Adversarial Robustness Guarantees for Random Deep Neural Networks

Giacomo De Palma [1 2 3] **Bobak T. Kiani** [3 4] **Seth Lloyd** [2 3]

Abstract

The reliability of deep learning algorithms is fundamentally challenged by the existence of adversarial examples, which are incorrectly classified inputs that are extremely close to a correctly classified input. We explore the properties of adversarial examples for deep neural networks with random weights and biases, and prove that for any $p \geq 1$, the ℓ^p distance of any given input from the classification boundary scales as one over the square root of the dimension of the input times the ℓ^p norm of the input. The results are based on the recently proved equivalence between Gaussian processes and deep neural networks in the limit of infinite width of the hidden layers, and are validated with experiments on both random deep neural networks and deep neural networks trained on the MNIST and CIFAR10 datasets. The results constitute a fundamental advance in the theoretical understanding of adversarial examples, and open the way to a thorough theoretical characterization of the relation between network architecture and robustness to adversarial perturbations.

1. Introduction

Deep neural networks constitute an extremely powerful architecture for machine learning and have achieved an enormous success in several fields such as speech recognition, computer vision and natural language processing where they can often outperform human abilities (Mnih et al., 2015; LeCun et al., 2015; Radford et al., 2015; Schmidhuber, 2015; Goodfellow et al., 2016). In 2014, a very surprising property of deep neural networks emerged in the context of image classification (Szegedy et al., 2014; Goodfellow et al., 2014): an extremely small perturbation can change the label of a correctly classified image. This property poses serious challenges to the reliability of deep learning algorithms since it may be exploited by a malicious adversary to fool a machine learning algorithm by steering its output. For this reason, methods to find perturbed inputs or adversarial examples have been named adversarial attacks. This problem further captured the attention of the deep learning community when it was discovered that real-world images taken with a camera can also constitute adversarial examples (Kurakin et al., 2018; Sharif et al., 2016; Brown et al., 2017; Eykholt et al., 2018). To study adversarial attacks, two lines of research have been developed: one aims at developing efficient algorithms to find adversarial examples (Su et al., 2019; Athalye et al., 2018; Liu et al., 2016), and the other aims at making deep neural networks more robust against adversarial attacks (Madry et al., 2018; Tsipras et al., 2019; Nakkiran, 2019; Lecuyer et al., 2019; Gilmer et al., 2019); algorithms to compute the robustness of a given trained deep neural network against adversarial attacks have also been developed (Li et al., 2019a; Jordan et al., 2019).

Several theories have been proposed to explain the phenomenon of adversarial examples (Raghunathan et al., 2018; Wong & Kolter, 2018; Xiao et al., 2019; Cohen et al., 2019; Schmidt et al., 2018; Tanay & Griffin, 2016; Kim et al., 2019; Fawzi et al., 2016; Shamir et al., 2019; Bubeck et al., 2019; Ilyas et al., 2019). One of the most prominent theories states that adversarial examples are an unavoidable feature of the high-dimensional geometry of the input space: Refs. (Gilmer et al., 2018; Fawzi et al., 2018; Shafahi et al., 2019; Mahloujifar et al., 2019) show that, whenever the classification error is finite, the label of a correctly classified input can be changed with an adversarial perturbation of size $O\left(1/\sqrt{n}\right)$ times the norm of the input, where n is the dimension of the input space.

In this paper, we explore the properties of adversarial examples for deep neural networks with random weights and biases in the limit of infinite width of the hidden layers. Our main result, presented in section 3, is a probabilistic robustness guarantee on the ℓ^1 distance of a given input from the closest classification boundary, which we later extend to all the ℓ^p distances[1]. We prove that the ℓ^1 distance

[1]Scuola Normale Superiore, Pisa, Italy [2]Department of Mechanical Engineering, MIT, Cambridge MA, USA [3]Research Laboratory of Electronics, MIT, Cambridge MA, USA [4]Department of Electrical Engineering & Computer Science, MIT, Cambridge MA, USA. Correspondence to: Giacomo De Palma <giacomo.depalma@sns.it>, Bobak T. Kiani <bkiani@mit.edu>, Seth Lloyd <slloyd@mit.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

[1]the ℓ^p norm of a vector $x \in \mathbb{R}^n$ is $\|x\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{\frac{1}{p}}$.

from the closest classification boundary of any given input $x \in \mathbb{R}^n$ whose entries are $O(1)$ is with high probability at least $\tilde{\Omega}\left(\sqrt{n}\right)$ (the tilde means that logarithmic factors are hidden), *i.e.*, the distance of any adversarial example from x is larger than $\tilde{\Omega}\left(\sqrt{n}\right)$. Since $\|x\|_1 = \Theta(n)$, our result implies that the size of any adversarial perturbation is at least $\tilde{\Omega}\left(1/\sqrt{n}\right)$ times the norm of the input. This lower bound to the size of adversarial perturbations matches the upper bound imposed by the high-dimensional geometry proven in Refs. (Gilmer et al., 2018; Fawzi et al., 2018; Shafahi et al., 2019; Mahloujifar et al., 2019). Therefore, our result proves that $1/\sqrt{n}$ is the universal scaling of the minimum size of adversarial perturbations with respect to the norm of the input. We also prove that, for any given unit vector $v \in \mathbb{R}^n$, with high probability all the inputs $x + t\,v$ with $0 \leq t \leq O\left(\sqrt{n}\right)$ have the same classification as x. Since $\|x\|_2 = \Theta\left(\sqrt{n}\right)$, a remarkable consequence of this result is that a finite fraction of the ℓ^2 distance to the origin can be traveled without encountering any classification boundary.

Our results encompass a wide variety of network architectures, namely any combination of convolutional or fully connected layers with nonlinear activation, skipped connections and pooling (see section 2). Our proof builds on the recently proved equivalence between deep neural networks with random weights and biases in the limit of infinite width and Gaussian processes (Lee et al., 2018; Yang, 2019b). We prove that the same probabilistic robustness guarantees for the adversarial distance also apply to a broad class of Gaussian processes when the variance is lower bounded by the Euclidean square norm of the input and the feature map of the kernel associated to the Gaussian process has an $O(1)$ Lipschitz constant, a result that can be of independent interest.

In section 4, we experimentally validate our theoretically predicted scaling of the adversarial distance for random deep neural networks, and we find a very good agreement between theory and experiments starting from $n \gtrsim 100$. In subsection 4.1, we perform experiments on the adversarial distance for deep neural networks trained on the MNIST and CIFAR10 datasets. In both cases, the training does not change the order of magnitude of the adversarial distance. While for MNIST the adversarial distances for random and trained networks are very close, in the case of CIFAR10 the training decreases the adversarial distance by roughly half order of magnitude. As better discussed in subsection 4.1, this can be ascribed to the different nature of the CIFAR10 with respect to the MNIST data.

Our adversarial robustness guarantee applies also to deep neural networks trained with Bayesian inference under the hypothesis that the target function f is generated by the same random deep neural network employed for the training. Indeed, given the training inputs $x^{(1)}, \ldots, x^{(n)}$ (which do not need to be random) and the corresponding random training labels $y^{(1)} = f\left(x^{(1)}\right), \ldots, y^{(n)} = f\left(x^{(n)}\right)$, the Bayesian classifier is a function g randomly drawn from the posterior probability distribution q obtained by conditioning on the observation of $y^{(1)}, \ldots, y^{(n)}$ the prior probability distribution p generated by the random deep neural network. If we forget the values of the training labels, the probability distribution of g becomes the average of q over the possible values of the training labels $y^{(1)}, \ldots, y^{(n)}$ induced by the random target function f. Such average has the effect of removing the conditioning on $y^{(1)}, \ldots, y^{(n)}$. Therefore, after such average, the probability distribution of the Bayesian classifier g coincides with the prior probability distribution p regardless of the choice of p and of $x^{(1)}, \ldots, x^{(n)}$. Therefore, the properties of the adversarial distance for a given random deep neural network and for the Bayesian classifier associated to the same network coincide under the hypothesis that the target function is also generated by the same random network.

