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

Andrew C. Cullen\textsuperscript{1*} Paul Montague\textsuperscript{2} Shijie Liu\textsuperscript{1}
Sarah M. Erfani\textsuperscript{1} Benjamin I.P. Rubinstein\textsuperscript{1}
\textsuperscript{1}School of Computing and Information Systems, University of Melbourne, Parkville, Australia
\textsuperscript{2}Defence Science and Technology Group, Adelaide, Australia
andrew.cullen@unimelb.edu.au

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. By considering the smallest distance to points on the boundary of a set of certifications this approach improves certifications for more than 80\% of Tiny-Imagenet instances, yielding an on average 5\% increase in the associated certification. When incorporating training time processes that enhance the certified radius, our technique shows even more promising results, with a uniform 4 percentage point increase in the achieved certified radius.

1 Introduction

Learned models, including neural networks, are well known to be susceptible having the output changed by crafted perturbations to an input, that preserve the inputs semantic properties \cite{goodfellow2015explaining}. 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 \cite{lecun1998gradient}.

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 \cite{carlini2017towards}, and can be used to target class changes, to attack all classes \cite{madry2017towards}, or even introduce backdoors into model behaviour \cite{moosavi2016deepfool}. 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.

To circumvent these limitations, certified guarantees of adversarial robustness can be constructed to identify class-constant regions around an input instance, that guarantee that all instances within a norm-bounded distance (typically $L_2$) are not adversarial examples. Certifications based on randomised smoothing of classifiers around an input point are in a sense optimal \cite{sagawa2018adversarial}: based only on the prediction
class scores at the input point, no better 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 radius large enough to completely enclose that of a sample point, improving the radius of certification. This process can be extended to 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 produces certifications exceeding those of the hitherto best-case guaranteed approach of Cohen *et al.* (2019) [8].

2 Background and literature review

**Bounding mechanisms** Conservative bounds upon the impact of norm-bounded perturbations can be constructed by way of either Interval Bound Propagation (IBP) which propagates interval bounds through the model; or Convex Relaxation, which utilise linear relaxation to construct bounding output polytopes over input bounded perturbations [34, 26, 41, 45, 46, 37, 28], in a manner that generally provides tighter bounds than IBP [25]. In contrast to Randomised Smoothing, bounding mechanisms employ augmented loss functions during training, which promote tight output bounds [42] at the cost of decreased applicability. Moreover they both exhibit a time and memory complexity that makes them infeasible for complex model architectures or high-dimensional data [40, 6, 21].

**Randomised smoothing** Outside of bounding mechanisms, another common framework for developing certifications leverages randomised smoothing [20], in which noise is applied to input instances to smooth model predictions, subject to a sampling distribution that is tied to the $L_p$-norm of adversarial perturbations being certified against. In contrast to other robustness mechanisms, this application of the noise is the only architectural change that is required to achieve certification. In the case of $L_2$-norm bounded attacks, Gaussian sampling of the form

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

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

$$E_Y[\arg \max f_\theta(x + Y) = i] \approx \frac{1}{N} \sum_{j=1}^{N} \mathbb{I}[\arg \max f_\theta(x_j) = i].$$

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 $r$, with recent work also considering rotational and/or translational semantic attacks [23, 7]. $L_p$-norm certifications were first demonstrated by way of differential privacy [20, 11], with more recent approaches employing Rényi divergence [22], and parametrising worst-case behaviours [8, 33]. By considering the worst-case $L_2$-perturbations, Cohen *et al.* (2019) purports that the largest achievable pointwise certification is

$$r = \frac{\sigma}{2} \left( \Phi^{-1}(E_0|x) - \Phi^{-1}(E_1|x) \right).$$

Here $\{E_0, E_1\}$ are the two largest class expectations (as per Equation (2)), $\sigma$ is the noise, and $\Phi^{-1}$ is the inverse normal CDF, or Gaussian quantile function.

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 [12]. Thus, constructing larger certifications in such a context is inherently valuable.
While it is provably true that Equation (3) is the largest achievable certification for any point $x$ with Figure 1a providing a two-dimensional exemplar. Crucially, the above formalism does not require obtaining a global optima, as any $r' > r_1$ yields an improved certification at $x_1$.

