Double Bubble, Toil and Trouble: Enhancing Certified Robustness through Transitivity

Anonymous Author(s)
Affiliation
Address
email

Abstract

In response to subtle adversarial examples flipping classifications of neural network models, recent research has promoted certified robustness as a solution. There, invariance of predictions to all norm-bounded attacks is achieved through randomised smoothing of network inputs. Today's state-of-the-art certifications make optimal use of the class output scores at the input instance under test: no better radius of certification (under the L_2 norm) is possible given only these score. However, it is an open question as to whether such *lower bounds* can be improved using local information around the instance under test. In this work, we demonstrate how today's "optimal" certificates can be improved by exploiting both the transitivity of certifications, and the geometry of the input space, giving rise to what we term *Geometrically-Informed Certified Robustness*. Essentially we search for, and merge with, larger certifiable balls nearby to an initial instance under test. Our approach improves certifications for more than 80% of instances in experiments with Tiny-Imagenet, with an on average 5% increase in the associated radius of certification.

1 Introduction

Learned models, including neural networks, are well known to be susceptible to crafted perturbations inducing changes in a model's output without changing semantic properties of the input [2]. Concerningly, neural networks not only misclassify these perturbations—known as *adversarial examples*—but they also assign high confidence to these incorrect predictions. These behaviours have been observed across a wide range of models and datasets, and appear to be a product of piecewise-linear interactions within neural networks [9].

Crafting these adversarial examples typically involves gradient-based optimisation to construct *small* perturbations. These attacks have been applied to both black- and white-box models [21], and can be used to target class changes, to attack all classes [6], or even introduce backdoors into model behaviour [4]. To mitigate the influence of these attacks, defences have typically been designed to minimise the effect of a specific attack (or attacks). Such defences are known as *best response* strategies in a Stackelberg security game where the defender leads the attacker. Best response defences inherently favour the attacker, as deployed mitigations can be defeated by identifying undefended attack frameworks. Moreover, the defender typically has to incorporate the defence at training time, and as such cannot response reactively to newly developed attacks.

In an attempt to circumvent these limitations, certified guarantees of adversarial robustness can be constructed, to identify class-constant regions around an input instance. Irrespective of the attack, all instances within a certain distance from an input (typically L_2 distance) are guaranteed not to be adversarial examples. Certifications based on randomised smoothing of classifiers around an input point are in a sense optimal [5]: based only on the prediction class scores at the input point, no better

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

radius is in general possible. Despite this, such certifications fail to use readily available—yet still local—information: the certifiability of points nearby to the input of interest. The key insight of this work is that these neighbourhood points may generate certified radii large enough that we may enlarge the certified region based solely on classifier information at the initial point. This process is not just limited to a single neighbouring point either, as we can also use the intersection of the regions of certification of multiple points, and the nature of the input domain itself to generate larger certifications. This leads to our main contribution—*Geometrically-Informed Certified Robustness*—that enjoys certifications exceeding those of the hitherto best-case guaranteed approach of Cohen *et al.* (2019) [5].

2 Background and literature review

Bounding mechanisms An alternate perspective on bounding the impact of norm-bounded perturbations has been to construct conservative bounds on the output of a model [19, 29, 32, 33, 25]. Framed as convex relaxation, these techniques utilise linear relaxation to construct bounding polytopes of a model's outputs over input bounded perturbations [24]. These approaches have been leveraged by adopting augmented loss functions during training, to promote tight output bounds [30]. While bounding mechanisms can produce valid certifications, their time and memory complexity makes them infeasible to address complex model architectures or high-dimensional data [28].

Randomised smoothing Outside of bounding mechanisms, another common framework for developing certifications leverages *randomised smoothing* [17], in which noise is applied to input instances to smooth model predictions. In contrast to other robustness mechanisms, randomised smoothing only requires minimal changes to a classifier, specifically the repeated sampling of noise-based perturbation of the inputs, where the choice of sampling distribution is tied to the L_P-norm of adversarial perturbations being certified against. In the case of L_2-norm bounded attacks, Gaussian sampling of the form

$$\mathbf{x}_i' = \mathbf{x} + \mathbf{y}_i \qquad \text{where } \mathbf{y}_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2) \;\; \forall i \in \{1, \ldots, N\} \tag{1}$$

is employed for all test-time instances. These N samples are then used to estimate the expected output of the predicted class of $\mathbf{x}$ by way of the Monte-Carlo estimator

$$E_{\mathbf{Y}}[\arg\max f_\theta(\mathbf{x} + \mathbf{Y}) = i] \approx \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}[\arg\max f_\theta(\mathbf{x}_j) = i] \;\; . \tag{2}$$

While this Monte Carlo estimation of output expectations under randomised smoothing is a test-time process, model sensitivity to random perturbations may be decreased by performing adversarial training on such random perturbations. To mitigate the computational expense of large N sample sizes during each training update, training typically employs single draws from the noise distribution.

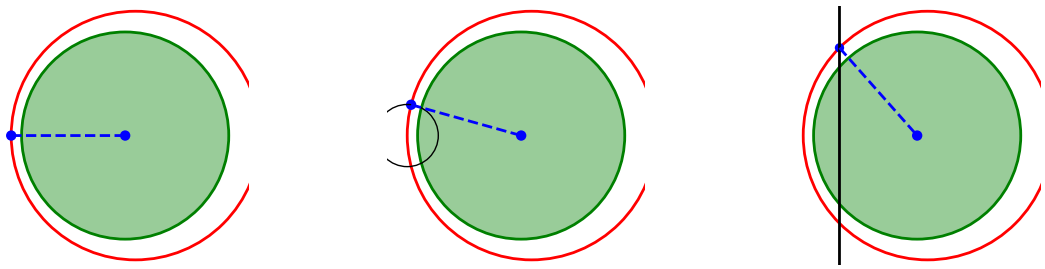
Smoothing-based certifications Based on randomised smoothing, certified robustness can guarantee classification invariance for *additive* perturbations up to some L_p-norm r. Certified robustness has not, however, extended to rotational or translational modifications [27], nor functional attacks [15].

That such guarantees are possible was first shown [17] by way of differential privacy [7], with more recent works using Rényi divergence [18], and parametrising worst-case behaviours [5, 23]. In this paper, we focus on the influential work of Cohen *et al.* (2019), and their lower-bound guarantee of classification invariance

$$r = \frac{\sigma}{2} \left(\Phi^{-1}\left(E_0[\mathbf{x}]\right) - \Phi^{-1}\left(E_1[\mathbf{x}]\right) \right) \;\; , \tag{3}$$

where Φ^{-1} is the inverse normal CDF (the Gaussian quantile function), and E_0 and E_1 denote the two largest class expectations (of the winning class and runner). These expectations capture the likelihood that the $\arg\max$ of a single draw under noise would return a given class, as per Equation (2). By parameterising the worst-case behaviour, their work purports to be the largest achievable pointwise certification for any given instance [5].

