Et Tu Certifications: Robustness Certificates Yield Better Adversarial Examples

Andrew C. Cullen [1] **Shijie Liu** [1] **Paul Montague** [2] **Sarah M. Erfani** [1] **Benjamin I.P. Rubinstein** [1]

Abstract

In guaranteeing the absence of adversarial examples in an instance's neighbourhood, certification mechanisms play an important role in demonstrating neural net robustness. In this paper, we ask if these certifications can compromise the very models they help to protect? Our new *Certification Aware Attack* exploits certifications to produce computationally efficient norm-minimising adversarial examples 74% more often than comparable attacks, while reducing the median perturbation norm by more than 10%. While these attacks can be used to assess the tightness of certification bounds, they also highlight that releasing certifications can paradoxically reduce security.

1. Introduction

A troubling property of learned models is that semantically indistinguishable samples can trigger different model outputs (Biggio et al., 2013). Known as *adversarial examples* when constructed deliberately, these samples can be incredibly difficult to detect, especially when the distance to clean examples is minimised. While *adversarial defences* have been proposed as a best-response countermeasure, the security they provide is often illusory as they can be exploited or evaded by motivated attackers.

As a response to this dynamic, guarantees that no adversarial examples exist within a calculable, bounded region—through techniques known as Certified Robustness—have emerged (Weng et al., 2018; Zhang et al., 2018; Li et al., 2019; Salman et al., 2019b; Cullen et al., 2022).

However while these certifications are typically framed as providing security, they cannot intrinsically distinguish between clean or adversarial examples. This is a consequence of *adversarial examples also being able to be certified*, and that practical attacks still exist *outside* the region of certification. While the existence of norm-minimising adversarial examples that satisfy this condition should not be controversial, we strongly argue that the certification community has been underestimating the impact of how these techniques are framed on non-expert users. Presenting certification techniques as being robust to adversarial examples has the potential to induce a false sense of security in users who are not familiar with the inherent risks facing these models.

Within this work we demonstrate that this risk is amplified by the fact that *certifications can be exploited to construct smaller adversarial perturbations* than prior approaches. This attack, which we will henceforth refer to as a *Certification Aware Attack* exploits the very nature of certifications to assist in rapidly identifying adversarial examples—a process that is demonstrated in Figure 1. Exploiting certifications allows our new attack framework to *(i) speed up the initial stages of the search with larger and more informative jumps, and (ii) to reduce the total adversarial perturbation*, which ensures that these adversarial examples have a higher chance of avoiding detection (Gilmer et al., 2018). As part of this, we also introduce a cohesive framework for attacking certified models.

The utility of such an attack extends to improving our understanding of the tightness of calculated certifications—for if a certification is a guaranteed lower bound on the location of the nearest adversarial example, a norm-minimising attack corresponds to an upper bound of the same quantity. However, while there is beneficial utility in these attacks, their existence still demonstrates that certification researchers must fundamentally reconsider how we think about the security implications of certifications. We believe that the only appropriate defence is to treat the certifications and class expectations as highly confidential information that must not be released, lest it be used to help compromise the very models the certification mechanisms purport to defend.

2. Background: Certification Mechanisms

Certification mechanisms eschew the responsive view of adversarial defences in favour of bounding the region in which adversarial examples $\mathbf{x}'$ can exist for a given input sample $\mathbf{x}$—typically $B_p(\mathbf{x}, r)$ a $\mathbf{x}$-centred p-norm ball of some radius r. To be *sound*, a radius must be strictly less

[1]School of Computing and Information Systems, University of Melbourne, Parkville, Australia [2]Defence Science and Technology Group, Adelaide, Australia. Correspondence to: Andrew C. Cullen <andrew.cullen@unimelb.edu.au>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Figure 1: Illustrative example of an evasion attack for a binary classifier, that changes the output from blue to red. Our new attack framework exploits knowledge of the certifications (circles) to minimise the number of iterative steps required.

than

$$r^\star = \inf\left\{\|\mathbf{x}-\mathbf{x}'\|_p \colon \mathbf{x}' \in \mathcal{S}, F(\mathbf{x}) \neq F(\mathbf{x}')\right\} \quad (1)$$
$$\text{where } F(\cdot) = \mathbb{1}\left(\arg\max_{i\in\mathcal{K}} f_i(\cdot)\right) \ .$$

Here $\mathbb{1}$ is a one-hot encoding of the predicted class in $\mathcal{K} = \{1,\ldots,K\}$, and $\mathcal{S}$ is the permissible input space, for example $[0,1]^d$ is typical in computer vision. The size of $B_p(\mathbf{x}, r)$ can be considered a reliable proxy for both the *detectability* of adversarial examples (Gilmer et al., 2018) and the *cost* to the attacker (Huang et al., 2011).

These certificates can be constructed through either exact or statistical-sampling based methods. Exact methods typically construct their bounds by way of either interval bound propagation (IBP), which propagates interval bounds through the model; or convex relaxation, which utilises linear relaxation to construct bounding output polytopes over input bounded perturbations (Salman et al., 2019b; Mirman et al., 2018; Weng et al., 2018; Zhang et al., 2018; Singh et al., 2019; Mohapatra et al., 2020), in a manner that generally provides tighter bounds than IBP (Lyu et al., 2021). However, these techniques exhibit a time or memory complexity that makes them infeasible for complex model architectures or high-dimensional data (Wang et al., 2021; Chiang et al., 2020; Levine & Feizi, 2022). While Lipschitz certified mechanisms (Tsuzuku et al., 2018; Leino et al., 2021) have recently been proposed as a less computationally intensive alternative than bound propagation mechanisms, they still exhibit scalability issues for larger and more semantically complex models.

Statistical-sampling based methods build upon *randomised smoothing* (Lecuyer et al., 2019), in which the Monte Carlo estimator of the expectation under repeatedly perturbed sampling

$$\frac{1}{N}\sum_{j=1}^{N} F(\mathbf{X}_j) \quad \approx \quad \mathbb{E}_{\mathbf{X}}[F(\mathbf{X})] \qquad \forall i \in \mathcal{K} \quad (2)$$
$$\mathbf{X}_1,\ldots,\mathbf{X}_N,\mathbf{X} \quad \overset{i.i.d.}{\sim} \quad \mathbf{x}+\mathcal{N}(0,\sigma^2) \ ,$$

can be exploited to provide guarantees of invariance under *additive* perturbations. In forming this aggregated classification, the model is re-construed as a *smoothed classifier*, which in turn is certified. Approaches for constructing such certifications include differential privacy (Lecuyer et al., 2019; Dwork et al., 2006), Rényi divergence (Li et al., 2019), and parameterising worst-case behaviours (Cohen et al., 2019; Salman et al., 2019a; Cullen et al., 2022). The latter has proved the most performant, and yields certifications of

$$r = \frac{\sigma}{2}\left(\Phi^{-1}\left(\check{E}_0[\mathbf{x}]\right) - \Phi^{-1}\left(\hat{E}_1[\mathbf{x}]\right)\right) \ . \quad (3)$$

Here Φ^{-1} is the inverse standard normal CDF, $(E_0, E_1) = \text{topk}\left(\{\mathbb{E}_{\mathbf{X}}[F(\mathbf{X})]\}, 2\right)$, and $(\check{E}_0, \hat{E}_1)$ are the lower and upper confidence bounds of these quantities to some confidence level α (Goodman, 1965).

3. Attacking Certified Defences

That certifications are framed as being robust to adversarial manipulation has troubling security implications, as it may induce a false sense of security in those unfamiliar with these techniques limitations. This is especially true when as a certification cannot tell us if a sample has already been compromised, and does not obviate the existence of semantically indistinguishable adversarial examples. However, as Table 1 shows, to date only Cohen et al. (2019) has consider test-time attack against certified models. However, they explicitly note that their tested attack does not align with the concept of attacking a certified model, and present no quantitative measures of performance. While other works have considered certifications from an adversarial lens, their focused has been on manipulating the training corpus to improve test time performance—a process that does not require reliable, hard to detect norm-minimising attacks. Thus there currently is clearly an insufficient understanding of both the tightness of certification bounds, and the risks facing certified models.

