# RANCER: Non-Axis Aligned Anisotropic Certification with Randomized Smoothing

Taras Rumezhak
Ukrainian Catholic University
SoftServe
rumezhak@ucu.edu.ua trume@softserveinc.com

Francisco Girbal Eiras
University of Oxford
eiras@robots.ox.ac.uk

Philip H.S. Torr
University of Oxford
philip.torr@eng.ox.ac.uk

Adel Bibi
University of Oxford
adel.bibi@eng.ox.ac.uk

## Abstract

*As modern networks have been proven to be unprotected from adversarial attacks and are applied in safety-critical applications, defense against them is very crucial. Many works were dedicated to this topic, but randomized smoothing has been recently proven to be an effective approach for the certified defense of deep neural networks and getting robust classifiers. Some prior results were obtained utilizing the techniques of adding extra parameters to extend the limits of the certification regions. In this way, sample-wise optimization was proposed to maximize the certification radius per input. The idea was further extended with the generalized anisotropic counterparts of $\ell_1$ and $\ell_2$ 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 the CIFAR-10 dataset to report the improved performance.*

## 1. Introduction

Deep Neural Networks (DNNs) for image classification have been shown to perform well in a variety of different fields, even outperforming humans in some medical imagery tasks [20]. However, they are also known to be vulnerable to *adversarial attacks* - small, imperceptible perturbations at the input level that lead to misclassification of the image [10]. This is particularly problematic in safety-critical applications of DNNs, such as autonomous driving [3], where guarantees are required before deployment. This motivates the need for certifiably robust classifiers, *i.e.*, classifiers that are provably robust over a certain input region. Therefore, many approaches were proposed to build the truly robust models [5, 29]. However, a lot of the current certification techniques have a scalability (verification

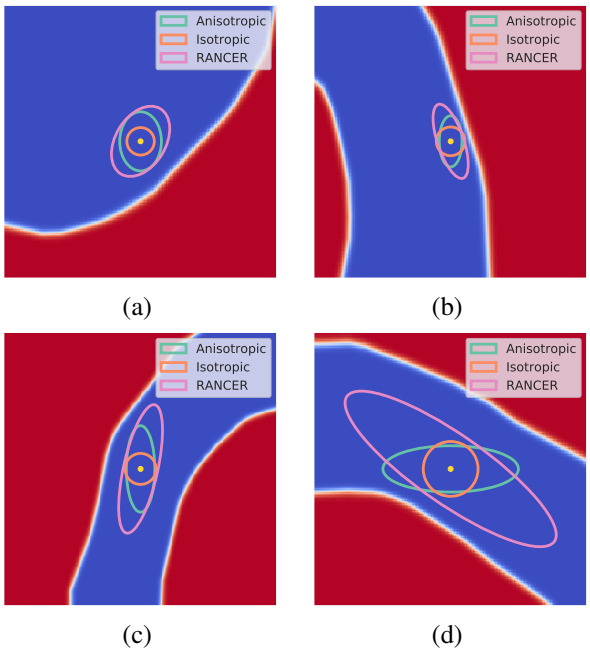



Figure 1. **Going beyond axis-aligned certification.** Example of the $\ell_2$ certification regions presented on the $2D$ dataset, where the blue and red pixels correspond to different data classes. Ellipsoids represent certification regions. Orange: Data dependent isotropic region [1], Blue: Anisotropic (ANCER) region [9], Pink: region obtained with our proposed solution - RANCER.

based methods [8, 27]) or computation (conservative certification [28, 12]) issues making them unlikely to be used in production pipelines.

Randomized smoothing is a recent approach that allows one to obtain such classifiers in a scalable manner [17, 5]. While several works in this field have introduced a variety of $\ell_p$ certificates [17, 5, 18, 7, 1], most of them have focused on *isotropic* certificates, *i.e.*, certificates that have perfect radial symmetry with respect to the input. As mentioned by Eiras *et al.* in [9], this is sub-optimal, as it considers only *worst-case* perturbations, ignoring other safe regions around the point. To tackle this inefficiency, in [9] the authors introduce the first *anisotropic* randomized smoothing framework for generalized $\ell_p$ norm certificates - ANCER. However, the data-dependent regions obtained by the procedure outlined in the paper are limited to axis-aligned ones, which can severely hinder the safe regions obtained, as motivated by the 2D images in Figure 1. In some specific cases when the input data distribution is concentrated along fixed axis direction, ANCER may show good results. Nonetheless, when the data is rotated, the performance is expected to drop by the axis constraint in the design. In this work, we extend the ANCER framework to the non-axis aligned setting, allowing for robust accuracy gains and a better characterization of the safe regions of the classifier.

**Contributions** can be summarized as follows:

- We provide a general analysis of non-axis aligned anisotropic certification, while preserving previous approaches as special cases.
- We conduct experiments on the CIFAR-10 dataset to validate out approach and show that our generalized framework outperforms existing approaches in terms of $\ell_2$ certified accuracy.

**Paper Structure.** In Section 2, you will find the general overview of the different techniques for the defense against adversarial attacks. Later, Section 3 is dedicated to the detailed description of the proposed approach - RANCER. Then, in Section 4, we report the conducted experiments and show the improved performance of our algorithm comparing to previous SOTA, analyze time and initialization. Finally, in Sections 5 and 6 we discuss the limitations and sum up the key takeaways of our work respectively.