This specific $L_p$ space is of interest due to both its viability as a defence model, and the provable guarantee that Cohen et al. produces the largest possible certification for any instance [8]. 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 $x$, it is possible to exploit the behaviour of points in the neighbourhood of $x$ in order to enhance certifiable radius. To achieve this, consider the case of a second point $x'$, that exists within the certifiable radius of $x$. 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).** A radius of certification $r_i$ about $x_i$ can be calculated by evaluating Equation (3) at $x_i$. This certification guarantees that no point $x : \|x - x_i\|_p \leq r_i$ can induce a change in the predicted class. That this shape is a $d$-dimensional hypersphere for input data $x \in \mathbb{R}^d$ allows us to introduce the notational shorthand

$$B_P(x_i, r_i) = \{x \in \mathbb{R}^d | \|x - x_i\|_p \leq r_i\} \quad S_i = \{x \in \mathbb{R}^d | \|x - x_i\|_p = r_i\} \quad (4)$$

to represent the region covered by the hypersphere and its surface. It follows from this definition that if $B_P(x_1, r_1) \cap B_P(x_2, r_2) \neq \emptyset$, which ensures that the class predictions at $x_1$ and $x_2$ match, then the region of certification about $x_1$ can be expressed as $B_P(x_1, r_1) \cup B_P(x_2, r_2)$.

However typically we are concerned not with the size of the region of classification invariance, but rather the distance to the nearest adversarial example. If it is possible to find some $x'$ such that its region of certification completely encircles that of the certification at $x$, the following definition demonstrates that the certification radius about $x$ can be increased.

**Lemma 3.2 (Set Unions Certified Radius).** If $x_1$ and $x_2$ have the same class associated with them and $B_P(x_1, r_1) \subset B_P(x_2, r_2)$, then the nearest possible adversarial example—and thus, the certifiable radius—exists at a distance $r' \geq r$ from $x_1$, where

$$r' = r_2 - \|x_2 - x_1\|_p \quad (5)$$

**Proof.** The closest point on the surface of $B_P(x_2, r_2)$ to $x_1$ must exist on the vector between $x_1$ and $x_2$. Thus $r' = \min (r_2 \pm \|x_2 - x_1\|_p)$, which takes the form of Equation (5). \qed

As such, we can recast the task of constructing a certification from being a strictly analytic function to the nonlinear optimisation problem in terms of a second ball with certified radius $r_2$ centred at $x_2$.

$$r' = \max_{x_2 \in [0,1]^d} r_2 - \|x_2 - x_1\|_p \quad (6)$$

with Figure 1b providing a two-dimensional exemplar. Crucially, the above formalism does not require obtaining a global optima, as any $r' > r_1$ yields an improved certification at $x_1$. 

---

Figure 1: Transitive certification exemplars. The Green, Red, and Black circles represent hyperspheres of radius $r_i$ (by Equation 3) about points $x_i, \forall i \in \{1, 2, 3\}$. The resulting certifications $r', r''$, and $\hat{r}$ are described within Equations 5, 12, and 16. The Black line represents the domain boundary.
3.2 Multiple transitivity

To further enhance our ability to certify, let us consider the set of points and their associated certifica-
tions \( \{x_1, r_1, x_2, r_2, \ldots, x_n, r_n\} \). If the union of \( B_P = \bigcup_{i=1,\ldots,n} B_P(x_i, r_i) \) is simply-connected, then the certification across this set can be expressed as \( r^{(n)'} = \min_{x \in \partial B_P} \|x - x_1\|_p \), where \( \partial B_P \) is the boundary of \( B_P \). This can be further simplified by imposing that \( x_j \in B_P(x_2, r_2) \forall j > 2 \) and that \( B_P(x_j, r_j) \not\subseteq B_P(x_k, r_k) \forall (j > 2, k > 2) \) to ensure that hyperspheres exist near the boundary of \( S_2 \) and yielding a certification of

\[
r^{(n)'} = \min_{x \in S_n} \|x - x_1\| \quad \text{where} \quad S_n = S_2 \cap \left( \bigcup_{i=3}^{n} S_i \right) \quad \text{for} \quad n \geq 2 .
\]

Here \( S_n \) is a \((d-1)\)-dimensional manifold embedded in \( \mathcal{R}^d \).

