Towards Strong Certified Defense with Universal Asymmetric Randomization

Hanbin Hong, *University of Connecticut*
Ashish Kundu, *Cisco Research*
Ali Payani, *Cisco Research*
Binghui Wang, *Illinois Institute of Technology*
Yuan Hong, *University of Connecticut*

***Abstract*—Randomized smoothing has become essential for achieving certified adversarial robustness in machine learning models. However, current methods primarily use isotropic noise distributions that are uniform across all data dimensions, such as image pixels, limiting the effectiveness of robustness certification by ignoring the heterogeneity of inputs and data dimensions. To address this limitation, we propose UCAN: a novel technique that Universally Certifies adversarial robustness with Anisotropic Noise. UCAN is designed to enhance any existing randomized smoothing method, transforming it from symmetric (isotropic) to asymmetric (anisotropic) noise distributions, thereby offering a more tailored defense against adversarial attacks. Our theoretical framework is versatile, supporting a wide array of noise distributions for certified robustness in different ℓ_p-norms and applicable to any arbitrary classifier by guaranteeing the classifier's prediction over perturbed inputs with provable robustness bounds through tailored noise injection. Additionally, we develop a novel framework equipped with three exemplary noise parameter generators (NPGs) to optimally fine-tune the anisotropic noise parameters for different data dimensions, allowing for pursuing different levels of robustness enhancements in practice. Empirical evaluations underscore the significant leap in UCAN's performance over existing state-of-the-art methods, demonstrating up to 182.6% improvement in certified accuracy at large certified radii on MNIST, CIFAR10, and ImageNet datasets.**[1]

***Index Terms*—Certified Robustness, Adversarial Robustness**

I. Introduction

Deep learning models have demonstrated remarkable performance across a wide range of applications. However, they are notoriously vulnerable to adversarial perturbations—carefully crafted minor modifications to inputs can lead to severe misclassification or misrecognition [9, 24]. These adversarial examples pose significant threats in real-world scenarios, such as autonomous driving [53], medical diagnosis [43], and face recognition systems [15], where even small prediction errors can result in catastrophic consequences.

To mitigate these vulnerabilities, robust defense mechanisms with certified guarantees are highly desirable. While empirical defense methods, including adversarial training [44, 50], perturbation destruction [60, 64], and feature regularization [61, 66], have shown promise, they often fall short against adaptive and stronger adversaries [4, 14, 62]. These methods typically cannot provide formal guarantees of robustness, as their defenses can be broken by more sophisticated attacks.

In contrast, certified robustness methods [12, 39, 58] offer provable guarantees by ensuring that no adversarial perturbation within a specified boundary—typically defined by an ℓ_p-norm ball (e.g., ℓ_1, ℓ_2, or ℓ_∞)—can alter the model's prediction. Among these, randomized smoothing has emerged as a state-of-the-art (SOTA) technique due to its universal applicability to any arbitrary classifier [12, 39, 54]. By injecting noise to data during both the training and inference phases, randomized smoothing transforms any classifier into a smoothed classifier with certified robustness guarantees.

Despite its success, existing randomized smoothing methods predominantly rely on isotropic noise distributions, where the same noise parameters are uniformly applied across all data dimensions (e.g., all pixels in an image). This uniform approach overlooks the inherent heterogeneity of different data dimensions, potentially limiting the effectiveness of robustness certification. For instance, using identical noise for all pixels might not be optimal, as certain pixels may be more critical to the model's decision-making process than others. Consequently, the *accuracy* of the smoothed classifier on both perturbed and clean inputs may be compromised, and the *certified radius* based on the ℓ_p-norm ball might not fully capture the true robustness potential. In reality, the ℓ_p-norm ball serves as a *sufficient but not necessary* condition for certification, as some regions outside the ℓ_p-norm ball can still maintain robustness.

To address these limitations, we introduce UCAN: a novel technique that Universally Certifies adversarial robustness with Anisotropic Noise. UCAN significantly enhances any existing randomized smoothing method by transitioning from symmetric (isotropic) to asymmetric (anisotropic) noise distributions. This tailored noise injection allows for more effective and adaptive defenses against adversarial attacks by assigning different noise parameters to various data dimensions based on their specific characteristics, importance, and vulnerability.

Developing UCAN involves overcoming two primary and complex challenges:

1) **Universal Certification Guarantee**: Establishing a robust theoretical foundation that universally supports various anisotropic noise distributions. This framework must ensure strict and sound certified robustness guarantees when different means and variances are assigned to different data dimensions.
2) **Optimal Noise Parameterization**: Designing mechanisms to derive appropriate means and variances for generating anisotropic noise tailored to various data dimensions. This ensures maximal certification performance without

[1]Code is anonymously available at https://github.com/youbin2014/UCAN/

compromising robustness.

To tackle these challenges, UCAN offers the following key contributions:

1) **Universal Certification Theory for Anisotropic Noise**. We present a universal theory for certifying the robustness of randomized smoothing with any anisotropic noise distribution. This theory seamlessly transforms existing certifications based on isotropic noise into those with anisotropic noise, supporting various ℓ_p-norm perturbations (e.g., ℓ_1, ℓ_2, ℓ_∞) and ensuring sound certified robustness across different noise distributions.
2) **Customizable Anisotropic Noise Parameter Generators (NPGs)**. We design three distinct NPGs, including two novel neural network-based methods, to efficiently generate the element-wise hyper-parameters (mean and variance) of the anisotropic noise distributions for all data dimensions (e.g., image pixels). These NPGs provide different efficiency-optimality trade-offs, significantly amplifying the certification performance while ensuring certified robustness.
3) **Significantly Boosted Certification Performance**. Empirical evaluations on benchmark datasets (MNIST, CIFAR10, and ImageNet) demonstrate that UCAN drastically outperforms SOTA randomized smoothing-based certified robustness methods. Specifically, UCAN achieves up to 182.6% improvement in certified accuracy at large certified radii compared to existing methods.

In summary, UCAN represents a significant advancement in certified robustness by introducing anisotropic noise into randomized smoothing. This approach not only enhances theoretical robustness guarantees but also provides practical mechanisms for optimizing noise parameters, leading to substantial improvements in certification across various datasets.

The remainder of this paper is organized as follows. Section II introduces the preliminaries, including the threat model and the isotropic randomized smoothing method for anisotropic certified robustness. Section III presents the proposed universal theory and metrics for robustness region. Section IV details the three different methods for customizing anisotropic noise to enhance certification performance. Section V discusses and proves the soundness of the certification-wise anisotropic noise. Section VI provides experimental results demonstrating UCAN's superior performance. Finally, Sections VII and VIII discuss related work and conclude the paper, respectively.

II. Preliminaries

In this section, we provide a brief overview of randomized smoothing with isotropic noise for certified robustness, which forms the foundation for our proposed UCAN framework.

A. Randomized Smoothing and Certified Robustness

Randomized smoothing is a technique that constructs a smoothed classifier from an arbitrary base classifier by adding random noise to the inputs. The smoothed classifier inherits certain robustness properties, allowing for certified guarantees against adversarial perturbations within a specific norm ball.

Consider a classification task where inputs $x \in \mathbb{R}^d$ are mapped to labels in a finite set of classes $\mathcal{C}$. Given any base classifier $f : \mathbb{R}^d \to \mathcal{C}$, randomized smoothing defines a smoothed classifier g that, for any input x, outputs the class most likely to be predicted by f when noise is added to x. Specifically, let $\epsilon \in \mathbb{R}^d$ be noise drawn from an arbitrary isotropic probability distribution ϕ (e.g., $\mathcal{N}(0,1)$), and define the random variable $X = x + \epsilon$. The smoothed classifier g is then defined as:

$$g(x) = \arg\max_{c \in \mathcal{C}} \mathbb{P}(f(X) = c) \tag{1}$$

B. Certified Robustness via Isotropic Randomized Smoothing

Randomized smoothing provides a way to certify the robustness of the smoothed classifier g against adversarial perturbations measured in terms of the ℓ_p-norm. The core idea is that if the smoothed classifier predicts a class c_A with high probability, and all other classes have significantly lower probabilities, then small perturbations to the input will not change the predicted class. We summarize existing results on certified robustness via randomized smoothing with isotropic noise in the following unified theorem.

Theorem 1 (**Certified Robustness via Randomized Smoothing with Isotropic Noise**). *Let $f : \mathbb{R}^d \to \mathcal{C}$ be any (possibly randomized) base classifier, and let ϕ be an isotropic probability distribution used to generate noise $\epsilon \in \mathbb{R}^d$. Define the smoothed classifier g as in Equation* (1)*. For a specific input $x \in \mathbb{R}^d$, let $X = x + \epsilon$, and suppose that there exists a class $c_A \in \mathcal{C}$ and bounds $\underline{p_A}, \overline{p_B} \in [0,1]$ such that:*

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

Then, for any perturbation $\delta \in \mathbb{R}^d$ where $\|\delta\|_p < R(\underline{p_A}, \overline{p_B})$, the smoothed classifier g will consistently predict class c_A at $x + \delta$, i.e., $g(x + \delta) = c_A$. Here, $\|\cdot\|_p$ denotes the ℓ_p-norm (with $p = 1, 2, \infty$), and $R(\underline{p_A}, \overline{p_B})$ is the certified radius, which depends on the noise distribution ϕ and the norm ℓ_p.

The certified radius $R(\underline{p_A}, \overline{p_B})$ quantifies the robustness of the smoothed classifier g around the input x. It ensures that any adversarial perturbation δ with $\|\delta\|_p < R(\underline{p_A}, \overline{p_B})$ cannot change the predicted class. The exact form of R varies depending on the noise distribution ϕ and the norm ℓ_p. For example, when ϕ is an isotropic Gaussian distribution with standard deviation σ, and the ℓ_2 norm is considered, the certified radius is given by [12]:

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

where Φ^{-1} is the inverse cumulative distribution function (CDF) of the standard normal distribution.

C. Limitations of Isotropic Noise in Randomized Smoothing

While randomized smoothing with isotropic noise provides a powerful tool for certified robustness, it applies the same noise distribution uniformly across all data dimensions. This isotropic

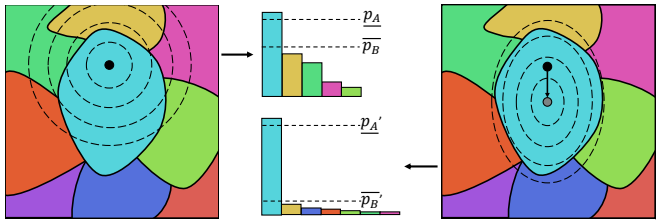


Fig. 1: Anisotropic noise vs isotropic noise. The decision regions of f are denoted in different colors. The dashed lines are the level sets of the noise distribution. The left figure shows the RS with isotropic Gaussian noise $\mathcal{N}(0, \lambda^2\mathbf{I})$ in [12] whereas the right figure shows the RS with anisotropic Gaussian noise $\mathcal{N}(\mu, \Sigma)$, where $\Sigma = \lambda^2 diag(\sigma_1^2, \sigma_2^2, ..., \sigma_d^2)$. Certification can be improved by enlarging the gap between $\underline{p_A}$ and $\overline{p_B}$.

approach may not fully exploit the potential robustness, as it ignores the heterogeneity and varying importance of different input features. Certain dimensions (e.g., pixels in an image) may be more sensitive or critical to the classification task, and treating them uniformly can limit the defense performance.

To address these limitations, our work extends randomized smoothing to anisotropic noise distributions, where different noise parameters can be assigned to different data dimensions. This extension poses significant challenges in developing universal theoretical guarantees and in designing efficient methods to optimize the noise parameters. In the following sections, we present our proposed UCAN framework, which overcomes these challenges to enhance certified robustness.

III. Certified Robustness with Anisotropic Noise

In this section, we establish a universal theory for the certification via randomized smoothing with *anisotropic* noise. Given any isotropic randomized smoothing methods, our method can universally transform them to anisotropic randomized smoothing for enhanced certified robustness.

A. General Anisotropic Noise

Given an arbitrary isotropic noise ϵ, we define the corresponding anisotropic noise ϵ' as:

$$\epsilon' = \epsilon^\top \Sigma + \mu \tag{4}$$

where $\Sigma \in \mathbb{R}^{d\times d}$ is an *invertible* covariance matrix, and $\mu \in \mathbb{R}^d$ is the mean offset vector. The covariance matrix Σ introduces dependencies between different dimensions of the noise, capturing potential correlations, while μ allows for mean shifts in the noise distribution. See Figure 8 for examples.

Then, certified robustness with anisotropic noise can be ensured per Theorem 2.