Prior works have been able to generate improvements to Cohen *et al.* for binomial classifiers [20], in which it is assumed that $E_1 = 1 - E_0$. While any multivariate model can be mapped to a binary output for the purposes of constructing a certification, artificially inflating E_1 inherently suppresses the size of the certification, as per Equation (3). In this paper we demonstrate a novel approach that can generate improvements to the certified robustness of multivariate classifiers.



(a) Transitivity (b) Multiple Transitivity (c) Boundary Treatment

Figure 1: Exemplars of how transitivity can be used to enhance the level of achieved certification. Blue represents the instance and nearest possible adversarial example; Green circle the original radii of certification; Red the certification region incorporating an eccentric hypersphere; and Black represents either a second eccentric hypersphere or the domain boundary.

3 Geometrically-informed certified robustness

While the work contained within this paper can be applied generally, for this work we will focus upon certifications of robustness about L_2-norm bounded adversarial perturbations, for which we assume that the difficulty of attacking a model is proportional to the size of the certification, based upon the need to evade both human *and* machine scrutiny [8]. Thus, constructing larger certifications in such a context is inherently valuable.

This specific L_P space is of interest due to both its viability as a defence model, and the existence of a guarantee of adversarial robustness that is provably as large as can be achieved for any instance [5]. Over the remainder of this section we will document how it is possible to improve upon this provably best-case guarantee by exploiting several properties of certified robustness.

3.1 Exploiting transitivity

While it is provably true that Equation (3) is the largest achievable certification for any point $\mathbf{x}$, it is possible to exploit the behaviour of points in the neighbourhood of $\mathbf{x}$ in order to enhance certifiable radius. To achieve this, consider the case of a second point $\mathbf{x}'$, that exists within the certifiable radii of $\mathbf{x}$. As both points must correspond to the same class, it then follows that the union of their regions of certification can be also be considered as a region of certification, leading to Definition 3.1.

Definition 3.1 (Overlap Properties of Certification). If $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same class associated with them, then the region of certification can be described as $B_P(\mathbf{x}_1, r_1) \cup B_P(\mathbf{x}_2, r_2)$, where r_1 and r_2 are respectively the radii of certification for $\mathbf{x}_1$ and $\mathbf{x}_2$.

However, when considering certification, we are often concerned not with the size of the region of classification invariance, but rather the distance to the nearest possible adversarial example. If it is possible to find some $\mathbf{x}'$ such that its region of certification completely encircles that of the certification at $\mathbf{x}$, the following definition demonstrates that the certification radius can be increased by way of Equation (4).

Lemma 3.2 (Set Unions Certified Radius). *If $\mathbf{x}_1$ and $\mathbf{x}_2$ have the same class associated with them and $B_P(\mathbf{x}_1, r_1) \subset B_P(\mathbf{x}_2, r_2)$, then the nearest possible adversarial example—and thus, the certifiable radius—exists at a distance of*

$$r' = r_2 - \|\mathbf{x}_2 - \mathbf{x}_1\|_P \quad , \tag{4}$$

from $\mathbf{x}_1$, where $r' \geq r_1$.

Proof. The closest point on the surface of $B_P(\mathbf{x}_2, r_2)$ to $\mathbf{x}_1$ must exist on the vector between $\mathbf{x}_1$ and $\mathbf{x}_2$. Thus $r' = \min\left(r_2 \pm \|\mathbf{x}_2 - \mathbf{x}_1\|_P\right)$, which takes the form of Equation (4). □

As such, we can recast the certification problem from being a strictly analytic function, to the nonlinear optimisation problem

$$r' = \max_{\mathbf{x}_2 \in [0,1]^d} r_2(\mathbf{x}_2) - \|\mathbf{x}_2 - \mathbf{x}_1\|_P \quad , \tag{5}$$

with Figure 1a providing a two-dimensional representation of this process. Crucially, the above formalism does not require obtaining a global optima, as any $r' > r_1$ is guaranteed to be an improvement to the certification at $\mathbf{x}_1$.

3.2 Multiple transitivity

The concept of transitivity can be further enhanced by considering the set $\{\mathbf{x}_1, r_1, \mathbf{x}_2, r_2, \ldots \mathbf{x}_n, r_n\}$. If $x_j \in B_P(\mathbf{x}_1, r_1)$ for all $j > 1$, then the overall region of certification is now the the union of all $B_P(x_i, r_i)$, the exterior boundary of which we describe as $\mathcal{S}_n(\mathbf{s})$, where $\mathbf{s} \in \mathcal{R}^{d-1}$. Based upon the construction of $\mathcal{S}_n$, it follows that the radius of certification is $r_n = \min_{\mathbf{s}} \|S_n(\mathbf{s}) - \mathbf{x}_1\|$.

Definition 3.3 (**Optimal positioning of $\mathbf{x}_3$ in the case of $n = 3$).** In the case where $n = 3$, let us consider two points $\mathbf{x}_3$ and $\mathbf{x}_3'$, where

$$\mathbf{x}_3(s) = \mathbf{x}_1 + sr' \frac{\mathbf{x}_1 - \mathbf{x}_2}{\|\mathbf{x}_1 - \mathbf{x}_2\|} \quad . \tag{6}$$

If no pair of certification hyperspheres exists such that $B_P(\mathbf{x}_3, r_3) \subset B_P(\mathbf{x}_3', r_3')$, then the optimal positioning of $\mathbf{x}_3$ must satisfy Equation (6).

Assuming that the above definition is true allows us to define a new certification using the following Lemma, which aligns with the visual sketch outlined in Figure 1b.

Lemma 3.4 (**Certification from two Eccentric Circles**). *In the case of where two additional eccentric circles are used to construct a bounding region about $\mathbf{x}_1$, Definition 3.3 allows us to demonstrate that the largest achievable radii of certification takes the form*

$$\begin{aligned} r'' &= \max_{\mathbf{x}_3 \in [0,1]^d} \sqrt{\frac{d_2(r_3^2 - d_3^2) + d_3(r_2^2 - d_2^2)}{d_2 + d_3}} \quad \textit{where} \\ d_2 &= \|\mathbf{x}_2 - \mathbf{x}_1\| \\ d_3 &= \|\mathbf{x}_3 - \mathbf{x}_1\| \end{aligned} \tag{7}$$

and where r_3 is the radius of certification about $\mathbf{x}_3$.

Proof. To begin with, we introduce the co-ordinate translation $\mathbf{y} = f(\mathbf{x} - \mathbf{x}_1)$, subject to a strictly rotational mapping $f(\mathbf{x})$ that allows us to state that the j-th components of $\mathbf{y}_2$ and $\mathbf{y}_3$ in this space are defined such that

$$y_{2,j} \neq 0 \text{ and } y_{3,j} \neq 0 \text{ only when } j = k \quad . \tag{8}$$

In these coordinates the intersection of the hyperspheres centred about $\mathbf{x}_2$ and $\mathbf{x}_3$ occurs at

$$\begin{aligned} r_3^2 - r_2^2 &= \|\mathbf{y} - \mathbf{y}_3\|^2 - \|\mathbf{y} - \mathbf{y}_2\|^2 \\ &= 2y_k(d_2 + d_3) + d_3^2 - d_2^2 \\ 2y_k &= \frac{r_3^2 - r_2^2 + d_2^2 - d_3^2}{d_2 + d_3} \quad . \end{aligned} \tag{9}$$

This is possible as our chosen mapping preserve distances under rotation, giving that $y_{2,k} = \|\mathbf{x}_2 - \mathbf{x}_1\| = d_2$, and with the equivalent also holding for $y_{3,k}$.

