}), \arg\max_{c \neq y} G^c(\boldsymbol{x}'; \boldsymbol{\theta})) \Big\}. \tag{12}$$

To be specific, instead of maximizing the standard cross entropy loss of smoothed classifier $l_{CE}(G(\boldsymbol{x}; \boldsymbol{\theta}), y)$, we maximize it together with ADRE regularization. To interpret this, when we maximize over $l_{CE}(G(\boldsymbol{x}; \boldsymbol{\theta}), y)$, we generate an adversarial example with respect to the smoothed classifier that leads to high loss and thus wrong prediction. In our scenario, however, $\tilde{\boldsymbol{x}}$ tends to be either 1) correctly classified but non-robust or 2) misclassified, potentially by a large margin. Therefore, the proposed attack is more versatile under the framework of randomized smoothing, and potentially leads to a smoothed classifier with a better balance between accuracy and robustness, when adversarial training based on this attack is employed. The proposed attack is an extension of the SMOOTHADV attack (Salman et al. 2019) when (12) is implemented with plug-in estimate (8). We also note that similar to SMOOTHADV we use $l_{CE}(G(\boldsymbol{x}; \boldsymbol{\theta}), y)$ in (12) instead of $\mathbb{E}_\delta(F(\boldsymbol{x} + \boldsymbol{\delta}; \boldsymbol{\theta}))$, where the latter one was found to be ineffective in practice.

Since exact evaluation of the above maximization problem is intractable, we will follow the widely used iterative first-order methods. For concreteness, in this paper we focus on non-targeted ℓ_2 projected gradient descent (PGD) attack (Madry et al. 2017), but other approaches can be applied as well. Specifically, we approximate the inner maximizer by iteratively solving

$$\boldsymbol{x}^{t+1} = \mathcal{P}_{2,\epsilon}\bigg(\boldsymbol{x}^t + \alpha \cdot \nabla\{l_{CE}(\hat{G}(\boldsymbol{x}^t;\boldsymbol{\theta}), y) \\ - \lambda l_{CE}(\hat{G}(\boldsymbol{x}^t;\boldsymbol{\theta}), \arg\max_{c \neq y} \hat{G}^c(\boldsymbol{x}^t;\boldsymbol{\theta}))\}\bigg). \quad (13)$$

where $\mathcal{P}_{2,\epsilon}$ is the projection operator into an ℓ_2 ball with radius ϵ and α is the step size.

The detailed implementation of the proposed adversarial training based on the above PGD attack, referred as ADRE$_{\text{ADV}}$, is described in Algorithm 1 where the helper function PGD($\boldsymbol{x}, \hat{G}, \alpha, \epsilon$) runs a single step of PGD iteration (13). We also reuse the same set of noise samples for each training example at each PGD iteration to stabilize the attack, as suggested in (Salman et al. 2019).

Robustness Certification

Certifying the robustness radius of a smoothed classifier g for a given input $\boldsymbol{x}$ requires evaluating g exactly for p_A and p_B. In practice, we may only estimate a lower bound $\underline{p_A}$ and an upper bound $\overline{p_B}$ that hold with high probability. In this section, we propose a Monte Carlo algorithm that guarantees a lower bound of the true certified robustness that holds with probability greater than $1-\alpha$, where α is a pre-specified significance level. This method independently estimates $\underline{p_A}, \overline{p_B}$ and thus can leverage the regularized smoothed classifier to provide tighter robustness guarantee.

We now describe the certification procedure. Given a base classifier f and input $\boldsymbol{x}$, we firstly sample $\boldsymbol{\delta}_i \overset{i.i.d}{\sim} N(0, \sigma^2 \boldsymbol{I})\ \forall\ i \in [n]$ and then evaluate each $f(\boldsymbol{x} + \boldsymbol{\delta}_i)$. Suppose we get a sequence of ordered counts $\hat{N}_{R_1} \geq \hat{N}_{R_2} \geq \ldots \geq \hat{N}_{R_C}$, where each $R_i \in [C]$ is an ordered class label. For a given significance level α and $\alpha' \in [0, \alpha]$, suppose for now $R_1 = g(\boldsymbol{x})$. Consider