Theorem 2 (Asymmetric Randomized Smoothing via Universal Transformation). *Let $f : \mathbb{R}^d \to \mathcal{C}$ be any deterministic or randomized function. Suppose that for the multivariate random variable with isotropic noise $X = x + \epsilon$ in Theorem 1, the certified radius function is $R(\cdot)$. Then, for the corresponding anisotropic input $Y = x + \epsilon^\top \Sigma + \mu$, if there exist $c_A' \in \mathcal{C}$ and $\underline{p_A}', \overline{p_B}' \in [0, 1]$ such that:*

$$\mathbb{P}\left(f(Y) = c_A'\right) \geq \underline{p_A}' \geq \overline{p_B}' \geq \max_{c\neq c_A'} \mathbb{P}\left(f(Y) = c\right) \tag{5}$$

then for the anisotropic smoothed classifier $g'(x + \delta') \equiv \arg\max_{c\in\mathcal{C}} \mathbb{P}\left(f(Y + \delta') = c\right)$, we can guarantee $g'(x+\delta') = c_A'$ for all perturbations $\delta' \in \mathbb{R}^d$ such that:

$$\left\|\Sigma^{-1}\delta'\right\|_p \leq R\left(\underline{p_A}', \overline{p_B}'\right) \tag{6}$$

provided that Σ is invertible.

Proof. See the detailed proof in Appendix A. □

The general anisotropic noise with covariance allows for capturing correlations between different dimensions of the input data. However, adopting this generalized anisotropic noise introduces certain practical limitations: 1) **Invertibility of Σ**: Our theory requires the covariance matrix Σ to be invertible. If Σ is not invertible, a small regularization term can be added to its diagonal to ensure invertibility and retain the validity of our certification guarantees. 2) **Computational Complexity**: Optimizing or learning a full covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$ is computationally intensive, especially for high-dimensional data (e.g., images with $d = 150{,}528$ for ImageNet). The memory and computational requirements scale quadratically with the input dimension, making it impractical for large-scale applications.

Given these limitations, in practice, one might consider structured covariance matrices that balance expressiveness with computational efficiency, such as low-rank approximations, block-diagonal matrices, or sparse covariance matrices. In this paper, we focus on the independent anisotropic noise where noise parameters are independent along dimensions.

Special Case: Diagonal Covariance Matrix. The case where Σ is a diagonal matrix corresponds to anisotropic noise with independent dimensions (i.e., no covariance between dimensions). Let $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$, where $\sigma_i > 0$ for all i. In this case, $\Sigma^{-1} = \text{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_d}\right)$.

Corollary 3 (Asymmetric Randomized Smoothing with Independent Anisotropic Noise). *Under the same conditions as Theorem 2 if $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_d)$, then the certified robustness guarantee becomes:*

$$\left\|\delta' \oslash \sigma\right\|_p \leq R\left(\underline{p_A}', \overline{p_B}'\right) \tag{7}$$

where $\sigma = [\sigma_1, \sigma_2, \ldots, \sigma_d]^\top$, $\oslash$ denotes element-wise division, and $\delta' \in \mathbb{R}^d$ is the perturbation in the anisotropic space.

Proof. Since Σ is diagonal, its inverse is also diagonal with entries $1/\sigma_i$. Therefore, we have:

$$\left\|\Sigma^{-1}\delta'\right\|_p = \left\|\delta' \oslash \sigma\right\|_p \tag{8}$$

Substituting this into Equation (6) yields the desired result. □

Theorem 2 provides a robustness guarantee with anisotropic noise, building on traditional isotropic RS theories. Equation (6) from this theorem is both explicit and widely applicable,

TABLE I: Certified radii (binary-case) for randomized smoothing with independent isotropic and anisotropic noise. d is the dimension size. Φ^{-1} is the inverse CDF of Gaussian distribution. λ is the scalar parameter of the isotropic noise. σ is the anisotropic scale multiplier.

Distribution	PDF	Adv.	Isotropic Guarantee	ℓ_p Radius for Iso. RS	Anisotropic Guarantee	ℓ_p Radius for Ani. RS
Gaussian [12]	$\propto e^{-\|\frac{z}{\lambda}\|_2^2}$	ℓ_2	$\|\|\delta\|\|_2 \leq \lambda(\Phi^{-1}(\underline{p_A}))$	$\lambda(\Phi^{-1}(\underline{p_A}))$	$\|\|\delta \oslash \sigma\|\|_2 \leq \lambda(\Phi^{-1}(\underline{p_A}'))$	$\min\{\sigma\}\lambda(\Phi^{-1}(\underline{p_A}))$
Gaussian [65]	$\propto e^{-\|\frac{z}{\lambda}\|_2^2}$	ℓ_1	$\|\|\delta\|\|_1 \leq \lambda(\Phi^{-1}(\underline{p_A}))$	$\lambda(\Phi^{-1}(\underline{p_A}))$	$\|\|\delta \oslash \sigma\|\|_1 \leq \lambda(\Phi^{-1}(\underline{p_A}'))$	$\min\{\sigma\}\lambda(\Phi^{-1}(\underline{p_A}))$
		ℓ_∞	$\|\|\delta\|\|_\infty \leq \lambda(\Phi^{-1}(\underline{p_A}))/\sqrt{d}$	$\lambda(\Phi^{-1}(\underline{p_A}))/\sqrt{d}$	$\|\|\delta \oslash \sigma\|\|_\infty \leq \lambda(\Phi^{-1}(\underline{p_A}'))/\sqrt{d}$	$\min\{\sigma\}\lambda(\Phi^{-1}(\underline{p_A}))/\sqrt{d}$
Laplace [54]	$\propto e^{-\|\frac{z}{\lambda}\|_1}$	ℓ_1	$\|\|\delta\|\|_1 \leq -\lambda\log(2(1-\underline{p_A}))$	$-\lambda\log(2(1-\underline{p_A}))$	$\|\|\delta \oslash \sigma\|\|_1 \leq -\lambda\log(2(1-\underline{p_A}'))$	$-\min\{\sigma\}\lambda\log(2(1-\underline{p_A}))$
Exp. ℓ_∞ [65]	$\propto e^{-\|\frac{z}{\lambda}\|_\infty}$	ℓ_1	$\|\|\delta\|\|_1 \leq 2d\lambda(\underline{p_A}-\frac{1}{2})$	$2d\lambda(\underline{p_A}-\frac{1}{2})$	$\|\|\delta \oslash \sigma\|\|_1 \leq 2d\lambda(\underline{p_A}'-\frac{1}{2})$	$2\min\{\sigma\}d\lambda(\underline{p_A}-\frac{1}{2})$
		ℓ_∞	$\|\|\delta\|\|_\infty \leq \lambda\log(\frac{1}{2(1-\underline{p_A})})$	$\lambda\log(\frac{1}{2(1-\underline{p_A})})$	$\|\|\delta \oslash \sigma\|\|_\infty \leq \lambda\log(\frac{1}{2(1-\underline{p_A}')})$	$\min\{\sigma\}\lambda\log(\frac{1}{2(1-\underline{p_A})})$
Uniform ℓ_∞ [41]	$\propto \mathbb{I}(\|z\|_\infty \leq \lambda)$	ℓ_1	$\|\|\delta\|\|_1 \leq 2\lambda(\underline{p_A}-\frac{1}{2})$	$2\lambda(\underline{p_A}-\frac{1}{2})$	$\|\|\delta \oslash \sigma\|\|_1 \leq 2\lambda(\underline{p_A}'-\frac{1}{2})$	$2\min\{\sigma\}\lambda(\underline{p_A}-\frac{1}{2})$
		ℓ_∞	$\|\|\delta\|\|_\infty \leq 2\lambda(1-\sqrt[d]{\frac{3}{2}-\underline{p_A}})$	$2\lambda(1-\sqrt[d]{\frac{3}{2}-\underline{p_A}})$	$\|\|\delta \oslash \sigma\|\|_\infty \leq 2\lambda(1-\sqrt[d]{\frac{3}{2}-\underline{p_A}'})$	$2\min\{\sigma\}\lambda(1-\sqrt[d]{\frac{3}{2}-\underline{p_A}})$
Power Law ℓ_∞ [65]	$\propto \frac{1}{(1+\|\frac{z}{\lambda}\|_\infty)^\alpha}$	ℓ_1	$\|\|\delta\|\|_1 \leq \frac{2d\lambda}{a-d}(\underline{p_A}-\frac{1}{2})$	$\frac{2d\lambda}{a-d}(\underline{p_A}-\frac{1}{2})$	$\|\|\delta \oslash \sigma\|\|_1 \leq \frac{2d\lambda}{a-d}(\underline{p_A}-\frac{1}{2})$	$\min\{\sigma\}\frac{2d\lambda}{a-d}(\underline{p_A}-\frac{1}{2})$

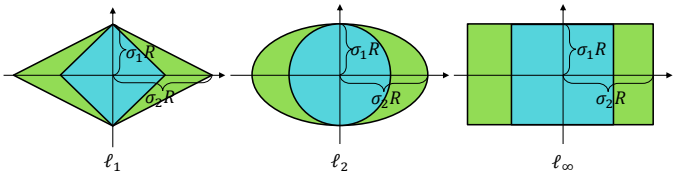


Fig. 2: An example illustration of the full certified region (green) and the certified radius (blue), where $\sigma_1 < \sigma_2$.

significantly enhancing existing RS frameworks' performance. In Table I, we outline some representative isotropic RS methods and their extension to anisotropic noise via Theorem 2. Our method can also be readily adapted to other RS methods like those in [29, 69] that lack explicit radius certifications, applying it directly to their numerical results. Details on the binary classifier version are discussed in Appendix A.

In Figure 1, we clearly illustrate the significant benefits that the anisotropic noise can bring to randomized smoothing. In Theorem 2, we observe that the mean offset μ does not affect the derivation of the certified robustness with anisotropic noise (Equation (6)). Thus, it is likely that the gap between probabilities $\underline{p_A}'$ and $\overline{p_B}'$ can be improved by a properly chosen mean offset of the anisotropic noise (without affecting the robustness guarantee). Additionally, with the heterogeneous variance, the anisotropic noise can better fit the different dimensions of the input without causing over-distortion.

Corollary 4. *For the anisotropic input Y in Theorem 2, if Equation (5) is satisfied, then $g'(x+\delta') \equiv \arg\max_{c\in\mathcal{C}} \mathbb{P}(f(Y+\delta')=c) = c'_A$ for all $||\delta'||_p \leq R'$ such that*

$$R' = \min\{\sigma\}R \tag{9}$$

where R is the certified radius of randomized smoothing via isotropic noise, and $\min\{\cdot\}$ denotes the minimum entry.

Proof. The guarantee in Theorem 2 holds for $||\delta' \oslash \sigma||_p \leq R$. Since $||\delta' \oslash \sigma||_p \leq ||\frac{\delta'}{\min\{\sigma\}}||_p$, if $||\frac{\delta'}{\min\{\sigma\}}||_p \leq R$, the guarantee still holds. This requires $||\delta'||_p \leq \min\{\sigma\}R$. □

B. Certified Region and Two Metrics

In Corollary 4, we derive the certified radii R' for asymmetric randomized smoothing in the formation of ℓ_p-ball, i.e., $||\delta'||_p \leq R'$. While the certified radius reflects the maximum size of the tolerated perturbation within the ℓ_p-ball. However, especially under asymmetric circumstances, the certified region can be highly asymmetric, and the ℓ_p-ball (which is symmetric) may only represent a subset of the full robustness region, which is given by $(\sum_i^d(\frac{\delta'_i}{\sigma_i})^p)^{\frac{1}{p}} \leq R(\underline{p_A}', \overline{p_B}')$, as illustrated in Figure 2. Therefore, to provide a more comprehensive and complementary assessment, we adopt an additional metric to measure the overall size of the certified region.

Radius and ALM for Certified Region. In addition to the ℓ_p radius, we also adopt the Alternative Lebesgue Measure (ALM[2]) for measuring the certified region. Specifically, we consider the certified region under the anisotropic guarantee as a d-dimensional super-ellipsoid, as defined in Definition 5. It is worth noting that the ℓ_p-norm ball is a sufficient but not necessary condition for certified robustness: while robustness within the ℓ_p ball is guaranteed, certain points outside this ball may also remain robust due to the true shape of the certified region. The super-ellipsoid formulation can represent this broader space.

Definition 5 (d-dimensional Generalized Super-ellipsoid of δ). *The d-dimensional generalized super-ellipsoid ball of δ is defined as*

$$S(d,p) = \{(\delta_1, \delta_2, ..., \delta_d) : \sum_{i=1}^{d} |\frac{\delta_i}{\sigma_i R}|^p \leq 1, p > 0\} \tag{10}$$

Theorem 6 (Lebesgue Measure of the Robust Perturbation Set $S(d,p)$). *Define $S(d,p)$ per Definition 5, then the Lebesgue measure of the robust perturbation set is given by*

$$V_S(d,p) = \frac{(2R\Gamma(1+\frac{1}{p}))^d \prod_{i=1}^d \sigma_i}{\Gamma(1+\frac{d}{p})} \tag{11}$$

where Γ is the Euler gamma function defined in Definition 10 in Appendix B.

Proof. See the detail proof in Appendix C. □

Recall that we aim to find an auxiliary metric that can measure the volume of this super-ellipsoid, so we derive the Lebesgue Measure of this super-ellipsoid as in Theorem 6 since the Lebesgue Measure quantifies the "volume" of high-dimensional space, and then we simplify it by removing the constants w.r.t. the dimension and p, as well as transforming it to the radius scale. To this end, the ALM can be simplified as: $ALM = \sqrt[d]{\prod_{i=1}^d \sigma_i} R$.