When it comes to attacking certified models, one may intuitively think of simply identifying a sample $\mathbf{x}'$ such that

Table 1: Extant attacks, distinguished by if their goal is to change the label of samples or to just improve robust accuracy; if they were deployed at train- or test-time; if the attack has direct applicability to certifications (where half-circles denote attempting to improve certified robustness); and if they exploit the certifications themselves.

Algorithms	Goal	Applicability			Exploits Certs.
		Train	Test	Certified	
PGD (Madry et al., 2018)	Label	●	●	○	○
(Carlini & Wagner, 2017)	Label	●	●	○	○
AutoAttack (Croce & Hein, 2020)	Label	●	●	○	○
DeepFool (Moosavi-Dezfooli et al., 2016)	Label	○	●	○	○
Training w/ Noise (Bishop, 1995)	Acc.	●	○	◐	○
(Salman et al., 2019a)	Acc.	●	○	◐	○
MACER (Zhai et al., 2020)	Acc.	●	○	◐	○
Cohen et al. (2019)	Label	○	●	●	○
Ours	Label	◐	●	●	●

$\|\mathbf{x}' - \mathbf{x}\| = r^\star$. However, in practice certification mechanisms are not able to construct tight bounds on $r^\star$, and even if they were, the search space for identifying $\mathbf{x}'$ would still be significant. And even if such a point could be identified, its certified radii would be 0, which would likely trigger further inspection in any operationalised certification system. As such, within this work, we introduce the idea of a *confident* adversarial attack against a certification mechanism being one in which a certification constructed at the adversarial example is non-zero. For randomised smoothing this condition is equivalent to

$$\arg\max \mathbb{E}_{\mathbf{X}}\left[F(\mathbf{x}')\right] \neq \arg\max \mathbb{E}_{\mathbf{X}}\left[F(\mathbf{x})\right] \quad \text{and} \quad \check{E}_0\left[F(\mathbf{x}')\right] > \hat{E}_1\left[F(\mathbf{x}')\right] \quad . \tag{4}$$

That these expectations are highly concentrated (for sufficiently high Monte Carlo sample sizes) enables any of the reference attacks within Table 1 to be effectively employed *against the class expectations, rather than the individual draws under noise*. This contrasts with approaches like Expectation Over Transformation (Athalye et al., 2018), in which each sample under noise is attacked in a numerically inefficient manner.

A complicating factor is that in randomised smoothing is that the final model layers can either be defined differentiable $\mathrm{softmax}$ or non-differentiable $\arg\max$ layers (Cullen et al., 2023). Most certification mechanisms assume the layer, including that of Equation 3. While it could naively be assumed that non-differentiable layers inherently defeat the gradient based attack mechanisms, in practice interventions like stochastic gradient estimation (Fu, 2006; Chen et al., 2019), surrogate modelling, and transfer attacks all providing potent mechanisms for a motivated attacker.

While we could test the performance of adversarial attacks under each of these interventions, distinguishing the impact of the attack against that of the intervention would be problematic. To both minimise the impact of ensuring differentiability and maximise the difficulty to the attacker, within this work we assume that the final $\arg\max$ layer can be replaced with a Gumbel Softmax (Jang et al., 2017)

$$y_i = \frac{\exp\left((\log(\pi_i) + g_i)/\tau\right)}{\sum_{j\in\mathcal{K}} \exp\left((\log(\pi_i) + g_i)/\tau\right)} \quad , \quad \forall i \in \mathcal{K} \quad . \tag{5}$$

We emphasise that this re-parametrisation is not necessary for the following attacks to work, as they can be applied to models with $\mathrm{softmax}$ outputs, or by way of any of the previously discussed interventions.

3.1. Threat model

Within this work we consider an attacker that attempts to construct a norm-minimising, confident (see Equation 4) adversarial example against a certified model. When the certification has been constructed through randomised smoothing, we assume the attacker has the ability to construct derivatives through $\arg\max$ layers, and knowledge of the level of additive noise σ. However, we note that even this last assumption is not strictly necessary, as Appendix F demonstrates that approximate values of σ still provide enhanced certification performance. When attacking models that do not incorporate randomised smoothing, the attacker is assumed to have white-box access to the models gradients, predictions, and certifications.

4. Certification Aware Attacks

While there is value in understanding the tightness of certification mechanisms by attacking them with extant attacks, within this work we are also interested in understanding how certifications may be exploited by a motivated attacker to minimise the size of the identified examples. Such a concept may seem contradictory, but it is important to consider that from an attacker's perspective a certification can be viewed as a *lower bound on the space where attacks may exist.*

Section 4.1 demonstrates how the existence of certifications at all points across the instance space (Cullen et al., 2022)

can be exploited to significantly reduce the search space for identifying adversarial examples. Once an adversarial example is identified, Section 4.2 then demonstrates how certifications associated with successful adversarial examples can be *exploited to minimise the perturbation norm of the sample*, as any norm-minimising step inside the certified radii still remains an adversarial example!

4.1. Step Size Control

We begin our attack by introducing the surrogate problem

$$\hat{\mathbf{x}} = \underset{\hat{\mathbf{x}} \in \mathcal{S}}{\arg\min} \left\{ |E_0(\hat{\mathbf{x}}) - E_1(\hat{\mathbf{x}})| \ : \ F(\hat{\mathbf{x}}) = F(\mathbf{x}) \right\} \ . \quad (6)$$

This formalism may seem counter-intuitive, as the constraint ensures that $\hat{\mathbf{x}}$ cannot be an adversarial example. However, consider the gradient-based solution of the previous problem

$$\mathbf{x}_{i+1} = P_{\mathcal{S}} \left(\mathbf{x}_i - \epsilon_i \left(\frac{\nabla_{\boldsymbol{x}_i} |E_0[\mathbf{x}_i] - E_1[\mathbf{x}_i]|}{\|\nabla_{\boldsymbol{x}_i} |E_0[\mathbf{x}_i] - E_1[\mathbf{x}_i]|\|} \right) \right) \ , \quad (7)$$

for a projection to the feasible space $P_{\mathcal{S}}$, and for which each $\mathbf{x}_i$ has associated certifications r_i. By imposing that $\epsilon_i > r_i$, we ensure that the new candidate solution $\mathbf{x}_{i+1}$ must exist outside the region of certification of the previous point, which is a *necessary but not sufficient* condition for identifying an adversarial example.

Ensuring that $\epsilon_i > r_i$ could be achieved by imposing that

$$\epsilon_i = \rho(\mathbf{x}_i)\,(1 + \delta) \ , \quad (8)$$

for some $\delta > 0$, and where $\rho(\mathbf{x}_i) = r_i$. However, doing so fails to account for the information gained from the certifications at all $\mathbf{x}_j$ for $j = 0, \ldots, i$. If we instead construct

$$\rho(\mathbf{x}_i) = \inf \left\{ \hat{\rho} \geq 0 : \mathbf{x}^\star(\hat{\rho}) \notin \bigcup_{j=0}^{i} B_P(\mathbf{x}_j, H\left[c_0 = c_j\right] r_j) \right\} \quad (9)$$

$$\mathbf{x}^\star(\hat{\rho}) = P_{\mathcal{S}} \left(\mathbf{x}_i - \hat{\rho} \left(\frac{\nabla_{\boldsymbol{x}_i} |E_0[\mathbf{x}_i] - E_1[\mathbf{x}_i]|}{\|\nabla_{\boldsymbol{x}_i} |E_0[\mathbf{x}_i] - E_1[\mathbf{x}_i]|\|} \right) \right) \ ,$$

then $\mathbf{x}_{i+1}$ remains strictly outside the certified radii region of all $\{\mathbf{x}_i | c_i = c_0\}$. Here c_i is the class prediction at step i of the iterative process, and $H_{c_0 = c_i}$ is an indicator function , and $\rho(\mathbf{x}_i)$ is solved for using a binary search.