$$\underline{p_A} = \sup\bigg\{p \big| \mathbb{P}(Bin(n,p) \geq \hat{N}_{R_1}) \leq \alpha'\bigg\},$$
$$\overline{p_B} = \inf\bigg\{p \big| \sum_{j=2}^{C} \mathbb{P}(Bin(n,p) \leq \hat{N}_{R_j}) \leq \alpha - \alpha'\bigg\}, \quad (14)$$

where the probability is over the binomial random variable $Bin(n,p)$ with n number of trials and success probability p. The lower bound of the top probability score $\underline{p_A}$ is given by the classic Clopper—Pearson method (Clopper and Pearson 1934) with one-sided significance level α'. For the upper bound, we generalize the Clopper—Pearson method to construct a one-sided confidence interval for p_B with significance level $\alpha - \alpha'$, where $\overline{p_B}$ is defined as its boundary point.

Proposition 1. *Following the certification procedure described above. For any fix $\boldsymbol{x}$, if the $R_1 = g(\boldsymbol{x})$ then with probability greater than $1-\alpha$, $g(\boldsymbol{x}+\gamma) = R_1\ \forall\ \|\gamma\|_2 \leq \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$.*

Proof Sketch. Without loss of generality, suppose the top label $R_1 = 1$. Write $G(\boldsymbol{x}) = (p_1, p_2, \cdots, p_C)$. It suffices to show that

$$\mathbb{P}(\underline{p_A} > p_1 \cup \overline{p_B} < \max_{c \neq 1} p_c) \leq \alpha. \quad (15)$$

Based on the definitions in (14), we know $\mathbb{P}(\underline{p_A} > p_1) \leq \alpha'$. On the other hand, write $\alpha_c = \mathbb{P}(Bin(n, \overline{p_B}) \leq \hat{N}_c)$, we know $\sum_{c=2}^{C} \alpha_c \leq \alpha - \alpha'$ and therefore

$$\mathbb{P}(\overline{p_B} < \max_{c \neq 1} p_c) \leq \sum_{c=2}^{C} \mathbb{P}(\overline{p_B} < p_c) \sum_{c=2}^{C} \alpha_c \leq \alpha - \alpha'.$$

This completes the proof by applying a union bound. □

Proposition 1 shows that, if we have knowledge about the top class then $\underline{p_A}, \overline{p_B}$ are proper bounds, and thus we can estimate a lower bound for the certified radius that holds with probability greater than $1-\alpha$. To obtain a tighter lower bound, we may maximize the radius $\frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$ over $\alpha' \in [0, \alpha]$. For practical implementation in which the top class is unknown, we propose T-CERTIFY, which extends CERTIFY (Cohen, Rosenfeld, and Kolter 2019) to provide a tighter certified robustness for a given input by estimating $\underline{p_A}, \overline{p_B}$ separately and searching over a grid of α's. The algorithm is as follows.

```
Algorithm 2 T-Certify
input      : base classifier $f$, input $\boldsymbol{x}$
parameter: variance $\sigma$, size $n_0, n > 0$, significance $\alpha$, grid $\mathcal{A}$
$\boldsymbol{c}_0 \leftarrow$ SampleUnderNoise($f, \boldsymbol{x}, n_0, \sigma$)
$R_1 \leftarrow$ top index in $\boldsymbol{c}_0$
$\boldsymbol{c} \leftarrow$ SampleUnderNoise($f, \boldsymbol{x}, n, \sigma$)
for $\alpha' \in \mathcal{A}$ do
    $\underline{p_A} \leftarrow$LowerConfBound($\boldsymbol{c}[R_1], n, 1-\alpha'$)
    $\overline{p_B} \leftarrow$ UpperConfBound($\boldsymbol{c}[-R_1], n, 1-(\alpha-\alpha')$)
    if $\underline{p_A} > 0.5$ then
        $r_{\alpha'} \leftarrow \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$.
    else
        $r_{\alpha'} \leftarrow 0$.
    end
end
if $\max_{\alpha' \in \mathcal{A}} r_{\alpha'} > 0$ return $(R_1, \max_{\alpha' \in \mathcal{A}} r_{\alpha'})$ else return
  ABSTAIN
```