## 2. Related Work

**Empirical Defenses.** In 2015 Goodfellow et al. [10] showed the way to attack DNNs and presented adversarial training. With the help of the fast gradient sign method, adversarial examples can be quickly generated during training making it possible to apply in practice. They trained the network on adversarial examples making the model more robust to adversarial attacks on MNIST dataset [16]. Then Kurakin et al. [14] in 2017 showed how to apply the explicit model training on the adversarial examples on a big scale. The authors applied adversarial training on ImageNet [6] and wrote the exact recommendations for how to successfully scale adversarial training to large models and datasets. One year later Madry et al. [19] started to think about fully resistant NNs which can be robust to a wide range of adversarial attacks. They discussed how to find more powerful attacks during training but 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 robustness methods and showed that all previous attacks can be defeated by carefully constructing new loss as well as new adversaries are harder to defeat. In that way, new attacks were breaking previously robust models and then new defenses were proposed. For example, Athalye et al. [2] analyzed that obfuscated gradients 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. As a result, there was less trust in empirical defenses, and interest increased in the defenses with formal guarantees.

**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 $\ell_1$, $\ell_2$ or $\ell_\infty$ ball. The certification works for both: generically trained NNs and robustly trained ones. For example, Wong and Kolter [8] proposed the method to learn deep ReLU-based classifiers that are provably robust against norm-bounded adversarial perturbations on the training data. There were some works proposing *exact certification*: take a smoothed classifier $g$, and check if there exists a perturbation with a norm lower than some $r$. The classifier is certifiably robust if the output corresponding to the perturbed input is the same as the output for the original input. For example, Ehlers [27] presented an approach for the verification of feed-forward neural networks in which all nodes have a piece-wise linear activation function. The problem with these methods is the lack of possibility to scale to large NNs. Tjeng et al. [27] 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 $\ell_\infty$ norm $\epsilon = 0.1$. But even this and some other recent achievements do not scale to the SOTA networks working with CIFAR10 [13] or ImageNet [6]. Then *conservative certification* 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. [28] presented an efficient calcula-

tion 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 [12] gave formal guarantees on the robustness of a classifier with instance-specific lower bounds on the norm of the input manipulation required to change the classifier decision and proposed the Cross-Lipschitz regularization functional, but calculations were still expensive.

**Randomized Smoothing.** In 2019 Lecuyer et al. [17] proposed the first certified defense based on differential privacy techniques and called it PixelDP. It scales to large networks and datasets (such as Google's Inception network [25] for ImageNet [6]) and applies broadly to arbitrary model types. In this work, the result was proved to be constant with average classifier prediction under Laplacian noise perturbations for $\ell_1$ certification. These ideas were later improved for the $\ell_2$ certification by Cohen et al. [5] for smoothing with Gaussian noise. They 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 $\ell_2$ norm. Later there were some other papers that showed the proofs for $\ell_1$ (Teng et al. [26]), $\ell_0$ (Levine and Feizi [18]), and even $\ell_p$ norm (Dvijotham et al, [7]). Those methods were proven to find near-optimal certification regions under different norms, but the certification was still very small. To resolve this, Mohapatra et al. [21] 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 did not provide a closed-form solution. In contrast to Cohen et al. [5] where the parameters of the model were set as a global hyperparameter, Alfarra et al. [1] 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 technique they achieved 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. [9] extend the isotropic randomized smoothing $\ell_1$ and $\ell_2$ certificates to their generalized anisotropic counterparts. The proposed framework called the ANCER achieves SOTA $\ell_1$ and $\ell_2$ performance on the CIFAR-10 and ImageNet utilizing the previous ideas of data dependent smoothing [1]. Previous approaches' certification regions were limited by the worst-case adversaries because of isotropic properties, but other (potentially large) areas can exist and be discovered by anisotropic counterparts. However the described anisotropic case is limited by the axis alignment and can extend only in a predefined set of direction, so we overcome this and propose an extended version that will not be aligned and can find larger safe regions in any direction.

# 3. RANCER: Non-Axis Aligned Anisotropic Certification

We extend ANCER [9] to provide a practical approach for certifying non-axis aligned anisotropic region. We dub our approach as RANCER, where the first "R" denotes rotation, i.e. certified regions rotated beyond the canonical basis. We first provide a theoretical intuition based on the general certification results of ANCER towards certifying non-axis aligned regions for a given fixed orthogonal transformation. However, since the learning of such orthogonal transformation is expensive as it is of the dimensionality of the input, we show one efficient approach for designing such a transformation. The detailed pipeline algorithm is proposed in Algorithm 1.

