UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Non-Axis Aligned Anisotropic Certification with Randomized Smoothing

Author:
Taras RUMEZHAK

Supervisor:
Adel BIBI

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Taras RUMEZHAK, declare that this thesis titled, "Non-Axis Aligned Anisotropic Certification with Randomized Smoothing" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"Attack is the secret of defense; defense is the planning of an attack."

Sun Tzu

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Non-Axis Aligned Anisotropic Certification with Randomized Smoothing

by Taras RUMEZHAK

Abstract

Modern classifiers have been proven to perform well in different domains as they achieve great performance in applied real-world tasks. State-of-the-art neural networks help humans to analyze medical images, make decisions about whether a bank should give a loan to a particular client, or control the self-driving car. Therefore we must be confident in their performance and if we can trust them. However every neural network was proven to be unprotected from adversarial attacks making wrong predictions or decisions in safety-critical applications. Therefore the defense against them is very crucial nowadays. Many works were dedicated to this topic, but randomized smoothing has been recently proven to be an effective state-of-the-art approach for the certification (guaranteed robustness) of deep neural networks and obtaining robust classifiers. Some prior results were obtained utilizing the techniques of adding extra parameters to extend the limits of the regions that can be certified. In this way, sample-wise optimization was proposed to maximize the certification radius per input. This idea was further extended with the generalized anisotropic counterparts of ℓ_1 and ℓ_2 certificates which allow achieving larger certified region volume avoiding worst-case certification near potentially larger safe regions. However, anisotropic certification is limited by the aligned axis lacking the freedom to extend in any direction. To mitigate this constraint, in this work, we **(i)** revisit the anisotropic certification, provide an analysis of its non-axis aligned counterpart and propose its rotation-free extension, **(ii)** conduct experiments on custom toy and academic CIFAR-10 datasets to prove the improved performance.

Acknowledgements

Probably there are no suitable words to describe the support of my supervisor Adel Bibi from the University of Oxford (Torr Vision Group) who inspired me to start this research and with his strong support pushed me to finish it. Thank you for sharing your huge experience with me, providing a lot of new ideas and potential improvements, and helping to overcome all technical and mental issues. The same gratitude goes to Francisco Eiras for his strong support and useful ideas on the main stage of the thesis experiments. I would also like to thank Oles Dobosevych and Rostyslav Hryniv for their additional support throughout the whole process of the research and for answering all my questions to which I was struggling to find the answer. Huge respect for my colleagues from SoftServe for helping me improve the quality of the thesis. I can't help mentioning my gratefulness to my family and my fiancee who motivated me the most and were asking the updates on my thesis every week. And last but not least I want to thank my Machine Learning Laboratory, Faculty of Applied Sciences, the Ukrainian Catholic University, its community, and lecturers for the best 4 years of my life and for giving me a strong background in the research field for writing this thesis and to learn how to learn.

Contents

List of Figures

List of Tables

List of Abbreviations

SOTA	**S**tate **O**f **T**he **A**rt
NN	**N**eural **N**etwork
DNN	**D**eep **N**eural **N**etwork
DDS	**D**ata **D**ependant **S**moothing
ANCER	**AN**isotropic **CER**tification via Sample-wise Volume Maximization
ReLU	**Re**ctified **L**inear **U**nit
SGA	**S**tochastic **G**radient **A**scent
FC	**F**ully **C**onnected
ACR	**A**verage **C**ertified **R**adius
AC$\tilde{R}$	**A**verage **C**ertified proxy **R**adius
ARI	**A**verage **R**adius **I**mprovement
A$\tilde{R}$I	**A**verage proxy **R**adius **I**mprovement
CDF	**C**umulative **D**istribution **F**unction

Dedicated to my beloved Viktoria Paliichuk

Chapter 1

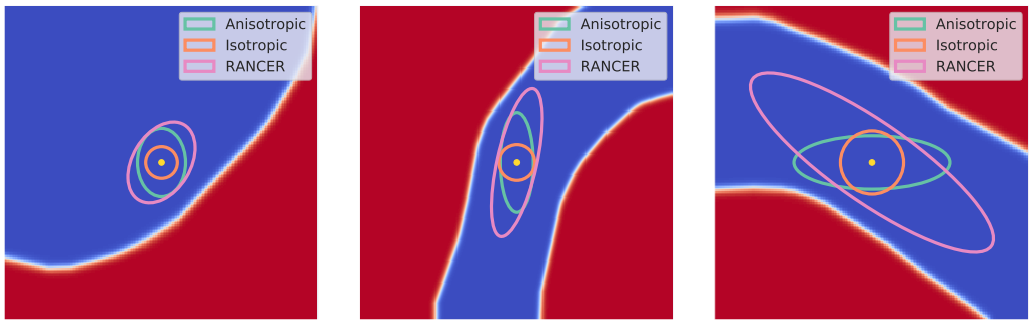
Introduction

1.1 Motivation

In the past decade, image classifiers have been proven to perform well in different fields. Modern SOTA approaches achieve great performance in classifying real-life images of different domains and sometimes can even outperform the human being for example in medical imaginary. This fact raises the discussion about the explainability of current NNs and the question if we can trust them. Every time when we train the network, we split our data to train, test and validation sets adding the verification on how the network is generalizable, stable and robust. But even networks with high results on test sets are vulnerable to small adversarially chosen perturbations. It means that you have an almost ideal classifier of cats and dogs with almost 100% accuracy on the test set and one can choose such perturbation that changes only a few pixels on the input image (you will not even recognize it) that classifier will swap the predictions and will be totally sure that cat image is a dog.

While fouling the cat/dog classifier seems harmless, living at the beginning of the era of autonomous vehicles forces us to rethink if we can trust NN for our safety. Even some small physical attack on the road sign (just put a sticker on it) can cause misprediction and instead of a "Stop" sign the car will interpret it as "Main Road". So the defense against adversarial attacks is very crucial nowadays because they are being used in safety-critical applications. Therefore we need to have certifiably robust classifiers that we know are not going to fail for budgeted adversaries, meaning if the adversary will have a certain budget (for example some ℓ_2 difference) we can guarantee that the robust classifier will not have worth performance than a certain value which can be directly computed.

Several papers have done a lot of progress in this direction and in our thesis we are trying to come up with ways to certify beyond the isotropic regions - non-axis aligned anisotropic regions. For example, there has been lots of work on providing certificates for ℓ_2 regions. By certificate we mean that prediction is constant over an ℓ_2 image change. Then there was an extension where certification was generalized for anisotropic regions providing the certificates over larger regions. And in our work, we look at generalization where those sets where a classifier is predicting correctly could be rotated differently and they do not need to be axis-aligned. This is important because with this setup the certificate is aware of the structure of the decision boundary and one will be able to certify more by this generalization compared to the previous works. And in this thesis, we worked with a randomized smoothing framework to do such extension and provide better certificates on that. We have proved our theoretical analysis by conducting experiments on the toy and CIFAR-10 [20] datasets.



(a) (b) (c)

FIGURE 1.1: Example of the ℓ_2 certificates regions presented on the $2D$ toy dataset, where the blue and red regions correspond to different data classes. **Orange:** Data dependent isotropic region [1], **Blue:** Anisotropic (ANCER) region [11], **Pink:** Certification region obtained with our proposed solution - RANCER (see Chapter 4)

1.2 Contributions

One can observe huge progress in the current state of randomized smoothing approaches but despite this, most of them provide the isotropic certification regions. Recent paper by Eiras et. al. [11] proposed a state-of-the-art framework with generalized anisotropic certificates giving the best certified accuracy on the CIFAR-10 dataset. In our work, we generalize anisotropic regions beyond axis-aligned counterparts and thereof we can summarize our contributions in:

- We provide a deep general analysis of non-axis aligned anisotropic certification while preserving previous approaches as special cases.
- We conduct experiments in both: toy and CIFAR-10 datasets. As a consequence of conducted experiments, we show that our generalized framework[1] outperforms existing approaches and shows better results in certified accuracy for ℓ_2 on CIFAR-10.

1.3 Structure Of The Thesis

1.3.1 Chapter 2. Related Work

In this chapter, we provide a general background of adversarial attacks and their types to familiarize with the research topic as well as a detailed overview of the two main categories of the defenses against previously mentioned adversarial attacks with a vast amount of research in this critical field.

1.3.2 Chapter 3. Theoretical Background On Randomized Smoothing

Here you can find the general principles of randomized smoothing as well as a description of the three main methods proposed recently for certified defenses against adversarial attacks.

[1] https://github.com/tarasrumezhak/RANCER

1.3.3 Chapter 4. Proposed Solution

This chapter describes our ideas on non-axis aligned anisotropic certification extension and brand-new methods to improve the performance of previous SOTA approaches.

1.3.4 Chapter 5. Experiments and Results

The chapter contains information about the detailed experiment setup together with used neural networks architectures, datasets and evaluation metrics, all the results, and a comparison of our algorithm with recent SOTA ones.

1.3.5 Chapter 6. Conclusion

Here we summarize all the provided theoretical background and experiments we did and reflect on the limitations of the proposed solution and potential improvements.

Chapter 2

Related Work

In this section we will provide a base introduction of adversarial attacks, how they can be performed and classified by different types. On the opposite, we will also overview the two main categories of defense against adversarial attacks. They are mainly divided into empirical which can be empirically robust to some particular types of attacks and certified defenses which are definitely robust to the adversarial perturbations which can be proved for some region.

2.1 Adversarial Attacks

In the 2013 Szegdy et al. [35] discussed the stability of NNs with respect to small perturbations to their input. Such intentionally designed changes cause the classifier to make mistakes and we call them adversarial attacks. In a study by Kloft et al. [19] it was shown that inserting malicious points in the training set could gradually shift the decision boundary of a classifier. Later Goodfellow et al. [13] introduced the adversarial examples (perturbed samples) together with the exact way how to generate them. In the case discussed in the thesis, we will overview only image data examples where adversaries look almost identical to the original image but still can lead to mispredictions.

With the vast popularity of NNs their security was doubted a long time ago, Bareno et al. [3] even in 2006 provided a taxonomy of different attacks and how we can protect NNs against them. Such techniques as regularization or randomization were supposed to increase robustness against particular attacks. In 2012 Ukrainian-born Alex Krizhevsky caused a rise in popularity of DNNs after winning the ImageNet challenge with his AlexNet architecture. Many people hoped that the fact of adding a high nonlinearity with deep layers will defend networks from attacks. But two years later Goodfellow et al. [13] demonstrated the fast gradient sign method to generate adversarial examples (see Figure 2.1) and broke all previous hopes.

Assume having parameters of a model θ, input image x with corresponding target y. And let the cost function which was used to train the classifier be $J(\theta, x, y)$. With this setup, we can linearize the cost function around the current value of θ, obtaining an optimal max-norm constrained perturbation of

$$\eta = \epsilon \operatorname{sign}\left(\nabla_x J(\theta, x, y)\right). \tag{2.1}$$

The authors have proven that this method causes a lot of different models to make the wrong classification output, for example, 99.9% error rate with shallow classifier on MNIST [22] with $\epsilon = 0.25$.

There are many ways to categorize adversarial attacks in the literature:

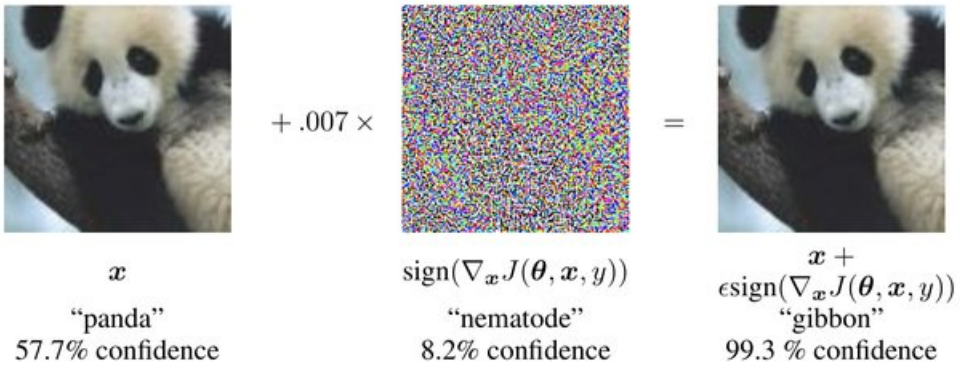


FIGURE 2.1: A demonstration of fast adversarial example generation applied to GoogLeNet on ImageNet. *Source:* [13]

2.1.1 By Target

The untargeted attack is the type described before in the thesis where the attacker's goal is to make the model return the incorrect prediction no matter which one. ([13]) Targeted attacks are more precise as they give us the opportunity to specify the exact output class which attacked models will predict. In this setting when creating an adversary we are changing it iteratively with the new loss function, which is constructed from two terms: original and target loss as we want to minimize the probability for the original true class and maximize it for the target class. This target loss term helps us to construct a targeted attack. The same method can be applied to both types of attacks, for example presented by Dong et al. [8].

2.1.2 By Access To The Network

White-box attack assumes that we have all the necessary information about attacked NN: architecture, weights, gradients, etc. This information helps to construct specific adversarial examples. ([13]) With black-box attacks our information is totally limited, the only thing we can get is the output of the model. In this setting, attackers try to investigate the dependence between input and output by giving a lot of samples as the input. A practical example of such attack was presented by Papernot et al. in [30].

