Scaleable input gradient regularization for adversarial robustness

Chris Finlay* **and Adam M Oberman†**

Department of Mathematics and Statistics
McGill University
Montréal QC, H3A 0B9

Abstract

Input gradient regularization is not thought to be an effective means for promoting adversarial robustness. In this work we revisit this regularization scheme with some new ingredients. First, we derive new per-image theoretical robustness bounds based on local gradient information, and curvature information when available. These bounds strongly motivate input gradient regularization. Second, we implement a scaleable version of input gradient regularization which avoids double backpropagation: adversarially robust ImageNet models are trained in 33 hours on four consumer grade GPUs. Finally, we show experimentally that input gradient regularization is competitive with adversarial training.

1 Introduction

Neural networks are vulnerable to *adversarial attacks*. These are small (imperceptible to the human eye) perturbations of an image which cause a network to misclassify the image [2, 15, 43]. The threat posed by adversarial attacks must be addressed before these methods can be deployed in error-sensitive and security-based applications [31].

Building adversarially robust models is an optimization problem with two objectives: (i) maintain test accuracy on clean unperturbed images, and (ii) be robust to large adversarial perturbations. The present state-of-the-art method for adversarial defence, adversarial training [15, 24, 25, 43, 44], in which models are trained on perturbed images, offers robustness at the expense of test accuracy [45]. It is not clear that multi-step adversarial training is scaleable to large datasets such as ImageNet-1k [10]. Previous attempts [20, 53] used hundreds of GPUs and took nearly a week to train, although recent work by Shafahi et al. [40] has offered a remedy.

Assessing the *empirical* effectiveness of an adversarial defence requires careful testing with multiple attacks [14]. Furthermore, existing defences are vulnerable to new, stronger attacks: Carlini and Wagner [5] and Athalye et al. [1] advocate designing specialized attacks to circumvent prior defences, while Uesato et al. [47] warn against using weak attacks to evaluate robustness. This has led the community to develop *theoretical* tools to certify adversarial robustness. Several certification approaches have been proposed: through linear programming [50, 51] or mixed-integer linear-programming [52]; semi-definite relaxation [32, 33]; randomized smoothing [8, 23]; or estimates of the local Lipschitz constant [17, 46, 49]. The latter two approaches have scaled well to ImageNet-1k.

In practice, certifiably robust networks often perform worse than adversarially trained models, which lack theoretical guarantees. In this article, we work towards bridging the gap between theoretically

*christopher.finlay@mail.mcgill.ca
†adam.oberman@mcgill.ca

Preprint. Under review.

robust networks and empirically effective training methods. Our approach relies on minimizing a loss regularized against large input gradients

$$\mathop{\mathbb{E}}_{(x,y)\sim\mathbb{P}}\left[\mathcal{L}(f(x;w),y)+\frac{\lambda}{2}\|\nabla_x\mathcal{L}(f(x;w),y)\|_*^2\right]$$

where $\|\cdot\|_*$ is dual to the one measuring adversarial attacks (for example the ℓ_1 norm for attacks measured in the ℓ_∞ norm). Heuristically, making loss gradients small should make gradient based attacks more challenging.

Drucker and LeCun [11] implemented gradient regularization using 'double backpropagation', which has been shown to improve model generalization [28]. It has been used to improve the stability of GANs [27, 36] and to promote learning robust features with contractive auto-encoders [34]. While it has been proposed for adversarial attacks robustness [17, 19, 35, 37, 42], experimental evidence has been mixed, in particular, input gradient regularization has so far not been competitive with multi-step adversarial training.

On non-smooth networks (such as those built of ReLUs) small gradients are no guarantee of adversarial robustness [30], and so it is thought input gradient regularization should not be effective on non-smooth networks. This raises the question, how often is the lack of smoothness an issue, in practice? In other words, when do Taylor approximations of the loss fail to predict adversarial robustness, and is smoothness only needed theoretically? The fact that first-order gradient-based attacks of the loss (like PGD [24]) are usually effective indicates that in many scenarios, non-smoothness is not an issue. However in a non-negligible minority of cases, attacks based on decision boundary information [4, 6, 7, 12] outperform gradient based attacks. This indicates the curvature near these points is large, and first-order information is not sufficient to guarantee robustness. We illustrate this point in Fig 1. In this work we overcome the limitation of gradient regularization for non-smooth networks by instead building networks of 'smooth' ReLUs. At the expense of a minor drop in test accuracy, we obtain tighter theoretical lower bounds on robustness, since we can better approximate the loss using local information.

Another drawback of input gradient regularization is that it is not presently tractable to update model weights using double backpropagation on large networks. We circumvent this limitation by differentiating the regularization term without double backpropagation.

Our main contributions are the following. First, we motivate using input gradient regularization *of the loss* by deriving new theoretical robustness bounds. These bounds show that small loss gradients and small curvature are sufficient conditions for adversarial robustness. Second, we empirically show that input gradient regularization is competitive with adversarial training, even on non-smooth networks, at a fraction of the training time. Finally, we scale input gradient regularization to ImageNet-1k by using finite differences to estimate the gradient regularization term, rather than double backpropagation. This allows us to train adversarially robust networks on ImageNet-1k in 33 hours on four consumer grade GPUs.

2 Adversarial robustness bounds from the loss

2.1 Background

Much effort has been directed towards determining theoretical lower bounds on the minimum sized perturbation necessary to perturb an image so that it is misclassified by a model. One promising approach, proposed by Hein and Andriushchenko [17] and Weng et al. [49], and which has scaled well to ImageNet-1k, is to use the Lipschitz constant of the model. In this section, we build upon these ideas: we propose using the Lipschitz constant of a suitable loss, designed to measure classification errors. In addition, when the loss is twice continuously differentiable, we propose a second-order bound based on the maximum curvature of the loss.