Recall that the early work of Cohen et al. [5] smooths a base classifier $f : \mathbb{R}^n \to \mathcal{P}(\mathcal{Y})$ where $\mathcal{P}(\mathcal{Y})$[1] is a probability simplex over classes $\mathcal{Y}$ with isotropic Gaussian distribution. In particular, the smooth classifier is given as $g_\sigma(x) = \mathbb{E}_\epsilon[f(x+\epsilon)]$ where $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The resultant smooth classifier thereafter enjoys an isotropic certified region parameterized by $\sigma$ and the top two predictions of $g_\sigma$. ANCER [9] then proposed a general extension where when one smooths the base classifier $f$ with a general positive definite covariance matrix $\mathbf{A}$, that is to say the smooth classifier is given as $g_{\mathbf{A}_{\text{ANCER}}}(x) = \mathbb{E}_\epsilon[f(x+\epsilon)]$ where $\epsilon \sim \mathcal{N}(0, \mathbf{A})$. However, the works limits $\mathbf{A}$ to be a diagonal matrix, i.e. $\mathbf{A}_{\text{ANCER}} = \Sigma$ where $\Sigma$ is a positive diagonal. This gives rise to certified regions that are anisotropic but axis aligned. In this work, we are interested in the certified regions that arise when using full dense covariance matrix $\mathbf{A}$, as well as approaches towards selecting $\mathbf{A}$. Note that this is of interest as it is a generalization to both Cohen et al [5] and ANCER [9]. First, we note that a proper covariance matrix is symmetric and therefore can be orthogonally diagonalzied, i.e. $\mathbf{A} = \mathbf{U}\hat{\Sigma}\mathbf{U}^\top$, where $\mathbf{U}$ and $\hat{\Sigma}$ are the set eigenvectors and eigenvalues, respectively. To that end, we consider the following smooth classifier instead:

$$g_{\mathbf{A}}(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \mathbf{A})}\left[f(x+\epsilon)\right]. \tag{1}$$

For a given set of eigenvectors $\mathbf{U}$ defining an orthogonal transformation, we introduce the following parameterization. Let $\epsilon' = \mathbf{U}\epsilon$ where $\epsilon \sim \mathcal{N}(0, \hat{\Sigma})$. Then, we have that $\epsilon' \sim \mathcal{N}(0, \mathbf{U}\hat{\Sigma}\mathbf{U}^\top)$ which is $\epsilon' \sim \mathcal{N}(0, \mathbf{A})$. Therefore, our smooth classifier in Equation (1) can equivalently be written as:

$$g_{\hat{\Sigma}}(x) = \mathbb{E}_{\epsilon \sim \mathcal{N}(0, \hat{\Sigma})}\left[f(x+\mathbf{U}\epsilon)\right]. \tag{2}$$

We observe that our new proposed smooth classifier is almost identical to that of ANCER where the smoothing distribution is with a diagonal covariance. The only exception

[1] We use here the soft smooth version for convenience following [24].

**Algorithm 1** Non-Axis Aligned Anisotropic Certification

**Input:** base classifier $f_\theta$, input point $x$, learning rate $\alpha$, initial sigma $\hat{\Sigma}^0$, number of noise samples $m$, number of iterations $k$, loss function $\mathcal{L}$, minimum and maximum clipping difference thresholds $\gamma_1$, $\gamma_2$

**Result:** Optimized sigmas $\hat{\Sigma}^k$ for input point $x$

$\mathbf{H} = \text{Hessian}(\mathcal{L})$;

$\mathbf{U}, \Lambda = \text{EigenDecomposition}(\mathbf{H})$;

**for** $i = 0 \dots k$ **do**

sample $\hat{\epsilon}_1, \dots \hat{\epsilon}_m \sim \mathcal{N}(0, \hat{\Sigma}^i)$

$\psi = \frac{1}{m} \sum_{j=1}^{n} f(x + \mathbf{U}\hat{\epsilon}_j)$

$E_A = \max_c \psi^c$; $y_A = \text{argmax}_c \psi^c$

$E_B = \max_{c \neq y_A} \psi^c$

$R(\hat{\Sigma}) = 1/2 \sqrt[n]{\prod_{j=1}^{n} \hat{\Sigma}_{jj}} \left(\Phi^{-1}(E_A) - \Phi^{-1}(E_B)\right)$

$\hat{\Sigma}_{jj}^{i+1} \leftarrow \hat{\Sigma}_{jj}^{i} + \alpha \nabla_{\Sigma_{jj}^i} R(\hat{\Sigma}^i)$

$\hat{\Sigma}_{jj}^{i+1} \leftarrow \min\left(\max\left(\hat{\Sigma}_{jj}^{i+1}, \sigma_0(1-\gamma_1)\right), \sigma_0(1+\gamma_2)\right)$

**end**

**return** $\hat{\Sigma}^k$

is with the rotated noise with $\mathbf{U}$. In the 2D case, i.e. $x \in \mathbb{R}^2$, we can view $\mathbf{U}\epsilon$ as a rotation to the sampling distribution of the ellipsoid with scale governed by $\hat{\Sigma}$. To that end, one can directly apply the certification result of ANCER which states that $g_{\hat{\Sigma}}(x) = g_{\hat{\Sigma}}(x + \delta)$ for all $\delta$ satisfying:

$$\sqrt{\delta^\top \hat{\Sigma}^{-1} \delta} \leq \frac{1}{2}\left(\Phi^{-1}(g_{\hat{\Sigma}}^{c_1}(x)) - \Phi^{-1}(g_{\hat{\Sigma}}^{c_2}(x))\right), \qquad (3)$$

where $c_1$ and $c_2$ are the top two predictions of $g_{\hat{\Sigma}}$ (from [9]). **A natural question arises here: how can one efficiently select U?** One potential approach is to directly optimize over symmetric positive definite matrices to learn $\mathbf{U}$ for every input $x$. This is in a similar spirit to Alfarra et al [1] who optimized directly for a scalar $\sigma$ for every input $x$ followed with a post-processing memory based certification. However, optimizing over $\mathbf{U}$ is generally very expensive, note that if $x \in \mathbb{R}^n$ then $\mathbf{U}$ is of size $n \times n$. Instead, we propose to directly estimate the local curvature by investigating the hessian of the loss function.