[2]This metric is mathematically equivalent to the "proxy radius" in [20].

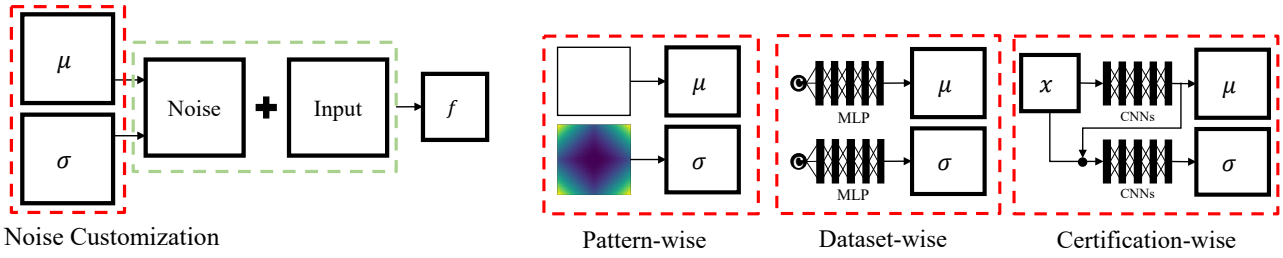


Fig. 3: Left: the framework for customizing anisotropic noise. Right: three example noise parameter generators (NPGs).

It serves as an additional measure for evaluating the robustness region. This is particularly useful because the ℓ_p radius (*as a more strict metric*) cannot fully capture the robustness region in certain dimensions due to its inherent symmetry, as illustrated in Figure 2. In summary, the ℓ_p radius serves as a conservative, worst-direction certificate—being controlled by the smallest anisotropic scale $\min_i \sigma_i$—whereas the ALM provides an auxiliary, volume-oriented characterization of the full certified region through the geometric mean $\left(\prod_{i=1}^{d} \sigma_i\right)^{1/d}$. It is worth noting that ALM subsumes the certified radius: in any dimension, the certified radius corresponds to the smallest semi-axis length of the certified region, while ALM (as the geometric mean of all semi-axes times R) is always greater than or equal to this value, and coincides with it in the isotropic case. Therefore, ALM serves as an auxiliary metric for the size of certified region instead of a formal guarantee. See Appendix B for detailed analysis and discussions for ALM.

IV. Customizing Anisotropic Noise

Theorem 2 formally guarantees robustness when heterogeneous noise parameters are assigned across different data dimensions. However, finding more optimal heterogeneous noise parameters rather than randomly assigning them remains a challenge. To this end, in UCAN, we design a unified framework to customize anisotropic noise for randomized smoothing (left in Figure 3), which includes three *noise parameter generators* (NPGs) with different scales of trainable parameters (right in Figure 3) and optimality levels.

Here, the "optimality" refers to the degree to which each NPG can optimize the noise parameters to maximize certified robustness (measured by either certified radius or ALM), while maintaining prediction accuracy.

- **Pattern-wise Anisotropic Noise** (Low optimality): Uses pre-defined patterns for noise variances, offering basic but non-adaptive and relatively lower optimal robustness.
- **Dataset-wise Anisotropic Noise** (Moderate optimality): Learns a global set of noise parameters optimized for the entire dataset, enabling some adaptation for different input data but still derived at the dataset level.
- **Certification-wise Anisotropic Noise** (High optimality): Generates noise parameters specifically tailored to each individual input at the certification time, achieving the most fine-grained and input-specific robustness optimization.

The optimality levels reflect a fundamental trade-off: higher-optimality approaches allow more precise and effective maximization of certified robustness, but require increased computational resources for training and/or inference. These three NPGs are representative examples—other designs are possible depending on application needs. In any case, the covariance matrix should be kept invertible (e.g., by adding a small positive constant to the diagonal if necessary). Implementation details and specific algorithms can be found in Appendix IV-D.

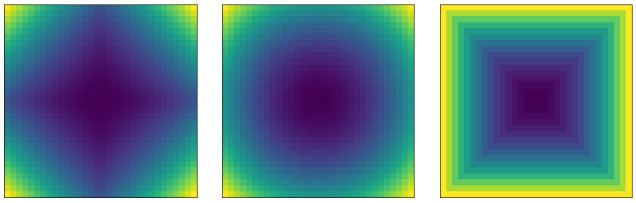


Fig. 4: Spatial distributions for noise variances (pattern-wise).

A. Pattern-wise Anisotropic Noise

The motivation for pattern-wise anisotropic noise in NPG is based on the understanding that different data regions affect predictions differently [22]. Typically, an image's center contains more critical visual information, requiring lower variance to preserve clarity, whereas the borders may accommodate higher variance without significantly impacting predictions. Thus, this NPG utilizes fixed spatial patterns for anisotropic noise.

Consider a function $\sigma(a, b)$ representing the variance distribution across an image, where (a, b) are the coordinates with the center at $(0, 0)$. The variance is intuitively smaller at the center than at the borders due to varying feature importance. We propose three distinct spatial distribution types:

$$\sigma(a, b) = \kappa||(a, b)||_p^2 + \iota, \quad p = 1, 2, \infty \tag{12}$$

where $||(a, b)||_p$ denotes the ℓ_p-norm distance between (a, b) and $(0, 0)$, κ is a constant parameter tuning the overall magnitude of the variance, ι denotes the variance of a center pixel since $\sigma(0, 0) = \iota$, and $\iota > 0$ such that $\sigma(a, b) > 0$. Figure 4 shows three example spatial distributions when $p = 1, 2, \infty$, where $\kappa = 1$ and $\iota = 1$. The noise mean is set as 0 to avoid unnecessary deviation in the images.

Pattern-wise anisotropic noise modifies predictions based on spatial contributions (e.g., pixels). Yet, without fine-tuning, this approach may impair performance, particularly with datasets' diverse characteristics. Then we propose an automated method to fine-tune the variance spatial distribution of anisotropic noise for each dataset.

B. Dataset-wise Anisotropic Noise

The NPG employs a constant-input neural network generator [13] to learn asymmetric variances during noise-based robust training (see Figure 3 right). The generator outputs tensors for anisotropic σ and μ, where μ_i and σ_i represent the mean offset and variance multiplier per pixel, respectively. Anisotropic noise $\epsilon' = \epsilon^\top \sigma + \mu$ is will be generated from base isotropic noise ϵ and added to the input for randomized smoothing.

Architecture. We adopt the generator architecture (a multi-layer perception) in Generative Adversarial Network (GAN) [23] to design a novel neural network generator. Different from GAN, this NPG does not depend on the input data but depends on the entire dataset. Therefore, we fixed the input as constants. Following [23], the NPG consists of 5 linear layers, and the first 4 of them are followed by activation layers. The output will be then transformed by a hyperbolic tangent function with an amplification factor γ, i.e., $\gamma tanh(\cdot)$. This amplified hyperbolic tangent layer limits the value of the variances since an infinite value in the noise parameters will crash the training. It is worth noting that for other tasks in Natural Language Processing or Audio Processing, other NPG structures, e.g., transformer, NNs, or even heuristic algorithms can be also designed to fit the targeted tasks.

Loss Function. The NPG and classifier can be trained together for optimal synergy. Designing an appropriate loss function is crucial for guiding NPG to desired convergence, targeting enhanced certified robustness via the ALM or certified radius. We aim to maximize either $\sqrt[d]{\prod \sigma} R$ or $\min\{\sigma\} R$, leading to two loss function variants that focus on increasing $\sqrt[d]{\prod \sigma}$ (equivalent to $mean\{\sigma\}$ in log scale) or $\min\{\sigma\}$, respectively. As the certified radius R is influenced by prediction accuracy against noise, enhancing this accuracy also boosts R, aligning with the smoothed classifier's training objectives.

$$\mathcal{L}(\theta_f, \theta_g) = \underbrace{-\mathit{mean/min}\{\sigma(\theta_g)\}}_{\textit{Variance Loss}} + \underbrace{\sum_{k=1}^{N} y_k \log \hat{y}_k(x + \epsilon^\top \sigma(\theta_g) + \mu(\theta_g), \theta_f, \theta_g)}_{\textit{Smoothing Loss}} \tag{13}$$

where the variance loss can be $\mathit{mean}\{\sigma(\theta_g)\}$ or $\mathit{min}\{\sigma(\theta_g)\}$ for improving ALM or certified radius, respectively. θ_f and θ_g denote the model parameters of the classifier and parameter generator, respectively, k denotes the prediction class, N represents the total number of classes, y_k denotes the label of input x, and $\hat{y}_k$ is the prediction of y_k. The training of the NPG for μ is also guided by the smoothing loss to improve the prediction over the dataset.

C. Certification-wise Anisotropic Noise

While dataset-wise anisotropic noise fine-tunes parameters during training for improved robustness, it doesn't fully account for the heterogeneity among different input samples. The certification also varies by input, being valid only for the certified input and the corresponding radius. Thus, we propose a certification-wise NPG that generates tailored anisotropic noise for each sample, considering heterogeneity across both inputs and data dimensions. Unlike dataset-wise noise, the parameter generators for μ and σ in certification-wise NPG are cascaded, not parallel (see right in Figure 3). The mean parameter generator first processes the input x to produce a μ map, followed by the variance generator using $x + \mu$ to generate a σ map. Then the smoothed classifier is based on the generated μ and σ through the classification.

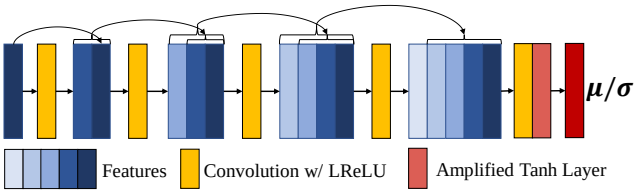


Fig. 5: Architecture of parameter generator for certification-wise anisotropic noise.

Architecture and Loss Function. This NPG learns the mapping from the image to the μ and σ maps, which is similar to the function of neural networks in image transformation tasks. Hence, inspired by the techniques used in image super-resolution [72], we also adopt the "dense blocks" [31] as the main architecture to design the NPG (see Figure 5). It consists of 4 convolutional layers followed by leaky-ReLU [63]. Similar to the generator in dataset-wise anisotropic noise, the output is rectified by the amplified hyperbolic tangent function to stabilize the overall training process. Note that our parameter generator is a relatively small network (5 layers), thus it can be plugged in before any classifier for generating the certification-wise anisotropic noise without consuming too many computing resources (see Section VI-E for a detailed discussion on running time). For both μ and σ, we train a corresponding parameter generator for each using the same architecture. The loss function is similar to that used for the dataset-wise anisotropic noise, but the NPG takes x as input and outputs μ and σ.

D. Practical Algorithms

Following Cohen et al. [12], we also use the Monte Carlo algorithm to bound the prediction probabilities of smoothed classifier and compute the ALM (certified region). Different from Cohen et al. [12], our noise distributions are either pre-assigned (as pattern-wise) or produced by the parameter generator (either dataset-wise or certification-wise). Our algorithms for certification and prediction using different noise generation methods are summarized in Algorithm 1 and 2 (w.l.o.g., taking the binary classifier as an example).

For simplicity of notations, the generation of anisotropic μ and σ are summarized by the noise generation method(s) M. In the case of pattern-wise anisotropic noise, M outputs pre-assigned fixed variance and zero-means; in case of dataset-wise and certification-wise anisotropic noise, M adopts the parameter generators to generate the mean and variance maps. In the certification (Algorithm 1), we select the top-1 class $\hat{c_A}$ by the CLASSIFYSAMPLES function, in which the base classifier outputs the prediction on the noisy input sampled

Algorithm 1 UCAN-Certification

Given: Base classifier f, anisotropic noise generation method M, input (e.g., image) x, number of Monte Carlo samples n_0 and n, confidence $1-\alpha$

μ, $\sigma \leftarrow M$
$counts\_select \leftarrow$ CLASSIFYSAMPLES(f, x, μ, σ, n_0)
$\hat{c}_A \leftarrow$ **top index in** $counts\_select$
$counts \leftarrow$ CLASSIFYSAMPLES(f, x, μ, σ, n)
$\underline{p_A}' \leftarrow$ LOWERCONFBOUND$(counts[\hat{c}_A], n, 1-\alpha)$
if $\underline{p_A}' > \frac{1}{2}$ **then**
    return prediction $\hat{c}_A$ and certified radius $\min\{\sigma\}$R/ALM $\sqrt[d]{\prod_i \sigma_i} R$
else
    return ABSTAIN
end if

Algorithm 2 UCAN-Prediction

Given: Base classifier f, anisotropic noise generation method M, input (e.g., image) x, number of Monte Carlo samples n, confidence $1-\alpha$

μ, $\sigma \leftarrow M$
$counts \leftarrow$ CLASSIFYSAMPLES(f, x, μ, σ, n)
$\hat{c}_A \leftarrow$ **top index in** $counts$
$n_A \leftarrow counts[\hat{c}_A]$
if BINOMIALPVALUE$(n_A, n, 0.5) \leq \alpha$ **then**
    return prediction $\hat{c}_A$
else
    return ABSTAIN
end if

from the noise distribution. Once the top-1 class is determined, classification will be executed on more samples and the LOWERCONFBOUND function will output the lower bound of the probability $\underline{p_A}'$ computed by the Binomial test. If $\underline{p_A}' > \frac{1}{2}$, we output the prediction class and the ALM (measuring the certified region). Otherwise, it outputs ABSTAIN. In the prediction (Algorithm 2), we also generate the noise and then compute the prediction counts over the noisy inputs. If the Binomial test succeeds, then it outputs the prediction class. Otherwise, it returns ABSTAIN.