2.1.3 By The Way Of Attack

Poisoning Attack - some deployed models are being retrained with the newly observed data, attackers can use this to poison the samples with wrong labels causing mispredictions later. ([36], [26]) Evasion Attack - attacker tries to foul previously trained and already deployed classifier. ([13])

2.2 Defences Against Adversaries

There have been several approaches towards building models to defend against adversaries and these methods can be generally divided into two categories:

2.2.1 Empirical Defenses

In 2015 Goodfellow et al. [13] showed the way to attack DNNs and in the same paper presented the adversarial training. With the help of the fast gradient sign method mentioned before (Equation 2.1), we can quickly generate adversarial examples during training which makes it possible to apply in practice. In the mentioned procedure they train the network on adversarial examples. The authors also found that training with an adversarial objective function based on the fast gradient sign method was an effective regularizer making the model more robust to adversarial attacks:

$$\bar{J}(\boldsymbol{\theta}, \boldsymbol{x}, y) = \alpha J(\boldsymbol{\theta}, \boldsymbol{x}, y) + (1 - \alpha) J\left(\boldsymbol{\theta}, \boldsymbol{x} + \epsilon \operatorname{sign}\left(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y)\right)\right). \tag{2.2}$$

Then Kurakin et al. [21] in 2017 showed how to apply the explicit model training on the adversarial examples on a big scale. The previously mentioned paper showed the result on MNIST dataset [22] while here the authors apply adversarial training on ImageNet [7] and wrote the exact recommendations for how to successfully scale adversarial training to large models and datasets. One year later Madry et al. [27] started to think about fully resistant NNs which can be robust to the wide range of adversarial attacks. They discussed how to find more powerful examples of attacking during training to make the model more robust. But in reality, it is impossible to prove if a prediction of a "robust" classifier trained with adversarial training is truly robust. It was shown later that such models were robust to only specific kinds of attacks and were successfully broken by stronger adversaries. Carlini and Wagner [4] surveyed ten recent methods and carried out additional experiments on them. They showed that all previous attacks can be defeated by carefully constructing new loss as well as new adversarial examples are harder to defeat than was mentioned in previous studies. They also provided some recommendations for defenses for instance to evaluate a model for harder attacks. And actually, the research community was pulling the rope: new attacks were breaking previously robust models and then new defenses were proposed and so on. For example, Athalye et al. [2] analyzed that obfuscated gradients (a special case of gradient masking) were a common occurrence, with 7 of 9 defenses relying on obfuscated gradients in ICLR 2018. The authors showed that their new attacks successfully circumvent 6 completely, and 1 partially, in the original threat model each paper (out of 7) considers.

With those studies and pulling the rope side by side with each new paper, there was less trust in empirical defenses and we can observe the increasing interest in papers focused on defenses with formal guarantees.

2.2.2 Certified Defenses

As Cohen et al. [5] defined a classifier is certifiably robust if for any input x, one can easily obtain a guarantee that the classifier's prediction is constant within some set around x, often ℓ_1, ℓ_2 or ℓ_∞ ball. Here we must mention that certification works for both: generically trained NNs and robustly trained ones. For example, Wong and Kolter [40] proposed the method to learn deep ReLU-based classifiers that are provably robust against norm-bounded adversarial perturbations on the training data.

Exact Certification

There were some works proposing exact certification which is very straightforward: just take a smoothed classifier g, and check if there exists a perturbation with a norm lower than some r. Then we report whether the output corresponding to perturbed input is the same as the output for original input. In that case, classifier is certifiably robust for r. An example of such an approach was presented by Ehlers [10]. His approach was used for the verification of feed-forward neural networks in which all nodes have a piece-wise linear activation function. But the problem with such kinds of methods is the lack of possibility to scale to large NNs. Tjeng et al. [38] formulated verification as a mixed-integer program and were able to speed up computations and certify networks with over 100 000 ReLUs to determine the exact adversarial accuracy on MNIST to perturbations with bounded ℓ_∞ norm $\epsilon = 0.1$. But even this and some other recent achievements do not scale to the SOTA networks working with harder datasets such as CIFAR10 [20] or ImageNet [7].

Conservative Certification

Then conservative methods come here which are usually utilizing the global or local Lipshitz constants of the network and are more scalable but they are computationally hard for modern networks. Tsuzuku et al. [39] presented an efficient calculation technique to lower-bound the size of adversarial perturbations that can deceive networks from the relationship between the Lipschitz constants and prediction margins. Hein and Andriushchenko [15] gave the formal guarantees on the robustness of a classifier by giving instance-specific lower bounds on the norm of the input manipulation required to change the classifier decision and proposed the Cross-Lipschitz regularization functional.

Generally, there are two corner problems for both mentioned categories of certification: poor scalability for exact methods and conservative methods are too expansive.

2.2.3 Randomized Smoothing

This type of certified adversarial robustness is used in many modern pieces of research as well as in our thesis. Everything began in 2019 with a paper proposed by Lecuyer et al. [23] and their defense called PixelDP. The authors proposed the first certified defense that both scales to large networks and datasets (such as Google's Inception network [34] for ImageNet [7]) and applies broadly to arbitrary model types. They successfully used the techniques from differential privacy where randomness is introduced into the computation. In this work, Laplacian noise was added to the inputs which enjoy ℓ_1 certification and the result was proved to be constant with average classifier prediction. These ideas were later improved for the ℓ_2 certification by Cohen et al. [6] for smoothing with Gaussian noise. In their work, the authors showed how to turn any classifier that classifies well under Gaussian noise into a new classifier that is certifiably robust to adversarial perturbations under the ℓ_2 norm. Later there were some other papers that showed the proofs for ℓ_1 (Teng et al. [37]), ℓ_0 (Levine and Feizi [25]) , ℓ_∞ (Zhang et al. [43]) and even ℓ_p norm (Yang et al. [41], Dvijotham et al, [9]). Those mentioned methods were proven to find near-optimal certification regions in different norms, but the problem is that the certification was still very small. To resolve this problem Mohapatra et al. [28] in 2020 proposed higher-order certification with a method to calculate the certified

safety region using zeroth-order and first-order information for Gaussian-smoothed classifiers, but still did not provide a closed-form solution.

In 2019 Cohen et al. [5] paper gave a good start for the deep research in the field of randomized smoothing. While a lot of recent papers achieved different good certification results with various smoothing distributions, their parameters were previously always set as a global hyperparameter. Alfarra et al. [1] have changed it and in their work showed that the variance of the Gaussian distribution can be optimized at each input so as to maximize the certification radius for the construction of the smooth classifier. With such revisited technique they were able to achieve 9% and 6% improvement over the certified accuracy of the strongest baseline for a radius of 0.5 on CIFAR10 and ImageNet respectively. Later Eiras et al. [11] went even further and extend the isotropic randomized smoothing ℓ_1 and ℓ_2 certificates to their generalized anisotropic counterparts. The proposed framework called the ANCER achieves SOTA ℓ_1 and ℓ_2 performance on the CIFAR-10 and ImageNet utilizing the previous ideas of data dependant smoothing. The main idea is that previous approaches' certification regions were limited by the worst-case adversaries because of isotropic properties, but there can be other (potentially large) areas that can be discovered by anisotropic counterparts. However the described anisotropic case is still limited by the axis alignment and can extend only in a predefined set of direction, so in our thesis, we overcome this and propose an improved extended version that will not be aligned and can find larger safe regions in any direction.

Chapter 3

Theoretical Background On Randomized Smoothing

3.1 Randomized Smoothing

3.1.1 Smoothed Classifier

Assuming we have a usual classification setup. Let x be the input ($x \in R^d$) and y be the corresponding labels ($y \in \mathcal{Y} = 1, \ldots, k$). Our base classifier f is parameterized by θ, $f_\theta : \mathbb{R}^d \to \mathcal{P}(\mathcal{Y})$ where $\mathcal{P}(\mathcal{Y})$ is a probability simplex over k labels. And our goal is to construct a smoothed classifier g out of base classifier such that it will return the label which the base classifier is most likely to return when the input where perturbed by the Gaussian noise $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$:

$$g_\theta(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)} \left[f_\theta(x + \epsilon) \right] \tag{3.1}$$

In a practical setup, it is impossible to evaluate the exact prediction of the smoothed classifier, so as proposed in Cohen et al. [5] we will use the Monte Carlo algorithm for sampling the output which is by definition incorrect with some (small in our case) probability.

3.1.2 Robustness Guarantee

The key theorem for the proof of the robustness guarantee was presented and proved by Cohen et al. [5]:

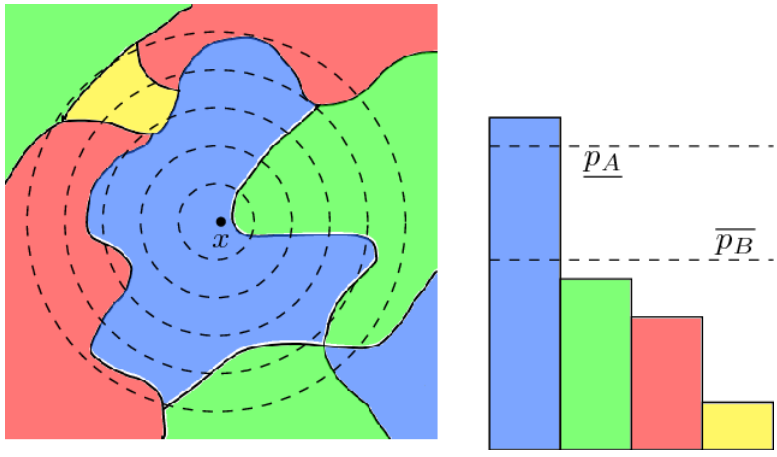


FIGURE 3.1: Evaluating the smoothed classifier at an input x. *Source:* [5]

Theorem 1 Source: *[5]. Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function, and let $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$. Let g be defined as in (3.1). Suppose $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy:*

$$\mathbb{P}\left(f(x+\varepsilon) = c_A\right) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(x+\varepsilon) = c) \tag{3.2}$$

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

$$R = \frac{\sigma}{2}\left(\Phi^{-1}\left(\underline{p_A}\right) - \Phi^{-1}\left(\overline{p_B}\right)\right) \tag{3.3}$$

Utilizing this theorem, suppose that the smoothed classifier g returned the prediction c_A with a probability p_A and the second top class c_A with probability p_B:

$$\mathbb{E}_\epsilon\left[f_\theta^{c_A}(x+\epsilon)\right] = p_A \geq p_B = \max_{c \neq c_A} \mathbb{E}_\epsilon\left[f_\theta^{c}(x+\epsilon)\right]. \tag{3.4}$$

The Φ^{-1} is the inverse of the standard Gaussian cumulative distribution function (CDF). A smoothed classifier g is robust in the area bounded by some radius $g(x+\delta) = g(x) \forall \|\delta\|_2 \leq R$ with R calculated by the formula mentioned in Theorem 1. In the original theorem lower bound for p_A and upper bound for p_B was used, meaning $\underline{P_A} \leq P_A$ and $\overline{P_B} \geq P_B$ because they are only approximations by Monte Carlo sampling, theoretically result still holds for both settings.

3.2 Data Dependent Smoothing

In previously mentioned approaches σ was set as a global hyperparameter of the smoothed classifier and played a huge role in the accuracy/robustness trade-off. One can argue that increasing σ will significantly increase the certification radius R and will be correct as it is in the numerator in Equation 3.3. But the main goal is to increase the certified accuracy for some σ instead of radius. For the classifier g and a similar to previous examples dataset with samples $x \in R^d$ and labels $y \in \mathcal{Y} = 1, ..., k$ and a radius R. Each ith sample will have the identifier if the prediction is not only robust but also correct, lets call this correctness identifier c_i and define it as follows:

$$c_i = \mathbf{1}\left[g\left(x_i + \delta\right) = y_i \forall \|\delta\|_2 < R\right]. \tag{3.5}$$

With such setting certified accuracy will be calculated $\frac{1}{dl}\sum_{i=1}^{dl} c_i$ where dl is the length of the dataset. And if the radius R will be huge enough the smoothed classifier will be more robust but at the same time may change the prediction causing a drop in the certified accuracy. This is what the accuracy/robustness trade-off means.

Alfarra et al. [1] analyzed the existing knowledge and utilized a logical property that the certification region R varies for different samples when σ is fixed as a global hyperparameter, so for a given f_θ different samples x can enjoy different optimal σ_x^* as a local optimal value. With this idea, they proposed a framework that optimizes a variance of Gaussian distribution σ at each sample to maximize the certification radius and increase certified accuracy (because previously some samples close to decision boundaries were wrongly classified with too big σ leading to drop in accuracy.