**Lemma 3.3 (Optimal positioning of \( x_3 \) in the case of \( n = 3 \)).** Consider the addition of a new hypersphere at some point \( x_3 \) with associated radius \( r_3 \), which has an associated boundary \( S_3 \). If it is true that

\[
B_P(x_2, r_2) \not\subset B_P(x_2, r_2) \quad \text{and} \quad B_P(x_3, r_3) \cap B_P(x_2, r_2) \neq \emptyset \tag{8}
\]

\[
B_P(x_1, r_1) \not\subset B_P(x_1, r_1) \quad \forall \{x \in [0, 1]^d | x \neq x_3\} \quad \text{with} \quad r = \frac{d}{2} \left( \Phi^{-1}(E_0[x]) - \Phi^{-1}(E_1[x]) \right) , \tag{9}
\]

then the largest possible certification \( r'' \) by Equation\( 7 \) is achieved at

\[
x_3(s) = x_1 + sr' \frac{\|x_1 - x_2\|}{\|x_1 - x_2\|_2} \quad \text{for some} \quad s \in [0, 1] . \tag{10}
\]

**Proof.** The closest point to \( x_1 \) upon \( S_2 \) is located at

\[
\hat{x} = x_1 + r' \frac{x_1 - x_2}{\|x_1 - x_2\|_2} , \tag{11}
\]

where \( r' \) is defined by Equation\( 6 \). Thus any improved radius of certification \( r'' > r' \) is only achievable if \( B_P(x_3, r_3) \) satisfies \( \hat{x} \in B_P(x_3, r_3) \) and Equation\( 8 \). Then by symmetry, \( r'' \) is the maximally achievable radius of certification if Equation\( 9 \) hold and if \( x_3 \) is defined by Equation\( 10 \) \( \square \)

While finding some \( x_3 \) satisfying Equations\( 8 \) and \( 10 \) is trivial, proving Equation\( 9 \) would require an exhaustive search of the input space \([0, 1]^d \). However, even in the absence of such a search, Equation\( 10 \) still provides the framework for a simple search for \( x_3 \), which follows Figure\( 11 \).

**Lemma 3.4 (Certification from two eccentric hyperspheres).** If \( x_3 \) is defined by Equation\( 10 \) in a fashion that satisfies Equation\( 8 \) then an updated certification can be achieved in terms of some \( x_3(s) \) defined by Equation\( 10 \) by way of

\[
r'' = \max_{s \in [0, 1]} \frac{d_2(r_3^2 - d_3^2) + d_3(r_2^2 - d_2^2)}{d_2 + d_3} \quad \text{where}
\]

\[
d_2 = \|x_2 - x_1\| \quad d_3 = \|x_3(s) - x_1\| \quad r_3 = \frac{d}{2} \left( \Phi^{-1}(E_0[x_3(s)]) - \Phi^{-1}(E_1[x_3(s)]) \right) .
\]

If Equation\( 9 \) holds, then this is the largest achievable certification for \( n = 3 \).

**Proof.** By symmetry we can define the arbitrary rotational mapping \( f : \mathcal{R}^d \rightarrow \mathcal{R}^d \) from \( x_i \rightarrow y_i \) by way of \( y_i = f(x_i - x_1) \), subject to the condition

\[
y_{i,j} \neq 0 \quad \text{for} \quad i \in \{1, 2\} \quad \text{and} \quad \forall j \neq k, \quad \text{for some} \quad k \in \{1, 2, \ldots, d\} \tag{13}
\]

then the intersection of the hyperspheres centred about \( x_2 \) and \( x_3 \) occurs at

\[
r_2^2 - r_3^2 = \|y - y_3\|^2 - \|y - y_2\|^2 = 2y_{2,k}(d_2 + d_3) + d_3^2 - d_2^2
\]

\[
2y_{2,k} = \frac{r_2^2 - r_3^2 + d_2^2 - d_3^2}{d_2 + d_3} . \tag{14}
\]
This is a consequence of our mapping \( y = f(x - x_1) \) preserving distances under rotation, giving that \( y_{2,k} = \|x_2 - x_1\| = d_2 \), and with the equivalent also holding for \( y_{3,k} \).

