Proceedings of Machine Learning Research vol ???:1–12, 2025

Existence of Adversarial Examples for Random Convolutional Networks via Isoperimetric Inequalities on $\mathbb{SO}(d)$

Amit Daniely AMIT.DANIELY@MAIL.HUJI.AC.IL
Hebrew University and Google Research

Abstract

We show that adversarial examples exist for various random convolutional networks, and furthermore, that this is a relatively simple consequence of the isoperimetric inequality on the special orthogonal group $\mathbb{SO}(d)$. This extends and simplifies a recent line of work (Daniely and Shacham, 2020; Bubeck et al., 2021; Bartlett et al., 2021; Montanari and Wu, 2023) which shows similar results for random fully connected networks.

Keywords: Adversarial Examples, Convolutional Networks, Isoperimetric Inequalities

1. Introduction

Adversarial examples, first observed by Szegedy et al. (2014), were studied extensively in recent years, with several attacks (e.g. Athalye et al. (2018); Carlini and Wagner (2017, 2018); Goodfellow et al. (2014); Grosse et al. (2017)) and defense methods (e.g. Papernot et al. (2016, 2017); Madry et al. (2017); Wong and Kolter (2018); Feinman et al. (2017)) being developed, as well as various attempts to explain why they exists (e.g. Fawzi et al. (2018); Shafahi et al. (2018); Shamir et al. (2019); Schmidt et al. (2018); Bubeck et al. (2019)). In particular, one line of work aims at explaining the phenomenon of adversarial examples by proving that they exists and can be found in random networks. Daniely and Shacham (2020) show that adversarial examples exists in random constant depth fully connected ReLU networks in which each layer reduces the width. Moreover, they showed that gradient flow, as well as gradient decent with sufficiently small step size, will find these adversarial examples. Bartlett et al. (2021) following Bubeck et al. (2021) improved on Daniely and Shacham (2020) as they replaced the assumption that the dimension decreases with a very mild assumption that there is no exponential gap between the width of different layers. They also showed that gradient decent with a single step will find an adversarial example. Montanari and Wu (2023) further improved these results, as they completely dropped the width requirement.

We continue this line of work. Our contribution is twofold. First, we extend the family of architectures for which these results are applicable. We show the existence of adversarial examples in random *convolutional* ReLU networks of any constant depth, with no restriction on the width. For odd activations (such as sigmoids like $\sigma(x) = \tan^{-1}(x)$ and $\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$) we show that adversarial examples exists for a broader class of architectures: We show that adversarial examples exists provided that the first layer is convolutional. This result is valid for any layered architecture in the remaining layers, with no restriction on the width nor on the depth. Our second contribution is that we substantially simplify the proofs of such results, and show that the existence of adversarial examples is a relatively simple consequence of the isoperimetric inequality on the special orthogonal group $\mathbb{SO}(d)$. On the flip side as opposed to previous papers in this line of research, our techniques are less constructive, and we do not present an algorithm for finding an adversarial perturbation.

1.1. Related Work

Several recent theoretical studies have explored the fundamental reasons behind the existence of adversarial examples in machine learning. Schmidt et al. (2018) demonstrate that training adversarially robust classifiers can require a significantly larger sample complexity compared to standard training, while Bubeck et al. (2019) highlight scenarios where adversarially robust training is computationally more demanding.

Fawzi et al. (2018); Mahloujifar et al. (2019) leverage concentration of measure results to show that for various subsets of $\mathbb{R}^d$, such as the sphere, ball, or cube, any partition of the space into a small number of subsets with non-negligible measure (with respect to the uniform distribution) will necessarily lead to an abundance of adversarial examples. In other words, most points will have a nearby example that belongs to a different subset of the partition, implying that *any* classifier implementing such a partition will be susceptible to adversarial examples. Shafahi et al. (2018) extend these findings to classification tasks where examples are generated by specific generative models.

Further, Vardi et al. (2022) and Melamed et al. (2023) establish the existence of adversarial examples in trained depth-two neural networks. Lastly, Shamir et al. (2019) investigate adversarial vulnerability with respect to the ℓ^0 norm.