V. SOUNDNESS ANALYSIS FOR CERTIFICATION-WISE ANISOTROPIC RANDOMIZED SMOOTHING

In this section, we address the potential soundness pitfall in existing input-dependent randomized smoothing methods and demonstrate how our certification-wise approach provides enhanced robustness guarantees. Specifically, we provide formal definitions, theorems, and detailed proofs to establish the soundness of our method. Additionally, we carefully explain why randomized smoothing inherently depends on the clean input for reliable certification.

A. Potential Soundness Pitfall in Existing Input-Dependent Randomized Smoothing Methods

Recent attempts to enhance randomized smoothing have introduced input-dependent noise parameters that vary with both the input x and potential adversarial perturbations δ [20, 47]. In these methods, the noise parameters $\mu(x, \delta)$ and $\sigma(x, \delta)$ are functions of both the clean input and the perturbation, leading to noise distributions that change based on the adversary's actions.

While such approaches aim to tailor the noise to each input and perturbation, it introduce potential concerns in the certified robustness guarantee. Specifically, when certifying a sample x, the noise is generated as $\epsilon(x)$, depending on x. However, when applying the guarantee to a potentially perturbed sample $x+\delta$, the noise changes to $\epsilon(x+\delta)$, which likely follows a different distribution from $\epsilon(x)$. Randomized smoothing requires that the prediction on the certified input x and the perturbed input $x+\delta$ be based on the same smoothed classifier with the same noise distribution [12]. Therefore, this input-dependent noise may affect the soundness of randomized smoothing. As noted in [20, 47], an additional component (e.g., fixing the distribution or adopting memory-based certification) has been utilized to maintain valid guarantees by sacrificing the system performance, as the "price paid for soundness".

B. Certification-wise Anisotropic Randomized Smoothing

Different from [20, 47], in our *certification-wise* anisotropic randomized smoothing method, the noise parameters $\mu(x)$ and $\sigma(x)$ depend solely on the clean input x and remain the same when applied to any potential perturbed samples during certification. This design ensures that the smoothed classifier remains consistent across all perturbations applied to x, inherently preserving the soundness of the certified robustness guarantee.

After generating the fine-tuned noise on the clean input x (to be certified), our method constructs a robustness region that conceptually ensures robustness for x against any perturbation δ within the certified radius, rather than actually injecting noise into the perturbed inputs. It is worth noting that the theorems in this paper universally work for isotropic and anisotropic noise independent of the noise generation method, which is sound for any randomized smoothing method. Furthermore, the "pattern-wise" and "dataset-wise" noise in Section IV have no potential concerns regarding this soundness problem.

C. Certified Robustness Guarantee on Certification-wise Noise

We now formally establish the soundness of our method by proving that the certified robustness guarantee holds under our certification-wise anisotropic noise.

Theorem 7 (Certified Robustness of Certification-Wise Anisotropic Randomized Smoothing)**.** *Let $f : \mathbb{R}^d \to \mathcal{C}$ be any deterministic or randomized base classifier. Suppose that for the random variable $X = x + \epsilon$, where ϵ is drawn from an isotropic distribution, the certified radius function is $R(\cdot)$. Define the anisotropic random variable $Y = x + \epsilon^\top \sigma(x) + \mu(x)$, where $\sigma(x) \in \mathbb{R}^d$ and $\mu(x) \in \mathbb{R}^d$ are functions of x only. If there exist $c_A \in \mathcal{C}$ and bounds $\underline{p_A}, \overline{p_B} \in [0,1]$ such that:*

$$\mathbb{P}_\epsilon\left(f(Y) = c_A\right) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}_\epsilon\left(f(Y) = c\right) \tag{14}$$

then, for any perturbation $\delta \in \mathbb{R}^d$ satisfying:

$$\|\delta \oslash \sigma(x)\|_p \leq R\left(\underline{p_A}, \overline{p_B}\right) \tag{15}$$

the smoothed classifier g will consistently predict class c_A at $x+\delta$, i.e., $g(x+\delta)=c_A$. Here, $\oslash$ denotes element-wise division, and $\|\cdot\|_p$ denotes the ℓ_p norm.

D. Proof of Theorem 7

Proof. Let $X = x + \epsilon$, where $\epsilon \in \mathbb{R}^d$ follows an isotropic noise distribution. Define the anisotropic random variable $Y = x + \epsilon^\top \sigma(x) + \mu(x)$.

Define a transformation $h_x : \mathbb{R}^d \to \mathbb{R}^d$ specific to input x:

$$h_x(z) = z^\top \sigma(x) + \mu(x) \tag{16}$$

where the multiplication and addition are element-wise.

Given any deterministic or randomized function $f : \mathbb{R}^d \to \mathcal{C}$, consider the composed classifier $f' = f \circ h_x$, mapping $z \mapsto f(h_x(z))$.

Under the transformation, the condition on the class probabilities becomes:

$$\mathbb{P}_\epsilon\left(f'(X) = c_A\right) = \mathbb{P}_\epsilon\left(f(h_x(X)) = c_A\right) \tag{17}$$

$$= \mathbb{P}_\epsilon\left(f(Y) = c_A\right) \geq \underline{p_A} \tag{18}$$

$$\max_{c \neq c_A} \mathbb{P}_\epsilon\left(f'(X) = c\right) = \max_{c \neq c_A} \mathbb{P}_\epsilon\left(f(Y) = c\right) \leq \overline{p_B} \tag{19}$$

Thus, the probability bounds required for certification are satisfied by f' under isotropic noise ϵ.

From the standard randomized smoothing theory (e.g., Theorem 1), since the transformed classifier f' satisfies the probability bounds with respect to the isotropic noise ϵ, we have that for any perturbation $\delta' \in \mathbb{R}^d$ satisfying $\|\delta'\|_p \leq R(\underline{p_A}, \overline{p_B})$, the prediction of f' remains constant:

$$\arg\max_{c \in \mathcal{C}} \mathbb{P}_\epsilon\left(f'(x + \delta' + \epsilon) = c\right) = c_A \tag{20}$$

Now, consider a perturbation $\delta \in \mathbb{R}^d$ in the original input space such that $\delta' = \delta \oslash \sigma(x)$. Then, the perturbed input after transformation is:

$$h_x(x+\delta) = (x+\delta)^\top \sigma(x) + \mu(x) \tag{21}$$

$$= x^\top \sigma(x) + \delta^\top \sigma(x) + \mu(x) \tag{22}$$

$$= h_x(x) + \delta^\top \sigma(x) \tag{23}$$

Substituting back, we have:

$$g(x+\delta) = \arg\max_{c \in \mathcal{C}} \mathbb{P}_\epsilon\left(f\left(x + \delta + \epsilon^\top \sigma(x) + \mu(x)\right) = c\right) \tag{24}$$

$$= \arg\max_{c \in \mathcal{C}} \mathbb{P}_\epsilon\left(f\left(h_x(x) + \delta^\top \sigma(x) + \epsilon^\top \sigma(x)\right) = c\right) \tag{25}$$

Since $\delta^\top \sigma(x) + \epsilon^\top \sigma(x) = (\delta + \epsilon)^\top \sigma(x)$, we have:

$$g(x+\delta) = \arg\max_{c \in \mathcal{C}} \mathbb{P}_\epsilon\left(f\left(h_x(x) + (\delta + \epsilon)^\top \sigma(x)\right) = c\right) \tag{26}$$

$$= \arg\max_{c \in \mathcal{C}} \mathbb{P}_\epsilon\left(f'\left(x + \delta' + \epsilon\right) = c\right) \tag{27}$$

By the robustness of f' under isotropic noise (Equation (20)), we have $g(x+\delta) = c_A$ whenever $\|\delta'\|_p = \|\delta \oslash \sigma(x)\|_p \leq R(\underline{p_A}, \overline{p_B})$.

Therefore, the smoothed classifier g maintains its prediction c_A within the certified region defined by the anisotropic noise parameters, establishing the soundness of our method. □

E. Why Randomized Smoothing Depends on the Clean Input

Randomized smoothing inherently depends on the clean input x because the certification process aims to guarantee the classifier's robustness for that specific input. The certified radius $R(\underline{p_A}, \overline{p_B})$ is computed based on the class probabilities at x, which are estimated using noise added to x. Consequently, the smoothed classifier g and the corresponding robustness guarantee are tied to the clean input.

In our method, the noise parameters $\sigma(x)$ and $\mu(x)$ are functions of x, further emphasizing this dependency. By designing the noise to be input-specific but independent of perturbations, we ensure that the certification process accurately reflects the classifier's behavior around the clean input, providing a meaningful and sound robustness guarantee.

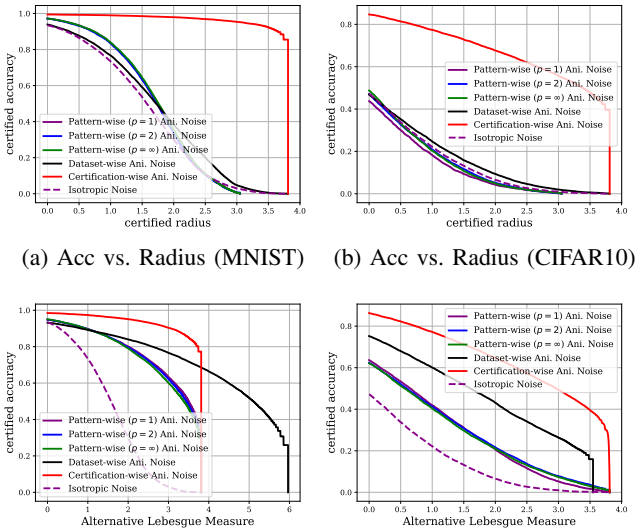
VI. Experiments

We present a comprehensive evaluation of UCAN in this section. Specifically, in Section VI-A, UCAN's performance with three anisotropic noise types is tested against isotropic noise baselines. Section VI-B assesses UCAN's universality regarding noise distributions and resistance to various ℓ_p perturbations. In Section VI-C, we compare UCAN's top performance with state-of-the-art randomized smoothing methods. Section VI-D and VI-E present the visualization and the efficiency. Additional experiments are detailed in Appendix D.

Metrics. Existing randomized smoothing methods often adopt the *approximate certified test set accuracy* from [12], defined as the proportion of the test set correctly certified above a radius R. Besides, we also report certified accuracy based on the ALM, representing the fraction of the test set certified correctly with *at least ALM*. Formally, for asymmetric RS, we define $Acc(\min\{\sigma\}R) = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{[g'(x^j+\delta)=y^j]}, \forall \quad ||\delta||_p \leq \min\{\sigma\}R$ and $Acc(ALM) = \frac{1}{N}\sum_{j=1}^{N} \mathbf{1}_{[g'(x^j+\delta)=y^j]}, \forall\ ||\delta||_p \leq \sqrt[d]{\prod \sigma_i} R$. In isotropic RS, these metrics converge to $Acc(R)$.

To fairly position our methods, when compared to the SOTA methods, we present the certified accuracy w.r.t. the certified radius and optionally w.r.t. ALM.

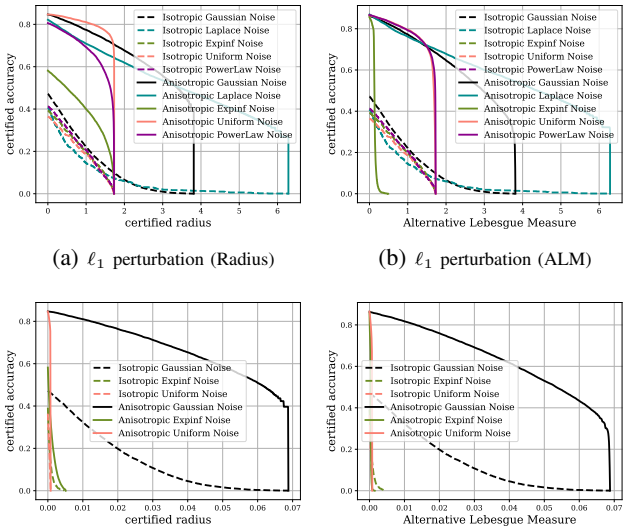
Experimental Settings. All the experiments are performed on three datasets: MNIST [38], CIFAR10, [37] and ImageNet [48]. Following [12], we obtain the certified accuracy on the entire test set in CIFAR10 and MNIST while randomly picking 500 samples in the test set of ImageNet; we set $\alpha = 0.001$ and the numbers of Monte Carlo samples $n_0 = 100$ and $n = 100,000$. We use the original size of the images in MNIST and CIFAR10, i.e., 28×28 and $3 \times 32 \times 32$, respectively. For the ImageNet dataset, we resize the images to $3 \times 224 \times 224$. In the training, we train the base classifier and the parameter generator (if needed) with all the training set in three datasets. For the MNIST dataset, we use a simple two-layer CNN as the base classifier. For the CIFAR10 and ImageNet datasets, we use the ResNet110 and ResNet50 [27] as the base classifier, respectively. Dataset-wise NPG uses a 5-layer MLP, and certification-wise NPG uses a 4-layer CNN, both trained with Adam optimizer (learning rate



(a) Acc vs. Radius (MNIST) (b) Acc vs. Radius (CIFAR10)

(c) Acc vs. ALM (MNIST) (d) Acc vs. ALM (CIFAR10)

Fig. 6: RS with anisotropic noise vs. baseline [12] with isotropic noise (Gaussian for ℓ_2) – UCAN gives significantly better certified accuracy and larger certified radius/ALM.



