Provable Robust Learning Based on Transformation-Specific Smoothing

Linyi Li
University of Illinois, Urbana-Champaign
27 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Maurice Weber
ETH Zurich
14 PUBLICATIONS **128** CITATIONS
SEE PROFILE

Tao Xie
University of Illinois, Urbana-Champaign
365 PUBLICATIONS **13,795** CITATIONS
SEE PROFILE

Ce Zhang
ETH Zurich
233 PUBLICATIONS **6,050** CITATIONS
SEE PROFILE

Provable Robust Learning Based on Transformation-Specific Smoothing

Linyi Li[†] [*] Maurice Weber[‡] [*] Xiaojun Xu[†] [*] Luka Rimanic[‡] Tao Xie[§]
Ce Zhang[‡] Bo Li[†]

[†] University of Illinois at Urbana-Champaign, USA {linyi2, xiaojun3, lbo}@illinois.edu
[‡] ETH Zurich, Switzerland {webermau, luka.rimanic, ce.zhang}@inf.ethz.ch
[§] Peking University, China {taoxie}@pku.edu.cn

Abstract

As machine learning (ML) systems become pervasive, safeguarding their security is critical. Recent work has demonstrated that motivated adversaries could add adversarial perturbations to the test data to mislead ML systems. So far, most research has focused on providing provable robustness guarantees for ML models against a specific ℓ_p norm bounded adversarial perturbation. However, in practice previous work has shown that there are other types of realistic adversarial transformations whose semantic meaning has been leveraged to attack ML systems. In this paper, we aim to provide *a unified framework for certifying ML robustness against general adversarial transformations*. First, we identify the semantic transformations as different categories: *resolvable* (e.g., Gaussian blur and brightness) and *differentially resolvable* transformations (e.g., rotation and scaling). We then provide sufficient conditions and strategies for certifying certain transformations. For instance, we propose a novel sampling-based interpolation approach with estimated Lipschitz upper bound to certify the robustness against differentially resolvable transformations. In addition, we theoretically optimize the smoothing strategies for certifying the robustness of ML models against different transformations. For instance, we show that smoothing by sampling from exponential distribution provides a tighter robustness bound than Gaussian. Extensive experiments on 7 semantic transformations show that our proposed unified framework significantly outperforms the state-of-the-art certified robustness approaches on several datasets including ImageNet.

1 Introduction

Recent advances in machine learning (ML) have vastly improved the capabilities of computational reasoning in complex domains, exceeding human-level performance in tasks such as image recognition [14] and game playing [33, 28]. Despite all of these advances, there are significant vulnerabilities inherent in these systems: image recognition systems can be easily misled [35, 12, 42], and malware detection models can be evaded [38, 44].

The current practice of security in ML has fallen into the trap that every month new attacks are identified [43, 12, 9], followed by new countermeasures [26, 39], which are subsequently broken [2], and so on *ad infinitum*. As a result, recent investigations have been made to provide *provable or certifiable robustness* guarantees for existing learning models. Such certification usually follows the form that when the perturbation is within a certain threshold, the ML model is provably robust against arbitrary adversarial attacks as long as the added perturbation satisfies the threshold. Different certifiable defenses and robustness verification approaches have provided non-trivial robust guarantees especially when the perturbation is bounded by ℓ_p norm [20, 37, 25, 5].

However, only certifying adversarial examples within the ℓ_p norm is not sufficient toward certifying learning

[*] The first three authors contribute equally to this work.

robustness against practical semantic transformation attacks. For instance, it has been shown that image rotations, scaling and other semantic transformations are able to mislead ML models [8, 11, 43]. Previous work [17] has shown that brightness/contrast attacks can achieve 91.6% attack success on CIFAR-10 and on ImageNet, brightness/contrast attacks achieve 71%-100% attack success rate [16]. In practice, there exists vulnerability to brightness attacks in autonomous driving scenarios [29]. Facing these semantic-transformation-based adversarial attacks, a natural question arises: *Can we provide provable robustness guarantees for these semantic transformations?*

In this paper, we propose a series of theoretic and empirical analyses to certify the model robustness against general semantic transformations beyond the ℓ_p norm bounded adversarial perturbations. The theoretical analysis is non-trivial and our empirical results set new state-of-the-arts for a range of different semantic transformations.

We first propose a general framework based on function smoothing to provide provable robustness for ML models against a range of different adversarial transformations (Figure 1). Our framework is two-fold. We first provide results for transformations that are *resolvable*, which include brightness, contrast (and their composition), translation, and Gaussian blur. However, there are many transformations that are not resolvable, including transformations such as rotation and scaling. We further define the notion of *differentially resolvable* transformations and develop novel certification techniques for this type of transformations, using as a building block what we have developed for resolvable transformations.

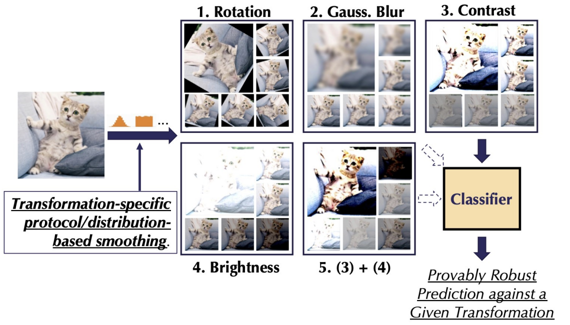

1. Rotation
2. Gauss. Blur
3. Contrast
Transformation-specific protocol/distribution-based smoothing.
4. Brightness
5. (3) + (4)
Classifier
Provably Robust Prediction against a Given Transformation

Figure 1: An illustration of the proposed general model smoothing framework against different semantic transformations. We develop a range of different transformation-specific smoothing protocols with different smoothing distributions to provide substantially better certified robustness bounds than state-of-the-art approaches.

However, a general framework that can be applied to all these transformations is just the very first step. Our main contribution is a series of transformation-specific techniques that improve the model robustness against each type of transform. We obtain these techniques by *jointly* reasoning about (1) function smoothing under different smoothing distributions and (2) the properties inherent to each transformation. To our best knowledge, it is the first time that such analyses have been conducted in the context of function smoothing and semantic transformations.

For *smoothing distributions* that we leverage to perform function smoothing, we analyze sampling distributions beyond the isotropic Gaussian distribution that previous work [5, 10] relies on— we explore the non-isotropic Gaussian and other distributions such as uniform, exponential, and Laplace. Although there has been recent concurrent work that provides certification radius for different distributions [45], we are the first to put these results into the context of semantic transformations.

Digging deeper into each transformation leads to a collection of interesting results that go significantly beyond existing techniques. For example, we show that against certain adversarial transformations such as Gaussian blur, smoothing by sampling from the exponential distribution is better than isotropic Gaussian. The composition of brightness and contrast belongs to a broader class of resolvable transforms and we provide a novel robustness certification strategy (based on non-isotropic Gaussian smoothing) that achieves state-of-the-art robustness guarantee. Rotation and scaling are both differentially resolvable; however, to use our theory, non-trivial algorithms need to be designed.

Empirically, we conduct extensive experiments to evaluate the proposed certification framework and show that it outperforms the state-of-the-art approaches substantially on different datasets against a series of practical semantic transformations.

This paper makes the following technical contributions: (1) We propose a *general* model smoothing framework to certify the model robustness against generic semantic transformations. (2) We categorize most adversarial semantic transformations in the literature into *resolvable* (e.g., brightness, contrast, their composition, and Gaussian blur) and *differentially resolvable* (e.g., rotation and scaling) transformations. We show that our framework is general enough to support both types of transformations. (3) We theoretically explore different smoothing strategies by sampling from different distributions including non-isotropic Gaussian, uniform, and Laplace. We show that for specific transformations, such as Gaussian blur, smoothing with exponential distribution is better. (4) We conduct extensive experiments and make them open-source and reproducible. We show that our framework can provide substantially higher certified robustness compared with the state-of-the-art, against a range of semantic transformations on MNIST, CIFAR-10, and ImageNet.

2 Related Work

Certified robustness against ℓ_p norm bounded perturbation. Certified adversarial robustness training and verification approaches have been proposed to demonstrate their effectiveness. In particular, interval bounding [13], linear relaxations [40, 20, 41], and semidefinite programming [30] are techniques that have been applied to certify the model robustness. Recently, randomized smoothing is shown to be scalable and effective by smoothing the model with Gaussian noise [5, 23, 24]. With improvements on optimizing the smoothing distribution [45, 36, 7] and better training mechanism [4, 31, 46], the performance of randomized smoothing can be further improved. However, these certifiable approaches are only able to provide robustness guarantees for the ℓ_p norm bounded perturbations, while in practice the semantic-transformation-based perturbation would create more stealthy and realistic adversarial instances.

Certified robustness against semantic transformations. Although previous work has shown the vulnerability of adversarial semantic transformations and defend them, provable robustness against adversarial semantic transformations is a relatively novel topic. The interval-propagation-based bounding strategy provides the first verification approach against rotation [34]. The linear-programming-based strategy has been utilized to certify model robustness against geometry transformations [3, 27]. Recent work [10] has applied the function smoothing scheme to provide provable robustness against general transformations. However, this work can only certifiably defend against randomized attacks that draw the transformation parameter from a fixed distribution, whereas our framework can certifiably defend against exhaustive attacks that draw an arbitrary transformation parameter. Moreover, our framework outperforms theirs significantly in terms of certified robustness.

3 Function Smoothing for semantic Transformations

We now state the *general* theorem for certifying robustness under semantic transformations. This theorem is general—in the next sections we leverage this result for smoothing strategies with different distributions and against different semantic transformations.

3.1 Problem Setup

We denote the space of inputs as $\mathcal{X} \subseteq \mathbb{R}^d$, the set of labels as $\mathcal{Y} = \{1, \ldots, C\}$ (where $C \geq 2$ is the number of classes) and denote noise space by $\mathcal{Z} \subseteq \mathbb{R}^m$. We use the notation $\mathbb{P}_X$ to denote the probability measure induced by the random variable X and write f_X to denote the probability density function with respect to a measure μ. For a measurable set S we denote its probability by $\mathbb{P}_X(S)$. Finally, we refer to base classifiers as general deterministic functions $h\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$, mapping feature vectors to class probabilities.

Semantic Transformations We model semantic transformations as general deterministic functions $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$, mapping data points to a transformed version with a $\mathcal{Z}$-valued parameter α. Examples of such transformations include rotations and translations as discussed in Section 5.

Function Smoothing is a framework for constructing a new classifier from an arbitrary base classifier h by introducing randomness to input transformations. Given an input x, the smoothed classifier predicts the class that h is most likely to return when the input is perturbed by some random transformation. Using this notion of function smoothing and transformations, we can define a smoothed classifier:

Definition 1 (ε-Smoothed Classifier)**.** *Suppose we are given a transform $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$, a random variable $\varepsilon \sim \mathbb{P}_{\varepsilon}$ taking values in $\mathcal{Z}$ and a base classifier $h\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. We define the ε-smoothed classifier $g^{\varepsilon}\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ as the expectation with respect to the smoothing distribution ε*

$$g^{\varepsilon}(x) := \mathbb{E}(h(\phi(x,\, \varepsilon))). \tag{1}$$

The next definition allows us to quantify the confidence of a smoothed classifier in making a prediction at an input $x \in \mathcal{X}$.

Definition 2. *Let $x \in \mathcal{X}$, $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ an input transform, $h\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ a base classifier, $c_A \in \mathcal{Y}$ and $p_A,\, p_B \in [0,\, 1]$. We say that the ε-smoothed classifier g^{ε} is $(p_A,\, p_B)$-confident at x if*

$$g^{\varepsilon}(x)_{c_A} \geq p_A \geq p_B \geq \max_{c \neq c_A} g^{\varepsilon}(x)_c. \tag{2}$$

A further building block of our framework is the notion of level sets, which is connected to statistical hypothesis testing and constitute rejection regions of likelihood ratio tests.

Definition 3. *Let $\varepsilon_0 \sim \mathbb{P}_0$, $\varepsilon_1 \sim \mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a measure μ. For $t \geq 0$ we define strict lower and lower level sets as*

$$\underline{S_t} := \{z \in \mathcal{Z}\colon\ \Lambda(z) < t\}\,, \quad \overline{S_t} := \{z \in \mathcal{Z}\colon\ \Lambda(z) \leq t\}\,, \quad \textit{where} \quad \Lambda(z) := \frac{f_1(z)}{f_0(z)}. \tag{3}$$

3.2 Robustness Guarantee for resolvable Transformations

In this section, we provide a generic robustness guarantee for resolvable transformations. The succeeding section leverages this result and develops a framework to certifying robustness to a more general family of semantic transformations. Informally, we call a semantic transform resolvable if we can separate transformation parameters from inputs with a function that acts on parameters and satisfies certain regularity conditions. The next definition makes this notion precise.

Definition 4. *A transform $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ is called resolvable if for any $\alpha \in \mathcal{Z}$ there exists a resolving function $\gamma_{\alpha}\colon \mathcal{Z} \to \mathcal{Z}$ that is injective, continuously differentiable, has non-vanishing Jacobian and*

$$\phi(\phi(x,\, \alpha),\, \beta) = \phi(x,\, \gamma_{\alpha}(\beta)) \qquad x \in \mathcal{X},\, \beta \in \mathcal{Z}. \tag{4}$$

Given an input $x \in \mathcal{X}$, suppose that the ε-smoothed classifier predicts x to be of class c_A with probability at least p_A and the second most likely class with probability at most p_B. Our goal is to derive a robustness condition on transformation parameters α depending on p_A and p_B and ε such that whenever the parameter α statisfies this condition, it is guaranteed that

$$\arg\max_k g^{\varepsilon}_k(\phi(x,\, \alpha)) = \arg\max_k g^{\varepsilon}_k(x). \tag{5}$$

In other words, the prediction of the smoothed classifier can never be changed by applying the transform ϕ with parameters α that satisfy the robustness condition.

Theorem 1. *Let $\varepsilon_0 \sim \mathbb{P}_0$ and $\varepsilon_1 \sim \mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a measure μ on $\mathcal{Z}$. Let $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be a transform and suppose that the ε_0-smoothed classifier g^{ε_0} is $(p_A,\, p_B)$-confident at $x \in \mathcal{X}$ for $k_A \in \mathcal{Y}$. Let $\zeta\colon \mathbb{R}_{\geq 0} \to [0,\, 1]$ be the function defined by $\zeta(t) := \mathbb{P}_0(\overline{S}_t)$*

and denote by $\zeta^{-1}(p) := \inf\{t \geq 0\colon\ \zeta(t) \geq p\}$ *its generalized inverse. For* $t \geq 0$ *and* $p \in [0,\, 1]$ *we define the function* ξ *by*

$$\xi(t,\, p) := \sup\{\mathbb{P}_1(S)\colon\ \underline{S}_t \subseteq S \subseteq \overline{S}_t,\, \mathbb{P}_0(S) \leq p\}. \tag{6}$$

If the robustness condition

$$1 - \xi(\zeta^{-1}(1 - p_B),\, 1 - p_B) < \xi(\zeta^{-1}(p_A),\, p_A) \tag{7}$$

is satisfied, then it is guaranteed that $\arg\max_k g_k^{\varepsilon_1}(x) = \arg\max_k g_k^{\varepsilon_0}(x)$.

This theorem is a more general version of what is proved in Cohen et al. [5], and its generality allows us to analyze cases beyond what is supported by Cohen et al. [5] in Section 4 and Section 5. A detailed proof is provided in Appendix A. From this statement it is not immediately clear how one can obtain the robustness guarantee (5). However if the transform ϕ is resolvable, the following result is more intuitive.

Corollary 1. *Suppose that the transform* ϕ *in Theorem 1 is resolvable with resolving function* γ_α*. Let* $\alpha \in \mathcal{Z}$ *and set* $\varepsilon_1 := \gamma_\alpha(\varepsilon_0)$ *in the definition of the functions* ζ *and* ξ*. Then, if* α *satisfies condition (7) it is guaranteed that* $\arg\max_k g_k^{\varepsilon_0}(\phi(x, \alpha)) = \arg\max_k g_k^{\varepsilon_0}(x)$.

This corollary requires the transform ϕ to be *resolvable*. However, many transforms such as rotations do not inhibit this property. In the next section, we show how this result can be leveraged to get a usable robustness certificate for this more general class of transformations.

3.3 Robustness Guarantee for differentially resolvable Transformations

Common semantic transformations such as rotations and scaling do not fall into the category of resolvable transformations due to their use of interpolation. For this reason, in this section we provide a technique to certify this more general family of transforms, leveraging the condition from Theorem 1. We define a transform ϕ to be differentially resolvable if it can be written in terms of a resolvable transform ψ and a parameter mapping δ. The next definition makes this intuition precise.

Definition 5. *Let* $\phi\colon \mathcal{X} \times \mathcal{Z}_\phi \to \mathcal{X}$ *be a transform with noise space* $\mathcal{Z}_\phi$ *and let* $\psi\colon \mathcal{X} \times \mathcal{Z}_\psi \to \mathcal{X}$ *be a resolvable transform with noise space* $\mathcal{Z}_\psi$*. We say that* ϕ *can be resolved by* ψ *if for any* $x \in \mathcal{X}$ *there exists function* $\delta_x\colon \mathcal{Z}_\phi \times \mathcal{Z}_\phi \to \mathcal{Z}_\psi$ *such that for any* $\beta \in \mathcal{Z}_\phi$

$$\phi(x,\, \alpha) = \psi(\phi(x,\, \beta),\, \delta_x(\alpha,\, \beta)). \tag{8}$$

This definition leaves open a certain degree of freedom with regards to the choice of the resolvable transform ψ. For example, we can choose the resolvable transform corresponding to additive noise

$$\psi\colon \mathcal{X} \times \mathcal{X} \to \mathcal{X},\, (x,\, \delta) \mapsto x + \delta, \tag{9}$$

which lets us write any transform ϕ as $\phi(x,\, \alpha) = \phi(x,\, \beta) + (\phi(x,\, \alpha) - \phi(x,\, \beta)) = \psi(\phi(x,\, \beta),\, \delta)$ with $\delta = (\phi(x,\, \alpha) - \phi(x,\, \beta))$. The next theorem uses the result in Theorem 1 by bootstrapping the robustness condition arising from certifying the resolvable transform ψ given inputs transformed by ϕ with parameters $\{\alpha_i\}_{i=1}^N$ sampled from the set of parameters $\mathcal{S}$ that we wish to certify.

Theorem 2. *Let* $\phi\colon \mathcal{X} \times \mathcal{Z}_\phi \to \mathcal{X}$ *be a transform that is resolved by* $\psi\colon \mathcal{X} \times \mathcal{Z}_\psi \to \mathcal{X}$*. Let* $\varepsilon \sim \mathbb{P}_\varepsilon$ *be a* $\mathcal{Z}_\psi$*-valued random variable and suppose that the* g^ε*-smoothed classifier given by* $g^\varepsilon(x) = \mathbb{E}(h(\psi(x,\, \varepsilon)))$ *predicts* $k_A = \arg\max_k g_k^\varepsilon(x)$*. Let* $\mathcal{S} \subseteq \mathcal{Z}_\psi$ *and* $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$ *be a set of transformation parameters such that for any* i*,* g^ε *is* $(p_A^{(i)},\, p_B^{(i)})$*-confident at* $\phi(x,\, \alpha_i)$*. Then there exists a set* $\Delta^* \subseteq \mathcal{Z}_\psi$ *with the property that, if for any* $\alpha \in \mathcal{S}$*,* $\exists \alpha_i$ *with* $\delta_x(\alpha,\, \alpha_i) \in \Delta^*$ *it is guaranteed that*

$$g_{k_A}^\varepsilon(\phi(x,\, \alpha)) > \max_{k \neq k_A} g_k^\varepsilon(\phi(x,\, \alpha)) \tag{10}$$

The key of using this theorem for a specific transformation is to choose the resolvable transformation ψ that can enable a tight calculation of Δ^* under a specific way of sampling $\{\alpha_i\}_{i=1}^N$. In the next section, we show how to make these design decisions for rotation and scaling transformations.

4 Smoothing Strategies for resolvable Transforms

The robustness condition for resolvable transforms in Theorem 1 is generic and leaves two questions open: (1) *How can we instantiate this general theorem with different smoothing distributions?* (2) *How can we apply it to specific semantic transformations?* Here, we focus on the first question and discuss transformation-specific smoothing strategies in the next section.

Previous work mainly provide results for cases in which this distribution is Gaussian, while extending it to other distributions is non-trivial. In this section, we conduct a novel analysis and provide results for a range of distributions, and discuss their differences. As we will see, *for different scenarios, different distributions behave differently and can certify different radii.* Here, we instantiate Theorem 1 with an arbitrary transform ϕ and with $\varepsilon_1 := \alpha + \varepsilon_0$ where ε_0 is the smoothing distribution and α the transformation parameter. The robust radius is then derived by solving condition (7) for α.

Table 1: Comparison of certification radii. The variance is set to 1 and noise dimensionality is $m = 1$. For Exponential and Folded Gaussian, the perturbations are restricted to $\mathbb{R}_{\geq 0}$.

Distribution	Robust Radius
$\mathcal{N}(0,\ 1)$	$\Phi^{-1}(p_A)$
Exp(1)	$-\log(2 - 2p_A)$
$\mathcal{L}(0,\ 1/\sqrt{2})$	$-\log(2 - 2p_A)/\sqrt{2}$
$\mathcal{U}([-\sqrt{3},\ -\sqrt{3}])$	$2\sqrt{3} \cdot (p_A - 1/2)$
$\lvert\mathcal{N}(0,\ \sqrt{\frac{\pi}{\pi-2}})\rvert$	$\sqrt{\frac{\pi}{\pi-2}} \cdot \left(\Phi^{-1}\left(\frac{1+p_A}{2}\right) - \Phi^{-1}\left(\frac{3}{4}\right)\right)$

Due to space limit, we summarize only different certification radii in the main body while leaving the theorems and proofs to the Appendix. Note that the contribution of this work is not merely these results on different smoothing distributions but, more importantly, the *joint study between different smoothing mechanisms and different semantic transformations.*

Comparison of Smoothing Noise Distributions In order to compare the different radii for a fixed base classifier, we assume that *the smoothed classifier g^{ε} always has the same confidence p_A for noises with equal variance.* We summarize our findings while leaving details to the Appendix.

1. *Exponential noise can provide larger robust radius.* We notice that smoothing with exponential noise generally allows for larger adversarial perturbations than other distributions. We also observe that, while all distributions behave similar for low confidence levels, it is only non-uniform noise distributions that converge towards $+\infty$ when $p_A \to 1$ and exponential noise converges quickest.

2. *Additional knowledge can lead to larger robust radius.* When we have additional information on the transformation, e.g., all perturbations in Gaussian blur are positive, we can take advantage of it to certify larger radii. For example, under this assumption, we can use folded Gaussian noise for smoothing instead of a standard Gaussian, resulting in a larger radius.

5 Robustness against adversarial semantic Transformations

In this section, we provide approaches to certify a range of different semantic transformations building on our theoretical results in Sections 3.2 and 3.3. Some transformations can be certified by directly applying Theorem 1; however, many transformations need more engaged analysis and it is delicate and non-trivial to reason about which smoothing distributions fit a given transformation better. We state all results here and provide proofs in supplementary materials.

Gaussian Blur is a transformation that is widely used in image processing to reduce noise and image detail. Mathematically speaking, applying Gaussian blur amounts to convolving an image with a Gaussian function

$$G_{\alpha}(k) = \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{k^2}{2\alpha}\right) \tag{11}$$

where $\alpha > 0$ is the squared kernel radius. For $x \in \mathcal{X}$, we define Gaussian blur as $\phi_B \colon \mathcal{X} \times \mathbb{R}_{\geq 0} \to \mathcal{X}$:

$$\phi_B(x,\, \alpha) = x * G_\alpha \tag{12}$$

where $*$ denotes the convolution operator. The following Lemma shows that Gaussian blur is an *additive transform* and hence resolvable.

Lemma 1. *The gaussian blur transform is additive, $\phi_B(\phi_B(x,\, \alpha),\, \beta) = \phi_B(x,\, \alpha + \beta)$.*

We notice that the Gaussian blur transform only uses positive parameters. We therefore consider uniform noise on $[0,\, a]$ for $a > 0$, folded Gaussians and exponential distribution for smoothing.

Brightness and Contrast transformations first add a constant value $b \in \mathbb{R}$ to every pixel and then change the contrast by multiplying each pixel with a positive factor e^k, for some $k \in \mathbb{R}$. Given an image $x \in \mathcal{X}$, we define the brightness and contrast transform $\phi_{BC} \colon \mathcal{X} \times \mathbb{R}^2 \to \mathcal{X}$ as

$$\phi_{BC}(x,\, k,\, b) = e^k(x + b) \tag{13}$$

where $k,\, b \in \mathbb{R}$ are contrast and brightness parameters. We notice that in general ϕ_{BC} is not an additive transform and we cannot directly apply Theorem 1.[1] However, if the parameters k and b are sampled from independent Gaussians, we can circumvent the issue of non-additivity using the following idea. Given $\varepsilon_0 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, \tau^2))$, we derive a distribution ε_1 such that using the smoothed classifier g^{ε_0} to classifiy $\phi_{BC}(x,\, \alpha)$ is the same as using $g^{\alpha+\varepsilon_1}$ to classify the original input x. We then show a connection between the confidence of g^{ε_0} and g^{ε_1} such that we can apply Theorem 1 to the random variables ε_0 and ε_1. These relations allow us to certify the brightness and contrast transformation. The following Lemmas justify this approach.

Lemma 2. *Let $\varepsilon_0 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, \tau^2))$, $\alpha = (k,\, b)^T \in \mathbb{R}^2$ and $\varepsilon_1 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, e^{-2k}\tau^2))$. Then, for all $x \in \mathcal{X}$, it holds that $g^{\varepsilon_0}(\phi_{BC}(x,\, \alpha)) = g^{\alpha+\varepsilon_1}(x)$.*

Lemma 3. *Let $x \in \mathcal{X}$, $k \in \mathbb{R}$, $\varepsilon_0 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, \tau^2))$ and $\varepsilon_1 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, e^{-2k}\tau^2))$. Suppose that $g_c^{\varepsilon_0}(x) \geq p$ for some $p \in [0,\, 1]$ and $c \in \mathcal{Y}$. Then*

$$g_c^{\varepsilon_1}(x) \geq \begin{cases} 2\Phi\left(e^k \Phi^{-1}\left(\frac{1+p}{2}\right)\right) - 1 & k \leq 0 \\ 2\left(1 - \Phi\left(e^k \Phi^{-1}(1 - \frac{p}{2})\right)\right) & k > 0. \end{cases} \tag{14}$$

Now suppose that g^{ε_0} makes the prediction c_A at x with probability at least p_A. Then, the above Lemma tells us that g^{ε_1} is similarly confident in predicting the same class. Given this confidence levels, we instantiate Theorem 1 with the random variables ε_0 and ε_1 to get a robustness condition. The next Lemma makes this condition explicit.

Lemma 4. *Let ε_0 and ε_1 be the random variables given in Lemmas 2 and 3 and suppose that $g_{c_A}^{\varepsilon_1}(x) > g_{c_B}^{\varepsilon_1}(x) = \max_{c \neq c_A} g_c^{\varepsilon_1}(x)$. Then, it is guaranteed that $c_A = \arg\max_c g_c^{\varepsilon_0}(\phi_{BC}(x,\, \alpha))$ as long as $\alpha = (k,\, b)^T$ satisfies*

$$\sqrt{\left(\frac{k}{\sigma}\right)^2 + \left(\frac{b}{e^{-k}\tau}\right)^2} < \frac{1}{2}\left(\Phi^{-1}\left(g_{c_A}^{\varepsilon_1}(x)\right) - \Phi^{-1}\left(g_{c_B}^{\varepsilon_1}(x)\right)\right) \tag{15}$$

Translation Let $\bar{\phi}_T \colon \mathcal{X} \times \mathbb{Z}^2 \to \mathcal{X}$ be the transform moving an image k_1 pixels to the right and k_2 pixels to the bottom with reflection padding. In order to handle continuous noise distributions, we define the translation transform $\phi_T \colon \mathcal{X} \times \mathbb{R}^2 \to \mathcal{X}$ as $\phi_T(x,\, \alpha) = \bar{\phi}_T(x,\, [\alpha])$ where $[\cdot]$ denotes rounding to the nearest integer, applied element-wise. With this definition we note that ϕ_T is an *additive transform* and thus resolvable allowing us to directly apply Theorem 1 and derive robustness bounds. We note that if we use black-padding

[1] We remark that both brightness and contrast alone are additive, whereas their composition is not.

instead of reflection-padding, the transform is not additive. However, since the number of possible translations is finite, another possibility is to use a simple brute force approach that can handle black-padding.