2. Setting

NOTATION

We will use $M_{d\times n}$ to denote the space if $d\times n$ matrices. We will use the standard Euclidean/Frobenius norm on the spaces $\mathbb{R}^d$, $\left(\mathbb{R}^d\right)^n$ and $M_{d\times n}$. We will denote the spectral norm of $A \in M_{d\times n}$ by $\|A\|_{\text{sp}}$. Similarly, for a sequence $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \left(\mathbb{R}^d\right)^n$ of n vectors in $\mathbb{R}^d$ we will denote by $\|\mathbf{x}\|_{\text{sp}}$ the spectral norm of the $d \times n$ matrix whose i'th column is $\mathbf{x}_i$. For a function $\sigma : \mathbb{R} \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^d$ we denote by $\sigma(\mathbf{x})$ the vector $(\sigma(x_1), \ldots, \sigma(x_d))$. We will use $\gtrsim$ for denoting inequality up to a multiplicative constant.

CONVOLUTIONAL NETWORKS

A *layer* is a function $F : (\mathbb{R}^{d_1})^n \to \mathbb{R}^{d_2}$ of the form $F(\mathbf{x}) = \sigma(W\mathbf{x})$ for a $d_2 \times (d_1 n)$ matrix W and $\sigma : \mathbb{R} \to \mathbb{R}$ (that is called the *activation function* of the layer). A layered network is a composition of several layers. A layer is *convolutional of width w, stride s and d_1 channels*, for $w \leq n$ such that s divides $n - w$, if it is of the form $F_W(\mathbf{x}) = (\sigma(WT_{0,s,w}(\mathbf{x})), \ldots, \sigma(WT_{\frac{n-w}{s},s,w}(\mathbf{x})))$ where $T_{i,s,w}(\mathbf{x}) = (\mathbf{x}^{is+1}, \ldots, \mathbf{x}^{is+w})$ and W is a $d_2 \times (d_1 w)$ matrix. We note that for the sake of simplicity, we consider one-dimensional convolutions, despite that our results can be phrased for general convolutional layers. We also note that the definition of convolutional layer encompasses fully connected layers as well, by taking the width w to be n.

RANDOM CONVOLUTIONAL NETWORKS

A *random convolutional layer* is a random function F_W where W is a random $d_2 \times (d_1 w)$ matrix. We say that W is *regular*, if for any orthogonal $U \in M_{(d_1 w),(d_1 w)}$ the distribution of W and WU are identical. We say that W is *Xavier* Glorot and Bengio (2010) if its entries are i.i.d. centered Gaussians with variance $\frac{1}{d_1 w}$ (note that a Xavier matrix is necessarily regular). We say that F_W

is *regular/Xavier* if W is a regular/Xavier random matrix. We note that it is common to initialize neural networks with regular random matrices. For instance Xavier matrices and random orthogonal matrices are standard choices for initial weights.

3. Main Results

Our first result assumes that the activation functions is odd (that is, satisfy $\sigma(-x) = -\sigma(x)$). Examples to such activations are sigmoids like $\sigma(x) = \tan^{-1}(x)$ and $\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Under this assumption, we can show that adversarial examples exist in a quite general setting. That is, we show that adversarial examples exist if the random network is a regular random convolutional layer, followed by *any* layered network (with no restriction on the width, depth, etc.).

Theorem 1 *Let $f : \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ be a random layered network whose first layer is a regular random convolutional layer that is independent from the remaining layers. Assume that the activations in all layers are odd and that d is even. Fix $\mathbf{x}_0 \in \left(\mathbb{R}^d\right)^n$. Then, w.p. $1 - 2e^{-\frac{\tau^2}{32}}$ over the choice of f either $f(\mathbf{x}_0) = 0$ or there is $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}_0\| \leq \left(\frac{\tau}{\sqrt{d-2}} \cdot \frac{\|\mathbf{x}_0\|_{\text{sp}}}{\|\mathbf{x}_0\|}\right) \cdot \|\mathbf{x}_0\|$ and $\text{sign}(f(\mathbf{x})) \neq \text{sign}(f(\mathbf{x}_0))$*