As a consequence of our choice of coordinate system, it follows that \( r'' = \|y\| \) and
\[
(r'')^2 = \|y - y_0\|^2 + 2y_{2,k}d_2 - d_2^2
= r_2^2 - d_2^2 + \frac{d_2r''_2 - d_2r''_2^2 + d_2d_2^2 - d_2d_2^2}{d_2 + d_3}
\]
which is an equivalence to Equation (12). □

While Equation 7 holds for any \( n \geq 2 \), the certification radius beyond \( n = 3 \) cannot be enhanced by adding any one single additional hypersphere without contradicting Lemma 3.4. This is a result of \( S_3 \) being a \( (d - 1) \)-dimensional manifold in \( \mathbb{R}^d \), the entirety of which must be enclosed to improve the certification. An example of this can be seen with the two equidistant intersections between \( S_2 \) (in Red) and \( S_3 \) (in Black) in Figure 1b. While multiple spheres could be constructed to completely enclose \( S_3 \), the number required grows exponentially with \( d \) due to the sphere packing kissing number problem [9]. This growth in complexity makes adding additional spheres beyond \( n = 3 \) infeasible. Further details of this are contained within Appendix A.2.

### 3.3 Boundary treatments

Without loss of performance or accuracy, we can freely scale the inputs of neural networks such that \( x_1 \in [0, 1]^d \). However in the majority of cases a subset of \( B_P(x_1, L_1) \) will exist outside \( [0, 1]^d \). While this observation is trivially true, it has no influence on the radius of certification achieved by Equation 3 due to the symmetry of \( B_P(x_1, r) \). However, the asymmetric nature of \( B_P(x_2, r_2) \) about \( x_1 \) guarantees that if \( B_P(x_2, r_2) \) exceeds \( [0, 1]^d \), then the closest point to \( x_1 \) within the feasible domain must have an associated distance \( r > r' \), 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 \( B_P(x_2, L_2) \) as a bounding region about \( x_1 \), and the potential for a subset of \( B_P(x_2, L_2) \) to exist outside the feasible space for instances allows us to construct an updated region of certification where
\[
\hat{r} = \max_{k \in \{1, \ldots, d\}} \left[ r_2^2 - (x_{2,k} - z_k)^2 \geq 0 \right] \times
\sqrt{(z_k - x_{1,k})^2 + \left( r_2^2 - (z_k - x_{2,k})^2 - \sqrt{(x_2 - x_1)^2 - (x_{2,k} - x_{1,k})^2} \right)^2}
\]
where \( z_k = \begin{cases} 0 & \text{if } x_{1,k} < 0.5 \\ 1 & \text{if } x_{1,k} > 0.5 \end{cases} \),

where \( 1 \) 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 \( x_n = \{x_{n,1}, x_{n,2}, \ldots, x_{n,d}\} \). If we let \( x_k \in \{0, 1\} \), then the intersection between \( B_P(x_2, L_2) \) and the bounding surface in dimension \( k \) creates a bounding hypersphere of the form
\[
\sum_{i=1, i \neq k}^d (x_i - x_{2,i})^2 = r_2^2 - (x_k - x_{2,k})^2
\]
which yields an effective radius \( \hat{r}_2 = \sqrt{r_2^2 - (z_k - x_{2,k})^2} \).

By denoting the projection of \( x_1 \) and \( x_2 \) onto the bounding hyperplane in the \( k \)-th dimension as \( \hat{x}_1 \) and \( \hat{x}_2 \), then the distance from \( x_1 \) to Equation (17) must take the form
\[
\hat{r}_k = \sqrt{\|\hat{x}_1 - x_1\|^2 + (\hat{r}_2 - \|\hat{x}_2 - \hat{x}_1\|)^2}
= \sqrt{(z_k - x_{1,k})^2 + \left( r_2^2 - (z_k - x_{2,k})^2 - \sqrt{(x_2 - x_1)^2 - (x_{2,k} - x_{1,k})^2} \right)^2}
\]
By imposing that \( z_k = 1 \) when the \( k \)-th component of \( x_1 \) is greater than 0.5, and 0 otherwise, it follows that \( \max_{k \in \{0, 1, \ldots, d\}} z_k \) must be an improved radius of certification. □
Algorithm 1 Single Bubble Loop.