Rotations and Scaling We outline the basic principles to certifying robustness for these two transforms and leave detailed descriptions to Appendix F. Since both transforms are not resolvable due to their use of bilinear interpolation, we instantiate Theorem 2 presented in Section 3.3 in the following way. First, we observe that both transforms can be resolved by the additive transform $\psi\colon \mathcal{X} \times \mathcal{X} \to \mathcal{X}$ defined by $(x,\,\delta) \mapsto \psi(x,\,\delta) := x + \delta$. Choosing isotropic Gaussian noise $\varepsilon \sim \mathcal{N}(0,\,\sigma^2)$ as smoothing noise then leads the condition that the maximum ℓ_2-sampling error between the interval $\mathcal{S} = [a,\,b]$ (which is to be certified) and the sampled parameters α_i must be bounded by a radius r. The next corollary makes this intuition precise.

Corollary 2. *Let $\psi(x,\,\delta) = x + \delta$ and let $\varepsilon \sim \mathcal{N}(0,\,\sigma^2\mathbb{1}_d)$. Furthermore, let ϕ be a transform with parameters in $\mathcal{Z}_\phi \subseteq \mathbb{R}^m$ and let $\mathcal{S} \subseteq \mathcal{Z}_\phi$ and $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$. Let $k_A \in \mathcal{Y}$ and suppose that for any i, the ε-smoothed classifier $g^\varepsilon(x) := \mathbb{E}[h(x+\varepsilon)]$ is $(p_A^{(i)},\,p_B^{(i)})$-confident at $\phi(x,\,\alpha_i)$ for k_A. Then, the set Δ^* in Theorem 2 is given by the open ℓ_2-ball of radius R around the origin*

$$\Delta^* \equiv B_R(0) \subseteq \mathbb{R}^d \qquad \textit{with} \qquad R := \frac{\sigma}{2} \min_{1\leq i \leq N} \left(\Phi^{-1}\left(p_A^{(i)}\right) - \Phi^{-1}\left(p_B^{(i)}\right)\right) \tag{16}$$

and it holds that $\forall \alpha \in \mathcal{S}\colon\ k_A = \arg\max_k g_k^\varepsilon(\phi(x,\,\alpha))$ whenever

$$M_{\mathcal{S}} := \max_{\alpha \in \mathcal{S}} \min_{1\leq i\leq N} \|\phi(x,\,\alpha) - \phi(x,\,\alpha_i)\|_2 < R. \tag{17}$$

In essence, this corollary states that if the smoothed classifier classifies a collection of sampled transformations consistent with the original inputs, then it will be guaranteed to make the same prediction on any similar transformation of the same input.

In order to use this result for rotation and scaling, we first sample N parameters α_i and then compute the maximum sampling error, $M_{\mathcal{S}}$, on N sampled parameters. This is not trivial. Our approach is based on computing an upper bound on $M_{\mathcal{S}}$. For transformations such as rotations we compute this upper bound based on the Lipschitz constant. We refer the reader to Appendices F.4, F.2 and F.3 for details. In the second step, we compute the robustness radius for each of those samples and compare whether the smallest such radius is bigger than $M_{\mathcal{S}}$. If this is true, then the smoothed classifier is guaranteed to make a consistent prediction for all parameters in the set $\mathcal{S}$.

6 Experiments

We validate our framework to certifying robustness over semantic transformations experimentally on the publicly available MNIST, CIFAR-10 and ImageNet-1k datasets. We compare with state-of-the-arts for each transformation and highlight our main results.

Setup We refer the reader to Appendix H.2 for experimental details.[2] On ImageNet we use a pretrained ResNet-50 architecture [15] as base classifier while on CIFAR-10 we use ResNet-110 and on MNIST we use a small CNN.

Evaluation Procedure In adversarially robust classification we are interested in the *robust accuracy* at radius r. This metric is defined as the fraction of the test set, which is classified correctly with a prediction that is certifiably robust within a ball of radius r. However, since we use randomized smoothing classifiers for inference, computing this quantity exactly is not possible without further assumptions on the base classifier. Instead, we report the *approximate robust accuracy* on a subset of the test set using adapted versions of Predict and Certify presented in [5]. The error rate is set to $\alpha = 0.001$ such that the certification holds with probability at least $1 - \alpha$.

[2] Our implementation is publicly available at `https://github.com/AI-secure/semantic-randomized-smoothing`

Table 2: Overview of the best robust accuracy for different semantic transformations. The current state-of-the-art is **bolded**.

Transformation	Dataset	Robustness Radii	Robust Acc.	
			Ours	Literature
Gaussian Blur	MNIST	Kernel Rad. $\alpha \leq 9$	$\mathbf{90}.4\%$	-
	CIFAR-10	Kernel Rad. $\alpha \leq 9$	$\mathbf{52}.0\%$	-
	ImageNet	Kernel Rad. $\alpha \leq 9$	$\mathbf{47}.0\%$	-
Translation (Reflection Pad.)	MNIST	$\sqrt{\Delta x^2 + \Delta y^2} \leq 5$	$\mathbf{96}.8\%$	-
	CIFAR-10	$\sqrt{\Delta x^2 + \Delta y^2} \leq 5$	$\mathbf{84}.8\%$	-
	ImageNet	$\sqrt{\Delta x^2 + \Delta y^2} \leq 5$	$\mathbf{63}.0\%$	-
Brightness	MNIST	$b \pm 0.1({}^{25.5}/_{255})$	$\mathbf{98}.6\%$	-
	CIFAR-10	$b \pm 0.1({}^{25.5}/_{255})$	$\mathbf{84}.2\%$	-
	ImageNet	$b \pm 0.1({}^{25.5}/_{255})$	$\mathbf{64}.0\%$	64.0%[a] [10]
Contrast	MNIST	$c \pm 30\%$	$\mathbf{98}.6\%$	-
	CIFAR-10	$c \pm 30\%$	$\mathbf{76}.8\%$	-
	ImageNet	$c \pm 30\%$	$\mathbf{56}.0\%$	45.0%[a] [10]
Contrast and Brightness	MNIST	$c \pm 20\%, b \pm 0.2({}^{51}/_{255})$	$\mathbf{98}.6\%$	-
	CIFAR-10	$c \pm 20\%, b \pm 0.2({}^{51}/_{255})$	$\mathbf{77}.4\%$	-
	ImageNet	$c \pm 20\%, b \pm 0.2({}^{51}/_{255})$	$\mathbf{57}.0\%$	-
Rotation	MNIST	$\pm 30°$	$\mathbf{95}.6\%$	87.0% [3]
	CIFAR-10	$\pm 10°$	$\mathbf{63}.8\%$	62.5% [3]
	ImageNet	$\pm 10°$	$\mathbf{33}.0\%$	9.0%[a] [10]
Scaling	MNIST	$\pm 20\%$	$\mathbf{96}.8\%$	-
	CIFAR-10	$\pm 20\%$	$\mathbf{58}.4\%$	-
	ImageNet	$\pm 15\%$	$\mathbf{31}.0\%$	-

[a] For brightness on ImageNet, [10] uses a slightly smaller radius than $20.37/255$; for contrast on ImageNet, [10] uses a smaller radius than $\pm 21\%$; for rotation on ImageNet, [10] uses a smaller radius than $8.13°$.

Main Results As summarized in Table 2, we observe that across transformations, our framework significantly outperforms state-of-the-art, if present, in terms of robust accuracy. In particular, for Gaussian Blur, translation with reflection padding, scaling and the compositional contrast and brightness, to our best knowledge, we are the first to provide their certified robust accuracy on ImageNet.

Noise Distribution Tuning Our theoretical results allow us to derive robustness bounds for different types of noise distributions. Therefore we explore smoothing using exponential, uniform and Gaussian distributions for a fixed variance level. We use Gaussian blur to showcase our findings. We leave the detailed result to Appendix H.3 where we show that smoothing with exponential distribution typically performs best, confirming our theoretical analysis.

Summary of Results for specific Transformations In the following, we summarize the experimental setup and observations for our transformation-specific results. In addition to Table 2, we provide detailed results for each specific transform in the appendix.

1. For Gaussian Blur we choose the exponential distribution $\text{Exp}(1/\lambda)$ for smoothing. During training, we use the same distribution to sample blurring parameters for data augmentation. We repeat the same

experiment for different values of λ. We observe a pronounced trade-off between robust and clean accuracy: *with more noise, we achieve higher certification radii, which comes at the cost of decreased clean accuracy*. This phenomenon is in line with what has been reported in previous work.

2. For brightness and contrast transformations we use Gaussian noise for data augmentation during training and smoothing during inference. We report robust and clean accuracy for (1) brightness, (2) contrast, and (3) compositional brightness and contrast transformations in the Appendix.
3. We certify translation transformations with black-padding and reflection padding. For black-padding we only use brute-force enumeration for certification, while for reflection padding we also use the randomized smoothing framework with Gaussian noise varying the standard deviation. We observe that in general, randomized smoothing does slightly better than the enumeration approach.
4. To certify rotations we obtain base classifiers with data augmentation during training. Specifically, we rotate images with angles α_{train} sampled uniformly at random and add Gaussian noise with variance σ^2_{train}. During inference with smoothed classifiers we use isotropic Gaussian noise for smoothing with variance σ^2_{test}. We notice that, generally, an increase in σ^2_{train} hurts clean accuracy, while it benefits robust accuracy.
5. The procedure to certify scaling transformation is similar in nature to rotations. That is, we apply data augmentation during training and scale images with a scaling parameter s_{train} sampled uniformly at random. After scaling, we add iid Gaussian noise with variance σ^2_{train}. We then use randomized smoothing classifiers with additive Gaussian noise to perform inference and obtain robustness bounds. Although we observe a clear trade-off between clean accuracy and noise variance, the relation between robust accuracy and variance is not obvious.

7 Conclusion

In this paper, we have provided a unified framework for certifying ML robustness against general adversarial transformations using function smoothing. We have categorized the semantic transformations as *resolvable* and *differentially resolvable* transformations, and have shown that our theoretical results can be used to derive provable robustness bounds against all transformations. Extensive experiments show that our transformation-specific smoothing approaches significantly outperform the state-of-the-art or, if no previous work exists, set new baselines.

References

[1] M. Abramowitz and I. A. (Eds.) Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*. New York: Dover, 1972.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[3] Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. Certifying geometric robustness of neural networks. In *Advances in Neural Information Processing Systems*, pages 15287–15297, 2019.

[4] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*, 2019.

[5] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Krishnamurthy (Dj) Dvijotham, Jamie Hayes, Borja Balle, Zico Kolter, Chongli Qin, Andras Gyorgy, Kai Xiao, Sven Gowal, and Pushmeet Kohli. A framework for robustness certification of smoothed classifiers using f-divergences. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJlKrkSFPH.

[8] Logan Engstrom, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 1(2):3, 2017.

[9] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[10] Marc Fischer, Maximilian Baader, and Martin Vechev. Statistical verification of general perturbations by gaussian smoothing. *https://openreview.net/forum?id=B1eZweHFwr*, 2020.

[11] Amin Ghiasi, Ali Shafahi, and Tom Goldstein. Breaking certified defenses: semantic adversarial examples with spoofed robustness certificates. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJxdTxHYvB.

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[13] Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

[17] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619, 2018.

[18] Jörn-Henrik Jacobsen, Jens Behrmann, Nicholas Carlini, Florian Tramer, and Nicolas Papernot. Exploiting excessive invariance caused by norm-bounded adversarial robustness. *arXiv preprint arXiv:1903.10484*, 2019.

[19] J. Jacod and P.E. Protter. *Probability Essentials*. Springer, 2000.

[20] J Zico Kolter and Eric Wong. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017.

[21] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[22] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[23] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE, 2019.

[24] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems*, pages 9459–9469, 2019.

[25] Linyi Li, Zexuan Zhong, Bo Li, and Tao Xie. Robustra: training provable robust neural networks over reference adversarial space. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 4711–4717. AAAI Press, 2019.

[26] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018.

[27] Jeet Mohapatra, Pin-Yu Chen, Sijia Liu, Luca Daniel, et al. Towards verifying robustness of neural networks against semantic perturbations. *arXiv preprint arXiv:1912.09533*, 2019.

[28] Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisỳ, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. DeepStack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 356(6337):508–513, 2017.

[29] Kexin Pei, Yinzhi Cao, Junfeng Yang, and Suman Jana. Deepxplore: Automated whitebox testing of deep learning systems. In *proceedings of the 26th Symposium on Operating Systems Principles*, pages 1–18, 2017.

[30] Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10877–10887, 2018.

[31] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 11289–11300, 2019.

[32] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A convex relaxation barrier to tight robust verification of neural networks. *arXiv preprint arXiv:1902.08722*, 2019.

[33] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 10 2017.

[34] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL):41, 2019.

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[36] Jiaye Teng, Guang-He Lee, and Yang Yuan. ℓ_1 adversarial robustness certificates: a randomized smoothing approach, 2020. URL https://openreview.net/forum?id=H1lQIgrFDS.

[37] Vincent Tjeng, Kai Y Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. 2018.

[38] Liang Tong, Bo Li, Chen Hajaj, Chaowei Xiao, Ning Zhang, and Yevgeniy Vorobeychik. Improving robustness of ml classifiers against realizable evasion attacks using conserved features. In *Proceedings of the 28th USENIX Conference on Security Symposium*, SEC'19, page 285–302, USA, 2019. USENIX Association. ISBN 9781939133069.

[39] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[40] Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S Dhillon, and Luca Daniel. Towards fast computation of certified robustness for relu networks. *arXiv preprint arXiv:1804.09699*, 2018.

[41] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. *arXiv preprint arXiv:1805.12514*, 2018.

[42] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

[43] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

[44] Weilin Xu, Yanjun Qi, and David Evans. Automatically evading classifiers. In *Proceedings of the 2016 Network and Distributed Systems Symposium*, 2016.

[45] Greg Yang, Tony Duan, Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. *arXiv preprint arXiv:2002.08118*, 2020.

[46] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang. Macer: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rJx1Na4Fwr.

A Proof of Theorem 1 and Corollary 1

Here we provide the proof for Theorem 1. For that purpose, recall the following definitions from the main part of this paper:

Definition 1 (restated)**.** *Suppose we are given a transform $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$, a random variable $\varepsilon \sim \mathbb{P}_\varepsilon$ taking values in $\mathcal{Z}$ and a base classifier $h\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$. We define the ε-smoothed classifier $g^\varepsilon\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ as the expectation with respect to the smoothing distribution ε*

$$g^\varepsilon(x) := \mathbb{E}[h(\phi(x,\, \varepsilon))]. \tag{18}$$

Definition 2 (restated)**.** *Let $x \in \mathcal{X}$, $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ an input transform, $h\colon \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ a base classifier, $c_A \in \mathcal{Y}$ and $p_A,\, p_B \in [0,\, 1]$. We say that the ε-smoothed classifier g^ε is $(p_A,\, p_B)$-confident at x if*

$$g^\varepsilon(x)_{c_A} \geq p_A \geq p_B \geq \max_{c \neq c_A} g^\varepsilon(x)_c. \tag{19}$$

Definition 3 (restated)**.** *Let $\varepsilon_0 \sim \mathbb{P}_0$, $\varepsilon_1 \sim \mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a measure μ. For $t \geq 0$ we define lower and strict lower level sets as*

$$\underline{S}_t := \{z \in \mathcal{Z}\colon\ \Lambda(z) < t\}\,, \quad \overline{S}_t := \{z \in \mathcal{Z}\colon\ \Lambda(z) \leq t\}\,, \quad \textit{where} \quad \Lambda(z) := \frac{f_1(z)}{f_0(z)}. \tag{20}$$

Definition 4 (restated)**.** *A transform $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ is called resolvable if for any $\alpha \in \mathcal{Z}$ there exists a resolving function $\gamma_\alpha\colon \mathcal{Z} \to \mathcal{Z}$ that is injective, continuously differentiable, has non-vanishing Jacobian and*

$$\phi(\phi(x,\, \alpha),\, \beta) = \phi(x,\, \gamma_\alpha(\beta)) \qquad x \in \mathcal{X},\, \beta \in \mathcal{Z}. \tag{21}$$

A.1 Auxiliary Lemmas

Lemma A.1. *Let ε_0 and ε_1 be random variables taking values in $\mathcal{Z}$ and with probability density functions f_0 and f_1 with respect to a measure μ. Denote by Λ the likelihood ratio $\Lambda(z) = f_1(z)/f_0(z)$. For $p \in [0,\, 1]$ let $t_p := \inf\{t \geq 0\colon\ \mathbb{P}_0(\overline{S}_t) \geq p\}$. Then, it holds that*

$$\mathbb{P}_0\left(\underline{S}_{t_p}\right) \leq p \leq \mathbb{P}_0\left(\overline{S}_{t_p}\right). \tag{22}$$

Proof. We first show the RHS of inequality (22). This follows directly from the definition of t_p if we show that the function $t \mapsto \mathbb{P}_0\left(\overline{S}_t\right)$ is right-continuous. For that purpose, let $t \geq 0$ and let $\{t_n\}_n$ be a sequence in $\mathbb{R}_{\geq 0}$ such that $t_n \downarrow t$. Define the sets $A_n := \{z\colon\ \Lambda(z) \leq t_n\}$ and note that $A_{n+1} \subseteq A_n$. Clearly, if $z \in \overline{S}_t$, then $\forall\, n\colon\ \Lambda(z) \leq t \leq t_n$, thus $z \in \cap_n A_n$ and hence $\overline{S}_t \subseteq \cap_n A_n$. If on the other hand $z \in \cap_n A_n$, then $\forall n\colon\ \Lambda(z) \leq t_n \to t$ as $n \to \infty$ and thus $z \in \overline{S}_t$, yielding $\overline{S}_t = \cap_n A_n$. Hence for any $t \geq 0$ we have that

$$\lim_{n \to \infty} \mathbb{P}_0\left(A_n\right) = \mathbb{P}_0\left(\bigcap_n A_n\right) = \mathbb{P}_0\left(\underline{S}_t\right). \tag{23}$$

Thus, the function $t \mapsto \mathbb{P}_0\left(\overline{S}_t\right)$ is right continuous and in particular it follows that $\mathbb{P}_0\left(\overline{S}_{t_p}\right) \geq p$. We now show the LHS of inequality (22). Consider the sets $B_n := \{z\colon\ \Lambda(z) < t_p - {}^1\!/\!_n\}$ and note that $B_n \subseteq B_{n+1}$. Clearly, if $z \in \cup_n B_n$, then $\exists\, n$ such that $\Lambda(z) < t_p - {}^1\!/\!_n < t_p$ and thus $z \in \underline{S}_{t_p}$. If on the other hand $z \in \underline{S}_{t_p}$, then we can choose n large enough such that $\Lambda(z) < t_p - {}^1\!/\!_n$ and thus $z \in \cup_n B_n$ yielding $\underline{S}_{t_p} = \cup_n B_n$. Furthermore, by the definition of t_p and since for any $n \in \mathbb{N}$ we have that $\mathbb{P}_0(B_n) = \mathbb{P}_0\left(\underline{S}_{t_p - {}^1\!/\!_n}\right) < p$ it follows that

$$\mathbb{P}_0\left(\underline{S}_{t_p}\right) = \mathbb{P}_0\left(\bigcup_n B_n\right) = \lim_{n \to \infty} \mathbb{P}_0\left(B_n\right) \leq p \tag{24}$$

concluding the proof. □

Lemma A.2. *Let ε_0 and ε_1 be random variables taking values in $\mathcal{Z}$ and with probability density functions f_0 and f_1 with respect to a measure μ. Let $h\colon \mathcal{Z} \to [0,\,1]$ be a determinstic function. Then, for any $t \geq 0$ the following implications hold:*

(i) For any $S \subseteq \mathcal{Z}$ with $\underline{S_t} \subseteq S \subseteq \overline{S_t}$ it holds that $\mathbb{E}[h(\varepsilon_0)] \geq \mathbb{P}_0(S) \Rightarrow \mathbb{E}[h(\varepsilon_1)] \geq \mathbb{P}_1(S)$.

(i) For any $S \subseteq \mathcal{Z}$ with $\overline{S_t}^c \subseteq S \subseteq \underline{S_t}^c$ it holds that: $\mathbb{E}[h(\varepsilon_0)] \leq \mathbb{P}_0(S) \Rightarrow \mathbb{E}[h(\varepsilon_1)] \leq \mathbb{P}_1(S)$.

Proof. We first prove (i). For that purpose, consider

$$\mathbb{E}[f(\varepsilon_1)] - \mathbb{P}_1(S) = \int h f_1 \, d\mu - \int_S f_1 \, d\mu = \int_{S^c} h f_1 \, d\mu - \left(\int_S (1-h) f_1 \, d\mu\right) \tag{25}$$

$$= \int_{S^c} h \Lambda f_0 \, d\mu - \left(\int_S (1-h)\Lambda f_0 \, d\mu\right) \tag{26}$$

$$\geq t \cdot \int_{S^c} h f_0 \, d\mu - t \cdot \left(\int_S (1-h) f_0 \, d\mu\right) \tag{27}$$

$$= t \cdot \left(\int h f_0 \, d\mu - \int_S f_0 \, d\mu\right) = t \cdot (\mathbb{E}[f(\varepsilon_0)] - \mathbb{P}_0(S)) \geq 0. \tag{28}$$

The inequality in (27) follows from the fact that whenever $z \in S^c$, then $f_1(z) \geq t \cdot f_0(z)$ and if $z \in S$, then $f_1(z) \leq t \cdot f_0(z)$ since S is a lower level set. Finally, the inequality in (28) follows from the assumption. The proof of (ii) is analogous and omitted here. □

A.2 Proof of Theorem 1

Theorem 1 (restated). *Let $\varepsilon_0 \sim \mathbb{P}_0$ and $\varepsilon_1 \sim \mathbb{P}_1$ be $\mathcal{Z}$-valued random variables with probability density functions f_0 and f_1 with respect to a measure μ on $\mathcal{Z}$. Let $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ be a transform and suppose that the ε_0-smoothed classifier g^{ε_0} is $(p_A,\, p_B)$-confident at $x \in \mathcal{X}$ for $k_A \in \mathcal{Y}$. Let $\zeta\colon \mathbb{R}_{\geq 0} \to [0,\,1]$ be the function defined by $\zeta(t) := \mathbb{P}_0(\overline{S}_t)$ and denote by $\zeta^{-1}(p) := \inf\{t \geq 0\colon\ \zeta(t) \geq p\}$ its generalized inverse. For $t \geq 0$ and $p \in [0,\,1]$ we define the function ξ by*

$$\xi(t,\, p) := \sup\{\mathbb{P}_1(S)\colon\ \underline{S}_t \subseteq S \subseteq \overline{S}_t,\, \mathbb{P}_0(S) \leq p\}. \tag{29}$$

If the robustness condition

$$1 - \xi(\zeta^{-1}(1-p_B),\, 1-p_B) < \xi(\zeta^{-1}(p_A),\, p_A) \tag{30}$$

is satisfied, then it is guaranteed that $\arg\max_k g_k^{\varepsilon_1}(x) = \arg\max_k g_k^{\varepsilon_0}(x)$.

Proof. For ease of notation, let $t_A := \zeta^{-1}(p_A)$, $t_B := \zeta^{-1}(1-p_B)$, $\underline{S}_A := \underline{S}_{t_A}$, $\underline{S}_B := \underline{S}_{t_B}$, $\overline{S}_A := \overline{S}_{t_A}$ and $\overline{S}_B := \overline{S}_{t_B}$. We first show that $g_{k_A}^{\varepsilon_1}(x)$ is lower bounded by $\xi(\zeta^{-1}(p_A),\, p_A)$. For that purpose, note that by Lemma A.1 we have that $\zeta(t_A) = \mathbb{P}_0(\overline{S}_A) \geq p_A \geq \mathbb{P}_0(\underline{S}_A)$. Thus, the collection of sets

$$\mathcal{S}_A := \{S \subseteq \mathcal{Z}\colon \underline{S}_A \subseteq S \subseteq \overline{S}_A,\, \mathbb{P}_0(S) \leq p_A\} \tag{31}$$

is not empty. Pick some $A \in \mathcal{S}_A$ arbitrary and note that, since by assumption g^{ε_0} is $(p_A,\, p_B)$-confident at x it holds

$$\mathbb{E}(h_{k_A}(\phi(x,\, \varepsilon_0))) = g_{k_A}^{\varepsilon_0}(x) \geq p_A \geq \mathbb{P}_0(A). \tag{32}$$

Since $\underline{S}_A \subseteq A \subseteq \overline{S}_A$ we can apply part (i) of Lemma A.2 and obtain the lower bound

$$g_{k_A}^{\varepsilon_1}(x) = \mathbb{E}(h_{k_A}(\phi(x,\, \varepsilon_1))) \geq \mathbb{P}_1(A). \tag{33}$$

Since $A \in \mathcal{S}_A$ was arbitrary, we take the $\sup$ over all $A \in \mathcal{S}_A$ and obtain

$$g_{k_A}^{\varepsilon_1}(x) \geq \sup_{A \in \mathcal{S}_A} \mathbb{P}_1(A) = \xi(t_A,\, p_A) = \xi(\zeta^{-1}(p_A),\, p_A). \tag{34}$$

We now show that for any $k \neq k_A$ the prediction $g_k^{\varepsilon_1}(x)$ is upper bounded by $1 - \xi(\zeta^{-1}(1-p_B),\, 1-p_B)$. For that purpose, note that by Lemma A.1 we have that $\zeta(t_B) = \mathbb{P}_0(\overline{S}_A) \geq 1 - p_B \geq \mathbb{P}_0(\underline{S}_B)$. Thus, the collection of sets

$$\mathcal{S}_B := \{S \subseteq \mathcal{Z} \colon \underline{S}_B \subseteq S \subseteq \overline{S}_B,\, \mathbb{P}_0(S) \leq 1 - p_B\} \tag{35}$$

is not empty. Pick some $B \in \mathcal{S}_A$ arbitrary and note that, since by assumption g^{ε_0} is $(p_A,\, p_B)$-confident at x it holds

$$\mathbb{E}(h_k(\phi(x,\, \varepsilon_0))) = g_k^{\varepsilon_0}(x) \leq p_B = 1 - (1 - p_B) \leq 1 - \mathbb{P}_0(B). \tag{36}$$

Since $\underline{S}_B^c \subseteq B^c \subseteq \overline{S}_B^c$ we can apply part (ii) of Lemma A.2 and obtain the upper bound

$$g_k^{\varepsilon_1}(x) = \mathbb{E}(h_k(\phi(x,\, \varepsilon_1))) \leq 1 - \mathbb{P}_1(B). \tag{37}$$

Since $B \in \mathcal{S}_B$ was arbitrary, we take the $\inf$ over all $B \in \mathcal{S}_B$ and obtain

$$g_k^{\varepsilon_1}(x) \leq \inf_{B \in \mathcal{S}_B} (1 - \mathbb{P}_1(B)) = 1 - \xi(t_B,\, 1 - p_B) = 1 - \xi(\zeta^{-1}(1-p_B),\, 1-p_B). \tag{38}$$

combining together (38) and (34), we find the robustness condition, whenever

$$1 - \xi(\zeta^{-1}(1-p_B),\, 1-p_B) < \xi(\zeta^{-1}(p_A),\, p_A) \tag{39}$$

it is guaranteed that

$$g_{k_A}^{\varepsilon_1}(x) > \max_{k \neq k_A} g_k^{\varepsilon_1}(x) \tag{40}$$

which concludes the proof. □

A.3 Proof of Corollary 1

Corollary 1 (restated)**.** *Suppose that the transform ϕ in Theorem 1 is resolvable with resolving function γ_α. Let $\alpha \in \mathcal{Z}$ and set $\varepsilon_1 := \gamma_\alpha(\varepsilon_0)$ in the definition of the functions ζ and ξ. Then, if α satisfies condition (7) it is guaranteed that $\arg\max_k g_k^{\varepsilon_0}(\phi(x,\alpha)) = \arg\max_k g_k^{\varepsilon_0}(x)$.*

Proof. Since ϕ is a resolvable transform, by definition γ_α is injective, continuously differentiable and has non-vanishing jacobian. By Jacobi's transformation formula (see e.g. [19]) it follows that the denisty of ε_1 vanishes outside the image of γ_α and is elsewhere given by

$$f_1(z) = f_0(\gamma_\alpha^{-1}(z))|\det(J_{\gamma_\alpha^{-1}(z)})| \quad \text{for any } z \in \mathrm{Im}(\gamma_\alpha) \tag{41}$$

where $J_{\gamma_\alpha^{-1}(z)}$ is the Jacobian of $\gamma_\alpha^{-1}(z)$. Since f_1 is paramterized by α, it follows by Theorem 1 that if α satisfies (7) it is guaranteed that $\arg\max_k g_k^{\varepsilon_1}(x) = \arg\max_k g_k^{\varepsilon_0}(x)$. The statement of the corollary immediately follows from the observation

$$g^{\varepsilon_1}(x) = \mathbb{E}(h(\phi(x,\, \varepsilon_1))) = \mathbb{E}(h(\phi(x,\, \gamma_\alpha(\varepsilon_0)))) = \mathbb{E}(h(\phi(\phi(x,\, \alpha),\, \varepsilon_0))) = g^{\varepsilon_0}(\phi(x,\, \alpha)). \tag{42}$$