For a given smooth classifier with the initial σ_0, with a specific case where $\sigma_0 = 0$ leading to the base classifier f_θ in the initial setup the authors construct a new smoothed classifier with parameter σ_x^* unique for every sample which was picked

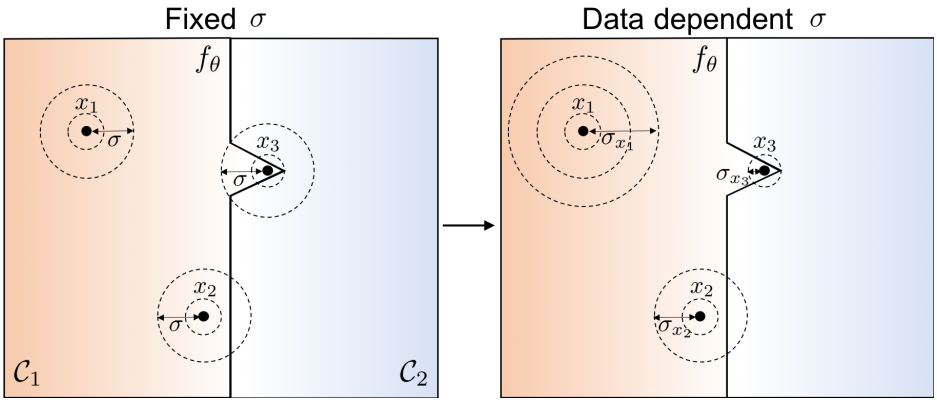


FIGURE 3.2: From fixed to data dependent smoothing. *Source:* [1]

in the manner to maximize the certification radius for that sample. With the initial σ_0 and $c_A = \arg\max_c \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma_0 I)}\left[f^c(x+\epsilon)\right]$ being the top prediction of $f_\theta(x)$, for every sample they apply an optimization procedure for σ directly optimizing the radius:

$$\sigma_x^* = \underset{\sigma}{\operatorname{argmax}} \frac{\sigma}{2}\left(\Phi^{-1}\left(\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}\left[f_\theta^{c_A}(x+\epsilon)\right]\right) - \Phi^{-1}\left(\max_{c\neq c_A}\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\sigma^2 I)}\left[f_\theta^{c}(x+\epsilon)\right]\right)\right). \tag{3.6}$$

While constructing the objective, they were inspired by Zhai et al. [42] who were also maximizing the certified radius (instead of setting it as a model hyperparameter) but as a global value for all samples. To optimize it we will also need Monte Carlo sampling for the expectation approximation and stochastic gradient ascent (SGA) as a solver. But as mentioned by the authors it suffers from high variance due to the dependence of the expectation on the optimization variable σ that parameterizes the smoothing distribution $\mathcal{N}\left(0,\sigma^2 I\right)$. Fortunately, Kingma and Welling [18] showed that a reparameterization of the variational lower bound yields a lower bound estimator that can be straightforwardly optimized using standard stochastic gradient methods. The key idea is to get rid of the optimization variable σ from the expectation (Monte Carlo sampling). For this we replace $\epsilon = \sigma\hat{\epsilon}$ where $\hat{\epsilon} \sim \mathcal{N}(0, I)$ (instead of $\mathcal{N}\left(0,\sigma^2 I\right)$) and the objective becomes:

$$\sigma_x^* = \underset{\sigma}{\operatorname{argmax}} \frac{\sigma}{2}\left(\Phi^{-1}\left(\mathbb{E}_{\hat{\epsilon}\sim\mathcal{N}(0,I)}\left[f_\theta^{c_A}(x+\sigma\hat{\epsilon})\right]\right) - \Phi^{-1}\left(\max_{c\neq c_A}\mathbb{E}_{\hat{\epsilon}\sim\mathcal{N}(0,I)}\left[f_\theta^{c}(x+\sigma\hat{\epsilon})\right]\right)\right). \tag{3.7}$$

With this setup, the gradient of the objective will have a lower variance than previously. The optimization is done iteratively with K steps of SGA, where K is a hyperparameter.

3.2.1 Memory-Based Certification

One of the serious drawbacks of the data dependent smoothing is caused by the variance of σ. It was constant for all samples in previous approaches and could be directly certified with a help of Monte Carlo sampling, as was proposed by Cohen et al. [5]. But with different sigmas the problem arises. For example, we have some random sample point x_i and $\mathcal{R}_i$ is the certification radius for it obtained with a smoothed classifier g_Θ and another point x_j with the corresponding radius $\mathcal{R}_j$.

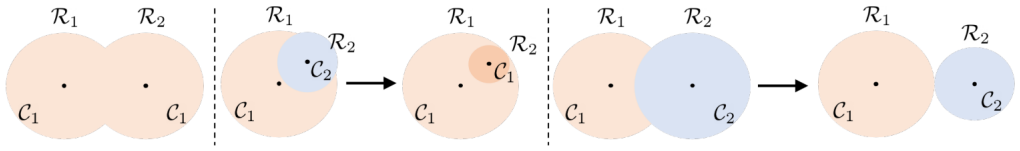


FIGURE 3.3: Memory-based certification of the data dependent classifier. *Source:* [1]

There is a possibility that point x_j is located inside the certification region, meaning $\|x_i - x_j\|_2 \leq \mathcal{R}_i$ and the data dependent smoothing does not take this into account. The problem occurs when g_Θ predicts x_j to belong to the class different from x_i and when the second belongs to the certification region of the first. This breaks the soundness of certification.

So in general the authors [1] propose a postprocessing step with a general memory where the reassigned regions will be added called memory-based certification with three possible scenarios when certified regions intersect. See Figure 3.3: (1) **Left:** Query sample point x_2 have the intersection of the corresponding certification region with the point x_1 already present in memory and the predictions are the same meaning $\|x_2 - x_1\|_2 \leq \mathcal{R}_1$ and $\mathcal{C}_1 = \mathcal{C}_2$. In this lucky case, we just add a new point to the memory. (2) **Middle:** A new sample is located inside the certification region of the point from memory and has a different class prediction. Here we will change the prediction of the query point x_2 to the $\mathcal{C}_1$ and will bound its certification region $\mathcal{R}_2$ to be the largest subset of the certification region $\mathcal{R}_1$, $\mathcal{R}_2 \subset \mathcal{R}_1$ and add it to memory. (3) **Right:** A query point is located outside the certification region of a point from memory, but their certification regions overlap and have different predictions. In this case, the query x_2 will preserve its original prediction $\mathcal{C}_2$, but the certification region $\mathcal{R}_2$ will be bounded to the largest region which does not intersect with $\mathcal{R}_1$. So with this setting, we will preserve initial results without intersection regions and change results with such problem resolving the issue with the soundness of certification.

3.3 ANCER

All previous approaches were built based on the knowledge that the nature of the certified region has to be isotropic. This fact was supported by the assumption that usual smoothing distributions are identically distributed. For example, Lee et al. [24] explicitly mentioned the addition of isotropic Gaussian noise to the input example. They were focusing on generalizing to the broader classes of the distributions (Uniform, Discrete) but still not mentioning the anisotropic counterparts. Sampling for example from ℓ_2 anisotropic distribution is not so hard but the interesting thing is how we can certify such regions. Eiras et al. [11] did an important contribution by providing an analysis on the way how we can certify the anisotropic regions characterized by any norm and holding prior art as a special case. The authors also introduced an evaluation framework to compare the methods that certify general regions and they proposed the framework called ANCER that utilizes Data Dependent Certification but in the anisotropic case.

3.3.1 Lipschitz Constant

We use the definition from the book by Houshang H. Sohrab called "Basic real analysis" [33] where Lipschitz function is defined as follows: Let $f : \mathbb{R}^n \to \mathbb{R}$. Then, we say that f is Lipschitz (or satisfies a Lipschitz condition) if there is a constant $L > 0$ such that

$$\left|f(x) - f\left(x'\right)\right| \leq L\|x - x'\| \quad \forall x, x' \in \mathbb{R}^n. \tag{3.8}$$

Geometrically, if $f : \mathbb{R} \to \mathbb{R}$ satisfies the Lipschitz condition then for any $x, x' \in I, x \neq x'$, the inequality

$$\frac{|f(x) - f(x')|}{\|x - x'\|} \leq L \tag{3.9}$$

indicates that the slope of the chord joining the points $(x, f(x))$ and $(x', f\left(x'\right))$ on the graph of f is bounded by L. Later, Jordan and Dimakis [16] published a paper related to the exact computation of the local Lipschitz constant, which is a very descriptive metric in robustness theory. And following those mentioned arguments, Eiras et al. [11] proposed a general certification analysis described as follows:

Proposition 1. ***Source:*** **[11]**. Consider a differentiable function $g : \mathbb{R}^n \to \mathbb{R}$. If $\sup \|_x\| g(x)\| \leq L$ where $\|\cdot\|_*$ has a dual norm $\|z\| = \max xz^\top x$ s.t. $\|x\| \leq 1$, then g is L-Lipschitz under norm $\|\cdot\|_*$, that is $|g(x) - g(y)| \leq L\|x - y\|$.

With a mentioned proposition the authors formalize $\|\cdot\|$ certification as follows:

Theorem 2 Source: *[11]. Let $g : \mathbb{R}^n \to \mathbb{R}^K, g^i$ be L-Lipschitz continuous under norm $\|\cdot\|_* \forall i \in \{1, \ldots, K\}$, and $c_A = \text{argmax}_i\, g^i(x)$. Then, we have arg $\max ig^i(x + \delta) = c_A$ for all δ satisfying:*

$$\|\delta\| \leq \frac{1}{2L}\left(g^{c_A}(x) - \max_c g^{c \neq c_A}(x)\right) \tag{3.10}$$

3.3.2 Certification

Now lets actually discuss how the certification is done for anisotropic regions. In previously mentioned Theorem 2 we have the exact definition of $\|\cdot\|$ norm of robustness certificates ready to use for any L-Lipschitz classifier under $\|\cdot\|_*$. This idea is a key part for the development of anisotropic certificates. For the example, consider the ellipsoid certification under ℓ_2^Σ norm defined as $\|\delta\|_{\Sigma,2} = \sqrt{\delta^\top \Sigma^{-1}\delta}$. It's dual norm is $\|\delta\|_{\Sigma^{-1},2}$. In general notation, assume that we have a vector space X with clearly defined norm $\|\cdot\|$, continuous linear function f and X^* is the dual space for X. The dual norm of f that belongs to X^* is defined as $\|f\| = \sup\{|f(x)| \quad : \|x\| \leq 1, x \in X\}$. In such setting, $\|\delta\|_{\Sigma,2} \leq r$ and $\|\delta\|_{\Sigma^{-1},2} \leq r$ define an ellipsoid where the smoothed classifier is defined as follows:

$$g_\Sigma(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\Sigma)}[f(x + \epsilon)]. \tag{3.11}$$

Utilizing the Theorem 2, the Lipschitz constant L is derived under $\|\cdot\|_{\Sigma^{-1},2}$ norm and $\Phi^{-1}\left(g_\Sigma(x)\right)$ is 1-Lipschitz meaning $L = 1$. For more details and proofs, see Section B in the supplementary materials of [11]. Then, the cetification for the anisotropic case follows naturally. Let $c_A = \text{argmax}_i g_\Sigma^i(x)$, then we have that $\text{argmax}_i g_\Sigma^i(x + \delta) = c_A$ for all δ satisfying:

$$\|\delta\|_{\Sigma,2} \leq \frac{1}{2}\left(\Phi^{-1}\left(g_{\Sigma}^{c_A}(x)\right) - \Phi^{-1}\left(\max_{c} g_{\Sigma}^{c\neq c_A}(x)\right)\right). \tag{3.12}$$

3.3.3 ANCER Objective

Similar to the previously mentioned data dependant objective, the key idea here is to maximize the volume of the certified region through the radius. Instead of calculating the radius directly, the authors proposed to find the proxy radius. To give the overview of the needed notations, lets take two certified regions: isotropic $\mathcal{R}_1$ and anisotropic $\mathcal{R}_2$ defined as:

$$\mathcal{R}_1 = \{\delta : \|\delta\|_2 \leq \tilde{\sigma} r_1\} \text{ and } \mathcal{R}_2 = \left\{\delta : \|\delta\|_{\Sigma,2} = \sqrt{\delta^\top \Sigma^{-1}\delta} \leq r_2\right\}, \tag{3.13}$$

where $r_1, r_2 > 0$. In 2D case the maximum enclosed circle in the ellipse is defined by the smaller of two radii. In larger dimensions we will have the maximum enclosed ℓ_2-ball defined as a minimum of all radii and in our specific case lets call it $\mathcal{R}_3$, where $\mathcal{R}_3 = \{\delta : \|\delta\|_2 \leq \min i\sigma_i r_2\}$ by the definition. This leads to $\mathcal{R}_2 \supseteq \mathcal{R}_3$ by construction. So if $\mathcal{R}_3 \supseteq \mathcal{R}_1$, then $\mathcal{R}_2 \supseteq \mathcal{R}_1$ and as authors called it, $\mathcal{R}_2$ is a superior certificate to the $\mathcal{R}_1$. With this setting proxy radius $\tilde{R}$ will be defined for $\mathcal{R}_2$ as:

$$\tilde{R} = r_2 \sqrt[n]{\prod_i^n \sigma_i}, \tag{3.14}$$

and maximizing $\tilde{R}$ will lead to maximizing the certification region $\mathcal{R}_2$ which we are interested in.