Our notation is as follows. Write $y = f(x;w)$ for a model which takes input vectors x to label probabilities, with parameters w. Let $\mathcal{L}(y_1,y_2)$ be the loss and write $\ell(x) := \mathcal{L}(f(x,w),y)$, for the loss of a model f.

Finding an adversarial perturbation is interpreted as a global minimization problem: find the closest image to a clean image, in some specified norm, that is is also misclassified by the model

$$\min_v \|v\| \quad \text{subject to } f(x+v) \text{ misclassified} \tag{1}$$

However, (1) is a difficult and costly non-smooth, non-convex optimization problem. Instead, Goodfellow et al. [15] proposed solving a surrogate problem: find a perturbation v of a clean image x that maximizes the loss, subject to the condition that the perturbation be inside a norm-ball of radius δ around the clean image. The surrogate problem is written

$$\max_v \ell(x+v) - c(v); \quad \text{where } c(v) = \begin{cases} 0 & \text{if } \|v\| \leq \delta \\ \infty & \text{otherwise} \end{cases} \tag{2}$$

The hard constraint $c(v)$ forces perturbations to be inside the norm-ball centred at the clean image x. Ideally, solutions of this surrogate problem (2) will closely align with solutions of the original more difficult global minimization problem. However, the hard constraint in (2) forces a particular scale: it may miss attacks which would succeed with only a slightly bigger norm. Additionally, the maximization problem (2) does not force misclassification, it only asks that the loss be increased.

The advantage of (2) is that it may be solved with gradient-based methods: present best-practice is to use variants of projected gradient descent (PGD), such as the iterative fast-signed gradient method [22, 24] when attacks are measured in the ℓ_∞ norm. However, gradient-based methods are not always effective: on non-smooth networks, such as those built of ReLU activation functions, a small gradient does not guarantee that the loss remains small locally. This deficiency was identified in [29]. See Figure 1: ReLU networks may increase rapidly with a very small perturbation, even when local gradients are small. PGD methods will fail to locate these worst-case perturbations, and give a false impression of robustness. Carlini and Wagner [6] avoid this scenario by incorporating decision boundary information into the loss; others solve (1) directly [4, 7, 12].

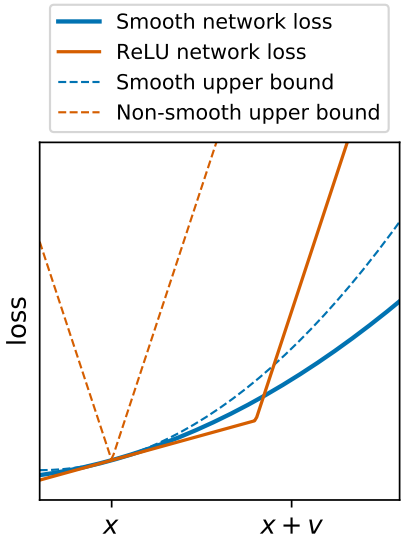


Figure 1: Illustration of upper bounds on the loss of two networks. For smooth networks (blue) with finite curvature, the loss is bounded above using $\ell(x)$ and $\nabla_x \ell(x)$. Non-smooth networks (orange) may have jumps in their gradients, which means robustness is not guaranteed by small local gradients.

2.2 Derivation of lower bounds

This leads us to consider the following compromise between (1) and (2). Consider the following modification of the Carlini and Wagner [6] loss $\ell(x) = \max_{i \neq c} f_i(x) - f_c(x)$, where c is the index of the correct label, and $f_i(x)$ is the model output for the i-th label. This loss has the appealing property the sign of the loss determines if the classification is correct. Adversarial attacks are found by minimizing

$$\min_v \|v\| \quad \text{subject to } \ell(x+v) \geq \ell_0 \tag{3}$$

The constant ℓ_0 determines when classification is incorrect; for the modified Carlini-Wagner loss, $\ell_0 = 0$. Problem (3) is closer to the true problem (1), and will always find an adversarial image. We use (3) to derive theoretical lower bounds on the minimum size perturbation necessary to misclassify an image. Suppose the loss is L-Lipschitz with respect to model input. Then we have the estimate

$$\ell(x+v) \leq \ell(x) + L\|v\| \tag{4}$$

Now suppose v *is* adversarial, with minimum adversarial loss $\ell(x+v) = \ell_0$. Then rearranging (4), we obtain the lower bound $\|v\| \geq \frac{1}{L}\left(\ell_0 - \ell(x)\right)$.

Unfortunately, the Lipschitz constant is a global quantity, and ignores local gradient information; see for example Huster et al. [18]. Thus this bound can be quite poor, even when networks have small Lipschitz constant. On the other hand, if the model is twice continuously differentiable, then the loss landscape is smoother. This allows us to achieve a tighter bound, using local gradient information, as illustrated in Figure 1. Let C be an upper bound on the maximum positive eigenvalue of the Hessian of the loss over all x

$$C := \left(\max_x \lambda_{\max}(\nabla_x^2 \ell(x))\right)^+ \tag{5}$$

This value will be estimated empirically by maximizing over the dataset. The constant C is a measure of the largest positive curvature of the network. Using a Taylor approximation about x, we may upper bound the perturbed loss with

$$\ell(x+v) \leq \ell(x) + \langle v, \nabla_x \ell(x) \rangle + \frac{C}{2}\|v\|_2^2 \tag{6}$$

These two bounds give us the following.

Proposition 2.1. *Suppose the loss $\ell(x)$ is Lipschitz continuous with respect to model input x, with Lipschitz constant L. Let ℓ_0 be such that if $\ell(x) < \ell_0$, the model is always correct. Then a lower bound on the minimum magnitude of perturbation v necessary to adversarially perturb an image x is*

$$\|v\| \geq \frac{\max\{\ell_0 - \ell(x), 0\}}{L} \tag{L-bound}$$

Suppose in addition that the loss is twice-differentiable, with maximum curvature C (defined as in (5)). Then

$$\|v\|_2 \geq \frac{1}{C}\left(-\|\nabla \ell(x)\|_2 + \sqrt{\|\nabla \ell(x)\|_2^2 + 2C\max\{\ell_0 - \ell(x), 0\}}\right) \tag{C-bound}$$

The proof of (L-bound) is given above; the proof of (C-bound) follows by rearranging (6) and solving for $\|v\|$.