Two remarks are in order. The first is that for any $\mathbf{x}_0 \in \left(\mathbb{R}^d\right)^n$ we have $\frac{1}{\sqrt{\min(d,n)}} \leq \frac{\|\mathbf{x}_0\|_{\text{sp}}}{\|\mathbf{x}_0\|} \leq 1$, and that for "typical" (e.g. random) $\mathbf{x}_0$ we have $\frac{\|\mathbf{x}_0\|_{\text{sp}}}{\|\mathbf{x}_0\|} \approx \frac{1}{\sqrt{\min(d,n)}}$. Thus, Theorem 1 guarantees that for any $\mathbf{x}_0$, w.h.p. there is an adversarial example $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{x}_0\| \leq \frac{1}{\sqrt{d}}\|\mathbf{x}_0\|$, which essentially matches state of the art (Daniely and Shacham, 2020; Bartlett et al., 2021; Bubeck et al., 2021; Montanari and Wu, 2023) for fully connected networks ($n = 1$). For a "typical" $\mathbf{x}_0$ the guarantee is stronger and implies that there is an adversarial example $\mathbf{x}$ with $\|\mathbf{x} - \mathbf{x}_0\| \leq \frac{1}{\sqrt{\min(d^2,dn)}}\|\mathbf{x}_0\|$. The second remark is about the possibility that $f(\mathbf{x}_0) = 0$. We note that for a "reasonable" model of random networks, this probability is $o_d(1)$ or even 0. For such models, the conclusion of Theorem 1 implies that w.p. $(1 - o_d(1)\left(1 - 2e^{-\frac{\tau^2}{32}}\right)$ over the choice of f there is $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}_0\| \leq \left(\frac{\tau}{\sqrt{d-2}} \cdot \frac{\|\mathbf{x}_0\|_{\text{sp}}}{\|\mathbf{x}_0\|}\right) \cdot \|\mathbf{x}_0\|$ and $\text{sign}(f(\mathbf{x})) \neq \text{sign}(f(\mathbf{x}_0))$.

Our second result considers constant depth random convolutional networks with ReLU activation.

Theorem 2 *Let $f : (\mathbb{R}^d)^n \to \mathbb{R}$ be a random layered network with l independent convolutional Xavier layers with ReLU activation, except the last layer which has linear activation. Assume that the number of channels in each layer, as well as d are all $\omega(\log(nd))$. Fix $\mathbf{x}_0 \in \left(R \cdot \mathbb{S}^{d-1}\right)^n$. Then, w.p. $(1 - o_d(1))\left(1 - 2e^{-\Omega(\tau^2/\log^2(d))}\right)$ over the choice of f, either $f(\mathbf{x}_0) = 0$ or there is $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}_0\| \leq \left(\frac{\tau}{\sqrt{d-2}} \cdot \frac{\|\mathbf{x}_0\|_{\text{sp}}}{\|\mathbf{x}_0\|}\right) \cdot \|\mathbf{x}_0\|$ and $\text{sign}(f(\mathbf{x})) \neq \text{sign}(f(\mathbf{x}_0))$. The asymptotic notations depend only on the depth l.*

4. Isoperimetric inequalities for $\mathbb{SO}(d)$ and $\mathbb{SO}(d)$-metric-spaces

Let $\mathbb{SO}(d)$ the special orthogonal group. That is, $\mathbb{SO}(d)$ is the group of orthogonal matrices with determinant 1. We will view $\mathbb{SO}(d)$ as a metric space w.r.t. the metric induced by the Frobenius

norm. We note that this metric is $\mathbb{SO}(d)$-invariant in the sense that $d(UV_1, UV_2) = d(V_1, V_2)$ for any $U, V_1, V_2 \in \mathbb{SO}(d)$. Let μ be the uniform (Haar) probability measure on $\mathbb{SO}(d)$. We note that μ is the unique probability measure on $\mathbb{SO}(d)$ that is $\mathbb{SO}(d)$-invariant. That is, it is the unique probability measure on $\mathbb{SO}(d)$ that satisfies $\mu(UA) = \mu(A) = \mu(AU)$ for any measurable $A \subseteq \mathbb{SO}(d)$ and any $U \in \mathbb{SO}(d)$.

The results in this paper will make a crucial use of isoperimetric inequalities on $\mathbb{SO}(d)$, which we introduce next. To this end, for $A \subseteq \mathbb{SO}(d)$, $U \in \mathbb{SO}(d)$ and $\epsilon > 0$ we let

$$d(A, U) = \min_{V \in A} \|V - U\|$$

and

$$A_\epsilon = \{U \in \mathbb{SO}(d) : d(A, U) \leq \epsilon\}$$

Isoperimetric inequalities shows that for relatively small $\epsilon > 0$, $\mu(A_\epsilon) \gg \mu(A)$. These inequalities are based on the concentration of measure property on $\mathbb{SO}(d)$.