□

B Proof of Theorem 2

Here we provide the proof of Theorem 2 which gives a robustness condition for differentially resolvable transformations. First, recall the definition:

Definition 5 (restated)**.** *Let $\phi \colon \mathcal{X} \times \mathcal{Z}_\phi \to \mathcal{X}$ be a transform with noise space $\mathcal{Z}_\phi$ and let $\psi \colon \mathcal{X} \times \mathcal{Z}_\psi \to \mathcal{X}$ be a resolvable transform with noise space $\mathcal{Z}_\psi$. We say that ϕ can be resolved by ψ if for any $x \in \mathcal{X}$ there exists function $\delta_x \colon \mathcal{Z}_\phi \times \mathcal{Z}_\phi \to \mathcal{Z}_\psi$ such that for any $\beta \in \mathcal{Z}_\phi$*

$$\phi(x,\, \alpha) = \psi(\phi(x,\, \beta),\, \delta_x(\alpha,\, \beta)). \tag{43}$$

Theorem 2 (restated)**.** *Let $\phi\colon \mathcal{X} \times \mathcal{Z}_\phi \to \mathcal{X}$ be a transform which is resolved by $\psi\colon \mathcal{X} \times \mathcal{Z}_\psi \to \mathcal{X}$. Let $\varepsilon \sim \mathbb{P}_\varepsilon$ be a $\mathcal{Z}_\psi$-valued random variable and suppose that the g^ε-smoothed classifier given by $g^\varepsilon(x) = \mathbb{E}(h(\psi(x,\, \varepsilon)))$ predicts $k_A = \arg\max_k g_k^\varepsilon(x)$. Let $\mathcal{S} \subseteq \mathcal{Z}_\psi$ and $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$ be a set of transformation parameters such that for any i, g^ε is $(p_A^{(i)},\, p_B^{(i)})$-confident at $\phi(x,\, \alpha_i)$. Then there exists a set $\Delta^* \subseteq \mathcal{Z}_\psi$ with the property that, if for any $\alpha \in \mathcal{S}$, $\exists \alpha_i$ with $\delta_x(\alpha,\, \alpha_i) \in \Delta^*$ it is guaranteed that*

$$g_{k_A}^\varepsilon(\phi(x,\, \alpha)) > \max_{k \neq k_A} g_k^\varepsilon(\phi(x,\, \alpha)) \tag{44}$$

Proof. We prove the theorem by explicitly constructing a region Δ^* with the desired property by applying Theorem 1. For that purpose let $\delta \in \mathcal{Z}_\psi$ and denote by $\gamma_\delta\colon \mathcal{Z}_\psi \to \mathcal{Z}_\psi$ the resolving function of ψ, i.e.

$$\psi(\psi(x,\, \delta),\, \delta') = \psi(x,\, \gamma_\delta(\delta')). \tag{45}$$

Let $\mathbb{P}_\gamma$ be the distribution of the random variable $\gamma := \gamma_\delta(\varepsilon)$ with density function f_γ and let

$$\underline{S}_t = \{z \in \mathcal{Z}_\psi \colon\ \Lambda(z) < t\}, \quad \overline{S}_t = \{z \in \mathcal{Z}_\psi \colon\ \Lambda(z) \leq t\}, \quad \text{where} \quad \Lambda(z) = \frac{f_\gamma(z)}{f_\varepsilon(z)}. \tag{46}$$

Furthermore, in the notation of Theorem 1 define the function $\zeta\colon \mathbb{R}_{\geq 0} \to [0,\, 1]$ by $t \mapsto \zeta(t) := \mathbb{P}_\varepsilon(\overline{S}_t)$ and denote by $\zeta^{-1}(p) := \inf\{t \geq 0\colon\ \zeta(t) \geq p\}$ its generalized inverse. For $t \geq 0$ and $p \in [0,\, 1]$ we define the function ξ by

$$\xi(t,\, p) := \sup\{\mathbb{P}_\gamma(S)\colon\ \underline{S}_t \subseteq S \subseteq \overline{S}_t,\, \mathbb{P}_\varepsilon(S) \leq p\}. \tag{47}$$

By assumption, for every $i = 1,\, \ldots,\, n$, the ε-smoothed classifier g^ε is $(p_A^{(i)},\, p_B^{(i)})$-confident at $\phi(x,\, \alpha_i)$. Denote by $\Delta_i \subseteq \mathcal{Z}_\psi$ the set of perturbations which satisfy the robustness condition in Theorem 1, i.e.

$$\Delta_i \equiv \{\delta \in \mathcal{Z}_\psi \colon\ 1 - \xi(\zeta^{-1}(1 - p_B^{(i)}),\, 1 - p_B^{(i)}) < \xi(\zeta^{-1}(p_A^{(i)}),\, p_A^{(i)})\} \tag{48}$$

Thus, by Theorem 1, we have that

$$\delta \in \Delta_i \Rightarrow g_{k_A}^\varepsilon(\psi(\phi(x,\, \alpha_i),\, \delta)) > \max_{k \neq k_A} g_k^\varepsilon(\psi(\phi(x,\, \alpha_i),\, \delta)). \tag{49}$$

Finally, note that for the set

$$\Delta^* \equiv \bigcap_{i=1}^N \Delta_i \tag{50}$$

it holds that, if for $\alpha \in \mathcal{S}$ there exists α_i with $\delta_x(\alpha,\, \alpha_i) \in \Delta^*$, then in particular $\delta_x(\alpha,\, \alpha_i) \in \Delta_i$ and hence, by Theorem 1 it is guaranteed that

$$g_{k_A}^\varepsilon(\phi(x,\, \alpha)) = g_{k_A}^\varepsilon(\psi(\phi(x,\, \alpha_i),\, \delta_x(\alpha,\, \alpha_i))) \tag{51}$$

$$> \max_{k \neq k_A} g_k^\varepsilon(\psi(\phi(x,\, \alpha_i),\, \delta_x(\alpha,\, \alpha_i))) = \max_{k \neq k_A} g_k^\varepsilon(\phi(x,\, \alpha)) \tag{52}$$

which concludes the proof. □

C Smoothing Distributions

C.1 Gaussian Smoothing

Corollary C.1. *Suppose $\mathcal{Z} = \mathbb{R}^m$, $\Sigma := \operatorname{diag}(\sigma_1^2, \ldots, \sigma_m^2)$ and $\varepsilon_0 \sim \mathcal{N}(0, \Sigma)$ and $\varepsilon_1 := \alpha + \varepsilon_0$ for some $\alpha \in \mathbb{R}^m$. Suppose that g^{ε_0} is (p_A, p_B)-confident at $x \in \mathcal{X}$ for some $k_A \in \mathcal{Y}$. Then, we have $g^{\varepsilon_1}_{k_A}(x) > \max_{k \neq k_A} g^{\varepsilon_1}_k(x)$ if α satisfies*

$$\sqrt{\sum_{i=1}^{m} \left(\frac{\alpha_i}{\sigma_i}\right)^2} < \frac{1}{2}\left(\Phi^{-1}(p_A) - \Phi^{-1}(p_B)\right). \tag{53}$$

Proof. By Theorem 1 we know that $g^{\varepsilon_1}_{k_A}(x) > \max_{k \neq k_A} g^{\varepsilon_1}_k(x)$ if ε_1 satisfies

$$1 - \xi(\zeta^{-1}(1 - p_B), 1 - p_B) < \xi(\zeta^{-1}(p_A), p_A). \tag{54}$$

The proof is thus complete if we show that (54) reduces to (53). For that purpose denote by f_0 and f_1 density functions of ε_0 and ε_1 respectively. Let $A := \Sigma^{-1}$ and note that the bilinear form $(z_1, z_2) \mapsto z_1^T A z_2 =: \langle z_1, z_2\rangle_A$ defines an inner product on $\mathbb{R}^m$. Let $z \in \mathbb{R}^m$ and consider

$$\Lambda(z) = \frac{f_1(z)}{f_0(z)} = \frac{\exp\left(-\frac{1}{2}\langle z - \alpha, z - \alpha\rangle_A\right)}{\exp\left(-\frac{1}{2}\langle z, z\rangle_A\right)} = \exp\left(\langle z, \alpha\rangle_A - \frac{1}{2}\langle \alpha, \alpha\rangle_A\right). \tag{55}$$

and thus

$$\Lambda(z) \leq t \iff \langle z, \alpha\rangle_A \leq \log(t) + \frac{1}{2}\langle \alpha, \alpha\rangle. \tag{56}$$

Let $Z \sim \mathcal{N}(0, 1)$ and notice that $\frac{\langle \varepsilon_0, \alpha\rangle_A}{\sqrt{\langle \alpha, \alpha\rangle_A}} \overset{d}{=} Z \overset{d}{=} \frac{\langle \varepsilon_1, \alpha\rangle_A - \langle \alpha, \alpha\rangle_A}{\sqrt{\langle \alpha, \alpha\rangle_A}}$. Let $\partial_t := \overline{S}_t \setminus \underline{S}_t = \{z \colon \Lambda(z) = t\}$ and notice that $\mathbb{P}_0\left(\partial_t\right) = \mathbb{P}_1\left(\partial_t\right) = 0$ and $\mathbb{P}_0(\underline{S}_t) = \mathbb{P}_0(\overline{S}_t)$. Similarly, it holds that $\mathbb{P}_1(\underline{S}_t) = \mathbb{P}_1(\overline{S}_t)$. The function $p \mapsto \xi(\zeta^{-1}(p), p)$ is thus given by

$$\xi(\zeta^{-1}(p), p) = \mathbb{P}_1\left(\overline{S}_{\zeta^{-1}(p)}\right). \tag{57}$$

We compute ζ as

$$\zeta(t) = \mathbb{P}\left(\Lambda(\varepsilon_0) \leq t\right) = \mathbb{P}\left(\langle \varepsilon_0, \alpha\rangle_A \leq \log(t) + \frac{1}{2}\langle \alpha, \alpha\rangle_A\right) \tag{58}$$

$$= \Phi\left(\frac{\log(t) + \frac{1}{2}\langle \alpha, \alpha\rangle_A}{\sqrt{\langle \alpha, \alpha\rangle_A}}\right) \tag{59}$$

and for $p \in [0, 1]$ its inverse

$$\zeta^{-1}(p) = \exp\left(\Phi^{-1}(p)\sqrt{\langle \alpha, \alpha\rangle_A} - \frac{1}{2}\langle \alpha, \alpha\rangle_A\right). \tag{60}$$

Thus

$$\mathbb{P}\left(\Lambda(\varepsilon_1) \leq \zeta^{-1}(p)\right) = \mathbb{P}\left(\frac{\langle \varepsilon_1, \alpha\rangle_A - \langle \alpha, \alpha\rangle_A}{\sqrt{\langle \alpha, \alpha\rangle_A}} \leq \frac{\log(\zeta^{-1}(p)) - \frac{1}{2}\langle \alpha, \alpha\rangle_A}{\sqrt{\langle \alpha, \alpha\rangle_A}}\right) \tag{61}$$

$$= \Phi\left(\frac{\left(\Phi^{-1}(p)\sqrt{\langle \alpha, \alpha\rangle_A} - \frac{1}{2}\langle \alpha, \alpha\rangle_A\right) - \frac{1}{2}\langle \alpha, \alpha\rangle_A}{\sqrt{\langle \alpha, \alpha\rangle_A}}\right) = \Phi\left(\Phi^{-1}(p) - \sqrt{\langle \alpha, \alpha\rangle_A}\right). \tag{62}$$

Finally, algebra shows that $1-\Phi\left(\Phi^{-1}(1-p_B)-\sqrt{\langle\alpha,\,\alpha\rangle_A}\right)<\Phi\left(\Phi^{-1}(p_A)-\sqrt{\langle\alpha,\,\alpha\rangle_A}\right)$ is equivalent to

$$\sqrt{\sum_{i=1}^{m}\left(\frac{\alpha_i}{\sigma_i}\right)^2}<\frac{1}{2}\left(\Phi^{-1}(p_A)-\Phi^{-1}(p_B)\right) \tag{63}$$

what concludes the proof. □

C.2 Exponential Smoothing

Corollary C.2. *Suppose $\mathcal{Z}=\mathbb{R}^m_{\geq 0}$, fix some $\lambda>0$ and let $\varepsilon_{0,i}\overset{\text{iid}}{\sim}\operatorname{Exp}(1/\lambda)$, $\varepsilon_0:=(\varepsilon_{0,1},\,\dots,\,\varepsilon_{0,m})^T$ and $\varepsilon_1:=\alpha+\varepsilon_0$ for some $\alpha\in\mathbb{R}^m_{\geq 0}$. Suppose that g^{ε_0} is $(p_A,\,p_B)$-confident at $x\in\mathcal{X}$ for some $k_A\in\mathcal{Y}$. Then, we have $g^{\varepsilon_1}_{k_A}(x)>\max_{k\neq k_A}g^{\varepsilon_1}_k(x)$ if α satisfies*

$$\|\alpha\|_1<-\frac{\log(1-p_A+p_B)}{\lambda}. \tag{64}$$

Proof. By Theorem 1 we know that $g^{\varepsilon_1}_{k_A}(x)>\max_{k\neq k_A}g^{\varepsilon_1}_k(x)$ if ε_1 satisfies

$$1-\xi(\zeta^{-1}(1-p_B),\,1-p_B)<\xi(\zeta^{-1}(p_A),\,p_A). \tag{65}$$

The proof is thus complete if we show that (65) reduces to (64). For that purpose denote by f_0 and f_1 density functions of ε_0 and ε_1 respectively and note that

$$f_1(z)=\begin{cases}\lambda\cdot\exp(-\lambda\|z-\alpha\|_1), & \min_i(z_i-\alpha_i)\geq 0,\\ 0, & \text{otherwise},\end{cases} \tag{66}$$

$$f_0(z)=\begin{cases}\lambda\cdot\exp(-\lambda\|z\|_1), & \min_i(z_i)\geq 0,\\ 0, & \text{otherwise},\end{cases} \tag{67}$$

and $\forall i,\ z_i-\alpha_i\leq z_i$ and hence $f_0(z)=0\Rightarrow f_1(z)=0$. Thus

$$\Lambda(z)=\frac{f_1(z)}{f_0(z)}=\begin{cases}\exp\left(\lambda\cdot\|\alpha\|_1\right) & \min_i(z_i-\alpha_i)\geq 0,\\ 0, & \text{otherwise}.\end{cases} \tag{68}$$

Let $S_0:=\{z\in\mathbb{R}^m_{\geq 0}\colon\ \min_i(z_i-\alpha_i)<0\}$ and note that due to independence

$$\mathbb{P}_0\left(S_0\right)=\mathbb{P}\left(\bigcup_{i=1}^{m}\{\varepsilon_{0,i}<\alpha_i\}\right)=1-\mathbb{P}\left(\bigcap_{i=1}^{m}\{\varepsilon_{0,i}\geq\alpha_i\}\right)=1-\prod_{i=1}^{m}\mathbb{P}\left(\varepsilon_{0,i}\geq\alpha_i\right) \tag{69}$$

$$=1-\prod_{i=1}^{m}\left(1-\left(1-\exp\left(-\lambda\,\alpha_i\right)\right)\right)=1-\exp\left(-\lambda\|\alpha\|_1\right). \tag{70}$$

Let $t_\alpha:=\exp(\lambda\|\alpha\|_1)$ and compute ζ as

$$\zeta(t)=\mathbb{P}\left(\Lambda(\varepsilon_0)\leq t\right)=\mathbb{P}\left(\mathbb{1}\{\min_i(\varepsilon_{0,i}-\alpha_i)\geq 0\}\leq t\cdot\exp\left(-\lambda\|\alpha\|_1\right)\right) \tag{71}$$

$$=\begin{cases}1-\exp\left(-\lambda\|\alpha\|_1\right) & t<t_\alpha,\\ 1 & t\geq t_\alpha.\end{cases} \tag{72}$$

Recall that $\zeta^{-1}(p):=\inf\{t\geq 0\colon\ \zeta(t)\geq p\}$ for $p\in[0,\,1]$ and hence

$$\zeta^{-1}(p)=\begin{cases}0 & p\leq 1-\exp(-\lambda\|\alpha\|_1),\\ \exp(\lambda\|\alpha\|_1) & p>1-\exp(-\lambda\|\alpha\|_1).\end{cases} \tag{73}$$

In order to evaluate ξ we compute the lower and strict lower level sets at $t = \zeta^{-1}(p)$. Recall that $\underline{S}_t = \{z \in \mathbb{R}^m_{\geq 0}\colon\ \Lambda(z) < t\}$ and $\overline{S}_t = \{z \in \mathbb{R}^m_{\geq 0}\colon\ \Lambda(z) \leq t\}$ and consider

$$\underline{S}_{\zeta^{-1}(p)} = \left(S_0^c \cap \left\{z \in \mathbb{R}^m_{\geq 0}\colon\ \exp(\lambda\|\alpha\|_1) < \zeta^{-1}(p)\right\}\right) \cup \left(S_0 \cap \left\{z \in \mathbb{R}^m_{\geq 0} \,|\, 0 < \zeta^{-1}(p)\right\}\right) \tag{74}$$

$$= \begin{cases} \varnothing & p \leq 1 - \exp(-\lambda\|\alpha\|_1), \\ S_0 & p > 1 - \exp(-\lambda\|\alpha\|_1) \end{cases} \tag{75}$$

and

$$\overline{S}_{\zeta^{-1}(p)} = \left(S_0^c \cap \left\{z \in \mathbb{R}^m_{\geq 0}\colon\ \exp(\lambda\|\alpha\|_1) \leq \zeta^{-1}(p)\right\}\right) \dot\cup \left(S_0 \cap \left\{z \in \mathbb{R}^m_{\geq 0}\colon\ 0 \leq \zeta^{-1}(p)\right\}\right) \tag{76}$$

$$= \begin{cases} S_0 & p \leq 1 - \exp(-\lambda\|\alpha\|_1), \\ \mathbb{R}^m_+ & p > 1 - \exp(-\lambda\|\alpha\|_1). \end{cases} \tag{77}$$

Suppose that $p \leq 1 - \exp(-\lambda\|\alpha\|_1)$. Then $\underline{S}_{\zeta^{-1}(p)} = \varnothing$ and $\overline{S}_{\zeta^{-1}(p)} = S_0$ and hence

$$p \leq 1 - \exp(-\lambda\|\alpha\|_1) \Rightarrow \xi(\zeta^{-1}(p),\, p) = \sup\{\mathbb{P}_1(S)\colon\ S \subseteq S_0 \wedge \mathbb{P}_0(S) \leq p\} = 0. \tag{78}$$

Condition (65) can thus only be satisfied, if $p_A > 1 - \exp(-\lambda\|\alpha\|_1)$ and $1 - p_B > 1 - \exp(-\lambda\|\alpha\|_1)$. In this case $\underline{S}_{\zeta^{-1}(p)} = S_0$ and $\overline{S}_{\zeta^{-1}(p)} = \mathbb{R}^m_{\geq 0}$. For $p \in [0, 1]$ let $\mathcal{S}_p = \{S \subseteq \mathbb{R}^m_{\geq 0}\colon\ S_0 \subseteq S \subseteq \mathbb{R}^m_{\geq 0},\, \mathbb{P}_0(S) \leq p\}$. Then

$$p > 1 - \exp(-\lambda\|\alpha\|_1) \Rightarrow \xi(\zeta^{-1}(p),\, p) = \sup_{S \in \mathcal{S}_p} \mathbb{P}_1(S). \tag{79}$$

We can write any $S \in \mathcal{S}_p$ as the disjoint union $S = S_0 \,\dot\cup\, T$ for some $T \subseteq \mathbb{R}^m_{\geq 0}$ such that $\mathbb{P}_0(S_0 \,\dot\cup\, T) \leq p$. Note that $\mathbb{P}_1\left(S_0\right) = 0$ and since $S_0 \cap T = \varnothing$ any $z \in T$ satisfies $0 \leq \min_i\left(z_i - \alpha_i\right) \leq \min_i\, z_i$ and hence $\Lambda(z) = \exp(\lambda\|\alpha\|_1)$. Thus

$$\mathbb{P}_1\left(S\right) = \mathbb{P}_1\left(T\right) = \int_T f_1(z)\, dz = \int_T \exp(\lambda\|\alpha\|_1) f_0(z)\, dz = \exp(\lambda\|\alpha\|_1) \cdot \mathbb{P}_0\left(T\right). \tag{80}$$

Thus, The supremum of the left hand side over all $S \in \mathcal{S}_p$ equals the supremum of the right hand side over all $T \in \{T' \subseteq S_0^c\colon\ \mathbb{P}_0(T') \leq 1 - \mathbb{P}_0(S_0)\}$

$$\sup_{S \in \mathcal{S}_p} \mathbb{P}_1\left(S\right) = \exp(\lambda\|\alpha\|_1) \cdot \sup\left\{\mathbb{P}_1(T')\colon\ T' \subseteq S_0^c,\, \mathbb{P}_0(T') \leq p - \mathbb{P}_0(S_0)\right\} \tag{81}$$

$$= \exp(\lambda\|\alpha\|_1) \cdot (p - \mathbb{P}_0(S_0)). \tag{82}$$

Computing ξ at $(\zeta^{-1}(p_A),\, p_A)$ yields

$$\xi(\zeta^{-1}\left(p_A\right), p_A) = \sup_{S \in \mathcal{S}_{p_A}} \mathbb{P}_1\left(S\right) = \exp(\lambda\|\alpha\|_1) \cdot \left(p_A - \mathbb{P}_0\left(S_0\right)\right) \tag{83}$$

$$= \exp(\lambda\|\alpha\|_1) \cdot \left(p_A - \left(1 - \exp\left(-\lambda\|\alpha\|_1\right)\right)\right) \tag{84}$$

$$= \exp(\lambda\|\alpha\|_1) \cdot \left(p_A + \exp\left(-\lambda\|\alpha\|_1\right) - 1\right) \tag{85}$$

where the third equality follows from (70). Similarly, computing $(\zeta^{-1}(1 - p_B),\, 1 - p_B)$ yields

$$\xi(\zeta^{-1}\left(1 - p_B\right), 1 - p_B) = \sup_{S \in \mathcal{S}_{1-p_B}} \mathbb{P}_1\left(S\right) = \exp(\lambda\|\alpha\|_1) \cdot \left(1 - p_B - \mathbb{P}_0\left(S_0\right)\right) \tag{86}$$

$$= \exp(\lambda\|\alpha\|_1) \cdot \left(1 - p_B - \left(1 - \exp\left(-\lambda\|\alpha\|_1\right)\right)\right) \tag{87}$$

$$= \exp(\lambda\|\alpha\|_1) \cdot \left(-p_B + \exp\left(-\lambda\|\alpha\|_1\right)\right). \tag{88}$$

Finally, condition (65) is satisfied whenever α satisfies

$$\begin{aligned} 1-\exp(\lambda\|\alpha\|_1)\cdot(-p_B+\exp(-\lambda\|\alpha\|_1)) &< \exp(\lambda\|\alpha\|_1)\cdot(p_A+\exp(-\lambda\|\alpha\|_1)-1) && (89)\\ \iff \exp(-\lambda\|\alpha\|_1)+p_B-\exp(-\lambda\|\alpha\|_1) &< p_A+\exp(-\lambda\|\alpha\|_1)-1 && (90)\\ \iff 1-p_A+p_B &< \exp(-\lambda\|\alpha\|_1) && (91)\\ \iff \|\alpha\|_1 &< -\frac{\log(1-p_A+p_B)}{\lambda} && (92)\end{aligned}$$

what completes the proof. □

C.3 Uniform Smoothing

Corollary C.3. *Suppose $\mathcal{Z}=\mathbb{R}^m$, and $\varepsilon_0\sim\mathcal{U}([a,\,b]^m)$ for some $a<b$. Set $\varepsilon_1:=\alpha+\varepsilon_0$ for $\alpha\in\mathbb{R}^m$. Suppose that g^{ε_0} is $(p_A,\,p_B)$-confident at $x\in\mathcal{X}$ for some $k_A\in\mathcal{Y}$. Then, we have $g^{\varepsilon_1}_{k_A}(x)>\max_{k\neq k_A}g^{\varepsilon_1}_k(x)$ if α satisfies*

$$1-\left(\frac{p_A-p_B}{2}\right)<\prod_{i=1}^{m}\left(1-\frac{|\alpha_i|}{b-a}\right)_+ \tag{93}$$

where $(x)_+:=\max\{x,\,0\}$.

Proof. By Theorem 1 we know that $g^{\varepsilon_1}_{k_A}(x)>\max_{k\neq k_A}g^{\varepsilon_1}_k(x)$ if ε_1 satisfies

$$1-\xi(\zeta^{-1}(1-p_B),\,1-p_B)<\xi(\zeta^{-1}(p_A),\,p_A). \tag{94}$$

The proof is thus complete if we show that (94) reduces to (93). For that purpose denote by f_0 and f_1 density functions of ε_0 and ε_1 respectively and let $I_0=[a,\,b]^m$ and $I_1:=\prod_{i=1}^m[a+\alpha_i,\,b+\alpha_i]$ bet the support of ε_0 and ε_1. Consider

$$f_0(z)=\begin{cases}(b-a)^{-m} & z\in I_0,\\ 0 & \text{otherwise}\end{cases}\qquad f_1(z)=\begin{cases}(b-a)^{-m} & z\in I_1,\\ 0 & \text{otherwise.}\end{cases} \tag{95}$$

Let $S_0:=I_0\setminus I_1$. Then, for any $z\in I_0\cup I_1$

$$\Lambda(z)=\frac{f_1(z)}{f_0(z)}=\begin{cases}0 & z\in S_0,\\ 1 & z\in I_0\cap I_1,\\ \infty & z\in I_1\setminus I_0.\end{cases} \tag{96}$$

Note that

$$\mathbb{P}_0\left(S_0\right)=1-\mathbb{P}_0\left(I_1\right)=1-\prod_{i=1}^{m}\mathbb{P}\left(a+\alpha_i\leq\varepsilon_{0,i}\leq b+\alpha_i\right)=1-\prod_{i=1}^{m}\left(1-\frac{|\alpha_i|}{b-a}\right)_+ \tag{97}$$

where $(x)_+=\max\{x,\,0\}$. We then compute ζ for $t\geq0$

$$\zeta(t)=\mathbb{P}\left(\Lambda(\varepsilon_0)\leq t\right)=\begin{cases}\mathbb{P}_0\left(S_0\right) & t<1,\\ \mathbb{P}_0\left(I_0\right) & t\geq1.\end{cases}=\begin{cases}1-\prod_{i=1}^{m}\left(1-\frac{|\alpha_i|}{b-a}\right)_+ & t<1,\\ 1 & t\geq1.\end{cases} \tag{98}$$

Recall that $\zeta^{-1}(p):=\inf\{t\geq0\colon\,\zeta(t)\geq p\}$ for $p\in[0,\,1]$ and hence

$$\zeta^1(p)=\begin{cases}0 & p\leq1-\prod_{i=1}^{m}\left(1-\frac{|\alpha_i|}{b-a}\right)_+,\\ 1 & p>1-\prod_{i=1}^{m}\left(1-\frac{|\alpha_i|}{b-a}\right)_+.\end{cases} \tag{99}$$

In order to evaluate ξ, we compute the lower and strict lower level sets at $t = \zeta^{-1}(p)$. Recall that $\underline{S}_t = \{z \in \mathbb{R}^m_{\geq 0}\colon\ \Lambda(z) < t\}$ and $\overline{S}_t = \{z \in \mathbb{R}^m_{\geq 0}\colon\ \Lambda(z) \leq t\}$ and consider

$$\underline{S}_{\zeta^{-1}(p)} = \begin{cases} \varnothing & p \leq 1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+, \\ S_0 & p > 1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \end{cases} \tag{100}$$

and

$$\overline{S}_{\zeta^{-1}(p)} = \begin{cases} S_0 & p \leq 1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+, \\ I_0 & p > 1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \end{cases} \tag{101}$$

Suppose $p \leq 1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+$. Then $\underline{S}_{\zeta^{-1}(p)} = \varnothing$ and $\overline{S}_{\zeta^{-1}(p)} = S_0$ and hence

$$p \leq 1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \Rightarrow \xi(\zeta^{-1}(p),\, p) = \sup\{\mathbb{P}_1(S)\colon\ S \subseteq S_0,\, \mathbb{P}_0(S) \leq p\} = 0. \tag{102}$$

Condition (94) can thus only be satisfied, if $p_A > 1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+$ and $1 - p_B > 1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+$. In this case $\underline{S}_{\zeta^{-1}(p)} = S_0$ and $\overline{S}_{\zeta^{-1}(p)} = I_0$. For $p \in [0,\, 1]$ let $\mathcal{S}_p = \{S \subseteq \mathbb{R}^m\colon\ S_0 \subseteq S \subseteq I_0,\, \mathbb{P}_0(S) \leq p\}$. Then

$$p > 1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \Rightarrow \xi(\zeta^{-1}(p),\, p) = \sup_{S \in \mathcal{S}_p} \mathbb{P}_1(S). \tag{103}$$