Moosavi-Dezfooli et al. [22] showed a connection between adversarial robustness and the loss of the curvature locally. In particular, under certain quadratic approximation of the loss function locally, they show that the higher curvature, measured as the maximum eigenvalue of the loss function at a point $x$, the less robust the classifier is around $x$. That is to say, if $\mathcal{L}$ is a suitable loss function for a given classifier, then locally around a point $x\mathbb{R}^n$ the loss function can be approximated as:

$$\mathcal{L}(x + \epsilon) \approx \mathcal{L}(x) + \epsilon^\top \nabla \mathcal{L}(x) + \frac{1}{2}\epsilon^\top \mathbf{H} \epsilon, \qquad (4)$$

where $\mathbf{H}$ is denotes the Hessian of size $n \times n$. Moosavi-Dezfooli et al [22] show that classifiers with larger $\lambda_{\text{max}}(\mathbf{H})$, where $\lambda_{\text{max}}$ denotes the largest eigenvalues, have a larger curvature resulting a less robust classifier. In particular, the classifier at $x$ is less robust particularly along the eigenvector of $\mathbf{H}$ corresponding to the largest eigenvalue. This motives our approach towards selecting the transformation $\mathbf{U}$ when smoothing the base classifier $f$. Note that the eigenvector space of the hessian of the loss function $\mathbf{H}$ indicates the space where the base classifier $f$ is robust and not robust. To that end, we seek to perform more smoothing, i.e. larger diagonal $\hat{\Sigma}$ along the eigencevector directions where the hessian have high eigenvalues. This implies that we are performing more "smoothing" to the non robust region and thereafter will will be robust, i.e. will require larger perturbation along those direction to flip the prediction of the smooth classifier. To that end, instead of learning $\mathbf{U}$ we set $\mathbf{U} = \mathbf{V}$ where $\mathbf{H} = \mathbf{V}\Lambda\mathbf{V}^\top$. That is to say, we fix the orthogonal transformation for the smoothing locally from the eigenvector space of the hessian of the loss function. We then following ANCER [9] optimize directly for $\hat{\Sigma}$, i.e. finding the the diagonal smoothing distribution under the predetermined transformed coordinates that is aligned with the non-robust directions.

As a summary, our approach can be summarized as follows: (**1**) compute the Hessian $\mathbf{H}$ of the loss at point to be certified $x$; (**2**) perform eigendecomposition of Hessian and set matrix $\mathbf{U}$ to be eigenvectors of the $H$, i.e. $\mathbf{V}$; (**3**) sample noise from the new non-axis aligned distribution and optimize directly following ANCER [9] the diagonal elements $\hat{\Sigma}$ that maximize the volume of the certified region. Note that in step (**3**), we sample the noise exactly the same as in ANCER, and then transform it by pre-multiplying it by $\mathbf{U}$, i.e. rotating it. Algorithm 1 summarizes our approach.

**Certification.** Our proposed smooth classifier in Equation (2) enjoys a certified radius as given in Equation 3 as shown by Eiras et al. However, one key limitations of input dependent smooth classifiers is that for every input $x$, we construct a new a classifier with a different smoothing distribution with covariance $\mathbf{A}$. Note that this is since $\mathbf{A}$ depends both on the eigenvectors of the hessian on the loss function at $x$ and that we optimize for the eigenvalues $\mathbf{A}$ for every $x$. To that end, still need to assure that the two classifiers at $x_1$ and $x_2$, i.e. $g_{\mathbf{A}_1}$ and $g_{\mathbf{A}_2}$ each with a certification region $\mathcal{S}_1$ and $\mathcal{S}_2$ centered at $x_1$ and $x_2$ are consistent. That is to say, that there exists no intersection of regions $\mathcal{S}_1 \cap \mathcal{S}_2 = \phi$ for any pair of smooth classifiers with two different predictions for $x_1$ and $x_2$. To that end, and following the works of [1] and [9], we extend memory based certification for the general non-axis aligned anisotropic regions.

**Memory-Based Certification.** To perform memory-based certification for RANCER, we can directly apply Algorithm (1) of Eiras et al with changes relating to the the computation of `Intersect`. This function returns TRUE

if two axis aligned ellipsoids intersect and FALSE otherwise. As was mentioned previously the test on intersecting ellipsoids is a computationally hard problem. In ANCER the authors were able to simplify the computation significantly as the ellipsoids were axis-aligned, i.e. the smoothing distribution defining an ellipse $\mathbf{A}$ is diagonal. However, we are interested in solving the general test problem of intersecting arbitrary ellipsoids. In particular, given two ellipsoids $\mathcal{S}_{\mathbf{Q}} = \{x \in \mathbb{R}^n : (x-q)^\top \mathbf{Q}(x-q) \leq 1\}$ and $\mathcal{S}_{\mathbf{R}} = \{x \in \mathbb{R}^n : (x-r)^\top \mathbf{R}(x-r) \leq 1\}$ where $\mathbf{Q}$ and $\mathbf{R}$ are general symmetric positive definite matrices, the problem of testing if two ellipsoids intersect reduces to checking whether there exists a $t \in (0,1)$ where:

$$K(t) = 1 - (r-q)^\top \left(\tfrac{1}{1-t}\mathbf{R}^{-1} + \tfrac{1}{t}\mathbf{Q}^{-1}\right)^{-1}(r-q) < 0. \tag{5}$$