To define the objective we need to present some other terms. Let Θ be a parameter for a smoothing distribution defined as $\Theta = \text{diag}\left(\{\theta_i\}_{i=1}^n\right)$, an ℓ_p-norm $(p \in \{1,2\})$, and a gap value of $r^p \in \mathbb{R}^+$. Here Θ is an anisotropic analogue of σ in previous approaches. And the idea is to utilize the Data Dependant approach not to set it as a global parameter but to optimize per each input sample. The goal for the anisotropic ℓ_p case is to maximize the region $\left\{\delta : \|\delta\|_p \leq \theta^x r^p\left(x, \theta^x\right)\right\}$, which can be achieved by maximizing radius $\theta^x r^p\left(x, \theta^x\right)$ through $\theta^x \in \mathbb{R}^+$, obtaining r_{iso}^*, and $r^p\left(x, \Theta^x\right)$ is the gap value. So with the described parameters the optimization from ANCER framework was defined as follows:

$$\arg\max_{\Theta^x} r^p\left(x, \Theta^x\right) \sqrt[n]{\prod_i \theta_i^x} \quad \text{s.t. } \min_i \theta_i^x r^p\left(x, \Theta^x\right) \geq r_{\text{iso}}^*. \tag{3.15}$$

3.3.4 Memory-Based Certification

Algorithm 1 Memory-Based Certification. *Source:* [11]

Input: point x_{N+1}, certified region $\mathcal{R}_{N+1}$, prediction $\mathcal{C}_{N+1}$, and memory $\mathcal{M}$
Result: Prediction for x_{N+1} and certified region at x_{N+1} that does not intersect with any certified region in $\mathcal{M}$.
for $(x_i, \mathcal{C}_i, \mathcal{R}_i) \in \mathcal{M}$ **do**
  if $\mathcal{C}_{N+1} \neq \mathcal{C}_i$ **then**
    if $x_{N+1} \in \mathcal{R}_i$ **then**
      return ABSTAIN, 0
    else if *MaxIntersect*$(\mathcal{R}_{N+1}, \mathcal{R}_i)$ *and* *Intersect*$(\mathcal{R}_{N+1}, \mathcal{R}_i)$ **then**
      $\mathcal{R}'_{N+1}$ = `LargestOutSubset`$(\mathcal{R}_i, \mathcal{R}_{N+1})$ $\mathcal{R}_{N+1} \leftarrow \mathcal{R}'_{N+1}$
end
add $(x_{N+1}, \mathcal{C}_{N+1}, \mathcal{R}_{N+1})$ to $\mathcal{M}$ **return** $\mathcal{C}_{N+1}, \mathcal{R}_{N+1}$

In ANCER the memory-based certification procedure was modified as we are now using ellipsoids. For the computational reasons they first check if the maximum of the radii of query point x_2 intersects with a maximum of radii of point already present in the memory x_1 (`MaxIntersect` from Algorithm 1). If there is no intersection of this minimum containing ℓ_2 ball then everything is fine. Only if the intersection of balls is present, do we check the direct intersection of ellipsoids (`Intersect`) which is more expansive in computations. The authors admit that the check if there exists an intersection between ellipsoids is not trivial and they rely on the fundamental works of Ros et al. [31], and Gilitschenski and Hanebeck [12]. Having two diagonal matrices A and B with the corresponding diagonal values $\{\mathbf{A}_{ii}\}_{i=1}^{n}$ and $\{\mathbf{B}_{ii}\}_{i=1}^{n}$ being the optimized sigmas (different radii of the ellipsoids), we calculate $K(t)$ defined as:

$$K(t) = 1 - \sum_{i=1}^{n} (b_i - a_i)^2 \frac{t(1-t)\mathbf{A}_{ii}\mathbf{B}_{ii}}{t\mathbf{A}_{ii} + (1-t)\mathbf{B}_{ii}}. \tag{3.16}$$

With this setting, the `Intersect` function will return *False* if there exists $t \in (0, 1)$ such that $K(t) < 0$ meaning the ellipsoids do not intersect, and return *True* otherwise. For more details and proofs check Appendix D in [11], or directly [31], [12].

After finding the intersection we have to decrease the query point radius $\mathcal{R}_2$ (similar to the original procedure in Figure 3.3). Still, finding the maximum ellipsoid that does not intersect with certification region $\mathcal{R}_1$ of a point x_1 present in memory with a different prediction is a difficult problem, so the authors found the maximum enclosing ℓ_2 ball such that it does not intersect with the ellipsoid region from memory and utilizes the same procedure as in [1]. For this the authors formulated the problem of the projection of a vector $y = b - a$, where b is the sample already present in memory and a is a query point. . It should be projected to the ellipsoid with the same shape as $\mathcal{R}_A$. Such a problem can be solved with the help of optimization analog for matrix A defined previously:

$$\min_x \frac{1}{2}\|x - y\|_2^2 \quad \text{s.t.} \quad x^\top \mathbf{A} x \leq 1. \tag{3.17}$$

With the help of Lagrangian formulation and utilizing the fact that A and B are diagonal, the authors proposed a simplified scalar optimization problem (for more details check supplementary D.4 from [11]):

$$f(\lambda) = \sum_{i=1}^{n} \frac{y_i^2 \mathbf{A}_{ii}}{(1 + 2\lambda \mathbf{A}_{ii})^2} - 1 = 0. \tag{3.18}$$

The authors were optimizing for x^* and after obtaining it, the maximum radius of the ℓ_2 ball with a center in b will be:

$$r^* = \|(x^* + a) - b\|_2 - \epsilon, \tag{3.19}$$

where ϵ is a very small number.

Chapter 4

Proposed Solution

4.1 RANCER: Non-Axis Aligned Anisotropic Certification

In this chapter, we propose a non-axis aligned anisotropic certification extension for the previous SOTA method - ANCER. We dub our new approach *RANCER* where "R" refers to rotations. We will first provide a general theoretical intuition on how to extend ramdomized smoothing to a generalized non-axis aligned region for a given fixed orthogonal transformation. Then, we will proceed by showing how to choose the such an orthogonal transformation matrix towards larger certified regions. The detailed pipeline algorithm is proposed in Algorithm 2.
Recall, in the smoothed classifier in Equation 3.1, the Gaussian noise was sampled as $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$, where $\sigma^2 I$ is the diagonal matrix with σ controlling the strength of the noise. For ANCER as per Equation 3.11, the noised was then sampled from a more general anisotropic Gaussian distribution $\varepsilon \sim \mathcal{N}\left(0, \Sigma\right)$, where Σ was a general diagonal matrix. Towards general non-axis aligned anisotropic certification, we construct a smooth classifier that samples Gaussian noise from a general Gaussian distribution. To wit, unlike in [1] where the gaussian covariance is $\mathcal{A}_{DDS} = \sigma^2 I$ and where $\mathcal{A}_{ANCER} = \Sigma$ for ANCER [11], we consider a general Gaussian smoothing with dense covariance positive-definite $\mathcal{A}$. Note that a proper covariance matrix is symmetric and therefore can be orthogonally diagonalzied, i.e. $\mathcal{A} = U\Sigma^{'}U^{\top}$, where U and $\Sigma^{'}$ are the set eigenvectors and eigenvalues, respectively. Note that $\mathcal{A}_{DDS}$ and $\mathcal{A}_{ANCER}$ are special cases of $\mathcal{A}$ for when U is an identity matrix. To that end, we propose the following new smooth classifier:

$$g_{\mathcal{A}}(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\mathcal{A})}\left[f(x+\epsilon)\right]. \tag{4.1}$$

We propose the following reparameterization. Let $\epsilon^{'} = U\epsilon$ where U is the eigenvectors matrix of $\mathcal{A}$ and similarly to ANCER $\epsilon \sim \mathcal{N}\left(0, \Sigma\right)$, then we have that $\epsilon^{'} \sim \mathcal{N}\left(0, U\Sigma U^{\top}\right)$ which is exactly $\epsilon^{'} \sim \mathcal{N}\left(0, A\right)$. Therefore, the summarized reparameterization is summarized as:

$$\epsilon \sim \mathcal{N}\left(0, \Sigma\right) \quad \rightarrow \quad \epsilon^{'} = U\epsilon \quad \rightarrow \quad \epsilon^{'} \sim \mathcal{N}\left(0, U\Sigma U^{\top}\right). \tag{4.2}$$

With previously defined matrix $\mathcal{A}$ and those simple transformations presented in Equation 4.2, currently, the smoothed classifier can be written as in Equation 4.1 which is equivalent to:

$$g_{\Sigma}(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0,\Sigma)}\left[f_{\theta}(x+U\epsilon)\right]. \tag{4.3}$$

In this setting, the only thing different from the ANCER is the noise multiplication by U. For the case where the input is two-dimensional, $n = 2$ from $\mathbb{R}^n$, it will simply

Algorithm 2 Non-Axis Aligned Anisotropic Certification

Function `OptimizeRANCERSigmas`(f_θ, x, α, σ_0, n, *num_iters*, *LossFunction*, *clip_diff_min*, *clip_diff_max*):

Initialize: $\sigma_x^0 \leftarrow \sigma_0$;
$base\_classifier\_output = f_\theta(x)$;
$H = CalculateHessian(LossFunction, base\_classifier\_output)$;
$\Sigma', U = EigenDecomposition(H)$;
for $i = 0 \dots num\_iters$ **do**
  sample $\hat{\epsilon}_1, \dots \hat{\epsilon}_n \sim \mathcal{N}(0, \Sigma)$
  $\psi(\sigma_x^i) = \frac{1}{n} \sum_{j=1}^{n} f_\theta(x + \sigma_x^i(\mathbf{U}\hat{\epsilon}_j))$
  $E_A(\sigma_x^i) = \max_c \psi^c; y_A = \operatorname{argmax}_c \psi^c$;
  $E_B(\sigma_x^i) = \max_{c \neq y_A} \psi^c$
  $R(\sigma_x^i) = \frac{\sigma_x^i}{2} \left(\Phi^{-1}(E_A) - \Phi^{-1}(E_B)\right)$
  $\sigma_x^{i+1} \leftarrow \sigma_x^i + \alpha \nabla_{\sigma_x^i} R(\sigma_x^i)$
  $\sigma_x^{i+1} \leftarrow \min\left(\max\left(\sigma_x^{i+1}, \sigma_0(1 - clip\_diff\_min)\right), \sigma_0(1 + clip\_diff\_max)\right)$
end
$\sigma_x^* \leftarrow \sigma_x^{num\_iters}$
return σ_x^*

be the rotation matrix that rotates the ellipsoid (ℓ_2) making it now not axis-aligned. For higher dimensions instead, it will be just a transformation matrix for the noise ϵ.

While doing the multiplication by U we will get our optimized δ, where $\delta^\top A \delta \leq r$ (r is a radius). We rewrite is as $\delta^\top U \Sigma U^\top \delta \leq r$ and with $y = U^\top \delta : y^\top \Sigma y \leq r$. In this setting the ellipsoid will be axis aligned in thy y coordinates and to interpret it as axis-aligned in the original domain: $\delta = Uy$. Let's also define the new certification for a non-axis aligned setting with the analogy from Equation 3.12 utilizing the notions defined above:

$$\|\delta\|_{\Sigma,2} \leq \frac{1}{2}\left(\Phi^{-1}\left(\mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma)}\left[f_\theta^{c_A}(x + U\epsilon)\right]\right) - \Phi^{-1}\left(\max_{c \neq c_A} \mathbb{E}_{\epsilon\sim\mathcal{N}(0,\Sigma)}\left[f_\theta^{c}(x + U\epsilon)\right]\right)\right). \tag{4.4}$$

In the setting with two-dimensional input, we used the rotation matrix as previously mentioned U. To make things handcrafted by visual approximation we can hardcode the value of rotation angle θ and use it to construct the rotation matrix U. Multiplication by it results to the new version of the noise:

$$U\epsilon = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix} = \begin{bmatrix} \epsilon_1 \cos\theta - \epsilon_2 \sin\theta \\ \epsilon_1 \sin\theta + \epsilon_2 \cos\theta \end{bmatrix}. \tag{4.5}$$

Such an approach will not work in higher dimensions and for the automated algorithm, so the key idea is to construct the matrix U automatically in a way that will maximize the "safe" certification region.

4.2 Transformation Matrix Construction

The first logical idea was to add the U term directly to the optimization procedure but due to its high dimension (square of the original input image), such an approach adds a lot of overhead to the certification and does not guarantee to converge to

the good suboptimal result. Therefore we focused on the direct estimation of the transformation matrix.

In 2019 Moosavi-Dezfooli et al. [29] presented a paper about adversarial robustness via curvature regularization where they provided theoretical evidence that there exists a strong relation between large robustness and a small curvature. For this, they used a quadratic approximation of a loss function, denoted as a Taylor expansion. Let $\mathcal{L}$ be the function representing the loss of our classifier or a general NN model, this function can be locally approximated with a quadratic function:

$$\mathcal{L}(x+\epsilon) \approx \mathcal{L}(x) + \nabla\mathcal{L}(x)^\top \epsilon + \frac{1}{2}\epsilon^\top H \epsilon, \tag{4.6}$$

where ϵ is a very small nearby region of the query point x, $\nabla\mathcal{L}(x)$ are gradients of $\mathcal{L}$ with respect to x and H is a Hessian matrix of $\mathcal{L}$ at x. While gradients are the first-order derivatives, Hessian is a square symmetric matrix of second-order partial derivatives of a scalar function (loss function in our case) and if we have sample image x such that $x \in \mathcal{R}^d$ then the Hessian matrix will be denoted as:

$$H = \left(\frac{\partial^2 \mathcal{L}}{\partial x_i \partial x_j}\right) \in \mathbb{R}^{d \times d}. \tag{4.7}$$

In the Equation 4.7, $x_i, \ldots, x_d$ are the corresponding image pixels or for the 2D input case just two features. So in the case of the CIFAR-10 [20] dataset with image sizes of $32x32x3$ the shape of the Hessian will be $32x32x3x32x32x3$ and for the 2D case, it will be $2x2$ matrix.