Remark 2.2. The second-order bound requires that the network and loss are smooth with respect to the input, but almost all image classification networks now use ReLUs, which are not smooth. We use the following smoothed ReLU

$$\sigma(x) = \begin{cases} \max(x, 0) & \text{if } |x| \geq \frac{1}{2} \\ -\frac{1}{2}\left(x + \frac{1}{2}\right)^4 + \left(x + \frac{1}{2}\right)^3 & \text{if } |x| < \frac{1}{2} \end{cases}$$

This activation function is twice continuously differentiable, and avoids the vanishing gradient problem of smooth sigmoidal activation functions. Moreover because it agrees with $\mathrm{ReLU}(x)$ outside of the interval $(-\frac{1}{2}, \frac{1}{2})$, it is fairly efficient during backpropagation. As for the loss, a smooth version of the Carlini-Wagner loss is available by using a soft maximum, rather than a strict $\max$.

Proposition 2.1 motivates the need for input gradient regularization. The Lipschitz constant L is the maximum gradient norm of *the loss* over all inputs. Therefore (L-bound) says that a regularization term encouraging small gradients (and so reducing L) should increase the minimum adversarial distance. This aligns with [17], who proposed the cross-Lipschitz regularizer, penalizing networks with large Jacobians in order to shrink the Lipschitz constant of *the network*.

However, this is not enough: the gap $\ell_0 - \ell(x)$ must be large as well. This explains one form of 'gradient masking' [30]. Shrinking the magnitude of gradients while also closing the gap $\ell_0 - \ell(x)$ effectively does nothing to improve adversarial robustness. For example, in defense distillation, the magnitude of the model Jacobian is reduced by increasing the temperature of the final softmax layer of the network. However, this has the detrimental side-effect of sending the model output to $(\frac{1}{N}, \cdots, \frac{1}{N})$, where N is the number of classes, which effectively shrinks the loss gap to zero. Thus with high distillation temperatures the lower bound provided by Proposition 2.1 approaches zero.

Moreover, even supposing the loss gradients are small and the gap $\ell_0 - \ell(x)$ is large, there may still be adversarially vulnerable images. For example, suppose we have two smooth networks, one with large curvature, and another with small curvature. Suppose that there is an image with zero gradient on both networks, each with identically large loss gaps $\ell_0 - \ell(x)$. The second-order bound (C-bound) says that the minimum adversarial distance here is bounded below by $\|v\| \geq \sqrt{\max\{\ell_0 - \ell(x), 0\}/C}$. In other words, the network with smaller curvature is more robust.

Taken together, Proposition 2.1 provides three sufficient conditions for training robust networks: (i) the loss gap $\ell_0 - \ell(x)$ should be large; (ii) the gradients of the loss should be small; and (iii) the curvature of the loss should also be small. The first point will be satisfied by default when the loss is minimized. The second point will be satisfied by training with a loss regularized to penalize large input gradients. Experimentally the third point is satisfied with input gradient regularization. When these conditions are satisfied, local information is enough to guarantee robustness.

Our robustness bounds are most similar in spirit to Weng et al. [49], who derive bounds using an estimate of the *local* Lipschitz constant of the model. Moosavi-Dezfooli et al. [26] have also used a second order approximation to derive approximate robustness bounds for binary classification, but they neglected higher order error terms. Cohen et al. [8] derive bounds by training with normally distributed input noise, then averaging model predictions normally sampled about the input image. It is well known that training with normal noise is equivalent to squared ℓ_2 norm gradient regularization [3]; thus Cohen et al. [8] achieve gradient regularization indirectly. Our bounds require at most one gradient and model evaluation per image once L and C have been estimated; whereas both Cohen et al. and Weng et al. require many hundreds of local model evaluations per image. Since L and C are globally estimated, our bounds could be improved using these local sampling techniques to obtain *local* values of L and C, with more computational effort.

3 Squared norm gradient regularization

Proposition 2.1 provides strong motivation for input gradient regularization as a method for promoting adversarial robustness. However, it does not tell us what form the gradient regularization term should take. In this section, we show how norm squared gradient regularization arises from a *quadratic cost*.

In adversarial training, solutions of (2) are used to generate images on which the network is trained. In effect, adversarial training seeks a solution of the minimax problem

$$\min_w \mathop{\mathbb{E}}_{x\sim\mathbb{P}} \left[\max_v \ell(x+v;w) - c(v) \right] \tag{7}$$

where $\mathbb{P}$ is the distribution of images. This is a robust optimization problem [38, 48]. The cost function $c(v)$ penalizes perturbed images from being too far from the original. When the cost function is the hard constraint from (2), perturbations must be inside a norm ball of radius δ. This leads to adversarial training with PGD [22, 24]. However this forces a particular scale: it is possible that no images are adversarial within radius δ, but that there are adversarial images with only a slightly larger distance. Instead of using a hard constraint, we can relax the cost function to be the quadratic cost $c(v) = \frac{1}{2\delta}\|v\|^2$. The quadratic cost allows attacks to be of any size, but penalizes larger attacks more than smaller attacks. With a quadratic cost, there is less of a danger that a local attack will be overlooked.

Solving (7) directly is expensive: on ImageNet-1k, both Kannan et al. [20] and Xie et al. [53] required large-scale distributed training with many dozens or hundreds of GPUs, and over a week of training time. Instead we take the view that (7) may be bounded above, and solved approximately. When the loss is smooth and $c(v) = \frac{1}{2\delta}\|v\|^2$, the optimal value of $\max_v \ell(x+v) - c(v)$ using the bound (6) is $\frac{\delta}{2(1-\delta C)}\|\nabla_x \ell(x)\|_*^2$, provided $\delta < \frac{1}{C}$. This gives the following proposition.