We can write any $S \in \mathcal{S}_p$ as the disjoint union $S = S_0 \,\dot\cup\, T$ for some $T \subseteq I_0 \cap I_1$ such that $\mathbb{P}_0(S_0 \,\dot\cup\, T) \leq p$. Note that $\mathbb{P}_1\left(S_0\right) = 0$ and for any $z \in T$, we have $f_0(z) = f_1(z)$. Hence

$$\mathbb{P}_1\left(S\right) = \mathbb{P}_1(T) = \mathbb{P}_0(T) \leq p - \mathbb{P}_0(S_0) = p - \left(1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right). \tag{104}$$

Thus, The supremum of the left hand side over all $S \in \mathcal{S}_p$ equals the supremum of the right hand side over all $T \in \{T' \subseteq I_0 \cap I_1\colon\ \mathbb{P}_0(T') \leq 1 - \mathbb{P}_0(S_0)\}$

$$\sup_{S \in \mathcal{S}_p} \mathbb{P}_1\left(S\right) = \sup\left\{\mathbb{P}_1(T')\colon\ T' \subseteq I_0 \cap I_1,\, \mathbb{P}_0(T') \leq p - \mathbb{P}_0(S_0)\right\} \tag{105}$$

$$= p - \left(1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right). \tag{106}$$

Hence, computing ξ at $(\zeta^{-1}(p_A),\, p_A)$ and $(\zeta^{-1}(1-p_B),\, 1-p_B)$ yields

$$\xi(\zeta^{-1}(p_A),\, p_A) = p_A - \left(1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right), \tag{107}$$

$$\xi((\zeta^{-1}(1-p_B),\, 1-p_B)) = 1 - p_B - \left(1 - \prod_{i=1}^m \left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right). \tag{108}$$

Finally, condition (94) is satisfied whenever α satisfies

$$1 - \left(1 - p_B - \left(1 - \prod_{i=1}^{m}\left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right)\right) < p_A - \left(1 - \prod_{i=1}^{m}\left(1 - \frac{|\alpha_i|}{b-a}\right)_+\right) \tag{109}$$

$$\iff p_B + 1 - \prod_{i=1}^{m}\left(1 - \frac{|\alpha_i|}{b-a}\right)_+ < p_A - 1 + \prod_{i=1}^{m}\left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \tag{110}$$

$$\iff 2 - p_A + p_B < 2 \cdot \prod_{i=1}^{m}\left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \tag{111}$$

$$\iff 1 - \left(\frac{p_A - p_B}{2}\right) < \prod_{i=1}^{m}\left(1 - \frac{|\alpha_i|}{b-a}\right)_+ \tag{112}$$

what concludes the proof. □

C.4 Laplacian Smoothing

Corollary C.4. *Suppose $\mathcal{Z} = \mathbb{R}$ and $\varepsilon_0 \sim \mathcal{L}(0, b)$ follows a Laplace distribution with mean 0 and scale parameter $b > 0$. Let $\varepsilon_1 := \alpha + \varepsilon_0$ for $\alpha \in \mathbb{R}$. Suppose that g^{ε_0} is (p_A, p_B)-confident at $x \in \mathcal{X}$ for some $k_A \in \mathcal{Y}$. Then, we have $g^{\varepsilon_1}_{k_A}(x) > \max_{k \neq k_A} g^{\varepsilon_1}_k(x)$ if α satisfies*

$$|\alpha| < \begin{cases} -b \cdot \log\left(4\, p_B\,(1 - p_A)\right) & (p_A = \frac{1}{2} \,\wedge\, p_B < \frac{1}{2}) \vee (p_A > \frac{1}{2} \wedge p_B = \frac{1}{2}), \\ -b \cdot \log\left(1 - p_A + p_B\right) & p_A > \frac{1}{2} \wedge p_B < \frac{1}{2}. \end{cases} \tag{113}$$

Proof. By Theorem 1 we know that $g^{\varepsilon_1}_{k_A}(x) > \max_{k \neq k_A} g^{\varepsilon_1}_k(x)$ if ε_1 satisfies

$$1 - \xi(\zeta^{-1}(1 - p_B),\, 1 - p_B) < \xi(\zeta^{-1}(p_A),\, p_A). \tag{114}$$

The proof is thus complete if we show that (114) reduces to (113). For that purpose denote by f_0 and f_1 density functions of ε_0 and ε_1 respectively and consider

$$f_0(z) = \frac{1}{2b}\exp\left(-\frac{|z|}{b}\right), \qquad f_1(z) = \frac{1}{2b}\exp\left(-\frac{|z-\alpha|}{b}\right). \tag{115}$$

Due to symmetry, assume without loss of generality that $\alpha \geq 0$. Then for $z \in \mathbb{R}$

$$\Lambda(z) = \frac{f_1(z)}{f_0(z)} = \exp\left(-\frac{|z-\alpha| - |z|}{b}\right) = \begin{cases} \exp\left(-\frac{\alpha}{b}\right) & z < 0, \\ \exp\left(\frac{2z-\alpha}{b}\right) & 0 \leq z < \alpha, \\ \exp\left(\frac{\alpha}{b}\right) & z \geq \alpha. \end{cases} \tag{116}$$

Note that the CDFs for ε_0 and ε_1 are given by

$$F_0(z) = \begin{cases} \frac{1}{2}\exp\left(\frac{z}{b}\right) & z \leq 0, \\ 1 - \frac{1}{2}\exp\left(-\frac{z}{b}\right) & z > 0, \end{cases} \qquad F_1(z) = \begin{cases} \frac{1}{2}\exp\left(\frac{z-\alpha}{b}\right) & z \leq \alpha, \\ 1 - \frac{1}{2}\exp\left(-\frac{z-\alpha}{b}\right) & z > \alpha. \end{cases} \tag{117}$$

Note that for $\exp\left(-\frac{\alpha}{b}\right) \leq t < \exp\left(\frac{\alpha}{b}\right)$ we have

$$\mathbb{P}_0\left(\exp\left(\frac{2\varepsilon_0 - \alpha}{b}\right) \leq t \,\wedge\, 0 \leq \varepsilon_0 < \alpha\right) = \mathbb{P}_0\left(\exp\left(-\frac{\alpha}{b}\right) \leq \exp\left(\frac{2\varepsilon_0 - \alpha}{b}\right) \leq t\right) \tag{118}$$

$$= \mathbb{P}_0\left(0 \leq \varepsilon_0 \leq \frac{b\log(t) + \alpha}{2}\right) = F_0\left(\frac{b\log(t) + \alpha}{2}\right) - F_0(0) \tag{119}$$

$$= \frac{1}{2} - \frac{1}{2}\exp\left(-\frac{1}{b}\left(\frac{b\log(t) + \alpha}{2}\right)\right) = \frac{1}{2} - \frac{1}{2\sqrt{t}}\exp\left(-\frac{\alpha}{2b}\right). \tag{120}$$

Computing ζ yields

$$\begin{aligned}\zeta(t) &= \mathbb{P}\left(\Lambda(\varepsilon_0) \leq t\right) && (121)\\ &= \mathbb{P}\left(\exp\left(-\frac{\alpha}{b}\right) \leq t \,\wedge\, \varepsilon_0 < 0\right) + && (122)\\ &\quad + \mathbb{P}\left(\exp\left(\frac{\alpha}{b}\right) \leq t \,\wedge\, \varepsilon_0 \geq \alpha\right) + \mathbb{P}\left(\exp\left(\frac{2\varepsilon_0 - \alpha}{b}\right) \leq t \,\wedge\, 0 \leq \varepsilon_0 < \alpha\right) && (123)\\ &= \begin{cases} 0 & t < \exp\left(-\frac{\alpha}{b}\right),\\ 1 - \frac{1}{2\sqrt{t}}\exp\left(-\frac{\alpha}{2b}\right) & \exp\left(-\frac{\alpha}{b}\right) \leq t < \exp\left(\frac{\alpha}{b}\right),\\ 1 & t \geq \exp\left(\frac{\alpha}{b}\right). \end{cases} && (124)\end{aligned}$$

The inverse is then given by

$$\zeta^{-1}(p) = \begin{cases} 0 & p < \frac{1}{2},\\ \frac{1}{4(1-p)^2}\exp\left(-\frac{\alpha}{b}\right) & \frac{1}{2} \leq p < 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}),\\ \exp\left(\frac{\alpha}{b}\right) & p \geq 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}). \end{cases} \tag{125}$$

In order to evaluate ξ we compute the lower and strict lower level sets at $t = \zeta^{-1}(p)$. Recall that $\underline{S}_t = \{z \in \mathbb{R}\colon\ \Lambda(z) < t\}$ and $\overline{S}_t = \{z \in \mathbb{R}\colon\ \Lambda(z) \leq t\}$ and consider

$$\underline{S}_{\zeta^{-1}(p)} = \begin{cases} \varnothing & p \leq \frac{1}{2},\\ \left(-\infty,\, b \cdot \log\left(\frac{1}{2(1-p)}\right)\right) & \frac{1}{2} < p < 1 - \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right),\\ (-\infty,\, \alpha\,], & p \geq 1 - \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right) \end{cases} \tag{126}$$

and

$$\overline{S}_{\zeta^{-1}(p)} = \begin{cases} \varnothing & p < \frac{1}{2},\\ \left(-\infty,\, b \cdot \log\left(\frac{1}{2(1-p)}\right)\right] & \frac{1}{2} \leq p < 1 - \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right),\\ \mathbb{R} & p \geq 1 - \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right). \end{cases} \tag{127}$$

Suppose $p < 1/2$. Then $\underline{S}_{\zeta^{-1}(p)} = \overline{S}_{\zeta^{-1}(p)} = \varnothing$ and hence $\xi(\zeta^{-1}(p),\, p) = 0$ and condition (114) cannot be satisfied. If $p = 1/2$, then $\underline{S}_{\zeta^{-1}(p)} = \varnothing$ and $\overline{S}_{\zeta^{-1}(p)} = (-\infty,\, 0]$. Note that for $z \leq 0$ we have $f_1(z) = f_0(z)\exp(-\alpha/b)$ and hence for any $S \subseteq \overline{S}_{\zeta^{-1}(1/2)}$ we have $\mathbb{P}_1(S) = \exp(-\alpha/b) \cdot \mathbb{P}_0(S)$. We can thus compute ξ at $(\zeta^{-1}(1/2),\, 1/2)$ as

$$p = \frac{1}{2} \Rightarrow \xi\left(\zeta^{-1}\left(\frac{1}{2}\right),\, \frac{1}{2}\right) = \sup\left\{\mathbb{P}_1(S)\colon\ S \subseteq (-\infty,\, 0],\, \mathbb{P}_0(S) \leq \frac{1}{2}\right\} = \frac{1}{2}. \tag{128}$$

Now suppose $1/2 < p < 1 - 1/2\exp(-\alpha/b)$. In this case, $\underline{S}_{\zeta^{-1}(p)} = (-\infty,\, b \cdot \log(1/2(1-p)))$ and $\overline{S}_{\zeta^{-1}(p)} = (-\infty,\, b \cdot \log\left(1/2(1-p)\right)]$. Since the singleton $\{b \cdot \log(1/2(1-p))\}$ has no probability mass under both $\mathbb{P}_0$ and $\mathbb{P}_1$, the function ξ is straight forward to compute: if $\frac{1}{2} < p < 1 - \frac{1}{2}\exp(-\frac{\alpha}{b})$, then

$$\begin{aligned}\xi(\zeta^{-1}(p),\, p) &= \mathbb{P}\left(\varepsilon_1 \leq b \cdot \log\left(\frac{1}{2(1-p)}\right)\right) = \frac{1}{2}\exp\left(\frac{b \cdot \log\left(\frac{1}{2(1-p)}\right) - \alpha}{b}\right) && (129)\\ &= \frac{1}{4(1-p)}\exp\left(-\frac{\alpha}{b}\right). && (130)\end{aligned}$$

Finally, consider the case where $p \geq 1 - {}^1\!/_2 \exp(-{}^\alpha\!/_b)$. Then $\underline{S}_{\zeta^{-1}(p)} = (-\infty,\, \alpha]$ and $\overline{S}_{\zeta^{-1}(p)} = \mathbb{R}$. Any $(-\infty,\, \alpha] \subseteq S \subseteq \mathbb{R}$ can then be written as $S = (-\infty,\, \alpha] \,\dot\cup\, T$ for some $T \subseteq (\alpha,\, \infty)$. Hence

$$\mathbb{P}_1(S) = \mathbb{P}(\varepsilon_1 \leq \alpha) + \mathbb{P}_1(T) = \frac{1}{2} + \exp\left(\frac{\alpha}{b}\right) \mathbb{P}_0(T), \tag{131}$$

$$\mathbb{P}_0(S) = \mathbb{P}(\varepsilon_0 \leq \alpha) + \mathbb{P}_0(T) = 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}) + \mathbb{P}_0(T). \tag{132}$$

Thus, if $p \geq 1 - \frac{1}{2}\exp(-\frac{\alpha}{b})$, then

$$\xi\left(\zeta^{-1}\left(p\right),\, p\right) = \sup\left\{\mathbb{P}_1(S) \colon\ (-\infty,\, \alpha] \subseteq S \subseteq \mathbb{R},\, \mathbb{P}_0(S) \leq p\right\} \tag{133}$$

$$= \frac{1}{2} + \sup\left\{\mathbb{P}_1(T) \colon\ T \subseteq (\alpha,\, \infty),\, \mathbb{P}_0(T) \leq p - 1 + \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right)\right\} \tag{134}$$

$$= \frac{1}{2} + \exp\left(\frac{\alpha}{b}\right)\left(p - 1 + \frac{1}{2}\exp\left(-\frac{\alpha}{b}\right)\right) = 1 - \exp\left(\frac{\alpha}{b}\right)(1 - p)\,. \tag{135}$$

In order to evaluate condition (114), consider

$$1 - \xi\left(\zeta^{-1}(1 - p_B),\, 1 - p_B\right) = \begin{cases} 1 & p_B > \frac{1}{2} \\ \frac{1}{2} & p_B = \frac{1}{2} \\ 1 - \frac{1}{4p_B}\exp\left(-\frac{\alpha}{b}\right) & \frac{1}{2} > p_B > \exp\left(-\frac{\alpha}{b}\right) \\ \exp\left(\frac{\alpha}{b}\right) p_B & \exp\left(-\frac{\alpha}{b}\right) \geq p_B, \end{cases} \tag{136}$$

$$\xi\left(\zeta^{-1}(p_A),\, p_A\right) = \begin{cases} 0 & p_A < \frac{1}{2} \\ \frac{1}{2} & p_A = \frac{1}{2} \\ \dfrac{1}{4(1 - p_A)}\exp\left(-\dfrac{\alpha}{b}\right) & \frac{1}{2} < p_A < 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}) \\ 1 - \exp\left(\dfrac{\alpha}{b}\right)(1 - p_A) & p_A \geq 1 - \frac{1}{2}\exp(-\frac{\alpha}{b}). \end{cases} \tag{137}$$

(138)

Note that the case $p_B > {}^1\!/_2$ can be ruled out, since by assumption $p_A \geq p_B$. If $p_A = {}^1\!/_2$, then we need $p_B < {}^1\!/_2$. Thus, if $p_A = {}^1\!/_2$, then condition (114) is satisfied if $p_B < {}^1\!/_2$ and

$$\max\left\{1 - \frac{1}{4p_B}\exp\left(-\frac{\alpha}{b}\right),\, \exp\left(\frac{\alpha}{b}\right) \cdot p_B\right\} < \frac{1}{2} \tag{139}$$

$$\iff\ p_B \cdot \exp\left(\frac{\alpha}{b}\right) < \frac{1}{2} \tag{140}$$

$$\iff\ \alpha < -b \cdot \log\left(2p_B\right). \tag{141}$$

Now consider the case where $p_A > {}^1\!/_2$. If $p_B = {}^1\!/_2$, then condition (114) is satisfied if

$$\frac{1}{2} < \min\left\{\frac{1}{4(1 - p_A)}\exp\left(-\frac{\alpha}{b}\right),\, 1 - \exp\left(\frac{\alpha}{b}\right)(1 - p_A)\right\} \tag{142}$$

$$\iff\ \frac{1}{2} < 1 - \exp\left(\frac{\alpha}{b}\right)(1 - p_A) \tag{143}$$

$$\iff\ \alpha < -b \cdot \log\left(2(1 - p_A)\right). \tag{144}$$

If on the other hand, $p_A > 1/2$ and $p_B < 1/2$, condition (114) is satisfied if

$$\max\left\{1 - \frac{1}{4p_B}\exp\left(-\frac{\alpha}{b}\right),\ \ \exp\left(\frac{\alpha}{b}\right)\cdot p_B\right\} < \tag{145}$$

$$< \min\left\{\frac{1}{4(1-p_A)}\exp\left(-\frac{\alpha}{b}\right),\ 1 - \exp\left(\frac{\alpha}{b}\right)(1-p_A)\right\} \tag{146}$$

$$p_B \cdot \exp\left(\frac{\alpha}{b}\right) < 1 - \exp\left(\frac{\alpha}{b}\right)(1-p_A) \tag{147}$$

$$\alpha < -b \cdot \log\left(1 - p_A + p_B\right). \tag{148}$$

Finally, we get that condition (114) is satisfied, if

$$|\alpha| < \begin{cases} -b \cdot \log\left(4\,p_B\,(1-p_A)\right) & (p_A = \frac{1}{2} \ \wedge\ p_B < \frac{1}{2}) \vee (p_A > \frac{1}{2} \ \wedge p_B = \frac{1}{2}) \\ -b \cdot \log\left(1 - p_A + p_B\right) & p_A > \frac{1}{2} \ \wedge p_B < \frac{1}{2} \end{cases} \tag{149}$$

what concludes the proof. □

C.5 Folded Gaussian Smoothing

Corollary C.5. *Suppose $\mathcal{Z} = \mathbb{R}_{\geq 0}$, $\varepsilon_0 \sim |\mathcal{N}(0,\sigma)|$ and $\varepsilon_1 := \alpha + \varepsilon_0$ for some $\alpha > 0$. Suppose that g^{ε_0} is (p_A, p_B)-confident at $x \in \mathcal{X}$ for some $k_A \in \mathcal{Y}$. Then, we have $g^{\varepsilon_1}_{k_A}(x) > \max_{k \neq k_A} g^{\varepsilon_1}_k(x)$ if α satisfies*

$$\alpha < \sigma \cdot \left(\Phi^{-1}\left(\frac{1 + \min\{p_A,\, 1 - p_B\}}{2}\right) - \Phi^{-1}\left(\frac{3}{4}\right)\right). \tag{150}$$

Proof. By Theorem 1 we know that $g^{\varepsilon_1}_{k_A}(x) > \max_{k \neq k_A} g^{\varepsilon_1}_k(x)$ if ε_1 satisfies

$$1 - \xi(\zeta^{-1}(1-p_B),\, 1-p_B) < \xi(\zeta^{-1}(p_A),\, p_A). \tag{151}$$

The proof is thus complete if we show that (151) reduces to (150). For that purpose denote by f_0 and f_1 density functions of ε_0 and ε_1 respectively and consider

$$f_0(z) = \begin{cases} \frac{2}{\sqrt{2\pi}\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right) & z \geq 0 \\ 0 & z < 0 \end{cases} \qquad f_1(z) = \begin{cases} \frac{2}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(z-\alpha)^2}{2\sigma^2}\right) & z \geq \alpha \\ 0 & z < \alpha. \end{cases} \tag{152}$$

Then, for $z \geq 0$,

$$\Lambda(z) = \frac{f_1(z)}{f_0(z)} = \begin{cases} 0 & z < \alpha, \\ \exp\left(\frac{z\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2}\right) & z \geq \alpha. \end{cases} \tag{153}$$

Let $t_\alpha := \exp\left(\frac{\alpha^2}{2\sigma^2}\right)$ and suppose $t < t_\alpha$. Then

$$\zeta(t) = \mathbb{P}\left(\Lambda(\varepsilon_0) \leq t\right] = \mathbb{P}\left(\varepsilon_0 < \alpha\right) = \int_0^\alpha \frac{2}{\sqrt{2\pi}\sigma}\exp\left(-\frac{z^2}{2\sigma^2}\right)dz \tag{154}$$

$$= 2 \cdot \int_0^{\alpha/\sigma} \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{s^2}{2}\right)ds = 2 \cdot \Phi\left(\frac{\alpha}{\sigma}\right) - 1. \tag{155}$$

If $t \geq t_\alpha$, then

$$\zeta(t) = \mathbb{P}\left(\Lambda(\varepsilon_0) \leq t\right) = \mathbb{P}\left(\frac{\varepsilon_0\,\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} \leq \log(t) \ \wedge\ \varepsilon_0 \geq \alpha\right) + \mathbb{P}\left(\varepsilon_0 < \alpha\right) \tag{156}$$

$$= \mathbb{P}\left(\varepsilon_0 \leq \frac{\sigma^2}{\alpha}\log(t) + \frac{1}{2}\alpha\right) = 2 \cdot \Phi\left(\frac{\sigma}{\alpha}\log(t) + \frac{\alpha}{2\sigma}\right) - 1 \tag{157}$$

and hence

$$\zeta(t) = \begin{cases} 2 \cdot \Phi\left(\frac{\alpha}{\sigma}\right) - 1 & t < t_\alpha \\ 2 \cdot \Phi\left(\frac{\sigma}{\alpha} \log(t) + \frac{\alpha}{2\sigma}\right) - 1 & t \geq t_\alpha. \end{cases} \tag{158}$$

Note that $\zeta(t_\alpha) = 2 \cdot \Phi\left(\frac{\alpha}{\sigma}\right) - 1$ and let $p_\alpha := \zeta(t_\alpha)$. Recall that $\zeta^{-1}(p) := \inf\{t \geq 0\colon\ \zeta(t) \geq p\}$ which yields

$$\zeta^{-1}(p) = \begin{cases} 0 & p \leq p_\alpha \\ \exp\left(\frac{\alpha}{\sigma}\Phi^{-1}\left(\frac{1+p}{2}\right) - \frac{\alpha^2}{2\sigma^2}\right) & p > p_\alpha. \end{cases} \tag{159}$$

In order to evaluate ξ we compute the lower and strict lower level sets at $t = \zeta^{-1}(p)$. Recall that $\underline{S}_t = \{z \in \mathbb{R}_{\geq 0}\colon\ \Lambda(z) < t\}$ and $\overline{S}_t = \{z \in \mathbb{R}_{\geq 0}\colon\ \Lambda(z) \leq t\}$. Let $S_0 := [0,\, \alpha)$ and note that if $p \leq p_\alpha$, we have $\zeta^{-1}(p) = 0$ and hence $\underline{S}_{\zeta^{-1}(p)} = \varnothing$ and $\overline{S}_{\zeta^{-1}(p)} = S_0$. If, on the other hand $p > p_\alpha$, then

$$\begin{aligned} \underline{S}_{\zeta^{-1}(p)} &= \left\{z \geq 0\colon\ \Lambda(z) < \zeta^{-1}(p)\right\} \qquad (160) \\ &= S_0 \cup \left\{z \geq \alpha\colon\ \frac{z\,\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} < \frac{\alpha}{\sigma}\Phi^{-1}\left(\frac{1+p}{2}\right) - \frac{\alpha^2}{2\sigma^2}\right\} \qquad (161) \\ &= S_0 \cup \left\{z \geq \alpha\colon\ z < \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right\} = S_0 \cup \left[\alpha,\, \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right) \qquad (162) \end{aligned}$$

and

$$\begin{aligned} \overline{S}_{\zeta^{-1}(p)} &= \left\{z \geq 0\colon\ \Lambda(z) \leq \zeta^{-1}(p)\right\} \qquad (163) \\ &= S_0 \cup \left\{z \geq \alpha\colon\ \frac{z\,\alpha}{\sigma^2} - \frac{\alpha^2}{2\sigma^2} \leq \frac{\alpha}{\sigma}\Phi^{-1}\left(\frac{1+p}{2}\right) - \frac{\alpha^2}{2\sigma^2}\right\} \qquad (164) \\ &= S_0 \cup \left\{z \geq \alpha\colon\ z \leq \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right\} = S_0 \cup \left[\alpha,\, \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right] \qquad (165) \\ &= \underline{S}_{\zeta^{-1}(p)} \cup \left\{\sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right\}. \qquad (166) \end{aligned}$$

In other words

$$\begin{aligned} \underline{S}_{\zeta^{-1}(p)} &= \begin{cases} \varnothing & p \leq p_\alpha, \\ S_0 \cup \left[\alpha,\, \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right) & p > p_\alpha, \end{cases} \qquad (167) \\ \overline{S}_{\zeta^{-1}(p)} &= \begin{cases} S_0 & p \leq p_\alpha, \\ S_0 \cup \left[\alpha,\, \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right] & p > p_\alpha. \end{cases} \qquad (168) \end{aligned}$$

Let $\mathcal{S}_{t,\,p} := \{S \subseteq \mathbb{R}_{\geq 0}\colon\ \underline{S}_t \subseteq S \subseteq \overline{S}_t,\, \mathbb{P}_0(S) \leq p\}$ and recall that $\xi(t,\, p) = \sup_{S \in \mathcal{S}_{t,\,p}} \mathbb{P}_1(S)$. Note that for $p \leq p_\alpha$, we have $\mathcal{S}_{\zeta^{-1}(p),\,p} = \{S \subseteq \mathbb{R}_{\geq 0}\colon\ S \subseteq S_0 \wedge \mathbb{P}_0(S) \leq p\}$ and for $S \subseteq S_0$, it holds that $\mathbb{P}_1(S) = 0$. Hence

$$p \leq p_\alpha \Rightarrow \xi\left(\zeta^{-1}(p),\, p\right) = \sup_{S \in \mathcal{S}_{\zeta^{-1}(p),\,p}} \mathbb{P}_1(S) = 0. \tag{169}$$

If $p > p_\alpha$, then

$$\begin{aligned} \mathcal{S}_{\zeta^{-1}(p),\,p} = \{S \subseteq \mathbb{R}_{\geq 0}\colon\ & S_0 \cup \left[\alpha,\, \sigma \cdot \Phi^{-1}\left(1+p/2\right)\right) \subseteq S \\ & \subseteq S_0 \cup \left[\alpha,\, \sigma \cdot \Phi^{-1}\left(1+p/2\right)\right],\, \wedge \mathbb{P}_0(S) \leq p\}. \end{aligned} \tag{170}$$

Since the singleton $\left\{\sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right\}$ has no mass under both $\mathbb{P}_0$ and $\mathbb{P}_1$, we find that if $p > p_\alpha$, then

$$\begin{aligned} \xi\left(\zeta^{-1}(p),\, p\right) &= \mathbb{P}\left(0 \leq \varepsilon_1 \leq \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right)\right) && (171) \\ &= \mathbb{P}\left(0 \leq \varepsilon_0 \leq \sigma \cdot \Phi^{-1}\left(\frac{1+p}{2}\right) - \alpha\right) && (172) \\ &= 2 \cdot \Phi\left(\Phi^{-1}\left(\frac{1+p}{2}\right) - \frac{\alpha}{\sigma}\right) - 1. && (173) \end{aligned}$$

Condition (151) can thus only be satisfied if $p_B < p_A$ and

$$2 \cdot \Phi\left(\frac{\alpha}{\sigma}\right) - 1 < \min\{p_A,\, 1 - p_B\} \,\wedge\, 1 - \xi\left(\zeta^{-1}(1 - p_B),\, 1 - p_b\right) < \xi\left(\zeta^{-1}(p_A),\, p_A\right) \tag{174}$$

which is equivalent to

$$\begin{aligned} \alpha < \sigma \cdot \Phi^{-1}\left(\frac{1 + \min\{p_A,\, 1 - p_B\}}{2}\right) \wedge& \\ \wedge\; \Phi\left(\Phi^{-1}\left(\frac{1 + (1 - p_B)}{2}\right) - \frac{\alpha}{\sigma}\right) + \Phi\left(\Phi^{-1}\left(\frac{1 + p_A}{2}\right) - \frac{\alpha}{\sigma}\right) &> \frac{3}{2}. \end{aligned} \tag{175}$$

Thus, the following is a sufficient condition for the two inequalities in (175) and hence (151) to hold

$$\alpha < \sigma \cdot \left(\Phi^{-1}\left(\frac{1 + \min\{p_A,\, 1 - p_B\}}{2}\right) - \Phi^{-1}\left(\frac{3}{4}\right)\right), \tag{176}$$

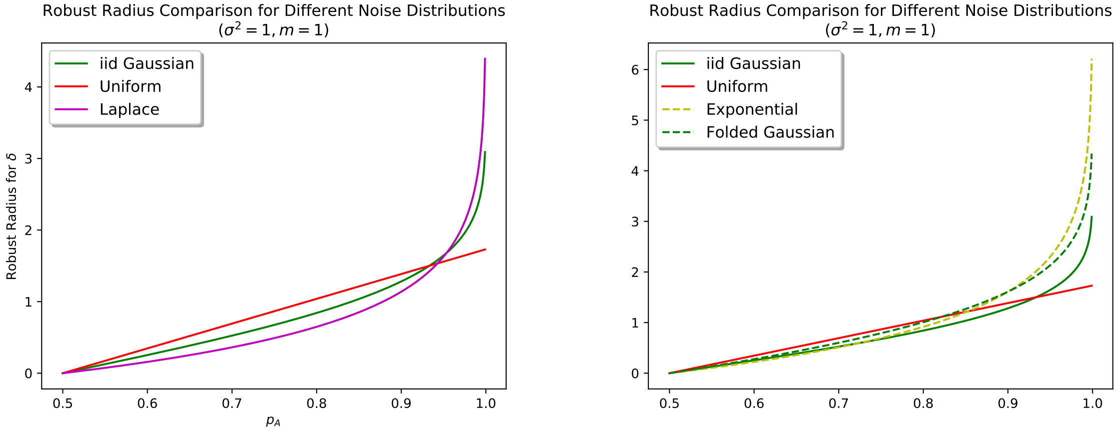
what completes the proof. □

D Comparison of smoothing noise distributions: Detailed Figures and Tables

In this section we provide more detailed graphical illustrations and tables accompanying our findings from section 4.1.