Moosavi-Dezfooli et al. [29] in the original work introduced the term curvature profile, so in our work, we will also operate it. And it will be calculated as a set of eigenvalues of the Hessian matrix. The curvature profile is the direct way to analyze a curvature in a small neighborhood region near the sample point. Small eigenvalues indicate a small curvature of the decision boundary near point x. By the way, by the term *small* we mean the absolute values because negative values with large absolute values also describe high curvature. For example, small values tell us that the classifier is almost linear in a small local area and when they are exactly zero it will result in a flat surface. And the large eigenvalues have the opposite effect. This intuition will be beneficial for our framework and will be discussed later.

The mentioned work analyzed the curvature profile which indicates the metric of how curvature the neighborhood area is. But our main idea was to find the particular transformation matrix which in the 2D case was rotation. So we need to find the directions of safe regions, "rotate" our sampling distribution to that region and expand in that particular direction. Remember previously mentioned matrix factorization. We had matrix $\mathcal{A}$ and did eigendecomposition of it: $\mathcal{A} = U\Sigma'U^\top$. The key indicator of curvature profile was Σ' in this case, but the main idea for our solution is to use matrix U. With described Hessian notations we set matrix $\mathcal{A}$ to be the Hessian $\mathcal{A} = H$ and with this U will be a matrix of eigenvectors of H describing the directions of the safe regions in a local area corresponding to the decision boundaries.

With those ideas the pipeline of the non-axis aligned part of our solution is pretty straightforward, and is divided into three main steps: 1) compute the Hessian matrix of the loss corresponding to the input; 2) do eigendecomposition of Hessian and set matrix U to be eigenvectors of the H; 3) sample noise from the new non-axis aligned distribution. The third step does the new sampling in two steps, first, we sample the noise exactly the same as in ANCER, and then we just multiply it by U transforming or "rotating" it.

4.3 Safe Region Directions Magnitude

4.3.1 Use Isotropic Sigmas

In this setup, we used only Hessian eigenvectors to find the directions of the safe regions and then used isotropic sigmas as the initial values for the optimization similarly as was proposed in ANCER. Based on our experiments it was the best way to increase the certified accuracy.

4.3.2 Use Hessian Eigenvalues As Initial Values

Here we were also trying to use the eigenvalues of the Hessian as the initial values for the optimization instead of using isotropic ones. This idea was tested based on the prior knowledge of Hessian eigenvalues properties describing the curvature profile for the sample. By utilizing this idea we were able to totally omit the step of calculating data dependant isotropic sigmas and save a lot of time but the drawback of such an approach is oversmoothing. See more about it in the section *Postprocessing* even when initial eigenvalues were very large and optimization would try to decrease them, it would not be able to decrease it to a small enough optimal value due to the lack of iterations.

4.3.3 Use Hessian Eigenvalues As Final Values For Sigmas

And the last logical test was to get rid of the optimization at all. Instead of optimizing sigmas with ANCER's objective to maximize the volume of the certification region, we were using eigenvalues directly to decrease the computation time and let the pipeline be more simple and straightforward. With this setup, we would get rid of not only data dependent computation but also optimization. But the drawback is also oversmoothing. In the previous setup, we had a chance that optimization will overcome it for some samples, but now we have not.

4.4 Postprocessing

4.4.1 Motivation

Similar to ANCER our main setup used the isotropic sigmas as the initial value for new non-axis aligned anisotropic sigmas computation. But in our research, we also discovered a problem of ANCER that sometimes it suffered from oversmoothing issue. This could happen when in the local neighborhood the safe direction seemed to have a large magnitude but actually, it was exceeding the bounds of the safe region and started sampling points from different classes leading to a large certification radius in that particular direction but in the same time the wrong prediction as we can potentially be unlucky and start sampling more points from the wrong class (see Figure 5.8). The original data dependant isotropic solution didn't have such an issue as it was bounded by the worst-case direction that didn't let it increase in the other direction and sample the wrong point.

So based on the trust in the isotropic sigmas but still preserving the anisotropic directions we introduce a postprocessing step that bounds the new non-axis aligned anisotropic sigmas to be not larger than some fixed percentage value of isotropic ones. Based on our experiments this step overcomes the issue of oversmoothing and still increases the certification region as it will definitely be not worth (or usually much better) than data dependent results. And with this setup, we slightly

decreased the potential of certification region volume but significantly increased the correctness of smoothed classifier prediction without the problem of oversmoothing.

4.4.2 Sigma Clipping During Optimization

Based on the experiments with getting the magnitude of the directions of the smoothing region, our main algorithm was chosen to be the data dependant isotropic values as initialization for optimization. With such a setup, we are able to utilize those values to do clipping based on them. We set two hyperparameters λ_1 and λ_2 to be the minimum and maximum percentage difference between the isotropic DDS result and our newly obtained σ. With the described setting the proposed procedure can be written as:

$$\sigma_{final} = \min\left(\max\left(\sigma_{optimized}, \sigma_{isotropic} * (1 - \lambda_1)\right), \sigma_{isotropic} * (1 + \lambda_2)\right), \tag{4.8}$$

where $\sigma_{optimized}$ is a value of σ after or during optimization procedure, $\sigma_{isotropic}$ is the value from DDS isotropic result. Another setting was to omit sigma clipping during the optimization and just do it in the final step before the return, but this approach gave worse certification results.

4.5 Certification

In our solution, we use exactly the same certification procedure as ANCER's. All the theorems and characteristics are preserved. The only difference is the noise sampling procedure. Recall Equation 3.12 which has a similar to original DDS structure with origins from Theorem 1, but smoothed classifier $g_\Sigma(x)$ is defined in a different way because of changed noise sampling distribution from $\epsilon \sim \mathcal{N}\left(0, \sigma^2 I\right)$ to $\epsilon \sim \mathcal{N}(0, \Sigma)$. And for our case it is the only thing we have changed to be $\epsilon \sim \mathcal{N}\left(0, A\right)$ (still in preserved format) and obtained the certification defined in Equation 4.4. Generally, changes were made only for noise sampling and certification remained exactly the same.

4.6 Memory-Based Certification

For RANCER the logic of memory based certification procedure is also exactly the same as in Algorithm 1 but the hard part is the intersection calculation in `Intersect` function. As was mentioned previously the test on ellipsoids intersection is a computationally hard problem. In ANCER the authors were able to simplify the calculations significantly by taking into account only diagonal values of A and B from Equation 3.16. We cannot afford the same thing as we are working with a general version of rotated ellipsoids and we should calculate $K(t)$ in a general way with possible identical forms:

$$\begin{aligned}
&(1) \quad K(t) = 1 - ta^\top \mathbf{A}a - (1-t)b^\top \mathbf{B}b + m^\top \mathbf{E}_t m \\
&(2) \quad K(t) = 1 - t(1-t)(b-a)^\top \mathbf{B}\mathbf{E}_t^{-1}\mathbf{A}(b-a) \\
&(3) \quad K(t) = 1 - (b-a)^\top \left(\tfrac{1}{1-t}\mathbf{B}^{-1} + \tfrac{1}{t}\mathbf{A}^{-1}\right)^{-1}(b-a)
\end{aligned} \tag{4.9}$$

In our implementation we used Equation 4.9 (3). Which was proved to be convex in the $t \in (0, 1)$ domain by [31] and [12]. Matrices A and B will also have a different form and not be diagonal any longer. As previously we will use optimized sigmas in diagonal matrix Σ, but will apply the found eigenvectors of Hessian to it. And in this

way $\mathbf{A} = U^\top \Sigma U$ where U are the eigenvectors of H. Another change was applied to *LargestOutSubset* from Algorithm 1 where we are now working with generalized ellipsoids forms, so diagonal simplification is not reasonable in our case (as was used in Equation 3.18). In this way the problem is defined as:

$$f(\lambda) = y^\top (2\lambda \mathbf{A} + I)^{-\top} \mathbf{A} (2\lambda \mathbf{A} + I)^{-1} y - 1 = 0, \tag{4.10}$$

and has a more computationally expensive optimization but solves the projection exactly the same as previously.

Chapter 5

Experiments and Results

In this chapter, we describe the datasets, evaluation metrics, and setup which was used to study the performance of the proposed algorithm.

5.1 Datasets

5.1.1 Intuition-Gathering 2D Dataset

"Only those who have patience to do simple things perfectly ever acquire the skill to do difficult things easily."

James J. Corbett

The best way to understand the performance of previous and proposed approaches is to analyze it in the simplest case in 2D. So for this, we proposed a bunch of different toy datasets so-called "intuition-gathering" because they helped a lot to understand the algorithm's behavior in different edge cases. To collect such data we used a visual creation tool drawdata[1] and its Jupyter extension [2] and just painted the interesting distributions of 2D data with positional coordinates features x and y and binary or 3-class classification labels 0 or 1, and possibly 2. Different variants include donut-shaped, axis-aligned, and non axis-aligned linearly-separable, and many other classes as you can see in the Figure 5.1 with 2 classes and 5.2 with 3 classes. In total, we have 30 different versions with 300 samples (binary) and 1200 samples (3 classes) each.

5.1.2 CIFAR-10

In the related literature overview, we mentioned that an important factor in the success of the defenses against adversarial attacks is the scalability of the method to deep models and large datasets. So it is critical to report algorithm performance on the more complex data and for this we used the CIFAR-10 dataset presented by Krizhevsky [20]. It consists of 60 thousand square colored images 32 by 32 pixels each. They are generally divided into 10 classes containing animals (dog, horse, etc.) and man-made transportation (car, ship, etc.), check Figure 5.3. The dataset is already divided into train and test split containing 50 thousand (split by 5 batches) and 10 thousand samples respectively. Another important thing is that all classes are totally mutually exclusive and as mentioned by the authors there is no overlap between them, for example, automobile and truck. The automobile contains sedans, and SUVs while the truck has only big trucks. There is no something in the middle

[1] https://drawdata.xyz/
[2] https://pypi.org/project/drawdata/

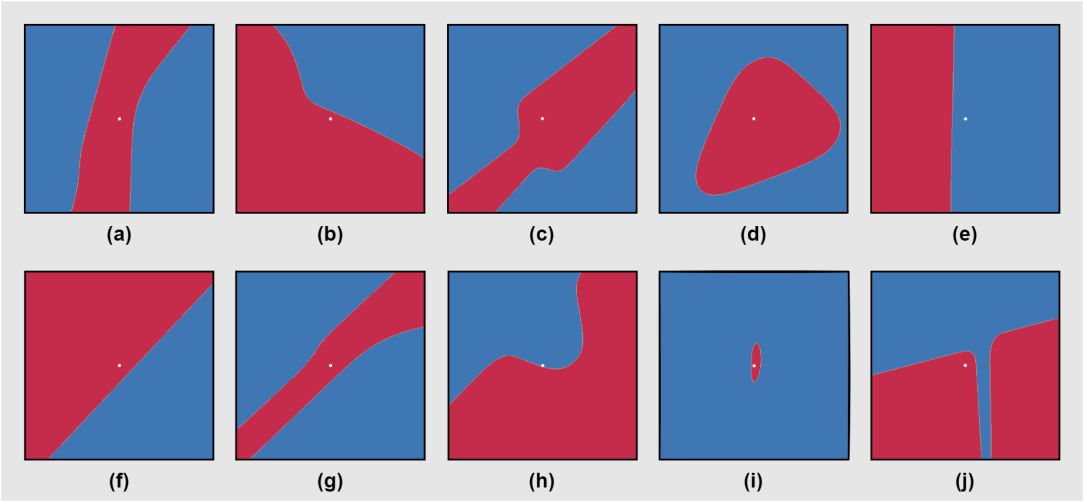


FIGURE 5.1: Examples of different versions of binary 2D Toy Dataset decision boundaries

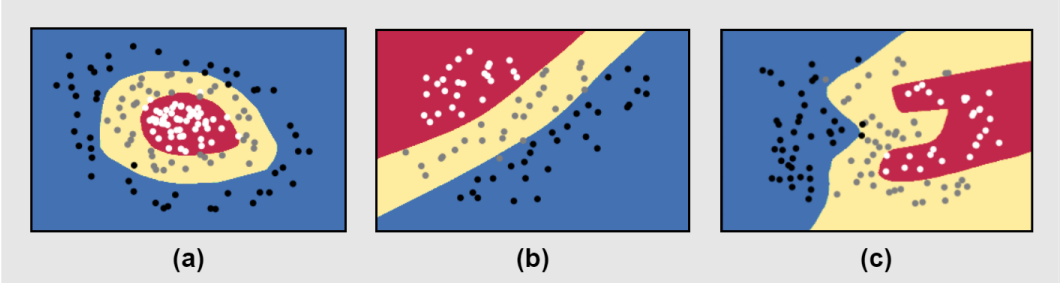


FIGURE 5.2: Examples of different versions of 2D Toy Dataset with 3 classes (decision boundaries and sample points)

like pickup trucks. This is an important fact for the adversarial robustness as we have prior knowledge that classes should not overlap.

