UniCR: Universally Approximated Certified Robustness via Randomized Smoothing

Hanbin Hong[1], Binghui Wang[2], and Yuan Hong[1]

[1] University of Connecticut, Storrs CT 06269, USA
hanbin.hong@uconn.edu, yuan.hong@uconn.edu
[2] Illinois Institute of Technology, Chicago IL 60616, USA
bwang70@iit.edu

Abstract. We study certified robustness of machine learning classifiers against adversarial perturbations. In particular, we propose the first universally approximated certified robustness (UniCR) framework, which can approximate the robustness certification of *any* input on *any* classifier against *any* ℓ_p perturbations with noise generated by *any* continuous probability distribution. Compared with the state-of-the-art certified defenses, UniCR provides many significant benefits: (1) the first universal robustness certification framework for the above 4 "any"s; (2) automatic robustness certification that avoids case-by-case analysis, (3) tightness validation of certified robustness, and (4) optimality validation of noise distributions used by randomized smoothing. We conduct extensive experiments to validate the above benefits of UniCR and the advantages of UniCR over state-of-the-art certified defenses against ℓ_p perturbations.

Keywords: Adversarial Machine Learning; Certified Robustness; Randomized Smoothing

1 Introduction

Machine learning (ML) classifiers are vulnerable to adversarial perturbations [36, 5, 7, 6]). Certified defenses [47, 27, 4, 19, 21, 38, 12, 37] were recently proposed to ensure provable robustness against adversarial perturbations. Typically, certified defenses aim to derive a certified radius such that an arbitrary ℓ_p (e.g., ℓ_1, ℓ_2 or ℓ_∞) perturbation, when added to a testing input, cannot fool the classifier, if the ℓ_p-norm value of the perturbation is within the radius. Among all certified defenses, randomized smoothing [35, 32, 11] based certified defense has achieved the state-of-the-art certified radius and can be applied to *any* classifier. Specifically, given a testing input and any classifier, randomized smoothing first defines a noise distribution and adds sampled noises to the testing input; then builds a smoothed classifier based on the noisy inputs, and finally derives certified radius for the smoothed classified, e.g., using the Neyman-Pearson Lemma [11].

However, existing randomized smoothing based (and actually all) certified defenses only focus on specific settings and cannot universally certify a classifier against *any* ℓ_p perturbation or *any* noise distribution. For example, the certified

Table 1. Comparison with highly-related works.

	Classifier	Smoothing Noise	Perturbations	Tightness	Optimizable	Analysis-free
Lecuyer et al. [32]	Any	Gaussian/Laplace	Any $\ell_p, p \in \mathbb{R}^+$	Loose	No	No
Cohen et al. [11]	Any	Gaussian	ℓ_2	Strictly Tight	No	No
Teng et al. [43]	Any	Laplace	ℓ_1	Strictly Tight	No	No
Dvijotham et al. [16]	Any	f-divergence-constrained	Any $\ell_p, p \in \mathbb{R}^+$	Loose	No	No
Croce et al. [12]	ReLU-based	No	Any ℓ_p for $p >= 1$	Loose	No	No
Yang et al. [51]	Any	Multiple types	Any $\ell_p, p \in \mathbb{R}^+$	Strictly Tight	No	No
Zhang et al. [52]	Any	ℓ_p-term-constrained	$\ell_1, \ell_2, \ell_\infty$	Strictly Tight	No	Yes
Ours (UniCR)	Any	Any continuous PDF	Any $\ell_p, p \in \mathbb{R}^+$	Approx. Tight	Yes	Yes

radius derived by Cohen et al. [11] is tied to the Gaussian noise and ℓ_2 perturbation. Recent works [51, 52, 12] propose methods to certify the robustness for multiple norms/noises, e.g., Yang et al. [51] propose the level set and differential method to derive the certified radii for multiple noise distributions. However, the certified radius derivation for different norms is still *subject to case-by-case theoretical analyses*. These methods, although achieving somewhat generalized certified robustness, are still lack of universality (See Table 1 for the summary).

In this paper, we develop the first Universally Approximated Certified Robustness (UniCR) framework based on *randomized smoothing*. Our framework can automate the robustness certification for any input on any classifier against any ℓ_p perturbation with noises generated by any *continuous probability density function (PDF)*. As shown in Figure 1, our UniCR framework provides four unique significant benefits to make certified robustness more universal, practical and easy-to-use with the above four "any"s. Our key contributions are as follows:

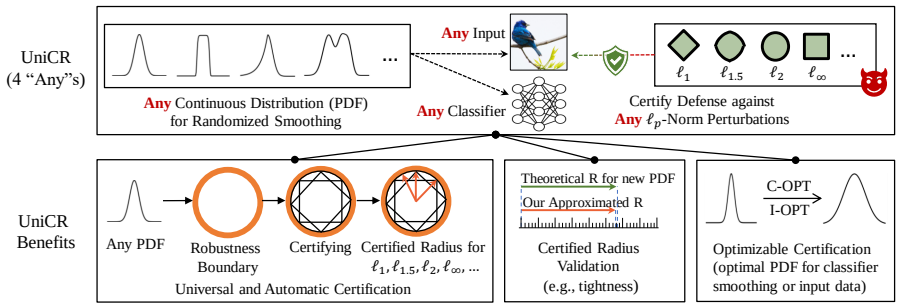


Fig. 1. Our Universally Approximated Certified Robustness (UniCR) framework.

1. **Universal Certification**. UniCR is the first universal robustness certification framework for the 4 "any"s.
2. **Automatic Certification**. UniCR provides an automatic robustness certification for all cases. It is easy-to-launch and avoids case-by-case analysis.
3. **Tightness Validation of Certified Radius**. It is also the first framework that can validate the *tightness* of the derived certified radius in existing

certification methods [35, 32, 11] or future methods based on any continuous noise PDF. In Section 3, we validate the tightness of the state-of-the-art certification methods (e.g., see Figure 4).

4. **Optimality Validation of Noise PDFs**. UniCR can also automatically tune the parameters in noise PDFs to strengthen the robustness certification against any ℓ_p perturbations. For instance, On CIFAR10 and ImageNet datasets, UniCR improves as high as 38.78% overall performance over the state-of-the-art certified defenses against all ℓ_p perturbations. In Section 5, we show that Gaussian noise and Laplace noise are not the optimal randomization distribution against the ℓ_2 and ℓ_1 perturbation, respectively.