(a) ℓ_1 perturbation (Radius) (b) ℓ_1 perturbation (ALM)

(c) ℓ_∞ perturbation (Radius) (d) ℓ_∞ perturbation (ALM)

Fig. 7: UCAN (certification-wise anisotropic) vs. isotropic randomized smoothing: different noise PDFs against various ℓ_p perturbations on CIFAR10.

1×10^{-2}, batch size 128, 200 epochs). Noise parameters are constrained to be positive.

Experimental Environment. All the experiments were performed on the NSF Chameleon Cluster [36] with Intel(R) Xeon(R) Gold 6230 CPU @ 2.10GHz, 128G RAM, and Tesla V100 SXM2 32GB.

A. Anisotropic vs Isotropic Noise in Randomized Smoothing

We first evaluate randomized smoothing with anisotropic noise generated by the three example NPGs. W.l.o.g., we adopt the most common setting as default: Gaussian distribution against ℓ_2 perturbations, compare with the isotropic Gaussian baseline [12] (with zero-mean), which derives the tight certified radius (under multi-class setting) against ℓ_2 perturbations. Other distributions against different ℓ_p perturbations (*universality* of UCAN) are evaluated in Section VI-B and VI-C.

Parameter Setting. For a fair comparison, we follow [12] to set the variance $\lambda = 1$ for isotropic Gaussian noise to benchmark with our methods. For our pattern-wise method, since the variance varies in different dimensions, we re-scale $\sigma(a, b)$ such that $mean\{\sigma\} \approx 1.0$. For the dataset-wise noise, we empirically select the γ to achieve the best trade-off on each dataset. For the certification-wise noise, the amplified factor γ in the parameter generator is set as 1.0 for all datasets.

Experimental Results. The results on MNIST and CIFAR10 are presented in Figure 6, and other results on ImageNet datasets are deferred to Appendix D. First, as shown in Figure 4, the certified accuracy of certification-wise anisotropic noise dominates all the noise customization methods across various settings. This indicates that optimizing anisotropic noise for each certification (of a specific input) can universally boost the performance w.r.t. the certified radius and the ALM since it achieves the best optimality on each certification (both the mean and variance can be optimized according to the input that will be certified). Second, dataset-wise anisotropic noise offers a modest improvement in certified accuracy w.r.t. the certified radius but significantly boosts certified accuracy w.r.t. the ALM. The reason is that the training of dataset-wise NPG can achieve a better trade-off between the prediction accuracy and the $mean\{\sigma\}$ since NPG can learn to assign small variance to the key data dimensions to improve the prediction while maintaining the $mean\{\sigma\}$ by increasing the variance in other data dimensions. However, it is hard to improve the trade-off between the prediction accuracy and the $\min\{\sigma\}$ since decreasing $\min\{\sigma\}$ drops the certified radius while improving the prediction (see some examples for the dataset-wise anisotropic noise in Appendix VI-D). Similarly, the pattern-wise anisotropic noise only improves the certified accuracy w.r.t. certified radius slightly (even reduces the performance on CIFAR10), but we also observe a better trade-off between the certified accuracy w.r.t. ALM. Finally, we also observe that the dataset-wise anisotropic noise can significantly improve the ALM on MNIST (as much as $ALM = 6$). These improvements stem from our method's ability to optimize noise parameters at different levels of optimality. Unlike fixed noise patterns, certification-wise anisotropic noise adapts both mean and variance to each input's characteristics, leading to better prediction accuracy while maintaining robustness guarantees. The significant ALM improvements (up to 6x on MNIST) result from our optimization targeting the overall certified volume rather than just the worst-case radius.

B. Universality (Noise Distributions for ℓ_p Perturbations)

In this section, we evaluate the universality of UCAN with the certification-wise anisotropic noise over different noise

distributions against different ℓ_p perturbations. Specifically, we evaluate the randomized smoothing methods with noise listed in Table I, and follow [65] to set the scalar parameter λ of different noise distributions with variance 1. For the anisotropic noise, we follow the aforementioned parameter settings for pattern-wise, dataset-wise, and certification-wise anisotropic noise. We present the results against ℓ_1 and ℓ_∞ perturbations due to the lack of existing RS theories for non-Gaussian noise against ℓ_2 perturbations.

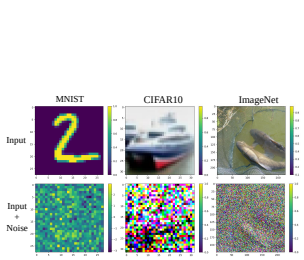
In Figure 7, for all settings, UCAN can universally amplify the certified robustness of isotropic RS. It also shows that the anisotropic Laplace noise and the anisotropic Gaussian noise achieve the best trade-offs between certified accuracy and certified radius/ALM against ℓ_1 perturbation and ℓ_∞ perturbation, respectively. This universal improvement across different ℓ_p norms and noise distributions validates our linear transformation theory, which can seamlessly convert any isotropic randomized smoothing method to its anisotropic counterpart. The consistent performance gains demonstrate that our approach captures fundamental properties of optimal noise distribution regardless of ℓ_p norms and noise types. More experiments with various NPG can be found in Appendix D.

C. Best Performance Comparison vs. SOTA Methods

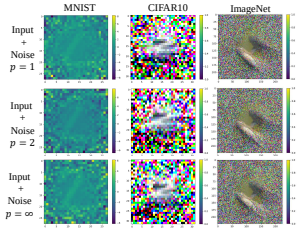
We also compare our best performance (certification with certification-wise anisotropic noise) with the best performance of 15 SOTA methods. Here we present the certified accuracy w.r.t. both *ALM* and ℓ_p *radius*. *Note that the ALM and radius are equivalent for SOTA methods with isotropic noise.*

Following the same settings in such existing randomized smoothing methods [2, 12, 52, 56], we focus on the Gaussian noise against ℓ_2 perturbations to benchmark with them. Results are shown in Table III, IV, and V. We also present the improvement of our method over the best baseline in percentage.

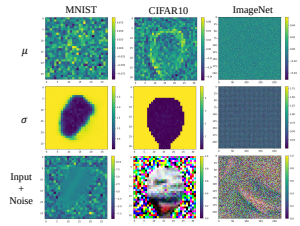
On all three datasets, UCAN significantly boosts the certified accuracy. For instance, it achieves the improvement of 142.5%, 182.6%, and 121.1% over the best baseline on MNIST, CIFAR10, and ImageNet, respectively. UCAN also achieves the best trade-off between certified accuracy and radius/ALM (*two complementary metrics*): 1) UCAN presents both larger radius/ALM and higher certified accuracy in general, and 2) On large radius/ALM, UCAN can still achieve high certified accuracy. We also observe that our certified accuracy is smaller than the SOTA performances at some low radius/ALM on MNIST and ImageNet, this is because of the lower training performance of the classifier. The substantial improvements (up to 182.6%) at larger radii demonstrate our method's strength in maintaining high certified accuracy where traditional methods degrade rapidly. This advantage stems from the combination of the linear transformation approach and the certification-wise noise, which preserves the statistical properties of the original smoothing while enabling dimension-specific noise adaptation. The larger the certified region required, the more pronounced our method's benefits become, as anisotropic noise can better accommodate the heterogeneity across input dimensions.



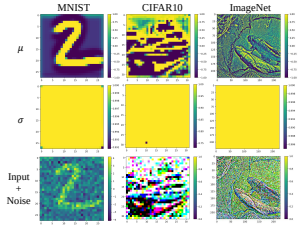
(a) Example of Isotropic Noise



(b) Examples of pattern-wise anisotropic noise



(c) Examples of dataset-wise anisotropic noise



(d) Examples of certification-wise anisotropic noise

Fig. 8: Visualization of anisotropic vs. isotropic noise, based on the input in (a).

Beyond ℓ_2 Norm. We also compare our certification-wise method's performance against ℓ_1 and ℓ_{inf} perturbation with existing baselines [3] (see Table II). Across both norms, UCAN consistently improves certified accuracy (CA) over isotropic RS. For ℓ_1, anisotropic Gaussian yields the highest CA at small–medium radii (e.g., 72% at R=1.5 vs. 55% in [65]), while anisotropic Laplace dominates at larger radii (e.g., 61% at R=2.0 vs. $\leq$25% in baselines), whereas anisotropic Uniform peaks early then collapses (0% beyond R=2.0). For ℓ_∞, anisotropic Gaussian is uniformly best (e.g., 73% at $6/255$ vs. 31–38% in [12, 69]), maintaining sizable margins up to $16/255$. These trends confirm our linear-transformation theory's universality and show that choosing the noise family to match the threat norm (Laplace for ℓ_1, Gaussian for ℓ_∞) and employing certification-wise NPGs yields the highest practical "optimality" under fixed budgets.

TABLE II: Certified accuracy vs. ℓ_1 and ℓ_{inf} perturbations (CIFAR10). Best and $\geq$ SOTA

Radius (ℓ_1 norm)	0.50	1.00	1.50	2.00	2.5	3.0	3.5	4.0
Teng et al.'s [54]	61%	39%	24%	16%	11%	7%	4%	3%
Yang et al.'s [65]	74%	62%	55%	48%	43%	40%	37%	33%
Ours (Ani. Uniform)	**84%**	**82%**	**76%**	0%	0%	0%	0%	0%
Ours (Ani. Gaussian)	**81%**	**77%**	**72%**	**66%**	**61%**	**56%**	**47%**	0%
Ours (Ani. Laplace)	**75%**	**70%**	**66%**	**61%**	**57%**	**52%**	**50%**	**46%**
Radius (ℓ_{inf} norm)	2/255	4/255	6/255	8/255	10/255	12/255	14/255	16/255
Cohen et al.'s [12]	58%	42%	31%	25%	18%	13%	–	–
Zhang et al.'s [69]	60%	47%	38%	32%	23%	17%	–	–
Ours (Ani. Gaussian)	**81%**	**78%**	**73%**	**70%**	**65%**	**60%**	**54%**	**48%**

D. Visualization

We present several examples of anisotropic (generated with ALM loss) and isotropic noise in Figure 8. All the proposed anisotropic noise generation methods find better

[3]Due to the limited number of prior studies on the ℓ_1 and ℓ_∞ norms, we compare our method with two available baselines for each case.

TABLE III: Certified accuracy vs. ℓ_2 perturbations (MNIST). Best and $\geq$ SOTA

Radius and ALM (equivalent for isotropic)	0.0	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25
Cohen's [12]	83%	61%	43%	32%	22%	17%	14%	9%	7%	4%
Sample-wise [56]	98%	97%	96%	93%	88%	81%	73%	57%	41%	25%
Input-depend [52]	99%	98%	97%	94%	88%	79%	58%	27%	0%	0%
MACER [68]	99%	99%	96%	95%	90%	83%	73%	50%	36%	28%
SmoothMix [33]	99%	99%	98%	97%	93%	89%	82%	71%	45%	37%
DRT [67]	99%	98%	98%	97%	93%	89%	83%	70%	48%	40%
Ours (certified accuracy w.r.t. ALM)	98%	98%	98%	98%	97%	97%	96%	96%	95%	94%
Improvement over the Best Baseline (%)	-1.0%	-1.0%	+0.0%	+1.0%	+4.3%	+9.0%	+15.7%	+35.2%	+98.0%	+135.0%
Ours (certified accuracy w.r.t. radius)	**99%**	**99%**	**99%**	**99%**	**99%**	**99%**	**98%**	**98%**	**98%**	**97%**
Improvement over the Best Baseline (%)	+0.0%	+0.0%	+1.0%	+2.1%	+6.5%	+11.2%	+18.1%	+38.0%	+104.2%	+142.5%

TABLE IV: Certified accuracy vs. ℓ_2 perturbations (CIFAR10). Best and $\geq$ SOTA