1.1. Related Works

The equivalence between Gaussian processes and neural networks with random weights and biases in the limit of infinite width of the hidden layers has been known for a long time in the case of fully connected neural networks with one hidden layer (Neal, 1996; Williams, 1997), and has recently been extended to multi-layer (Schoenholz et al., 2016; Pennington et al., 2018; Lee et al., 2018; Matthews et al., 2018; Poole et al., 2016; Schoenholz et al., 2016) and convolutional deep neural networks (Garriga-Alonso et al., 2019; Xiao et al., 2018; Novak et al., 2019). The equivalence is now proved for practically all the existing neural networks architectures (Yang, 2019b), and has been extended to trained deep neural networks (Jacot et al., 2018; Lee et al., 2019; Yang, 2019a; Arora et al., 2019; Huang & Yau, 2020; Li et al., 2019b; Wei et al., 2019; Cao & Gu, 2019) including adversarial training (Gao et al., 2019). Ref. (Cardelli et al., 2019b) proves a probabilistic robustness guarantee for Bayesian classifiers with prior probability distribution given by a Gaussian process in the same spirit of Theorem 1, and also this proof exploits the Borell–TIS inequality and Dudley's theorem. The results of (Cardelli et al., 2019b) have been expanded to probabilistic guarantees for neural networks in (Cardelli et al., 2019a; Wicker et al., 2020). The smoothness of the feature map of a kernel plays a key role in machine learning applications (Mallat, 2012; Oyallon & Mallat, 2015; Bruna & Mallat, 2013; Bietti & Mairal, 2019a) and kernels associated to deep neural networks have been studied from this point of view (Bietti & Mairal, 2019b). In the setup of binary classification of bit strings, the Hamming distance of a given input from the closest classification boundary has been theoretically studied in (De Palma et al., 2019), where the scaling $O\left(\sqrt{n/\ln n}\right)$ has been found.

2. Setup

Our inputs are D-dimensional images considered as elements of $\mathbb{R}^{n_C^{(0)} \times \mathcal{I}^{(0)}}$, where $n_C^{(0)}$ is the number of the input channels (e.g., $n_C^{(0)} = 3$ for Red-Green-Blue images) and $\mathcal{I}^{(0)} = \mathbb{Z}_{h_1} \times \ldots \times \mathbb{Z}_{h_D}$ is the set of the input pixels, assumed for simplicity to be periodic. $D = 2$ recovers standard 2D images. For the sake of a simpler notation, we will sometimes consider the input space as $\mathbb{R}^n$, with $n = n_C^{(0)} \ \mathcal{I}^{(0)}$.

Our architecture allows for any combination of convolutional layers, fully connected layers, skipped connections and pooling. For the sake of a simpler notation, we treat each of the above operations as a layer, even if it does not include any nonlinear activation. For simplicity, we assume that the nonlinear activation function is the ReLU $\tau(x) = \max(0, x)$. Our results can be easily extended to other activation functions.

For any $l = 1, \ldots, L+1$ and any input $x \in \mathbb{R}^{n_C^{(0)} \times \mathcal{I}^{(0)}}$, let $n_C^{(l)}$ be the number of channels and $\mathcal{I}^{(l)}$ the set of pixels of the output of the l-th layer $\phi^{(l)}(x) \in \mathbb{R}^{n_C^{(l)} \times \mathcal{I}^{(l)}}$. The layer transformations have the following mathematical expression:

- **Input layer:** We have $\mathcal{I}^{(1)} = \mathcal{I}^{(0)}$ and

$$\phi_{i,\alpha}^{(1)}(x) = b_i^{(1)} + \sum_{j=1}^{n_C^{(0)}} \sum_{\beta \in \mathcal{P}^{(1)}} W_{ij,\beta}^{(1)} \, x_{j,\alpha+\beta} \qquad (1)$$

  for any $i = 1, \ldots, n_C^{(1)}$ and any $\alpha \in \mathcal{I}^{(1)}$, where $\mathcal{P}^{(1)} \subseteq \mathcal{I}^{(1)} = \mathcal{I}^{(0)}$ is the convolutional patch of the first layer. We assume for simplicity that $-\mathcal{P}^{(1)} = \mathcal{P}^{(1)}$.

- **Nonlinear layer:** If the $(l+1)$-th layer is a nonlinear layer, we have $\mathcal{I}^{(l+1)} = \mathcal{I}^{(l)}$ and

$$\phi_{i,\alpha}^{(l+1)}(x) = b_i^{(l+1)} + \sum_{j=1}^{n_C^{(l)}} \sum_{\beta \in \mathcal{P}^{(l+1)}} W_{ij,\beta}^{(l+1)} \, \tau\left(\phi_{j,\alpha-\beta}^{(l)}(x)\right) \qquad (2)$$

  for any $i = 1, \ldots, n_C^{(l+1)}$ and any $\alpha \in \mathcal{I}^{(l+1)}$, where $\tau : \mathbb{R} \to \mathbb{R}$ is the activation function and $\mathcal{P}^{(l+1)} \subseteq \mathcal{I}^{(l+1)} = \mathcal{I}^{(l)}$ is the convolutional patch of the layer. We assume for simplicity that $-\mathcal{P}^{(l+1)} = \mathcal{P}^{(l+1)}$. Fully connected layers are recovered by $\mathcal{I}^{(l)} = \mathcal{I}^{(l+1)} = \mathcal{P}^{(l+1)} = 1$.

- **Skipped connection:** If the $(l+1)$-th layer is a skipped connection, we have $n_C^{(l+1)} = n_C^{(l)}$, $\mathcal{I}^{(l+1)} = \mathcal{I}^{(l)}$ and

$$\phi_{i,\alpha}^{(l+1)}(x) = \phi_{i,\alpha}^{(l)}(x) + \phi_{i,\alpha}^{(l-k)}(x) \qquad (3)$$

  for any $i = 1, \ldots, n_C^{(l+1)}$ and any $\alpha \in \mathcal{I}^{(l+1)}$, where $k \in \{1, \ldots, l-2\}$ is such that the sum in (3) is well defined, i.e., $n_C^{(l-k)} = n_C^{(l)}$ and $\mathcal{I}^{(l-k)} = \mathcal{I}^{(l)}$. For the sake of a simple proof, we assume that the l-th layer is either a convolutional or a fully connected layer.

- **Pooling:** If the $(l+1)$-th layer is a pooling layer, we have $n_C^{(l+1)} = n_C^{(l)}$, and $\mathcal{I}^{(l+1)}$ is a partition of $\mathcal{I}^{(l)}$, i.e., the elements of $\mathcal{I}^{(l+1)}$ are disjoint subsets of $\mathcal{I}^{(l)}$ whose union is equal to $\mathcal{I}^{(l)}$. We assume for simplicity that the l-th layer is a convolutional layer and that all the elements of $\mathcal{I}^{(l+1)}$ have the same cardinality, which is therefore equal to $\mathcal{I}^{(l)} \ / \ \mathcal{I}^{(l+1)}$. We have

$$\phi_{i,\alpha}^{(l+1)}(x) = \sum_{\beta \in \alpha} \phi_{i,\beta}^{(l)}(x) \qquad (4)$$

  for any $i = 1, \ldots, n_C^{(l+1)}$ and any $\alpha \in \mathcal{I}^{(l+1)}$.