Here SampleUnderNoise($f, \boldsymbol{x}, n, \sigma$) samples the noise $\boldsymbol{\delta}'_i\ \forall\ i \in [n]$, evaluate $f(\boldsymbol{x}+\boldsymbol{\delta}'_i)$ and get counts for each class. Function LowerConfBound($c[R_1], n, 1-\alpha$) calculate $\underline{p_A}$ following (14) based on the Clopper—Pearson confidence interval (Clopper and Pearson 1934), and similarly for UpperConfBound. Similar to CERTIFY, T-CERTIFY abstains from making a prediction when the lower bound at significance level α' is no larger than a half, which guarantees the correctness of the top class prediction.

Theorem 1. *If T-CERTIFY does not abstain and returns a label c with radius r, then with probability at least $1-\alpha$, $g(\boldsymbol{x}+\gamma) = c\ \forall\ \|\gamma\|_2 \leq r$, where r is the returned radius in T-CERTIFY.*

Table 1: Certified top-1 accuracy on CIFAR-10 and ImageNet at various radii.

	Method	ℓ_2 Radius	0.0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25	2.5	2.75	3.0
CIFAR-10	Basic	$\sigma = 0.12$	0.81	0.59	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.25$	0.75	0.60	0.43	0.27	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.50$	0.65	0.55	0.41	0.32	0.23	0.15	0.09	0.05	0.00	0.00	0.00	0.00	0.00
		$\sigma = 1.00$	0.47	0.39	0.34	0.28	0.22	0.17	0.14	0.12	0.10	0.08	0.05	0.04	0.02
	ADRE$_{\text{REG}}$	$\sigma = 0.12, \lambda = 0.1$	0.83	0.65	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.12, \lambda = 0.2$	**0.85**	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.12, \lambda = 0.3$	0.83	**0.68**	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.25, \lambda = 0.1$	0.78	0.64	**0.50**	0.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.25, \lambda = 0.2$	0.74	0.60	0.48	0.35	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.25, \lambda = 0.3$	0.73	0.62	0.49	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.50, \lambda = 0.1$	0.67	0.57	0.48	0.38	**0.30**	0.23	0.17	0.11	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.50, \lambda = 0.2$	0.65	0.57	0.47	0.35	0.27	0.20	0.13	0.09	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.50, \lambda = 0.3$	0.64	0.55	0.46	**0.38**	0.30	**0.23**	**0.17**	0.11	0.00	0.00	0.00	0.00	0.00
		$\sigma = 1.00, \lambda = 0.1$	0.49	0.43	0.36	0.29	0.22	0.19	0.15	0.13	0.11	0.08	0.05	0.03	0.02
		$\sigma = 1.00, \lambda = 0.2$	0.48	0.41	0.35	0.28	0.22	0.18	0.16	0.14	0.11	**0.09**	0.06	0.05	0.02
		$\sigma = 1.00, \lambda = 0.3$	0.47	0.39	0.33	0.29	0.24	0.20	0.17	**0.14**	**0.12**	0.09	**0.07**	**0.05**	**0.03**
ImageNet	Basic	$\sigma = 0.25$	0.67	0.58	0.49	0.37	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.50$	0.57	0.52	0.46	0.42	0.37	0.33	0.29	0.22	0.00	0.00	0.00	0.00	0.00
		$\sigma = 1.00$	0.44	0.41	0.38	0.35	0.33	0.29	0.26	0.22	0.19	0.17	0.15	0.13	0.12
	ADRE$_{\text{REG}}$	$\sigma = 0.25, \lambda = 0.05$	**0.70**	**0.64**	**0.57**	0.45	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.25, \lambda = 0.10$	0.69	0.63	0.55	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.50, \lambda = 0.05$	0.61	0.56	0.51	0.46	0.40	**0.36**	0.30	0.25	0.00	0.00	0.00	0.00	0.00
		$\sigma = 0.50, \lambda = 0.10$	0.62	0.57	0.52	**0.47**	**0.42**	0.36	0.29	0.24	0.00	0.00	0.00	0.00	0.00
		$\sigma = 1.00, \lambda = 0.05$	0.48	0.45	0.41	0.37	0.36	0.32	0.30	0.26	0.23	**0.22**	**0.18**	0.15	**0.14**
		$\sigma = 1.00, \lambda = 0.10$	0.47	0.44	0.40	0.38	0.36	0.33	**0.30**	**0.27**	**0.24**	0.20	0.18	**0.16**	0.13