Radius and ALM (equivalent for isotropic)	0.0	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25
Cohen's [12]	83%	61%	43%	32%	22%	17%	14%	9%	7%	4%
SmoothAdv[49]	–	81%	63%	52%	37%	33%	29%	25%	18%	16%
MACER [68]	81%	71%	59%	47%	39%	33%	29%	23%	19%	17%
Consistency [34]	78 %	69%	58%	49%	38%	34%	30%	25%	20%	17%
SmooothMix [33]	77%	68%	58%	48%	37%	32%	26%	20%	17%	15%
Boosting [30]	83%	71%	60%	52%	39%	34%	30%	25%	20%	17%
DRT [67]	73 %	67%	60%	51%	40%	36%	30%	24%	20%	–
Black-box [69]	–	61%	46%	37%	25%	19%	16%	14%	11%	9%
Data-depend [2]	82%	68%	53%	44%	32%	21%	14%	8%	4%	1%
Sample-wise [56]	84%	74%	61%	52%	45%	41%	36%	32%	27%	23%
Input-depend [52]	83%	62%	43%	27%	18%	11%	5%	2%	0%	0%
Denoise 1 [8]	80%	70%	55%	48%	37%	32%	29%	25%	15%	14%
Denoise 2 [71]	85%	76%	66%	57%	44%	37%	31%	25%	22%	20%
ANCER [20]	84%	80%	67%	–	34%	–	15%	–	11%	–
RANCER [47]	–	81%	48%	28%	11%	1%	–	–	–	–
Ours (certified accuracy w.r.t. ALM)	**86%**	**84%**	**82%**	**80%**	74%	71%	68%	65%	61%	57%
Improvement over the Best Baseline (%)	+1.2%	+3.7%	+6.5%	+40.4%	+39.6%	+73.2%	+88.9%	+103.1%	+125.9%	+147.8%
Ours (certified accuracy w.r.t. radius)	85%	83%	81%	80%	**77%**	**75%**	**73%**	**70%**	**68%**	**65%**
Improvement over the Best Baseline (%)	+0.0%	+2.5%	+5.2%	+40.4%	+45.3%	+82.9%	+102.8%	+118.8%	+151.9%	+182.6%

TABLE V: Certified accuracy vs. ℓ_2 perturbations (ImageNet). Best and $\geq$ SOTA

Radius and ALM (equivalent for isotropic)	0.00	0.50	1.00	1.50	2.00	2.50	3.00	3.50
Cohen's [12]	67%	49%	37%	28%	19%	15%	12%	9%
SmoothAdv [49]	67%	56%	45%	38%	28%	26%	20%	17%
MACER [68]	68%	57%	43%	37%	27%	25%	20%	–
Consistency [34]	57%	50%	44%	34%	24%	21%	17%	–
SmoothMix [33]	55%	50%	43%	38%	26%	24%	20%	–
Boosting [30]	68%	57%	45%	38%	29%	25%	21%	19%
DRT[67]	50%	47%	44%	39%	30%	29%	23%	–
Black-box [69]	–	50%	39%	31%	21%	17%	13%	10%
Data-depend [2]	62%	59%	48%	43%	31%	25%	22%	19%
Denoise 1 [8]	48%	41%	30%	24%	19%	16%	13%	–
Denoise 2 [71]	66%	59%	48%	40%	31%	25%	22%	–
ANCER [20]	66%	66%	62%	58%	44%	37%	32%	–
Ours (certified accuracy w.r.t. ALM)	**71%**	**66%**	**62%**	**58%**	**54%**	**51%**	**47%**	**42%**
Improvement over the Best Baseline (%)	+4.4%	+0.0%	+0.0%	+0.0%	+22.7%	+37.8%	+46.9%	+121.1%
Ours (certified accuracy w.r.t. radius)	65%	62%	58%	53%	50%	46%	43%	38%
Improvement over the Best Baseline (%)	-4.4%	-6.1%	-6.5%	-8.6%	+13.6%	+24.3%	+34.4%	+100.0%

spatial distribution to generate the anisotropic noise. Both the pattern-wise and dataset-wise anisotropic reduce the variance on the key area, except the dataset-wise anisotropic noise on ImageNet (it seems the parameter generator does not find a constant key area on ImageNet due to the complicated data distribution in ImageNet). We also observe that the parameter generator for certification-wise anisotropic noise generates large mean offsets to compensate for the high σ values. It turns out that with high variance ($\sigma_i \approx 1$), the object is still recognizable in the certification-wise anisotropic noise.

E. *Efficiency*

UCAN is a universal framework that can be readily integrated into existing randomized smoothing to boost performance. Whether the extra neural network components (parameter generator) in UCAN will degrade the efficiency of existing randomized smoothing is an important question. We show that the running time overhead resulting from the parameter

generator is negligible compared to the running time of the certification, since for each input, the classifier needs to evaluate N noise samples while μ and σ are generated once. Typically, $N = 100,000$. UCAN can be trained offline and tested online to boost the performance of randomized smoothing. We evaluate the online certification running time for certification-wise anisotropic noise generation and traditional randomized smoothing [12] on ImageNet with four Tesla V100 GPUs and $2,000$ batch size, the average runtimes over 500 samples are 27.43s and 27.09s per sample for our method and [12]'s method, respectively. Thus, the NPG will only slightly increase the overall runtime. Also, Training a certification-wise NPG requires 200 epochs, taking 31.67 minutes on an H100 GPU.

VII. Related Work

In this section, we review the related works for certified defenses against evasion attacks on machine learning models.

Certified Defenses. Certified defenses guarantee robustness against adversarial perturbations within a specified boundary (e.g., ℓ_1, ℓ_2, or ℓ_∞ ball of radius R). Existing approaches fall into two categories: exact and conservative certified defenses. Exact methods leverage satisfiability modulo theories [7, 19, 32, 35] or mixed-integer linear programming [6, 10, 21, 42] to precisely determine the existence of adversarial examples within R, but are not scalable to large networks. Conservative methods use global/local Lipschitz constants [3, 11, 25, 28, 55], optimization-based certification [16, 46, 58, 59], or layer-by-layer approaches [26, 45, 51, 57, 70], which scale better but typically yield conservative or architecture-specific guarantees. Neither can certify arbitrary classifiers, a gap addressed by randomized smoothing.

Randomized Smoothing. Randomized smoothing was first explored by Lecuyer et al. [39], who derived a loose theoretical robustness bound via Differential Privacy [17, 18]. Cohen et al. [12] later provided the first tight guarantee, showing that adding Gaussian noise transforms any classifier into a smoothed classifier with certified ℓ_2 robustness. Subsequent work extended smoothing to other ℓ_p norms: Teng et al. [54] analyzed ℓ_1 robustness with Laplace noise, and Lee et al. [40] studied ℓ_0 robustness with uniform noise. Unified frameworks have also been proposed: Zhang et al. [69] presented an optimization-based approach certifying ℓ_1, ℓ_2, and ℓ_∞ robustness; Yang et al. [65] introduced methods based on level sets and differentials to bound certified radii for various distributions and norms; Hong et al. [29] proposed a framework to approximately certify any ℓ_p-norm adversary with any noise. However, existing randomized smoothing methods use fixed noise distributions (e.g., Gaussian, Laplace) applied uniformly to all input dimensions, overlooking data heterogeneity and limiting optimal protection for different inputs.

Data-Dependent Randomized Smoothing. Data-dependent randomized smoothing improves certified robustness by optimizing noise distributions for individual inputs. Most existing approaches focus on isotropic noise, typically optimizing the variance to enlarge the certified radius. For instance, Alfarra et al. [2] optimize the Gaussian variance via gradient methods; Sukenik et al. [52] model variance as an input-dependent function; Wang et al. [56] use grid search for variance selection. However, these methods still inject noise with identical variance across all dimensions, due to the absence of anisotropic theory.

Asymmetric Randomized Smoothing. Prior work has already shown that randomized smoothing can certify *anisotropic* regions: ANCER [20] introduces axis-aligned ellipsoidal certificates by scaling per-dimension noise, and RANCER [47] generalizes this to full (rotated) covariance structures. Both frameworks derive guarantees by enforcing Lipschitz constraints on the smoothed classifier and by (optionally) adapting noise parameters per input to enlarge the certified set.

Our approach departs from the Lipschitz-based route and instead relies on an explicit *linear transformation* between isotropic and anisotropic noise distributions. Concretely, let the isotropic guarantee certify x within radius R under noise ϵ. For any invertible covariance matrix Σ (cf. our notation in Eq. (4)), the same argument transfers to the anisotropic setting by the change of variables $\delta' = \Sigma\delta$: $\|\delta\|_p \leq R \iff \|\Sigma^{-1}\delta'\|_p \leq R$, so the certified region becomes an ellipsoid (or generalized ellipsoid) parameterized by Σ. This algebraic reduction preserves the original Neyman–Pearson optimality of isotropic smoothing and avoids the typically looser (or at least not tighter) bounds introduced by gradient/Lipschitz upper bounds used in [20, 47].

[20, 47] allow input-dependent noise but must cache those parameters to apply the same distribution for all evaluations during certification; otherwise the proof assumptions change. Our "certification-wise" design fixes the noise after observing the clean input and reuses it throughout the procedure, eliminating the memory-based workaround while retaining input adaptivity at x. By replacing Lipschitz bounds with a linear change of variables, we (i) simplify the theoretical pipeline, (ii) obtain tighter certificates (no gradient over-approximation), and (iii) remove the need for parameter memorization. Empirically (Table IV), at radius 1.25, our method certifies 75% accuracy, far exceeding ANCER (31%) and RANCER (1%). Across all large radii, we achieve up to 182.6% improvement over these baselines.

VIII. Conclusion

In this paper, we introduce UCAN (Universally Certifies adversarial robustness with Anisotropic Noise), a novel and flexible randomized smoothing framework that transforms any isotropic noise-based scheme into schemes utilizing anisotropic noise, providing strict and theoretically grounded robustness guarantees. We also propose a unified and customizable noise customization framework with three methods for fine-tuning anisotropic noise in classifier smoothing. Extensive and comprehensive evaluation of the proposed method on MNIST, CIFAR10, and ImageNet confirms that UCAN substantially enhances the certified robustness of existing randomized smoothing methods. By enabling flexible and dimension-aware noise design, UCAN opens new possibilities for certified defense in more complex and heterogeneous data domains.

References

[1] Ahmed, S., Saleeby, E.G.: On volumes of hyper-ellipsoids. Mathematics Magazine (2018)

[2] Alfarra, M., Bibi, A., Torr, P.H., Ghanem, B.: Data dependent randomized smoothing. arXiv preprint arXiv:2012.04351 (2020)

[3] Anil, C., Lucas, J., Grosse, R.: Sorting out lipschitz function approximation. In: ICML (2019)

[4] Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: ICML (2018)

[5] Bartle, R.G.: The elements of integration and Lebesgue measure. John Wiley & Sons (2014)

[6] Bunel, R.R., Turkaslan, I., Torr, P.H., Kohli, P., Mudigonda, P.K.: A unified view of piecewise linear neural network verification. In: NeurIPS (2018)

[7] Carlini, N., Katz, G., Barrett, C., Dill, D.L.: Provably minimally-distorted adversarial examples. arXiv preprint arXiv:1709.10207 (2017)

[8] Carlini, N., Tramer, F., Dvijotham, K.D., Rice, L., Sun, M., Kolter, J.Z.: (certified!!) adversarial robustness for free! ICLR (2023)

[9] Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: IEEE S&P (2017)

[10] Cheng, C.H., Nührenberg, G., Ruess, H.: Maximum resilience of artificial neural networks. In: ATVA (2017)

[11] Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y.N., Usunier, N.: Parseval networks: Improving robustness to adversarial examples. In: ICML (2017)

[12] Cohen, J., Rosenfeld, E., Kolter, Z.: Certified adversarial robustness via randomized smoothing. In: ICML (2019)

[13] Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A.: Generative adversarial networks: An overview. IEEE signal processing magazine (2018)

[14] Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML (2020)

[15] Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., Zhu, J.: Efficient decision-based black-box adversarial attacks on face recognition. In: CVPR (2019)

[16] Dvijotham, K., Stanforth, R., Gowal, S., Mann, T.A., Kohli, P.: A dual approach to scalable verification of deep networks. In: UAI (2018)

[17] Dwork, C.: Differential privacy. In: ICALP (2006)

[18] Dwork, C.: Differential privacy: A survey of results. In: TAMC (2008)

[19] Ehlers, R.: Formal verification of piece-wise linear feed-forward neural networks. In: ATVA (2017)

[20] Eiras, F., Alfarra, M., Torr, P., Kumar, M.P., Dokania, P.K., Ghanem, B., Bibi, A.: ANCER: Anisotropic certification via sample-wise volume maximization. TMLR (2022)

[21] Fischetti, M., Jo, J.: Deep neural networks and mixed integer linear optimization. Constraints (2018)

[22] Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. In: DSAA (2018)

[23] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM (2020)

[24] Goodfellow, I., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015)

[25] Gouk, H., Frank, E., Pfahringer, B., Cree, M.J.: Regularisation of neural networks by enforcing lipschitz continuity. Machine Learning (2021)

[26] Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T.A., Kohli, P.: On the effectiveness of interval bound propagation for training verifiably robust models (2018), http://arxiv.org/abs/1810.12715

[27] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

[28] Hein, M., Andriushchenko, M.: Formal guarantees on the robustness of a classifier against adversarial manipulation. In: NeurIPS (2017)

[29] Hong, H., Wang, B., Hong, Y.: Unicr: Universally approximated certified robustness via randomized smoothing. In: ECCV (2022)

[30] Horváth, M.Z., Müller, M.N., Fischer, M., Vechev, M.: Boosting randomized smoothing with variance reduced classifiers. arXiv preprint arXiv:2106.06946 (2021)