Proposition 3.1. *Suppose both the model and the loss are twice continuously differentiable. Suppose attacks are measured with quadratic cost $\frac{1}{2\delta}\|v\|^2$. Then the optimal value of (7) is bounded above by*

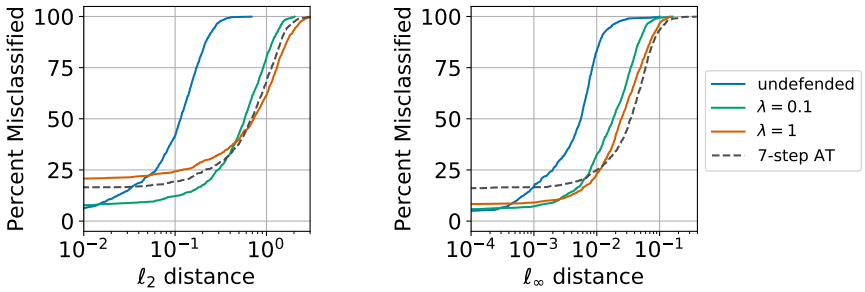
$$\min_w \mathop{\mathbb{E}}_{x\sim\mathbb{P}} \left[\ell(x;w) + \frac{\lambda}{2}\|\nabla_x \ell(x)\|_*^2 \right] \tag{8}$$

where $\lambda = \frac{\delta}{1-\delta C}$.

That is, we may bound the solution of the adversarial training problem (7) by solving the gradient regularization problem (8), when the cost function is quadratic. It is not necessary to know δ or compute C; they are absorbed into λ. In the adversarial robustness literature, input gradient regularization using the squared ℓ_2 norm was proposed by Ross and Doshi-Velez [35]. It was expanded by Roth et al. [37] to use a Mahalanobis norm with the correlation matrix of adversarial attacks. When $c(v)$ is the hard constraint forcing attacks inside the δ norm ball and C is small, supposing the curvature term is negligible, we can estimate the maximum in (7) by $\ell(x) + \frac{1}{\delta}\|\nabla_x \ell(x)\|_*$, using the dual norm for the gradient. This is norm gradient regularization (not squared), and was recently used for adversarial robustness on both CIFAR-10 [42], and MNIST [39].

3.1 Finite difference implementation

Norm squared input gradient regularization has long been used as a regularizer in neural networks: Drucker and LeCun [11] first showed its effectiveness for generalization. Drucker and LeCun



(a) ℓ_2 norm adversarial attacks (b) ℓ_∞ norm adversarial attacks

Figure 2: Adversarial attacks on the CIFAR-10 dataset, on networks built with standard ReLUs. Regularized networks attacked in ℓ_2 are trained with squared ℓ_2 norm gradient regularization; networks attacked in ℓ_∞ are trained with squared ℓ_1 norm regularization.

implemented gradient regularization with 'double backpropagation' to compute the derivatives of the penalty term with respect to the model parameters w, which is needed to update the parameters during training. Double backpropagation involves two passes of automatic differentiation: one pass to compute the gradient of the loss with respect to the inputs x, and another pass on the output of the first to compute the gradient of the penalty term with respect to model parameters w. In neural networks, double backpropagation is the standard technique for computing the parameter gradient of a regularized loss. However, it is not currently scaleable to large neural networks. Instead we approximate the gradient regularization term with finite differences.

Proposition 3.2 (Finite difference approximation of squared ℓ_2 gradient norm)**.** *Let d be the normalized input gradient direction: $d = \nabla_x \ell(x) / \| \nabla_x \ell(x) \|_2$ when the gradient is nonzero, and set $d = 0$ otherwise. Let h be the finite difference step size. Assume further that the loss is twice continuously differentiable. Then, the squared ℓ_2 gradient norm is approximated by*

$$\| \nabla_x \ell(x) \|_2^2 \approx \left(\frac{\ell(x + hd) - \ell(x)}{h} \right)^2 \tag{9}$$

The vector d is normalized to ensure the accuracy of the finite difference approximation, which is of order h, as can be seen by a Taylor approximation. The finite differences approximation (9) allows the computation of the gradient of the regularizer (with respect to model parameters w) to be done with only two regular passes of backpropagation, rather than with double backpropagation. On the first, the input gradient direction d is calculated. The second computes the gradient with respect to model parameters by performing backpropagation on the right-hand-side of (9). Double backpropagation is avoided by detaching d from the computational graph after the first pass. In practice, for large networks, we have found that the finite difference approximation of the regularization term is considerably more efficient than using double backpropagation.

The proposed training algorithm, with squared Euclidean input gradient regularization, is presented in Algorithm 1 of the appendix. Other gradient penalty terms can be approximated as well. For example, when defending against attacks measured in the ℓ_∞ norm, the squared ℓ_1 norm penalty can approximated by setting instead $d = \text{sign}(\nabla_x \ell(x)) / \sqrt{N}$ when the gradient is nonzero.

4 Experimental results

In this section we provide empirical evidence that input gradient regularization is an effective tool for promoting adversarial robustness, *even on non-smooth networks* built with standard ReLU activation functions.

We train networks on the CIFAR-10 dataset [21], and ImageNet-1k [10]. On the CIFAR dataset we use the ResNeXt architecture[3] [54]; on ImageNet-1k we use a ResNet-50 [16]. The CIFAR networks were trained with standard data augmentation and learning rate schedules on a single GeForce GTI 1080 Ti.