As the instance is the origin of our new coordinate system, it follows that $r'' = \|\mathbf{y}\|$ and

$$\begin{aligned} (r'')^2 &= \|\mathbf{y}\|^2 \\ &= \|\mathbf{y} - \mathbf{y}_2\|^2 + 2y_k d_2 - d_2^2 \\ &= r_2^2 - d_2^2 + \frac{d_2 r_3^2 - d_2 r_2^2 + d_2 d_2^2 - d_2 d_3^2}{d_2 + d_3} \end{aligned} \tag{10}$$

which is equivalent to Equation (7). □

Intuitively we can assume that this logic should extend to $\mathcal{S}_n$ for $n > 3$. However, while the nearest point from $\mathbf{x}_1$ on $\mathcal{S}_3$ is a point, if we consider $\mathbf{x}_i$ for $i \leq 3$ are fixed, then the nearest point to $\mathbf{x}_1$ is no longer a point, but a surface in $\mathcal{R}^{d-1}$. Thus, finding additional hyperspheres that completely cover this $\mathcal{R}^{d-1}$ surfaces is non-trivial. Similarly, any other arrangement of bounding hyperspheres scales nonlinearly with the intrinsic dimensionality of the space, resulting in exponential growth in the number of hyperspheres that need to be solved for, and thus, in the overall computational cost. This effectively imposes that it is infeasible to solve for any $\mathcal{S}_n$ where $n > 3$.

3.3 Boundary treatments

Without loss of performance or accuracy, we can freely scale the inputs of neural networks such that $\mathbf{x}_1 \in [0,1]^d$. However, for the majority of certifications, a subset of $B_P(\mathbf{x}_1, L_1)$ will be outside the feasible domain for $\mathbf{x}_1$. While this observation is trivially true, we are unable to leverage it to improve the size of the certification, as the distance from $B_P(\mathbf{x}_1, L_1)$ to any point of intersection with $\{0,1\}^d$ will still be of length L_1 away from $\mathbf{x}_1$.

This property also holds for the intersection of $B_P(\mathbf{x}_2, L_2)$ with the bounds of the feasible space. However, as the distance from $\mathbf{x}_1$ to the intersection of $B_P(\mathbf{x}_2, L_2)$ and the feasible space is larger than that from $\mathbf{x}_1$ to the boundary of the hypersphere about $\mathbf{x}_2$, as is demonstrated within Figure 1c. This allows us to make the following observation about improving the feasible radius of certification.

Lemma 3.5 (**Boundary Certifications by way of Eccentric Circles**). *The eccentricity of $B_P(\mathbf{x}_2, L_2)$ as a bounding region about $\mathbf{x}_1$, and the potential for a subset of $B_P(\mathbf{x}_2, L_2)$ to exist outside the feasible space for instances allows us to construct an updated region of certification where*

$$\begin{aligned}\hat{r} = &\max_{k \in \{1,\ldots,d\}} \mathbb{1}\left[r_2^2 - (x_{2,k} - z_k)^2 \geq 0\right] \times \\ &\sqrt{(z_k - x_{1,k})^2 + \left(\sqrt{r_2^2 - (z_k - x_{2,k})^2} - \sqrt{\|\mathbf{x}_2 - \mathbf{x}_1\|^2 - (x_{2,k} - x_{1,k})^2}\right)^2} \\ &\textit{where } z_k = \begin{cases} 0 & \textit{if } x_{1,k} < 0.5 \\ 1 & \textit{if } x_{1,k} > 0.5 \end{cases},\end{aligned} \tag{11}$$

where $\mathbb{1}$ is an indicator function acting upon its operator.

Proof. In contrast to the prior proof, for this problem we retain the coordinate system of the input space. To support this, we introduce the notation that $\mathbf{x}_n$ is defined as as $\{x_{1,n}, x_{2,n}, \ldots, x_{d,n}\}$. If we let $x_k \in \{0,1\}$, then the intersection between $B_P(\mathbf{x}_2, L_2)$ and the bounding surface in dimension k creates a bounding hypersphere of the form

$$\sum_{i=1, i \neq k}^{d} (x_i - x_{2,i})^2 = r_2^2 - (x_k - x_{2,k})^2 \ . \tag{12}$$

which yields an effective radius $\widetilde{r_2} = \sqrt{r_2^2 - (z_k - x_{2,k})^2}$.

By denoting the projection of $\mathbf{x}_1$ and $\mathbf{x}_2$ onto the bounding hyperplane in the k-th dimension as $\widetilde{\mathbf{x}_1}$ and $\widetilde{\mathbf{x}_2}$, then the distance from $\mathbf{x}_1$ to Equation (12) must take the form

$$\begin{aligned}\hat{r_k} &= \sqrt{\|\widetilde{\mathbf{x}_1} - \mathbf{x}_1\|^2 + \left(\widetilde{r_2} - \|\widetilde{\mathbf{x}_2} - \widetilde{\mathbf{x}_1}\|\right)^2} \\ &= \sqrt{(z_k - x_{1,k})^2 + \left(\sqrt{r_2^2 - (z_k - x_{2,k})^2} - \sqrt{\|\mathbf{x}_2 - \mathbf{x}_1\|^2 - (x_{2,k} - x_{1,k})^2}\right)^2} \ .\end{aligned} \tag{13}$$

By imposing that $z_k = 1$ when the k-th component of $\mathbf{x}_1$ is greater than 0.5, and 0 otherwise, it follows that $\max_{k \in \{0,1,\ldots,d\}} z_k$ must be an improved radii of certification. □

3.4 Algorithms

To demonstrate how the above certification approaches can be applied in practice, Algorithm 1 demonstrates the application of Equation (4) through a simple, gradient based solver. Such a solver is

Algorithm 1 Single Bubble Loop.

Input: data $\mathbf{x}$, samples N, iterations M, true-label i, cutoff c, multiplicative Lagrange multiplier scaling factor λ, penalty parameter $\Gamma \gg 1$
for 1 **to** M **do**
  $\widehat{E}_0, \widehat{E}_1, L_2, j = \text{Model}\left(\mathbf{x}' + \mathcal{N}\left(0, \sigma^2 \mathcal{I}_N, N\right)\right)$ {Details of this can be found in Algorithm 2}
  if $j = i$ **then**
    $r' = L_2 - \|\mathbf{x}' - \mathbf{x}\|$
    if $r' > r'_o$ **then**
      $r'_o = r'$
    end if
    $\mathbf{x}' = \mathbf{x}' + \gamma \frac{\nabla_{\mathbf{x}'} r'}{\|\nabla_{\mathbf{x}'} r'\|}$ {Step size γ determined by Barzilai-Borwein [1].}
  else
    $\mathbf{x}' = \mathbf{x}' - \gamma \frac{\mathbf{x}' - \mathbf{x}}{\|\mathbf{x}' - \mathbf{x}\|}$
  end if
end for

highly applicable for solving such problems, due to the inherent smoothing nature of randomised smoothing being applicable both to the function space and its derivatives. To implement the multiple transitivity based approach of Section 3.2, Algorithm 1 can trivially be adapted to evaluate derivatives with respect to Equation (7). The boundary treatment of Section 3.3 does not require any additional calculations, but instead is simply the result of applying Equation (11) to the output of Algorithm 1.