Theorem 3 (Measure concentration for $\mathbb{SO}(d)$. E.g. Meckes (2019) section 5.3) *Let $f : \mathbb{SO}(d) \to \mathbb{R}$ be L-Lipschitz w.r.t. the Frobenius metric. We have $\mu\left(f \geq \mathbb{E}_\mu f + \epsilon\right) \leq e^{-\frac{(d-2)\epsilon^2}{8L^2}}$*

Corollary 4 (Isoperimetric inequality for $\mathbb{SO}(d)$) *For $A \subseteq \mathbb{SO}(d)$ we have $\mu(A_\epsilon) \geq 1 - e^{-\frac{(d-2)\epsilon^2(\mu(A))^2}{8}}$*

Proof Define $f(x) = \max(0, \epsilon - d(A, x))$. We note that f is 1-Lipschitz. Also, $f(x) \geq \epsilon \cdot 1_A(x)$. Hence, we have $\mathbb{E}_\mu f \geq \epsilon \mathbb{E}_\mu 1_A = \mu(A)\epsilon$. This implies that

$$\mu\left(A_\epsilon^{\complement}\right) = \mu(f = 0) \leq \mu\left(f \leq \mathbb{E}_\mu f - \mu(A)\epsilon\right) \leq e^{-\frac{(d-2)\epsilon^2(\mu(A))^2}{8}}$$

■

We next extend the $\mathbb{SO}(d)$-isoperimetric inequalities to metric spaces on which $\mathbb{SO}(d)$ act. Let X be a metric space. We say that $\mathbb{SO}(d)$ *acts on* X if there is a mapping $(U, \mathbf{x}) \mapsto U\mathbf{x}$ such that for any $U, V \in \mathbb{SO}(d)$ and $\mathbf{x}, \mathbf{y} \in X$ we have (i) $V(U\mathbf{x}) = (VU)\mathbf{x}$, (ii) $I\mathbf{x} = \mathbf{x}$, and (iii) $d(U\mathbf{x}, U\mathbf{y}) = d(\mathbf{x}, \mathbf{y})$. We will refer to a metric space on which $\mathbb{SO}(d)$ acts as an $\mathbb{SO}(d)$-*metric-space*.

Examples of $\mathbb{SO}(d)$-metric-space are $\mathbb{R}^d$ an $\mathbb{S}^{d-1}$ (which are the input spaces for fully connected networks). Here, the action of $\mathbb{SO}(d)$ is standard matrix multiplication. Additional examples are $\left(\mathbb{R}^d\right)^n$ and $\left(\mathbb{S}^{d-1}\right)^n$ (which are the input spaces for convolutional networks). Here, the action is

$$U \cdot (\mathbf{x}_1, \ldots, \mathbf{x}_n) := (U\mathbf{x}_1, \ldots, U\mathbf{x}_n)$$

We say that an $\mathbb{SO}(d)$-metric-space is *transitive* if for any $\mathbf{x}, \mathbf{y} \in X$ there is $U \in \mathbb{SO}(d)$ such that $\mathbf{y} = U\mathbf{x}$. It is clear $\mathbb{S}^{d-1}$ is transitive and that for any $\mathbb{SO}(d)$-metric-space X and any $\mathbf{x} \in X$, the *orbit* of $\mathbf{x}$, i.e. $\mathcal{C}(\mathbf{x}) = \{U\mathbf{x} : U \in \mathbb{SO}(d)\}$, is transitive. On the other hand, $\mathbb{R}^d$ as well as $\left(\mathbb{R}^d\right)^n$ and $\left(\mathbb{S}^{d-1}\right)^n$ are not transitive. We define the Haar probability measure on a transitive $\mathbb{SO}(d)$-metric-space X as follows. Choose some $\mathbf{x} \in X$. For any $A \subseteq X$ we let

$$\mu_X(A) = \mu(\{U \in \mathbb{SO}(d) : U\mathbf{x} \in A\})$$

Note that μ_X in the l.h.s. denotes the Haar measure on X while in the r.h.s. μ denotes the Haar measure on $\mathbb{SO}(d)$. We note that the definition of μ_X does not depend on the choice of $\mathbf{x}$. Indeed, for any $\mathbf{y} \in X$, since X is transitive we have $\mathbf{y} = V\mathbf{x}$ for some $V \in \mathbb{SO}(d)$. Hence,

$$\begin{aligned}\mu(\{U \in \mathbb{SO}(d) : U\mathbf{y} \in A\}) &= \mu(\{U \in \mathbb{SO}(d) : UV\mathbf{x} \in A\}) \\ &= \mu(\{U \in \mathbb{SO}(d) : U\mathbf{x} \in A\} \cdot V^{-1}) \\ &\overset{\mu \text{ is } \mathbb{SO}(d)\text{-invariant}}{=} \mu(\{U \in \mathbb{SO}(d) : U\mathbf{x} \in A\})\end{aligned}$$