2 Universally Approximated Certified Robustness

In this section, we propose the theoretical foundation for universally certifying a testing input against any ℓ_p perturbations with noise from any continuous PDF.

2.1 Universal Certified Robustness

Consider a general classification problem that classifies input data in $\mathbb{R}^d$ to a class belonging to a set of classes $\mathcal{Y}$. Given an input $x \in \mathbb{R}^d$, an *any* (base) classifier f that maps x to a class in $\mathcal{Y}$, and a random noise ϵ from *any* continuous PDF μ_x. We define a smoothed classifier g as the most probable class over the noise-perturbed input:

$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(x+\epsilon) = c) \tag{1}$$

Then, we show that the input has a certified accurate prediction against any l_p perturbation and its certified radius is given by the following theorem.

Theorem 1. ***(Universal Certified Robustness)*** *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random classifier, and let ϵ be drawn from an arbitrary continuous PDF μ_x. Denote g as the smoothed classifier in Equation (1), the most probable and second probable classes for predicting a testing input x via g as $c_A, c_B \in \mathcal{Y}$, respectively. If the lower bound of the class c_A's prediction probability $\underline{p_A} \in [0,1]$, and the upper bound of the class c_B's prediction probability $\overline{p_B} \in [0,1]$ satisfy:*

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

Then, we guarantee that $g(x+\delta) = c_A$ for all $||\delta||_p \leq R$, where R is called the **certified radius** *and it is the minimum ℓ_p-norm of all the adversarial perturbations δ that satisfies the* **robustness boundary conditions** *as below:*

$$\begin{aligned} &\mathbb{P}(\frac{\mu_x(x-\delta)}{\mu_x(x)} \leq t_A) = \underline{p_A}, \quad \mathbb{P}(\frac{\mu_x(x-\delta)}{\mu_x(x)} \geq t_B) = \overline{p_B}, \\ &\mathbb{P}(\frac{\mu_x(x)}{\mu_x(x+\delta)} \leq t_A) = \mathbb{P}(\frac{\mu_x(x)}{\mu_x(x+\delta)} \geq t_B) \end{aligned} \tag{3}$$

where t_A and t_B are auxiliary parameters to satisfy the above conditions.

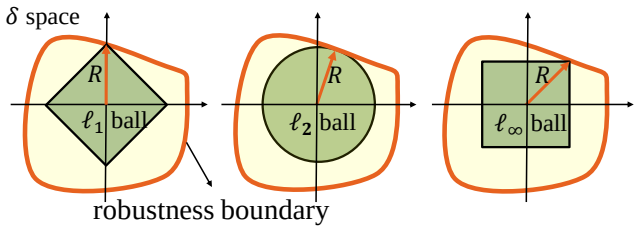


Fig. 2. An illustration to Theorem 1. The conditions in Theorem 1 construct a "**Robustness Boundary**" in δ space. In case of a perturbation inside the robustness boundary, the smoothed prediction can be certifiably correct. From left to right, the figures show that the minimum $||\delta||_1$, $||\delta||_2$ and $||\delta||_\infty$ on the robustness boundary are exactly the certified radius R in ℓ_1, ℓ_2 and ℓ_∞-norm, respectively.

Proof. See the detailed proof in Appendix B.1. □

Robustness Boundary. Theorem 1 provides a novel insight that meeting certain conditions is equivalent to deriving the certified robustness. The conditions in Equation (3) construct a boundary in the perturbation δ space, which is defined as the "*robustness boundary*". Within this robustness boundary, the prediction outputted by the smoothed classifier g is certified to be consistent and correct. The robustness boundary, rather than the certified radius, is actually more general to measure the certified robustness since the space constructed by each certified radius (against any specific ℓ_p perturbation) is only a subset of the space inside the robustness boundary. Figure 2 illustrates the relationship between certified radius and the robustness boundary against ℓ_1, ℓ_2 and ℓ_∞ perturbations.

Notice that, given any continuous noise PDF, the corresponding robustness boundary for all the ℓ_p-norms would naturally exist. Each maximum ℓ_p ball is a subspace of the robustness boundary, and gives the certified radius for that specific ℓ_p-norm. Thus, all the certified radii can be universally derived, and Theorem 1 provides a theoretical foundation to certify any input against any ℓ_p perturbations with any continuous noise PDF.

All ℓ_p Perturbations. Although we mainly introduce UniCR against ℓ_1, ℓ_2 and ℓ_∞ perturbations, our UniCR is not limited to these three norms. We emphasize that any $p \in \mathbb{R}^+$ (See Appendix D.5) can be used and our UniCR can derive the corresponding certified radius since our robustness boundary gives a general boundary in the δ perturbation space.

2.2 Approximating Tight Certified Robustness

The tight certified radius can be derived by finding a perturbation δ on the robustness boundary that has a minimum $\|\delta\|_p$ (for any $p \in \mathbb{R}^+$). However, it is challenging to either find a perturbation δ that is exactly on the robustness boundary, or find the minimum $||\delta||_p$. Here, we design an alternative two-phase

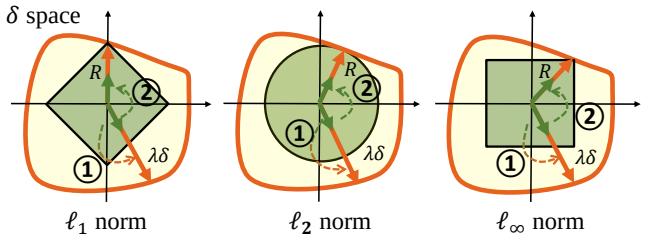


Fig. 3. An illustration to estimating the certified radius. The scalar optimization (①) and direction optimization (②) effectively find the minimum $||\delta||_p$ within the robustness boundary, which is the certified radius R.

optimization scheme to accurately approximate the tight certification in practice. In particular, Phase I is to suffice the conditions such that δ is on the robustness boundary, and Phase II is to minimize the ℓ_p-norm.