5.2 Evaluation Metrics

The evaluation procedure for robust classifiers was crystallized in the pioneering works such as Cohen et al. [5], Salman et al. [32] and Zhai et al. [42]. It was proposed to use a pretrained network (usual training without adversarial) therefore we use ResNet-18 [14] as a baseline classifier for CIFAR-10 and custom small network for toy dataset. Our proposed generalized framework can be applied to different (ℓ_1, ℓ_2, ℓ_p) norms but in the proposed thesis we concentrated only on ℓ_2 certification. In that way our main metric is ℓ_2 certified accuracy. Usually for some fixed values of radius R and proxy radius $\tilde{R}$ we compute a certified accuracy as the portion of dataset samples which were correctly classified by the smoothed classifier and in the same time had an ℓ_2 certification radius bigger than R or $\tilde{R}$. We follow the definition of proxy radius from ANCER. With the Definition 1 from [11] describing the superior certificates and the formula to calculate the volume of certified region: $\mathcal{V}(\mathcal{R}) = r^n \sqrt{\pi^n} / \Gamma\left(n/2 + 1\right) \prod_{i=1}^{n} \sigma_i$ [17], the authors directly defined proxy radius for certification the same as in Equation 3.14 for given r. This holds because for larger ellipsoids volumes we will obtain larger $\tilde{R}$. In that case $\tilde{R}$ can be assumed as a generalization to the certified radius R.

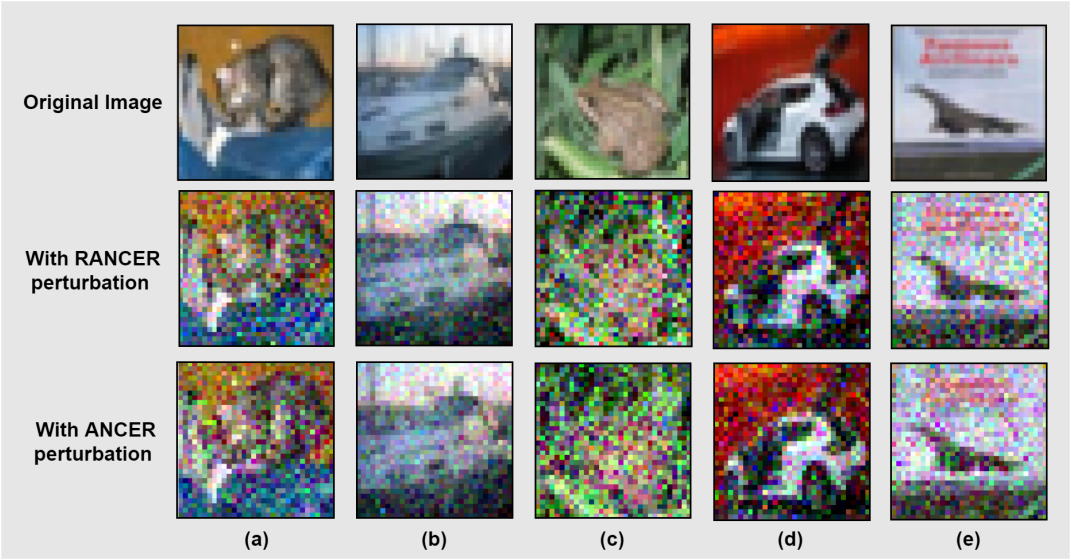


FIGURE 5.3: Examples of dataset images perturbed with random noise parameterized by optimal sigmas for RANCER and ANCER.

The other two important metrics are the average certified radius (ACR) and average certified proxy radius ($AC\tilde{R}$) defined as following:

$$\mathbb{E}_{x,y\sim\mathcal{D}_t}\left[R_x\mathbb{1}(g(x)=y)\right] \text{ and } \mathbb{E}_{x,y\sim\mathcal{D}_t}\left[\tilde{R}_x\mathbb{1}(g(x)=y)\right], \tag{5.1}$$

where $\mathcal{D}_t$ is a test set R_x, $\tilde{R}_x$ are the corresponding radius and proxy radius at sample point x with corresponding ground truth label y and $\mathbb{1}$ is the indicator function.

For our specific method, we also report two more metrics. The first is called average radius improvement (ARI) and indicates how much the RANCER certification radius improved from the previous SOTA approach and another called average proxy radius improvement ($A\tilde{R}I$) indicates the proxy radius improvement. The proposed additional metrics will be also calculated based only on the correctly predicted samples.

5.3 Certified Architectures

In our research, we concentrated on two model architectures. For the case with the toy dataset, we trained our own shallow classifier and for the CIFAR-10 dataset we used pre-trained ResNet-18 architecture.

5.3.1 ResNet-18

In 2015, He et al. [14] provided an analysis of the difficulties of training deep NNs and proposed a residual learning framework to ease the training of networks that are substantially deeper than those used previously. In the original work, the authors performed training on CIFAR-10. The inputs are 32x32 images which are passed to the 3x3 convolutions in the first layer followed by a bunch of 6n layers with the same convolutions performed on the feature maps with size 32,16,8 with 2n layers for each size of the feature map. convolutions with stride 2 were used for the subsampling and we have global average pooling, a 10-classes FC layer and a $Softmax$ activation function in the end. Exactly the same architecture (Figure 5.4) was used in our work with the original pre-trained weights.

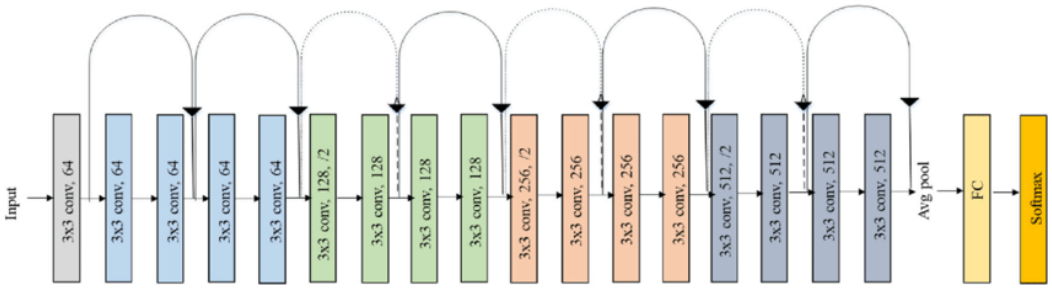


FIGURE 5.4: ResNet-18 architecture. *Source:* [14]

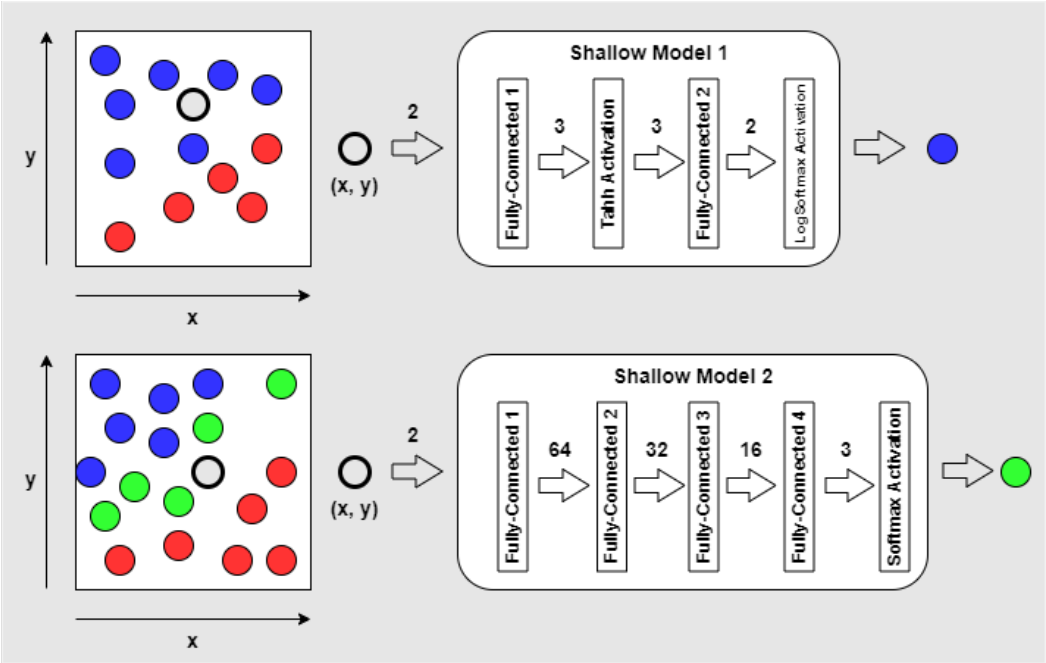


FIGURE 5.5: Custom shallow classifiers architectures

5.3.2 Custom Shallow Classifiers Architectures

For the proof of concept, we used previously described versions of toy datasets with 2 or 3 possible classes. Consequently, we trained two shallow architectures for them, described in Figure 5.5. For the binary classification task, the simple classifier with two hidden layers was used. We used the $LogSoftmax$ activation function and trained it with the $NLLLoss$ criterion and stochastic gradient descent optimizer. For the harder dataset, we used four hidden layers with more neurons, $Softmax$ activation function and $Adam$ optimizer as can be seen in the figure. This helped us to practically test two simple models with different activation functions, criteria, and optimizers.

5.4 Safe Directions Approximation

In our experiments, we calculate the Hessian of the loss function corresponding to the inputs following the similar setup as proposed in [29]. For the 2D case, we conducted different experiments on how good the safe approximation is based on the difference between the original and second-order approximation of the loss function.

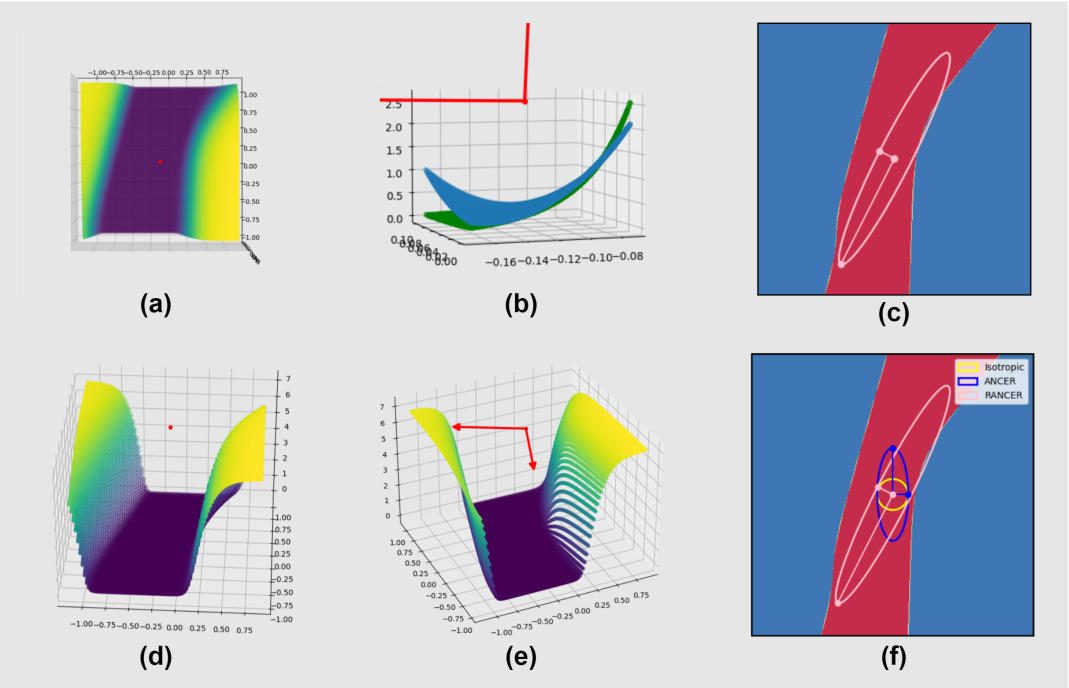


FIGURE 5.6: Safe directions approximation via Hessian eigenvectors

An example of safe directions approximation can be seen in Figure 5.6. A sample point with a "red" class was used here, it is located in the middle of the certification ellipsoid on (c) and is marked as a red dot in (a). The violet "valley" (a, b, d, e) represents the points where the loss function value is small for this class, while yellow points correspond to higher values of loss (standard "viridis" colormap from Matplotlib Python library [3]). On (b) and (e) we can see the safe directions represented as red arrows. It seems that the violet region is flat and cannot be representative of local curvature, but in reality, there is a small curvature, which is better visible in a small neighborhood area presented in (b).

The main idea is to have a representative (can be very small) curvature in a small region near the sample point. The approximation correctness is not critical here, because for example in the case with a flat region (far away from the decision boundary) we will have a close to perfect estimation of the plate, but in that case, Hessian safe directions will not be useful.