Proof Sketch. By Proposition 1 we know that with probability at least $1 - \alpha$ $\underline{p_A} \leq p_1$ and $\overline{p_B} \geq \max_{c \neq 1} p_c$ hold, where again suppose without loss of generality $R_1 = 1$. On this event, since T-CERTIFY does not abstain only when $\underline{p_A} > 0.5$, we know the top class is correctly predicted, i.e., $g(\boldsymbol{x}) = R_1$. This completes the proof. □

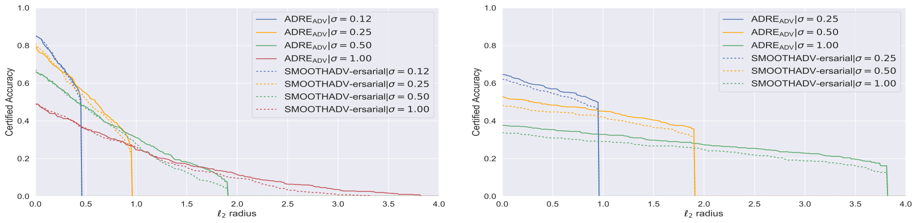
Experiments

In this section, we evaluate the effectiveness of ADRE regularization and T-CERTIFY algorithm for randomized smoothed classifier. For the training procedure, we mainly compare with the basic single Gaussian perturbation augmented training, referred as Basic training(Cohen, Rosenfeld, and Kolter 2019), and SMOOTHADV-ersarial training (Salman et al. 2019), as these two approaches achieve state-of-the-art ℓ_2 robustness under standard (non-adversarial) and adversarial training scheme, respectively. For the certification algorithm, we mainly compare with CERTIFY.

To evaluate robustness, we focus on the approximate certified accuracy at radius r, defined as the fraction of samples which are classified correctly by the certification algorithm along with a certified radius being at least r. When comparing ADRE with other training methods, for direct comparison we only apply CERTIFY for robustness certification with significance level $\alpha = 0.001$ and number of samples $n_0 = 100, n = 100,000$. This means that we use 100 Monte Carlo samples to predict the output of smoothed classifier, and $100,000$ to calculate a lower bound of certified radius for each sample that will hold with probability being at least 99.9%. Note that the approximate certified accuracy is not equivalent to the lower bound of the true accuracy that holds with probability at least $1 - \alpha$ over the randomness of the CERTIFY algorithm, but the difference is negligible when α is small. We refer the reader to (Cohen, Rosenfeld, and Kolter 2019) for details. For T-CERTIFY, we search over $\alpha = 0.1, 0.2, \ldots, 1.0$, where at $\alpha = 1.0$ it returns the same certified radius as in CERTIFY.

We firstly assess the performance of ADRE$_{\text{REG}}$ and ADRE$_{\text{ADV}}$ training. We run experiments on CIFAR-10 (Krizhevsky and others 2009) and ImageNet (Deng et al. 2009) datasets. Consistent to compared work, we employ a 110-layer residual network and ResNet-50 as the base classifier for CIFAR-10 and ImageNet, respectively. For adversarial training, we used a constant step size $\alpha = 2\epsilon/M$ with M being the number of attack iterations, and ϵ being the ℓ_2 attack radius. On CIFAR-10, we trained the classifier using SGD on a single NVIDIA Tesla V100 GPU. We used a batch size of 400 with initial learning rate 0.1 which drops by a factor of 10 every 50 epochs, in the total 150 epochs. On ImageNet, we trained the classifier on 4 NVIDIA Tesla V100 GPU using synchronous SGD with batch size 256 when $k = 1$ and 64 when $k = 4$, where k is the number of Gaussian perturbations for plug-in estimates. We also used momentum (0.875), weight decay (1/32768), label smoothing (0.1) and cosine learning rate schedule for 50 epochs in total, where we set $0.1 \cdot epoch/8$ for warm-up and $0.05 \cdot (1 + cos(\pi \cdot epoch/(50 - 8)))$ afterwards. For both datasets, we trained the base classifier with random horizontal flips and random crops. Similar to compared work, the certified radii are with respect to original coordinate for direct comparison. We also added a centering layer as the first layer of the base classifier, which performed a channel-wise standardization, as implemented in (Salman et al. 2019).