- **Flattening layer:** Let the (L_f+1)-th layer be the flattening layer. We notice that we include a fully connected layer directly after the flattening as part of this layer. We have $\mathcal{I}^{(L_f+1)} = 1$ and

$$\phi_i^{(L_f+1)}(x) = b_i + \sum_{j=1}^{n_C^{(L_f)}} \sum_{\alpha \in \mathcal{I}^{(L_f)}} W_{ij,\alpha}^{(L_f+1)} \, \tau\left(\phi^{(L_f)}(x)\right) \qquad (5)$$

  for any $i = 1, \ldots, n_C^{(L_f+1)}$.

- **Output layer:** The final output of the network is $\phi(x) = \phi_1^{(L+1)}(x)$, and the output label is $\operatorname{sign} \phi(x)$. We introduce the other components of $\phi^{(L+1)}$ for the sake of a simpler notation in the proof of Theorem 2.

Our random deep neural networks draw all the weights $W_{ij,\alpha}^{(l)}$ and the biases $b_i^{(l)}$ from independent Gaussian probability distributions with zero mean and variances ${\sigma_W^{(l)}}^2 \Big/ n_C^{(l-1)}$ and ${\sigma_b^{(l)}}^2$, respectively. The variances are allowed to depend on the layer.

3. Theoretical Results

A recent series of works (Schoenholz et al., 2016; Pennington et al., 2018; Lee et al., 2018; Matthews et al., 2018; Poole et al., 2016; Garriga-Alonso et al., 2019; Xiao et al., 2018; Novak et al., 2019; Yang, 2019b) has proved that in the limit $n_C^{(1)}, \ldots, n_C^{(L+1)} \to \infty$ the random deep neural networks defined in section 2 are centered Gaussian processes, i.e., for any $M \in \mathbb{N}$ and any set of M inputs $x^1, \ldots, x^M \in \mathbb{R}^{n_C^{(0)} \times \mathcal{I}^{(0)}}$, the joint probability distribution of the corresponding outputs $\phi\left(x^1\right), \ldots, \phi\left(x^M\right) \in \mathbb{R}$ is

Gaussian with zero mean and covariance given by a kernel $K(x,y) = \mathbb{E}(\phi(x)\,\phi(y))$ that depends on the architecture of the deep neural network. Therefore, the properties of adversarial perturbations for random deep neural networks are equivalent to the properties of adversarial perturbations for the corresponding Gaussian processes. First, we prove in Theorem 1 an adversarial robustness guarantee for a broad class of Gaussian processes. We then prove in Theorem 2 that the guarantee applies to the Gaussian processes generated by random deep neural networks, and therefore it applies to random deep neural networks.

We recall that we can associate to any kernel K on $\mathbb{R}^n$ a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with scalar product and norm denoted by $\cdot$ and $\|\cdot\|$, respectively, and a feature map $\Phi : \mathbb{R}^n \to \mathcal{H}$ such that for any $x,\, y \in \mathbb{R}^n$ (Rasmussen & Williams, 2006)

$$K(x,y) = \Phi(x) \cdot \Phi(y)\,. \tag{6}$$

The kernel K induces on the input space the RKHS distance

$$\begin{aligned} d(x,y)^2 &= \|\Phi(x) - \Phi(y)\|^2 \\ &= K(x,x) - 2\,K(x,y) + K(y,y)\,. \end{aligned} \tag{7}$$

We can now state our main result.

Theorem 1 (ℓ^1 adversarial robustness guarantee for Gaussian processes)**.** *Let ϕ be a Gaussian process on $\mathbb{R}^n$ with zero mean and covariance K, and let d be the associated RKHS distance. Let $C,\, M > 0$ be such that for any $x,\, y \in \mathbb{R}^n$*

$$\sqrt{K(x,x)} \geq C\,\|x\|_2\,, \quad d(x,y) \leq M\,C\,\|x-y\|_2\,. \tag{8}$$

Let $x_0 \in \mathbb{R}^n$, and for any $r > 0$ let

$$\mathcal{B}_r^1 = \{x \in \mathbb{R}^n : \|x - x_0\|_1 < r\} \tag{9}$$

be the ℓ^1 ball with center x_0 and radius r. Then, for any $0 < \delta < 1$ and any

$$0 < r \leq \frac{\|x_0\|_2\,\delta\sqrt{\pi}}{M\left(12\sqrt{\ln 4n} + 8\ln n\,\sqrt{\ln 2n} + 2\sqrt{\pi}\right)} \tag{10}$$

we have

$$\mathbb{P}\left(\exists\, x \in \mathcal{B}_r^1 : \phi(x) = 0\right) \leq \delta\,. \tag{11}$$

Moreover, let v be a unit vector in $\mathbb{R}^n$, and for any $r > 0$ let $\mathcal{L}_r = \{x_0 + t\,v : 0 \leq t \leq r\}$ be the segment starting in x_0, parallel to v and with length r. Then, for any

$$0 < r \leq \pi\,\|x_0\|_2\,\delta\,/(2\,M + \pi) \tag{12}$$

we have

$$\mathbb{P}\left(\exists\, x \in \mathcal{L}_r : \phi(x) = 0\right) \leq \delta\,. \tag{13}$$

We prove the first part of Theorem 1 in subsection 3.1, and we refer to the Supplementary Manuscript for the proof of the second part.

Remark 1. Since our classifier is $\operatorname{sign}\phi(x)$, we have $\phi(x) = 0$ for some x in $\mathcal{B}_r^1$ iff $\mathcal{B}_r^1$ is crossed by a classification boundary, *i.e.*, iff there exists $x \in \mathcal{B}_r^1$ such that $\phi(x)\,\phi(x_0) < 0$.

Remark 2. The prefactor in (10) is not sharp. Indeed, the proof of Theorem 1 relies on Dudley's theorem (Bartlett, 2013), which provides an upper bound to the expectation value of the maximum of a Gaussian process over a given region, and on an estimate of the covering number of the ℓ^1 unit ball (Theorem 3). Despite employing the best state-of-the-art tools, the prefactors of both these results are not sharp (Ledoux & Talagrand, 2013; Price, 2016).

The following Theorem 2, which we prove in the Supplementary Manuscript, states that the kernels of the Gaussian processes associated to random deep neural networks satisfy the hypotheses of Theorem 1.

Theorem 2 (smoothness of the DNN Gaussian processes)**.** *The kernel associated to the output of a random deep neural network as in section 2 satisfies (8) with*

$$M = \sqrt{\mathcal{I}^{(0)}\,/\,\mathcal{I}^{(L_f)}}\,, \tag{14}$$

where $\mathcal{I}^{(0)}$ and $\mathcal{I}^{(L_f)}$ are sets of the pixels of the input and of the layer immediately before the flattening, respectively.