As taking large steps may be disadvantageous in certain contexts, in practice we define ϵ_i in terms of pre-defined lower- and upper-bounds

$$\tilde{\epsilon}_i = \text{clip}\left(\epsilon_i, \epsilon_{\text{min}}, \epsilon_{\text{max}}\right) \ . \quad (10)$$

4.2. Refining Adversarial Examples

Once we have identified an adversarial example, we switch to the second stage of our iterative process, in which we minimise the perturbation norm of any identified examples, in order to decrease their detectability. At this stage, the attack iterates $\mathbf{x}_i$ now produces a class prediction of $c_i \neq c_0$. Thus, any $\mathbf{x}_i$ must also be an adversarial attack if the difference between the two points is less than or equal to r_i. Thus our iterator can be defined as

$$\mathbf{x}_{i+1} = P\left(\mathbf{x}_i - \min\{\rho, \epsilon_{\text{max}}\}(1 - \delta) \left(\frac{\mathbf{x}_0 - \mathbf{x}_i}{\|\mathbf{x}_0 - \mathbf{x}_i\|} \right) \right)$$

$$\rho = \sup \left\{ \hat{\rho} \geq 0 : \mathbf{x}^\star(\hat{\rho}) \in \bigcup_{j=0}^{i} B_P(\mathbf{x}_j, H\left[c_0 \neq c_j\right] r_j) \right\} \quad (11)$$

$$\mathbf{x}^\star(\hat{r}) = P_{\mathcal{S}} \left(\mathbf{x}_i - \hat{\rho} \left(\frac{\mathbf{x}_0 - \mathbf{x}_i}{\|\mathbf{x}_0 - \mathbf{x}_i\|} \right) \right) \ .$$

Similar to Section 4.1, a simpler variant of the above simply involves setting that $\rho = r_i$, however doing so discards the potential for prior certifications to help refine the search space. This framing ensures that $c_i = c_j \ \forall \ j > i$—i.e., that all adversarial examples share the same class as the first identified adversarial example. As such, it may be true that there exists some adversarial example $\mathbf{x}''$

$$\|\mathbf{x}'' - \mathbf{x}_0\| < \|\mathbf{x}_i - \mathbf{x}_0\| \qquad \forall \ i \in \mathbb{N} \ .$$

However, even with this limitation, the following section demonstrates that this process still produces significantly smaller adversarial examples than other techniques.

Algorithms detailing the aforementioned processes can be found within Appendices B and C, and the code associated with this work can be found at `https://github.com/andrew-cullen/Attacking-Certified-Robustness`.

5. Results

To demonstrate the performance of our new Certification Aware Attack, we test our attack relative to a range of other comparable approaches. We emphasise that both our new attack and the reference attacks are *deployed against certified models*, rather than the associated base classifiers.

To achieve this, our experiments consider attacks against MNIST (LeCun et al., 1998) (GNU v3.0 license), CIFAR-10 (Krizhevsky et al., 2009) (MIT license), and the Large Scale Visual Recognition Challenge variant of ImageNet (Deng et al., 2009; Russakovsky et al., 2015) (which uses a custom, non-commercial license). In the case of models defended by randomised smoothing, each model was trained in PyTorch (Paszke et al., 2019) using a ResNet-18 architecture, with experiments considering two distinct levels of smoothing noise scale σ. Additional experiments on the MACER (Zhai et al., 2020) certification framework and a ResNet-110 architecture can be found in Appendix G.

The confidence intervals of expectations in all experiments were set according to the $\alpha = 0.005$ significance level. To demonstrate the generality of our identified threat model, Section 5.3 eschews randomised smoothing to attack certifications constructed using IBP. Due to the inherent computational cost associated with constructing solutions with IBP, our results were limited to MNIST models solved using a sequential model of two convolutional layers followed by two linear layers, with ReLU activation functions. All calculations were constructed using one NVIDIA A100 GPUs for MNIST and CIFAR-10, while Imagenet test and training time evaluations employed two.

To explore the performance of our attacks, we will rely upon a handful of key metrics, including some that we have constructed specifically for this work. These include the success rate of successful attacks; the proportion of samples for which an attack produces smaller radii perturbations than its competitors; the percentage increase *above* the certified radius $\%$-C; the median ℓ_2 attack size r_{50}; and the attack time, which includes the time for certifying each sample. Further details can be found in Appendix D

5.1. Attacking Randomised Smoothing

Attack Size vs Success Rate To establish the performance of our Certification Aware Attack framework against certifications employing randomised smoothing, Figure 2 explores the relationship between success rates and the size of the identified adversarial examples. As each technique may be successfully attacking a different subset of samples, the size of the identified adversarial examples is normalised by their associated certified radii in $\%$-C to control for the difficulty of identifying examples. This was achieved through a broad sweep of each model's respective parameter spaces, as detailed within Table 3 and Appendix C. Exemplars of some of these attacks can be seen in Appendix H.

These results demonstrate the existence of a quasi-exponential relationship between the attack size and the location of the smallest *potential* adversarial example, which serves as a proxy for detectability. This growth is in part a product of the range of the explored parameter space (see Appendix C), which extends to attacks so large that they remove all semantic features from the inputs—producing almost guaranteed attack successes, at the cost of clearly detectable perturbations.

Our approach consistently identifies significantly smaller adversarial examples than both PGD and Carlini-Wagner for all success rates, with a 20 percentage point difference in the distance to Cohen et al. seen for CIFAR-10 at $\sigma = 1.0$, and an over 30 percentage point difference for MNIST and ImageNet. While it is true that our technique produces slightly smaller maximum success rates than PGD in MNIST and CIFAR-10 (with a larger, notable gap for ImageNet), we emphasise the significant differences between the observed percentages distances to Cohen, especially as the success rate grows for MNIST and CIFAR-10. The the adversarial examples identified by our technique are consistently closer to the certified radii confirms that our technique is reliably producing smaller, more difficult to detect adversarial examples.

Relative Performance To enable representative comparisons, for the remainder of this work we will assume that an attacker that can control its position in parameter space will choose hyper-parameters such that it minimises the median percentage difference to Cohen et al. for a success rate above 90%. If such a success rate is not achievable, the attacker will instead maximise their success rate.

To more comprehensively examine the suite of potential attack frameworks, we now expand our suite of comparisons to also include DeepFool (Moosavi-Dezfooli et al., 2016) and AutoAttack (Croce & Hein, 2020), which were excluded from broader parameter sweeps due to their relative performance. While AutoAttack has the ability to specify a ℓ_2 norm perturbation magnitude, the associated computational cost makes a broader parameter space exploration infeasible. In contrast, while DeepFool is the fastest of all tested attacks, its failure to successfully identify norm minimising adversarial examples led to its exclusion from a broader parameter exploration.

Across our full set of experiments, Figure 3 and Table 2 demonstrate that our new attack framework consistently constructs smaller adversarial perturbations than any other tested technique. On a sample-by-sample basis, in the most challenge experiment for our technique—Imagenet at $\sigma = 1.0$—our technique produces the smallest adversarial example for 54% of the time (denoted by the *Best* column), for samples able to be attacked. This result is particularly striking given the relatively low success rate for our approach in Imagenet, relative to the other experiments, which suggests that the range of parameter space tested over may need further modification for datasets of the size and complexity of Imagenet. In the remainder of the tested experiments, as the success rate of our technique increases, so too does the proportion of attacked samples for which our technique produces the smallest possible adversarial attack, demonstrating the viability of our approach as a framework for constructing minimal norm adversarial examples.

Our approach produces a median certification that is on average 11% smaller for MNIST, 12% for CIFAR-10, and 52% smaller in the case of Imagenet. When controlling for the size of the certified radii $\%$-C demonstrates that our technique produces an on average 24% reduction in the median attack size relative to the next best attack.