[31] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)

[32] Huang, X., Kwiatkowska, M., Wang, S., Wu, M.: Safety verification of deep neural networks. In: CAV (2017)

[33] Jeong, J., Park, S., Kim, M., Lee, H.C., Kim, D.G., Shin, J.: Smoothmix: Training confidence-calibrated smoothed classifiers for certified robustness. NeurIPS (2021)

[34] Jeong, J., Shin, J.: Consistency regularization for certified robustness of smoothed classifiers. NeurIPS (2020)

[35] Katz, G., Barrett, C., Dill, D.L., Julian, K., Kochenderfer, M.J.: Reluplex: An efficient smt solver for verifying deep neural networks. In: CAV (2017)

[36] Keahey, K., Anderson, J., Zhen, Z., Riteau, P., Ruth, P., Stanzione, D., Cevik, M., Colleran, J., Gunawi, H.S., Hammock, C., Mambretti, J., Barnes, A., Halbach, F., Rocha, A., Stubbs, J.: Lessons learned from the chameleon testbed. In: USENIX ATC (2020)

[37] Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)

[38] LeCun, Y., Cortes, C., Burges, C.: Mnist handwritten digit database. ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist (2010)

[39] Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., Jana, S.: Certified robustness to adversarial examples with differential privacy. In: IEEE S&P (2019)

[40] Lee, G., Yuan, Y., Chang, S., Jaakkola, T.S.: Tight certificates of adversarial robustness for randomly smoothed classifiers. In: NeurIPS (2019)

[41] Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. NeurIPS (2018)

[42] Lomuscio, A., Maganti, L.: An approach to reachability analysis for feed-forward relu neural networks. arXiv preprint arXiv:1706.07351 (2017)

[43] Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition (2021)

[44] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)

[45] Mirman, M., Gehr, T., Vechev, M.: Differentiable abstract interpretation for provably robust neural networks. In: ICML (2018)

[46] Raghunathan, A., Steinhardt, J., Liang, P.: Certified defenses against adversarial examples. arXiv preprint arXiv:1801.09344 (2018)

[47] Rumezhak, T., Eiras, F.G., Torr, P.H., Bibi, A.: Rancer: Non-axis aligned anisotropic certification with randomized smoothing. In: WACV (2023)

[48] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV (2015)

[49] Salman, H., Li, J., Razenshteyn, I., Zhang, P., Zhang, H., Bubeck, S., Yang, G.: Provably robust deep learning via adversarially trained smoothed classifiers. NeurIPS (2019)

[50] Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G., Goldstein, T.: Adversarial training for free! In: NeurIPS (2019)

[51] Singh, G., Gehr, T., Mirman, M., Püschel, M., Vechev, M.: Fast and effective robustness certification. In: NeurIPS (2018)

[52] Súkeník, P., Kuvshinov, A., Günnemann, S.: Intriguing properties of input-dependent randomized smoothing. arXiv preprint arXiv:2110.05365 (2021)

[53] Sun, J., Cao, Y., Chen, Q.A., Mao, Z.M.: Towards robust lidar-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In: USENIX Security (2020)

[54] Teng, J., Lee, G.H., Yuan, Y.: ℓ_1 adversarial robustness certificates: a randomized smoothing approach (2020), https://openreview.net/forum?id=H1lQIgrFDS

[55] Tsuzuku, Y., Sato, I., Sugiyama, M.: Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In: NeurIPS (2018)

[56] Wang, L., Zhai, R., He, D., Wang, L., Jian, L.: Pretrain-to-finetune adversarial training via sample-wise randomized smoothing (2020)

[57] Weng, T., Zhang, H., Chen, H., Song, Z., Hsieh, C., Daniel, L., Boning, D.S., Dhillon, I.S.: Towards fast computation of certified robustness for relu networks. In: ICML (2018)

[58] Wong, E., Kolter, J.Z.: Provable defenses against adversarial examples via the convex outer adversarial polytope. In: ICML (2018)

[59] Wong, E., Schmidt, F.R., Metzen, J.H., Kolter, J.Z.: Scaling provable adversarial defenses. arXiv preprint arXiv:1805.12514 (2018)

[60] Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: ICLR (2018)

[61] Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K.: Feature denoising for improving adversarial robustness. In: CvPR (2019)

[62] Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: CVPR (2019)
[63] Xu, B., Wang, N., Chen, T., Li, M.: Empirical evaluation of rectified activations in convolutional network. arXiv preprint arXiv:1505.00853 (2015)
[64] Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)
[65] Yang, G., Duan, T., Hu, J.E., Salman, H., Razenshteyn, I., Li, J.: Randomized smoothing of all shapes and sizes. In: ICML (2020)
[66] Yang, S., Guo, T., Wang, Y., Xu, C.: Adversarial robustness through disentangled representations. In: AAAI (2021)
[67] Yang, Z., Li, L., Xu, X., Kailkhura, B., Xie, T., Li, B.: On the certified robustness for ensemble models and beyond. arXiv preprint arXiv:2107.10873 (2021)
[68] Zhai, R., Dan, C., He, D., Zhang, H., Gong, B., Ravikumar, P., Hsieh, C.J., Wang, L.: Macer: Attack-free and scalable robust training via maximizing certified radius. arXiv preprint arXiv:2001.02378 (2020)
[69] Zhang, D., Ye, M., Gong, C., Zhu, Z., Liu, Q.: Black-box certification with randomized smoothing: A functional optimization based framework (2020)
[70] Zhang, H., Weng, T., Chen, P., Hsieh, C., Daniel, L.: Efficient neural network robustness certification with general activation functions. In: NeurIPS (2018)
[71] Zhang, J., Chen, Z., Zhang, H., Xiao, C., Li, B.: {DiffSmooth}: Certifiably robust learning via diffusion models and local smoothing. In: USENIX Security (2023)
[72] Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y.: Residual dense network for image super-resolution. In: CVPR (2018)

APPENDIX A
PROOF OF THEOREM 2

We restate Theorem 2:

Theorem 2 (Asymmetric Randomized Smoothing via Universal Transformation). *Let $f : \mathbb{R}^d \to \mathcal{C}$ be any deterministic or randomized function. Suppose that for the multivariate random variable with isotropic noise $X = x + \epsilon$ in Theorem 1, the certified radius function is $R(\cdot)$. Then, for the corresponding anisotropic input $Y = x + \epsilon^\top \Sigma + \mu$, if there exist $c'_A \in \mathcal{C}$ and $\underline{p_A}', \overline{p_B}' \in [0, 1]$ such that:*

$$\mathbb{P}\left(f(Y) = c'_A\right) \geq \underline{p_A}' \geq \overline{p_B}' \geq \max_{c \neq c'_A} \mathbb{P}\left(f(Y) = c\right) \quad (5)$$

then for the anisotropic smoothed classifier $g'(x + \delta') \equiv \arg\max_{c \in \mathcal{C}} \mathbb{P}\left(f(Y + \delta') = c\right)$, we can guarantee $g'(x+\delta') = c'_A$ for all perturbations $\delta' \in \mathbb{R}^d$ such that:

$$\left\|\Sigma^{-1}\delta'\right\|_p \leq R\left(\underline{p_A}', \overline{p_B}'\right) \quad (6)$$

provided that Σ is invertible.

Proof. Let $X = x + \epsilon$, where $\epsilon \in \mathbb{R}^d$ follows an isotropic noise distribution. Let $Y = x + \Sigma\epsilon + \mu$, where $\Sigma \in \mathbb{R}^{d \times d}$ is an invertible covariance matrix, and $\mu \in \mathbb{R}^d$ is a mean offset vector.

Given the input x, covariance matrix Σ, and mean offset μ, define:

$$\mu' = \mu + x - \Sigma x \quad (28)$$

Then, the anisotropic input can be written as:

$$Y = \Sigma(x + \epsilon) + \mu' = \Sigma X + \mu' \quad (29)$$

Define a transformation $h : \mathbb{R}^d \to \mathbb{R}^d$ as:

$$h(z) = \Sigma z + \mu' \quad (30)$$

This transformation maps the isotropic input z to the anisotropic space.

Given any deterministic or randomized function $f : \mathbb{R}^d \to \mathcal{C}$, define the anisotropic smoothed classifier as:

$$g'(x) = \arg\max_{c \in \mathcal{C}} \mathbb{P}\left(f\left(x + \Sigma\epsilon + \mu\right) = c\right) \quad (31)$$

Suppose that for the anisotropic random variable Y, there exist $c'_A \in \mathcal{C}$ and $\underline{p_A}', \overline{p_B}' \in [0, 1]$ such that:

$$\mathbb{P}\left(f(Y) = c'_A\right) \geq \underline{p_A}' \geq \overline{p_B}' \geq \max_{c \neq c'_A} \mathbb{P}\left(f(Y) = c\right) \quad (32)$$

By the transformation h, the condition in Equation (32) is equivalent to:

$$\begin{aligned}
\mathbb{P}\left(f(Y) = c'_A\right) &= \mathbb{P}\left(f\left(\Sigma X + \mu'\right) = c'_A\right) && (33)\\
&= \mathbb{P}\left(f\left(h(X)\right) = c'_A\right) && (34)\\
&\geq \underline{p_A}' \geq \overline{p_B}' \geq \max_{c \neq c'_A} \mathbb{P}\left(f\left(h(X)\right) = c\right) && (35)
\end{aligned}$$

Consider a new classifier $f' : \mathbb{R}^d \to \mathcal{C}$ defined as:

$$f'(z) = f\left(h(z)\right) \quad (36)$$

This classifier maps the isotropic input z to the class space $\mathcal{C}$ through the transformation h.

From Equations (34) and (35), we have:

$$\mathbb{P}\left(f'(X) = c'_A\right) \geq \underline{p_A}' \geq \overline{p_B}' \geq \max_{c \neq c'_A} \mathbb{P}\left(f'(X) = c\right) \quad (37)$$

This is the prerequisite condition in the standard isotropic randomized smoothing theory (e.g., Theorem 1).

Therefore, we can obtain the robustness guarantee with isotropic certified radius R such that:

$$\arg\max_{c \in \mathcal{C}} \mathbb{P}\left(f'(X + \delta) = c\right) = c'_A \quad (38)$$

$$\text{subject to } \left\|\delta\right\|_p \leq R\left(\underline{p_A}', \overline{p_B}'\right) \quad (39)$$

By the definition of h in Equation (30) and the fact that $Y = h(X)$, Equation (38) is equivalent to:

$$\begin{aligned}
&\arg\max_{c \in \mathcal{C}} \mathbb{P}\left(f'\left(X + \delta\right) = c\right) && (40)\\
&\quad = \arg\max_{c \in \mathcal{C}} \mathbb{P}\left(f\left(h\left(X + \delta\right)\right) = c\right) && (41)\\
&\quad = \arg\max_{c \in \mathcal{C}} \mathbb{P}\left(f\left(\Sigma\left(X + \delta\right) + \mu'\right) = c\right) && (42)\\
&\quad = \arg\max_{c \in \mathcal{C}} \mathbb{P}\left(f\left(\Sigma X + \Sigma\delta + \mu'\right) = c\right) && (43)\\
&\quad = \arg\max_{c \in \mathcal{C}} \mathbb{P}\left(f\left(Y + \Sigma\delta\right) = c\right) = c'_A && (44)
\end{aligned}$$

Define $\delta' = \Sigma\delta$, with $\delta = \Sigma^{-1}\delta'$ since Σ is invertible.

Substituting $\delta = \Sigma^{-1}\delta'$ into the condition $\left\|\delta\right\|_p \leq R\left(\underline{p_A}', \overline{p_B}'\right)$, we obtain:

$$\left\|\Sigma^{-1}\delta'\right\|_p \leq R\left(\underline{p_A}', \overline{p_B}'\right) \quad (45)$$

Therefore, the guarantee in Equations (38) and (39) is equivalent to:

$$\arg\max_{c \in \mathcal{C}} \mathbb{P}\left(f\left(Y + \delta'\right) = c\right) = c'_A \quad (46)$$

$$\text{subject to } \left\|\Sigma^{-1}\delta'\right\|_p \leq R\left(\underline{p_A}', \overline{p_B}'\right) \quad (47)$$

By Equation (31), we have:

$$g'\left(x+\delta'\right)=\arg\max_{c\in\mathcal{C}}\mathbb{P}\left(f\left(Y+\delta'\right)=c\right)=c'_A \tag{48}$$

for all $\delta'\in\mathbb{R}^d$ satisfying $\left\|\Sigma^{-1}\delta'\right\|_p\leq R\left(\underline{p_A}',\overline{p_B}'\right)$.

This completes the proof. □

BINARY CASE OF THEOREM 2

Theorem 8 (Universal Transformation for Anisotropic Noise). *Let $f:\mathbb{R}^d\to\mathcal{C}$ be any deterministic or random function. Suppose that for the isotropic input X in Theorem 1, the certified radius function for the binary case is $R(\cdot)$. Then, for the corresponding anisotropic input Y such that $Y=x+\epsilon^\top\sigma+\mu$, if there exist $c'_A\in\mathcal{C}$ and $\underline{p_A}'\in(1/2,1]$ such that:*

$$\mathbb{P}(f(Y)=c'_A)\geq\underline{p_A}'\geq\frac{1}{2} \tag{49}$$

Then $g'(x+\delta')=\arg\max_{c\in\mathcal{C}}\mathbb{P}(f(Y)=c)=c'_A$ for all $||\delta'\oslash\sigma||_p\leq R(\underline{p_A}')$ where g' denotes the anisotropic smoothed classifier, δ' denotes the perturbation injected to g'.