Corollary 1 (ℓ^1 adversarial robustness guarantee for random deep neural networks)**.** *Let ϕ be a random deep neural network as in section 2. For any input $x_0 \in \mathbb{R}^{n_C^{(0)} \times \mathcal{I}^{(0)}}$ and any $r > 0$ let*

$$\mathcal{B}_r^1 = \left\{x \in \mathbb{R}^{n_C^{(0)} \times \mathcal{I}^{(0)}} : \|x - x_0\|_1 < r\right\}. \tag{15}$$

Then, in the limit $n_C^{(1)}, \ldots, n_C^{(L+1)} \to \infty$, for any $0 < \delta < 1$ and any

$$0 < r \leq \frac{\|x_0\|_2\,\delta\sqrt{\pi\;\mathcal{I}^{(L_f)}\,/\,\mathcal{I}^{(0)}}}{12\sqrt{\ln 4n} + 8\ln n\,\sqrt{\ln 2n} + 2\sqrt{\pi}}\,, \tag{16}$$

where $n = n_C^{(0)}\;\mathcal{I}^{(0)}$, we have

$$\mathbb{P}\left(\exists\, x \in \mathcal{B}_r^1 : \phi(x) = 0\right) \leq \delta\,. \tag{17}$$

Moreover, let v be a unit vector in $\mathbb{R}^{n_C^{(0)} \times \mathcal{I}^{(0)}}$, and for any $r > 0$ let $\mathcal{L}_r = \{x_0 + t\,v : 0 \leq t \leq r\}$. Then, for any

$$0 < r \leq \frac{\pi\,\|x_0\|_2\,\delta}{2\sqrt{\mathcal{I}^{(0)}\,/\,\mathcal{I}^{(L_f)}} + \pi} \tag{18}$$

we have

$$\mathbb{P}\left\{\exists\, x \in \mathcal{L}_r : \phi(x) = 0\right\} \leq \delta\,. \tag{19}$$

Remark 3. The bounds of Corollary 1 do not depend on the choice of the variances of weights and biases.

Remark 4 (asymptotic scaling). Theorem 1 and Corollary 1 hold for any choice of n, $n_C^{(0)}$ and $\mathcal{I}^{(0)}$. In the limit $n \to \infty$, if all the entries of x_0 are $\Theta(1)$ we have $\|x_0\|_2 = \Theta\left(\sqrt{n}\right)$, and therefore both (10) and (12) become, up to logarithmic factors,

$$0 < r \le \tilde{O}\left(\delta\sqrt{n}/\,M\right) . \tag{20}$$

Analogously, in the limit $n = n_C^{(0)}\ \mathcal{I}^{(0)}\ \to \infty$ both (16) and (18) become

$$0 < r \le \tilde{O}\left(\delta\sqrt{n\ \mathcal{I}^{(L_f)}\ /\ \mathcal{I}^{(0)}}\right) . \tag{21}$$

Remark 5 (ℓ^p adversarial robustness guarantees). Let us assume for simplicity that $\mathcal{I}^{(L_f)}\ /\ \mathcal{I}^{(0)}$ does not scale with n. For any $p \ge 1$ and any $r > 0$, let

$$\mathcal{B}_r^p = \left\{x \in \mathbb{R}^{n_C^{(0)} \times \mathcal{I}^{(0)}} : \|x - x_0\|_p < r\right\} \tag{22}$$

be the ℓ^p ball with center x_0 and radius r. Since

$$\|x - x_0\|_1 \le n^{\frac{p-1}{p}} \|x - x_0\|_p \tag{23}$$

for any $x \in \mathbb{R}^{n_C^{(0)} \times \mathcal{I}^{(0)}}$, we trivially have from Remark 4 that in the limit $n \to \infty$,

$$\mathbb{P}\left(\exists\, x \in \mathcal{B}_r^p : \phi(x) = 0\right) \le \delta \tag{24}$$

for $0 < r \le \tilde{O}\left(\delta\, n^{\frac{1}{p}-\frac{1}{2}}\right)$. In particular, the ℓ^2 and ℓ^∞ distances from the closest classification boundary scale at least as $\tilde{\Omega}(1)$ and $\tilde{\Omega}\left(1/\sqrt{n}\right)$, respectively. If all the entries of x_0 are $\Theta(1)$, then $\|x_0\|_p = \Theta(n^{\frac{1}{p}})$, and the ratio between the ℓ^p distance to the classification boundary and $\|x_0\|_p$ scales at least as $\tilde{\Omega}\left(1/\sqrt{n}\right)$ regardless of p, *i.e.*, regardless of the choice of the norm, a fraction $1/\sqrt{n}$ of the input must be changed to change the label.

To summarize, we have proved that the ℓ^1 distance of any given input from the closest classification boundary is with high probability at least $\Omega(\sqrt{n})$, where n is the dimension of the input. Moreover, for any $p \ge 1$, the ℓ^p distance of any given input from the closest classification boundary is with high probability at least $\Omega\left(1/\sqrt{n}\right)$ times the ℓ^p norm of the input. This result applies to both smooth Gaussian processes and deep neural networks with almost any architecture and random weights and biases.

3.1. Proof of Theorem 1, Part I

Let

$$p_r = \mathbb{P}\left(\exists\, x \in \mathcal{B}_r^1 : \phi(x) = 0\right) , \tag{25}$$

and for any $\phi_0 > 0$ let

$$p_r(\phi_0) = \mathbb{P}\left(\exists\, x \in \mathcal{B}_r^1 : \phi(x) = 0 \,|\, \phi(x_0) = \phi_0\right) . \tag{26}$$

Conditioning on $\phi(x_0) = \phi_0$, ϕ becomes the Gaussian process with average

$$\mu(x) = K(x, x_0)\, \phi_0 \,/ K(x_0, x_0) \tag{27}$$

and covariance

$$\hat{K}(x, y) = K(x, y) - \frac{K(x, x_0)\, K(x_0, y)}{K(x_0, x_0)} . \tag{28}$$

We put $\phi(x) = \mu(x) - \varphi(x)$ for any $x \in \mathbb{R}^n$, such that φ is a centered Gaussian process with covariance $\hat{K}$. Let

$$\begin{aligned} K_r &= \inf_{x \in \mathcal{B}_r^1} \frac{K(x, x_0)}{K(x_0, x_0)} , \\ \varphi_r &= \mathbb{E} \sup_{x \in \mathcal{B}_r^1} \varphi(x) , \qquad \sigma_r^2 = \sup_{x \in \mathcal{B}_r^1} \hat{K}(x, x) , \end{aligned} \tag{29}$$

and let us assume that $K_r\, \phi_0 > \varphi_r$. The Borell–TIS inequality (Adler & Taylor, 2009) provides an upper bound to $p_r(\phi_0)$:

Theorem (Borell–TIS inequality)**.** *Let φ be a centered Gaussian process on $\Omega \subset \mathbb{R}^n$, and let $\hat{K}$ be the associated kernel. Then, for any $t > 0$*

$$\mathbb{P}\left(\sup_{x \in \Omega} \varphi(x) \ge \mathbb{E} \sup_{x \in \Omega} \varphi(x) + t\right) \le \mathrm{e}^{-\frac{t^2}{2\,\sigma^2}} , \tag{30}$$

where $\sigma^2 = \sup_{x \in \Omega} \hat{K}(x, x)$.