5.5 Toy Dataset Certification Results

To calculate our main set of previously described evaluation metrics we are using already mentioned Toy Datasets from Figures 5.1 and 5.2. Based on them we are conducting the experiments in exactly the same setup as in [1] and [11]. The comparison of fixed and data dependent ℓ_2 isotropic balls, ℓ_2^Σ anisotropic ellipsoids and rotated anisotropic ellipsoids certificates is done with a Gaussian smoothing framework with the custom (same for all methods) networks. In Figure 5.7 we report the distribution of certified accuracy of the proposed approach compared to previous SOTA methods (a, b) - without memory-based certification, (c, d) - with. The corresponding tabular results are presented in Table 5.1 and 5.2 (for proxy radius).

[3] https://matplotlib.org/

We conducted these experiments on all proposed toy dataset versions but in the thesis, we report metrics only for the dataset from Figure 5.2 (a) as it is the fairest for all approaches. Compared to other cases, for example, we have extremely beneficial (g) and worst-performance (e) datasets from Figure 5.1. The first has constant rotation for decision boundary and the other has no rotation. As the main goal of toy experiments was to test different hypotheses and gain intuition, we will not report the metrics for all other versions as it will be the overhead for the thesis. The proposed version, according to experiments, is the most representative.

Based on the conducted experiments we analyzed that fixed σ has significantly better certified accuracy results with memory-based certification. We must admit that the results in Figure 5.7 (a) and (b) are not representative, because they were obtained without memory-based certification. However, the problem of a huge difference in results with and without memory-based certification is present only for the low dimensional data because there is a high chance of different samples regions' intersection. In practice, with images represented as high-dimensional vectors, this problem neglects, so for CIFAR-10 experiments the results have to be much better.

The main goal of our setup was to improve the performance of ANCER and as reported in Tables 5.1 and 5.2 the proposed solution achieves better results than the previous best method - ANCER for all tested radii.

RANCER Run Clarification

As was mentioned in the analysis of the safe directions, RANCER method depends on the curvature profile of the loss function, when there is no curvature or it increases very fast (faster than quadratic), safe directions approximation may fail, leading to worse than ANCER performance. To mitigate this issue we run both approaches and return better results. In this manner, RANCER performance will be at least as good as ANCER and in Figure 5.7 we can see that it always has better or the same result as ANCER.

Method	$r = 0.1$	$r = 0.3$	$r = 0.5$
Fixed	**0.96**	**0.90**	0.00
Isotropic	0.33	0.33	**0.32**
ANCER	0.18	0.12	0.04
RANCER	0.27	0.17	0.06

TABLE 5.1: Comparison of top-1 certified accuracy at different ℓ_2 radii

Method	$r = 0.1$	$r = 0.3$	$r = 0.5$
Fixed	**0.96**	**0.90**	0.00
Isotropic	0.32	0.30	**0.26**
ANCER	0.18	0.13	0.07
RANCER	0.27	0.19	0.11

TABLE 5.2: Comparison of top-1 certified accuracy at different ℓ_2 proxy radii

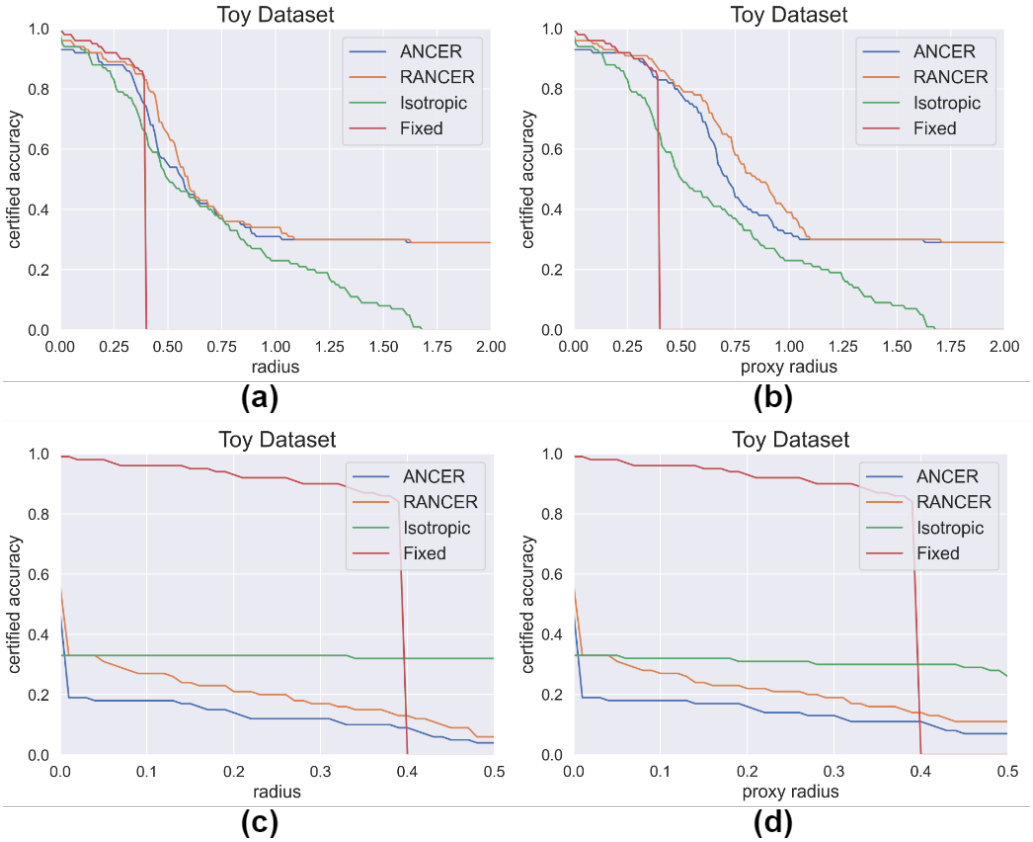


FIGURE 5.7: Distribution of top-1 certified accuracy as a function of ℓ_2 (a,c) radius, (b,d) proxy radius obtained with different approaches (a, b) without and (c, d) with memory based certification.

5.5.1 Optimization Convergence

In Figure 5.8 we present the optimization convergence analysis for four different examples from the same dataset. (a) final RANCER certification region is close to ANCER, (b, c) ANCER result has bigger certification region, (d) RANCER has bigger region.

For the data dependent isotropic method [1] the goal is to maximize the isotropic ℓ_2 region via the gap. It is calculated as R in the formula from Theorem 1. At the same time, ANCER's objective is to maximize the proxy volume of the certified region which can be done by maximizing the proxy radius. In the Figure 5.8 we can see the changes in the three mentioned terms through the iteration procedure. Sigma volume is computed as a product of all sigmas of certification ellipsoid and proxy radius is the sigma volume multiplied by a gap. The optimizer tries to maximize that radius by optimizing sigmas.

In the shown examples we intentionally included (b) and (c) to note the problem of oversmoothing where ANCER included the red region inside the certification region for the blue point. The figures ideally describe why the result of memory-based certificates is much worse than without this procedure. If we sample this blue point first, then all the points in the red region will have to change the prediction to the blue (by the design of memory-based certification) leading to a significant accuracy drop. RANCER is less vulnerable to this (c) due to the postprocessing with clipping, but it still oversmooths the regions (b) and has also dropped in certified accuracy performance.

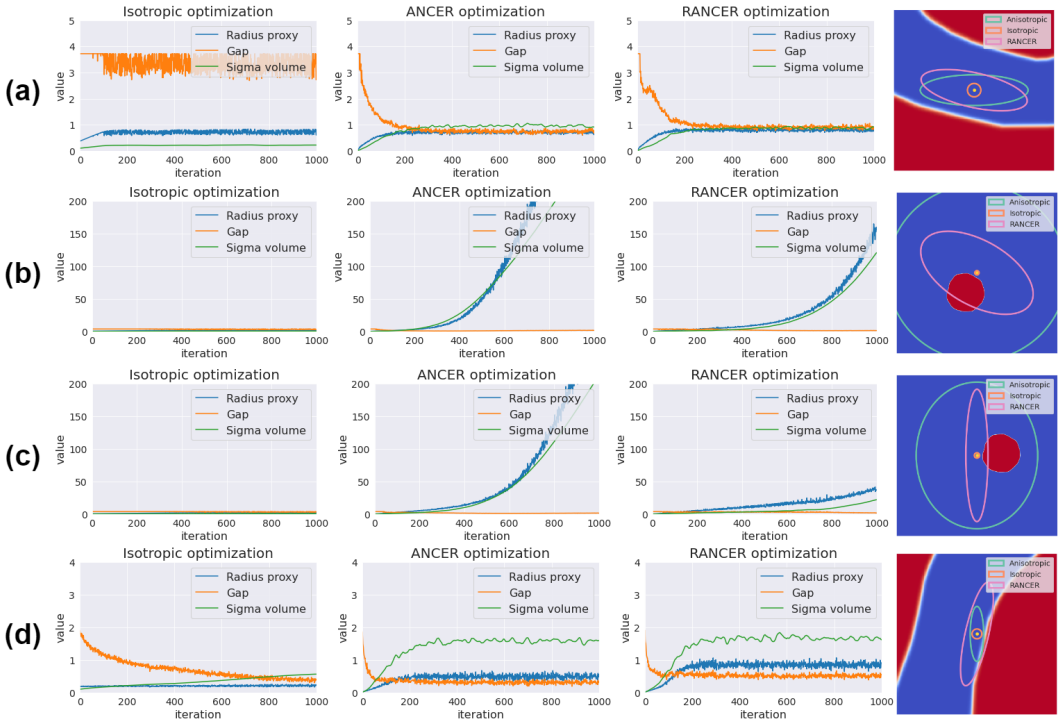


FIGURE 5.8: Optimization convergence analysis for the compared approaches based on 4 examples.

5.6 CIFAR-10 Certification Results

The experimental setup for CIFAR-10 experiments is exactly the same as described previously. We report the final certified accuracy metrics (after memory-based certification) in Table 5.3 and 5.4 for proxy radius. The corresponding visual comparison is present in Figure 5.9. The important fact is that we used 1000 uniformly distributed random samples from CIFAR-10 and obtained the same results as in [11] where the experiments were performed on 5000 samples. The reported results follow the same logic as previously stated by [11].

The newly proposed metrics ARI and $A\tilde{R}I$ create the additional value as we are now able to analyze the exact improvement in terms not only of the certified accuracy but also of the certification region itself. With the conducted experiments we obtained $ARI = 0.0517$ and $A\tilde{R}I = 0.0395$ compared to ANCER. Thus, on average RANCER has bigger certification radii letting it increase the certified accuracy.

Finally, based on the experiments and intuition gained from toy datasets we were able to perform certification for CIFAR-10. As expected, RANCER has better results in terms of all proposed metrics (see also Table 5.5 for ACR and $A\tilde{C}R$) and outperforms previous SOTA in certified accuracy for CIFAR-10.

Method	$r = 0.25$	$r = 0.5$	$r = 0.75$	$r = 1.0$	$r = 1.25$	$r = 1.5$
Fixed	0.55	0.00	0.00	0.00	0.00	0.00
DDS	0.39	0.18	0.08	0.02	0.00	0.00
ANCER	0.77	0.45	0.24	0.08	**0.00**	**0.00**
RANCER	**0.81**	**0.49**	**0.27**	**0.10**	0.00	0.00

TABLE 5.3: Comparison of top-1 certified accuracy at different ℓ_2 radii on CIFAR-10 dataset

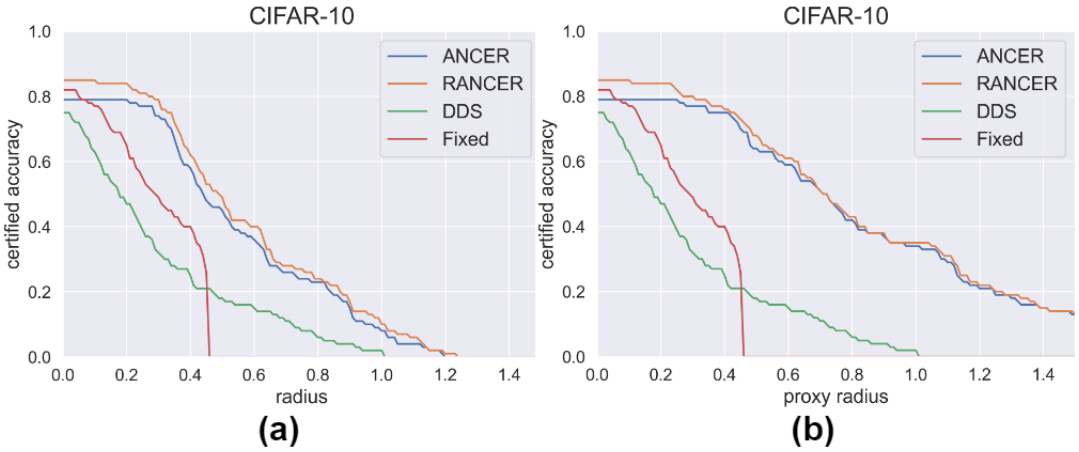


FIGURE 5.9: Distribution of top-1 certified accuracy as a function of ℓ_2 (a) radius, (b) proxy radius obtained with different approaches

Method	$r = 0.25$	$r = 0.5$	$r = 0.75$	$r = 1.0$	$r = 1.25$	$r = 1.5$
Fixed	0.55	0.00	0.00	0.00	0.00	0.00
DDS	0.39	0.18	0.08	0.02	0.00	0.00
ANCER	0.79	0.64	0.46	0.34	0.19	**0.13**
RANCER	**0.82**	**0.68**	**0.47**	**0.35**	**0.20**	0.13

TABLE 5.4: Comparison of top-1 certified accuracy at different ℓ_2 proxy radii on CIFAR-10 dataset

Method	ACR	$AC\tilde{R}$
Fixed	0.27	0.27
DDS	0.26	0.26
ANCER	0.48	0.76
RANCER	**0.52**	**0.79**

TABLE 5.5: Comparison of ACR and $AC\tilde{R}$ on CIFAR-10 dataset for different methods

Fixed	DDS	ANCER	RANCER
4.2s	4.9s	7.65s	19.1s

TABLE 5.6: Certification time comparison per sample for each method.