Table D.3: Comparison of certification radii. The variance is normalized to 1 and dimensionality of the noise space is $m = 1$. Note that for Exponential and Folded Gaussian distributions the perturbations are restricted to $\mathbb{R}_{\geq 0}$.

Distribution	Robust Radius
$\mathcal{N}(0,\, 1)$	$\Phi^{-1}(p_A)$
$\mathcal{L}(0,\, \frac{1}{\sqrt{2}})$	$-\frac{1}{\sqrt{2}} \cdot \log(2 - 2p_A)$
$\mathcal{U}([-\sqrt{3},\, -\sqrt{3}])$	$2\sqrt{3} \cdot (p_A - \frac{1}{2})$
Exp(1)	$-\log(2 - 2p_A)$
$\lvert\mathcal{N}(0,\, \sqrt{\frac{\pi}{\pi-2}})\rvert$	$\sqrt{\frac{\pi}{\pi-2}} \cdot \left(\Phi^{-1}\left(\frac{1+p_A}{2}\right) - \Phi^{-1}\left(\frac{3}{4}\right)\right)$



(a) Two-sided noise.

(b) One-Sided noise. Attacker uses positive parameters.

Figure 2: Robust radius comparison for different noise distributions. Additional knowledge on attack model leads to higher radii.

E Adversarial Semantic Transforms: Proofs

E.1 Gaussian Blur

Recall that the Gaussian blur transform is given by a convolution with a Gaussian kernel

$$G_\alpha(k) = \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{k^2}{2\alpha}\right) \tag{177}$$

where $\alpha > 0$ is the squared kernel radius. Here we show that the transform $x \mapsto \phi_B(x) := x * G$ is additive.

Lemma 1 (restated)**.** *The gaussian blur transform is additive,* $\phi_B(\phi_B(x,\, \alpha),\, \beta) = \phi_B(x,\, \alpha + \beta)$.

Proof. Note that associativity of the convolution operator implies that

$$\phi_B(\phi_B(x,\, \alpha),\, \beta) = (\phi_B(x,\, \alpha) \, * \, G_\beta) = ((x \, * \, G_\alpha) \, * \, G_\beta) = (x \, * \, (G_\alpha \, * \, G_\beta)). \tag{178}$$

The claim thus follows, if we can show that $(G_\alpha \, * \, G_\beta) = G_{\alpha+\beta}$. Let $\mathcal{F}$ denote the Fourier transform and $\mathcal{F}^{-1}$ the inverse Fourier transform and note that by the convolution Theorem $(G_\alpha \, * \, G_\beta) = \mathcal{F}^{-1}\{\mathcal{F}(G_\alpha) \cdot \mathcal{F}(G_\beta)\}$. Therefore we have to show that $\mathcal{F}(G_\alpha) \cdot \mathcal{F}(G_\beta) = \mathcal{F}(G_{\alpha+\beta})$. For that purpose, consider

$$\begin{aligned}
\mathcal{F}(G_\alpha)(\omega) &= \int_{-\infty}^{\infty} G_\alpha(y) \exp(-2\,\pi i \omega y)\, dy & (179)\\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\alpha}} \exp\left(-\frac{y^2}{2\alpha}\right) \exp\left(-2\,\pi i \omega y\right)\, dy & (180)\\
&= \frac{1}{\sqrt{2\pi\alpha}} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2\alpha}\right) \left(\cos\left(2\pi\omega y\right) + i \sin\left(2\pi\omega y\right)\right)\, dy & (181)\\
&\overset{(i)}{=} \frac{1}{\sqrt{2\pi\alpha}} \int_{-\infty}^{\infty} \exp\left(-\frac{y^2}{2\alpha}\right)\, \cos\left(2\pi\omega y\right)\, dy \overset{(ii)}{=} \exp\left(-\omega^2\pi^2 2\alpha\right), & (182)
\end{aligned}$$

where (i) follows from the fact that the second term is an integral of an odd function over a symmetric range and (ii) follows from $\int_{-\infty}^{\infty} \exp\left(-a\, y^2\right)\, \cos\left(2\pi\omega y\right)\, dy = \sqrt{\frac{\pi}{a}} \exp(\frac{-(\pi\omega)^2}{a})$ with $a = \frac{1}{2\alpha}$ (see p. 302, eq. 7.4.6 in [1]). This concludes our proof since

$$\begin{aligned}
(\mathcal{F}(G_\alpha) \cdot \mathcal{F}(G_\beta))(\omega) &= \exp\left(-\omega^2\pi^2 2\alpha\right) \cdot \exp\left(-\omega^2\pi^2 2\beta\right) & (183)\\
&= \exp\left(-\omega^2\pi^2 2(\alpha+\beta)\right) = \mathcal{F}(G_{\alpha+\beta})(\omega) & (184)
\end{aligned}$$

and hence

$$(G_\alpha \, * \, G_\beta) = \mathcal{F}^{-1}\{\mathcal{F}(G_\alpha) \cdot \mathcal{F}(G_\beta)\} = \mathcal{F}^{-1}\{\mathcal{F}(G_{\alpha+\beta})\} = G_{\alpha+\beta}. \tag{185}$$

□

Remark E.1. *We notice that the above Theorem naturally extends to higher dimensional Gaussian kernels of the form*

$$G_\alpha(k) = \frac{1}{(2\pi\alpha)^{\frac{m}{2}}} \exp\left(-\frac{\|k\|^2}{2\alpha}\right), \qquad k \in \mathbb{R}^m. \tag{186}$$

Consider

$$\begin{aligned}
\mathcal{F}(G_\alpha)(\omega) &= \int_{\mathbb{R}^m} G_\alpha(y) \exp\left(-2\pi i \langle \omega,\, y\rangle\right)\, dy = \frac{1}{(2\pi\alpha)^{\frac{m}{2}}} \int_{\mathbb{R}^m} \exp\left(-\frac{\|y\|_2^2}{2\alpha} - 2\pi i \langle \omega,\, y\rangle\right)\, dy & (187)\\
&= \prod_{j=1}^{m} \left(\frac{1}{\sqrt{2\pi\alpha}} \int_{\mathbb{R}} \exp\left(-\frac{y_j^2}{2\alpha} - 2\pi i \omega_j y_j\right)\, dy_j\right) = \exp\left(-\, \|\omega\|_2^2\, \pi^2 2\alpha\right) & (188)
\end{aligned}$$

which leads to $(G_\alpha * G_\beta) = G_{\alpha+\beta}$*, and hence additivity.*

E.2 Brightness and contrast

Recall that the brightness and contrast transform is defined as

$$\phi_{BC}\colon \mathcal{X} \times \mathbb{R}^2 \to \mathcal{X}, \quad (x,\, \alpha) \mapsto e^{\alpha_1}(x + \alpha_2). \tag{189}$$

Lemma 2 (restated). *Let $\varepsilon_0 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, \tau^2))$, $\alpha = (k,\, b)^T \in \mathbb{R}^2$ and $\varepsilon_1 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, e^{-2k}\tau^2))$. Then, for all $x \in \mathcal{X}$, it holds that $g^{\varepsilon_0}(\phi_{BC}(x,\, \alpha)) = g^{\alpha+\varepsilon_1}(x)$.*

Proof. Let $x \in \mathcal{X}$, and write $\varepsilon_i = (\varepsilon_{i,1},\, \varepsilon_{i,2})^T$ for $i = 0,\, 1$. Note that

$$\begin{aligned} \phi_{BC}(\phi_{BC}(x,\, \alpha),\, \varepsilon_0) &= e^{\varepsilon_{0,1}}\left(\phi_{BC}(x,\, \alpha) + \varepsilon_{0,2}\right) = e^{\varepsilon_{0,1}}\left(e^{k}\,(x+b) + \varepsilon_{0,2}\right) \qquad (190)\\ &= e^{\varepsilon_{0,1}+k}\left(x + \left(b + e^{-k}\varepsilon_{0,2}\right)\right) = \phi_{BC}(x, \alpha + \tilde{\varepsilon}_0) \qquad (191) \end{aligned}$$

where $\tilde{\varepsilon}_0 = (\varepsilon_{0,1},\, e^{-k}\varepsilon_{0,2})^T$. Note that $\tilde{\varepsilon}_0$ follows a Gaussian distribution since

$$\tilde{\varepsilon}_0 = A \cdot \varepsilon_0, \quad A = \begin{pmatrix} 1 & 0 \\ 0 & e^{-k} \end{pmatrix} \tag{192}$$

and hence $\mathbb{E}\left(\tilde{\varepsilon}_0\right) = A \cdot \mathbb{E}\left(\varepsilon_0\right) = 0$ and

$$\mathrm{Cov}\left(\tilde{\varepsilon}_0\right) = \mathbb{E}\left(\varepsilon_0\, A\, A^T\, \varepsilon_0^T\right) = A^2 \cdot \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & e^{-2k}\tau^2 \end{pmatrix}. \tag{193}$$

The choice $\varepsilon_1 \equiv \tilde{\varepsilon}_0 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma_1^2,\, e^{-2k}\sigma_2^2))$ yields

$$g^{\varepsilon_0}(\phi_{BC}(x,\, \alpha)) = \mathbb{E}\left(\phi(\phi(x,\, \alpha),\, \varepsilon_0)\right) = \mathbb{E}\left(\phi(x,\, \alpha + \varepsilon_1)\right) = g^{\alpha+\varepsilon_1}(x) \tag{194}$$

what concludes the proof. □

Lemma 3 (restated). *Let $x \in \mathcal{X}$, $k \in \mathbb{R}$, $\varepsilon_0 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, \tau^2))$ and $\varepsilon_1 \sim \mathcal{N}(0,\, \mathrm{diag}(\sigma^2,\, e^{-2k}\tau^2))$. Suppose that $g^{\varepsilon_0}(x)_c \geq p$ for some $p \in [0,\, 1]$ and $c \in \mathcal{Y}$. Then*

$$g_c^{\varepsilon_1}(x) \geq \begin{cases} 2\Phi\left(e^k\Phi^{-1}\left(\frac{1+p}{2}\right)\right) - 1 & k \leq 0 \\ 2\left(1 - \Phi\left(e^k\Phi^{-1}(1 - \frac{p}{2})\right)\right) & k > 0. \end{cases} \tag{195}$$

Proof. Note that $\varepsilon_0 \sim \mathcal{N}(0,\, \Sigma)$ and $\varepsilon_1 = A\,\varepsilon_0 \sim \mathcal{N}(0,\, A^2\,\Sigma)$ where

$$A = \begin{pmatrix} 1 & 0 \\ 0 & e^{-k} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 \\ 0 & \tau^2 \end{pmatrix} \tag{196}$$

and denote by f_0 and f_1 the probability density functions of ε_0 and ε_1 respectively and denote by $\mathbb{P}_0$ and $\mathbb{P}_1$ the corresponding probability measures. Recall Definition 3, for $t \geq 0$, (strict) lower level sets are defined as

$$\underline{S_t} := \{z \in \mathcal{Z}\colon\ \Lambda(z) < t\}, \qquad \overline{S_t} := \{z \in \mathcal{Z}\colon\ \Lambda(z) \leq t\}, \qquad \Lambda(z) = \frac{f_1(z)}{f_0(z)}. \tag{197}$$

By assumption we know that $\mathbb{E}[h_c(\phi(x,\, \varepsilon_0))] = g_c^{\varepsilon_0}(x)_k \geq p$. Let $\zeta(t) = \mathbb{P}_0(\overline{S}_t)$ be defined as in Theorem 1 and note that by Lemma A.1, for any $p \in [0,\, 1]$ we have that $\mathbb{P}_0(\underline{S}_{\zeta^{-1}(p)}) \leq p$. Let $\underline{S}_{\zeta^{-1}(p)} \subseteq S \subseteq \overline{S}_{\zeta^{-1}(p)}$ be such that $\mathbb{P}_0(S) \leq p$. Then, from part (i) of Lemma A.2, it follows that $\mathbb{E}(h_c(\phi(x,\, \varepsilon_1))) = g_c^{\varepsilon_1}(x) \geq \mathbb{P}_1(S)$. Note that

$$\begin{aligned} \Lambda(z) = \frac{f_1(z)}{f_0(z)} &= \frac{\left((2\pi)^2|A^2\,\Sigma|\right)^{-\frac{1}{2}}\exp(-\frac{1}{2}(z^T(A^2\Sigma)^{-1}z))}{((2\pi)^2|\Sigma|)^{-\frac{1}{2}}\exp(-\frac{1}{2}(z^T(\Sigma)^{-1}z))} \qquad (198)\\ &= \frac{1}{|A|}\exp\left(-\frac{1}{2}z^T\left((A^2\,\Sigma)^{-1} - \Sigma^{-1}\right)z\right) = \exp\left(k - \frac{z_2^2}{2\tau^2}\left(e^{2k} - 1\right)\right). \qquad (199) \end{aligned}$$

Note that, if $k = 0$, then $\varepsilon_1 = \varepsilon_0$. Thus, suppose that $k \neq 0$ because otherwise the Lemma is satisfied trivially. Suppose that $k > 0$ and consider

$$\zeta(t) = \mathbb{P}_0\left(\overline{S}_t\right) = \mathbb{P}\left(\exp\left(k - \frac{\varepsilon_{0,2}^2}{2\tau^2}\left(e^{2k} - 1\right)\right) \leq t\right) \tag{200}$$

$$= 1 - \mathbb{P}\left(\left(\frac{\varepsilon_{0,2}}{\tau}\right)^2 \leq 2 \cdot \frac{k - \log(t)}{e^{2k} - 1}\right) \tag{201}$$

$$= 1 - F_{\chi^2}\left(2 \cdot \frac{k - \log(t)}{e^{2k} - 1}\right) \tag{202}$$

$$= \begin{cases} 0 & t = 0, \\ 1 - F_{\chi^2}\left(2 \cdot \frac{k - \log(t)}{e^{2k}-1}\right) & 0 < t < e^k, \\ 1 & t \geq e^k, \end{cases} \tag{203}$$

where F_{χ^2} denotes the CDF of the χ^2-distribution with one degree of freedom. Note that for any $t \geq 0$ we have that $\mathbb{P}_0(\overline{S}_t) = \mathbb{P}_0(\underline{S}_t)$ and thus the inverse $\zeta^{-1}(p) = \inf\{t \geq 0\colon\ \zeta(t) \geq p\}$ is given by

$$\zeta^{-1}(p) = \begin{cases} 0 & p = 0 \\ \exp\left(k - F_{\chi^2}^{-1}(1-p) \cdot \frac{e^{2k}-1}{2}\right) & 0 < p < 1 \\ e^k & p = 1. \end{cases} \tag{204}$$

Thus, for any $p \in [0,\, 1]$, we find that $\mathbb{P}_0(\overline{S}_{\zeta^{-1}(p)}) = \mathbb{P}_0(\underline{S}_{\zeta^{-1}(p)}) = \zeta(\zeta^{-1}(p)) = p$. Then, $\mathbb{E}_0(h_c(\phi(x,\, \varepsilon_0))) = g_c^{\varepsilon_0}(x) \geq p = \mathbb{P}_0(\overline{S}_{\zeta^{-1}(p)})$ and part (i) of Lemma A.2 implies that $g_c^{\varepsilon_1}(x) \geq \mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)})$. Computing $\mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)})$ yields

$$g_c^{\varepsilon_1}(x) \geq \mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)}) = 1 - \mathbb{P}\left(\left(\frac{\varepsilon_{1,2}}{\tau^2}\right)^2 \leq (k - \log(\zeta^{-1}(p)))\frac{2}{e^{2k} - 1}\right] \tag{205}$$

$$= 1 - \mathbb{P}\left(\left(\frac{\varepsilon_{0,2}}{\tau^2}\right)^2 \leq (k - \log(\zeta^{-1}(p)))\frac{2e^{2k}}{e^{2k} - 1}\right] \tag{206}$$

$$= 1 - F_{\chi^2}\left((k - \log(\zeta^{-1}(p)))\frac{2e^{2k}}{e^{2k} - 1}\right) \tag{207}$$

$$= 1 - F_{\chi^2}\left(\left(k - \left(k - \frac{e^{2k} - 1}{2}F_{\chi^2}^{-1}(1-p)\right)\right)\frac{2e^{2k}}{e^{2k} - 1}\right) \tag{208}$$

$$= 1 - F_{\chi^2}\left(e^{2k}F_{\chi^2}^{-1}(1-p)\right). \tag{209}$$

If, on the other hand, $k < 0$, then

$$\zeta(t) = \mathbb{P}_0\left(\overline{S}_t\right) = \mathbb{P}\left(\exp\left(k + \frac{\varepsilon_{0,2}^2}{2\tau^2}\left|e^{2k} - 1\right|\right) \leq t\right) \tag{210}$$

$$= \mathbb{P}\left(\left(\frac{\varepsilon_{0,2}}{\tau}\right)^2 \leq 2 \cdot \frac{\log(t) - k}{|e^{2k} - 1|}\right) \tag{211}$$

$$= F_{\chi^2}\left(2 \cdot \frac{\log(t) - k}{|e^{2k} - 1|}\right) = \begin{cases} 0 & t \leq e^k, \\ F_{\chi^2}\left(2 \cdot \frac{\log(t) - k}{|e^{2k}-1|}\right) & t > e^k. \end{cases} \tag{212}$$

A similar computation as in the case where $k > 0$ leads to an expression for the inverse $\zeta^{-1}(p) = \inf\{t \mid \zeta(t) \geq$

$p\}$

$$\zeta^{-1}(p) = \begin{cases} 0 & p = 0, \\ \exp\left(k + F_{\chi^2}^{-1}(p) \cdot \frac{|e^{2k}-1|}{2}\right) & p > 0. \end{cases} \tag{213}$$

Thus, for any $p \in [0,\, 1]$, we find that $\mathbb{P}_0(\overline{S}_{\zeta^{-1}(p)}) = \mathbb{P}_0(\underline{S}_{\zeta^{-1}(p)}) = \zeta(\zeta^{-1}(p)) = p$. Then, $\mathbb{E}_(h_c(\phi(x,\, \varepsilon_0))) = g_c^{\varepsilon_0}(x) \geq p = \mathbb{P}_0(\overline{S}_{\zeta^{-1}(p)})$ and part (i) of Lemma A.2 implies that $g_c^{\varepsilon_1}(x) \geq \mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)})$. Computing $\mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)})$ yields

$$\begin{aligned} g_c^{\varepsilon_1}(x) \geq \mathbb{P}_1(\overline{S}_{\zeta^{-1}(p)}) &= \mathbb{P}\left(\left(\frac{\varepsilon_{1,2}}{\tau}\right)^2 \leq 2 \cdot \frac{\log(\zeta^{-1}(p)) - k}{|e^{2k}-1|}\right) && (214) \\ &= \mathbb{P}\left(\left(\frac{\varepsilon_{0,2}}{\tau}\right)^2 \leq 2e^{2k} \cdot \frac{\log(\zeta^{-1}(p)) - k}{|e^{2k}-1|}\right) && (215) \\ &= F_{\chi^2}\left(\left(\left(k + F_{\chi^2}^{-1}(p)\frac{|e^{2k}-1|}{2}\right) - k\right)\frac{2\,e^{2k}}{|e^{2k}-1|}\right) && (216) \\ &= F_{\chi^2}\left(e^{2k} F_{\chi^2}^{-1}(p)\right). && (217) \end{aligned}$$

Finally, consider the following relation between $\chi^2(1)$ distribution and the standard normal distribution. Let $Z \sim \mathcal{N}(0,1)$ and denote by Φ the CDF of Z. Then, for any $z \geq 0$, $F_{\chi^2}(z) = \mathbb{P}(Z^2 \leq z) = \mathbb{P}(-\sqrt{z} \leq Z \leq \sqrt{z}) = \Phi(\sqrt{z}) - \Phi(-\sqrt{z}) = 2\Phi(\sqrt{z}) - 1$ and the inverse is thus given by $F_{\chi^2}^{-1}(p) = (\Phi^{-1}(\frac{1+p}{2}))^2$. Thus

$$g_c^{\varepsilon_1}(x) \geq \begin{cases} 2\Phi\left(e^k \Phi^{-1}\left(\frac{1+p}{2}\right)\right) - 1 & k \leq 0, \\ 2\left(1 - \Phi\left(e^k \Phi^{-1}(1 - \frac{p}{2})\right)\right) & k > 0, \end{cases} \tag{218}$$

what concludes the proof. □

E.3 Rotations and Scaling

Corollary 2 (restated)**.** *Let $\psi(x,\, \delta) = x + \delta$ and let $\varepsilon \sim \mathcal{N}(0,\, \sigma^2 \mathbb{1}_d)$. Furthermore, let ϕ be a transform with parameters in $\mathcal{Z}_\phi \subseteq \mathbb{R}^m$ and let $\mathcal{S} \subseteq \mathcal{Z}_\phi$ and $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$. Let $k_A \in \mathcal{Y}$ and suppose that for any i, the ε-smoothed classifier $g^\varepsilon(x) := \mathbb{E}[h(x+\varepsilon)]$ is $(p_A^{(i)},\, p_B^{(i)})$-confident at $\phi(x,\, \alpha_i)$ for k_A. Then, the set Δ^* in Theorem 2 is given by the open ℓ_2-ball of radius R around the origin*

$$\Delta^* \equiv B_R(0) \subseteq \mathbb{R}^d \qquad \textit{with} \qquad R := \frac{\sigma}{2} \min_{1 \leq i \leq N} \left(\Phi^{-1}\left(p_A^{(i)}\right) - \Phi^{-1}\left(p_B^{(i)}\right)\right) \tag{219}$$

and it holds that $\forall \alpha \in \mathcal{S}\colon\ k_A = \arg\max_k g_k^\varepsilon(\phi(x,\, \alpha))$ whenever

$$M_{\mathcal{S}} := \max_{\alpha \in \mathcal{S}} \min_{1 \leq i \leq N} \|\phi(x,\, \alpha) - \phi(x,\, \alpha_i)\|_2 < R. \tag{220}$$

Proof. Since the resolvable transform is given by $\psi(x,\, \delta) = x + \delta$ we can write

$$\phi(x,\, \alpha) = \phi(x,\, \alpha_i) + \underbrace{(\phi(x,\, \alpha) - \phi(x,\, \alpha_i))}_{=: \delta_x(\alpha,\, \alpha_i)}. \tag{221}$$

Furthermore, by assumption $\varepsilon \sim \mathcal{N}(0,\, \sigma^2 \mathbb{1}_d)$ and g^ε is $(p_A^{(i)},\, p_B^{(i)})$-confident at $\phi(x,\, \alpha_i)$ for k_A for all i. Thus, by Corollary C.1 we have the robustness condition

$$\|\delta\|_2 < R_i := \frac{\sigma}{2}\left(\Phi^{-1}\left(p_A^{(i)}\right) - \Phi^{-1}\left(p_B^{(i)}\right)\right) \Rightarrow k_A = \arg\max_k g_k^\varepsilon(\phi(x,\, \alpha_i) + \delta). \tag{222}$$

Let $\Delta_i := B_{R_i}(0)$ and notice that $R \equiv \min_i R_i$ and thus

$$\bigcap_{i=1}^{N} B_{R_i}(0) = B_R(0) = \Delta^*. \tag{223}$$

To see that Δ^* has the desired property, consider

$$\forall \alpha \in \mathcal{S}\, \exists \alpha_i \colon\ \delta_x(\alpha,\, \alpha_i) \in \Delta^* \tag{224}$$

$$\iff \forall \alpha \in \mathcal{S}\, \exists \alpha_i \colon\ \left\| \phi(x,\, \alpha) - \phi(x,\, \alpha_i) \right\|_2 < R. \tag{225}$$

Since $R \leq R_i$ it follows that for $\delta_i = \phi(x,\, \alpha) - \phi(x,\, \alpha_i)$ it is guaranteed that

$$k_A = \arg\max_k g_k^{\varepsilon}(\phi(x,\, \alpha_i) + \delta_i) = \arg\max_k g_k^{\varepsilon}(\phi(x,\, \alpha)). \tag{226}$$

Thus, the set Δ^* has the desired property. In particular, since

$$\forall \alpha \in \mathcal{S}\, \exists \alpha_i \colon\ \left\| \phi(x,\, \alpha) - \phi(x,\, \alpha_i) \right\|_2 < R \tag{227}$$

$$\iff \max_{\alpha \in \mathcal{S}} \min_{1 \leq i \leq N} \left\| \phi(x,\, \alpha) - \phi(x,\, \alpha_i) \right\|_2 < R \tag{228}$$

the statement follows. □

F Transformation Details for Rotation and Scaling

In this section we explain rotation and scaling transformations in greater detail. Due to bilinear interpolation, a more in-depth analysis is required. For the sequel, we define images to be real-valued tensors $x \in \mathbb{R}^{K \times W \times H}$. For a given image x, we use the notation $x(k,\, i,\, j) := x_{k,\, i,\, j}$ for the index function, retrieving the pixel value of x at position $(k,\, i,\, j)$. For a real number $y \in \mathbb{R}$ we denote by $\lfloor y \rfloor$ the nearest smaller and by $\lceil y \rceil$ the nearest larger integer.