1: **Input:** data \(x\), samples \(N\), iterations \(M\), true-label \(i\)
2: for \(1\) to \(M\) do
3: \(\bar{E}_0, \bar{E}_1, L_2, j = \text{Algorithm}[2](x' + N(0, \sigma^2 I_N), N, \sigma)\)
4: if \(j = i\) then
5: \(r' = L_2 - \|x' - x\|\)
6: if \(r' > r'_o\) then
7: \(r'_o = r'\)
8: end if
9: \(x' = x' \pm \gamma \frac{\nabla x' r'}{\|\nabla x' r'\|}\) \{\(\gamma\) calculated by Barzilai-Borwein [1]. Positive branch is selected if \(j = i\), otherwise the negative branch brings \(x'\) towards the region in which \(j = i\).\}
10: end if
11: end for

3.4 Algorithms

To demonstrate how the above certification approaches can be applied in practice, Algorithm 1 demonstrates the application of Equation (5) through a simple, gradient based solver. Such a solver is 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 (12). The boundary treatment of Section 3.3 does not require any additional calculations, but instead is simply the result of applying Equation (16) to the output of Algorithm 1.

4 Extracting gradient information from non-differentiable functions

Implementing the aforementioned 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 \(\text{arg max}\) operation with a Gumbel-Softmax [15]. In doing so, the class expectations are rendered differentiable.

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

\[
E[\text{arg max}_{\theta} f_\theta(x) = k] = \int \mathbb{1}[\text{arg max}_{\theta} f_\theta(y) = k] \omega(y, x) dy ,
\]

where \(\omega(y, x)\) is the multivariate-Normal probability distribution centred around \(x\). This formalism, and the symmetry of the underlying space allows for the construction of undifferentiable gradients by

\[
\nabla x E[\text{arg max}_{\theta} f_\theta(x) = k] = \nabla x \int \mathbb{1}[\text{arg max}_{\theta} f_\theta(y) = k] \omega(y, x) dy
\]

\[
= \frac{1}{\sigma^2} \int \mathbb{1}[\text{arg max}_{\theta} f_\theta(y) = k] (y - x) \omega(y, x) dy
\]

\[
= \frac{1}{N \sigma^2} \sum_{i=1}^{N} \mathbb{1}[\text{arg max}_{\theta} f_\theta(x_i) = k] (x_i - x)
\]

where \(x_i = x + N(0, \sigma^2)\).

While this derivation is novel, the resultant gradient operator reflects those seen in prior works [33]. 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 [27].

The relative performance of these two approaches—respectively labelled ‘Approximate’ and ‘Full’ for the above approach and the Gumbel-Softmax approaches—will be tested in the following section. 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 [19], CIFAR-10 [18], and Tiny-Imagenet [16], the latter of these is a 200-class variant of Imagenet [43] which downsamples images to 3 × 60 × 60. All datasets were modelled using the Resnet18 architecture in PyTorch [32], 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 [17], 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 https://github.com/andrew-cullen/DoubleBubble.

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 certified radius, 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 certified radius, 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 samples with a larger \( R \) would exhibit a decreased difference between our techniques and that of Cohen et al. would decrease, as smaller values of \( E_1 \) provide less search information. However, it appears that singularities in the derivatives of \( \Phi^{-1}[E] \) as \( E \rightarrow \{0, 1\} \) counteract the decreased information provided by the second highest class, leading to the contradictory performance best observed in the MNIST experiments of Figure 2 at \( \sigma = 0.5 \).
Figure 3: Percentage improvement in the Certified Radius of Tiny-Imagenet instances relative to Cohen et al., for varying \( r \). This measure presents the median improvement over \([r-0.075, r+0.075]\). Equivalent figures for MNIST and CIFAR-10 are found in Appendix A.3.

Figure 4: Proportion of correctly predicted instances for which each approach yields the highest certification for Tiny-Imagenet. Red, Green, and Orange represent the boundary treatment, Double transitivity, and Single transitivity, with both ties Cohen et al. in Blue. MNIST and CIFAR-10 results can be seen in Appendix A.3.

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 certification for \( \sigma = 1.0 \) has significantly more information about the second class (due to the multiplicative influence of \( \sigma \)), 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 (20). That the approximate technique is still applicable verifies the utility of our approach even when the final layer is a strict \( \text{arg max} \) function, rather than a Gumbel-Softmax.