We have from the Borell–TIS inequality

$$\begin{aligned} p_r(\phi_0) &\le \mathbb{P}\left(\exists\, x \in \mathcal{B}_r^1 : \varphi(x) \ge \mu(x)\right) \\ &\le \mathbb{P}\left(\exists\, x \in \mathcal{B}_r^1 : \varphi(x) \ge K_r\, \phi_0\right) \le \mathrm{e}^{-\frac{(K_r\, \phi_0 - \varphi_r)^2}{2\,\sigma_r^2}} . \end{aligned} \tag{31}$$

Recalling that $\phi(x_0)$ is a centered Gaussian random variable with variance $K(x_0, x_0)$, we have

$$\begin{aligned} p_r &= 2 \int_0^\infty p_r(\phi_0)\, \mathrm{e}^{-\frac{\phi_0^2}{2\,K(x_0, x_0)}} \frac{\mathrm{d}\phi_0}{\sqrt{2\pi\, K(x_0, x_0)}} \\ &\le 2 \int_0^{\frac{\varphi_r}{K_r}} \mathrm{e}^{-\frac{\phi_0^2}{2\,K(x_0, x_0)}} \frac{\mathrm{d}\phi_0}{\sqrt{2\pi\, K(x_0, x_0)}} \\ &\quad + 2 \int_{\frac{\varphi_r}{K_r}}^\infty \mathrm{e}^{-\frac{(K_r\, \phi_0 - \varphi_r)^2}{2\,\sigma_r^2} - \frac{\phi_0^2}{2\,K(x_0, x_0)}} \frac{\mathrm{d}\phi_0}{\sqrt{2\pi\, K(x_0, x_0)}} \\ &\le \frac{\sqrt{\frac{2}{\pi}}\, \varphi_r + \sigma_r}{K_r \sqrt{K(x_0, x_0)}} . \end{aligned} \tag{32}$$

We get an upper bound on φ_r from Dudley's theorem (Bartlett, 2013).

Theorem (Dudley's theorem)**.** *Let φ be a centered Gaussian process on $\Omega \subset \mathbb{R}^n$, and let $\hat{d}$ be the RKHS distance of the associated kernel. For any $\epsilon > 0$, let $N(\epsilon)$ be the minimum number of balls of $\hat{d}$ with radius ϵ that can cover Ω. Then,*

$$\mathbb{E} \sup_{x \in \Omega} \varphi(x) \le 8\sqrt{2} \int_0^\infty \sqrt{\ln N(\epsilon)}\, \mathrm{d}\epsilon . \tag{33}$$

We directly get from Dudley's theorem

$$\varphi_r \leq 8\sqrt{2}\int_0^\infty \sqrt{\ln N_r(\epsilon)}\,\mathrm{d}\epsilon\,, \tag{34}$$

where $N_r(\epsilon)$ is the minimum number of balls of $\hat{d}$ with radius ϵ that can cover $\mathcal{B}_r^1$. Let $N(\epsilon)$ be the minimum number of balls of the Euclidean distance with radius ϵ that can cover the unit ℓ^1 ball. In the Supplementary Manuscript we prove the following Lemma 1:

Lemma 1. *$\hat{d}(x,y) \leq d(x,y)$ for any $x,\, y \in \mathbb{R}^n$.*

From Lemma 1 and (8) we get $\hat{d}(x,y) \leq M\,C\,\|x-y\|_2$ for any $x,\, y \in \mathbb{R}^n$, therefore $N_r(\epsilon) \leq N\left(\epsilon/(M\,C\,r)\right)$ and (34) implies

$$\varphi_r \leq 8\sqrt{2}\,M\,C\,r\int_0^\infty \sqrt{\ln N(\epsilon)}\,\mathrm{d}\epsilon\,. \tag{35}$$

In the Supplementary Manuscript we prove the following Theorem 3:

Theorem 3. *For any $\epsilon > 0$, the open unit ball $\mathcal{B}_1$ of the ℓ^1 norm in $\mathbb{R}^n$ can be covered with*

$$N(\epsilon) \leq \begin{cases} 1 & \epsilon \geq 1 \\ (2n)^{\frac{1}{\epsilon^2}} & \frac{1}{\sqrt{n}} < \epsilon < 1 \\ \left(1+\frac{2}{\epsilon}\right)^n & 0 < \epsilon \leq \frac{1}{\sqrt{n}} \end{cases} \tag{36}$$

balls of the Euclidean distance with radius ϵ and centers in $\mathcal{B}_1$.

We get from Theorem 3

$$\begin{aligned}
&\int_0^\infty \sqrt{\ln N(\epsilon)}\,\mathrm{d}\epsilon \leq \\
&\int_0^{\frac{1}{\sqrt{n}}} \sqrt{n\ln\left(1+\frac{2}{\epsilon}\right)}\,\mathrm{d}\epsilon + \sqrt{\ln 2n}\int_{\frac{1}{\sqrt{n}}}^1 \frac{\mathrm{d}\epsilon}{\epsilon} \\
&= \sqrt{\frac{\ln 4n}{2}}\int_0^1 \sqrt{\ln\left(\frac{1}{2\sqrt{n}}+\frac{1}{x}\right)}\,\mathrm{d}x + \frac{\ln n}{2}\sqrt{\ln 2n} \\
&\leq \sqrt{\frac{\ln 4n}{2}}\int_0^1 \sqrt{\ln\left(\frac{1}{2\sqrt{2}}+\frac{1}{x}\right)}\,\mathrm{d}x + \frac{\ln n}{2}\sqrt{\ln 2n} \\
&\leq \frac{3}{4}\sqrt{\ln 4n} + \frac{\ln n}{2}\sqrt{\ln 2n} = a_n\,,
\end{aligned} \tag{37}$$

where in the second line we made the change of variable $x = \epsilon\sqrt{n}$. In the Supplementary Manuscript, we prove the following Lemma 2 and Lemma 3:

Lemma 2. *We have*

$$K_r \geq 1 - M\,C\,r\,\Big/\sqrt{K(x_0,x_0)}\,. \tag{38}$$

Lemma 3. *We have $\sigma_r \leq M\,C\,r$.*

Putting together (37), (35), (32), Lemma 2 and Lemma 3 we get

$$p_r \leq \frac{\frac{16}{\sqrt{\pi}}\,a_n + 1}{\frac{\sqrt{K(x_0,x_0)}}{M\,C\,r} - 1} \leq \frac{\frac{16}{\sqrt{\pi}}\,a_n + 1}{\frac{\|x_0\|_2}{M\,r} - 1}\,, \tag{39}$$

where the last inequality follows from (8). Therefore, we have $p_r \leq \delta$ for

$$\frac{M\,r}{\|x_0\|_2} \leq \frac{\delta}{\frac{16}{\sqrt{\pi}}\,a_n + 1 + \delta}\,, \tag{40}$$

and the claim follows.

4. Experiments

To experimentally validate Corollary 1 and Remark 5, we performed adversarial attacks on random inputs for various network architectures with randomly chosen weights [2]. As shown in the Supplementary Manuscript, experimental findings were consistent across a variety of networks. For sake of brevity, in this section we only provide figures and results for a simplified residual network. Figure 1 plots the median distance of adversarial examples for a residual network similar to the first proposed residual network (He et al., 2016). This network contains three residual blocks and does not contain a global average pooling layer before the final output (its complete architecture is given in the Supplementary Manuscript). Attacks were performed on 2-dimensional images with three channels and pixel values chosen randomly from the standard uniform distribution.

Results from Figure 1 plotting median adversarial distances as a function of the input dimension are consistent with the expected theoretical scaling in Remark 5. Namely, adversarial distances in the ℓ^1, ℓ^2, and ℓ^∞ norms scale with the dimension of the input n proportionally to $\sqrt{n}$, a constant C (not dependent on n), and $1/\sqrt{n}$, respectively (up to logarithmic factors). Adversarial distances relative to the average starting norm of an input are plotted in Figure 2. This adjusted metric named relative distance provides a convenient means of understanding the scaling of adversarial distances, since relative adversarial distances scale proportionally to $1/\sqrt{n}$ in all norms.