4 Extracting gradient information from non-differentiable functions

Implementing the aformentioned process requires the ability to evaluate the gradient of the class expectations. This is problematic, as each class expectation is described in terms of a finite sum of non-differentiable indicator functions, as is seen in Equation (2). Within this work we have implemented two mechanisms to circumvent these conditions. The first substitutes the $\arg\max$ operation with a Gumbel-Softmax [11]. In doing so, the class expectations are rendered differentiable.

The second approach involves recasting the Monte-Carlo estimators as integrals of the form

$$E[\arg\max f_\theta(\mathbf{x}) = k] = \int \mathbb{1}[\arg\max f_\theta(\mathbf{y}) = i]\omega(\mathbf{y}, \mathbf{x}) d\mathbf{y} \quad , \tag{14}$$

where $\omega(\mathbf{y}, \mathbf{x})$ is the multivariate-Normal probability distribution centred around $\mathbf{x}$. From this formalism, the symmetry of the underlying space allows gradient free derivatives of $\nabla_{\mathbf{x}} E$ to be calculated by way of

$$\begin{aligned}
\nabla_{\mathbf{x}} E[\arg\max f_\theta(\mathbf{x}) = k] &= \nabla_{\mathbf{x}} \int \mathbb{1}[\arg\max f_\theta(\mathbf{y}) = k]\omega(\mathbf{y}, \mathbf{x}) d\mathbf{y} \\
&\approx \int \mathbb{1}[\arg\max f_\theta(\mathbf{y}) = i]\nabla_{\mathbf{x}}\omega(\mathbf{y}, \mathbf{x}) d\mathbf{y} \\
&= \frac{1}{\sigma^2} \int \mathbb{1}[\arg\max f_\theta(\mathbf{y}) = k](\mathbf{y} - \mathbf{x})\omega(\mathbf{y}, \mathbf{x}) d\mathbf{y} \\
&\approx \frac{1}{N\sigma^2} \sum_{i=1}^{N} \mathbb{1}[\arg\max f_\theta(\mathbf{x}_i) = k](\mathbf{x}_i - \mathbf{x}) \\
&\qquad \text{where } \mathbf{x}_i = \mathbf{x} + \mathcal{N}(0, \sigma^2) \quad .
\end{aligned} \tag{15}$$

While this derivation is novel, the resultant gradient operator reflects those seen in prior works [23]. It is important to note that such a sampling process is inherently noisy, and it has previously suggested that the underlying uncertainty scales with the input dimensionality [20].

To explore the relative performance of these two approaches, both of these approaches will be tested in the following section, with the Gumbel-Softmax based approach labelled as 'Full', and the above gradients as 'Approximate', which will be used to consider the transitivity approach of Section 3.1

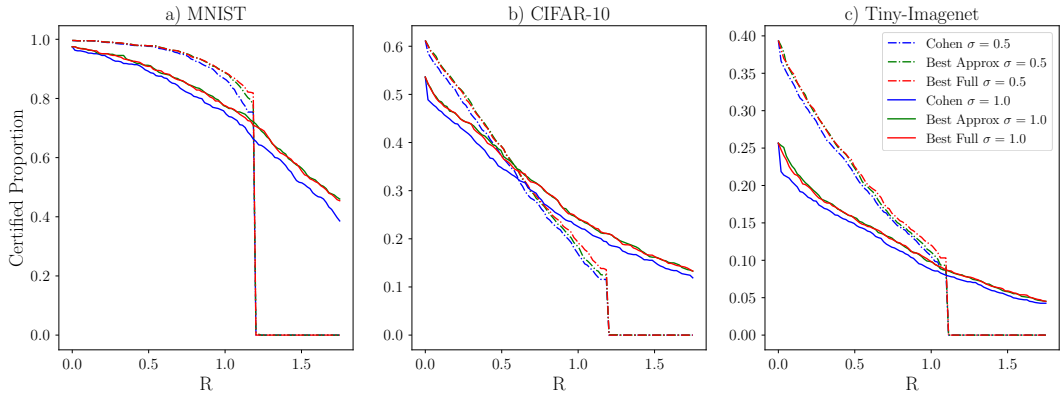


Figure 2: The Certified Proportion: the proportion of samples correctly predicted and with a certified radius greater than R. Blue represents Cohen, while Green and Red respectively represent the best sub-variant utilising either the Approximate or Full derivative approaches. Dashed lines for $\sigma = 0.5$, solid lines for $\sigma = 1.0$.

under the label 'Single', and the boundary consideration of Section 3.3 as 'Boundary'. For the case of the double transitivity of Section 3.2 our experiments suggest that uncertainty in the analytic derivatives produces deleterious results. As such derivatives for the multiple transitivity approach are exclusively considered through autograd for both the Full and Approximate methods.

5 Experiments

Configuration To evaluate the performance of our proposed certification improvements, we considered the certified radius produced for MNIST [16], CIFAR-10 [14], and Tiny-Imagenet [12], the latter of these is a 200-class variant of Imagenet [31] which downsamples images to $3 \times 60 \times 60$. All datasets were modelled using the Resnet18 architecture in PyTorch [22], with Tiny-Imagenet also utilising 2D adaptive average pooling. For both MNIST and CIFAR-10, our experimentation utilised a single NVIDIA P100 GPU core with 12 GB of GPU RAM, with expectations estimated over 1500 samples. Training employed Cross Entropy loss with a batch size of 128 over 50 epochs. Each epoch involved every training example was perturbed with a single perturbation drawn from $\mathcal{N}(0, \sigma^2)$, which was added prior to normalisation. Parameter optimisation was performed with Adam [13], with the learning rate set as 0.001. Tiny-Imagenet training and evaluation utilised 3 P100 GPU's and utilised 1000 samples. Training occurred using SGD over 80 epochs, with a starting learning rate of 0.1, decreasing by a factor of 10 after 30 and 60 epochs, and momentum set to 0.9.

The full code to implement our experiments can be found at *anonymised link*.

Certified accuracy To explore the performance advantage provided by our technique, we begin by considering the performance of Cohen *et al.* against the best of our approaches using both the approximate and full derivatives, as seen in Figure 2. While there are only minor differences between the two derivative approaches, there are clear regions of out performance relative to Cohen across all tested datasets. The proportion of this increase appears to be tied to the semantic complexity of the underlying dataset, with decreases in predictive accuracy (visible at $R = 0$) appearing to elicit greater percentage changes in the achieved certification radii, as evidenced by the results for Tiny-Imagenet.