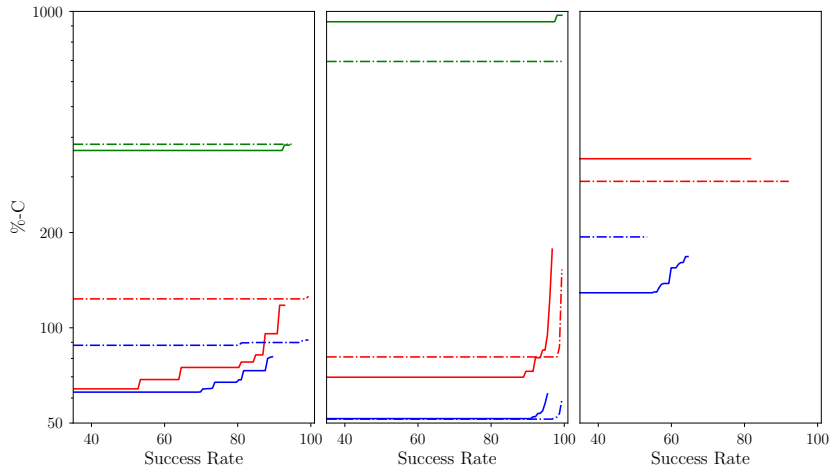


Figure 2: Minimum achievable average percentage difference between the attack radii and the certified guarantee of Cohen et al. (Equation 3) for a given success rate for our technique (Blue), PGD (Red), and Carlini-Wagner (Green), when tested against MNIST, CIFAR-10 and Imagenet. Solid and dashed lines represents $\sigma = \{0.5, 1.0\}$ for parameter space of Table 3.

Contextualising Performance One feature noted within Section 4.2 was that all adversarial examples identified by our Certification Aware Attack framework must share the same class prediction as the first identified adversarial example. Intuitively it would appear that such a drawback would induce a disproportionate increase in the median certified radii for the 1000-class ImageNet, as compared to MNIST or CIFAR-10. In practice this is more than counterbalanced by our attacks increased efficiency in exploring the search space. This efficiency in exploring the search space is evident in the computational cost of identifying attacks, with our approach requiring significantly less computational time to identify norm minimising adversarial examples relative to all of the other techniques. The exception to this is DeepFool, however its performance delta relative to the Cohen et al. certified radius emphasises that adding additional iterative steps would likely be a forlorn task.

5.2. Tightness of Certified Guarantees

While attacks against certifications are valuable in their own right, they also serve as an upper-bound on the location of the nearest adversarial example, counterbalancing the lower-bound provided by certification mechanisms. In doing so, they provide a proxy for both utility and the potential scope for future improvements in certification schemes.

When considering these bounds in the contexts of the tested datasets, Figure 3 suggests that MNIST—which is often perceived as being the simplest of datasets—demonstrates a more significant delta between the certified radii and the attack performance. This may be a consequence of the simpler semantic properties of the dataset being more difficult to attack, relative to Cohen et al. style certifications. Due to its role as a multiplicative constant in Equation 3, increasing σ inherently increases the size of the certifications, an effect that is partially offset by a decrease in the observed class expectations. However, from the perspective of an attacker increasing σ should also increase the smoothness of the gradients, which theoretically should make the model *easier to attack*. In practice, Figure 3 and Table 2 demonstrate that increasing σ leads to a small increase in the size of identified attacks, relative to certified guarantees. While this may at first appear contradictory, it suggests that the ease in identifying adversarial attacks for larger σ is offset by decreases in the tightness of the certified bound.

5.3. Performance against other certification mechanisms

To demonstrate the generality of our identified threat model, Figure 4 demonstrates the relative performance of our technique and PGD when tested for a model certified using IBP. While we have not fully explored the parameter spaces of both attacks, nor the broader suite of attacks in the context of this framework, these results reinforce the *information advantage* an attacker has when attempting to compromise models employing randomised smoothing *if they incorporate the certification into their attack*, irrespective of the certification mechanism. That this is true confirms that all certification mechanisms should assess their risk to adversarial attack in light of our Certification Aware Attacks.

6. Discussion: Impact and Mitigation

We emphasise that our CAA does not compromise the certification associated with individual samples. Rather, we establish that it is possible to leverage certifications to con-

Table 2: MNIST (M), CIFAR-10 (C), and ImageNet (I) attack performance across σ for Carlini-Wagner (C-W), AutoAttack (Auto) and DeepFool (DeepF). Metrics are the proportion of samples attacked (*Suc.*), smallest attack proportion (*Best*), median attack size (r_{50}), time (*Time* [s]), and percentage difference to the Cohen et al. (%-C). All bar the success rate are only calculated over *successful attacks*.$\star$ denotes solutions selected following Appendix C, bolded values represent the best performing metric (excluding the success rate, as it is a control parameter), and arrows denote if a metric is more favourable with increased or decreased values.

		$\sigma = 0.5$					$\sigma = 1.0$				
	Type	Suc.↑	Best ↑	r_{50} ↓	%-C↓	Time ↓	Suc.↑	Best ↑	r_{50} ↓	%-C↓	Time ↓
M	Ours*	90%	**73%**	**2.02**	**82**	0.34	97%	**97%**	**2.23**	**90**	1.22
	PGD*	91%	19%	2.17	96	2.04	99%	3%	2.62	123	2.03
	C-W*	93%	7%	5.46	364	3.03	95%	0%	5.36	380	3.02
	Auto	92%	1%	5.44	393	27.32	97%	0%	5.65	386	26.50
	DeepF	9%	0%	14.43	2417	**0.07**	51%	0%	17.10	2143	**0.07**
C	Ours*	91%	**87%**	**0.83**	**56**	0.53	96%	**92%**	**1.26**	**56**	0.86
	PGD*	92%	4%	0.92	72	2.17	99%	3%	1.46	77	2.15
	C-W*	98%	5%	3.13	432	3.18	99%	1%	3.65	352	3.14
	Auto	94%	3%	4.00	493	28.37	91%	2%	5.61	492	28.40
	DeepF	88%	2%	2.44	504	**0.08**	98%	3%	3.42	462	**0.08**
I	Ours*	63%	**84%**	**1.14**	**127**	4.49	52%	**63%**	**1.43**	**157**	5.21
	PGD*	82%	13%	2.05	211	51.46	91%	35%	3.42	188	50.49
	C-W*	53%	0%	33.22	4747	26.71	56%	0%	32.42	3451	27.34
	DeepF	56%	3%	2.89	654	**2.88**	71%	3%	4.98	647	**2.93**

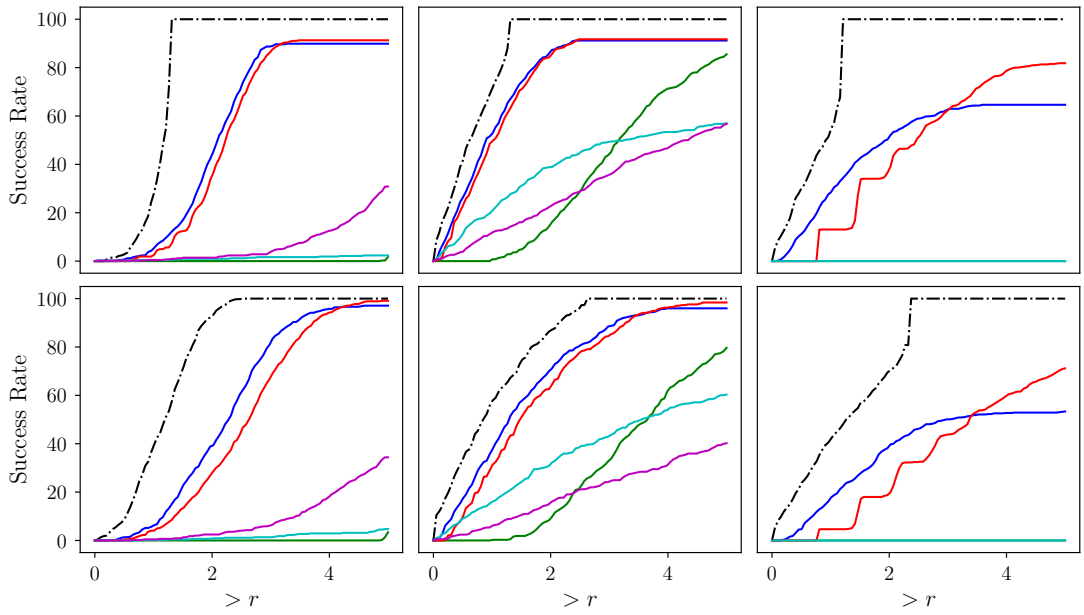


Figure 3: Best achieved Attack Proportion for our new Certification Aware Attack (blue), PGD (red), DeepFool (cyan), Carlini-Wagner (green), and AutoAttack (magenta); where the rows correspond to $\sigma = \{0.5, 1.0\}$ and the columns correspond to MNIST, CIFAR-10 and Imagenet. Black dotted line represents the best case performance as per Equation 3.