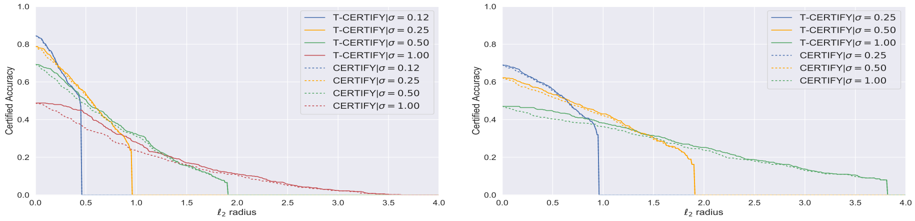
Table 1 reports the approximate top-1 certified accuracy on CIFAR-10 and ImageNet comparing ADRE$_{REG}$ and Basic training. On CIFAR-10, we train the base classifier with number of perturbations $k = 8$ and regularization $\lambda \in \{0.1, 0.2, 0.3\}$ for different magnitude of perturbations $\sigma \in \{0.12, 0.25, 0.50, 1.00\}$. On ImageNet, we train with $k = 1, \lambda \in \{0.05, 0.1\}$ for $\sigma \in \{0.25, 0.50, 1.00\}$. For a direct comparison, we slightly change the implementation of ADRE$_{REG}$ training on CIFAR-10. Specifically, instead of calculating $l_{per}^{(i)}$ following (7) as described in Algorithm 1, in this experiment we only randomly sample a single pertur-



(a)

(b)

Figure 1: Certified accuracy of smoothed classifier trained with ADRE$_{\mathrm{ADV}}$ (solid line) vs SMOOTHADV-ersarial (dashed line) on (a) CIFAR-10 and (b) ImageNet.



(a)

(b)

Figure 2: Certified accuracy of one representative base classifier certified by our T-CERTIFY (solid line) vs CERTIFY (dashed line) on (a) CIFAR-10 and (b) ImageNet

bation for $l_{per}^{(i)}$, i.e., we let $l_{per}^{(i)} = l_{CE}(F(\boldsymbol{x}_i+\boldsymbol{\delta}_{ij};\boldsymbol{\theta}), y_i)$ for a random index $i \in [k]$. By doing this, the only difference between ADRE$_{REG}$ and Basic training lies in ADRE regularization for both datasets. The results from Table 1 suggests that ADRE regularization indeed improves the accuracy and robustness of smoothed classifier, where the certified robustness at zero radius is just the standard accuracy of the smoothed classifier. In particular, with a proper hyperparameter λ, for each perturbation σ we can improve the certified radius up to 9% on CIFAR-10 and 8% on ImageNet without sacrificing the standard accuracy. We point out that on ImageNet, there is little additional computation compared to Basic training. We also run the original ADRE$_{REG}$ with $k = 8$ on CIFAR-10 and $k = 4$ for ImageNet. As is expected, we observe even stronger robustness at various radii when the base classifier is smoothed.

In the next experiment, we compare SMOOTHADV-ersarial and the proposed ADRE$_{\mathrm{ADV}}$ training. For demonstration, we focus on 2-step PGD adversarial training on CIFAR-10 with $k = 8$ and 1-step PGD on ImageNet with $k = 1$. Figure 1 plots the approximate certified accuracy of representative models on (a) CIFAR-10 and (b) ImageNet. Each solid line depicts the certified accuracy of a model trained by ADRE$_{\mathrm{ADV}}$ and the dashed line depicts the certified accuracy of SMOOTHADV-ersarially trained model with the same k and ϵ, in which multiple Gaussian perturbation was applied for each training example on CIFAR-10. The results from Figure 1 suggest that ADRE regularization is also useful under adversarial training scheme.