We also note that μ_X is $\mathbb{SO}(d)$-invariant. Indeed, for $A \subseteq X$, $V \in \mathbb{SO}(d)$ and $\mathbf{x} \in X$ we have

$$\begin{aligned}\mu_X(VA) &= \mu(\{U \in \mathbb{SO}(d) : U\mathbf{x} \in VA\}) \\ &= \mu(\{U \in \mathbb{SO}(d) : V^{-1}U\mathbf{x} \in A\}) \\ &= \mu(V \cdot \{U \in \mathbb{SO}(d) : U\mathbf{x} \in A\} \cdot) \\ &\overset{\mu \text{ is } \mathbb{SO}(d)\text{-invariant}}{=} \mu(\{U \in \mathbb{SO}(d) : U\mathbf{x} \in A\}) \\ &= \mu_X(A)\end{aligned}$$

And similarly $\mu_X(AV) = \mu_X(A)$.

We say that the action of $\mathbb{SO}(d)$ on an $\mathbb{SO}(d)$-metric-space X is *L-Lipschitz* (or that X is L-Lipschitz) if for any $\mathbf{x}$ the mapping $U \mapsto U\mathbf{x}$ is L-Lipschitz. We note that in the case that X is transitive, the action is L-Lipschitz if the mapping $U \mapsto U\mathbf{x}_0$ is L-Lipschitz for some $\mathbf{x}_0 \in X$. Indeed, in this case, given any $\mathbf{x} \in X$, there is $V \in \mathbb{SO}(d)$ such that $\mathbf{x} = V\mathbf{x}_0$. Now, the mapping $U \mapsto U\mathbf{x}$ is L-Lipschitz as the composition of the 1-Lipschitz mapping $U \mapsto UV$ followed by the 1-Lipschitz mapping $U \mapsto U\mathbf{x}_0$. We will use the following Lemma

Lemma 5 *Let $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \left(\mathbb{R}^d\right)^n$. We have that the action of $\mathbb{SO}(d)$ on $\mathcal{C}(\mathbf{x})$ is $\|\mathbf{x}\|_{\text{sp}}$-Lipschitz.*

Proof Let X be $d \times n$ the matrix whose i'th column in $\mathbf{x}_i$. Since $\mathcal{C}(\mathbf{x})$ is transitive it is enough to show that $U \mapsto U\mathbf{x}$ is $\|\mathbf{x}\|_{\text{sp}}$-Lipschitz. Indeed, for $U, V \in \mathbb{SO}(d)$ we have

$$\begin{aligned}\|U\mathbf{x} - V\mathbf{x}\| &= \|((U-V)\mathbf{x}_1, \ldots, (U-V)\mathbf{x}_n)\| \\ &= \|(U-V)X\| \\ &\leq \|U-V\| \cdot \|X\|_{\text{sp}}\end{aligned}$$

■

We will use the following generalization of the isoperimetric inequality on $\mathbb{SO}(d)$. To this end, for $A \subseteq X$, $\mathbf{x} \in X$ and $\epsilon > 0$ we let $d(A, \mathbf{x}) = \min_{\mathbf{y} \in A} d(\mathbf{y}, \mathbf{x})$ and $A_\epsilon = \{\mathbf{x} \in X : d(A, \mathbf{x}) \leq \epsilon\}$.

Theorem 6 (Isoperimetric inequality for $\mathbb{SO}(d)$-metric-spaces) *Let X be transitive $\mathbb{SO}(d)$-metric-space. Assume that the action of $\mathbb{SO}(d)$ is L-Lipschitz. Then, for any $A \subseteq X$ we have $\mu_X(A_\epsilon) \geq 1 - e^{-\frac{(d-2)\epsilon^2(\mu_X(A))^2}{8L^2}}$*