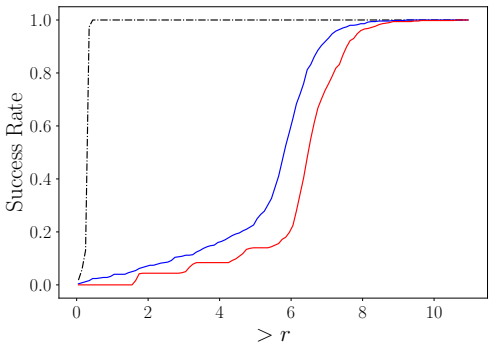


Figure 4: Success rates for our attack (blue) and PGD (red) for an IBP certified MNIST model.

struct samples that are semantically identical to clean samples (see Figure 8), but that are able to trick a certified classifier into changing its class prediction. That this is possible demonstrates that the security provided by certification mechanisms is illusory if the attacker knows (or can reconstruct) the robustness certificate. This observation runs contrary to likely user expectations, where a certification or the class expectations is likely to be be seen as metric for demonstrating model confidence to be released alongside the predicted class.

While uncovering new attacks has the potential to compromise deployed systems, there is a prima facie argument that any security provided by ignoring new attack vectors is illusory. Such a perspective has uncovered new attack vectors including data poisoning, backdoor attacks, model stealing, and transfer attacks, all of which can now be protected against. Our work similarly reveals the paradoxical observation that the very mechanisms that we rely upon to protect models introduce new attack surfaces, enhancing the ability for attackers to construct norm minimising attacks. Moreover, our CAA also provides a framework to better explore the tightness of constructed certification bounds in real world systems.

It would appear that restricting the publication of the certification could nullify the potential of certification aware attacks. However, the certifications can be trivially reconstructed from the class expectations and σ. Thus securing systems against these attacks requires that only the class predictions are released, without any associated expectations. While this would not prevent a suitably motivated attacker developing a surrogate model, it would significantly increase the cost and difficulty associated with a successful attack, relative to attacking a model that releases its certifications. Figure 6b and Appendix F demonstrate that our attack still outperforms other frameworks even when σ is estimated, however we leave exploring the effectiveness of this mitigation to future work.

While this work has specifically considered ℓ_2 norm bounded evasion attacks against randomised smoothing and IBP based mechanisms, it is important to note that other threat models threat models (Liu et al., 2023), and certification frameworks exist, like Lipschitz certification (Tsuzuku et al., 2018; Leino et al., 2021). However, these approaches still involve constructing a certificate, and if these certificates are published then this work has clearly demonstrated that this release can be exploited by a motivated attacker. As such, we believe that any future certification—for evasion, backdoor, or other attacks—must seriously consider the risk associated with publishing certifications.

7. Related Work

This work presents both a framework for attacking certifiably robust models, and a demonstration of how such certification can be exploited to improve attack efficacy. While we formalise the concept of attacking certifications, prior works have considered the impact of corrupting the inputs of both undefended and certified models. One common framework involves corrupting input samples with additive noise or adversarial examples, in order to improve robustness (Bishop, 1995; Salman et al., 2019a; Cohen et al., 2019). Of these, Salman et al. (2019a) is closest to our work, although their attacks only considered a small number of draws from randomised smoothing (rather than the full expectations), and employed a $\mathrm{softmax}$ in place of the $\arg\max$ operator. All three of these approaches are un-targeted, un-directed, training time modifications attempting to improve generalisation by increasing training loss. In contrast, our focus was placed upon both constructing a definition of test time adversarial attacks against certified models, and then exploiting the nature of certifications themselves to improve the performance of adversarial attacks against certified models.

8. Conclusion

Within this work we have demonstrated the counter-intuitive concept that certifications can be exploited to attack the very models they were designed to guard. Through our novel Certification Aware Attack framework, we exploit this observation to significantly decrease the size of the identified adversarial perturbations relative to state-of-the-art test-time attacks, leading to an up to 55% decrease in the size of adversarial perturbations relative to the next best performing technique. Being able to reliably, and repeatedly generate such norm-minimising adversarial examples would better allow for developers to analyse the performance of certification mechanisms. However, this same benefit would also allow an attacker to reliably influence more samples before potentially being detected. These results underscore that significant consideration must be placed upon the safety of releasing robustness certificates.

Impact Statement

This work explores a heretofore undiscovered security risk associated with publishing the certifications associated with certifiable robust mechanisms. While such attacks can induce negative societal consequences, there is a clear consensus within the computer security community that responsible disclosure of attacks is a crucial part of improving systemic security. Training-time, test-time and backdoor attacks are always published in the public domain, and the risk associated with the publication of our new attack should be viewed within this context. The new attack vector contained within this work allows for a fundamental reevaluation of the security associated with systems designed to demonstrate the resistance of machine learning models to adversarial manipulation, and provides a new mechanism that can be used to better study the tightness of the bounds provided by robustness mechanisms.

References

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing Robust Adversarial Examples. In *International Conference on Machine Learning*, ICML, pp. 284–293. PMLR, 2018.

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*, ECMLPKDD, pp. 387–402. Springer, 2013.

Chris M Bishop. Training with Noise is Equivalent to Tikhonov Regularization. *Neural Computation*, 7(1): 108–116, 1995.

Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *2017 IEEE Symposium on Security and Privacy (S & P)*, pp. 39–57. IEEE, 2017.

Beidi Chen, Yingchen Xu, and Anshumali Shrivastava. Fast and Accurate Stochastic Gradient Estimation. In *Advances in Neural Information Processing Systems*, volume 32, pp. 12349–12359. NeurIPS, 2019.

Ping-yeh Chiang, Renkun Ni, Ahmed Abdelkader, Chen Zhu, Christoph Studer, and Tom Goldstein. Certified Defenses for Adversarial Patches. In *International Conference on Learning Representations*, ICLR, 2020.

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

Francesco Croce and Matthias Hein. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks. In *International Conference on Machine Learning*, ICML, pp. 2206–2216. PMLR, 2020.

Andrew C. Cullen, Paul Montague, Shijie Liu, Sarah Monazam Erfani, and Benjamin I.P. Rubinstein. Double Bubble, Toil and Trouble: Enhancing Certified Robustness through Transitivity. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19099–19112. NeurIPS, 2022.

Andrew C Cullen, Paul Montague, Shijie Liu, Sarah M Erfani, and Benjamin I.P. Rubinstein. It's Simplex! Disaggregating Measures to Improve Certified Robustness. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 65–65. IEEE Computer Society, 2023.

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*, CVPR, pp. 248–255. IEEE, 2009.

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*, CVPR, pp. 9185–9193, 2018.

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

Michael C Fu. Gradient Estimation. *Handbooks in Operations Research and Management Science*, 13:575–616, 2006.

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.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, ICLR, 2015.

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

Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, and J. D. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, pp. 43–58, 2011.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical Reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, ICLR, 2017.

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

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.

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 (S & P)*, pp. 656–672. IEEE, 2019.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-Robust Nneural Networks. In *International Conference on Machine Learning*, pp. 6212–6222. PMLR, 2021.

Alexander Levine and Soheil Feizi. (de)Randomized Smoothing for Certifiable Defense against Patch Attacks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6465–6475. NeurIPS, 2022.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified Adversarial Robustness with Additive Noise. In *Advances in Neural Information Processing Systems*, volume 32, pp. 9459–9469. NeurIPS, 2019.

Shijie Liu, Andrew C Cullen, Paul Montague, Sarah M Erfani, and Benjamin IP Rubinstein. Enhancing the Antidote: Improved Pointwise Certifications against Poisoning Attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 8861–8869, 2023.