4.1. Adversarial Attacks on Trained Neural Networks

Results from section 4 indicate that adversarial attacks on networks with randomly chosen weights empirically conform with our main findings presented in section 3. In this section, we extend our experimental analysis to networks trained on MNIST and CIFAR10 data. We trained networks

[2] code to replicate experiments published at https://github.com/bkiani/Adversarial-robustness-guarantees-for-random-deep-neural-networks

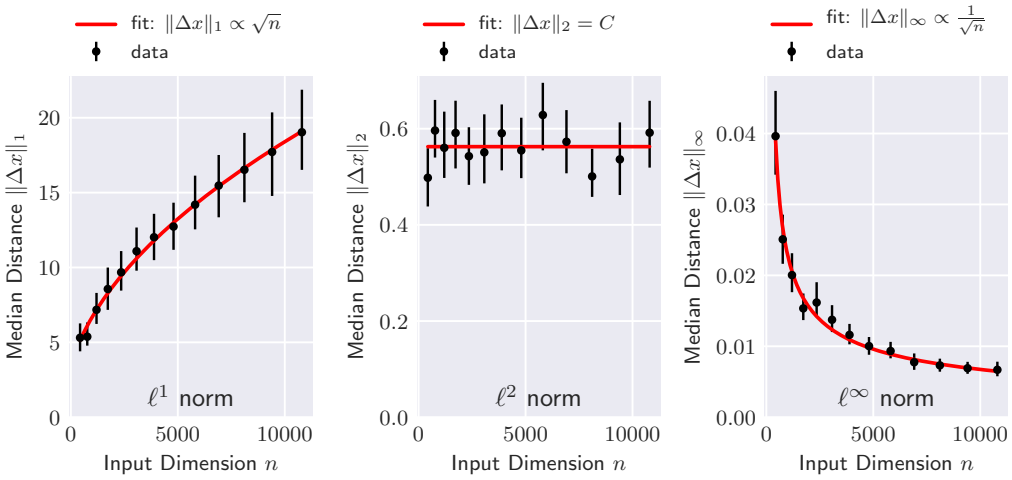

Figure 1. **Random untrained residual networks:** The median ℓ^p distances of closest adversarial examples from their respective inputs for $p = 1,\ 2,\ \infty$ scale as predicted in Remark 5 for a residual network (see the Supplementary Manuscript for full description of network). Error bars span ± 5 percentiles from the median. For each input dimension, results are calculated from 2000 samples (200 random networks each attacked at 10 random points). See the Supplementary Manuscript for further details on how experiments were performed.

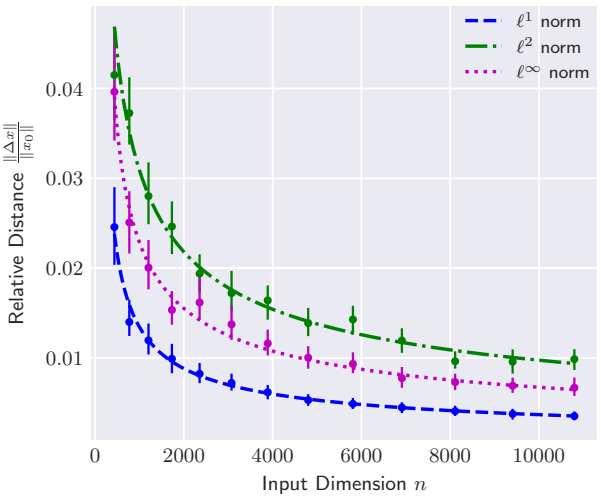

Figure 2. **Random untrained networks:** Median relative distance of closest adversarial examples $\|\Delta x\|_p/\|x_0\|_p$ from their respective inputs ($p \in \{1,\ 2,\ \infty\}$) scale with the input dimension n as $O(1/\sqrt{n})$ in all norms for a residual network with random weights (see the Supplementary Manuscript for full description of network), confirming the theoretical predictions of Remark 5. Results plotted here are for residual networks with random weights. Error bars span ± 5 percentiles from the median. For each input dimension, results are calculated from 2000 samples (200 random networks each attacked at 10 random points).

with the same residual network architecture given in the prior section on MNIST and CIFAR10 data under the task of binary classification. For the case of MNIST, the binary classification task was determining if a digit is odd or even. For CIFAR10, image classes were assigned to binary categories of either {airplane, bird, deer, frog, ship} or {automobile, cat, dog, horse, truck}. Networks were trained for 15 and 25 epochs for the MNIST and CIFAR10 datasets respectively achieving greater than 98% training set accuracy in all cases. We refer to the Supplementary Manuscript for full details on the training of the networks.

Properties of trained neural networks, especially as they relate to adversarial robustness and generalization, are dependent on the properties of the data used to train them. For example, since neural networks can be trained to "memorize" data (Choromanska et al., 2015), Corollary 1 can be forced to fail if the network is trained on a dataset which contains very close inputs with different labels. From Figure 4, the networks trained on CIFAR10 data show a smaller adversarial distance with respect to random networks on both random images and images taken from the training or test set. In the case of MNIST, training decreases the adversarial distance for random images, but does not significantly change it for training or test images. A possible explanation for this discrepancy is the conspicuous geometric and visual structure inherent in the MNIST dataset relative to CIFAR10. Digits in MNIST all have the same uniform black background and geometry and roughly fill the whole image, while in CIFAR10 the background and the relative size of the relevant part of the image can vary significantly, and pictures are taken from various different angles (*e.g.,* different orientations of a dog or car). Thus, when trained on MNIST, networks can more easily embed training and

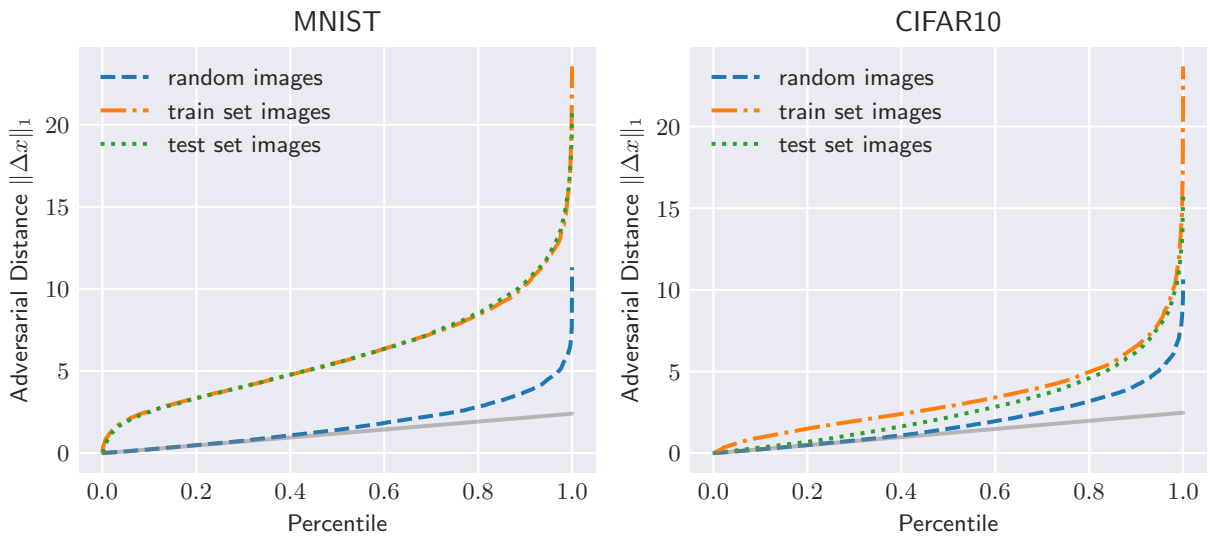

Figure 3. **Trained networks:** Adversarial distance by percentile for random images (images with randomly chosen pixel values) and images in the training and test sets. The expected linear relationship between distance and percentile is observed for random images apart from the highest percentiles as is evident from the linear fit over percentiles ranging from 0 to 0.25 shown as solid line. Adversarial attacks are performed on the ℓ^1 norm. Network architecture is a simplified residual network (see Supplementary Manuscript).

test points within areas far from classification boundaries. More generally, networks trained on MNIST data achieve low generalization error and increased adversarial distances are correlated with those lower errors (though adversarial robustness can sometimes be at odds with generalization (Raghunathan et al., 2019; Tsipras et al., 2019)). On the other hand, the networks trained on CIFAR10 slightly suffer from overfitting, since they have a 16.3% discrepancy between the performances on the training and the test data. Therefore, the lower adversarial distance might be due to noise-like signals employed for prediction.