Proof Choose some $\mathbf{x}_0 \in X$ and let $f : \mathbb{SO}(d) \to X$ be the function $f(U) = U\mathbf{x}_0$. We note that f is L-Lipschitz and that $\mu(f^{-1}(C)) = \mu_X(C)$ for any $C \subseteq X$. Now we have

$$\begin{array}{rcl} \mu_X(A_\epsilon) & = & \mu(f^{-1}(A_\epsilon)) \\ & \overset{(f^{-1}(A))_{\epsilon/L} \subseteq f^{-1}(A_\epsilon)}{\geq} & \mu((f^{-1}(A))_{\epsilon/L}) \\ & \overset{\mathbb{SO}(d)\text{-isoperimetric inequality}}{\geq} & 1 - e^{-\frac{(d-2)\epsilon^2(\mu(f^{-1}(A)))^2}{8L^2}} \\ & = & 1 - e^{-\frac{(d-2)\epsilon^2(\mu_X(A))^2}{8L^2}} \end{array}$$

■

5. Proof of the Main Results

Let X be an L-Lipschitz $\mathbb{SO}(d)$-metric-space. For $f : X \to Y$ and $U \in \mathbb{SO}(d)$ we define the function $Uf : X \to Y$ by $(Uf)(\mathbf{x}) = f(U^{-1}\mathbf{x})$. Let f be a random function from X to $\mathbb{R}$. We say that f is $\mathbb{SO}(d)$-*invariant* if for any $U \in \mathbb{SO}(d)$ the distribution of Uf and f are identical. An example for an $\mathbb{SO}(d)$-invariant random function is a random convolutional or fully connected network.

Lemma 7 *Let $f : \left(\mathbb{R}^d\right)^n \to \mathbb{R}$ be a random layered network whose first layer is a regular random convolutional layer that is independent from the remaining layers. Then, f is $\mathbb{SO}(d)$-invariant.*

Proof We have $f = g \circ F_W$ where $F_W(\mathbf{x}) = (\sigma(WT_{1,s,w}(\mathbf{x})), \ldots, \sigma(WT_{\frac{n-w}{s},s,w}(\mathbf{x})))$ is a convolutional layer for a regular random matrix W that is independent of g. For $U \in \mathbb{SO}(d)$ We have

$$\begin{array}{rcl} Uf(\mathbf{x}) & = & g \circ F_W(U^{-1}\mathbf{x}) \\ & = & g\left(\sigma(WT_{1,s,w}(U^{-1}\mathbf{x})), \ldots, \sigma(WT_{\frac{n-w}{s},s,w}(U^{-1}\mathbf{x}))\right) \\ & = & g\left(\sigma(WU^{-1}T_{1,s,w}(\mathbf{x})), \ldots, \sigma(WU^{-1}T_{\frac{n-w}{s},s,w}(\mathbf{x}))\right) \\ & = & g \circ F_{WU^{-1}}(\mathbf{x}) \end{array}$$

That is $Uf = g \circ F_{WU^{-1}}$. This implies that Uf and $f = g \circ F_W$ are identically distributed: Since W is regular, W and WU^{-1} are identically distributed and hence also F_W and $F_{WU^{-1}}$. Since g is independent of W, we conclude that f and Uf are identically distributed as well. ■

We say that a random function $f : X \to \mathbb{R}$ is (q, p)-balanced if w.p. $\geq q$ over the choice of f we have that $\mu_X(f \geq 0) \geq p$ and $\mu_X(f \leq 0) \geq p$. We note that if f is a layered network with an odd activation, and $X = \mathcal{C}(\mathbf{x}_0)$ for some $\mathbf{x}_0 \in (\mathbb{R}^d)^n$ then f is $(1, 1/2)$-balanced. Indeed, if $A = \{\mathbf{x} : f(\mathbf{x}) \geq 0\}$ then, since f is odd, $A^{\complement} = -I \cdot A$. Hence,

$$\mu_X(f \geq 0) = \mu_X(A) \overset{\mu_X \text{ is } \mathbb{SO}(d) \text{ invariant}}{=} \mu_X(-I \cdot A) = \mu_X(A^{\complement}) = \mu_X(f \leq 0)$$

In the second equality from the left we relied on the fact that d is even and hence $-I \in \mathbb{SO}(d)$. The following Theorem shows that a random function that is balanced and $\mathbb{SO}(d)$-invariant has adversarial examples w.h.p.