Proof. The proof for the binary case is similar to the proofs for the multiclass-case (see Appendix A). □

APPENDIX B
ADDITIONAL METRIC FOR CERTIFIED REGION (BINARY-CASE)

How to develop a general metric for evaluating the robustness region for randomized smoothing with anisotropic noise is an important but challenging problem. We observe that the guarantee in Theorem 2 forms a certified region, within which the perturbation is certifiably safe to the smoothed classifier. The ℓ_p-norm bounding on the scaled perturbation $\delta'\oslash\sigma$ results in the anisotropy of the certified region around the input. We illustrate the asymmetric certified region for different ℓ_p-norm guarantees in Figure 2. It shows that if the δ space is a 2-dimension space, then the guarantee of asymmetric RS draws a rhombus, ellipse, and rectangle in ℓ_1, ℓ_2, and ℓ_∞ norms, respectively. Within the asymmetric region, we can find an isotropic region that also satisfies the robustness guarantee (a subset of the anisotropic region), which represents an explicit certified radius as depicted in Corollary 4.

However, evaluating the performance of randomized smoothing with anisotropic noise via Eq. (9) in Corollary 4, although formally guaranteed with robustness, fails to capture the full certified regions.

Specifically, Eq. (9) evaluates the performance only based on the blue region in Figure 2, but the Eq. (6) in Theorem 2 actually guarantees that all the δ (perturbations) within the green region do not change the perturbation. Therefore, to fairly and accurately evaluate the performance of certification via anisotropic noise, we need to develop an auxiliary metric that can cover the entire certified region in highly-dimensional δ space as a complement besides the certified radius.

From another perspective, evaluating the performance of randomized smoothing can be considered as evaluating the size of the robust perturbation set $S(n,p)$.

Consider the Euclidean structure, $S(d,p)$ is a finite set in d-dimensional Euclidean space. Therefore, we leverage the Lebesgue measure [5] to compute the size of $S(d,p)$ (see Theorem 6).

We observe that for a fixed d and p, the $\frac{(2\Gamma(1+\frac{1}{p}))^d}{\Gamma(1+\frac{d}{p})}$ factor in the Lebesgue measure is a constant. Then, when comparing the Lebesgue measure in the same norm ℓ_p and the same space $\mathbb{R}^d$, the constant term can be ignored. Also, the R^d factor can lead to infinite numeral computation, thus we also scale the Lebesgue measure by calculating the d-th root. As a result, we define the Alternative Lebesgue Measure (ALM) of the robust perturbation set with the same d and p as:

$$V'_S=\sqrt[d]{\prod_{i=1}^{d}\sigma_i R} \tag{50}$$

Table VI presents the ALM formulas of common RS methods.

Alternative Lebesgue Measure[4] vs Radius. When the multipliers of the scale parameter for anisotropic noise $\sigma_1=\sigma_2=...=1$, the noise turns into the isotropic noise and the alternative Lebesgue measure turns into the certified radius R. Therefore, the alternative Lebesgue measure can be treated as a generalized metric compared to the certified radius. This generalization based on the certified radius also enables us to fairly compare the randomized smoothing based on anisotropic noise with isotropic noise.

Note that the new metric ALM is not developed to bound the perturbation, but to accurately measure the certified guarantees of randomized smoothing with anisotropic noise.

APPENDIX C
PROOF OF THEOREM 6

Proof.

Definition 9 (d-dimensional Generalized Super-ellipsoid). *The d-dimensional generalized super-ellipsoid ball is defined as*

$$E(d,p)=\{(\delta_1,\delta_2,...,\delta_d):\sum_{i=1}^{d}|\frac{\delta_i}{c_i}|^{p_i}\leq 1,p_i>0\} \tag{51}$$

Definition 10 (Euler Gamma Function). *The Euler gamma function is defined by*

$$\Gamma(\beta)=\int_0^\infty\alpha^{\beta-1}e^{-\alpha}d\alpha \tag{52}$$

There are some properties of Γ: 1) For all $\beta>0$, $\beta\Gamma(\beta)=\Gamma(\beta+1)$, 2) For all positive integers n, $\Gamma(n)=(n-1)!$, and 3) $\Gamma(1/2)=\sqrt{\pi}$.

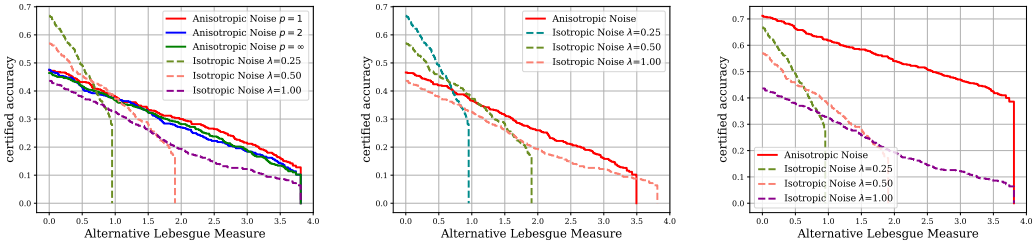
Lemma 11 (Lebesgue Measure of Generalized Super-ellipsoids [1]). *The Lebesgue measure of the generalized super-ellipsoids defined in Definition 9 is given by*

$$V_E(d,p)=2^d\frac{\prod_{i=1}^{d}c_i\Gamma(1+\frac{1}{p_i})}{\Gamma(1+\sum_{i=1}^{d}\frac{1}{p_i})};p_i>0,d=1,2,3,... \tag{53}$$

[4]ALM is like the normalized radius of the certified region in all d dimensions.

TABLE VI: ALM (binary-case) for randomized smoothing with independent anisotropic noise. d is the dimension size. Φ^{-1} is the inverse CDF of Gaussian distribution. λ is the scalar parameter of the isotropic noise. σ is the anisotropic scale multiplier.

Distribution	PDF	Adv.	Anisotropic Guarantee	ALM
Gaussian [12]	$\propto e^{-\|\|\frac{z}{\lambda}\|\|_2^2}$	ℓ_2	$\|\|\delta \oslash \sigma\|\|_2 \leq \lambda(\Phi^{-1}(\underline{p_A}'))$	$\sqrt[d]{\prod_{i=1}^d \sigma_i}\lambda(\Phi^{-1}(\underline{p_A}'))$
Gaussian [65]	$\propto e^{-\|\|\frac{z}{\lambda}\|\|_2^2}$	ℓ_1	$\|\|\delta \oslash \sigma\|\|_1 \leq \lambda(\Phi^{-1}(\underline{p_A}'))$	$\sqrt[d]{\prod_{i=1}^d \sigma_i}\lambda(\Phi^{-1}(\underline{p_A}'))$
		ℓ_∞	$\|\|\delta \oslash \sigma\|\|_\infty \leq \lambda(\Phi^{-1}(\underline{p_A}'))/\sqrt{d}$	$\sqrt[d]{\prod_{i=1}^d \sigma_i}\lambda(\Phi^{-1}(\underline{p_A}'))/\sqrt{d}$
Laplace [54]	$\propto e^{-\|\|\frac{z}{\lambda}\|\|_1}$	ℓ_1	$\|\|\delta \oslash \sigma\|\|_1 \leq -\lambda\log(2(1-\underline{p_A}'))$	$\sqrt[d]{\prod_{i=1}^d \sigma_i}\lambda\log(2(1-\underline{p_A}'))$
Exp. ℓ_∞ [65]	$\propto e^{-\|\|\frac{z}{\lambda}\|\|_\infty}$	ℓ_1	$\|\|\delta \oslash \sigma\|\|_1 \leq 2d\lambda(\underline{p_A}' - \frac{1}{2})$	$2\sqrt[d]{\prod_{i=1}^d \sigma_i}d\lambda(\underline{p_A}' - \frac{1}{2})$
		ℓ_∞	$\|\|\delta \oslash \sigma\|\|_\infty \leq \lambda\log(\frac{1}{2(1-\underline{p_A}')})$	$\sqrt[d]{\prod_{i=1}^d \sigma_i}\lambda\log(\frac{1}{2(1-\underline{p_A}')})$
Uniform ℓ_∞ [41]	$\propto \mathbb{I}(\|\|z\|\|_\infty \leq \lambda)$	ℓ_1	$\|\|\delta \oslash \sigma\|\|_1 \leq 2\lambda(\underline{p_A}' - \frac{1}{2})$	$2\sqrt[d]{\prod_{i=1}^d \sigma_i}\lambda(\underline{p_A}' - \frac{1}{2})$
		ℓ_∞	$\|\|\delta \oslash \sigma\|\|_\infty \leq 2\lambda(1 - \sqrt[d]{\frac{3}{2} - \underline{p_A}'})$	$2\sqrt[d]{\prod_{i=1}^d \sigma_i}\lambda(1 - \sqrt[d]{\frac{3}{2} - \underline{p_A}'})$
Power Law ℓ_∞ [65]	$\propto \frac{1}{(1+\|\|\frac{z}{\lambda}\|\|_\infty)^a}$	ℓ_1	$\|\|\delta \oslash \sigma\|\|_1 \leq \frac{2d\lambda}{a-d}(\underline{p_A} - \frac{1}{2})$	$2\sqrt[d]{\prod_{i=1}^d \sigma_i}\frac{d\lambda}{a-d}(\underline{p_A} - \frac{1}{2})$



(a) Pattern-wise (ImageNet) (b) Dataset-wise (ImageNet) (c) Certification-wise (ImageNet)

Fig. 9: Comparison of randomized smoothing performance on ImageNet with anisotropic noise and that with isotropic noise (Gaussian distribution for certified defense against ℓ_2 perturbations, comparing with [12]).

We leverage Lemma 11 to prove Theorem 2. Let $p_i = p$, the d-dimensional robust perturbation set is equivalent to

$$S(d,p) = \{(\delta_1, \delta_2, ..., \delta_d) : \sum_{i=1}^{d} |\frac{\delta_i}{\sigma_i R}|^p \leq 1\} \tag{54}$$

The Lebesgue measure in Lemma 11 will be

$$V_E(d,p) = 2^d \frac{\prod_{i=1}^d c_i R\Gamma(1+\frac{1}{p})}{\Gamma(1+\sum_{i=1}^d \frac{1}{p})} = \frac{(2R\Gamma(1+\frac{1}{p}))^d \prod_{i=1}^d \sigma_i}{\Gamma(1+\frac{d}{p})}; \quad p > 0, d = 1, 2, 3, ... \tag{55}$$

Thus, this completes the proof. □

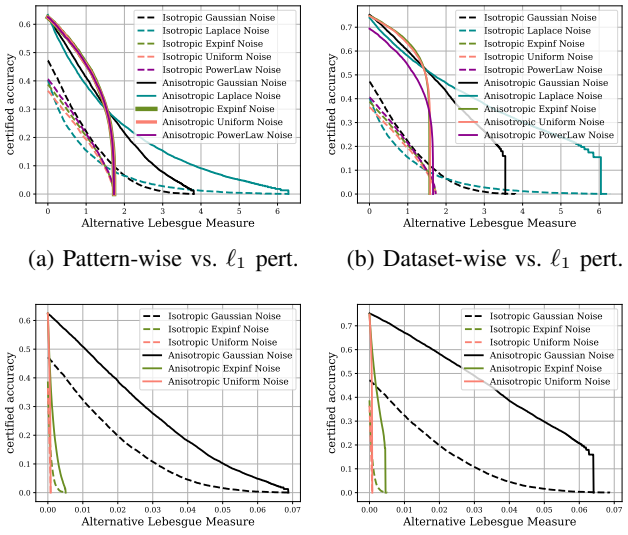
APPENDIX D
ADDITIONAL EXPERIMENTS

A. Anisotropic noise vs. isotropic noise

We also present the performance comparison of the anisotropic and isotropic RS on ImageNet in Figure 9. It shows that similar to the results in the MNIST and CIFAR10, the pattern-wise and dataset-wise anisotropic noise can improve the performance of certified accuracy w.r.t. the ALM moderately, while the certification-wise anisotropic noise can significantly boost the performance with the best optimality.

B. Universality of Pattern-wise and Dataset-wise Methods

Similar to the certification-wise noise, pattern-wise and dataset-wise noise are also universal to different ℓ_p norms and different PDFs. As shown in Figure 10, the pattern-wise and dataset-wise anisotropic noise can also significantly boost the performance of isotropic RS on various settings.



(a) Pattern-wise vs. ℓ_1 pert. (b) Dataset-wise vs. ℓ_1 pert.

(c) Pattern-wise vs. ℓ_∞ pert. (d) Dataset-wise vs. ℓ_∞ pert.

Fig. 10: UCAN with pattern-wise and dataset-wise anisotropic noise vs. RS with isotropic noise – different noise PDFs against different ℓ_p perturbations (universality) on CIFAR10.