We perform Phase I by the "scalar optimization", where any perturbation δ will be λ-scaled to the robustness boundary (see ① in Figure 3). We perform Phase II by the "direction optimization", where the direction of δ will be optimized towards a minimum $\|\lambda\delta\|_p$ (see ② in Figure 3). In the two-phase optimization, the direction optimization will be iteratively executed until finding the minimum $||\lambda\delta||_p$, where the perturbation δ will be scaled to the robustness boundary beforehand in every iteration. Thus, the intractable optimization problem in Equation 3 can be converted to:

$$
\begin{aligned}
&R = ||\lambda\delta||_p, \\
&s.t. \quad \delta \in \arg\min_{\delta} ||\lambda\delta||_p, \quad \lambda = \arg\min_{\lambda} |K|, \\
&\qquad \mathbb{P}(\frac{\mu_x(x-\lambda\delta)}{\mu_x(x)} \leq t_A) = \underline{p_A}, \quad \mathbb{P}(\frac{\mu_x(x-\lambda\delta)}{\mu_x(x)} \geq t_B) = \overline{p_B}, \\
&\qquad K = \mathbb{P}(\frac{\mu_x(x)}{\mu_x(x+\lambda\delta)} \leq t_A) - \mathbb{P}(\frac{\mu_x(x)}{\mu_x(x+\lambda\delta)} \geq t_B).
\end{aligned} \tag{4}
$$

The scalar optimization in Equation (4) aims to find the scale factor λ that scales a perturbation δ to the boundary so that $|K|$ approaches 0. With the scalar λ for ensuring that the scaled δ is nearly on the boundary, the direction optimization optimizes the perturbation δ's direction to find the certified radius $R = ||\lambda\delta||_p$. We also present the theoretical analysis on the certification confidence and the optimization convergence in Appendix B.4 and B.5, respectively.

3 Deriving Certified Radius within Robustness Boundary

In this section, we will introduce how to universally and automatically derive the certified radius against any ℓ_p perturbations within the robustness boundary constructed by any noise PDF. In particular, we will present practical algorithms for solving the two-phase optimization problem to approximate the certified radius, empirically validate that our UniCR approximates the tight certified radius

derived by recent works [11, 51, 43], and finally discuss how to apply UniCR to validate the radius of existing certified defenses.

3.1 Calculating Certified Radius in Practice

Following the existing randomized smoothing based defenses [11, 43], we first use the Monte Carlo method to estimate the probability bounds ($\underline{p_A}$ and $\overline{p_B}$). Then, we use them in our two-phase optimization scheme to derive the certified radius.

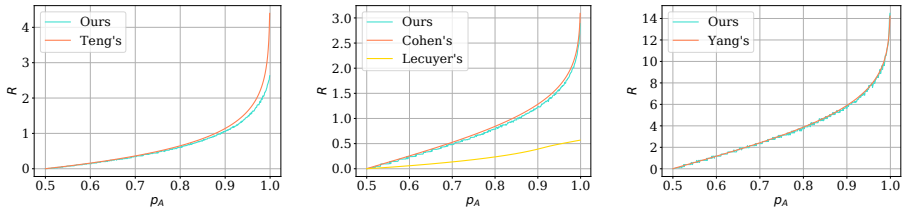
Estimating Probability Bounds. The two-phase optimization needs to estimate the probabilities bounds $\underline{p_A}$ and $\overline{p_B}$ and compute two auxiliary parameters t_A and t_B (required by the certified robustness based on the Neyman-Pearson Lemma in Appendix A). Identical to existing works [11, 43], the probabilities bounds $\underline{p_A}$ and $\overline{p_B}$ are commonly estimated by the Monte Carlo method [11]. Given the estimated $\underline{p_A}$ and $\overline{p_B}$ as well as any given noise PDF and a perturbation δ, we also use the Monte Carlo method to estimate the cumulative density function (CDF) of fraction $\mu_x(x-\lambda\delta)/\mu_x(x)$. Then, we can compute the auxiliary parameters t_A and t_B. Specifically, the auxiliary parameters t_A and t_B can be computed by $t_A = \Phi^{-1}(\underline{p_A})$ and $t_B = \Phi^{-1}(\overline{p_B})$, where Φ^{-1} is the inverse CDF of the fraction $\mu_x(x-\lambda\delta)/\mu_x(x)$. The procedures for computing t_A and t_B are detailed in Algorithm 1 in Appendix C.

Scalar Optimization. Finding a perturbation δ that is exactly on the robustness boundary is computationally challenging. Thus, we alternatively scale the δ to approach the boundary. We use the binary search algorithm to find a scale factor that minimizes $|K|$ (*the distance between δ and the robustness boundary*). The algorithm and detailed description are presented in Appendix C.2.

Direction Optimization. We use the Particle Swarm Optimization (PSO) method [29] to find δ that minimizes the ℓ_p-norm after scaling to the robustness boundary. In each iteration of PSO, the particle's position represents δ, and the cost function is $f_{PSO}(\delta) = ||\lambda\delta||_p$, where the scalar λ is found by the scalar optimization. The PSO aims to find the position δ that can minimize the cost function. To pursue convergence, we choose some initial positions in symmetry for different ℓ_p-norms. Empirical results show that the radius obtained by PSO with these initial positions can accurately approximate the tight certified radius. We show how to set the initial positions in Appendix C.3.

In our experiments, the certification (deriving the certified radius) can be efficiently completed on MNIST [31], CIFAR10 [30] and ImageNet [40] datasets (less than 10 seconds per image), as shown in Appendix D.4.

Certified Radius Comparison with State-of-The-Arts. We compare the certified radius obtained by our two-phase optimization method and that by the state-of-the-arts [11, 51, 43] and the comparison results are shown in Figure 4. Note that the certified radius is a function of p_A (the prediction probability of the top-1 class). The p_A-R curve can well depict the certified radius R w.r.t. p_A. We observe that our p_A-R curve highly approximates the tight theoretical curves in existing works, e.g., the Gaussian noise against ℓ_2 and ℓ_∞ perturbations [11, 51], Laplace noise against ℓ_1 perturbations [43], as well as General Normal noise and General Exponential noise derived by Yang et al. [51]'s method.