Theorem 8 *Let X be transitive L-Lipschitz $\mathbb{SO}(d)$-metric-space and let $\mathbf{x}_0 \in X$. Let $f : X \to \mathbb{R}$ be a random function that is $\mathbb{SO}(d)$-invariant and (q,p)-balanced. Then, w.p. $q(1-2e^{-\frac{(d-2)\epsilon^2 p^2}{8L^2}})$ either $f(\mathbf{x}_0) = 0$ or there is $\mathbf{x}$ such that $d(\mathbf{x},\mathbf{x}_0) \leq \epsilon$ and $\operatorname{sign}(f(\mathbf{x})) \neq \operatorname{sign}(f(\mathbf{x}_0))$*

Before proving Theorem 8 we note that it implies Theorem 1. Indeed, we can take $X = \mathcal{C}(\mathbf{x}_0)$ which is the $\|\mathbf{x}_0\|$-Lipschitz (lemma 5), $\epsilon = \frac{\tau\|\mathbf{x}_0\|_{\text{sp}}}{\sqrt{d-2}}$ and f to be a random layered network with odd activations, whose first layer is a regular random convolutional layer that is independent from the remaining layers. Since f is $\mathbb{SO}(d)$-invariant and $(1,1/2)$-balanced, Theorem 8 implies that w.p. $1-2e^{-\frac{\tau^2}{32}}$ either $f(\mathbf{x}_0) = 0$ or there is $\mathbf{x}$ such that $d(\mathbf{x},\mathbf{x}_0) \leq \frac{\tau\|\mathbf{x}_0\|_{\text{sp}}}{\sqrt{d-2}}$ and $\operatorname{sign}(f(\mathbf{x})) \neq \operatorname{sign}(f(\mathbf{x}_0))$. This implies Theorem 1.

Theorem 8 also implies Theorem 2 via a similar argument. Take again $X = \mathcal{C}(\mathbf{x}_0)$, $\epsilon = \frac{\tau\|\mathbf{x}_0\|_{\text{sp}}}{\sqrt{d-2}}$, and let f to be a random convolutional ReLU random network as in Theorem 2. We have that f is $\mathbb{SO}(d)$-invariant (lemma 7). However, as opposed to random network with odd activations it is not immediate that f is balanced. In Lemma 10 below we do show that this is the case that f is $(1-o_d(1), 1/\log(d))$-balanced. Hence, Theorem 8 implies that w.p. $(1-o_d(1))\left(1-2e^{-\Omega(\tau^2/\log^2(d))}\right)$ either $f(\mathbf{x}_0) = 0$ or there is $\mathbf{x}$ such that $d(\mathbf{x},\mathbf{x}_0) \leq \frac{\tau\|\mathbf{x}_0\|_{\text{sp}}}{\sqrt{d-2}}$ and $\operatorname{sign}(f(\mathbf{x})) \neq \operatorname{sign}(f(\mathbf{x}_0))$. This implies Theorem 2.

Proof (of Theorem 8) Let $U \in \mathbb{SO}(d)$ be a random matrix. We can assume w.l.o.g. that $f = Ug$ where g is distributed as f and independent of U. Now, by conditioning on g, and since w.p. $\geq q$ we have $\mu_X(g \geq 0) \geq p$ and $\mu_X(g \leq 0) \geq p$, the Theorem follows from Lemma 9 below. ■

Lemma 9 *Let X be transitive L-Lipschitz $\mathbb{SO}(d)$-metric-space and let $\mathbf{x}_0 \in X$. Let $f : X \to \mathbb{R}$ be a function such that $\mu_X(f \geq 0) \geq p$ and $\mu_X(f \leq 0) \geq p$. Let $U \in \mathbb{SO}(d)$ be a random matrix. Then, w.p. $1-2e^{-\frac{(d-2)\epsilon^2 p^2}{8L^2}}$ either $Uf(\mathbf{x}_0) = 0$ or there is $\mathbf{x}$ such that $d(\mathbf{x},\mathbf{x}_0) \leq \epsilon$ and $\operatorname{sign}(Uf(\mathbf{x})) \neq \operatorname{sign}(Uf(\mathbf{x}_0))$*

Proof Let $A^+ = \{\mathbf{x} : f(\mathbf{x}) > 0\}$ and $A^- = \{\mathbf{x} : f(\mathbf{x}) < 0\}$. By Theorem 6 and the fact that $\mu_X(A^+) \geq p$ and $\mu_X(A^-) \geq p$ we have

$$\mu_X(A^+_\epsilon \cap A^-_\epsilon) \geq 1 - 2e^{-\frac{(d-2)\epsilon^2 p^2}{8L^2}}$$

Hence, w.p. $1-2e^{-\frac{(d-2)\epsilon^2 p^2}{8L^2}}$ over the choice of U we have that $U^{-1}\mathbf{x}_0 \in A^+_\epsilon \cap A^-_\epsilon$. It is enough to show that in this case either $Uf(\mathbf{x}_0) = 0$ or there is $\mathbf{x}$ such that $d(\mathbf{x},\mathbf{x}_0) \leq \epsilon$ and $\operatorname{sign}(Uf(\mathbf{x})) \neq \operatorname{sign}(Uf(\mathbf{x}_0))$.