[3]ResNeXt34-2x32 on CIFAR-10; ResNeXt34-2x64 on CIFAR-100

Table 1: Adversarial robustness statistics, measured in the ℓ_∞ norm. Top1 error is reported on CIFAR-10; Top5 error on ImageNet-1k.

	smooth ReLU?	% clean error	% error at $\varepsilon = \frac{2}{255}$	% error at $\varepsilon = \frac{8}{255}$	mean distance	improvement ratio	training time (hours)
CIFAR-10							
Undefended		**4.36**	70.82	98.94	6.62e−3	-	2.06
Madry et al (7-step AT)		16.33	22.86	**46.02**[5]	**4.07e−2**	1.88	12.10
squared ℓ_1 norm, $\lambda = 0.1$		6.45	24.92	70.41	2.35e−2	**5.31**	5.22
squared ℓ_1 norm, $\lambda = 1$		9.02	**18.47**	58.69	3.34e−2	3.78	5.15
ImageNet-1k							
Undefended		**6.94**	90.21	98.94	3.94e−3	-	20.30
Undefended	✓	9.39	82.03	95.42	9.74e−3	4.17	23.46
squared ℓ_2 norm, $\lambda = 0.1$		7.66	70.56	97.53	7.96e−3	**9.83**	32.60
squared ℓ_2 norm, $\lambda = 0.1$	✓	9.49	63.23	**94.21**	**1.24e−2**	5.84	52.47
squared ℓ_2 norm, $\lambda = 1$		10.26	**52.79**	95.93	9.95e−3	3.19	33.87

On ImageNet-1k, we modified the training code of Shaw et al.'s [41] submission to the DAWNBench competition [9] and train with four GPUs. Training code and trained model weights are available on GitHub.[4]

We train an undefended network as a baseline to compare various types of regularization. On CIFAR-10, networks are trained with squared ℓ_2 and squared ℓ_1 gradient norm regularization. The former is appropriate for defending against attacks measured in ℓ_2; the latter for attacks measured in ℓ_∞. We set the regularization strength to be either $\lambda = 0.1$ or 1; and set finite difference discretization $h = 0.01$. We compare each network with the current state-of-the-art form of adversarial training, with models trained using the hyperparameters in Madry et al. [24] (7-steps of FGSM, ℓ_∞ step size $\frac{2}{255}$, projected onto an ℓ_∞ ball of radius $\frac{8}{255}$). On ImageNet-1k we only train adversarially robust models with squared ℓ_2 regularization.

On each dataset, we attack 1000 randomly selected images. We perturb each image with attacks in both the Euclidean and ℓ_∞ norms, with a suite of current state-of-the-art attacks: the Carlini-Wagner attack [6]; the Boundary attack [4]; the LogBarrier attack [12]; and PGD [24] (in both the ℓ_∞ norm or the ℓ_2 norm). The former three attacks are effective at evading gradient masking defences; the latter is very good at finding images close to the original when gradients are not close to zero. We record the best adversarial distance *on a per image basis*, for each norm.

Adversarial robustness results for networks attacked in the ℓ_∞ norm are presented in Table 1. These results are for networks built of standard ReLUs. Table 1 and Figure 2 demonstrate a clear trade-off between test accuracy and adversarial robustness, as the strength of the regularization is increased. On CIFAR-10, the undefended network achieves test error of 4.36%, but is not robust to attacks even at ℓ_∞ distance $\frac{2}{255}$. However with a strong regularization parameter ($\lambda = 1$), test error increases to 9.02% on clean images, and only 18.47% test error at attack distance $\frac{2}{255}$. In contrast, the network trained with 7-steps of adversarial training appears to be over-regularized: on clean images, the adversarially trained network achieves 16.33% test error, but 22.86% error at distance $\frac{2}{255}$. To be fair, at the commonly reported ℓ_∞ of $\frac{8}{255}$, the adversarially trained network outperforms the best gradient regularized networks by about 12%, but at over twice the training time of the regularized networks. On ImageNet, we see a reduction of nearly 40% at distance $\frac{2}{255}$.

It has been noted that adversarial robustness comes with a cost of degraded test error [45]. This trade-off may be quantified. We measure the relative improvement in adversarial robustness against the cost of degraded test error with the following metric. Suppose an undefended network has test error e_0, and let a regularized network's network test error be denoted e_λ. Define the relative degradation in test error to be $R_e = (e_\lambda - e_0)/e_0$. Similarly define the relative improvement in robustness (measured by mean adversarial distance μ) to be $R_\mu = (\mu_\lambda - \mu_0)/\mu_0$. We define the *adversarial improvement ratio* to be R_μ/R_e. This measures the improvement in adversarial robustness against

[4] https://github.com/cfinlay/tulip

[5] Madry et al report 54.2% error at $\varepsilon = \frac{8}{255}$ with the WRN-28x10 architecture; our results are obtained with ResNeXt34 (2x32).

Table 2: Adversarial robustness statistics, measured in ℓ_2. Top1 error is reported on CIFAR-10; Top5 error on ImageNet-1k.

	smooth ReLU?	% clean error	mean adversarial distance			improvement ratio	training time (hours)
			L-bound	C-bound	empirical		
CIFAR-10							
Undefended		**4.36**	5.57e−3	-	0.12	-	2.06
Undefended	✓	6.84	1.01e−2	1.19e−2	0.11	−0.20	3.78
Madry et al (7-step AT)		16.33	0.18	-	0.74	1.81	12.10
squared ℓ_2 norm, $\lambda = 0.1$		8.03	0.14	-	0.63	**4.86**	5.18
squared ℓ_2 norm, $\lambda = 0.1$	✓	11.68	0.13	0.17	0.59	2.25	9.46
squared ℓ_2 norm, $\lambda = 1$		20.31	**0.30**	-	**0.81**	1.52	5.08
ImageNet-1k							
Undefended		**6.94**	3.63e−2	-	0.55	-	20.30
Undefended	✓	9.39	2.56e−2	3.40e−2	0.56	0.12	23.46
squared ℓ_2 norm, $\lambda = 0.1$		7.66	0.13	-	1.14	**10.23**	32.60
squared ℓ_2 norm, $\lambda = 0.1$	✓	9.49	9.23e−2	7.52e−2	1.09	2.64	52.47
squared ℓ_2 norm, $\lambda = 1$		10.26	**0.26**	-	**1.75**	4.52	33.87

the expense of poorer test error: high values mean the defended model is much more robust and has not lost significant test accuracy. Values close to zero imply the model is more robust but has a much worse test accuracy relative to the undefended model. The improvement ratio is non-dimensional, and so it allows for comparison between datasets.