Another possible explanation of the discrepancy in the size of the adversarial perturbations between random and trained deep neural networks is that networks trained with stochastic gradient descent are known to be less robust to adversarial perturbations than networks trained with Bayesian inference (Duvenaud et al., 2016; Bekasov & Murray, 2018; Carbone et al., 2020). As shown in section 1, under the hypothesis that the target function is generated by a given random deep neural network, the classifier obtained from Bayesian training of the same network has the same properties as a function generated by the random network. Therefore we expect that, for the size of the adversarial perturbations, random deep neural networks are closer to deep neural networks trained with Bayesian inference than to deep neural networks trained with gradient descent.

From Corollary 1, we expect the portion of images that have at least one adversarial example within a given ℓ^1 distance to increase linearly with the distance. This finding is validated by results shown in Figure 3 which plots the adversarial distance by percentile (sorted smallest to largest distance). In the case of random images, the linear increase in adversarial distance by percentile is evident throughout most of the percentiles in the chart conforming closely to the linear fit (dotted line). Interestingly, this linear correlation is even observed in images in the training and test sets outside of the smallest and highest percentiles. For training and test set images, networks usually predicted labels with high confidence thus limiting the percentage of images falling at small distances from a classification boundary.

5. Discussion

We have studied the properties of adversarial examples for deep neural networks with random weights and biases and have proved that for any $p \geq 1$, the ℓ^p distance from the closest classification boundary of any given input is with high probability at least $\tilde{\Omega}\left(1/\sqrt{n}\right)$ times the ℓ^p norm of the input, where n is the dimension of the input space (Corollary 1 and Remark 5). This lower bound matches the upper bound of (Gilmer et al., 2018; Fawzi et al., 2018; Shafahi et al., 2019; Mahloujifar et al., 2019), and our result determines the universal scaling of the minimum size of adversarial perturbations. Under the hypothesis that the target function is generated by a given random deep neural network, our probabilistic robustness guarantee also applies to the same network trained with Bayesian inference.

We have validated our theoretical results with experiments

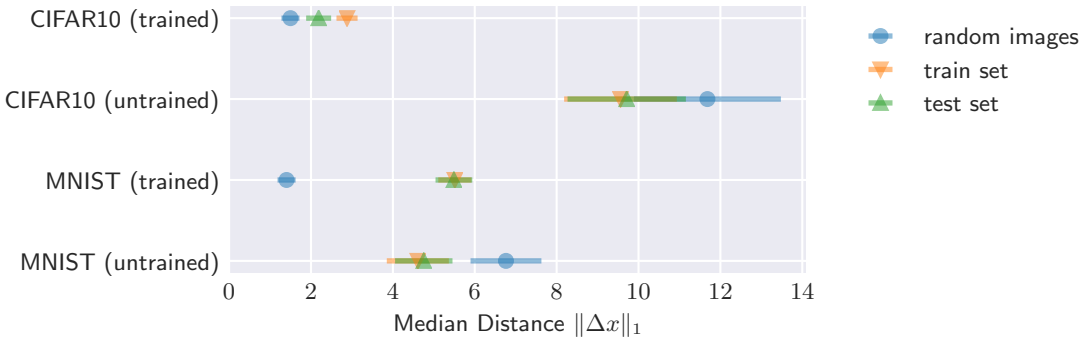

Figure 4. **Random vs trained networks:** Median distance of adversarial examples (in ℓ^1 norm) for random neural networks and neural networks of same architecture trained on MNIST and CIFAR10 data Figure 1. Analysis is performed for random images (images with randomly chosen pixel values) and images in the training and test sets. Network architecture is a simplified residual network (see Supplementary Manuscript).

on both random deep neural networks and deep neural networks trained with stochastic gradient descent on the MNIST and CIFAR10 datasets. The experiments on random networks are in complete agreement with our theoretical predictions. Networks trained on MNIST and CIFAR10 data are mostly consistent with our main findings, and we conjecture that the proof of our adversarial robustness guarantee can be extended to trained deep neural networks. Indeed, (Jacot et al., 2018; Lee et al., 2019; Yang, 2019a; Arora et al., 2019; Huang & Yau, 2020; Li et al., 2019b; Wei et al., 2019; Cao & Gu, 2019) have proved that the training of deep neural networks via stochastic gradient descent is similar to the training of Gaussian processes with Bayesian inference, and the classifier generated by a trained deep neural network still behaves as a Gaussian process. Therefore, we expect that the robustness of deep neural networks trained with stochastic gradient descent can be studied with similar techniques, and we believe that a probabilistic robustness guarantee similar to the guarantee proved in this paper can be proved. While the robustness guarantee for random deep neural networks does not require any assumption on the inputs, extensions to trained deep neural networks will definitely require assumptions on the training data. Given our results on random untrained networks extend directly to random networks trained via Bayesian inference under the hypothesis that the target classifier is drawn from the same random distribution, we conjecture that extending our results more broadly to trained networks will require that training labels look "typical" with respect to the probability distribution generated by the random initialization of the deep neural network at the beginning of the training.

The robustness guarantee of Corollary 1 depends on the architecture of the deep neural network through the ratio between the number of pixels of the output and of the input layer, favoring the case where such ratio is not small. We expect that with similar techniques the dependence of the bound on the architecture can be refined, thus allowing for a systematic study of how the choice of the model affects the adversarial robustness. Moreover, the extension of our results to deep neural networks trained with stochastic gradient descent would provide an analytic lower bound to the size of adversarial perturbations in terms of their architecture, and would therefore open the way to the first thorough theoretical understanding of the relationship between the network architecture and its robustness to adversarial attacks.

Finally, our methods can be employed to study the robustness of deep neural networks with respect to adversarial perturbations that keep the data manifold invariant, such as smooth deformations of the input image (Mallat, 2012; Oyallon & Mallat, 2015; Bruna & Mallat, 2013; Bietti & Mairal, 2019a;b).

Acknowledgements

We thank Milad Marvian, Dario Trevisan and Laurent Bétermin for useful discussions.

This work was supported by the USA Air Force Office of Scientific Research, the USA Army Research Office under the Blue Sky program, DOE and IARPA.

References

Adler, R. and Taylor, J. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer New York, 2009. ISBN 9780387481166.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely wide neural net. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and

Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/dbc4d84bfcfe2284ba11beffb853a8c4-Paper.pdf.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 284–293, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Bartlett, P. Uniform laws of large numbers. metric entropy. *Theoretical Statistics, Lecture Notes*, 14, March 2013.

Bekasov, A. and Murray, I. Bayesian adversarial spheres: Bayesian inference and adversarial examples in a noiseless setting. *arXiv preprint arXiv:1811.12335*, 2018.

Bietti, A. and Mairal, J. Group invariance, stability to deformations, and complexity of deep convolutional representations. *The Journal of Machine Learning Research*, 20(1):876–924, 2019a.