(a) Laplace vs. ℓ_1 (b) Gaussian vs. ℓ_2 (c) Gaussian vs. ℓ_∞

(d) Pareto vs. ℓ_1 (e) General Normal vs. ℓ_1 (f) Exponential Mix. vs. ℓ_1

Fig. 4. p_A-R curve comparison of our method and state-of-the-arts (i.e., Teng et al. [43], Cohen et al. [11], Lecuyer et al. [32] and Yang et al. [51]). We observe that the certified radius obtained by our UniCR is close to that obtained by the state-of-the-arts. These results demonstrate that our UniCR can approximate the tight certification to any input in any ℓ_p norm with any continuous noise distribution. We also evaluate our UniCR's defense accuracy against a diverse set of attacks, including universal attacks [10], white-box attacks [13, 48], and black-box attacks [1, 6], and against ℓ_1, ℓ_2 and ℓ_∞ perturbations. The experimental results show that UniCR is as robust as the state-of-the-arts (100% defense accuracy) against all the types of the real attack. The detailed experimental settings and results are presented in Appendix D.2.

Tightness Validation of Certified Radius. Since our UniCR accurately approximates the tight certified radius, it can be used as an auxiliary tool to validate whether an obtained certified radius is tight or not. For example, the certified radius derived by PixelDP [32][3] is loose, because [32]'s p_A-R curve in Figure 4(b) is far below ours. Also, Yang et al. [51] derives a low bound certified radius for Pareto Noise (Figure 4(d))— It shows that this certified radius is not tight either since it is below ours. For those theoretical radii that are slightly above our radii, they are likely to be tight.

Moreover, due to the high university, our UniCR can even derive the certified radii for complicated noise PDFs, e.g., mixture distribution in which the certified radii are difficult to be theoretically derived. In Section 5.2, we show some examples of deriving radii using UniCR on a wide variety of noise distributions in Figure 6-8. In most examples, the certified radii have not been studied before.

[3] PixelDP [32] adopts differential privacy [17], e.g., Gaussian mechanism to generate noises for each pixel such that certified robustness can be achieved for images.

4 Optimizing Noise PDF for Certified Robustness

UniCR can derive the certified radius using any continuous noise PDF for randomized smoothing. This provides the flexibility to optimize a noise PDF for enlarging the certified radius. In this section, we will optimize the noisy PDF in our UniCR framework for obtaining better certified robustness.

4.1 Noise PDF Optimization

All the existing randomized smoothing methods [11, 43, 51, 52] use the same noise for training the smoothed classifier and certifying the robustness of testing inputs. The motivation is that: the training can improve the lower bound of the prediction probability over the the same noise as the certification. Here, the question we ask is: Must we necessarily use the same noise PDF to train the smoothed classifier and derive the certified robustness? Our answer is No.

We study the master optimization problem that uses UniCR as a function to maximize the certified radius by tuning the noise PDF (different randomization), as shown in Figure 5. To defend against certain ℓ_p perturbations for a classifier, we consider the noise PDF as a variable (Remember that UniCR can provide a certified radius for each noise PDF), and study the following two master optimization problems with two different strategies:

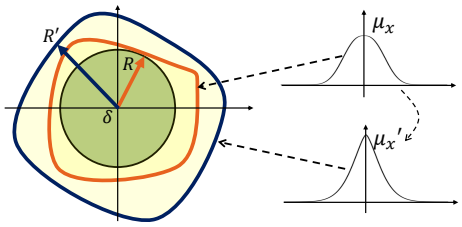


Fig. 5. An illustration to noise PDF optimization (take ℓ_2-norm perturbation as an example). The noise distribution is tuned from μ_x to μ_x', which enlarges the robustness boundary. Thus, UniCR can find a larger certified radius R'.

1. *Classifier-Input Noise Optimization* ("**C-OPT**"): finding the optimal noise PDF and injecting *the same noise* from this noise PDF into both the training data to train a classifier and testing input to build a smoothed classifier.
2. *Input Noise Optimization* ("**I-OPT**"): Training a classifier with the standard noise (e.g., Gaussian noise), while finding the optimal noise PDF for the testing input and injecting noise from this PDF into the testing input only.

4.2 C-OPT and I-OPT

Before optimizing the certified robustness, we need to define metrics for them. First, since I-OPT only optimizes the noise PDF when certifying each testing input, a "better" randomization in I-OPT can be directly indicated by a larger

certified radius for a specific input. Second, since C-OPT optimizes the noise PDF for the entire dataset in both training and robustness certification, a new metric for the performance on the entire dataset need to be defined.

Existing works [52, 51] draw several certified accuracy vs. certified radius curves computed by noise with different variances (See Figure 10 in Appendix D.1). These curves represent the certified accuracy at a range of certified radii, where the certified accuracy at radius R is defined as the percent of the testing samples with a derived certified radius larger than R (and correctly predicted by the smoothed classifier). To simply measure the overall performance, we use the area under the curve as an overall metric to the certified robustness, namely "robustness score". Then, we design the C-OPT method based on this metric. Specifically, the robustness score R_{score} is formally defined as below:

$$R_{score} = \int_0^{+\infty} \max_{\sigma}(Acc_{\sigma}(R))dR, \sigma \in \Sigma, \tag{5}$$

where $Acc_{\sigma}(R)$ is the certified accuracy at radius R computed by the noises with variance σ, and Σ is a set of candidate σ.

Notice that our UniCR can automatically approximate the certified radius and compute the robustness score w.r.t. different noise PDFs, thus we can tune the noise PDF towards a better robustness score. From the perspective of optimization, denoting the noise PDF as μ, the C-OPT and the I-OPT problems are defined as $\max_{\mu} R_{score}$ for a classifier and $\max_{\mu} R$ for an input, respectively.

Algorithms for Noise PDF Optimization. We use grid-search in C-OPT to search the best parameters of the noise PDF. We use Hill-Climbing algorithms in I-OPT to find the best parameters of the noise PDF around the noise distribution used in training while maintaining the certified accuracy.

Optimality Validation of Noise PDF. Finding an optimal noise PDF against a specific ℓ_p perturbation is important. Although Gaussian distribution can be used for defending against ℓ_2 perturbations with tight certified radius, there is no evidence showing that Gaussian distribution is the optimal distribution against ℓ_2 perturbations. Our UniCR can also somewhat validate the optimality of using different noise PDFs against different ℓ_p perturbations. For instance, Cohen et al. [11]'s certified radius is tight for Gaussian noise against ℓ_2 perturbations (see Figure 4(b)). However, it is validated as not-optimal distribution against ℓ_2 perturbations in our experiments (see Table 2).