Semantic complexity drives a decrease in the overall confidence of models, inducing a decrease in the separation between the highest class expectations E_0 and E_1. While this process shrinks the achievable certification radii, a higher E_1 provides more information to inform our gradient based search process, allowing for the identification of larger certifications. This property would suggest that as R is increased, the difference in the certifications between our techniques and that of Cohen *et al.* would decrease, as smaller values of E_1 provide less information for the search. However, the very opposite appears to occur, a property that is particularly visible in the MNIST experiments of Figure 2 at $\sigma = 0.5$. This appears to be a product of the singularities in the derivatives of $\Phi^{-1}[E]$ as $E \to \{0, 1\}$, which counteracts the decreased information provided by the second highest class.

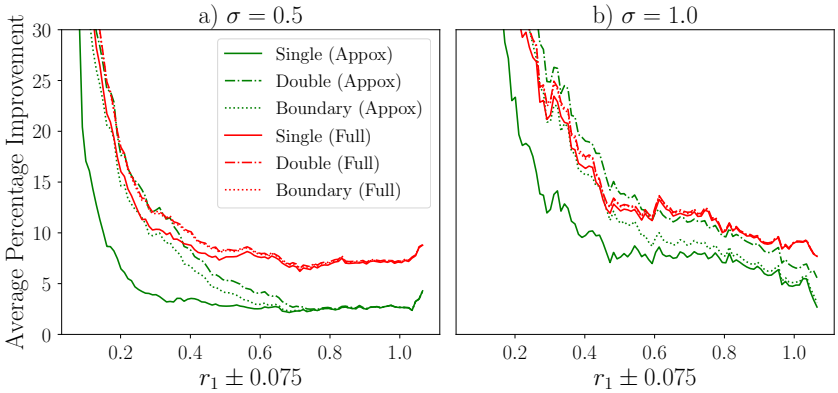


Figure 3: Average percentage improvement of Tiny-Imagenet instances relative to Cohen *et al.*. Equivalent figures for MNIST and CIFAR-10 are found in Appendix A.2

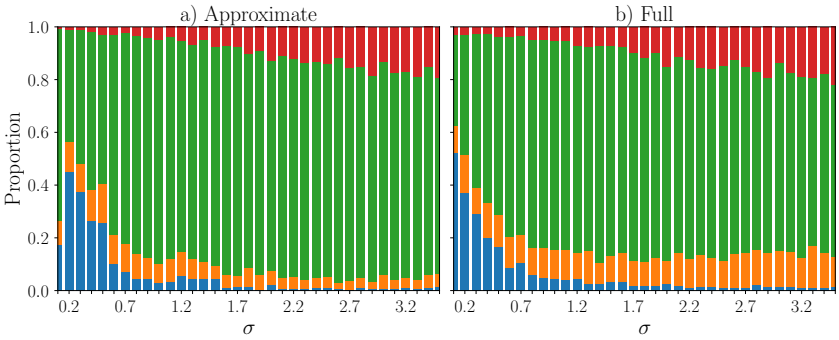


Figure 4: Proportion of correctly predicted instances for which each approach yields the highest certification across σ for Tiny-Imagenet. Red represents the proportion for which the boundary treatment produces the largest certification, with Green, Orange, and Blue representing the same for Double transitivity, Single transitivity, or Cohen *et al.*. Variants for MNIST and CIFAR-10 can be found in Appendix A.2

Rather than strictly considering the best performing of the Approximate and Full solvers, we can also delve into the relative performance of the underlying solvers. Notably there is a moderate increase in the average percentage improvement of Figure 3 between the $\sigma = 0.5$ and $\sigma = 1.0$ cases. This would appear to belie our previous statement regarding larger certifications yielding smaller improvements, due to the asymmetry of class information. However, an equivalent radii for $\sigma = 1.0$ has significantly more information about the second class (due to the multiplicative influence of σ), allowing for greater improvements from our gradient based search. That there is a clear demarcation between the Full and Approximate solver variants reflects the uncertainties introduced by Equation (15). However, that the approximate technique is still applicable verifies the utility of our approach even when the final layer is a strict $\arg\max$ function, rather than our Gumbel-Softmax based approach.

The trends in performance across σ are further explored in Figures 4 and 5, the latter of which demonstrates that the median performance improvement increases quasi-linearly with σ. This is driven by both an increase in the performance of the certifications themselves, and in the number of instances able to be certified in an improved fashion. This later property stems from the smoothing influence of σ, with larger levels of added noise inducing decreases in the difference between the highest class expectations, improving the ability for our search based mechanisms to identify performance improvements. Here increases in the performance of the boundary treatment are correlated with larger radii of certification, due to the multiplicative influence of σ upon Equation (3).

Numerical performance The numerical optimisation process at the core of our certification process inherently induces increases in the computational cost of finding these improved certifications, as

Table 1: Average wall clock time (in seconds) for each computational technique.

Dataset		Approx.			Full		
	Cohen	Single	Boundary	Double	Single	Boundary	Double
M	0.08	0.66	0.66	2.94	2.17	2.17	3.97
C-10	0.08	0.76	0.76	3.83	2.08	2.08	4.62
T-I	0.12	1.13	1.13	5.61	2.87	2.86	7.95

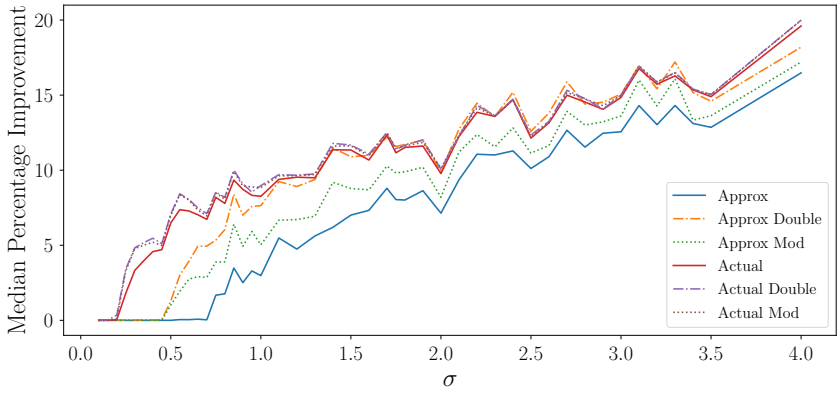


Figure 5: Median percentage improvement relative to Cohen *et al.* for Tiny-Imagenet across σ.

shown in Table 1. While analytically approximating the derivatives for the first eccentric hypersphere yields lower Certified Proportions, the fact that the corresponding computational cost decreases by a factor of more than 4 emphasises the value of this approach. Interestingly, while the Approximate method does also utilise auto-differentiation for the Double variant, the increase in computational cost from the Single to Double variants is significantly higher than for the Actual approach. This is surprising, as the Approx variant derives smaller values of $\mathbf{r}'$, which should in turn lead to a smaller, easier to navigate search space for $\mathbf{r}''$. Instead it counter-intuitively appears that a smaller $\mathbf{r}'$ induces a search space for $\mathbf{r}''$ which is less convex, and more difficult to converge upon.