Zhaoyang Lyu, Minghao Guo, Tong Wu, Guodong Xu, Kehuan Zhang, and Dahua Lin. Towards Evaluating and Training Rerifiably Robust Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 4308–4317, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, ICLR, 2018.

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

Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Towards Verifying Robustness of Neural Networks Against A Family of Semantic Perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 244–252, 2020.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, CVPR, pp. 2574–2582, 2016.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035. NeurIPS, 2019.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

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*, volume 32, pp. 11292–11303. NeurIPS, 2019a.

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*, volume 32, pp. 9835–9846. NeurIPS, 2019b.

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.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing Properties of Neural Networks. In *International Conference on Learning Representations*, ICLR, 2014.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-Margin Training: Scalable Certification of Per-

turbation Invariance for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. NeurIPS, 2018.

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. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29909–29921. NeurIPS, 2021.

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*, ICML, pp. 5276–5285. PMLR, 2018.

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*, ICLR, 2020.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 31, pp. 4939–4948. NeurIPS, 2018.

A. Background: Adversarial Examples

The existence of highly confident but incorrect adversarial examples in neural networks has been documented extensively (Szegedy et al., 2014; Goodfellow et al., 2015). We provide an overview of the topic in this appendix for completeness. Formally, adversarial examples are perturbations $\boldsymbol{\gamma} \in \mathcal{S}$ to the input $\mathbf{x} \in \mathcal{S}$ of a learned model $\mathbf{f}(\cdot)$, for which $F(\mathbf{x}+\boldsymbol{\gamma}) \neq F(\mathbf{x})$.

The p-norm of this perturbation can be considered a reliable proxy for both the *detectability* of adversarial examples (Gilmer et al., 2018) and the *cost* to the attacker (Huang et al., 2011).

The process for identifying such attacks commonly involves gradient descent over the input space. A prominent example is the Iterative Fast Gradient Sign Method (Madry et al., 2018; Dong et al., 2018), which we will henceforth refer to as PGD. This technique attempts to converge upon an adversarial example by way of the iterative scheme

$$\mathbf{x}_{k+1} = P_{\mathcal{S}}\left(\mathbf{x}_k - \epsilon\left(\frac{\nabla_{\mathbf{x}} J(\mathbf{x}, y)}{\|\nabla_{\mathbf{x}} J(\mathbf{x}, y)\|_2}\right)\right) \quad . \tag{12}$$

This process exploits gradients of the loss $J(\mathbf{x}, y)$ relative to a target label y to form each attack iteration, with the step size ϵ and a projection operator P ensuring that $\mathbf{x}_{k+1}$ is restricted to $\mathcal{S}$.

Carlini & Wagner (2017)—henceforth known as C-W—demonstrated the construction of adversarial perturbations by employing gradient descent to solve

$$\arg\min_{\mathbf{x}'} \left\{ \|\mathbf{x}' - \mathbf{x}\|_2^2 + c \cdot g \right\} \tag{13}$$
$$g = \max\left\{ \max\{ f_\theta(\mathbf{x}')_j : j \neq i \} - f_\theta(\mathbf{x}')_i, -\kappa \right\} \quad .$$

We note that while one-shot variants of these attacks have historically been used as a baseline for the performance of iterative attacks to be assessed against, we believe that by their nature such attacks always poorly represent the success-rate and attack-size trade off. Instead, we have performed our comparisons against the certified guarantee of Cohen et al. at the sample point, which provides an absolute lower bound on the size of possible adversarial attacks. We feel that this form of comparison more appropriately captures how these techniques perform, rather than attempting to compare one-shot with iterative attacks, which fundamentally incorporate different access level threat models.

B. Algorithms

Within Algorithm 1, lines 6–11 cover the processes outlined within Section 4 and 4.1, with lines 13–17 covering the materials of Section 4.2.

One important piece of detail relates to the case where $\check{E}_0 < \hat{E}_1$, which is equivalent to $r = 0$. Under both of these circumstances, the model is unable to construct a confident prediction, so the algorithm induces minimal size-steps either away from the origin—if an adversarial example has not yet been identified—or towards the most recent point, if that point was an adversarial example.

In order to calculate the class expectations and associated certifications for a given input $\mathbf{x}'$, Algorithm 2 performs the Monte-Carlo sampling and then corrects for sampling uncertainties. We note here that for the purposes of constructing derivatives, the lower and upper-bounding processes are treated as if they were perturbations to the expectations, and as such they are not considered as a part of the differentiation process. While this has the potential to slightly perturb the derivatives, our experiments have demonstrated that any $\delta > 0$ is sufficient to more than compensate

C. Parameter Space

As was discussed in Section 5, understand the relative performance of techniques requires a consideration of how an attacks parameter space influences its performance metrics.

Algorithm 1 Certification Aware Attack Algorithm.

Input: data $\mathbf{x}$, level of additive noise σ, samples N, iterations M, true-label i, minimum and maximum step size $(\epsilon_{\text{min}}, \epsilon_{\text{max}})$, scaling factor $\delta \in [0, 1]$
$\mathbf{x}', \mathbf{x}'_s$ Successful $= \mathbf{x}, \mathbf{x}$, False
for 1 **to** M **do**
  $\mathbf{y}, \breve{E}_0, \widehat{E}_1, r = \text{Model}(\mathbf{x}'; \sigma, N)$ ▷ Detailed in Algorithm 2
  if $\arg\max_{i \in \mathcal{K}} y = i$ **then** ▷ Adversarial Example not yet identified.
    if $\breve{E}_0 > \widehat{E}_1$ **then**
      $\epsilon =$ Equation 8 $(\mathbf{x}', \delta, \epsilon_{\text{min}}, \epsilon_{\text{max}})$
    else
      $\epsilon = \epsilon_{\text{min}}$
    end if
    $\mathbf{x}' =$ Equation 7 $(\mathbf{x}', \epsilon)$
  else
    if $r = 0$ **then** ▷ Attempting to improve confidence of adversarial examples
      $\mathbf{x}' = P_{\mathcal{S}}\left(\mathbf{x}' + \epsilon_{\text{min}} \frac{\nabla_{\mathbf{x}'}\left(\breve{E}_0 - \widehat{E}_1\right)}{\|\nabla_{\mathbf{x}'}\left(\breve{E}_0 - \widehat{E}_1\right)\|_2}\right)$
    else ▷ Examples are refined while staying inside the certified radii
      $\mathbf{x}'_s$, Successful = $\mathbf{x}'$, True
      $\mathbf{x}' =$ Equation 11$(\mathbf{x}', \delta, \epsilon_{\text{max}})$
    end if
  end if
end for
return $\mathbf{x}'_s$, Successful

In aide of this, for our three most highly performant attack frameworks, for each dataset we performed a parameter sweep over the parameters outlined within Figure 3. From this, for each attack we selected a representative position in parameter space that either exhibited the minimal %-C for a success rate over 90%, or, if such a success rate was not achievable, the maximum achievable success rate. In doing so, we attempted to construct fair comparisons that accurately reflected the performance of the techniques.

We note that within this table the parameter ϵ_{max} (for our technique) and ϵ for PGD extend to a level close to that for which the semantic features of the inputs would entirely be destroyed by the attacker. While this choice was made in the interests of validating the performance of our attacks, we emphasise that none of these results ended up factoring into our key reported results.

One complicating factor of such parameter sweeps is the computational cost associated with the exploration, especially in the case for Imagenet—as can be seen in Table 2. As such while we endeavoured to select our representative attacks based upon 500 randomly selected samples, it was only possible to consider 50 samples for PGD and Carlini-Wagner for Imagenet due to the computational time associated with these parameter sweeps.

To explore the influence of the step-size control parameters of Equation 10, Figure 5 considers the influence of a range of these parameters upon key attack metrics, based upon the parameter space explored over Appendix C. Based upon this , it is clear that the primary driver of the success-rate and certification size trade off (as explored in Figure 2) is the parameter ϵ_{max}, that controls the largest possible step size that the Certification Aware Attack framework is allowed to make. Thus further exploring the parameter space in this direction would likely be a critical factor in increasing the success rate observed for Imagenet.