Moreover, $K(t)$ was shown earlier by [23] to be convex in $t$ and thus the problem can be easily solved by any convex optimization solver solving $t^* = \text{argmin}_t K(t)$ and checking whether $K(t^*) < 0$. Moreover, another change to the memory-based certification is the one concerning `LargestOutSubset` from Algorithm (1) in Eiras et al. This function should, given two general ellipsoids described above that do intersect, find the the largest isotropic ball of one ellipsoid such that it does not intersect with the other ellipsoid. In particular, consider the problem of reducing $\mathcal{S}_{\mathbf{R}}$ to the smallest $\ell_2$ ball that does not intersect with the ellipsoids $\mathcal{S}_{\mathbf{Q}}$. To that end, we solve the more general problem of solving the following scalar non-linear Equation:

$$(r-q)^\top (2\lambda\mathbf{Q} + \mathbf{I})^{-\top}\mathbf{Q}(2\lambda\mathbf{Q} + \mathbf{I})^{-1}(r-q) = 1. \tag{6}$$

This is as opposed to solving the simplified problem where $\mathbf{Q}$ and $\mathbf{R}$ are diagonal resulting in a reduced more efficient algorithm as proposed by Eiras et al [9].

# 4. Experiments

We evaluate RANCER and show that it outperforms the previous state-of-the-art ANCER [9] on CIFAR-10 dataset for $\ell_2$ and $\ell_2^\Sigma$ in certified robustness. The following subsections are dedicated to the detailed information about experimental setup, evaluation metrics discussion, and obtained final certification results.

**Experimental Setup.** Following the previous evaluation procedures for robust classifiers which were established in previous works, we used CIFAR-10 dataset [13] and pretrained *ResNet-18* architecture proposed in 2015 by He et al. [11] (the weight were taken from [9] repository). We also use isotropic $\sigma$ as initialization for optimization similarly to ANCER. An important factor in the success of the defenses against adversarial attacks is the scalability to deep models and large datasets. Therefore such data and model choice was made to prove the advantage against verification-based methods. Our experimental setup matches those of prior art for a fair comparison. Specifically, we compare our work against fixed $\sigma$, DDS [1] and ANCER following Cohen *et al*. [5] and [9] for isotropic and anisotropic certification, respectively. We compare all methods by reporting the certified accuracy at multiple radii, the average certified radius proposed by MACER [30], and the proxy radius for anisotropic regions. Additionally, we propose two new metrics to track the exact certified radius and proxy radius improvement comparing to ANCER.

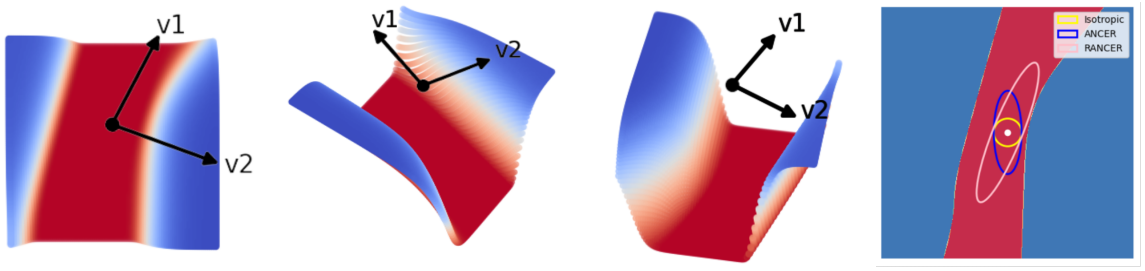
## 4.1. Evaluation Metrics

**Certified Accuracy.** A classifier $f$ is said to be $\ell_p$ certifiably accurate with radius $r$ at $x$ if the classifier predicts $x$ correctly and the prediction is constant for all perturbations $\delta$ in an $\ell_p$ ball of radius $r$. That is to say, $\arg\max_c f^c(x) = \arg\max_c f^c(x+\delta) = y \; \forall \|\delta\|_p \leq r$, where $f^c$ is the $c^{\text{th}}$ element of $f$. We compute the certified accuracy as the portion of the test set correctly classified by $g_{\hat{\Sigma}}$ and has an $\ell_p$ ($p = 2$ in our case) certified radius at least $r$.

**Anisotropic Certified Accuracy.** We use the definition of the anisotropic certified accuracy from the specialization of Definition 1 from [9]. Following that, our ellipsoid certified region $\mathcal{R}_2$ is "superior certificate" to the isotropic region $\mathcal{R}_1$ ($\ell_2$-ball). To compare the regions we can compute the volumes, which in our case will be $\mathcal{V}(\mathcal{R}_2) = r_2^n\sqrt{\pi^n}/\Gamma(n/2+1)\prod_{i=1}^n \hat{\Sigma}_{ii}$ [15]. Instead of calculating volume directly, we compute *proxy radius* for $\mathcal{R}_2$ as $\tilde{R} = r_2 \sqrt[n]{\prod_i^n \hat{\Sigma}_{ii}}$, because larger $\tilde{R}$ will give regions with larger volumes. Based on this, the anisotropic certified accuracy is computed as the portion of the correctly classified samples with an $\ell_2^\Sigma$ proxy radius that is at least $\tilde{R}$.