F.1 Bilinear Interpolation

Let $\Omega_K := \{0,\, \ldots,\, K-1\}$ and $\Omega := [0,\, W-1] \times [0,\, H-1]$. We define bilinear interpolation to be the map $Q\colon \mathbb{R}^{K \times W \times H} \to L^2(\Omega_K \times \mathbb{R}^2,\, \mathbb{R})$, $x \mapsto Q(x) =: Q_x$ where Q_x is given by

$$(k,\, i,\, j) \mapsto Q_x(k,\, i,\, j) := \begin{cases} 0 & (i,\, j) \notin \Omega \\ x_{k,i,j} & (i,\, j) \in \Omega \cap \mathbb{N}^2 \\ \tilde{x}_{k,\, i,\, j} & (i,\, j) \in \Omega \setminus \mathbb{N}^2. \end{cases} \tag{229}$$

where

$$\begin{aligned} \tilde{x}_{k,i,j} := \quad & (1 - (i - \lfloor i \rfloor)) \cdot \big((1 - (j - \lfloor j \rfloor)) \cdot x_{k, \lfloor i \rfloor, \lfloor j \rfloor} + (j - \lfloor j \rfloor) \cdot x_{k, \lfloor i \rfloor, \lfloor j \rfloor + 1} \big) \\ & + (i - \lfloor i \rfloor) \cdot \big((1 - (j - \lfloor j \rfloor)) \cdot x_{k, \lfloor i \rfloor + 1, \lfloor j \rfloor} + (j - \lfloor j \rfloor) \cdot x_{k, \lfloor i \rfloor + 1, \lfloor j \rfloor + 1} \big) . \end{aligned} \tag{230}$$

F.2 Details: Rotation Transformation

The rotation transformation is denoted as $\phi_R\colon \mathbb{R}^{K \times W \times H} \times \mathbb{R} \to \mathbb{R}^{K \times W \times H}$ and acts on an image in three steps which we will highlight in greater detail. First, it rotates the image by α degrees counter-clockwise. After rotation, pixel values are determined using bilinear interpolation (229). Finally, we apply black-padding to all pixels $(i,\, j)$ whose ℓ_2-distance to the center pixel is larger than half of the length of the shorter side, and denote this operation by P. Let c_W and c_H be the center pixels

$$c_W := \frac{W-1}{2}, \qquad c_H := \frac{H-1}{2}. \tag{231}$$

and

$$d_{i,j} = \sqrt{(i - c_W)^2 + (j - c_H)^2}, \qquad g_{i,j} = \operatorname{arctan2}(j - c_H,\, i - c_W)\,. \tag{232}$$

We write $\tilde{\phi}_R$ for the rotation transform before black padding and decompose ϕ_R as $\phi_R = P \circ \tilde{\phi}_R$, where $\tilde{\phi}_R\colon \mathbb{R}^{K \times W \times H} \times \mathbb{R} \to \mathbb{R}^{K \times W \times H}$, $(x,\, \alpha) \mapsto \tilde{\phi}_R(x,\, \alpha)$ is defined by

$$\tilde{\phi}_R(x,\, \alpha)_{k,i,j} := Q_x\left(k,\, c_W + d_{i,j}\, \cos(g_{i,j} - \alpha),\, c_H + d_{i,j}\, \sin(g_{i,j} - \alpha)\right) \tag{233}$$

and $P\colon \mathbb{R}^{K \times W \times H} \to \mathbb{R}^{K \times W \times H}$ by

$$f \mapsto P(f)_{k,i,j} = \begin{cases} f(k,\, i,\, j) & d_{i,j} < \min\{c_W,\, c_H\} \\ 0 & \text{otherwise} \end{cases} . \tag{234}$$

F.3 Details: Scaling Transformation

The scaling transformation is denoted as $\phi_S\colon \mathbb{R}^{K \times W \times H} \times \mathbb{R} \to \mathbb{R}^{K \times W \times H}$. Similar as for rotations, ϕ_S acts on an image in three steps. First, it stretches height and width by a fixed ratio $\alpha \in \mathbb{R}$. Secondly, we determine missing pixel values with bilinear interpolation. Finally, we apply black-padding to regions with missing pixel values if the image is scaled by a factor smaller than 1. Let c_W and c_H be the center pixels

$$c_W := \frac{W-1}{2}, \qquad c_H := \frac{H-1}{2}. \tag{235}$$

We notice that black padding is naturally applied during bilinear interpolation in cases where the scaling factor is smaller than 1 (that is, when we make images smaller). We can thus write the scaling operation as $\phi_S \colon \mathbb{R}^{K\times W\times H} \times \mathbb{R}_{>0} \to \mathbb{R}^{K\times W\times H}$, $(x,\, \alpha) \mapsto \phi(x,\, \alpha)$ where

$$\phi_S(x,\, \alpha)_{k,i,j} := Q_x\left(k,\, c_W + \frac{i - c_W}{\alpha},\, c_H + \frac{j - c_H}{\alpha}\right). \tag{236}$$

F.4 Rotation and Scaling

In this section we state definitions and Lemmas needed to justify our approach to certifying rotations and scaling transformations using randomized smoothing. In addition, we provide the details on computing the maximum ℓ_2-sampling error for both rotation and scaling transforms. We first define the maximum ℓ_2-sampling error.

Definition F.6. *Let $x \in \mathcal{X}$, $\phi\colon \mathcal{X} \times \mathcal{Z} \to \mathcal{X}$ a transform, $\mathcal{S} \subseteq \mathcal{Z}$, $N \in \mathbb{N}$ and suppose $\{\alpha_i\}_{i=1}^N \subseteq \mathcal{S}$. We define the maximum ℓ_2 sampling error as*

$$M_{\mathcal{S}} := \max_{a\leq\alpha\leq b}\, \min_{1\leq i\leq N} \|\phi(x,\, \alpha) - \phi(x,\, \alpha_i)\|_2. \tag{237}$$

Furthermore, recall the definitions of the coordinate sets

$$\Omega_K = \{0,\, \dots,\, K-1\} \qquad \text{and} \qquad \Omega = [0,\, W-1] \times [0,\, H-1]. \tag{238}$$

Definition F.7. *For pixels $(i,\, j) \in \Omega$, we define the grid pixel generator G_{ij} as*

$$G_{ij} := \{(i,\, j),\, (i+1,\, j),\, (i,\, j+1),\, (i+1,\, j+1)\}. \tag{239}$$

Definition F.8 (max-color extractor)**.** *We define the operator which extracts the channel-wise maximum pixel wise on a grid $S \subseteq \Omega$ as*

$$\begin{aligned} \bar{m}\colon \mathbb{R}^{K\times W\times H} &\times \Omega_K \times 2^{\Omega} \to \mathbb{R} \\ &(x,\, k,\, S) \mapsto \max_{(i,j)\in S}\left(\max_{(r,s)\in G_{ij}} x_{k,r,s}\right) \end{aligned} \tag{240}$$

Definition F.9 (max-color difference extractor)**.** *We define the operator which extracts the channel-wise maximum change in color on a grid $S \subseteq \Omega$ as*

$$\begin{aligned} m_{\Delta}\colon \mathbb{R}^{K\times W\times H} &\times \Omega_K \times 2^{\Omega} \to \mathbb{R} \\ &(x,\, k,\, S) \mapsto \max_{(i,j)\in S}\left(\max_{(r,s)\in G_{ij}} x_{k,r,s} - \min_{(r,s)\in G_{ij}} x_{k,r,s}\right) \end{aligned} \tag{241}$$

Lemma F.3. *Let $x \in \mathbb{R}^{K\times W\times H}$, $-\infty < t_1 < t_2 < \infty$ and suppose $\rho\colon [t_1,\, t_2] \to [0,\, W-1] \times [0,\, H-1]$ is a curve of class C^1. Let*

$$\psi_k\colon [t_1,\, t_2] \to \mathbb{R},\ \ \psi_k(t) := Q_x(k,\, \rho_1(t),\, \rho_2(t)) \tag{242}$$

where $k \in \Omega_K$ and Q_x denotes bilinear interpolation. Then ψ_k is L_k-Lipschitz continuous with constant

$$L_k = \max_{t\in[t_1,t_2]}\left(\sqrt{2}\,\|\dot{\rho}(t)\|_2 \cdot m_{\Delta}(x,\, k,\, \lfloor\rho(t)\rfloor)\right) \tag{243}$$

Proof of Lemma F.3. Note that the function $t \mapsto \lfloor \rho(t) \rfloor$ is piecewise constant and let $t_1 =: u_1 < u_2 < \ldots < u_{N_0} := t_2$ such that $\lfloor \rho(t) \rfloor$ is constant on $[u_i,\, u_{i+1})$ for all $1 \leq i \leq N_0 - 1$ and $\dot{\cup}_{i=1}^{N_0} [u_i,\, u_{i+1}) = [t_1,\, t_2)$. We notice that ψ_k is a continuous real-valued function since it is the composition of the continuous Q_x and C^1-curve ρ. L_k-Lipschitz continuity on $[t_1,\, t_2)$ thus follows if we show that ψ_k is L_k-Lipschitz on each interval in the partition. For that purpose, let $1 \leq i \leq N_0$ be arbitrary and fix some $t \in [u_i,\, u_{i+1})$. Let $(w,\, h) := \lfloor \rho(t) \rfloor$ and $\gamma(t) := \rho(t) - \lfloor \rho(t) \rfloor$ and notice that $\gamma(t) \in [0,\, 1)^2$. Let

$$V_1 := x_{k,w,h},\, V_2 := x_{k,w,h+1},\, V_3 := x_{k,w+1,h},\, V_4 := x_{k,w+1,h+1}, \tag{244}$$

Then, for any $u \in [u_i,\, u_{i+1})$

$$\begin{aligned} \psi_k(u) &= Q_x(k,\, \rho_1(u),\, \rho_2(u)) && (245) \\ &= (1 - \gamma_1(u)) \cdot ((1 - \gamma_2(u)) \cdot V_1 + \gamma_2(u) \cdot V_2) + \gamma_1(u) \cdot ((1 - \gamma_2(u) \cdot V_3 + \gamma_2(u) \cdot V_4)\,. && (246) \end{aligned}$$

Let $m_\Delta := m_\Delta(x,\, k \lfloor \rho(t) \rfloor)$ and notice that by definition

$$m_\Delta = \max_i V_i - \min_i V_i \tag{247}$$

and in particular

$$|V_i - V_j| \leq m_\Delta \quad \forall\, i,\, j. \tag{248}$$

Since V_i is constant for each i and γ is differentiable, ψ_k is differentiable on $[u_i,\, u_{i+1})$ and hence

$$\dot{\psi}_k(u) = (\dot{\gamma}_1(u)\gamma_2(u) + \gamma_1(u)\dot{\gamma}_2(u))(V_1 - V_2 - V_3 + V_4) + \dot{\gamma}_1(u)(V_3 - V_1) + \dot{\gamma}_2(u)(V_2 - V_1). \tag{249}$$

Note that the derivative $\dot{\psi}_k$ is linear in γ_1 and γ_2 and hence its extreme values are bounded when evaluated at extreme values of γ, that is $(\gamma_1,\, \gamma_2) \in \{0,\, 1\}^2$. We treat each case separately:

- $\gamma_1 = \gamma_2 = 0$. Then,

$$\left|\dot{\psi}_k\right| \leq |\dot{\gamma}_1(V_3 - V_1) + \dot{\gamma}_2(V_2 - V_1)| \leq |\dot{\gamma}_1| \cdot |V_3 - V_1| + |\dot{\gamma}_2| \cdot |V_2 - V_1| \leq m_\Delta(|\dot{\gamma}_1| + |\dot{\gamma}_2|) \tag{250}$$

- $\gamma_1 = \gamma_2 = 1$. Then,

$$\left|\dot{\psi}_k\right| \leq |\dot{\gamma}_1(V_4 - V_2) + \dot{\gamma}_2(V_4 - V_3)| \leq |\dot{\gamma}_1| \cdot |V_4 - V_2| + |\dot{\gamma}_2| \cdot |V_4 - V_3| \leq m_\Delta(|\dot{\gamma}_1| + |\dot{\gamma}_2|) \tag{251}$$

- $\gamma_1 = 0,\, \gamma_2 = 1$. Then,

$$\left|\dot{\psi}_k\right| \leq |\dot{\gamma}_1(V_4 - V_2) + \dot{\gamma}_2(V_2 - V_1)| \leq |\dot{\gamma}_1| \cdot |V_4 - V_2| + |\dot{\gamma}_2| \cdot |V_2 - V_1| \leq m_\Delta(|\dot{\gamma}_1| + |\dot{\gamma}_2|) \tag{252}$$

- $\gamma_1 = 1,\, \gamma_2 = 0$. Then,

$$\left|\dot{\psi}_k\right| \leq |\dot{\gamma}_1(V_3 - V_1) + \dot{\gamma}_2(V_4 - V_3)| \leq |\dot{\gamma}_1| \cdot |V_3 - V_1| + |\dot{\gamma}_2| \cdot |V_4 - V_3| \leq m_\Delta(|\dot{\gamma}_1| + |\dot{\gamma}_2|) \tag{253}$$

Hence, for any $u \in [u_i,\, u_{i+1})$, the modulus of the derivative is bounded by $m_\Delta(|\dot{\gamma}_1| + |\dot{\gamma}_2|)$. We can further bound this by observing the following connection between ℓ_1 and ℓ_2 distance

$$\forall\, x \in \mathbb{R}^n: \quad \|x\|_1 = |\langle |x|\,,\, \mathbf{1} \rangle| \leq \|x\|_2 \, \|\mathbf{1}\|_2 = \sqrt{n}\, \|x\|_2 \tag{254}$$

and hence $\forall\, u \in [u_i,\, u_{i+1})$

$$|\psi_k(u)| \leq m_\Delta \, \|\dot{\gamma}(u)\|_1 \leq m_\Delta \sqrt{2} \, \|\dot{\gamma}(u)\|_2 = m_\Delta \sqrt{2} \, \|\dot{\rho}(u)\|_2 \,. \tag{255}$$

Since ψ_k is differentiable on $[u_i,\, u_{i+1})$, its Lipschitz constant is bounded by the maximum absolute value of its derivative. Hence

$$\max_{u\in[u_i,\, u_{i+1})} m_\Delta \sqrt{2}\,\|\dot{\rho}(u)\|_2 = \max_{u\in[u_i,\, u_{i+1})} m_\Delta(x,\, k,\, \lfloor\rho(u)\rfloor)\sqrt{2}\,\|\dot{\rho}(u)\|_2 \tag{256}$$

$$\leq \max_{u\in[t_1,\, t_2)} m_\Delta(x,\, k,\, \lfloor\rho(u)\rfloor)\sqrt{2}\,\|\dot{\rho}(u)\|_2 = L_k \tag{257}$$

is a Lipschitz constant for ψ_k on $[u_i,\, u_{i+1})$. Note that L_k does not depend on i. Furthermore, i was chosen arbitrarily and hence L_k is a Lipschitz constant for ψ_k on $[t_1,\, t_2)$ and due to continuity on $[t_1,\, t_2]$, concluding the proof. □

In order to apply Corollary 2 to certify rotations and scaling we have to compute the maximum ℓ_2-sampling error (237). In the next sections, we provide detailed explanations on how we can upper bound this quantity for both transforms based on their Lipschitz constant.

F.4.1 Computing an upper bound on $M_{a,\,b}$ for rotations

In order to compute an upper bound on (237) for rotations, we are interested in finding $M \geq 0$ such that

$$M_{a,\,b}^2 \leq M \tag{258}$$

in which case we can replace condition (220) by $\sqrt{M} < r$. For that purpose, consider sampling the α_i equally spaced from the interval $[a,\, b]$, that is

$$\alpha_i := a + (b-a)\frac{i-1}{N-1} \tag{259}$$

and note that $a = \alpha_1 < \alpha_2 < \ldots < \alpha_N = b$. Furthermore, let g_i be the set of functions

$$\begin{aligned} g_i \colon [a,\, b] &\to \mathbb{R}_{\geq 0} \\ \alpha \mapsto g_i(\alpha) &:= ||\phi_R(x,\, \alpha) - \phi_R(x,\, \alpha_i)||_2^2 \end{aligned} \tag{260}$$

where $1 \leq i \leq N$. Note that $\forall \alpha \in [a,\, b]\, \exists\, i$ such that $\alpha \in [\alpha_i,\, \alpha_{i+1}]$ and let

$$M_i := \max_{\alpha_i \leq \alpha \leq \alpha_{i+1}} \min\{g_i(\alpha),\, g_{i+1}(\alpha)\}. \tag{261}$$

Note that

$$\max_{\alpha_i \leq \alpha \leq \alpha_{i+1}} \left(\min_{1\leq j\leq N} g_j(\alpha)\right) \leq \max_{\alpha_i \leq \alpha \leq \alpha_{i+1}} \left(\min\left\{g_i(\alpha),\, g_{i+1}(\alpha)\right\}\right) \quad \forall\, 1 \leq i \leq N-1 \tag{262}$$

and thus

$$M_{a,b}^2 = \max_{a\leq\alpha\leq b}\left(\min_{1\leq j\leq N} g_j(\alpha)\right) = \max_{1\leq i\leq N-1}\left(\max_{\alpha_i\leq\alpha\leq\alpha_{i+1}}\left(\min_{1\leq j\leq N} g_j(\alpha)\right)\right) \tag{263}$$

$$\leq \max_{1\leq i\leq N-1}\left(\max_{\alpha_i\leq\alpha\leq\alpha_{i+1}}\left(\min\left\{g_i(\alpha),\, g_{i+1}(\alpha)\right\}\right)\right) = \max_{1\leq i\leq N-1} M_i \tag{264}$$

We now further divide each interval $[\alpha_i,\, \alpha_{i+1}]$ by sampling $R \in \mathbb{N}$ equally spaced points $\{\gamma_{i,j}\}_{j=1}^R$ given by

$$\gamma_{i,j} := \alpha_i + (\alpha_{i+1} - \alpha_i)\frac{j-1}{R-1} \tag{265}$$

and define

$$m_{i,j} := \max_{\gamma_{i,j}\leq\gamma\leq\gamma_{i,j+1}} \min\{g_i(\gamma),\, g_{i+1}(\gamma)\} \tag{266}$$

and thus

$$M_i \leq \max_{1 \leq j \leq R-1} m_{i,j}. \tag{267}$$

Once we can compute an upper bound on each $m_{i,j}$, we have found an upper bound on $M^2_{a,\,b}$. Now, suppose $\exists\, L \geq 0$ such that

$$\max\left\{ \max_{c,\,d \in [\alpha_i,\,\alpha_{i+1}]} \left| \frac{g_i(c) - g_i(d)}{c-d} \right|,\ \max_{c,\,d \in [\alpha_i,\,\alpha_{i+1}]} \left| \frac{g_{i+1}(c) - g_{i+1}(d)}{c-d} \right| \right\} \leq L \quad \forall\, i. \tag{268}$$

Then, for any $c,\, d \in [\alpha_i,\, \alpha_{i+1}]$ we know that

$$g_i(d) \leq g_i(c) + L \cdot |d - c|\,, \tag{269}$$

$$g_{i+1}(d) \leq g_{i+1}(c) + L \cdot |d - c| \tag{270}$$

and hence for any $\gamma \in [\gamma_{i,j},\, \gamma_{i,j+1}]$

$$g_i(\gamma) \leq g_i(\gamma_{i,j}) + L \cdot |\gamma - \gamma_{i,j}|\,, \tag{271}$$

$$g_i(\gamma) \leq g_i(\gamma_{i,j+1}) + L \cdot |\gamma - \gamma_{i,j+1}|\,, \tag{272}$$

$$g_{i+1}(\gamma) \leq g_{i+1}(\gamma_{i,j}) + L \cdot |\gamma - \gamma_{i,j}|\,, \tag{273}$$

$$g_{i+1}(\gamma) \leq g_{i+1}(\gamma_{i,j+1}) + L \cdot |\gamma - \gamma_{i,j+1}|\,. \tag{274}$$

We can thus bound each g_i on the intervals $[\gamma_{i,j},\, \gamma_{i,j+1}]$

$$\max_{\gamma_{i,\,j} \leq \gamma \leq \gamma_{i,\,j+1}} g_i(\gamma) \tag{275}$$

$$\leq \max_{\gamma_{i,\,j} \leq \gamma \leq \gamma_{i,\,j+1}} \left(\min\left\{ g_i(\gamma_{i,j}) + L \cdot |\gamma - \gamma_{i,j}|\,,\, g_i(\gamma_{i,j+1}) + L \cdot |\gamma - \gamma_{i,j+1}| \right\}\right) \tag{276}$$

$$\leq \max_{\gamma_{i,\,j} \leq \gamma \leq \gamma_{i,\,j+1}} \left(\frac{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1})}{2} + L \cdot \frac{|\gamma - \gamma_{i,j}| + |\gamma - \gamma_{i,j+1}|}{2} \right) \tag{277}$$

$$= \frac{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1})}{2} + L \cdot \frac{\gamma_{i,j+1} - \gamma_{i,j}}{2} \tag{278}$$

$$= \frac{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1})}{2} + L \cdot \frac{\alpha_{i+1} - \alpha_i}{2(R-1)}. \tag{279}$$

Similarly, bounding g_{i+1} on the interval $[\gamma_{i,j},\, \gamma_{i,j+1}]$ yields

$$\max_{\gamma_{i,\,j} \leq \gamma \leq \gamma_{i,\,j+1}} g_{i+1}(\gamma) \tag{280}$$

$$\leq \max_{\gamma_{i,\,j} \leq \gamma \leq \gamma_{i,\,j+1}} \left(\min\left\{ g_{i+1}(\gamma_{i,j+1}) + L \cdot |\gamma - \gamma_{i,j+1}|\ ,\right.\right. \tag{281}$$

$$\left.\left. g_{i+1}(\gamma_{i,j+1}) + L \cdot |\gamma - \gamma_{i,j+1}| \right\}\right) \tag{282}$$

$$\leq \max_{\gamma_{i,\,j} \leq \gamma \leq \gamma_{i,\,j+1}} \left(\frac{g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})}{2} + L \cdot \frac{|\gamma - \gamma_{i,j}| + |\gamma - \gamma_{i,j+1}|}{2} \right) \tag{283}$$

$$= \frac{g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})}{2} + L \cdot \frac{\gamma_{i,j+1} - \gamma_{i,j}}{2} \tag{284}$$

$$= \frac{g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})}{2} + L \cdot \frac{\alpha_{i+1} - \alpha_i}{2(R-1)}. \tag{285}$$

We can thus bound $m_{i,j}$ for each i and j

$$m_{i,j} = \max_{\gamma_{i,j} \leq \gamma \leq \gamma_{i,j+1}} \min\{g_i(\gamma),\, g_{i+1}(\gamma)\} \tag{286}$$

$$\leq \min\left\{ \frac{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1})}{2} + L \cdot \frac{\alpha_{i+1} - \alpha_i}{2(R-1)}, \right. \tag{287}$$

$$\left. \frac{g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})}{2} + L \cdot \frac{\alpha_{i+1} - \alpha_i}{2(R-1)} \right\} \tag{288}$$

$$= \frac{1}{2}\left(\min\{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}),\, g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})\} + L \cdot \frac{\alpha_{i+1} - \alpha_i}{(R-1)} \right) \tag{289}$$

and note that $\alpha_{i+1} - \alpha_i = \frac{b-a}{N-1}$. Then

$$m_{i,j} \leq \frac{1}{2} \left(\min\{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}), \right. \tag{290}$$

$$\left. g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})\} + L \cdot \frac{b-a}{(N-1)(R-1)} \right) \tag{291}$$

leading to an expression for an upper bound on M_i

$$M_i \leq \max_{1 \leq j \leq R-1} m_{i,j} \tag{292}$$

$$\leq \max_{1 \leq j \leq R-1} \frac{1}{2} \left(\min\{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}), \right. \tag{293}$$

$$\left. g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})\} + L \cdot \frac{b-a}{(N-1)(R-1)} \right) \tag{294}$$

and hence setting

$$M := \max_{1 \leq i \leq N-1} \left(\max_{1 \leq j \leq K-1} \frac{1}{2} \left(\min\{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}), \right. \right. \tag{295}$$

$$\left. \left. g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})\}\right)\right) + L \cdot \frac{b-a}{(N-1)(R-1)} \tag{296}$$

yields $\max_{1 \leq i \leq N-1} M_i \leq M$. Equation (295) thus provides us with a computable upper bound of the maximum ℓ_2-sampling error.

Computing the Lipschitz bound L We now need to find the Lipschitz bound L satisfying (268), which is defined to be the maximum of the Lipschitz bound for g_i and g_{i+1}. Recall that ϕ_R acts on images $x \in \mathbb{R}^{K \times W \times H}$ and that g_i is defined as

$$g_i(\alpha) = \|\phi_R(x,\, \alpha) - \phi_R(x,\, \alpha_i)\|_2^2 = \sum_{k=0}^{K-1} \sum_{r=0}^{W-1} \sum_{s=0}^{H-1} \left(\phi_R(x,\, \alpha)_{k,r,s} - \phi_R(x,\, \alpha_i)_{k,r,s}\right)^2 \tag{297}$$

Let c_W and c_H denote the center pixels

$$c_W := \frac{W-1}{2}, \qquad c_H := \frac{H-1}{2}. \tag{298}$$

and recall the following quantities from the definition of ϕ_R (see Section F.2):

$$d_{r,s} = \sqrt{\left(r - c_W\right)^2 + \left(s - c_H\right)^2}, \qquad g_{r,s} = \arctan 2\left(s - c_H,\, r - c_W\right) \tag{299}$$

Note that

$$d_{r,s} \geq \min\{c_W,\, c_H\} \;\;\Rightarrow\;\; \phi_R(x,\, \alpha)_{k,r,s} = 0. \tag{300}$$

We thus only need to consider pixels which lie inside the centered disk. We call the collection of such pixels *valid* pixels, denoted by V:

$$V := \left\{(r,\, s) \in \mathbb{N}^2 \mid d_{r,s} < \min\{c_W,\, c_H\}\right\}. \tag{301}$$

Let $f_1^{r,s} \colon \mathbb{R} \to \mathbb{R}$ and $f_2^{r,s} \colon \mathbb{R} \to \mathbb{R}$ be functions defined as

$$f_1^{r,s}(\alpha) := c_W + d_{r,s}\, \cos(g_{r,s} - \alpha), \;\; f_2^{r,s}(\alpha) = c_H + d_{r,s}\, \sin(g_{r,s} - \alpha). \tag{302}$$

Then for any valid pixel $(r,\, s)$, the value of the rotated image $\phi_R(x,\, \alpha)$ is given by

$$\phi_R(x,\, \alpha)_{k,r,s} = Q_x(k,\, f_1^{r,s}(\alpha),\, f_2^{r,s}(\alpha)) \tag{303}$$

where Q_x denotes bilinear interpolation. We define the shorthand

$$g_i^{k,r,s}(\alpha) := \left(\phi_R(x,\, \alpha)_{k,r,s} - \phi_R(x,\, \alpha_i)_{k,r,s}\right)^2 \tag{304}$$

and denote by $L_i^{k,r,s}$ and $L_{i+1}^{k,r,s}$ the Lipschitz constants of $g_i^{k,r,s}$ and $g_{i+1}^{k,r,s}$ on $[\alpha_i,\, \alpha_{i+1}]$. We can write (297) as

$$g_i(\alpha) = \sum_{k=0}^{K-1} \sum_{(r,\, s) \in V} g_i^{k,r,s}(\alpha), \quad g_{i+1}(\alpha) = \sum_{k=0}^{K-1} \sum_{(r,\, s) \in V} g_{i+1}^{k,r,s}(\alpha) \tag{305}$$

and note that Lipschitz constants of g_i and g_{i+1} on $[\alpha_i,\, \alpha_{i+1}]$ are given by

$$\max_{c,\, d \in [\alpha_i,\, \alpha_{i+1}]} \frac{|g_i(c) - g_i(d)|}{|c - d|} \leq \left(\sum_{k=0}^{K-1} \sum_{(r,\, s) \in V} L_i^{k,r,s} \right) =: L_i \tag{306}$$

$$\max_{c,\, d \in [\alpha_i,\, \alpha_{i+1}]} \frac{|g_{i+1}(c) - g_{i+1}(d)|}{|c - d|} \leq \left(\sum_{k=0}^{K-1} \sum_{(r,\, s) \in V} L_{i+1}^{k,r,s} \right) =: L_{i+1} \tag{307}$$

We can hence determine L according to equation (268) as

$$L = \max_i \left\{\max\left\{L_i,\, L_{i+1}\right\}\right\}. \tag{308}$$

Without loss of generality, consider $L_i^{k,r,s}$ and note that

$$\max_{c,d \in [\alpha_i,\, \alpha_{i+1}]} \left| \frac{g_i^{k,r,s}(c) - g_i^{k,r,s}(d)}{c - d} \right| \tag{309}$$

$$= \max_{c,d \in [\alpha_i,\, \alpha_{i+1}]} \left| \frac{\phi_R(x,\, c)_{k,r,s} - \phi_R(x,\, d)_{k,r,s}}{c - d} \right| \cdot \tag{310}$$

$$\cdot \left|\phi_R(x,\, c)_{k,r,s} + \phi_R(x,\, d)_{k,r,s} - 2\phi_R(x,\, \alpha_i)_{k,r,s}\right| \tag{311}$$

$$\leq \max_{c,d \in [\alpha_i,\, \alpha_{i+1}]} \underbrace{\left| \frac{\phi_R(x,\, c)_{k,r,s} - \phi_R(x,\, d)_{k,r,s}}{c - d} \right|}_{\text{(I)}} \cdot \tag{312}$$

$$\cdot\, 2 \max_{\theta \in [\alpha_i,\, \alpha_{i+1}]} \underbrace{\left|\phi_R(x,\, \theta)_{k,r,s} - \phi_R(x,\, \alpha_i)_{k,r,s}\right|}_{\text{(II)}}. \tag{313}$$

In order to compute a Lipschitz constant for $g_i^{k,r,s}$ on the interval $[\alpha_i,\, \alpha_{i+1}]$ we thus only need to compute a Lipschitz constant for $\phi_R(x,\, \cdot)$ on $[\alpha_i,\, \alpha_{i+1}]$ and an upper bound on (II). For that purpose, note that ϕ_R takes only positive values and consider

$$\text{(II)} \leq \max_{\theta \in [\alpha_i,\, \alpha_{i+1}]} \{\phi_R(x,\, \theta)_{k,r,s},\, \phi_R(x,\, \alpha_i)_{k,r,s}\} = \max_{\theta \in [\alpha_i,\, \alpha_{i+1}]} \phi_R(x,\, \theta)_{k,r,s} \tag{314}$$

Notice that now both $L_i^{k,r,s}$ and $L_{i+1}^{k,r,s}$ share the same upper bound. Recall (303), that is

$$\phi_R(x,\, \theta)_{k,r,s} = Q_x(k,\, f_1^{r,s}(\theta),\, f_2^{r,s}(\theta)). \tag{315}$$