5.7 Computational Resources And Runtime Analysis

We used NVIDIA GeForce RTX 2080 Super with Max-Q Design for our experiments. It has a total memory of 8 Gb and 48 multiprocessors. Previous SOTA ([1], [11]) approaches have significantly improved certified accuracy but with a cost of runtime. Those methods are based on sample-wise optimization which requires additional time complexity compared to older approaches with fixed σ. Please check the time comparison in Table 5.6. The main complexity of RANCER optimization comes from the hard procedure of safe directions calculation as it involves calculating Hessian

and performing eigendecomposition. Another overhead is caused due to a more general check of ellipsoids intersection compared to the simplified version in ANCER. We believe that improved certification results are more significant than runtime performance but we will still work on the ways to optimize it.

Chapter 6

Conclusion

6.1 Results Summary

In our bachelor thesis, we presented the deep analysis of certified adversarial robustness for the shallow and large-scale neural networks. With the help of already acquired theoretical foundations of the world's most promising works in the relevant field, we succeeded in providing the theoretical extension to the anisotropic data dependant randomized smoothing and presenting its generalized counterpart - non-axis aligned anisotropic certification framework. With the obtained results we provide a practical clue of the performance of our solution. It was experimentally proven that our generalized framework outperforms recent works and shows better results in the l_2 certified accuracy for the academic CIFAR-10 dataset. Thereby, all the goals which were set for this thesis were achieved at the same time giving the vast research opportunity for future work.

6.2 Algorithm Limitations

As was already mentioned, the main drawback of data dependent certification is the variance of σ which breaks the soundness of certification. The original solution to this problem (memory-based certification) was proposed in [1] and later modified in [11]. But such a solution raises a new problem - memory and runtime complexity by its definition. In our framework, the complexity increases even more in three bottleneck places. Firstly. we were able to remove the transformation matrix optimization procedure and replace it with the straightforward hessian eigenvectors computation but it is still a computationally hard procedure. Secondly, the runtime of some particular functions was increased for example checking the intersection of rotated ellipsoids. And finally, we need to store the transformation matrices for certification. For CIFAR-10 with $32x32x3$ image sizes, the transformation matrix will have a size of $32x32x3x32x32x3$. There are some potential improvement methods to reduce the memory consumption and speed up the process (for example k-d trees) but they were not in the scope of this thesis.

Notwithstanding these limitations, we believe that our approach will be beneficial for the safety-critical applications where a high robustness guarantee is needed and inference time complexity is not critical.

6.3 Future Improvements

The theoretical part of our thesis has proved that we can obtain a more robust classifier by utilizing the ideas of non-axis aligned anisotropic certification. At the same

time there are two more directions of further research that are being tested by us and can be extended by other researchers:

Theoretical Improvements

This direction includes providing the certification regions for different norms, for example l_1 or l_p. And another important potential theoretical improvement is an automated selection of the samples for which the Hessian safe directions are definitely well-approximated.

Practical Improvements

Here the researchers can aim to optimize the time and memory complexity of the algorithm to be more suitable for the inference time dependant applications. For this, one can use previously mentioned k-d trees or propose some new ideas. And another important task is to discover a faster way to find the safe directions for example do not compute the Hessian directly but find its approximation.

Bibliography

[1] Motasem Alfarra et al. *Data-Dependent Randomized Smoothing*. 2022. URL: https://openreview.net/forum?id=ZFIT_sGjPJ.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. *Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples*. 2018. DOI: 10.48550/ARXIV.1802.00420. URL: https://arxiv.org/abs/1802.00420.

[3] Marco Barreno et al. "Can Machine Learning Be Secure?" In: *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*. ASIACCS '06. Taipei, Taiwan: Association for Computing Machinery, 2006, 16–25. ISBN: 1595932720. DOI: 10.1145/1128817.1128824. URL: https://doi.org/10.1145/1128817.1128824.

[4] Nicholas Carlini and David Wagner. *Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods*. 2017. DOI: 10.48550/ARXIV.1705.07263. URL: https://arxiv.org/abs/1705.07263.

[5] Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. *Certified Adversarial Robustness via Randomized Smoothing*. 2019. DOI: 10.48550/ARXIV.1902.02918. URL: https://arxiv.org/abs/1902.02918.

[6] Jeremy M Cohen, Elan Rosenfeld, and J. Zico Kolter. *Certified Adversarial Robustness via Randomized Smoothing*. 2019. DOI: 10.48550/ARXIV.1902.02918. URL: https://arxiv.org/abs/1902.02918.

[7] Jia Deng et al. "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[8] Yinpeng Dong et al. *Boosting Adversarial Attacks with Momentum*. 2017. DOI: 10.48550/ARXIV.1710.06081. URL: https://arxiv.org/abs/1710.06081.

[9] Krishnamurthy (Dj) Dvijotham et al. "A FRAMEWORK FOR ROBUSTNESS CERTIFICATION OF SMOOTHED CLASSIFIERS USING F-DIVERGENCES". In: *International Conference on Learning Representations*. 2020. URL: https://openreview.net/forum?id=SJlKrkSFPH.

[10] Ruediger Ehlers. *Formal Verification of Piece-Wise Linear Feed-Forward Neural Networks*. 2017. DOI: 10.48550/ARXIV.1705.01320. URL: https://arxiv.org/abs/1705.01320.

[11] Francisco Eiras et al. *ANCER: Anisotropic Certification via Sample-wise Volume Maximization*. 2022. URL: https://openreview.net/forum?id=UFYYol-bRq.

[12] Igor Gilitschenski and Uwe D. Hanebeck. "A robust computational test for overlap of two arbitrary-dimensional ellipsoids in fault-detection of Kalman filters". In: *2012 15th International Conference on Information Fusion*. 2012, pp. 396–401.

[13] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. 2014. DOI: 10.48550/ARXIV.1412.6572. URL: https://arxiv.org/abs/1412.6572.

[14] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. DOI: 10.48550/ARXIV.1512.03385. URL: https://arxiv.org/abs/1512.03385.

[15] Matthias Hein and Maksym Andriushchenko. *Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation*. 2017. DOI: 10.48550/ARXIV.1705.08475. URL: https://arxiv.org/abs/1705.08475.

[16] Matt Jordan and Alexandros G. Dimakis. *Exactly Computing the Local Lipschitz Constant of ReLU Networks*. 2020. DOI: 10.48550/ARXIV.2003.01219. URL: https://arxiv.org/abs/2003.01219.

[17] M.G. Kendall. *A Course in the Geometry of N Dimensions*. Dover books on mathematics. Dover Publications, 2004. ISBN: 9780486439273. URL: https://books.google.com.ua/books?id=vfWpT1X38S8C.

[18] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. DOI: 10.48550/ARXIV.1312.6114. URL: https://arxiv.org/abs/1312.6114.

[19] Marius Kloft and Pavel Laskov. "Online Anomaly Detection under Adversarial Impact". In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, 2010, pp. 405–412. URL: https://proceedings.mlr.press/v9/kloft10a.html.

[20] Alex Krizhevsky. *Learning multiple layers of features from tiny images*. Tech. rep. 2009.

[21] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. *Adversarial Machine Learning at Scale*. 2016. DOI: 10.48550/ARXIV.1611.01236. URL: https://arxiv.org/abs/1611.01236.

[22] Yann LeCun and Corinna Cortes. "MNIST handwritten digit database". In: (2010). URL: http://yann.lecun.com/exdb/mnist/.

[23] Mathias Lecuyer et al. *Certified Robustness to Adversarial Examples with Differential Privacy*. 2018. DOI: 10.48550/ARXIV.1802.03471. URL: https://arxiv.org/abs/1802.03471.

[24] Guang-He Lee et al. *Tight Certificates of Adversarial Robustness for Randomly Smoothed Classifiers*. 2019. DOI: 10.48550/ARXIV.1906.04948. URL: https://arxiv.org/abs/1906.04948.

[25] Alexander Levine and Soheil Feizi. *Robustness Certificates for Sparse Adversarial Attacks by Randomized Ablation*. 2019. DOI: 10.48550/ARXIV.1911.09272. URL: https://arxiv.org/abs/1911.09272.

[26] Jing Lin et al. "ML Attack Models: Adversarial Attacks and Data Poisoning Attacks". In: *CoRR* abs/2112.02797 (2021). arXiv: 2112.02797. URL: https://arxiv.org/abs/2112.02797.

[27] Aleksander Madry et al. *Towards Deep Learning Models Resistant to Adversarial Attacks*. 2017. DOI: 10.48550/ARXIV.1706.06083. URL: https://arxiv.org/abs/1706.06083.

[28] Jeet Mohapatra et al. *Higher-Order Certification for Randomized Smoothing*. 2020. DOI: 10.48550/ARXIV.2010.06651. URL: https://arxiv.org/abs/2010.06651.

[29] Seyed-Mohsen Moosavi-Dezfooli et al. *Robustness via curvature regularization, and vice versa*. 2018. DOI: 10.48550/ARXIV.1811.09716. URL: https://arxiv.org/abs/1811.09716.

[30] Nicolas Papernot et al. "Practical Black-Box Attacks against Deep Learning Systems using Adversarial Examples". In: *CoRR* abs/1602.02697 (2016). arXiv: 1602.02697. URL: http://arxiv.org/abs/1602.02697.

[31] L. Ros, A. Sabater, and F. Thomas. "An ellipsoidal calculus based on propagation and fusion". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 32.4 (2002), pp. 430–442. DOI: 10.1109/TSMCB.2002.1018763.

[32] Hadi Salman et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers". In: *CoRR* abs/1906.04584 (2019). arXiv: 1906.04584. URL: http://arxiv.org/abs/1906.04584.

[33] H.H. Sohrab. *Basic Real Analysis*. Birkhäuser Boston, 2003. ISBN: 9780817642112. URL: https://books.google.com.ua/books?id=gBPI_oYZoMMC.

[34] Christian Szegedy et al. "Going Deeper with Convolutions". In: *Computer Vision and Pattern Recognition (CVPR)*. 2015. URL: http://arxiv.org/abs/1409.4842.

[35] Christian Szegedy et al. *Intriguing properties of neural networks*. 2013. DOI: 10.48550/ARXIV.1312.6199. URL: https://arxiv.org/abs/1312.6199.

[36] Pooya Tavallali et al. "Adversarial Poisoning Attacks and Defense for General Multi-Class Models Based On Synthetic Reduced Nearest Neighbors". In: *CoRR* abs/2102.05867 (2021). arXiv: 2102.05867. URL: https://arxiv.org/abs/2102.05867.

[37] Jiaye Teng, Guang-He Lee, and Yang Yuan. *ℓ_1 Adversarial Robustness Certificates: a Randomized Smoothing Approach*. 2020. URL: https://openreview.net/forum?id=H1lQIgrFDS.

[38] Vincent Tjeng, Kai Xiao, and Russ Tedrake. *Evaluating Robustness of Neural Networks with Mixed Integer Programming*. 2017. DOI: 10.48550/ARXIV.1711.07356. URL: https://arxiv.org/abs/1711.07356.

[39] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. *Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks*. 2018. DOI: 10.48550/ARXIV.1802.04034. URL: https://arxiv.org/abs/1802.04034.

[40] Eric Wong and J. Zico Kolter. *Provable defenses against adversarial examples via the convex outer adversarial polytope*. 2017. DOI: 10.48550/ARXIV.1711.00851. URL: https://arxiv.org/abs/1711.00851.

[41] Greg Yang et al. *Randomized Smoothing of All Shapes and Sizes*. 2020. DOI: 10.48550/ARXIV.2002.08118. URL: https://arxiv.org/abs/2002.08118.

[42] Runtian Zhai et al. *MACER: Attack-free and Scalable Robust Training via Maximizing Certified Radius*. 2020. DOI: 10.48550/ARXIV.2001.02378. URL: https://arxiv.org/abs/2001.02378.

[43] Dinghuai Zhang* et al. *Filling the Soap Bubbles: Efficient Black-Box Adversarial Certification with Non-Gaussian Smoothing*. 2020. URL: https://openreview.net/forum?id=Skg8gJBFvr.