D. Metrics

To help explore the relative performance of the tested techniques we consider a series of metrics which, in aggregate, reflect the overall performance of the technique. To explain these metrics in additional detail, the *Success Rate* represents the proportion of correctly predicted samples for which a technique is able to construct a successful attack, and can be calculated as

$$\text{Suc.}_i = \frac{1}{N} \sum_{j=1}^{N} (r_{i,j} > 0) \ . \tag{14}$$

Here the subscript i denotes a particular attack drawn from the set of attacks $\mathcal{I}$, and $r_{i,j} = \|\mathbf{x}'_j - \mathbf{x}_j\|$ is the attack radii, which for notational simplicity is set to 0 in the case of a failed attack. The set of samples (of size N) has been filtered to ensure that each is correctly predicted by the model in the absence of an adversarial attack.

The *Best* is then the proportion of samples that a particular technique produces an attack radii smaller than any other correctly identified adversarial attack is calculable as

$$\text{Best}_i = \frac{\sum_{j=1} r_{i,j} \leq r_{i',j} \qquad \forall\, (i' \neq i) \in \mathcal{I}}{\sum_{j=1} r_{i,j} > 0 \qquad \forall\, i \in \mathcal{I}} \tag{15}$$

Increases to both of these metrics are advantageous, although as was noted in Appendix C each result within Table 2 must be contextualised against the decision to attempt to control the success rate to approximately 90%, if such an success rate was achievable for the technique in light of the tested parameter space.

The measure %-C represents the median percentage difference between the attack radii and the certified guarantee of Cohen et al., which takes the form

$$\%\text{-C} = \text{med}_{r_{i,j} > 0} \left(\frac{r_{i,j} - C(\mathbf{x}_j)}{C(\mathbf{x}_j)} \right) \ . \tag{16}$$

Here $\text{med}(\cdot)$ is the median over the set of successfully attacked samples, and $C(\mathbf{x}_j)$ is the certified radii for an ℓ_2

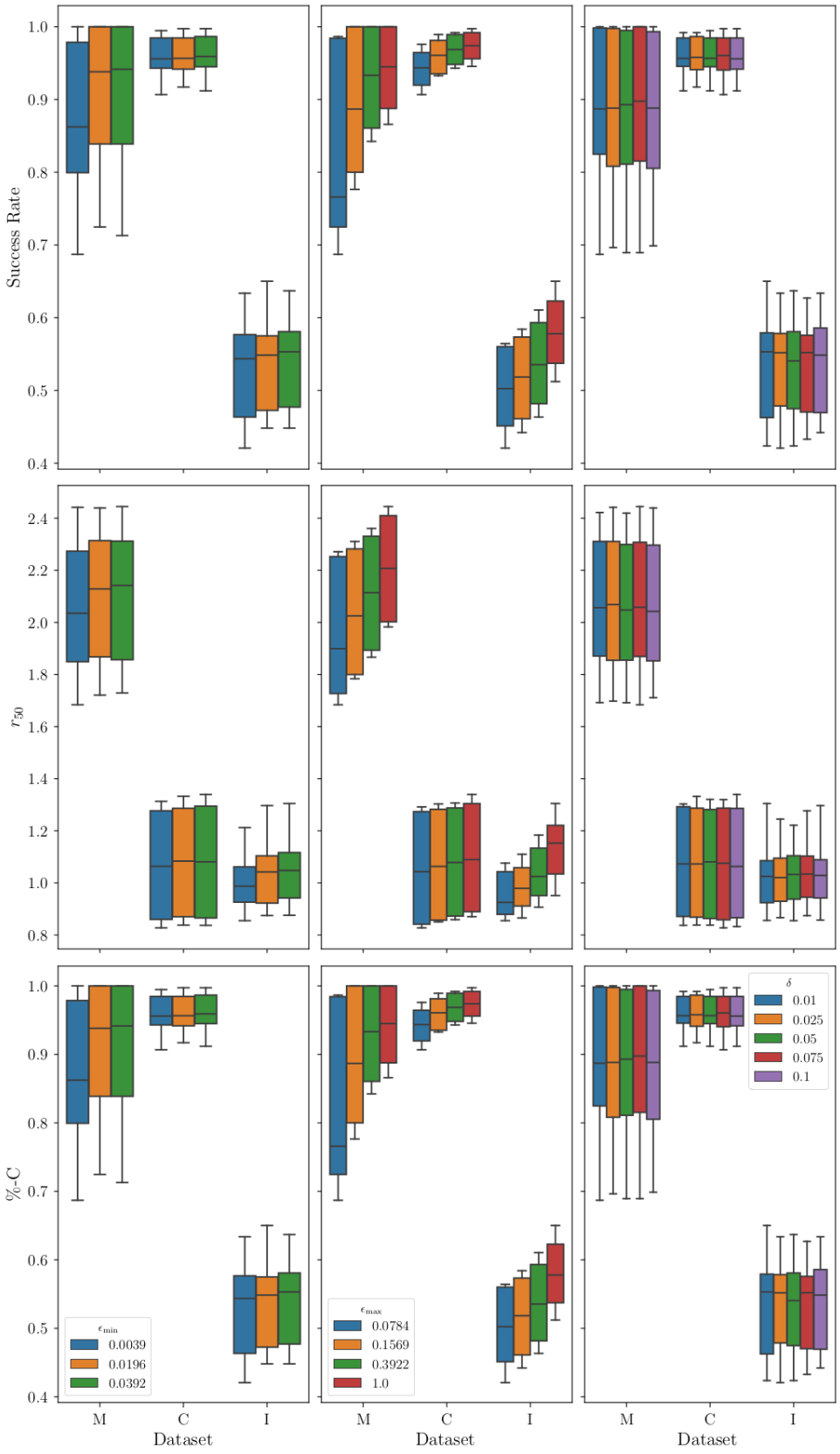


Figure 5: Response of key metrics for our Certification Aware Attack to changes in $\epsilon_{\min}$, $\epsilon_{\max}$ and δ.

Algorithm 2 Class prediction and certification for the Certification Aware Attack algorithm of Algorithm 1.

Input: Perturbed data $\mathbf{x}'$, samples N, level of added noise σ
$\mathbf{y} = \mathbf{0}$
for i = 1:N **do**
    $\mathbf{y} = \mathbf{y} + GS\left(f_\theta\left(\mathbf{x}' + \mathcal{N}(0, \boldsymbol{\sigma}^2)\right)\right)$
end for
$\mathbf{y} = \frac{1}{N}\mathbf{y}$
$(z_0, z_1) = \text{topk}(\mathbf{y}, k = 2)$ ▷ topk is used as it is differentiable, $z_0 > z_1$
$\left(\breve{E}_0, \widehat{E}_1\right) = (\text{lowerbound}(\mathbf{y}, z_0), \text{upperbound}(\mathbf{y}, z_1))$ ▷ Calculated via Goodman (1965)
$R = \frac{\sigma}{2}\left(\Phi^{-1}(\breve{E}_0) - \Phi^{-1}(\widehat{E}_1)\right)$
return $\mathbf{y}, \breve{E}_0, \widehat{E}_1, R$

Table 3: Parameter space employed for our Certification Aware Attack, PGD (see Equation 12 for details), and Carlini-Wagner (see Equation 13).

Ours	$\epsilon_{\text{min}} \times 255$	$=$	$\{1, 5, 10\}$
	$\epsilon_{\text{max}} \times 255$	$=$	$\{20, 40, 100, 255\}$
	δ	$=$	$\{0.01, 0.025, 0.05, 0.075, 0.1\}$
PGD	$\epsilon \times 255$	$=$	$\{1, 4, 8, 10, 20, 30, 40, 50, 100, 200\}$
C-W	c	$=$	$\{10^{-5}, 10^{-4}, 10^{-2}, 10^{-1}, 1, 2, 3\}$

norm, as calculable by Equation 3. Beyond this, r_{50} is the median certified radii of the samples able to be successfully attacked by a given technique, and Time represents the median attack time (in seconds) across all tested samples. All three of these latter metrics demonstrate favourable performance with decreasing values.