Bietti, A. and Mairal, J. On the inductive bias of neural tangent kernels. In *Advances in Neural Information Processing Systems*, pp. 12873–12884, 2019b.

Brown, T., Mané, D., Roy, A., Abadi, M., and Gilmer, J. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017.

Bruna, J. and Mallat, S. Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1872–1886, 2013.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 831–840, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Cao, Y. and Gu, Q. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 10835–10845, 2019.

Carbone, G., Wicker, M., Laurenti, L., Patane', A., Bortolussi, L., and Sanguinetti, G. Robustness of bayesian neural networks to gradient-based attacks. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 15602–15613. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/b3f61131b6eceeb2b14835fa648a48ff-Paper.pdf.

Cardelli, L., Kwiatkowska, M., Laurenti, L., Paoletti, N., Patane, A., and Wicker, M. Statistical guarantees for the robustness of bayesian neural networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5693–5700. International Joint Conferences on Artificial Intelligence Organization, 7 2019a. doi: 10.24963/ijcai.2019/789. URL https://doi.org/10.24963/ijcai.2019/789.

Cardelli, L., Kwiatkowska, M., Laurenti, L., and Patane, A. Robustness guarantees for bayesian inference with gaussian processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 7759–7768, 2019b.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. The loss surfaces of multilayer networks. In *Artificial Intelligence and Statistics*, pp. 192–204, 2015.

Cohen, J., Rosenfeld, E., and Kolter, Z. Certified adversarial robustness via randomized smoothing. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

De Palma, G., Kiani, B., and Lloyd, S. Random deep neural networks are biased towards simple functions. In *Advances in Neural Information Processing Systems*, pp. 1962–1974, 2019.

Duvenaud, D., Maclaurin, D., and Adams, R. Early stopping as nonparametric variational inference. In *Artificial Intelligence and Statistics*, pp. 1070–1077. PMLR, 2016.

Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., and Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. Robustness of classifiers: from adversarial to random noise. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 1632–1640. Curran Associates, Inc., 2016.

Fawzi, A., Fawzi, H., and Fawzi, O. Adversarial vulnerability for any classifier. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 1178–1187. Curran Associates, Inc., 2018.

Gao, R., Cai, T., Li, H., Hsieh, C.-J., Wang, L., and Lee, J. D. Convergence of adversarial training in overparametrized neural networks. In *Advances in Neural Information Processing Systems*, pp. 13009–13020, 2019.

Garriga-Alonso, A., Rasmussen, C. E., and Aitchison, L. Deep convolutional networks as shallow gaussian processes. In *International Conference on Learning Representations*, 2019.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. Adversarial spheres, 2018.

Gilmer, J., Ford, N., Carlini, N., and Cubuk, E. Adversarial examples are a natural consequence of test error in noise. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2280–2289, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. Adaptive computation and machine learning. MIT Press, 2016. ISBN 9780262035613.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Huang, J. and Yau, H.-T. Dynamics of deep neural networks and neural tangent hierarchy. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4542–4551. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/huang20l.html.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 125–136. Curran Associates, Inc., 2019.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8580–8589, 2018.

Jordan, M., Lewis, J., and Dimakis, A. G. Provable certificates for adversarial examples: Fitting a ball in the union of polytopes. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 14082–14092. Curran Associates, Inc., 2019.

Kim, B., Seo, J., and Jeon, T. Bridging adversarial robustness and gradient interpretability. *arXiv preprint arXiv:1903.11626*, 2019.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pp. 99–112. Chapman and Hall/CRC, 2018.

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436, 2015.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672. IEEE, 2019.

Ledoux, M. and Talagrand, M. *Probability in Banach Spaces: Isoperimetry and Processes*. Classics in Mathematics. Springer Berlin Heidelberg, 2013. ISBN 9783642202124.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide neural networks of any depth evolve as linear models under gradient descent. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/0d1a9651497a38d8b1c3871c84528bd4-Paper.pdf.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 32*, pp. 9464–9474. Curran Associates, Inc., 2019a.

Li, Z., Wang, R., Yu, D., Du, S. S., Hu, W., Salakhutdinov, R., and Arora, S. Enhanced convolutional neural tangent kernels. *arXiv preprint arXiv:1911.00809*, 2019b.

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4536–4543, 2019.

Mallat, S. Group invariant scattering. *Communications on Pure and Applied Mathematics*, 65(10):1331–1398, 2012.

Matthews, A. G. d. G., Rowland, M., Hron, J., Turner, R. E., and Ghahramani, Z. Gaussian process behaviour in wide deep neural networks. *arXiv preprint arXiv:1804.11271*, 2018.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529, 2015.

Nakkiran, P. Adversarial robustness may be at odds with simplicity. *arXiv preprint arXiv:1901.00532*, 2019.

Neal, R. M. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, 1996.

Novak, R., Xiao, L., Bahri, Y., Lee, J., Yang, G., Abolafia, D. A., Pennington, J., and Sohl-dickstein, J. Bayesian deep convolutional networks with many channels are gaussian processes. In *International Conference on Learning Representations*, 2019.

Oyallon, E. and Mallat, S. Deep roto-translation scattering for object classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2865–2873, 2015.

Pennington, J., Schoenholz, S., and Ganguli, S. The emergence of spectral universality in deep networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 1924–1932, 2018.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pp. 3360–3368, 2016.

Price, E. Maurey's empirical method. *CS395T: Sublinear Algorithms, Lecture Notes*, 13, October 2016.

Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018.

Raghunathan, A., Xie, S. M., Yang, F., Duchi, J. C., and Liang, P. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Rasmussen, C. and Williams, C. *Gaussian Processes for Machine Learning*. Adaptative computation and machine learning series. University Press Group Limited, 2006. ISBN 9780262182539.

Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31*, pp. 5014–5026. Curran Associates, Inc., 2018.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep information propagation. *arXiv preprint arXiv:1611.01232*, 2016.

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. Are adversarial examples inevitable? In *International Conference on Learning Representations*, 2019.

Shamir, A., Safran, I., Ronen, E., and Dunkelman, O. A simple explanation for the existence of adversarial examples with small hamming distance. *arXiv preprint arXiv:1901.10861*, 2019.

Sharif, M., Bhagavatula, S., Bauer, L., and Reiter, M. K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 acm sigsac conference on computer and communications security*, pp. 1528–1540, 2016.

Su, J., Vargas, D. V., and Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Tanay, T. and Griffin, L. A boundary tilting persepective on the phenomenon of adversarial examples. *arXiv preprint arXiv:1608.07690*, 2016.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

Wei, C., Lee, J. D., Liu, Q., and Ma, T. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pp. 9709–9721, 2019.

Wicker, M., Laurenti, L., Patane, A., and Kwiatkowska, M. Probabilistic safety for bayesian neural networks. In *Conference on Uncertainty in Artificial Intelligence*, pp. 1198–1207. PMLR, 2020.

Williams, C. K. Computing with infinite networks. In *Advances in neural information processing systems*, pp. 295–301, 1997.

Wong, E. and Kolter, Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5286–5295, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

Xiao, K. Y., Tjeng, V., Shafiullah, N. M. M., and Madry, A. Training for faster adversarial robustness verification via inducing reLU stability. In *International Conference on Learning Representations*, 2019.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5393–5402. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/xiao18a.html.

Yang, G. Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation. *arXiv preprint arXiv:1902.04760*, 2019a.

Yang, G. Wide feedforward or recurrent neural networks of any architecture are gaussian processes. In *Advances in Neural Information Processing Systems*, pp. 9947–9960, 2019b.