Limitations While our focus was on L_2-norm bounded perturbations was driven by the Cohen *et al.* guarantee of best possible certifications, the the principles of our Geometrically Informed Certified Robustness are extensible to other spaces. However, as with prior certification approaches we have only considered bounded perturbations without extending to translations, rotations, or other functional attacks.

Finally we note that the generation of guarantees of certified robustness have the potential to counter beneficial applications of adversarial attacks, such as those used to promote stylometric privacy [3]. However, we believe that these drawbacks are significantly outweighed by the advantage of being able to understand the behaviour of models for which incentives exist for adversarial behaviour. However, achieving this does increase the computational cost and associated energy use, however we believe that the increased computational cost is not intractable, especially when these techniques are incorporated into human-in-the-loop certification systems.

6 Conclusions

This work has presented mechanisms that can be exploited to improve achievable levels of certified robustness, based upon exploiting underlying geometric properties of robust neural networks. In doing so our Geometrically Informed Certified Robustness approach has been able to generate certifications that exceed prior guarantees by on average more than 5% at $\sigma = 0.5$, with the percentage increase improving quasi-linearly with σ. Being able to improve upon the size of these guarantees inherently increases the cost of constructing adversarial attacks against systems leveraging machine learning, especially in the case where the attacker has no ability to observe the size of the robustness certificate.

References

[1] Jonathan Barzilai and Jonathan M Borwein. Two-Point Step Size Gradient Methods. *IMA Journal of Numerical Analysis*, 8(1):141–148, 1988.

[2] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion Attacks Against Machine Learning at Test Time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 387–402. Springer, 2013.

[3] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22, 2012.

[4] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[5] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

[6] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018.

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography Conference*, pages 265–284. Springer, 2006.

[8] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the Rules Of The Game for Adversarial Example Research. *arXiv preprint arXiv:1807.06732*, 2018.

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. *arXiv preprint arXiv:1412.6572*, 2014.

[10] Leo A Goodman. On Simultaneous Confidence Intervals for Multinomial Proportions. *Technometrics*, 7(2):247–254, 1965.

[11] Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[12] Justin Johnson, Fei-Fei Li, and Andrej Karpathy. Tiny ImageNet Visual Recognition Challenge, Accessed 2022-01-10.

[13] Diederik P Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

[15] Cassidy Laidlaw and Soheil Feizi. Functional Adversarial Attacks. *arXiv preprint arXiv:1906.00001*, 2019.

[16] 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.

[17] 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.

[18] 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.

[19] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable Abstract Interpretation for Provably Robust Neural Networks. In *International Conference on Machine Learning*, pages 3578–3586. PMLR, 2018.

[20] Jeet Mohapatra, Ching-Yun Ko, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Higher-Order Certification for Randomized Smoothing. *Advances in Neural Information Processing Systems*, 33:4501–4511, 2020.

[21] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks against Machine Learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519, 2017.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[23] 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*, 2019.

[24] Hadi Salman, Greg Yang, Huan Zhang, Cho-Jui Hsieh, and Pengchuan Zhang. A Convex Relaxation Barrier to Tight Robustness Verification of Neural Networks. In *Advances in Neural Information Processing Systems*, 2019.

[25] 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):1–30, 2019.

[26] Cristina P Sison and Joseph Glaz. Simultaneous Confidence Intervals and Sample Size Determination for Multinomial Proportions. *Journal of the American Statistical Association*, 90(429):366–369, 1995.

[27] Shixin Tian, Guolei Yang, and Ying Cai. Detecting Adversarial Examples through Image Transformation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-CROWN: Efficient Bound Propagation with Per-Neuron Split Constraints for Neural Network Robustness Verification. *Advances in Neural Information Processing Systems*, 34, 2021.

[29] Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards Fast Computation of Certified Robustness for ReLU Networks. In *International Conference on Machine Learning*, pages 5276–5285. PMLR, 2018.

[30] Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic Perturbation Analysis for Scalable Certified Robustness and Beyond. *Advances in Neural Information Processing Systems*, 33, 2020.

[31] Kaiyu Yang, Jacqueline Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A Study of Face Obfuscation in Imagenet. *arXiv preprint arXiv:2103.06191*, 2021.

[32] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient Neural Network Robustness Certification with General Activation Functions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4939–4948. Curran Associates, Inc., 2018.

[33] Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient Neural Network Robustness Certification with General Activation Functions. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] In the Limitations section of Section 5
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] In the Limitations section of Section 5
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] This has been included in the supplementary materials, with license details included therein.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] In Section 5
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] As all experimental results are based upon metrics calculated over more than 1000 samples, the size of the underlying error bars are small enough to be negligible.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In Section 5
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes] In Appendix A.1.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Algorithm 2 Class prediction and certification, as required for Algorithm 1

1: **Input:** Perturbed data $\mathbf{x}'$, samples N, level of added noise σ
2: $\mathbf{y} = \mathbf{0}$
3: **for** i = 1:N **do**
4: $\quad y_j = y_j + 1$ if $GS\left(f_\theta\left(\mathbf{x}' + \mathcal{N}(0, \boldsymbol{\sigma}^2)\right)\right) = j$ {Here GS is the Gumbel-Softmax}
5: **end for**
6: $\mathbf{y} = \frac{1}{N}\mathbf{y}$
7: $z_0, z_1 = \mathrm{topk}(\mathbf{y}, k = 2)$ {topk is used as it is differentiable, $z_0 > z_1$}
8: $\check{E}_0, \widehat{E}_1 = \mathrm{lowerbound}(\mathbf{y}, z_0), \mathrm{upperbound}(\mathbf{y}, z_)1$ {Calculated by way of Goodman *et al.* [10]}
9: $R = \frac{\sigma}{2}\left(\Phi^{-1}(\check{E}_0) - \Phi^{-1}(\widehat{E}_1)\right)$
10: **return** $\check{E}_0, \widehat{E}_1, R, j$

A Appendix

A.1 Algorithmic details

Algorithm 2 supports Algorithm 1 by demonstrating how the class prediction and expectations are calculated. Of note are two minor changes from prior implementations of this certification regime. The first is the addition of the Gumbel-Softmax on line 4, although this step is only required for the 'Full' derivative approach. In contrast th 'Approximate' techniques able to circumvent this limitation and can be applied directly to the case where the class election is determined by an $\arg\max$.

The second difference to prior works is the calculation of the lower and upper bounds on z_0 and z_1 on line 8. Our initial testing revealed that when we employed either Sison-Glaz [26] or Goodman *et al.* [10] to estimate the multivariate class uncertainties, some Tiny-Imagenet samples devoted more than 95% of their computational time of the process to evaluating the confidence intervals, significantly outweighing even the costly process of model sampling. Further investigation revealed that this was occurring when there were a significant number of classes reporting counts of approximately 0, the likelihood for which was higher in Tiny-Imagenet due to the increased class count relative to MNIST and CIFAR-10. To resolve this, we coalesced all classes where $y_j < 5$ into one single meta-class with an associated class-count $y' = \max(5, \sum_{y_j<5} y_j$, which conforms with the requirements of Goodman *et al.* [10] that all class counts must be greater than 5. Our testing demonstrated that while this process slightly decreased the resulting radii of certification (due to small changes in z_0 and z_1), the associated decrease in computational time was significant enough to justify this modification.