5 Experiments

In this section, we thoroughly evaluate our UniCR framework, and benchmark with state-of-the-art certified defenses. First, we evaluate the *universality* of UniCR by approximating the certified radii w.r.t. the probability p_A using a variety of noise PDFs against ℓ_1, ℓ_2 and ℓ_∞ perturbations. Second, we validate the certified radii in existing works (results have been discussed and shown in Section 3). Third, we evaluate our noise PDF optimization on three real-world datasets. Finally, we compare our best certified accuracy on CIFAR10 [30] and ImageNet [40] with the state-of-the-art methods.

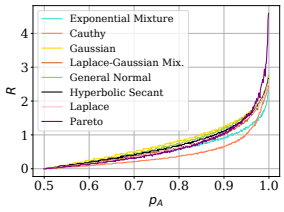


Fig. 6. R-p_A curve vs. ℓ_1

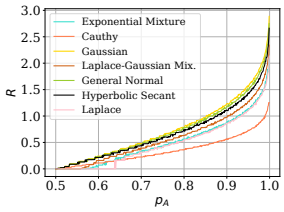


Fig. 7. R-p_A curve vs. ℓ_2

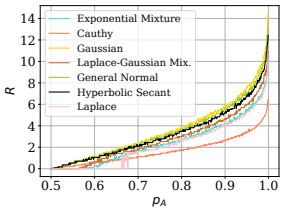


Fig. 8. R-p_A curve vs. ℓ_∞

5.1 Experimental Setting

Datasets. We evaluate our performance on MNIST [31], CIFAR10 [30] and ImageNet [40], which are common datasets to evaluate the certified robustness.
Metrics. We use certified accuracy[11] and the robustness score (Equation (5)) to evaluate the performance of proposed methods.
Experimental Environment. All the experiments were performed on the NSF Chameleon Cluster [28] with Intel(R) Xeon(R) Gold 6126 2.60GHz CPUs, 192G RAM, and NVIDIA Quadro RTX 6000 GPUs.

5.2 Universality Evaluation

As randomized smoothing derives certified robustness for any input and any classifier, our evaluation targets "any noise PDF" and "any ℓ_p perubations".

The certified radii of some noise PDFs, e.g., Gaussian noise against ℓ_2 perturbations [11], Laplace noise against ℓ_1 perturbations [43], Pareto noise against ℓ_1 perturbations [51], have been derived. These distributions have been verified by our UniCR framework in Figure 4, where our certified radii highly approximate these theoretical radii. However, there are numerous noise PDFs of which the certified radii have not been theoretically studied, or they are difficult to derive. It is important to derive the certified radii of these distributions in order to find the optimal PDF against each of the ℓ_p perturbations. Therefore, we use our UniCR to approximately compute the certified radii of numerous distributions (including some mixture distributions, see Table 7 in Appendix D.3), some of which have not been studied before. Specifically, we evaluate different noise PDFs with the same variance, i.e., $\sigma = \mathbb{E}_{\epsilon\sim\mu}[\sqrt{\frac{1}{d}||\epsilon||_2^2}] = 1$. For those PDFs with multiple parameters, we set β as 1.5, 1.0 and 0.5 for General Normal, Pareto, and mixture distributions, respectively. Following Cohen et al. [11], and Yang et al. [51], we consider the binary case (Theorem 3) and only compute the certified radius when $p_A \in (0.5, 1.0]$.

In Figure 6-8, we plot the R-p_A curves for the noise distributions listed in Table 7 in Appendix D.3 against ℓ_1, ℓ_2 and ℓ_∞ perturbations. Specifically, we present the ℓ_∞ radius scaled by $\times 255$ to be consistent with the existing works [52]. We observe that for all ℓ_p perturbations, the Gaussian noise generates the largest certified radius for most of the p_A values. All the noise distribution has

Table 2. Classifier-input noise optimization (C-OPT). We show the Robustness Score w.r.t. different β settings of General Normal distribution ($\propto e^{-|x/\alpha|^\beta}$). The σ is set to 1.0 for all distributions by adjusting the α parameter in General Normal.

β	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3	4.00	5.00
vs. ℓ_1	1.8999	2.6136	**2.8354**	2.7448	2.5461	2.4254	2.3434	2.2615	2.2211	2.1730	2.1081	2.0679	1.9610	1.8925
vs. ℓ_2	0.0000	0.0003	1.0373	1.5954	1.9255	2.0882	2.1746	2.1983	**2.2081**	2.1771	2.1184	2.0655	1.8857	1.7296
vs. ℓ_∞	0.0000	0.0109	0.0420	0.0641	0.0771	0.0839	0.0871	0.0879	**0.0880**	0.0870	0.0847	0.0825	0.0758	0.0693

very close R-p_A curves except the Cauthy distribution. We also notice that when p_A is low against ℓ_2 and ℓ_∞ perturbations, our UniCR cannot find the certified radius for the Laplace-based distributions, e.g., Laplace distribution, and Gaussian-Laplace mixture distribution. This matches the findings on injecting Laplace noises for certified robustness in Yang et al. [51]—The certified radii for Laplace noise against ℓ_2 and ℓ_∞ perturbations are difficult to derive.

We also conduct experiments to illustrate UniCR's universality in deriving ℓ_p norm certified radius for any real number $p > 0$ in Appendix D.5. Besides, we also conduct fine-grained evaluations on General Normal, Laplace-Gaussian Mixture, and Exponential Mixture noises with various β parameters (See Figure 13 in Appendix D.6), and we can draw similar observations from such results.

5.3 Optimizing Certified Radius with C-OPT

We next show how C-OPT uses UniCR to improve the certification against any ℓ_p perturbations. Recall that tight certified radii against ℓ_1 and ℓ_2 perturbations can be derived by the Laplace [43] and Gaussian [11] noises, respectively. However, there does not exist any theoretical study showing that Laplace and Gaussian noises are the optimal noises against ℓ_1 and ℓ_2 perturbations, respectively. [51, 52] have identified that there exists other better noise for ℓ_1 and ℓ_2 perturbations. Therefore, we use our C-OPT to explore the optimal distribution for each ℓ_p perturbation. Since the commonly used noise, e.g., Laplace and Gaussian noises, are only special cases of the General Normal Distribution ($\propto e^{-|x/\alpha|^\beta}$), we will find the optimal parameters α and β that generate the best noises for maximizing certified radius against each ℓ_p perturbation.