Measured in this metric, the tradeoff between test accuracy and adversarial robustness is clear. On both ImageNet-1k and CIFAR-10, models regularized with $\lambda = 0.1$ offer the best trade-off between robustness and test error. If test accuracy is not of foremost concern, then stronger regularization parameters may be chosen. If neither training time nor test accuracy are important factors, then adversarial training is competitive with gradient regularization.

In Table 2 we report results on models trained for attacks in the ℓ_2 norm. On CIFAR-10, the most robust model is trained with regularization strength $\lambda = 1$, and outperforms even the adversarially trained model. On ImageNet-1k, we see the same pattern: the model trained with $\lambda = 1$ offers the best protection against adversarial attacks. Due to the long training time, we were not able to train ImageNet-1k with multi-step adversarial training.

In Table 2 we also report our theoretical bounds on the minimum distance required to adversarially perturb, using the Carlini-Wagner loss.[6] Figures 4 and 5 of the appendix show these bounds on a per-image basis. The theoretical bounds require calculating constants L and C, which are not readily available. Instead, we estimate L as the maximum gradient norm over test images; for smooth models we estimate C as the maximum spectral norm of the Hessian.[7] These estimates are reported in Table 3 of the appendix. Gradient regularization reduces L and C, by *one to two orders of magnitude*. Table 3 shows adversarial training also reduces L: effectively adversarial training is a regularizer. Because L and C are estimated, and not exact, one would expect that our bounds would sometimes fail. However, on CIFAR-10, the bounds reliable held on all attacked images. On ImageNet-1k, the bounds failed on about 9% of attacked test images, which indicates that C and L could be estimated more accurately, for example using by estimating these constants locally like in [49].

5 Conclusion

We have provided motivation for training adversarially robust networks through input gradient regularization, by bounding the minimum adversarial distance with gradient and curvature statistics of the loss. We have shown empirically that gradient regularization is scaleable to ImageNet-1k, and provides adversarial robustness competitive with adversarial training. We gave theoretical per-image bounds on the minimum adversarial distance, for non-smooth models (using the Lipschitz constant of

[6] This loss can be modified for Top-5 mis-classification as well.

[7] We compute the spectral norm of the Hessian using the Lanczos algorithm [13, §10.1] on Hessian-vector products (computed via automatic differentiation).

the loss), and augmented these bounds using smooth models with a second-order bound based on model curvature. These bounds were empirically validated against state-of-the-art attacks.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/athalye18a.html.

[2] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 387–402, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-40994-3.

[3] Christopher M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995. doi: 10.1162/neco.1995.7.1.108. URL https://doi.org/10.1162/neco.1995.7.1.108.

[4] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=SyZI0GWCZ.

[5] Nicholas Carlini and David A. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec@CCS 2017, Dallas, TX, USA, November 3, 2017*, pages 3–14, 2017. doi: 10.1145/3128572.3140444. URL https://doi.org/10.1145/3128572.3140444.

[6] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017. URL https://doi.org/10.1109/SP.2017.49.

[7] Jianbo Chen and Michael I. Jordan. Boundary attack++: Query-efficient decision-based adversarial attack. *CoRR*, abs/1904.02144, 2019. URL http://arxiv.org/abs/1904.02144.

[8] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. *CoRR*, abs/1902.02918, 2019. URL http://arxiv.org/abs/1902.02918.

[9] Cody Coleman, Daniel Kang, Deepak Narayanan, Luigi Nardi, Tian Zhao, Jian Zhang, Peter Bailis, Kunle Olukotun, Christopher Ré, and Matei Zaharia. Analysis of dawnbench, a time-to-accuracy machine learning performance benchmark. *CoRR*, abs/1806.01427, 2018. URL http://arxiv.org/abs/1806.01427.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255, 2009. doi: 10.1109/CVPRW.2009.5206848. URL https://doi.org/10.1109/CVPRW.2009.5206848.

[11] Harris Drucker and Yann LeCun. Double backpropagation increasing generalization performance. In *IJCNN-91-Seattle International Joint Conference on Neural Networks*, volume 2, pages 145–150. IEEE, 1991.

[12] Chris Finlay, Aram-Alexandre Pooladian, and Adam M. Oberman. The LogBarrier adversarial attack: making effective use of decision boundary information. *CoRR*, abs/1903.10396, 2019. URL http://arxiv.org/abs/1903.10396.

[13] Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2012.

[14] Ian Goodfellow, Patrick McDaniel, and Nicolas Papernot. Making machine learning robust against adversarial inputs. *Communications of the ACM*, 61(7):56–66, June 2018. URL http://dl.acm.org/citation.cfm?doid=3234519.3134599.

[15] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2014. URL http://arxiv.org/abs/1412.6572.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pages 630–645, 2016. URL https://doi.org/10.1007/978-3-319-46493-0_38.

[17] Matthias Hein and Maksym Andriushchenko. Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2263–2273, 2017. URL http://papers.nips.cc/paper/6821-formal-guarantees-on-the-robustness-of-a-classifier-against-adversarial-manipulation.