We also note that all the experiments contained within this work have been conducted against publically releaed datasets with established licenses. MNIST exists under a GNU v3.0 license; CIFAR-10 employs a MIT license; and Imagenet employs a BSD $3-$Clause license.

A.2 Relative performance for MNIST and CIFAR-10

While Figure 2 presents the best performing certified proportion, it is important to understand the relative performance of the Single Transitivity, Double Transitivity, and Boundary treatments, in a similar fashion to Figures 3 and 4. In the case of MNIST, while the percentage increases exhibited in Figure 6 as $r_1 \to 1$ are broadly similar to their Tiny-Imagenet counterpart for the Approximate solver, the difference between those results and the Full derivative treatment is significantly smaller, especially at $\sigma = 1$. This may, in part, be driven by the 1500 samples employed when using MNIST and CIFAR-10, in contrast to 1000 for Tiny-Imagenet, which should decrease the uncertainty of the gradient estimation steps.

However, the fact that this decreased difference in performance holds for CIFAR-10 at $\sigma = 1.0$ but not $\sigma = 0.5$ suggests that the performance difference between the techniques is also dependent upon the semantic complexity of the prediction task. While CIFAR-10 is a more complex predictive environment than MNIST, which should increase the complexity of the gradient based search routine employed within this work, the increased level of noise at $\sigma = 1.0$ has a smoothing influence that decreases the complexity of the search task, and it would appear that this is the primary driver of the

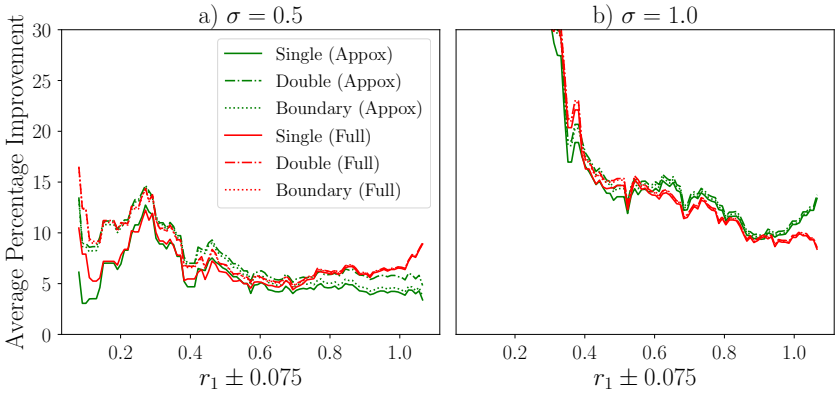


Figure 6: Average percentage improvement of MNIST instances relative to Cohen *et al.*. Equivalent figures for Tiny-Imagenet can be seen in Figure 3

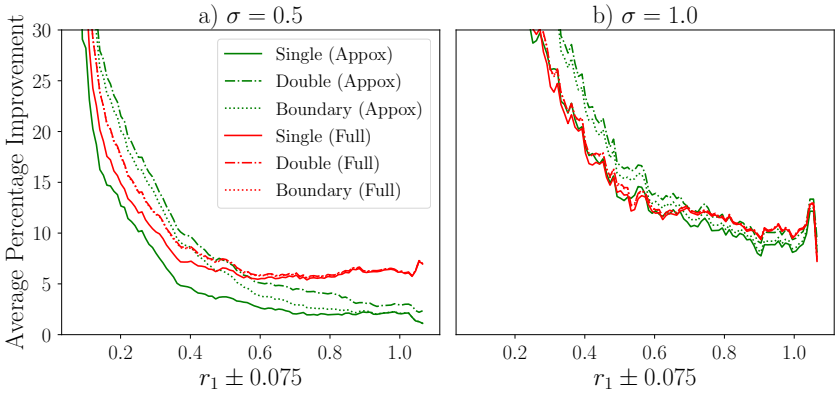


Figure 7: Average percentage improvement of CIFAR-10 instances relative to Cohen *et al.*. Equivalent figures for Tiny-Imagenet can be seen in Figure 3

relative under performance of the Approximate derivatives in both Tiny-Imagenet and CIFAR-10 when $\sigma = 0.5$.

When considering the median percentage improvement (relative to Cohen *et al.*) of these techniques, MNIST again reveals interesting properties when we consider Figure 10. When compared to CIFAR-10 and Tiny-Imagenet (in Figures 11 and 5) it becomes apparent that the Approximate approach only produces consistently larger certifications in MNIST. Given the increased uncertainty in the derivatives calculated by the Approximate technique, this would suggest that the approximate solver may be improved by considering common improvements to gradient descent methods like momentum or the addition of calibrated noise.

While MNIST may be the simplest of all the prediction tasks, Figure 8 demonstrates that at low $\sigma < 0.5$ the majority of samples cannot be improved upon by any of the certification enhancements developed within this paper. Given that this does not hold for CIFAR-10 nor Tiny-Imagenet (in Figure 9 and 4 respectively) this would suggest that the potential for Cohen *et al.* to be improved upon in low semantic complexity datasets is smaller. That this behaviour is predominantly seen for small σ also suggests that our initial step size may be too large in these particular cases.

A.3 Influence of the starting step size

The one heretofore un-considered feature is the influence of the initial step-size γ within Algorithm 1. As is shown in Figures 12 and 13, while the Full solver only exhibits sensitivity to γ when $\gamma > 0.1$, the approximate solvers are far more sensitive, with deleterious performance being observed for $\gamma > 0.01$

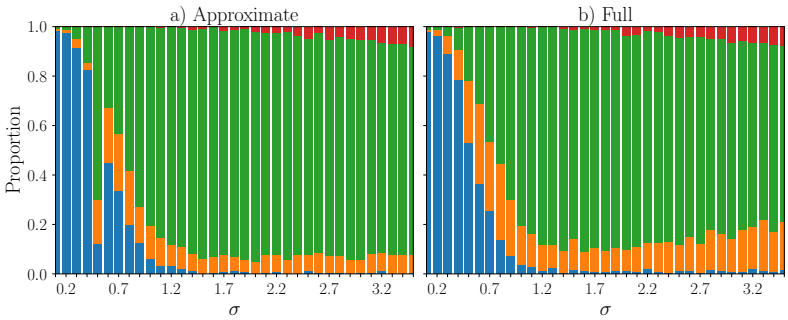


Figure 8: Proportion of correctly predicted instances for which each approach yields the highest certification across σ for MNIST. Red represents the proportion for which the boundary treatment produces the largest certification, with Green, Orange, and Blue representing the same for Double transitivity, Single transitivity, or Cohen *et al.*.