In the experiments, we use the grid search method to search the best parameters. We choose β as the main parameter, and α will be set to satisfy $\sigma = 1$. Specifically, we evaluate C-OPT on the MNIST dataset, where we train a model on the training set for each round of the grid search and certify $1,000$ images in the test set. Specifically, for each pair of parameters α and β in the grid search, we train a Multiple Layer Perception on MNIST with the smoothing noise. Then, we compute the robustness score over a set of $\sigma = [0.12, 0.25, 0.50, 1.00]$. When approximating the certified radius with UniCR, we set the sampling number as $1,000$ in the Monte Carlo method. The results are shown in Table 2.

We observe that the best β for ℓ_1-norm is 0.75 in the grid search. It indicates that the Laplace noise ($\beta = 1$) is not the optimal noise against ℓ_1 perturbations. A slightly smaller β can provide a better trade-off between the certified radius

Table 3. Average Certified Radius with Input Noise Optimization (I-OPT) against ℓ_1, ℓ_2 and ℓ_∞ perturbations on ImageNet.

Top ℓ_1 radius	20%	40%	60%	80%	100%
Yang's Gaussian [51]	**2.44**	2.10	1.59	1.19	0.95
Ours with I-OPT	2.36	**2.11**	**1.64**	**1.23**	**0.98**
Top ℓ_2 radius	20%	40%	60%	80%	100%
Cohen's Gaussian [11]	**2.43**	2.10	1.58	1.19	0.95
Ours with I-OPT	2.36	**2.11**	**1.64**	**1.23**	**0.98**
Top ℓ_∞ radius $\times 255$	20%	40%	60%	80%	100%
Yang's Gaussian [51]	1.60	1.38	1.04	0.78	0.63
Ours with I-OPT	**1.75**	**1.54**	**1.20**	**0.90**	**0.72**

and accuracy (measured by the robustness score). When $\beta < 1.0$, the radius is observed to be larger than the radius derived with Laplace noise at $p_A \approx 1$ (see Figure 13(a)). Since p_A on MNIST is always high, the noise distribution with $\beta = 0.75$ will give a larger radius at most cases. Furthermore, we observe that the best performance against ℓ_2 and ℓ_∞ are given by $\beta = 2.25$, showing that the Gaussian noise is not the optimal noise against ℓ_2 and ℓ_∞ perturbations, either.

5.4 Optimizing Certified Radius with I-OPT

The optimal noises for different inputs are different. We customize the noise for each input using the I-OPT. Specifically, we adapt the hyper-parameters in the noise PDF to find the optimal noise distribution for each input (the classifier is smoothed by a standard method such as Cohen's [11]).

We perform I-OPT for noise PDF optimization with a Gaussian-trained ResNet50 classifier ($\sigma = 1$) on ImageNet. We compare our derived radius with the theoretical radius in [51, 11]. We use the General Normal distribution to generate the noise for input certification since it provides a new parameter dimension for tuning, and tune the parameters α and β in $e^{-|x/\alpha|^\beta}$. The Gaussian distribution is only a specific case of the General Normal distribution with $\beta = 2$. In the two baselines [51, 11], they set $\sigma = 1$ and $\beta = 2$, respectively. In the I-OPT, we initialize the noise with the same setting, but optimize the noise for each input. The Monte Carlo sample is set to $1,000$ for ImageNet.

Table 3 presents the average values of the top 20%-100% certified radius (the higher the better). It shows that our I-OPT significantly improves the certified radius over the tight certified radius since it provides a personalized noise optimization to each input (see Figure 14 in Appendix E for the illustration).

5.5 Best Performance Comparison

In this section, we compare our best performance with the state-of-the-art certified defense methods on the CIFAR10 and ImageNet datasets. Following the setting in [11], we use a ResNet110 [23] classifier for the CIFAR10 dataset and a ResNet50 [23] classifier for the ImageNet dataset. We evaluate the certification performance with the noise PDF of a range of variances σ. The σ is set to vary in $[0.12, 0.25, 0.5, 1.0]$ for CIFAR10 and $[0.25, 0.5, 1.0]$ for ImageNet. We also

Table 4. Certified accuracy and robustness score against ℓ_1, ℓ_2 and ℓ_∞ perturbations on CIFAR10. Ours: General Normal with I-OPT.

ℓ_1 radius	0.50	1.00	1.50	2.00	2.50	R_{score}
Teng's Laplace [43]	39.2	17.2	10.0	6.0	2.8	0.5606
Ours	**45.8**	**22.4**	**14.8**	**8.2**	**3.6**	**0.7027**
ℓ_2 radius	0.50	1.00	1.50	2.00	2.50	R_{score}
Cohen's Gaussian [11]	38.6	17.4	8.6	3.4	1.6	0.5392
Ours	**48.4**	**26.8**	**16.6**	**6.8**	**2.0**	**0.7141**
ℓ_∞ radius	$\frac{2}{255}$	$\frac{4}{255}$	$\frac{6}{255}$	$\frac{8}{255}$	$\frac{10}{255}$	R_{score}
Yang's Gaussian [51]	43.6	21.8	10.8	5.6	2.6	0.0098
Ours	**53.4**	**30.4**	**21.2**	**13.2**	**5.6**	**0.0136**

Table 5. Certified accuracy and robustness score against ℓ_1, ℓ_2 and ℓ_∞ perturbations on ImageNet (Teng's Laplace [43] is not available). Ours: General Normal with I-OPT.

ℓ_1 radius	0.50	1.00	1.50	2.00	2.50	R_{score}
Yang's Gaussian [51]	58.8	45.6	34.6	27.0	0.0	1.0469
Ours	**63.4**	**49.6**	**36.8**	**29.6**	**6.6**	**1.1385**
ℓ_2 radius	0.50	1.00	1.50	2.00	2.50	R_{score}
Cohen's Gaussian [11]	58.8	44.2	34.0	27.0	0.0	1.0463
Ours	**62.6**	**49.0**	**36.6**	**28.6**	**2.0**	**1.0939**
ℓ_∞ radius	$\frac{0.25}{255}$	$\frac{0.50}{255}$	$\frac{0.75}{255}$	$\frac{1.00}{255}$	$\frac{1.25}{255}$	R_{score}
Yang's Gaussian [51]	63.6	52.4	39.8	34.2	28.0	0.0027
Ours	**69.2**	**57.4**	**47.2**	**38.2**	**33.0**	**0.0031**

present the Robustness Score based on this set of variances. We use the General Normal distribution and perform the I-OPT. The distribution is initialized with the same setting in the baselines, e.g., $\beta = 1$ (or 2) for Laplace (Gaussian) baseline. We benchmark it with the Laplace noise [43] on CIFAR10 when against ℓ_1 perturbations; and the Gaussian noise [11, 51] on both CIFAR10 and ImageNet against all ℓ_p perturbations. For both our method and baselines, we use $1,000$ and $4,000$ Monte Carlo samples on ImageNet and CIFAR10, respectively, due to different scales, and the certified accuracy is computed over the certified radius of 500 images randomly chosen in the test set for both CIFAR10 and ImageNet.