**Average Certified (Proxy) Radius.** Following the previous evaluation baselines, we also report two more metrics proposed by Zhai *et al*. [30], also extended by ANCER [9] for anisotropic regions, namely $ACR = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\left[R_x \mathbb{1}(g_{\hat{\Sigma}}(x) = y)\right]$, Average Certified Radius, and $AC\tilde{R} = \mathbb{E}_{(x,y)\sim\mathcal{D}_t}\left[\tilde{R}_x \mathbb{1}(g_{\hat{\Sigma}}(x) = y)\right]$, Average Certified Proxy Radius, where $R_x$ and $\tilde{R}_x$ are the radius and proxy radius at sample $x$, respectively, with corresponding ground truth label $y$. $\mathbb{1}$ is the indicator function.

**Average (Proxy) Radius Improvement.** To summarily analyze the improvement of RANCER with respect to ANCER [9], we introduce two new metrics. The first metric is average radius improvement $ARI = \mathbb{E}_{x,y\sim\mathcal{D}_t}[R_x \mathbb{1}(g_{\hat{\Sigma}}(x) = y) - \mathbb{R}_x \mathbb{1}(g_{\mathbf{A}}(x) = y)]$ and the second is average proxy radius improvement $A\tilde{R}I = \mathbb{E}_{x,y\sim\mathcal{D}_t}[\tilde{R}_x \mathbb{1}(g_{\hat{\Sigma}}(x) = y) - \tilde{\mathbb{R}}_x \mathbb{1}(g_{\mathbf{A}}(x) = y)]$ where $\mathbb{R}_x$ and $\tilde{\mathbb{R}}_x$ are the corresponding radius and proxy radius obtained with ANCER. According to the definition, a positive $ARI/A\tilde{R}I$ indicates RANCER on average outperforms the radii/proxy radii obtained by ANCER in $\mathcal{D}_t$.




(a) (b) (c) (d)

Figure 2. **Safe directions approximation via Hessian eigenvectors.** Figures (a,b,c) show the loss function value in the the directions $\mathbf{v_1}$ and $\mathbf{v_2}$ which are the Hessian eigenvectors corresponding to largest and second largest eigen values. The red region corresponds to small loss value while the blue region corresponds to the high loss value. Figure (d) shows a point classified as class "red" along with the certified regions 3 different methods.

### 4.2. Safe Directions Approximation

To illustrate the practical soundness of using the eigenvectors of the Hessian of the loss function as the covariance of our smoothing distribution following the similar setup as proposed in [22], we conduct different experiments with 2D data to show how good the safe approximation is based on the difference between the original and second-order approximation of the loss function. An example of safe directions approximation can be seen in Figure 2. The center point classified as class "red" from (**d**) is used to show the hessian approximation. This point is located in the middle of the certified ellipsoid and is marked as a black dot in (**a**,**b**,**c**). The red "valley" represents the points where the loss value is small for this class, while blue points correspond to higher values of the loss. Based on the representative curvature of the "valley" dependent on the loss values we obtain a Hessian with a good directions approximation.

As observed in **(a)**, **(b)**, and **(c)**, the first eigenvector $\mathbf{v_1}$ points to the safe potential direction to expand. During optimization vector $\mathbf{v_1}$ will become larger and $\mathbf{v_2}$ shorter and they will form the radii of the final pink ellipsoid in **(d)**. As a conclusion, due to our experiments, the safe directions are reasonably aligned with the eigenvectors of the Hessian of the loss function, so it is a good choice for approximation.

### 4.3. CIFAR-10 Certification Results

We report the previously discussed final evaluation metrics obtained with Gaussian smoothing procedure described in the experimental setup and clearly defined in Algorithm 1. Following the evaluation implementation from [5, 9], we calculated the top-1 certified accuracy for different $\ell_2$ radii for CIFAR-10 dataset. The obtained values were taken after memory-based certification following the same logic as previously stated by [9], see Tables 1 and 2 for radius and proxy radius results correspondingly. 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 the certification region. With the conducted experiments we obtained $ARI = 0.0673$ and $A\tilde{R}I = 0.0792$ compared to ANCER. Thus, on average RANCER has bigger certification radii letting it to increase the certified accuracy.

Additionally, we show the plots of certified accuracy and anisotropic certified accuracy as a function of radius and proxy radius, respectively, for the different methods in Figure 3. From the visual comparison we can see that both Isotropic DD and ANCER achieve better performance than Fixed $\sigma$ for radii bigger then $0.5$. This coincides with the findings reported in [1, 9] because fixed $\sigma$ struggles with the robustness/accuracy trade-off and leads to the rapid drop in accuracy. ANCER shows better certification results by certifying larger regions in terms of volume utilizing the anisotropic generalization. However, our proposed approach achieves even better performance giving the potential to extend to bigger non-axis aligned regions. Analyzing the visual results, we see that the improvement of certified accuracy is almost stable for all radii $R$ and $\tilde{R}$.

We observe that RANCER improves certified accuracy to $82\%$ (from $75\%$) and $48\%$ (from $43\%$) for $\ell_2$ radii $0.25$ and $0.5$, respectively. In that way, the proposed method outperforms the previous state-of-the-art approach for $\ell_2$ certified robustness. Moreover, in the same manner, RANCER improves $ACR$ and $AC\tilde{R}$ significantly (see Tables 1 and 2). As expected, we were able to obtain much larger certification regions by utilizing the non-axis aligned anisotropic generalization which leads to higher $ACR$ and $AC\tilde{R}$.