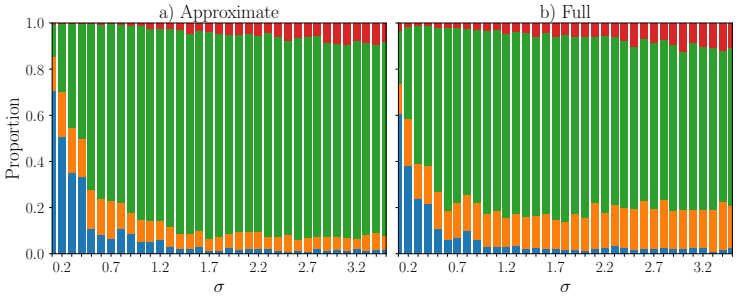


Figure 9: Proportion of correctly predicted instances for which each approach yields the highest certification across σ for CIFAR-10. Red represents the proportion for which the boundary treatment produces the largest certification, with Green, Orange, and Blue representing the same for Double transitivity, Single transitivity, or Cohen *et al.*.

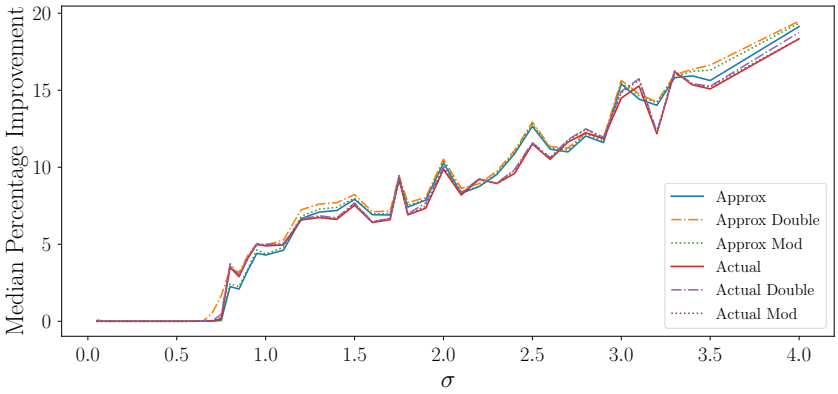


Figure 10: Median percentage improvement relative to Cohen *et al.* for MNIST across σ.

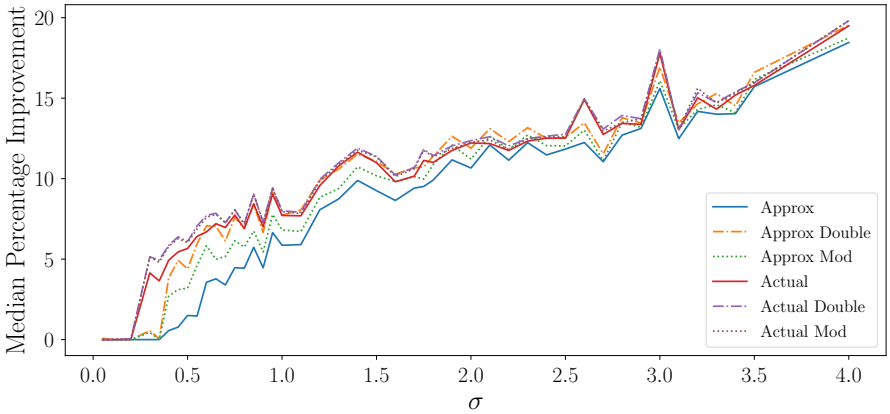


Figure 11: Median percentage improvement relative to Cohen *et al.* for CIFAR-10 across σ.

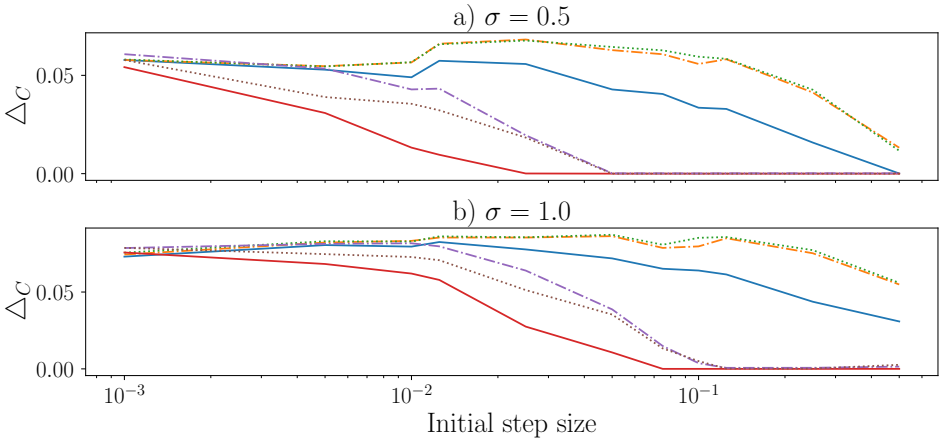


Figure 12: Average delta to Cohen ($\triangle_C$ for CIFAR-10 as a function of the initial step size γ. Here Blue, Green, and Yellow represent the Full Single transitivity, Double transitivity, and Boundary treatments; with Blue, Purple, and Brown representing the same for the Approximate solver.

in CIFAR-10, and even earlier for Tiny-Imagenet. This is likely due to the added uncertainty in the Approximate derivatives leading to convergence upon local sub-optima in the more semantically complex datasets. Based upon these results, the starting step size was uniformly set to 0.01 for all experiments.

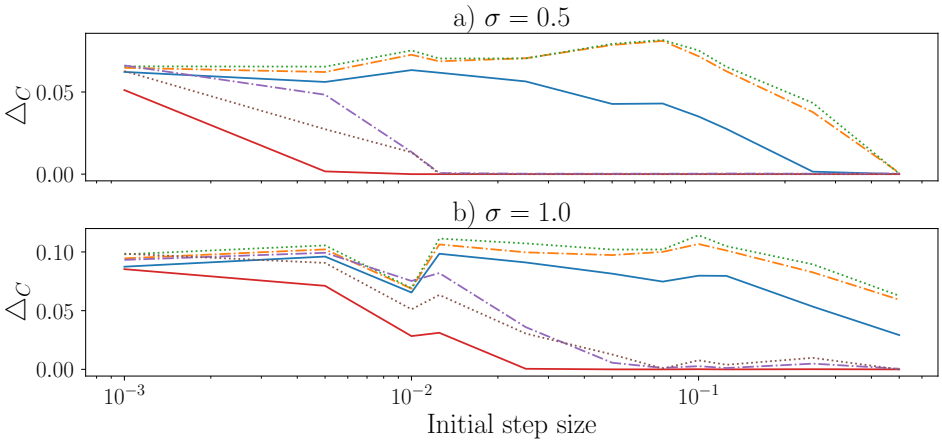


Figure 13: Average delta to Cohen ($\triangle_C$ for Tiny-Imagenet as a function of the initial step size γ. Here Blue, Green, and Yellow represent the Full Single transitivity, Double transitivity, and Boundary treatments; with Blue, Purple, and Brown representing the same for the Approximate solver.