Robustness Certification In this section, we evaluate the effectiveness of T-CERTIFY algorithm. We use the same α, n_0 and n as applied in CERTIFY. When certifying a given example $\boldsymbol{x}$, we firstly generate a set of perturbations, and then use the same set of perturbed inputs to estimate $g(\boldsymbol{x})$ and calculate the certified radius. This helps reduce uncertainty when comparing two approaches.

Figure 2 depicts the certified accuracy from both approaches. We can observe that at each radius, T-CERTIFY yields higher certified robustness. In addition, we notice that the improvement gets more significant when σ is larger. This is reasonable since with a larger perturbation, the confidence of the smoothed classifier may become lower. In this case, it becomes more important to estimate $\underline{p_A}$ and $\overline{p_B}$ separately in order to provide tighter lower bound for certified radius.

Conclusion

In this paper, we introduced a novel training procedure and certification algorithm for randomized smoothed classifier. We derived ADRE regularized risk and discussed how it can be implemented in both standard and iterative first-order adversarial training scheme. For certifying the (probabilistic) robustness of a smoothed classifier, we introduced T-CERTIFY to estimate lower bound for the ℓ_2 robustness radius that will hold with high probability. We showed through experiments on CIFAR-10 and ImageNet datasets that ADRE regularization can improve the accuracy and ℓ_2 robustness of the smoothed classifier, whose base classifier was trained under both standard and adversarial training scheme. We also demonstrated that T-CERTIFY can further improve the robustness guarantee based on the proposed regularization.

References

Anil, C.; Lucas, J.; and Grosse, R. 2018. Sorting out lipschitz function approximation. *arXiv preprint arXiv:1811.05381*.

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*.

Carlini, N., and Wagner, D. 2017. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 3–14. ACM.

Cisse, M.; Bojanowski, P.; Grave, E.; Dauphin, Y.; and Usunier, N. 2017. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 854–863. JMLR. org.

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

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Ding, G. W.; Sharma, Y.; Lui, K. Y. C.; and Huang, R. 2018. Max-margin adversarial (mma) training: Direct input space margin maximization through adversarial training. *arXiv preprint arXiv:1812.02637*.

Ehlers, R. 2017. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, 269–286. Springer.

Elsayed, G.; Krishnan, D.; Mobahi, H.; Regan, K.; and Bengio, S. 2018. Large margin deep networks for classification. In *Advances in neural information processing systems*, 842–852.

Fawzi, A.; Fawzi, O.; and Frossard, P. 2018. Analysis of classifiers' robustness to adversarial perturbations. *Machine Learning* 107(3):481–508.

Fischetti, M., and Jo, J. 2017. Deep neural networks as 0-1 mixed integer linear programs: A feasibility study. *arXiv preprint arXiv:1712.06174*.

Hein, M., and Andriushchenko, M. 2017. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *Advances in Neural Information Processing Systems*, 2266–2276.

Huang, X.; Kwiatkowska, M.; Wang, S.; and Wu, M. 2017. Safety verification of deep neural networks. In *International Conference on Computer Aided Verification*, 3–29. Springer.

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

Katz, G.; Barrett, C.; Dill, D. L.; Julian, K.; and Kochenderfer, M. J. 2017. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, 97–117. Springer.

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

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Lécuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2018. Certified robustness to adversarial examples with differential privacy. In *IEEE S&P 2019*.

Li, B. H.; Chen, C.; Wang, W.; and Carin, L. 2018. Certified adversarial robustness with additive gaussian noise. *arXiv preprint arXiv:1809.03113*.

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

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Raghunathan, A.; Steinhardt, J.; and Liang, P. S. 2018. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, 10877–10887.

Salman, H.; Yang, G.; Li, J.; Zhang, P.; Zhang, H.; Razenshteyn, I.; and Bubeck, S. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. *arXiv preprint arXiv:1906.04584*.

Sison, C. P., and Glaz, J. 1995. Simultaneous confidence intervals and sample size determination for multinomial proportions. *Journal of the American Statistical Association* 90(429):366–369.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*.

Tsuzuku, Y.; Sato, I.; and Sugiyama, M. 2018. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, 6541–6550.

Uesato, J.; O'Donoghue, B.; Kohli, P.; and Oord, A. 2018. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*, 5032–5041.

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

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*.