### 4.4. Runtime Analysis

We compare the runtime complexity of RANCER against previous methods. All experiments are conducted on NVIDIA GeForce RTX 2080 GPU with a total memory of 8Gb. We report the time results in Table 3. Previous SOTA approaches [1, 9] have significantly improved certified accuracy, but with a cost of runtime because of sample-wise optimization in contrast to Fixed $\sigma$. RANCER

| Certification Method | Accuracy @ $\ell_2$ radius (%) | | | | | | $\ell_2$ $ACR$ |
|---|---|---|---|---|---|---|---|
| | 0.25 | 0.5 | 0.75 | 1.0 | 1.25 | 1.5 | |
| Fixed $\sigma$ | 56 | 1 | 0 | 0 | 0 | 0 | 27 |
| ISOTROPIC DD [1] | 38 | 15 | 7 | 1 | 0 | 0 | 25 |
| ANCER [9] | 75 | 43 | 22 | 7 | 0 | 0 | 46 |
| **RANCER** | **81** | **48** | **28** | **11** | **1** | 0 | **53** |

Table 1. **Certified accuracy comparison at different $\ell_2$ radii and $\ell_2$ and ACR on CIFAR-10.** We compare top-1 certified accuracy and $ACR$ obtained by using the isotropic $\sigma$ used during training of NNs (Fixed $\sigma$); the isotropic data-dependent optimization procedure from [1] (Isotropic DD); and anisotropic approach from [9] (ANCER).

| Certification Method | Accuracy @ $\ell_2^\Sigma$ proxy radius (%) | | | | | | $\ell_2^\Sigma$ $AC\tilde{R}$ |
|---|---|---|---|---|---|---|---|
| | 0.25 | 0.5 | 0.75 | 1.0 | 1.25 | 1.5 | |
| Fixed $\sigma$ | 56 | 1 | 0 | 0 | 0 | 0 | 27 |
| ISOTROPIC DD [1] | 38 | 15 | 7 | 1 | 0 | 0 | 25 |
| ANCER [9] | 77 | 64 | 45 | 29 | 17 | 12 | 72 |
| **RANCER** | **82** | **68** | **48** | **36** | **21** | **13** | **80** |

Table 2. **Certified accuracy comparison at different $\ell_2^\Sigma$ proxy radii and $\ell_2^\Sigma$ and AC$\tilde{\text{R}}$ on CIFAR-10.** We compare top-1 certified accuracy and $AC\tilde{R}$ obtained by using the isotropic $\sigma$ used during training of NNs (Fixed $\sigma$); the isotropic data-dependent optimization procedure from [1] (Isotropic DD); and anisotropic approach from [9] (ANCER).

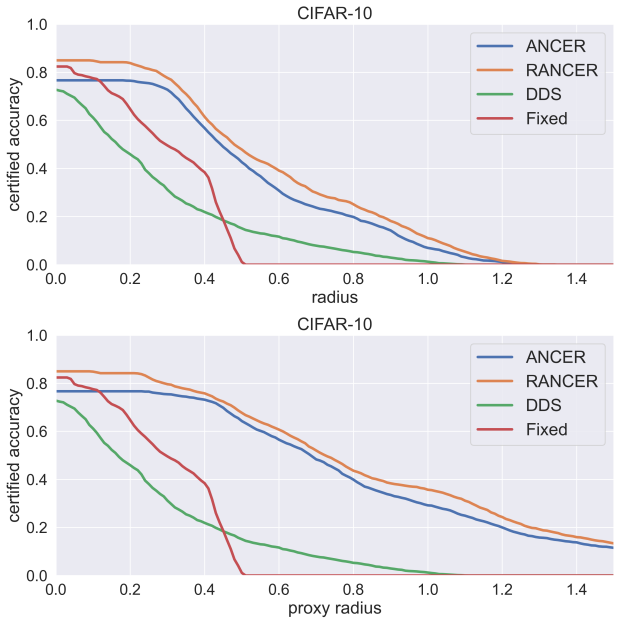



Figure 3. **Certified accuracies over multiple radii on CIFAR-10.** Distribution of top-1 certified accuracy as a function of $\ell_2$ radius (on the top) and $\ell_2^\Sigma$ proxy radius (bottom) obtained with different approaches. RANCER line is on the top for every value of radii showing the improved performance across all values of $r$.

is approximately 2.5 times slower than ANCER. The additional complexity comes from the hard procedure of safe directions calculation as it involves calculating Hessian and performing eigendecomposition. This affects a batch size significantly and we had to set it at most of 4 to be able to run the experiments (it was 128 in [9]). Another overhead is caused due to a general check of ellipsoids intersection mentioned in Section 3. As such, we observe RANCER trades off certified accuracy for runtime efficiency.

| Fixed $\sigma$ | Isotropic DD | ANCER | RANCER |
|---|---|---|---|
| 4.2s | 4.9s | 7.65s | 19.1s |

Table 3. **Per sample certification time for each method.** The results of average certification time for each sample are reported here. The measures were done on the same hardware mentioned in the Section 4.4.