The results are shown in Table 4 and 5. Both on CIFAR10 and ImageNet, we observe a significant improvement on the certified accuracy and robustness score. Specifically, on CIFAR10, our robustness score outperforms the state-of-the-arts by 25.34%, 32.44% and 38.78% against ℓ_1, ℓ_2 and ℓ_∞ perturbations, respectively. On ImageNet, our robustness score outperforms the state-of-the-arts by 8.75%, 4.55% and 14.81% against ℓ_1, ℓ_2 and ℓ_∞ perturbations, respectively.

6 Related Work

Certified Defenses. Existing certified defenses methods can be classified into leveraging Satisfiability Modulo Theories [41, 4, 18, 27], mixed integer-linear programming [8, 19, 3], linear programming [47, 49], semidefinite programming [38, 39], dual optimization [14, 15], global/local Lipschitz constant methods [21, 44, 2, 9, 24], abstract interpretation [20, 37, 42], and layer-wise certification [37, 42, 22, 46, 53], etc. However, none of these methods is able to scale to large models (e.g., deep neural networks) or is limited to specific type of network architec-

ture, e.g., ReLU based networks. Randomized smoothing was recently proposed certified defenses [32, 35, 11, 25, 45] that is scalable to large models and applicable to arbitrary classifiers. Lecuyer et al. [32] proposed the first randomized smoothing-based certified defense via differential privacy [17]. Li et al. [35] proposed a stronger guarantee for Gaussian noise using information theory. The first tight robustness guarantee against l_2-norm perturbation for Gaussian noise was developed by Cohen et al. [11]. After that, a series follow-up works have been proposed for other ℓ_p-norms, e.g., ℓ_1-norm [43], ℓ_0-norm [34, 33, 26], etc. However, all these methods are limited to guarantee the robustness against only a specific ℓ_p-norm perturbation.

Universal Certified Defenses. More recently, several works [52, 51] aim to provide more universal certified robustness schemes for all ℓ_p-norms. Yang et al. [51] proposed a level set method and a differential method to derive the upper bound and lower bound of the certified radius, while the derivation is relying on the case-by-case theoretical analysis. Zhang et al. [52] proposed a black-box optimization scheme that automatically computes the certified radius, but the solvable distribution is limited to ℓ_p-norm-constrained. Our UniCR framework can automate the robustness certification for any classifier against any l_p-norm perturbation with any noise PDF.

Certified Defenses with Optimized Noise PDFs/Distributions. Yang et al. [51] proposed to use the Wulff Crystal theory [50] to find optimal noise distributions. Zhang et al. [52] claimed that the optimal noise should have a more central-concentrated distribution from the optimization perspective. However, no existing works provide quantitative solutions to find optimal noise distributions. We propose the **C-Opt** and **I-Opt** schemes to quantitatively optimize the noisy PDF in our UniCR framework and provide better certified robustness. Table 1 summarizes the differences in all the closely-related works.

7 Conclusion

Randomize smoothing has achieved great success in certifying the adversarial robustness. However, the state-of-the-art methods lack universality to certify robustness. We propose the first randomized smoothing-based universal certified robustness approximation framework against any ℓ_p perturbations with any continuous noise PDF. Extensive evaluations on multiple image datasets demonstrate the effectiveness of our UniCR framework and its advantages over the state-of-the-art certified defenses against any ℓ_p perturbations.

Acknowledgement

This work is partially supported by the National Science Foundation (NSF) under the Grants No. CNS-2046335 and CNS-2034870, as well as the Cisco Research Award. In addition, results presented in this paper were obtained using the Chameleon testbed supported by the NSF. Finally, the authors would like to thank the anonymous reviewers for their constructive comments.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: European Conference on Computer Vision. pp. 484–501. Springer (2020)
2. Anil, C., Lucas, J., Grosse, R.: Sorting out lipschitz function approximation. In: International Conference on Machine Learning. pp. 291–301. PMLR (2019)
3. Bunel, R.R., Turkaslan, I., Torr, P.H., Kohli, P., Mudigonda, P.K.: A unified view of piecewise linear neural network verification. In: NeurIPS (2018)
4. Carlini, N., Katz, G., Barrett, C., Dill, D.L.: Provably minimally-distorted adversarial examples. arXiv preprint arXiv:1709.10207 (2017)
5. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017)
6. Chen, J., Jordan, M.I., Wainwright, M.J.: Hopskipjumpattack: A query-efficient decision-based attack. In: 2020 IEEE Symposium on Security and Privacy (2020)
7. Chen, P.Y., Zhang, H., Sharma, Y., Yi, J., Hsieh, C.J.: Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In: 10th ACM workshop on artificial intelligence and security (2017)
8. Cheng, C.H., Nührenberg, G., Ruess, H.: Maximum resilience of artificial neural networks. In: International Symposium on Automated Technology for Verification and Analysis. pp. 251–268. Springer (2017)
9. Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y.N., Usunier, N.: Parseval networks: Improving robustness to adversarial examples. In: Proceedings of the 34th International Conference on Machine Learning (2017)
10. Co, K.T., Muñoz-González, L., de Maupeou, S., Lupu, E.C.: Procedural noise adversarial examples for black-box attacks on deep convolutional networks. In: ACM SIGSAC conference on computer and communications security (2019)
11. Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: International Conference on Machine Learning (2019)
12. Croce, F., Hein, M.: Provable robustness against all adversarial l_p-perturbations for $p\geq 1$. In: ICLR. OpenReview.net (2020)
13. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
14. Dvijotham, K., Gowal, S., Stanforth, R., et al.: Training verified learners with learned verifiers. arXiv (2018)
15. Dvijotham, K., Stanforth, R., Gowal, S., Mann, T.A., Kohli, P.: A dual approach to scalable verification of deep networks. In: UAI (2018)
16. Dvijotham, K.D., Hayes, J., Balle, B., Kolter, J.Z., Qin, C., György, A., Xiao, K., Gowal, S., Kohli, P.: A framework for robustness certification of smoothed classifiers using f-divergences. In: ICLR (2020)
17. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Found. Trends Theor. Comput. Sci. **9**(3-4), 211–407 (2014)
18. Ehlers, R.: Formal verification of piece-wise linear feed-forward neural networks. In: International Symposium on Automated Technology for Verification and Analysis. pp. 269–286. Springer (2017)
19. Fischetti, M., Jo, J.: Deep neural networks and mixed integer linear optimization. Constraints **23**(3), 296–309 (2018)
20. Gehr, T., Mirman, M., Drachsler-Cohen, D., Tsankov, P., Chaudhuri, S., Vechev, M.: Ai2: Safety and robustness certification of neural networks with abstract interpretation. In: IEEE S & P (2018)