[18] Todd Huster, Cho-Yu Jason Chiang, and Ritu Chadha. Limitations of the lipschitz constant as a defense against adversarial examples. In *ECML PKDD 2018 Workshops - Nemesis 2018, UrbReas 2018, SoGood 2018, IWAISe 2018, and Green Data Mining 2018, Dublin, Ireland, September 10-14, 2018, Proceedings*, pages 16–29, 2018. doi: 10.1007/978-3-030-13453-2_2. URL https://doi.org/10.1007/978-3-030-13453-2_2.

[19] Daniel Jakubovitz and Raja Giryes. Improving DNN robustness to adversarial attacks using jacobian regularization. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XII*, pages 525–541, 2018. doi: 10.1007/978-3-030-01258-8_32. URL https://doi.org/10.1007/978-3-030-01258-8_32.

[20] Harini Kannan, Alexey Kurakin, and Ian J. Goodfellow. Adversarial logit pairing. *CoRR*, abs/1803.06373, 2018. URL http://arxiv.org/abs/1803.06373.

[21] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

[22] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *CoRR*, abs/1607.02533, 2016. URL http://arxiv.org/abs/1607.02533.

[23] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Second-order adversarial attack and certifiable robustness. *CoRR*, abs/1809.03113, 2018. URL http://arxiv.org/abs/1809.03113.

[24] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017. URL http://arxiv.org/abs/1706.06083.

[25] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[26] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. *CoRR*, abs/1811.09716, 2018. URL http://arxiv.org/abs/1811.09716.

[27] Vaishnavh Nagarajan and J. Zico Kolter. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5591–5600, 2017. URL http://papers.nips.cc/paper/7142-gradient-descent-gan-optimization-is-locally-stable.

[28] Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=HJC2SzZCW.

[29] Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pages 372–387, 2016. URL https://doi.org/10.1109/EuroSP.2016.36.

[30] Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017*, pages 506–519, 2017. URL http://doi.acm.org/10.1145/3052973.3053009.

[31] Richard Potember. Perspectives on research in artificial intelligence and artificial general intelligence relevant to DoD. Technical report, The MITRE Corporation McLean United States, 2017. URL https://fas.org/irp/agency/dod/jason/ai-dod.pdf.

[32] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=Bys4ob-Rb.

[33] Aditi Raghunathan, Jacob Steinhardt, and Percy S. Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 10900–10910, 2018. URL http://papers.nips.cc/paper/8285-semidefinite-relaxations-for-certifying-robustness-to-adversarial-examples.

[34] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, pages 833–840, 2011. URL https://icml.cc/2011/papers/455_icmlpaper.pdf.

[35] Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1660–1669, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17337.

[36] Kevin Roth, Aurélien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Stabilizing training of generative adversarial networks through regularization. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2015–2025, 2017. URL http://papers.nips.cc/paper/6797-stabilizing-training-of-generative-adversarial-networks-through-regularization.

[37] Kevin Roth, Aurélien Lucchi, Sebastian Nowozin, and Thomas Hofmann. Adversarially robust training through structured gradient regularization. *CoRR*, abs/1805.08736, 2018. URL http://arxiv.org/abs/1805.08736.

[38] Peter J Rousseeuw and Annick M Leroy. *Robust regression and outlier detection*, volume 1. Wiley Online Library, 1987.

[39] Ismaïla Seck, Gaëlle Loosli, and Stephane Canu. L1-norm double backpropagation adversarial defense. *arXiv preprint arXiv:1903.01715*, 2019.

[40] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *CoRR*, abs/1904.12843, 2019. URL http://arxiv.org/abs/1904.12843.

[41] Andrew Shaw, Yaroslav Bulatov, and Jeremy Howard. ImageNet in 18 minutes. URL https://github.com/diux-dev/imagenet18.

[42] Carl-Johann Simon-Gabriel, Yann Ollivier, Bernhard Schölkopf, Léon Bottou, and David Lopez-Paz. Adversarial vulnerability of neural networks increases with input dimension. *CoRR*, abs/1802.01421, 2018. URL http://arxiv.org/abs/1802.01421.

[43] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. URL http://arxiv.org/abs/1312.6199.

[44] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

[45] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *CoRR*, abs/1805.12152, 2018. URL http://arxiv.org/abs/1805.12152.

[46] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 6542–6551, 2018. URL http://papers.nips.cc/paper/7889-lipschitz-margin-training-scalable-certification-of-perturbation-invariance-for-deep-neural-networks.

[47] Jonathan Uesato, Brendan O'Donoghue, Pushmeet Kohli, and Aäron van den Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5032–5041, 2018. URL http://proceedings.mlr.press/v80/uesato18a.html.

[48] Abraham Wald. Statistical decision functions which minimize the maximum risk. *Annals of Mathematics*, pages 265–280, 1945.

[49] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=BkUHlMZ0b.

[50] Eric Wong and J. Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5283–5292, 2018. URL http://proceedings.mlr.press/v80/wong18a.html.

[51] Eric Wong, Frank R. Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada.*, pages 8410–8419, 2018. URL http://papers.nips.cc/paper/8060-scaling-provable-adversarial-defenses.

[52] Kai Y. Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing relu stability. *CoRR*, abs/1809.03008, 2018. URL http://arxiv.org/abs/1809.03008.

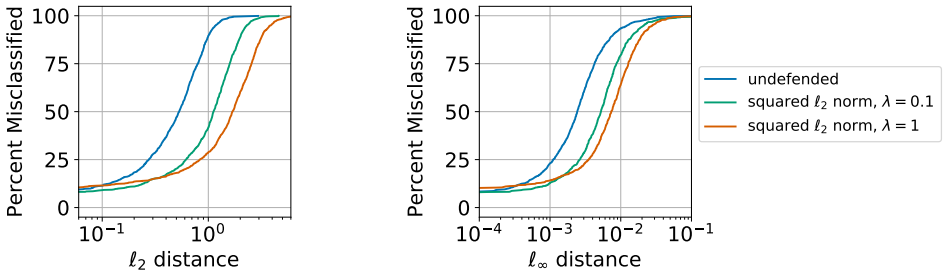
[53] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *CoRR*, abs/1812.03411, 2018. URL http://arxiv.org/abs/1812.03411.