Now, we upper bound (314) by finding all integer grid pixels that are covered by the trajectory $(f_1^{r,s}(\theta),\, f_2^{r,s}(\theta))$. Specifically, let

$$\mathcal{P}_{r,s} := \bigcup_{\theta \in [\alpha_i,\, \alpha_{i+1}]} \left(\lfloor f_1^{r,s}(\theta) \rfloor,\, \lfloor f_2^{r,s}(\theta) \rfloor\right). \tag{316}$$

Since ϕ_R is interpolated from integer pixels, we can consider the maximum over $\mathcal{P}_{r,s}$ in order to upper bound (314):

$$\begin{aligned} \max_{\theta \in [\alpha_i,\, \alpha_{i+1}]} \phi_R(x,\, \theta)_{k,r,s} &= \max_{\theta \in [\alpha_i,\, \alpha_{i+1}]} Q_x(k,\, f_1^{r,s}(\theta),\, f_2^{r,s}(\theta)) & (317)\\ &\leq \max_{(i,j) \in \mathcal{P}_{r,s}} \max \{x(k,\, i,\, j),\, x(k,\, i+1,\, j),\, x(k,\, i,\, j+1),\, x(k,\, i+1,\, j+1)\} & (318)\\ &= \bar{m}(x,\, k,\, \mathcal{P}_{r,s}). & (319) \end{aligned}$$

We now have to find an upper bound of (I), that is, a Lipschitz constant of $\phi_R(x,\, \cdot)_{k,r,s}$ on the interval $[\alpha_i,\, \alpha_{i+1}]$. For that purpose, consider the following. Note that the curve $\rho\colon [\alpha_i,\, \alpha_{i+1}] \to \mathbb{R}^2$, $\rho(t) := (f_1^{r,s}(t),\, f_2^{r,s}(t))$ is of class C^1 and

$$\begin{aligned} \frac{df_1^{r,s}(t)}{dt} &= \frac{d}{dt}\left(c_W + d_{r,s}\, \cos(g_{r,s} - t)\right) = d_{r,s}\, \sin(g_{r,s} - t) & (320)\\ \frac{df_2^{r,s}(t)}{dt} &= \frac{d}{dt}\left(c_H + d_{r,s}\, \sin(g_{r,s} - t)\right) = -d_{r,s}\, \cos(g_{r,s} - t) & (321) \end{aligned}$$

and hence

$$\|\dot{\rho}(t)\|_2 = \sqrt{\left(\frac{df_1^{r,s}(t)}{dt}\right)^2 + \left(\frac{df_2^{r,s}(t)}{dt}\right)^2} = \sqrt{2}\, d_{r,s}. \tag{322}$$

By Lemma F.3 a Lipschitz constant for the function $\phi_R(x,\, \cdot)_{k,r,s}$ is thus given by

$$\max_{c,d \in [\alpha_i,\, \alpha_{i+1}]} \left| \frac{\phi_R(x,\, c)_{k,r,s} - \phi_R(x,\, d)_{k,r,s}}{c - d} \right| \leq 2\, d_{r,s} \cdot m_\Delta(x,\, k,\, \mathcal{P}_{r,s}). \tag{323}$$

We can thus upper bound (I) and (II) in (313) yielding a Lipschitz constant for $g_i^{k,r,s}$ and $g_{i+1}^{k,r,s}$ on $[\alpha_i,\, \alpha_{i+1}]$

$$\begin{aligned} \max_{c,d \in [\alpha_i,\, \alpha_{i+1}]} \left| \frac{g_i^{k,r,s}(c) - g_i^{k,r,s}(d)}{c - d} \right| &\leq 2\, d_{r,s} \cdot m_\Delta(x,\, k,\, \mathcal{P}_{r,s}) \cdot \bar{m}(x,\, k,\, \mathcal{P}_{r,s}) & (324)\\ &= L_i^{k,r,s} = L_{i+1}^{k,r,s}. & (325) \end{aligned}$$

Finally, we can compute L in (268) as

$$L = \max_i \sum_{k=0}^{K-1} \sum_{(r,s) \in V} L_i^{k,r,s} \tag{326}$$

F.4.2 Computing an upper bound on $M_{a,\,b}$ for scaling

Recall the Definition of the Scaling transformation $\phi_S \colon \mathbb{R}^{K\times W\times H} \times \mathbb{R} \to \mathbb{R}^{K\times W\times H}$, $(x,\,\alpha) \mapsto \phi_S(x,\,\alpha)$, where

$$\phi_S(x,\,\alpha)_{k,r,s} := Q_x\left(k,\, c_W + \frac{r - c_W}{s},\, c_H + \frac{s - c_H}{s}\right). \tag{327}$$

Recall that the set Ω is given by $\Omega = [0,\,W-1] \times [0,\,H-1] = \{1,\,\dots,\,K\}$ and let

$$\Omega_{\mathbb{N}} := \Omega \cap \mathbb{N}^2 \tag{328}$$

be the set of integers in Ω. Let $f_1^r \colon [a,\,b] \to \mathbb{R}$ and $f_2^{r,s} \colon [a,\,b] \to \mathbb{R}$ be functions defined as

$$f_1^r(\alpha) := c_W + \frac{r - c_W}{\alpha},\;\; f_2^s(\alpha) = c_H + \frac{s - c_H}{\alpha}. \tag{329}$$

Then, the value of the scaled image $\phi_S(x,\,\alpha)$ is given by

$$\phi_S(x,\,\alpha)_{k,r,s} = Q_x(k,\, f_1^r(\alpha),\, f_2^s(\alpha)) \tag{330}$$

where Q_x denotes bilinear interpolation. Let

$$\psi_k \colon [a,\,b] \to \mathbb{R} \tag{331}$$

$$\alpha \mapsto Q_x(k,\, f_1^r(\alpha),\, f_2^s(\alpha)). \tag{332}$$

We notice that – in contrast to rotations – ψ_k is *not* continuous at every $\alpha \in \mathbb{R}_{>0}$. Namely, when considering scaling factors in $(0,\,1)$, bilinear interpolation applies black padding to some $(r,\,s) \in \Omega$ resulting in discontinuities of ψ_k. To see this, consider the following. The interval $[\alpha_{i+1},\,\alpha_i]$ contains a discontinuity of ψ_k, if

$$\begin{cases} \alpha_{i+1} < \dfrac{r - c_W}{c_W} < \alpha_i & r > c_W, \\ \alpha_{i+1} < \dfrac{c_W - r}{c_W} < \alpha_i & r < c_W \end{cases} \tag{333}$$

because then $\exists\, \alpha_0 \in [\alpha_{i+1},\,\alpha_i]$ such that $f_1^r(\alpha_0) \in \{0,\,W-1\} \subseteq \Omega$ and hence

$$\phi_S(x,\,\alpha_0)_{k,r,s} \neq 0 \tag{334}$$

but, for $r > c_W$,

$$\phi_S(x,\,\alpha_0 + \varepsilon)_{k,r,s} = 0 \quad \forall \varepsilon > 0 \tag{335}$$

or, when $r < c_W$,

$$\phi_S(x,\,\alpha_0 - \varepsilon)_{k,r,s} = 0 \quad \forall \varepsilon > 0. \tag{336}$$

A similar reasoning leads to a discontinuity in the s-coordinates. We can thus define the set of discontinuities of ψ_k as

$$\mathcal{D} := \left(\bigcup_{r=0}^{W-1} \mathcal{D}^r\right) \cup \left(\bigcup_{s=0}^{H-1} \mathcal{D}^s\right) \tag{337}$$

where

$$\mathcal{D}^r := \{\alpha_0 \in [a,\,b] \mid f_1^r(\alpha_0) \in \{0,\,W-1\}\} \tag{338}$$

$$\mathcal{D}^s := \{\alpha_0 \in [a,\,b] \mid f_2^s(\alpha_0) \in \{0,\,H-1\}\}. \tag{339}$$

We notice that $|\mathcal{D}| \leq H + W$ and hence for large enough N, each interval $[\alpha_i,\,\alpha_{i+1}]$ contains at most 1 discontinuity.

We now derive an upper bound on $M_{a,b}^2$. For that purpose, recall that fo $a < b$, and $\{\alpha_i\}_{i=1}^N$ the maximum L_2-sampling error $M_{a,b}$ is given by

$$M_{a,\,b} := \max_{a \leq \alpha \leq b} \min_{1 \leq i \leq N} ||\phi_S(x,\,\alpha) - \phi_S(x,\,\alpha_i)||_2\,. \tag{340}$$

In order to compute an upper bound on (340) for scaling, we are interested in finding $M \geq 0$ such that

$$M_{a,\,b}^2 \leq M \tag{341}$$

in which case we can replace condition (220) by $\sqrt{M} < r$. For scaling, we sample the α_i according to

$$\alpha_i = \frac{ab}{a + (b-a)\frac{i-1}{N-1}} \tag{342}$$

and note that $\alpha_1 = b$ and $\alpha_N = a$. For $1 \leq i \leq N$ Let g_i be the functions defined by

$$g_i \colon [a,\,b] \to \mathbb{R}_{\geq 0} \tag{343}$$

$$\alpha \mapsto g_i(\alpha) := \|\phi_S(x,\,\alpha) - \phi_S(x,\,\alpha_i)\|_2^2\,. \tag{344}$$

Note that $\forall\, \alpha \in [a,\,b]$, $\exists\, i$ such that $\alpha \in [\alpha_{i+1},\,\alpha_i]$. Suppose that N is large enough such that $\forall\, i\colon |\mathcal{D} \cap [\alpha_{i+1},\,\alpha_i]| \leq 1$ and denote the discontinuity in interval $[\alpha_{i+1},\,\alpha_i]$ by t_i if it exists. Let

$$M_i := \begin{cases} \max\limits_{\alpha_{i+1} \leq \alpha \leq \alpha_i} \min\{g_i(\alpha),\, g_{i+1}(\alpha)\} & [\alpha_{i+1},\,\alpha_i] \cap \mathcal{D} = \varnothing \\ \max\left\{ \max\limits_{\alpha_{i+1} \leq \alpha \leq t_i} g_{i+1}(\alpha),\; \max\limits_{t_i \leq \alpha \leq \alpha_i} g_i(\alpha) \right\} & [\alpha_{i+1},\,\alpha_i] \cap \mathcal{D} = \{t_i\} \end{cases} \tag{345}$$

Similarly as in the case for rotations, we find

$$M_{a,b}^2 \leq \max_{1 \leq i \leq N-1} M_i. \tag{346}$$

For simplicity, we assume for the sequel that $\mathcal{D} = \varnothing$. The case where discontinuities exist can be treated analogously. We further divide each interval $[\alpha_{i+1},\,\alpha_i]$ by sampling $R \in \mathbb{N}$ points $\{\gamma_{i,j}\}_{j=1}^R$ according to

$$\gamma_{i,j} := \frac{\alpha_i\,\alpha_{i+1}}{\alpha_i + (\alpha_{i+1} - \alpha_i)\frac{j-1}{R-1}} \tag{347}$$

and define

$$m_{i,j} := \max_{\gamma_{i,j+1} \leq \gamma \leq \gamma_{i,j}} \min\left\{g_i(\gamma),\, g_{i+1}(\gamma)\right\}. \tag{348}$$

We can thus upper bound each M_i by

$$M_i \leq \max_{1 \leq j \leq R-1} m_{i,j}. \tag{349}$$

In order to find an upper bound on $M_{a,b}^2$, we thus need to find upper bound on $m_{i,j}$ and can proceed analogously to rotations. Namely, setting

$$M := \max_{1 \leq i \leq N-1} \left(\max_{1 \leq j \leq R-1} \left(\frac{1}{2} \cdot (\min\left\{g_i(\gamma_{i,j}) + g_i(\gamma_{i,j+1}),\right. \right.\right. \tag{350}$$

$$\left.\left.\left. g_{i+1}(\gamma_{i,j}) + g_{i+1}(\gamma_{i,j+1})\right\}) + L \cdot \frac{\gamma_{i,\,j} - \gamma_{i,\,j+1}}{2} \right)\right) \tag{351}$$

yields a computable upper bound of the maximum ℓ_2 sampling error. Computing a Lipschitz constant for g_i and g_{i+1} is also analogous to rotations. The only difference lies in computing a Lipschitz constant for ϕ_S what we will explain in greater detail.

Recall that Lemma F.3 provides us with a way to compute a Lipschitz constant for the function $t \mapsto \psi_k(t) := Q_x(k,\, \rho_1(t),\, \rho_2(t))$ where ρ is a differentiable curve with values in $\mathbb{R}^2$. Namely, a Lipschitz constant for ψ_k is given by

$$L_k = \max_{t\in[t_1,t_2]} \left(\sqrt{2}\, \|\dot{\rho}(t)\|_2 \cdot m_\Delta(x,\, k,\, \lfloor \rho(t) \rfloor)\right). \tag{352}$$

Consider the curve

$$\rho(t) := (f_1^r(t),\, f_2^s(t)), \quad t > 0 \tag{353}$$

and note that it is differentiable with derivatives

$$\frac{df_1^r(t)}{dt} = \frac{d}{dt}\left(c_W + \frac{r - c_W}{t}\right) = \frac{c_W - r}{t^2} \tag{354}$$

$$\frac{df_2^s(t)}{dt} = \frac{d}{dt}\left(c_H + \frac{s - c_H}{t}\right) = \frac{c_H - s}{t^2} \tag{355}$$

and

$$\|\dot{\rho}(t)\|_2 = \frac{1}{t^2}\sqrt{(c_W - r)^2 + (c_H - s)^2}. \tag{356}$$

A Lipschitz constant for $\phi_S(x,\, \cdot)_{k,r,s}$ is thus given by

$$L_k^{r,s} = \max_{t\in[t_1,\, t_2]} \left(\frac{\sqrt{(c_W - r)^2 + (c_H - s)^2}}{t^2} \cdot \sqrt{2} m_\Delta(x,\, k,\, \lfloor \rho(t) \rfloor)\right) \tag{357}$$

$$\leq \frac{\sqrt{(c_W - r)^2 + (c_H - s)^2}}{t_1^2} \cdot \sqrt{2} \cdot m_\Delta(x,\, k,\, \mathcal{P}_{r,s}) \tag{358}$$

where

$$\mathcal{P}_{r,s} = \bigcup_{\alpha\in[t_1,\, t_2]} \{(\lfloor f_1^r(t) \rfloor,\, \lfloor f_2^s(t) \rfloor)\}. \tag{359}$$

Finally, setting

$$L_i^{k,r,s} := L_k^{r,s} \cdot \bar{m}(x,\, k,\, \mathcal{P}_{r,s}) \tag{360}$$

and

$$L := \sum_{k=0}^{K-1} \sum_{(r,s)\in\Omega_{\mathbb{N}}} L_i^{k,r,s} \tag{361}$$

yields a Lipschitz constant for g_i and g_{i+1} on $[\alpha_{i+1},\, \alpha_i]$.

G Robustness Certificates

When we focus on general transformations, instead of additive noises as in previous work, there is a subtle difference. In this paper, given a data example x, our algorithm tries to certify the following property

(C1) $g_h^\varepsilon(\phi(x, \alpha))$ predicts the same as $g_h^\varepsilon(x)$.

But what if we are given a data point $\phi(x, \beta)$ *with an unknown* β*?* Directly applying our algorithm would lead to

(C2) $g_h^\varepsilon(\phi(\phi(x, \beta), \alpha))$ predicts the same as $g_h^\varepsilon(\phi(x, \beta))$

The meaning of this can be confusing. However, for transformations that are *reversible*:

$$\forall \beta \in A. \, \forall x \in \mathcal{X}. \, \exists \alpha \in A. \, \phi(\phi(x, \beta), \alpha) = x$$

the system can also guarantee

(C3) $g_h^\varepsilon(\phi(x, \beta))$ predicts the same as $g_h^\varepsilon(x)$.

In the existing work on certifying robustness against ℓ_p norm bounded perturbation, both C1 and C3 are automatically satisfied since the additive noise is reversible. Moreover, for many transformations that we focused on in this paper, namely brightness, contrast, and translation, they are also reversible and thus can also certify both C1 and C3. However, for other transformations such as Gaussian blur, rotation, and scaling, the function smoothing-based technique can only certify C1. This illustrates one important difference between function smoothing over general transformations and specific additive noises, and we believe it is exciting future work to fully understand the differences, and application scenarios, of these three different types of certificates. Moreover, this would also shed light on future research on understanding the gap between $g_h^\varepsilon(x)$ and the ground truth as the utility depends on whether $g_h^\varepsilon(x)$ agrees with the ground truth. Some recent endeavors by other researchers already start to touch this question [18], which we believe will become increasingly important.

H Experiment Details

In this section, we provide our experimental setup, dataset specifications and analysis for each transformation in detail.

H.1 Datasets and Certification

H.1.1 Datasets

We run experiments on the three publicly available MNIST [22], CIFAR-10 [21] and ImageNet-1k datasets [6].

- **MNIST** contains $60,000$ training and $10,000$ testing images of handwritten digits from 0 to 9 corresponding to 10 classes. Each image is a gray scale 28×28 image normalized to $[0,1]$. Before feeding the images to the models, we follow common practice and scale the pixel values to $[-1,\ 1]$.
- **CIFAR-10** contains natural images of 10 classes with $50,000$ training and $10,000$ testing samples. Each image is a RGB-image of resolution 32×32 pixels. We normalize each image to $[0,1]$ and then scale each image by subtracting $(0.485,\ 0.456,\ 0.406)$ from each channel and divide by $(0.229,\ 0.224,\ 0.225)$ (also per channel).
- **ImageNet-1k**[3] contains over 10^6 training and $50,000$ validation images from $1,000$ classes. Since the image resolutions vary, we firstly scale each image so that the short edge is 256-pixel long, then apply center cropping to get 224×224-sized 3-channel colored images. Similarly, the pixel colors are normalized to range $[0,1]$. We then scale each image by subtracting $(0.4914,\ 0.4822,\ 0.4465)$ from each channel and divide by $(0.2023,\ 0.1994,\ 0.2010)$ (also per channel).

Note that preprocessing does not affect the measurement of perturbation - we report perturbation magnitude with respect to the original image.

H.1.2 Robustness Certification

For all randomized-smoothing based robust certification, we follow common practice and adopt $\alpha = 0.001$, such that the certified radii hold true with probability at least $1 - \alpha = 99.9\%$ (per sample).

The number of samples, N, is another important hyperparameter for randomized-smoothing. Larger N results in better robust radii but consumes longer running time as shown in [5]. In previous works, the common choice is $N = 10^5$. We also adopt $N = 10^5$ for MNIST and CIFAR-10 models, while for ImageNet, we use $N = 10^4$. Note that for rotation and scaling transformation, we adopt early-stop strategy once we get enough samples to certify robustness, so the actual N is much smaller than the setting. For predicted class guessing, we adopt $N_0 = 100$ throughout all experiments which is in line with previous works.

The reported robust accuracy is actually the *approximate robust accuracy* as mentioned in the main part of this paper. We compute this metric on a subset of the test set. Following common practice, we construct the subset as follows:

- On MNIST, pick one sample in every 20 samples. In total 500 samples.
- On CIFAR-10, pick one sample in every 20 samples. In total 500 samples.
- On ImageNet-1k, pick one sample in every 500 samples. In total 100 samples.

For each setting, we run the certification process 5 times. Due to the large number of N, the robust accuracies are mostly the same and always differ within 0.2% on MNIST and CIFAR-10, and within 1% on ImageNet-1k. When there are multiple numbers among repeated runs, we report the mode.

[3] http://image-net.org/challenges/LSVRC/2012/

H.1.3 Rotation and Scaling: Progressive Certification

According to Corolloary 2, once we have computed $M_{a,b}$, we need to certify that for every sampled parameter α_i, its robust radius satisfies $r > M$. Comparing with the normal usage, where we need to compute the exact r, here we just need to know $r > M$. We exploit this and adopt a progressive certification strategy in order to speed up certification. We first determine the batch size B. Then, we sample B noisy inputs in order to derive the current r_0 with confidence $1 - \alpha$. Once we have $r_0 > M$, we early-stop the process, since in this case we already know that $r > M$. Otherwise, we continue sampling another batch of size B until termination or failing to certify after N inputs. In practice, we set $B = 400$. We find that usually r is much larger than $M_{a,b}$, and hence we can obtain that $r > M_{a,b}$ after a relatively small number of batches. Comparing with generating and computing all N samples, this progressive certification strategy significantly improves speed.

H.2 Classification Models

H.2.1 Model Architectures

For each dataset, we use one neural network architecture, aligning with models used in the literature such as [41, 5, 32, 10].

- **MNIST.** Table H.4 shows the architecture used in greater detail. The model is the identical with the one used in [41].
- **CIFAR-10.** Here we use ResNet-110, a 110-layer ResNet model which is the same as [5, 31].
- **ImageNet-1k.** Here, we use ResNet-50 [15] aligning with the model chosen in [5, 32, 10].

On ImageNet we initialize our models with weights obtained from pretrained classifiers. Namely, for training with rotation and scaling transformations, we initialize the model with weights from the best trained models in [32]. For training with Gaussian blur, translation, and contrast and brightness transformations, we initialize the model weights from [15]. We then retrain the model with corresponding data augmentation. On both MNIST and CIFAR-10 we train each model from scratch.

Table H.4: MNIST model architecture. "Conv2d" stands for 2-dimensional convolutional layer. All convolutional layers use 1-pixel padding for both sides of each dimension. "FC" stands for a fully-connected layer.

Layer	Activation	in-channels	out-channels	kernel size	stride
Conv 2d	ReLU	1	32	3	1
Conv 2d	ReLU	32	32	4	2
Conv 2d	ReLU	32	64	3	1
Conv 2d	ReLU	64	64	4	2
Flatten	-	64	64×7×7	-	-
FC	ReLU	64×7×7	512	-	-
FC	ReLU	512	512	-	-
FC	Softmax	512	10	-	-

H.2.2 Training

We implement all models in Python using the PyTorch[4] Library and train them on a single NVIDIA GTX 1080Ti GPU.

Hyperparameters On *MNIST*, we train each model using SGD with weight decay 10^{-4}, batch size 400, and learning rate 0.1 for 20 epochs. On *CIFAR-10*, for the Gaussian blur transformation we use the Adam optimizer

[4] https://pytorch.org/

with the learning rate set to 0.1, while for other transformations we use SGD with learning rate 0.1 and weight decay 10^{-4}. The learning rate is multiplied by 0.1 for every 30 epochs. and we train each model for 90 epochs. On *ImageNet-1k*, we use SGD optimizer with a small learning rate set to 0.001 and weight decay 10^{-4}. We train the models only for 3 epochs, since the accuracy on the training set stabilizes after 3 epochs. On MNIST and CIFAR-10 we set the batch size to 400 and on ImageNet-1k to 80 due to the larger image resolution.

Data Augmentation During training, we first do random flipping (except on MNIST), then data augmentation by sampling transformation parameters from the corresponding noise distribution and apply the resulting transformation to the image. The specific distributions vary for different transformations.

H.3 Detailed Experimental Results for Noise Distribution Tuning

We explore smoothing using exponential, uniform and Gaussian distributions for a fixed variance level for Gaussian blur transformation on CIFAR-10. The results are shown in Table H.5. The "Robust Acc. for Radius $\alpha \leq$" stands for the robust accuracy permitting any Gaussian blur with parameter α smaller or equal than the corresponding threshold. For the case $\sigma^2_{\text{train}} = 100$, we show robust accuracy curves corresponding to different noise distributions in Figure 3.

Table H.5: Comparison of different noise distributions for *Gaussian Blur* on CIFAR-10. σ^2_{train} stands for the variance of distribution from which we sample noise in training data augmentation. We highlight best numbers in **bold** for the experiments with the same variance.

Variance	Setting	Clean	Robust Acc. for radius $\alpha \leq$							
		Acc.	1.0	4.0	9.0	16.0	25.0	36.0	49.0	64.0
$\sigma^2_{\text{train}} = 4.0$	Exp(1/2)	**88.2%**	**84.6%**	**76.6%**	**58.4%**	0.0%	0.0%	0.0%	0.0%	0.0%
	$\mathcal{U}(0, \alpha = 4\sqrt{3})$	83.2%	80.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	$\mathcal{N}(0, \sigma^2 = \frac{4\pi}{\pi-2})$	81.4%	77.6%	69.6%	47.0%	0.0%	0.0%	0.0%	0.0%	0.0%
$\sigma^2_{\text{train}} = 100.0$	Exp(1/10)	**76.4%**	**74.2%**	**70.0%**	**62.8%**	**55.4%**	**47.8%**	**40.8%**	**31.2%**	**20.2%**
	$\mathcal{U}(0, \alpha = 20\sqrt{3})$	66.8%	65.2%	60.4%	55.8%	45.6%	0.0%	0.0%	0.0%	0.0%
	$\mathcal{N}(0, \sigma^2 = \frac{100\pi}{\pi-2})$	69.8%	68.2%	64.2%	57.4%	50.6%	42.4%	35.4%	0.0%	0.0%
$\sigma^2_{\text{train}} = 400.0$	Exp(1/20)	**70.4%**	**69.8%**	**67.0%**	**63.4%**	**57.0%**	**49.4%**	**43.2%**	**37.4%**	**32.6%**
	$\mathcal{U}([0, 40\sqrt{3}])$	57.0%	56.8%	54.8%	51.0%	47.8%	40.8%	0.0%	0.0%	0.0%
	$\mathcal{N}(0, \frac{400\pi}{\pi-2})$	64.8%	63.6%	61.0%	56.6%	50.4%	43.6%	37.0%	32.4%	27.4%

From both the table and figure, we find that smoothing using exponential noise is much better than using uniform noise or normal distribution noise, when the noise variance is the same, which confirms our theoretical analysis.

H.4 Detailed Experimental Results for specific Semantic Transformations

H.4.1 Gaussian Blur

Table H.6 presents the robust accuracy for different kernel radii α for Gaussian blur transformation. We adopt exponential noise $\text{Exp}([0,\ 1/\lambda])$ for both training data augmentation and robust certification and in each setting, we use the same λ for training and certifying.

To compare the effect of noise variance, in Figure 4 we compare different λ on CIFAR-10. From both the table and figure, we observe that larger λ, i.e., larger variance brings better robust accuracy for large perturbation, but on the other hand hurts robust accuracy for small perturbations and clean accuracy. Previous works on ℓ_2 robustness [5] and our theoretical analysis (Corollaries C.1-C.5) indicate a similar phenomenon.

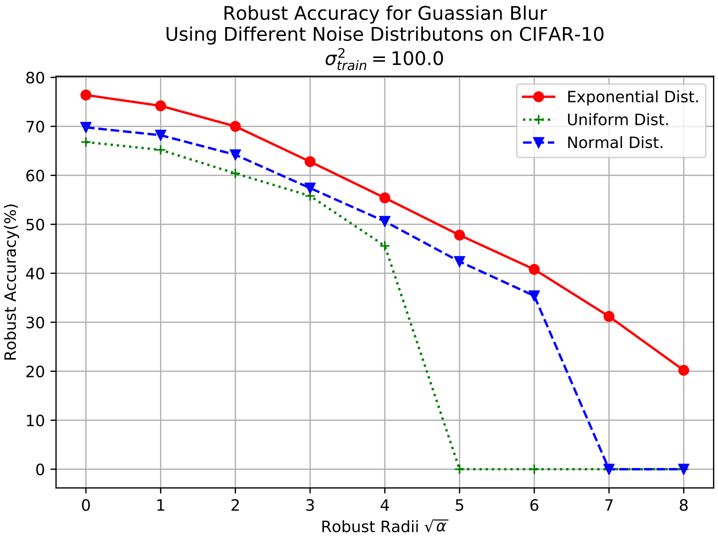


Figure 3: Comparison of different noise distributions for Gaussian blur on CIFAR-10 with $\sigma^2_{\text{train}} = 100$.

Table H.6: Clean and robust accuracy for different kernel radii for *Gaussian Blur*. Smoothing and data augmentation are performed with exponential noise $\text{Exp}(1/\lambda)$.

Dataset	$\text{Exp}(1/\lambda)$	Clean Acc.	Robust Acc. for radius $\alpha \leq$							
			1.0	4.0	9.0	16.0	25.0	36.0	49.0	64.0
MNIST	$\lambda = 2.0$	97.8%	**97.6%**	**95.8%**	**90.4%**	70.4%	0.0%			
	$\lambda = 10.0$	93.8%	93.6%	91.2%	89.2%	**85.8%**	**80.8%**	**73.2%**	**62.8%**	**44.8%**
CIFAR-10	$\lambda = 2.0$	88.2%	**84.6%**	**76.6%**	58.4%	0.0%				
	$\lambda = 10.0$	76.4%	74.2%	70.0%	62.8%	55.4%	47.8%	40.8%	31.2%	20.2%
	$\lambda = 20.0$	70.4%	69.8%	67.0%	**63.4%**	**57.0%**	**49.4%**	**43.2%**	**37.4%**	**32.6%**
ImageNet	$\lambda = 10.0$	60.0%	58.0%	58.0%	55.0%	51.0%	42.0%	37.0%	24.0%	20.0%

H.4.2 Brightness and Contrast

We present the clean and robust accuracy against brightness change, contrast change and the composition of brightness and contrast change in Tables H.7, H.8 and H.9. The parameter b stands for the brightness change, which adds b uniformly to every pixel value. Note that the pixel value is normalized to the range $[0, 1]$ so $b = +0.1$ stands for brightness change $+25.5/255$. The parameter k determines the change in contrast and is applied by uniformly multiplying all pixel values with e^k. For example, $e^k - 1 = -10\%$ means -10% contrast adjustment and multiplying all pixel values by 0.9. The training and certification settings are specified by σ_b and σ_s, which are the standard deviation of the Gaussian noise applied on brightness and contrast dimension respectively. We use the same noise magnitude during training and certification.