21. Gouk, H., Frank, E., Pfahringer, B., Cree, M.J.: Regularisation of neural networks by enforcing lipschitz continuity. Machine Learning **110**(2), 393–416 (2021)
22. Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T.A., Kohli, P.: On the effectiveness of interval bound propagation for training verifiably robust models. CoRR **abs/1810.12715** (2018), http://arxiv.org/abs/1810.12715
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
24. Hein, M., Andriushchenko, M.: Formal guarantees on the robustness of a classifier against adversarial manipulation. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 2266–2276 (2017)
25. Jia, J., Cao, X., Wang, B., Gong, N.Z.: Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In: International Conference on Learning Representations (2019)
26. Jia, J., Wang, B., Cao, X., Liu, H., Gong, N.Z.: Almost tight l0-norm certified robustness of top-k predictions against adversarial perturbations. In: ICLR (2022)
27. Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient smt solver for verifying deep neural networks. In: International Conference on Computer Aided Verification. pp. 97–117. Springer (2017)
28. Keahey, K., Anderson, J., Zhen, Z., Riteau, P., Ruth, P., Stanzione, D., Cevik, M., Colleran, J., Gunawi, H.S., Hammock, C., Mambretti, J., Barnes, A., Halbach, F., Rocha, A., Stubbs, J.: Lessons learned from the chameleon testbed. In: Proceedings of the 2020 USENIX Annual Technical Conference (USENIX ATC '20). USENIX Association (July 2020)
29. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks. IEEE (1995)
30. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
31. LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist **2** (2010)
32. Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., Jana, S.: Certified robustness to adversarial examples with differential privacy. In: 2019 IEEE Symposium on Security and Privacy (SP). pp. 656–672. IEEE (2019)
33. Lee, G., Yuan, Y., Chang, S., Jaakkola, T.S.: Tight certificates of adversarial robustness for randomly smoothed classifiers. In: NeurIPS. pp. 4911–4922 (2019)
34. Levine, A., Feizi, S.: Robustness certificates for sparse adversarial attacks by randomized ablation. In: AAAI. pp. 4585–4593. AAAI Press (2020)
35. Li, B., Chen, C., Wang, W., Carin, L.: Second-order adversarial attack and certifiable robustness. arXiv preprint arXiv:2006.00731 (2020)
36. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (2018)
37. Mirman, M., Gehr, T., Vechev, M.: Differentiable abstract interpretation for provably robust neural networks. In: International Conference on Machine Learning (2018)
38. Raghunathan, A., Steinhardt, J., Liang, P.: Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344 (2018)
39. Raghunathan, A., Steinhardt, J., Liang, P.S.: Semidefinite relaxations for certifying robustness to adversarial examples. In: NeurIPS (2018)

40. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
41. Scheibler, K., Winterer, L., Wimmer, R., Becker, B.: Towards verification of artificial neural networks. In: MBMV. pp. 30–40 (2015)
42. Singh, G., Gehr, T., Mirman, M., Püschel, M., Vechev, M.: Fast and effective robustness certification. In: Proceedings of the 32nd International Conference on Neural Information Processing Systems. pp. 10825–10836 (2018)
43. Teng, J., Lee, G.H., Yuan, Y.: ℓ_1 adversarial robustness certificates: a randomized smoothing approach (2020)
44. Tsuzuku, Y., Sato, I., Sugiyama, M.: Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In: NeurIPS (2018)
45. Wang, B., Cao, X., Gong, N.Z., et al.: On certifying robustness against backdoor attacks via randomized smoothing. In: CVPR 2020 Workshop on Adversarial Machine Learning in Computer Vision (2020)
46. Weng, T., Zhang, H., Chen, H., Song, Z., Hsieh, C., Daniel, L., Boning, D.S., Dhillon, I.S.: Towards fast computation of certified robustness for relu networks. In: Dy, J.G., Krause, A. (eds.) International Conference on Machine Learning (2018)
47. Wong, E., Kolter, J.Z.: Provable defenses against adversarial examples via the convex outer adversarial polytope. In: ICML (2018)
48. Wong, E., Schmidt, F., Kolter, Z.: Wasserstein adversarial examples via projected sinkhorn iterations. In: International Conference on Machine Learning (2019)
49. Wong, E., Schmidt, F.R., Metzen, J.H., Kolter, J.Z.: Scaling provable adversarial defenses. arXiv preprint arXiv:1805.12514 (2018)
50. Wul, G.: Zur frage der geschwindigkeit des wachstums und der auflosung der kristall achen. Z. Kristallogr **34**, 449–530 (1901)
51. Yang, G., Duan, T., Hu, J.E., Salman, H., Razenshteyn, I., Li, J.: Randomized smoothing of all shapes and sizes. In: International Conference on Machine Learning. pp. 10693–10705. PMLR (2020)
52. Zhang, D., Ye, M., Gong, C., Zhu, Z., Liu, Q.: Black-box certification with randomized smoothing: A functional optimization based framework (2020)
53. Zhang, H., Weng, T., Chen, P., Hsieh, C., Daniel, L.: Efficient neural network robustness certification with general activation functions. In: Neural Information Processing Systems (2018)