### 4.5. Sensitivity to Initialization

One of the bottlenecks of the DDS based approaches is the optimization procedure for $\sigma$. To simplify the computations, we experimented with different safe directions magnitude values. The first approach was to use Hessian eigenvalues $\Lambda$ as the magnitude for safe directions vectors without the optimization in Algorithm 1 when computing $\hat{\Sigma}$. This simplifies the pipeline and improves the time complexity, however, results in an oversmoothing distribution, large diagonals for $\hat{\Sigma}_{ii}$, leading to worse performance. The next approach was to set $\Lambda$ as an initialization for $\hat{\Sigma}$ to get rid of DDS initial $\sigma$ calculation. While this does decrease the time complexity, it also resulted in oversmoothing, large diagonals for $\hat{\Sigma}$, worsening the performance. So our final decision was to use isotropic $\sigma$ as initialization for optimization similarly to ANCER as it showed the best results.

## 5. Limitations

As was mentioned in the [1], the main drawback of data-dependent certification is the variance of $\sigma$ which breaks the soundness of certification. The original solution to this problem (memory-based certification) was proposed in [1] and later modified in [9]. 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 and replace it with the straightforward hessian eigenvectors computation, but it is still a computationally expensive procedure. Secondly, the runtime of some of the particular memory-based procedures increased, for example, checking the intersection of rotated ellipsoids. And finally, we need to store the transformation matrices for certification. For CIFAR-10 with $32 \times 32 \times 3$ image sizes, the transformation matrix will have a size of $32 \times 32 \times 3 \times 32 \times 32 \times 3$. There are some potential improvement methods to reduce the memory consumption and speed up the process (for example using $k$-d trees), but they are outside the scope of this paper.

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. Conclusion

We succeeded in providing the theoretical extension to the anisotropic data dependent randomized smoothing and presenting its generalized counterpart - non-axis aligned anisotropic certification. To that end, we introduced RANCER, a new practical framework that optimizes the parameters of a data dependent non-axis aligned anisotropic smoothing distribution in order to certify larger regions than the axis aligned case. We experimentally validated our approach by obtaining $\ell_2$ and $\ell_2^\Sigma$ certification results on the CIFAR-10 dataset, achieving a state-of-the-art memory-based randomized smoothing classifier in that setting.

## 7. Acknowledgments

Authors thanks the Machine Learning Laboratory at the Ukrainian Catholic University for providing computing resources. Taras also expresses gratitude to Rostyslav Hryniv and Oles Dobosevych for their support. This work is partially funded by the EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (EP/S024050/1) and Five AI. The work is also partially supported by the UKRI grant: Turing AI Fellowship (EP/W002981/1) and the EPSRC/MURI grant (EP/N019474/1).

## References

[1] Motasem Alfarra, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Data dependent randomized smoothing. *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2022.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine Learning (ICML)*, 2018.

[3] Edward W Ayers, Francisco Eiras, Majd Hawasly, and Iain Whiteside. Parot: A practical framework for robust deep neural network training. In *NASA Formal Methods Symposium*. Springer, 2020.

[4] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Workshop on Artificial Intelligence and Security*, 2017.

[5] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations (ICLR)*, 2020.

[8] Ruediger Ehlers. Formal verification of piece-wise linear feed-forward neural networks. In *International Symposium on Automated Technology for Verification and Analysis*, 2017.

[9] Francisco Eiras, Motasem Alfarra, Philip Torr, M. Pawan Kumar, Puneet K. Dokania, Bernard Ghanem, and Adel Bibi. ANCER: Anisotropic certification via sample-wise volume maximization. *Transactions on Machine Learning Research (TMLR)*, 2022.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[12] Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. *Neural Information Processing Systems (NeurIPS)*, 2017.

[13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[14] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations (ICLR)*, 2016.

[15] H. O. Lancaster and M. G. Kendall. *A Course in the Geometry of n Dimensions*. Charles Griffin, 1962.

[16] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[17] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*, 2019.

[18] Alexander Levine and Soheil Feizi. Robustness certificates for sparse adversarial attacks by randomized ablation. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018.

[20] Scott McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg Corrado, Ara Darzi, Mozziyar Etemadi, Florencia Garcia-Vicente, Fiona Gilbert, Mark Halling-Brown, Demis Hassabis, Sunny Jansen, Alan Karthikesalingam, Christopher Kelly, Dominic King, and Shravya Shetty. International evaluation of an ai system for breast cancer screening. *Nature*, 2020.

[21] Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Higher-order certification for randomized smoothing. *Neural Information Processing Systems (NeurIPS)*, 2020.

[22] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[23] Lluís Ros, Assumpta Sabater, and Federico Thomas. An ellipsoidal calculus based on propagation and fusion. *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 2002.

[24] Hadi Salman, Greg Yang, Jerry Li, Pengchuan Zhang, Huan Zhang, Ilya P. Razenshteyn, and Sébastien Bubeck. Provably robust deep learning via adversarially trained smoothed classifiers. *Neural Information Processing Systems (NeurIPS)*, 2019.

[25] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[26] Jiaye Teng, Guang-He Lee, and Yang Yuan. $\ell_1$ adversarial robustness certificates: a randomized smoothing approach, 2020.

[27] Vincent Tjeng, Kai Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. *International Conference on Learning Representations (ICLR)*, 2019.

[28] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Neural Information Processing Systems (NeurIPS)*, 2018.

[29] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning (ICML)*, 2018.

[30] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. *International Conference on Learning Representations (ICLR)*, 2020.