[54] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5987–5995, 2017. URL https://doi.org/10.1109/CVPR.2017.634.

A Additional methods and results

Algorithm 1 Training with squared ℓ_2-norm input gradient regularization, using finite differences

Input: Initial model parameters w_0
Hyperparameters: Regularization strength λ; batch size m; finite difference discretization h
while w_t not converged **do**
  sample minibatch of data $\left\{(x^{(i)}, y^{(i)})\right\}_{i=1,\ldots,m}$ from empirical distribution $\hat{\mathbb{P}}$
  for $i = 0$ **to** m **do**
    $g^{(i)} = \nabla_x \ell(x^{(i)}, y^{(i)}; w_t)$
    $d^{(i)} = \begin{cases} \frac{g^{(i)}}{\|g^{(i)}\|_2} & \text{if } g^{(i)} \neq 0 \\ 0 & \text{otherwise} \end{cases}$ ▷ for ℓ_1-norm use normalized signed gradient
    detach $d^{(i)}$ from computational graph
    $z^{(i)} = x^{(i)} + h d^{(i)}$
  end for
  $\mathcal{L}(w) = \frac{1}{m} \sum_{i=1}^{m} \ell(x^{(i)}, y^{(i)}; w)$
  $\mathcal{R}(w) = \frac{1}{m} \sum_{i=1}^{m} \frac{1}{h^2} \left(\ell(z^{(i)}, y^{(i)}; w) - \ell(x^{(i)}, y^{(i)}; w)\right)^2$
  $w_{t+1} \leftarrow w_t - \tau_t \nabla_w \left(\mathcal{L}(w_t) + \lambda \mathcal{R}(w_t)\right)$
end while



(a) ℓ_2-norm adversarial attacks (b) ℓ_∞-norm adversarial attacks

Figure 3: Adversarial attacks on ImageNet-1k with the ResNet-50 architecture. Top5 error reported.

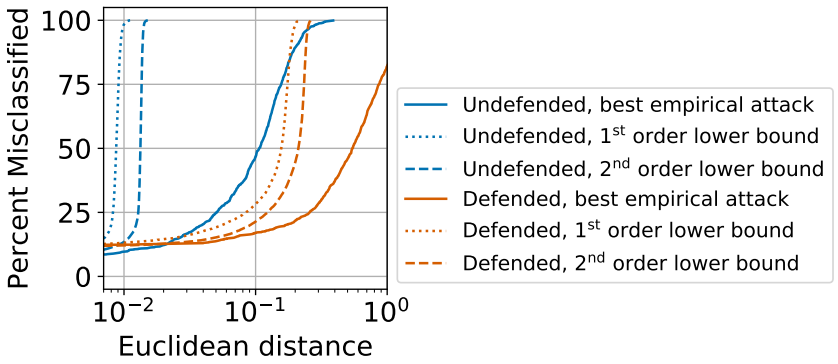


Figure 4: Theoretical minimum lower bound on adversarial distance for CIFAR-10, on networks with smooth ReLU activation functions. Defended networks trained with $\lambda = 0.1$, penalized with squared ℓ_2 norm gradient.

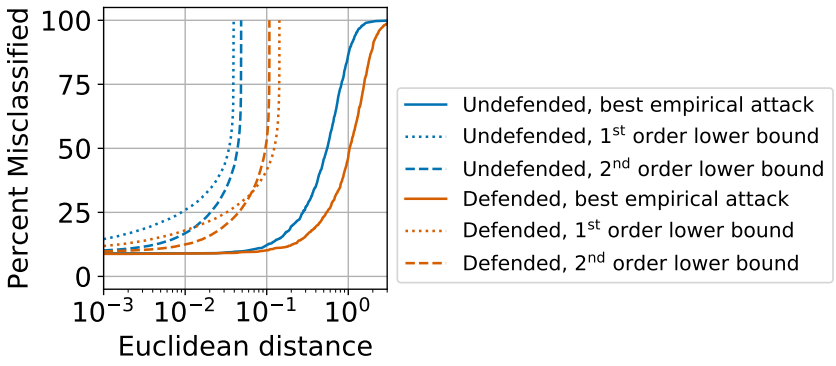


Figure 5: Theoretical minimum lower bound on adversarial distance for ImageNet-1k, on networks with smooth ReLU activation functions. Defended networks trained with $\lambda = 0.1$, penalized with squared ℓ_2 norm gradient.

Table 3: Regularity statistics on selected models, measured in the ℓ_2 norm. Statistics computed using modified loss $\max_{i \neq c} f_i(x) - f_c(x)$. A soft maximum is used for curvature statistics.

	soft ReLU?	mean		maximum	
		$\|\nabla \ell(x)\|$	$\|\nabla^2 \ell(x)\|$	$\|\nabla \ell(x)\|$	$\|\nabla^2 \ell(x)\|$
CIFAR-10					
Undefended		3.05	-	122.34	-
Undefended	✓	3.25	198.23	65.35	8134.26
Madry et al (7-step AT)		0.40	-	2.52	-
squared ℓ_2 norm, $\lambda = 0.1$		0.58	-	4.43	-
squared ℓ_2 norm, $\lambda = 0.1$	✓	0.65	2.08	4.52	27.05
squared ℓ_2 norm, $\lambda = 1$		0.35	-	1.33	-
ImageNet-1k					
Undefended		1.12	-	17.51	-
Undefended	✓	1.02	11.61	25.43	848.69
squared ℓ_2 norm, $\lambda = 0.1$		0.46	-	4.85	-
squared ℓ_2 norm, $\lambda = 0.1$	✓	0.45	1.87	6.99	171.98
squared ℓ_2 norm, $\lambda = 1$		0.27	-	2.12	-