The trends in performance across \( \sigma \) are further explored in Figures 4 and 6, the latter of which demonstrates that the median performance improvement increases quasi-linearly with \( \sigma \). 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 \( \sigma \), 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 \( \sigma \) 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 shown in Table 1. While analytically approximating the derivatives for the first eccentric hypersphere yields a lower certified accuracy, 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
Table 1: Average wall clock time (in seconds) for each computational technique, for a single sample evaluated over 1000 draws under noise.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Approx.</th>
<th>Full</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cohen</td>
<td>Single boundary</td>
</tr>
<tr>
<td>M</td>
<td>0.08</td>
<td>0.66</td>
</tr>
<tr>
<td>C-10</td>
<td>0.08</td>
<td>0.76</td>
</tr>
<tr>
<td>T-I</td>
<td>0.12</td>
<td>1.13</td>
</tr>
<tr>
<td></td>
<td>Single</td>
<td>Boundary</td>
</tr>
<tr>
<td></td>
<td>2.17</td>
<td>2.17</td>
</tr>
<tr>
<td></td>
<td>2.08</td>
<td>2.08</td>
</tr>
<tr>
<td></td>
<td>2.87</td>
<td>2.86</td>
</tr>
</tbody>
</table>

Figure 5: Certified Accuracy comparing the Cohen certification (when trained incorporating MACER) for CIFAR-10, as well as the best sub-variant utilising either the Approximate or Full derivative approaches. Dashed lines for $\sigma = 0.5$, solid lines for $\sigma = 1.0$.

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 $r'$, which should in turn lead to a smaller, easier to navigate search space for $r''$. Instead it counter-intuitively appears that a smaller $r'$ induces a search space for $r''$ which is less convex, and more difficult to converge upon.

Alternative training routines Recent work has considered the potential for enhancing certified robustness by modifying the underlying training regime to incentivise maximising the expectation gap between classes [33]. One such approach is MACER [44], which augments the training time behaviour by considering not just the classification loss, but also the $\epsilon$-robustness loss, which reflects proportion of training samples with robustness above a threshold level. Such a training time modification can increase the average certified radius by $10\% - 20\%$, however doing so does increase the overall training cost by more than an order of magnitude.

When applying Geometrically-Informed Certified Robustness to models trained with MACER, Figure 5 demonstrates that our modifications yield an even more significant improvement than those observed in Figure 2. Under training with MACER, the best performing of our techniques yielded an approximately 4 percentage point increase in the average certification. From this it is clear that while MACER does enhance the certified radii at a sample point, it also induces enough smoothing in the neighbourhood of the sample point to allow transitivity to yield even more significant improvements than are present without MACER.

However, we must emphasise that while such training time modifications do produce consistently larger certifications, doing so requires significantly more computational resources, both in terms of training time and GPU memory, as compared to the more traditional certification training regime. We also emphasise that training with MACER requires a degree of re-engineering. In contrast the training mechanism used for the remainder of this work only requires the addition of noise to samples prior to being passed through the model, and thus imposes significantly fewer engineering considerations.

Limitations While the principles of our Geometrically Informed Certified Robustness are extensible to $L_p$ spaces, our experimental work has so far only considered $L_2$-norm bounded perturbations due to the guarantee of best possible certification in this space provided by Cohen et al.. Further
experimentation could also consider both these general spaces and a broader range of training methods, which have been shown to be able to tighten the achievable radius of certification \cite{24}.

We note that enhanced robustness certification have the potential to counter beneficial applications of adversarial attacks, such as those used to promote stylometric privacy \cite{4}. However, we believe this drawbacks is significantly outweighed by the potential for enhanced confidence about models for which adversarial incentives exist.

Finally, we also emphasise that our approach requires $M$ evaluations of the certified robustness, which each require $N$ Monte-Carlo draws, resulting in time- and memory-complexity of $O(MN)$ and $O(NS)$ respectively, where $S$ is the size of the output logit vector. With respect to the memory-complexity, this is shared by any randomised smoothing based approach, and could be improved by implementing batching across the Monte-Carlo process. While the time cost can be problematic in some contexts, we emphasise that this framework is both requires both fewer adaptations to the training loop and significantly less training time relative to bound propagation approaches \cite{21,36}. We believe these costs can be reduced by performing the optimisation stages with $N' < N$ model draws, and by potentially reusing model draws across the iterative process to approach the $O(N')$ time-complexity of prior randomised smoothing based certifications. Even at present, we believe that the increased computational cost is not intractable, especially for human-in-the-loop certifications.

6 Acknowledgements

This research was undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. This work was also supported in part by the Australian Department of Defence Next Generation Technologies Fund, as part of the CSIRO/Data61 CRP AMLC project. Sarah Erfani is in part supported by Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) DE220100680.

7 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 $\sigma$. Incorporating training time modifications like MACER yields more promising results, with the best performing of our approaches yielding a 4 percentage point increase in the certified proportion at a given radius. 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


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]