H.4.3 Translation

The robust accuracy against translation with reflection-padding and black-padding is shown in Tables H.10 and H.11. We report the robust accuracy obtained from two methods: randomized smoothing and enumeration. Note that when we consider translation with black-padding, we can only use the enumeration approach. The models are trained with data augmentation using displacement parameters sampled from Gaussian noise. The same variance $\sigma$is used for both x- and y- axis. We use the same distribution when certifying with

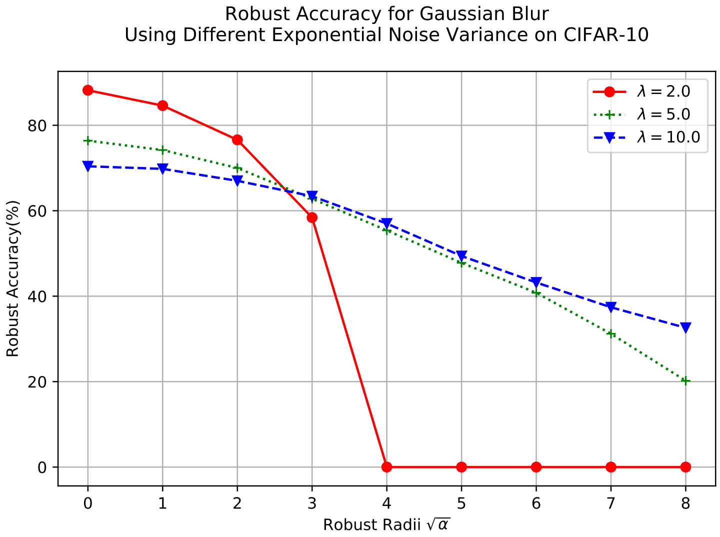


Figure 4: Comparison of different λ used in noise distribution $\mathrm{Exp}(1/\lambda)$ for Gaussian blur on CIFAR-10.

Table H.7: Clean and robust accuracy for *Brightness*. During inference with the smoothed classifier, we set $\sigma_s = 0$ to disable contrast changes which improves robust accuracy.

Dataset	Setting	Clean Acc.	Robust Acc. for $\lvert b \rvert \leq$					
			0.1	0.2	0.3	0.4	0.5	0.6
MNIST	$\sigma_b = \sigma_s = 0.3$	98.6%	98.6%	98.6%	98.6%	98.6%	98.6%	98.6%
CIFAR-10	$\sigma_b = \sigma_s = 0.2$	84.8%	84.2%	83.6%	83.0%	82.6%	82.0%	81.2%
ImageNet	$\sigma_b = \sigma_s = 0.2$	68.0%	64.0%	63.0%	61.0%	60.0%	60.0%	59.0%

Table H.8: Clean and robust accuracy for *Contrast*.

Dataset	Setting	Clean Acc.	Robust Acc. for $\lvert e^k - 1 \rvert \leq$				
			10%	20%	30%	40%	50%
MNIST	$\sigma_b = \sigma_s = 0.3$	98.6%	98.6%	98.6%	98.6%	97.8%	0.0%
CIFAR-10	$\sigma_b = \sigma_s = 0.2$	85.0%	82.8%	80.6%	76.8%	0.0%	
ImageNet	$\sigma_b = \sigma_s = 0.2$	67.0%	62.0%	60.0%	56.0%	0.0%	

Table H.9: Clean and robust accuracy for *Contrast and Brightness Composition*.

Dataset	Setting	Clean Acc.	Robust Acc. for $\lvert e^k - 1 \rvert \leq 20\% \wedge \lvert b \rvert \leq$					
			0.0	0.1	0.2	0.3	0.4	0.5
MNIST	$\sigma_b = \sigma_s = 0.3$	98.6%	98.6%	98.6%	98.6%	98.6%	98.2%	97.8%
CIFAR-10	$\sigma_b = \sigma_s = 0.2$	85.0%	80.0%	78.4%	77.4%	75.8%	72.0%	0.0%
ImageNet	$\sigma_b = \sigma_s = 0.2$	67.0%	60.0%	60.0%	57.0%	56.0%	54.0%	0.0%

randomized smoothing. The quantity $\sqrt{\Delta x^2 + \Delta y^2}$ in the tables specifies the displacement magnitude, where Δx corresponds to the x axis displacement and Δy to the y axis displacement.

Table H.10: Clean and robust accuracy for *Translation with reflection-padding*.

Dataset	$\mathcal{N}(0,\, \sigma^2)$	Method	Clean Acc.	Robust Acc. for $\sqrt{\Delta x^2 + \Delta y^2} \leq$								
				1.0	2.0	5.0	7.5	10.0	12.5	15.0	17.5	20.0
MNIST	$\sigma = 3.0$	Rand. Smooth	99.0%	99.0%	**98.6%**	96.0%	**92.2%**	59.0%	0.0%			
		Enumeration	**99.4%**	**99.4%**	98.2%	**96.8%**	91.8%	**74.8%**	**23.2%**	**2.8%**	**0.2%**	0.0%
	$\sigma = 10.0$	Rand. Smooth	**99.4%**	**99.4%**	**99.2%**	**99.2%**	**98.6%**	**97.8%**	**97.2%**	**96.2%**	**94.6%**	**93.2%**
		Enumeration	99.0%	99.0%	98.6%	96.4%	93.0%	90.0%	85.6%	84.2%	82.4%	81.8%
CIFAR-10	$\sigma = 3.0$	Rand. Smooth	**90.4%**	88.8%	**86.6%**	**76.0%**	**64.6%**	49.6%	0.0%			
		Enumeration	89.4%	**89.4%**	84.6%	74.4%	62.6%	**51.6%**	**35.2%**	**25.6%**	**19.6%**	**16.0%**
	$\sigma = 10.0$	Rand. Smooth	**88.0%**	87.2%	**86.4%**	**84.8%**	**82.0%**	**79.8%**	**76.0%**	**71.2%**	**67.0%**	**63.8%**
		Enumeration	87.6%	**87.6%**	82.0%	74.4%	68.2%	64.6%	57.4%	51.8%	49.2%	47.2%
ImageNet	$\sigma = 3.0$	Rand. Smooth	**63.0%**	**63.0%**	**62.0%**	**56.0%**	**54.0%**	0.0%				
		Enumeration	**63.0%**	**63.0%**	58.0%	**56.0%**	**54.0%**	**53.0%**	**53.0%**	**50.0%**	**48%**	**47.0%**
	$\sigma = 10.0$	Rand. Smooth	66.0%	66.0%	**66.0%**	**63.0%**	**61.0%**	**60.0%**	**57.0%**	**54.0%**	**54.0%**	**53.0%**
		Enumeration	**69.0%**	**69.0%**	59.0%	54.0%	53.0%	53.0%	53.0%	53.0%	52.0%	52.0%

Table H.11: Clean and robust accuracy for *Translation with black-padding*.

Dataset	$\mathcal{N}(0, \sigma^2)$	Clean Acc.	Robust Acc. for $\sqrt{\Delta x^2 + \Delta y^2} \leq$								
			1.0	2.0	5.0	7.5	10.0	12.5	15.0	17.5	20.0
MNIST	$\sigma = 3.0$	**99.2%**	**99.2%**	**98.8%**	93.6%	80.4%	45.8%	8.6%	3.2%	**2.6%**	**2.6%**
	$\sigma = 10.0$	99.0%	99.0%	97.6%	**94.8%**	**88.4%**	**75.2%**	**42.4%**	**12.4%**	0.8%	0.2%
CIFAR-10	$\sigma = 3.0$	**88.6%**	**88.6%**	**84.2%**	**77.0%**	64.6%	54.2%	37.0%	19.8%	6.0%	0.6%
	$\sigma = 10.0$	86.2%	86.2%	80.8%	74.6%	**67.6%**	**60.8%**	**47.4%**	**33.2%**	**21.2%**	**11.4%**
ImageNet	$\sigma = 3.0$	**71.0%**	**71.0%**	**66.0%**	58.0%	51.0%	50.0%	49.0%	48.0%	48.0%	47.0%
	$\sigma = 10.0$	67.0%	67.0%	**66.0%**	**61.0%**	**58.0%**	**57.0%**	**55.0%**	**53.0%**	**52.0%**	**50.0%**

Randomized Smoothing vs. Enumeration. To compare the performance of randomized smoothing and enumeration, we plot the robust accuracy curve with respect to the displacement $\sqrt{\Delta x^2 + \Delta y^2}$ for all $\sigma = 10$ models in Fig. 5. The randomized smoothing results are shown by solid lines, while enumeration results are shown by dotted lines. We observe that in most cases, randomized smoothing outperforms the enumeration approach in terms of robust accuracy. We attribute this to the power of smoothing, which flattens the extreme points of single-point prediction.

Reflection-Padding vs. Black Padding. In Figure 6 we compare the best achieved robust accuracy of reflection-padding and black-padding on all three datasets with models trained using $\sigma = 10$. We observe that the robust accuracy of reflection-padding is much higher than that of black-padding. We attribute this observation to two causes: 1) For reflection-padding, randomized smoothing can be used, which usually results in higher robust accuracy than enumeration; and 2) For reflection-padding, the out-of-margin pixels reappear at the opposite side, while for black-padding, they are replaced by black pixels, resulting in information loss.

H.4.4 Rotation

We present the clean and robust accuracy for rotations in Table H.12. During training, we first apply rotations with angles sampled uniformly at random from the interval $[\pm\alpha_{\text{train}}]$, then add additive Gaussian noise with variance σ^2_{train}. During certification, we do randomized smoothing with additive Gaussian noise with variance σ^2_{test}, and certify the rotation robustness against angles from the interval $[\pm\alpha_{\text{test}}]$ using our sampling-based approach. For certification, in our sampling-based approach, we use $N = 10,000$ and $R = 1,000$. In other words, we partition the whole rotation angle interval $[-180°,\, 180°]$ to $N = 10,000$ small intervals. We

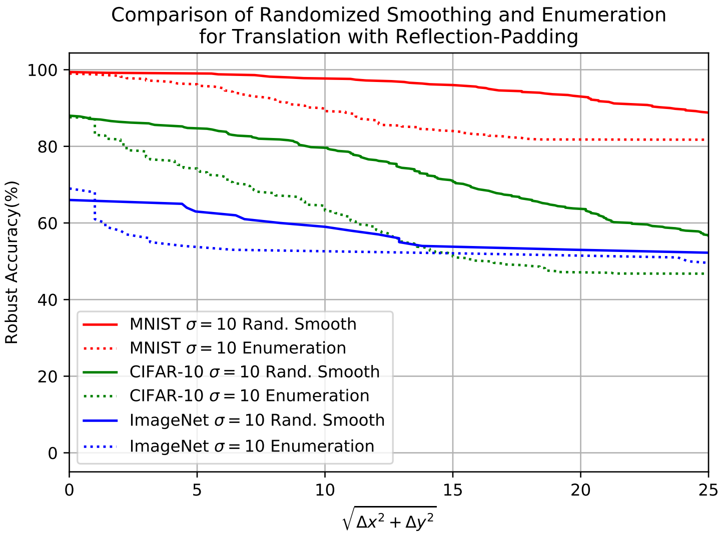


Figure 5: Comparison of Randomized Smoothing and Enumeration for Translation with Reflection-Padding. The models are trained using $\sigma = 10$ Gaussian noise data augmentation. Randomized smoothing results are shown by solid lines, and enumeration results are shown by dotted lines.

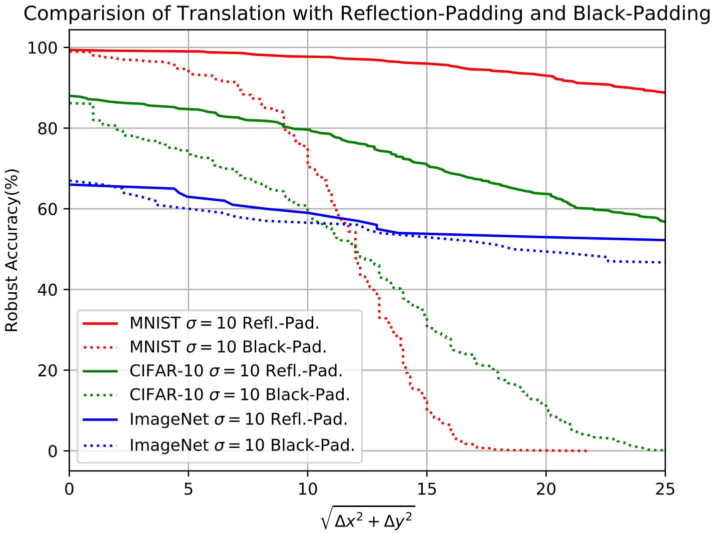


Figure 6: Camparison of Best Achieved Robust Accuracy of Reflection-Padding and Black-Padding. The models are trained using $\sigma = 10$ Guassian noise data augmentation. Reflection-padding results are shown by solid lines, and enumeration results are shown by dotted lines.

then further sample $R = 1,000$ points for the local max aliasing computation. Note that since we only certify $|\alpha_{\text{test}}| \leq 30°$, in our implementation we only consider $10,000 \times \frac{30}{180} \leq 2,000$ intervals. This selection of hyperparameters balances the efficiency and precision as shown in Section H.5. From the table, we find that the training and certification noise variance σ^2_{train} and σ^2_{test} does not play an important role in general. However, we find that not applying noise (see CIFAR-10 $\sigma_{\text{train}} = 0.00$ row) harms robust accuracy considerably.

Table H.12: Clean and robust accuracy for *rotations*. During training, data augmentation is applied with angles sampled uniformly at random from the interval $[\pm\alpha_{\text{train}}]$. In addition we use additive Gaussian noise with variance σ^2_{train}.

Dataset	$\lvert\alpha_{\text{train}}\rvert \leq$	σ_{train}	σ_{test}	Clean Acc.	$\lvert\alpha_{\text{test}}\rvert \leq$	Robust Acc.
MNIST	$35.0°$	0.10	0.10	$\mathbf{99.4}\%$	$30.0°$	92.8%
	$35.0°$	0.25	0.25	99.2%	$30.0°$	94.6%
	$35.0°$	0.50	0.50	98.6%	$30.0°$	$\mathbf{95.6}\%$
CIFAR-10	$12.5°$	0.00	0.05	63.6%	$10.0°$	16.0%
	$12.5°$	0.05	0.05	$\mathbf{84.2}\%$	$10.0°$	$\mathbf{63.8}\%$
	$12.5°$	0.10	0.10	78.4%	$10.0°$	63.2%
	$12.5°$	0.15	0.15	77.4%	$10.0°$	60.8%
ImageNet	$12.5°$	0.25	0.25	$\mathbf{64.0}\%$	$10.0°$	19.0%
	$12.5°$	0.50	0.50	56.0%	$10.0°$	$\mathbf{33.0}\%$

H.4.5 Scaling

We present the clean and robust accuracy for scaling in Table H.13. Similar as for rotation transformations, during training we first apply scaling with a factor s_{train} sampled uniformly at random, then add additive Gaussian noise with variance σ^2. During certification, we do randomized smoothing with additive Gaussian noise with variance σ^2 and certify the scaling robustness against factor s using our sampling-based approach. For certification, in our sampling-based approach, we use $N = 1,000$ and $R = 250$. In other words, we partition the whole scaling factor interval to $1,000$ small intervals and then further sample 250 points for the local max aliasing computation. This selection of hyperparameters balances the efficiency and precision as shown in Section H.5. Similar as rotation, the noise variance σ^2 has a small effect on the robust accuracy.

Table H.13: Clean and robust accuracy for *scaling transformations*. During training, data augmentation is applied with scaling factors sampeld uniformly at random from the interval $[1 \pm s_{\text{train}}]$. In addition we use additive Gaussian noise with variance σ^2.

Dataset	$\lvert s_{\text{train}} - 1\rvert \leq$	σ	Clean Acc.	$\lvert s - 1\rvert \leq$	Robust Acc.
MNIST	25.0%	0.10	$\mathbf{98.8}\%$	20.0%	$\mathbf{96.8}\%$
	25.0%	0.25	$\mathbf{98.8}\%$	20.0%	95.6%
	25.0%	0.50	98.6%	20.0%	95.8%
CIFAR-10	25.0%	0.05	$\mathbf{86.8}\%$	20.0%	57.2%
	25.0%	0.10	81.4%	20.0%	57.8%
	25.0%	0.15	82.0%	20.0%	$\mathbf{58.4}\%$
	25.0%	0.20	75.0%	20.0%	53.4%
	25.0%	0.25	78.6%	20.0%	58.2%
	25.0%	0.30	74.6%	20.0%	54.0%
ImageNet	15.0%	0.50	58.0%	15.0%	31.0%

H.5 For Sampling-Based Certification: Effects of Number of Samples

In Table H.14 we compare the effect of different sampling numbers. We find that the time needed to compute the sampling error M is roughly linearly proportional to $N \times R$. Increasing N and R significantly reduces the maximum sampling error upper bound M. For fixed $N \times R$, increasing N (and decreasing R) reduces M more effectively, opposed to increasing R and decreasing N. Furthermore, we observe that the robust accuracy is directly related with M. We observe that when M is smaller, a smaller ℓ_2 robustness margin is required to certify a sample. We furthermore find that models with larger σ usually have larger ℓ_2 robustness margin, and thus tolerate larger M (compare $\sigma = 0.05$ column and $\sigma = 0.15$ column). We also find that certification time does not scale linearly with N, while the running time to compute M does.

Table H.14: Effect of the number of samples for sampling-based certification for rotation transformation. The rotation angle interval $[\pm 180^\circ]$ is uniformly divided into N subintervals, then each interval is further divided R times in order to estimate the upper bound of the maximum ℓ_2 sampling error M. Note that we only certify $[\pm 10^\circ]$ while the outer intervals are skipped. The time needed to compute M for each sample is reported in the column "Computing Time". The robust accuracy and certification time of each model (specified by the noise std σ) is shown in the corresponding columns. The results are computed from 100 random samples from the CIFAR-10 test set.

Sampling Numbers	Sampling Err. M	Computing Time	$\sigma = 0.05$		$\sigma = 0.10$		$\sigma = 0.15$	
			Rob. Acc.	Certify Time	Rob. Acc.	Certify Time	Rob. Acc.	Certify Time
$N = 100, R = 100$	11.0	1.069 s	0%	13.5 s	0%	25.7 s	0%	11.9 s
$N = 100, R = 1000$	8.17	5.701 s	0%	13.2 s	0%	25.7 s	0%	11.3 s
$N = 1000, R = 100$	0.414	7.58 s	0%	14.0 s	0%	21.6 s	4%	13.9 s
$N = 1000, R = 1000$	0.122	80.35 s	2%	14.1 s	22%	73.9 s	28%	13.8 s
$N = 1000, R = 10000$	0.0932	660.2 s	3%	23.0 s	30%	56.2 s	32%	14.3 s
$N = 10000, R = 1000$	0.00417	669.1 s	59%	28.7 s	63%	37.5 s	59%	63.3 s

H.6 Running Time

Table H.15 records the certification time per sample in seconds. Our certification procedur uses one NVIDIA GTX 1080Ti GPU with a single core Intel Xeon E5-2650 CPU. Note that for most settings and most methods, the average time is less than 60 s, even for the large ImageNet ResNet-50 model, indicating that the approach is scalable.

Moreover, from the table we find that enumeration method runs the fastest, taking less than 1 s even for the slowest sample. Randomized smoothing method running time is very stable, mostly in the range from 10 s to 30 s. Note that randomized smoothing running time on ImageNet is comparable to those on MNIST and CIFAR-10 mainly because we set $N = 10,000$ on ImageNet and $N = 100,000$ on other datasets. Sampling randomized smoothing running time is on average the slowest on CIFAR-10 and ImageNet. We notice that the running time varies largely from sample to sample. For example, for rotations on ImageNet, the fastest sample takes only 0.152 s and the slowest sample takes 14 880 s. We attribute this to the effect of progressive certification, which significantly accelerates the certification process of easy-to-verify samples.

Table H.15: Certification time per sample for all types for transformations on all datasets.

Dataset	Transformation	Setting	Method	Avg.(s)	Min.(s)	Max.(s)
MNIST	Gaussian Blur	$\lambda = 2.0$	Rand. Smooth	16.9	14.7	20.5
		$\lambda = 10.0$	Rand. Smooth	17.3	14.0	21.7
	Brightness	$\sigma_b = \sigma_s = 0.3$	Rand. Smooth	10.3	9.33	11.9
	Contrast	$\sigma_b = \sigma_s = 0.3$	Rand. Smooth	10.5	9.25	12.7
	Contrast and Brightness	$\sigma_b = \sigma_s = 0.3$	Rand. Smooth	11.6	9.18	13.7
	Translation w/ Reflection-Pad.	$\sigma = 3$	Rand. Smooth	17.1	14.5	19.6
		$\sigma = 10$	Rand. Smooth	18.6	16.7	22.2
		$\sigma = 3$	Enumeration	0.132	0.108	0.496
		$\sigma = 10$	Enumeration	0.378	0.11	0.995
	Translation w/ Black-Pad.	$\sigma = 3$	Enumeration	0.0735	0.0506	0.632
		$\sigma = 10$	Enumeration	0.0676	0.0491	0.595
	Rotation	$\sigma_{\text{test}} = 0.10$	Sampling Rand. Smooth	4.59	0.00155	143
		$\sigma_{\text{test}} = 0.25$	Sampling Rand. Smooth	0.942	0.0016	2.63
		$\sigma_{\text{test}} = 0.50$	Sampling Rand. Smooth	0.851	0.00147	3.84
	Scaling	$\sigma = 0.10$	Sampling Rand. Smooth	4.78	0.00145	6.44
		$\sigma = 0.25$	Sampling Rand. Smooth	5.05	0.00151	24.3
		$\sigma = 0.50$	Sampling Rand. Smooth	5.32	0.00144	29.2
CIFAR-10	Gaussian Blur	$\lambda = 2.0$	Rand. Smooth	28.4	26.9	30.9
		$\lambda = 10.0$	Rand. Smooth	30.8	29.5	33.7
		$\lambda = 50.0$	Rand. Smooth	36.9	34.9	40.4
	Brightness	$\sigma_b = \sigma_s = 0.2$	Rand. Smooth	23.9	22.6	26.5
	Contrast	$\sigma_b = \sigma_s = 0.2$	Rand. Smooth	26.2	23.3	29.8
	Contrast and Brightness	$\sigma_b = \sigma_s = 0.2$	Rand. Smooth	27.6	22.7	42.4
	Translation w/ Reflection-Pad.	$\sigma = 3$	Rand. Smooth	26.8	25.0	28.8
		$\sigma = 10$	Rand. Smooth	28.2	25.6	29.7
		$\sigma = 3$	Enumeration	0.389	0.245	1.15
		$\sigma = 10$	Enumeration	0.591	0.222	1.21
	Translation w/ Black-Pad.	$\sigma = 3$	Enumeration	0.186	0.172	0.462
		$\sigma = 10$	Enumeration	0.213	0.172	0.526
	Rotation	$\sigma_{\text{train}} = 0.00, \sigma_{\text{test}} = 0.05$	Sampling Rand. Smooth	304	0.021	2852
		$\sigma_{\text{test}} = 0.05$	Sampling Rand. Smooth	431	0.0593	2633
		$\sigma_{\text{test}} = 0.10$	Sampling Rand. Smooth	413	0.0243	2300
		$\sigma_{\text{test}} = 0.15$	Sampling Rand. Smooth	353	0.0834	3501
	Scaling	$\sigma = 0.05$	Sampling Rand. Smooth	987	0.0381	13423
		$\sigma = 0.10$	Sampling Rand. Smooth	693	0.0184	3438
		$\sigma = 0.15$	Sampling Rand. Smooth	743	0.0188	3720
		$\sigma = 0.20$	Sampling Rand. Smooth	638	0.0533	2798
		$\sigma = 0.25$	Sampling Rand. Smooth	747	0.0367	4940
		$\sigma = 0.30$	Sampling Rand. Smooth	693	0.0189	6809
ImageNet	Gaussian Blur	$\lambda = 10.0$	Rand. Smooth	60.2	56.8	63.8
	Brightness	$\sigma_b = \sigma_s = 0.2$	Rand. Smooth	21.2	19.6	25.0
	Contrast	$\sigma_b = \sigma_s = 0.2$	Rand. Smooth	21.2	19.8	25.0
	Contrast and Brightness	$\sigma_b = \sigma_s = 0.2$	Rand. Smooth	21.9	21.2	26.7
	Translation w/ Reflection-Pad.	$\sigma = 3$	Rand. Smooth	20.1	19.6	25.0
		$\sigma = 10$	Rand. Smooth	20.2	19.6	25.4
		$\sigma = 3$	Enumeration	121	0.723	450
		$\sigma = 10$	Enumeration	134	0.719	460
	Translation w/ Black-Pad.	$\sigma = 3$	Enumeration	22.9	0.685	194
		$\sigma = 10$	Enumeration	25.9	0.685	211
	Rotation	$\sigma_{\text{test}} = 0.25$	Sampling Rand. Smooth	767	0.152	14880
		$\sigma_{\text{test}} = 0.50$	Sampling Rand. Smooth	336	0.153	2947
	Scaling	$\sigma = 0.50$	Sampling Rand. Smooth	280	0.152	1828

I Comparison to Other Semantic Transformation Certification Approaches

We compare our semantic transformation certification approaches with recent and concurrent works [34, 3, 27, 10].

(Singh et al) [34]. Within our knowledge, [34] proposes the first approach to certify neural network robustness against rotation transformations. In order to certify robustnes against rotations, they split the interval of rotation angles to small sub-intervals and subsequently compute each pixel's color range and perform interval bound propagation to derive the final output range. However, the certification is relatively loose and does not scale to larger datasets such as ImageNet. Our sampling-based approach for rotation and scaling is similar in that we also split the rotation angle or scaling factor interval. However, we do a rigorous analysis on the ℓ_2 aliasing and bound the maximum ℓ_2 sampling error instead. This allows us to circumvent the looseness of interval bound propagation and effectively exploit the strong ℓ_2-certification of randomized smoothing.

(Balunovic et al) [3]. [3] improves on [34] by alternating the interval bound for per pixel value by linear constraints. Similar to our method, they also first split the transformation parameter interval. After this, they optimize the per pixel linear constraints via sampling and optimizing from general Lipschitz optimization. The approach obtains substantially better robust accuracy compared to previous works. In our sampling-based approach for rotation and scaling, we analyze the concrete form of the Lipschitz upper bound and, rather than expressing the aliasing bounds using linear constraints, we use the ℓ_2 norm which significantly improves efficiency and scalability. Note that [3] also cover the composition of geometric transformations such as composition of scaling and shearing. Our sampling-based approach can also be extended to support those types of transformations but requires additional Lipschitz bound analysis.

(Mohapatra et al) [27]. [27] proposes to reduce robustness verification against semantic transformation to classical ℓ_p robustness verification problem by expressing semantic transformation using regular neural network layers. Therefore, existing ℓ_p-based deterministic verification approaches such as [40, 20] can be directly applied to certify robust region with respect to transformation parameters. However, the semantic verification is limited to transformations which can be precisely expressed by neural networks. Rotation and scaling, in their approach, requires large number of interval splits. They do not report robust accuracy (but only average certified radius) nor open-source their code within our best knowledge, therefore it is not comparable with our approach. Moreover, based on existing linear-relaxation based verification approaches, their approach cannot scale up to large neural networks on ImageNet.

(Fischer et al) [10]. [10] is the first to generalize randomized smoothing to semantic transformations and certify rotations and translation for ImageNet classification. They furthermore cover volume changes and pitch shifts on audio data and robustness to floating-point soundness. Their theoretical results generalize randomized smoothing to a richer class of transformations and go beyond isotropic Gaussian noise. However their result is limited to additive transformations while our theoretical result generalizes to a broader class of noise distributions and to transforms which are resolvable. In addition, we provide theoretical barriers on the randomized smoothing approach enabling a deeper understanding.

Since rotations are by nature non-additive transforms, [10] overcome the limitation to additive transforms by computing the maximum ℓ_2-error induced from random sampling and use a statistical upper bound. However, an attacker can deterministically select the transformation parameter with maximum ℓ_2-error in order to break the certification. Our sampling-based approach on the other hand provides us with a rigorous upper bound resulting in much stronger certification.