This broad set of metrics was deliberately chosen to reflect different aspects of performance. However, we call particular attention to %-C, as it is a measure of the size of the adversarial examples *relative to the location of the minimal possible adversarial example*—with the certification of Cohen et al. essentially providing what is in essence characteristic scale that can be used for normalisation. We emphasise that such a measure of relative importance is important to further illuminate performance in light of the fact that the other metrics may not all strictly consider the same samples, as they are often constructed over the set of samples an attack method is successfully able to manipulate.

E. Samplewise Performance

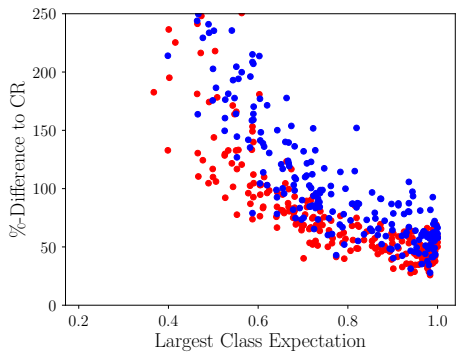
To further illuminate the nature of the performance of our attack, Figure 6a considers the sample-wise performance of both PGD and our Certification Aware Attack. Within this data there is a clear self-similar trend, in which the percentage difference to Equation 3 increases as the largest class expectation decreases. This difference could indicate the potential for improving the certification of samples within this region. There also appears to be a correlation between the outperformance of our approach and the semantic complexity of the prediction task, which suggests that tightening these guarantees could be increasingly relevant for complex datasets of academic and industry interest.

F. Accuracy of σ

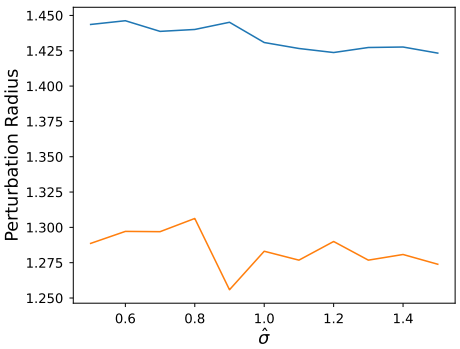
The white-box threat model assumes that the attacker has access to the full model and its parameters, including the level of additive noise σ. However, if the attacker only had access to the model and output class expectations, but was somehow prevented from directly accessing σ and r, it turns out that the Certification Aware Attack can still be applied subject to a sufficiently accurate guess of σ. As is shown by Figure 6b, even over-estimating σ by 50% can decrease the radius of the identified adversarial perturbation under certain experimental conditions. That this is possible is a product of the terms δ_1 and δ_2 in Algorithm 1, as both of these parameters set the idealised step size to try and either change or preserve the predicted class. While this does suggest that there is potentially additional scope for optimising δ_1 and δ_2, it also demonstrates the possibility of estimating σ as part of a surrogate model, in order to attack within a limited threat mode.

G. Training with MACER

Recent work has considered how certifications might be improved by augmenting the training objective to maximising the expectation gap between classes (Salman et al., 2019a).



(a) Percentage differences relative to E_0



(b) Performance for estimated σ

Figure 6: (a) captures the percentage difference between constructed adversarial perturbations and the certified radii of Equation 3 for CIFAR-10 at $\sigma = 0.5$, with Our technique in red and PGD in blue. (b) demonstrates that the blue mean and orange median performance of our technique are consistent even when $\sigma = 1.0$ is approximated by an estimated $\hat{\sigma}$.

A popular approach for this is MACER (Zhai et al., 2020), in which the training loss is augmented to incorporate what the authors dub the ϵ-robustness loss, which reflects proportion of training samples with robustness above a threshold level. In principle 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.

To test the performance of our new attack framework against models trained with MACER, Table 4 and Figure 7 recreate earlier results from within this work for CIFAR-10, subject to the same form of parameter exploration seen within Appendix C. We note that these calculations were performed with a ResNet-110 architecture, rather than the ResNet-18 architecture employed within the previous sections. While the broad qualitative feature of the success rates, best proportions, and median certifications broadly align with those seen within Table 2, we note that there is a significant difference in the %-C scores, which are a product of the ResNet-110 architecture (when trained under MACER) producing certifications that are an order of magnitude smaller than those observed within the main body of this work. That the attack radii are remaining constant while the certification radii decrease, strongly suggests that there would be significant scope for improving the performance of these results by varying the range of the parameter space exploration. One other notable feature is the improvement of DeepFool for MACER trained models, relative to the performance seen within the main body of this work, which we believe is a consequence of the changes in MACER's model decision space influencing the ability for DeepFool to converge upon successful evasion attacks.

H. Exemplar Attacks

The size of the associated adversarial perturbations has been established as a proxy of the risk of an adversarial attack evading human-or-machine-scrutiny (Gilmer et al., 2018). While considering metrics of performance are a more reliable measure of this adversarial risk, for completeness in Figure 8 provides a visual exemplar of the performance of both our attack and PGD. As both attacks share similar methodological features, the adversarial perturbations share similar semantic features, however our attack consistently requires smaller adversarial perturbations in order to trick the classifier—which in turn would have a higher probability of potentially evading any detection framework.

Table 4: CIFAR-10 attack performance across σ for a ResNet-110 architecture trained with MACER. Table features follow Table 4

σ	Type	Suc.↑	Best ↑	r_{50} ↓	%-C↓	Time ↓
0.25	Ours*	100%	76%	0.83	2308	9.66
	PGD*	100%	5%	1.03	2918	24.47
	C-W*	24%	0%	9.10	39952	24.57
	DeepF	100%	18%	1.32	3687	7.01
0.5	Ours*	77%	58%	1.09	2875	12.18
	PGD*	95%	18%	1.73	2294	24.74
	C-W*	43%	1%	11.35	19073	24.83
	DeepF	100%	23%	2.94	4377	7.58
1.0	Ours*	59%	43%	1.38	12654	14.06
	PGD*	98%	39%	2.86	3201	24.52
	C-W*	9%	0%	9.63	20597	24.60
	DeepF	100%	19%	5.29	5670	7.11

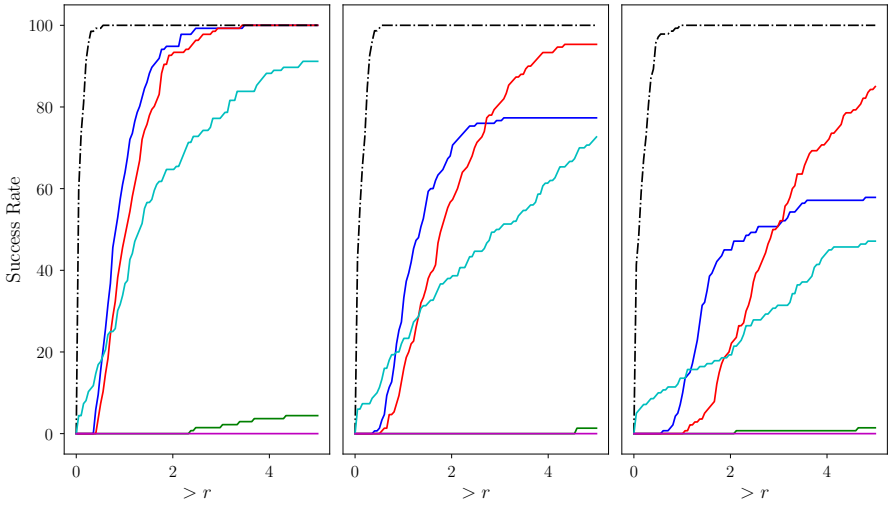


Figure 7: Attack and certification performance for a ResNet-110 model for CIFAR-10, when trained with MACER, covering our new Certification Aware Attack (blue), PGD (red), DeepFool (cyan), Carlini-Wagner (green), and AutoAttack (magenta). Similar to Figure 3, an ideal attack will approach the Cohen et al. (2019) radii suggested by the black dotted lines.

Figure 8: Illustrative examples of attack performance. Each column (from left to right) represents: the original image; the image under our attack; the adversarial perturbation associated with our attack; the image under PGD; the adversarial perturbation associated with PGD. The adversarial perturbations have been multiplied by 25 for visual clarity.