Indeed, suppose that $f(U^{-1}\mathbf{x}_0) > 0$ (the case $f(U^{-1}\mathbf{x}_0) < 0$ is similar and the case $f(U^{-1}\mathbf{x}_0) = 0$ is immediate). Since $U^{-1}\mathbf{x}_0 \in A^-_\epsilon$ there is $\mathbf{y} \in A^-$ such that $d(\mathbf{y}, U^{-1}\mathbf{x}_0)$. Let $\mathbf{x} = U\mathbf{y}$. We have

$$d(\mathbf{x},\mathbf{x}_0) = d(U^{-1}\mathbf{x}, U^{-1}\mathbf{x}_0) = d(U^{-1}U\mathbf{y}, U^{-1}\mathbf{x}_0) = d(\mathbf{y}, U^{-1}\mathbf{x}_0) \leq \epsilon$$

Likewise,

$$Uf(\mathbf{x}) = f(U^{-1}\mathbf{x}) = f(U^{-1}U\mathbf{y}) = f(\mathbf{y}) \leq 0$$

Hence, $\operatorname{sign}(Uf(\mathbf{x})) \neq \operatorname{sign}(Uf(\mathbf{x}_0))$ ■

6. Balance-ness of random ReLU networks

In this section we will prove the following Lemma

Lemma 10 *Fix a constant l. Let $f:(\mathbb{R}^d)^n\to\mathbb{R}$ be a random layered network with l independent convolutional Xavier layers with ReLU activation, except the last layer which has linear activation. Assume that the number of channels in each layer, as well as d are all $\omega(\log(nd))$. Fix $\mathbf{x}_0\in\left(R\cdot\mathbb{S}^{d-1}\right)^n$. Then, $f|_{\mathcal{C}(\mathbf{x}_0)}$ is $(1-o_d(1),1/\log(d))$-balanced.*

6.1. Proof of Lemma 10

Since the ReLU is homogeneous, we can assume w.l.o.g. that $R=\sqrt{d}$. Let $\mathbf{x}^1,\ldots,\mathbf{x}^m$ be i.i.d. uniform points in $X=\mathcal{C}(\mathbf{x}_0)$ with $m=\lfloor\sqrt{\log(d)}\rfloor$. We say that f *separates* $\mathbf{x}^1,\ldots,\mathbf{x}^m$ if there are $1\le i,j\le m$ such that $f(\mathbf{x}^i)<0<f(\mathbf{x}^j)$. Let A_m be the event that f separates $\mathbf{x}^1,\ldots,\mathbf{x}^m$. We will show that

Lemma 11 $\Pr_{\mathbf{x}^1,\ldots,\mathbf{x}^m,f}(A_m)=1-o_d(1)$.

Before proving Lemma 11, we will explain how it implies Lemma 10. . We have

$$\begin{aligned}\Pr_{\mathbf{x}^1,\ldots,\mathbf{x}^m,f}(A_m) &= \Pr(f\text{ is }\beta\text{-balanced})\mathbb{E}_f\left[\Pr_{\mathbf{x}^1,\ldots,\mathbf{x}^m}(A_m)|f\text{ is }\beta\text{-balanced}\right]\\ &\quad+\Pr(f\text{ is not }\beta\text{-balanced})\mathbb{E}_f\left[\Pr_{\mathbf{x}^1,\ldots,\mathbf{x}^m}(A_m)|f\text{ is not }\beta\text{-balanced}\right]\\ &\le \Pr(f\text{ is }\beta\text{-balanced})+\mathbb{E}_f\left[\Pr_{\mathbf{x}^1,\ldots,\mathbf{x}^m}(A_m)|f\text{ is not }\beta\text{-balanced}\right]\\ &\le \Pr(f\text{ is }\beta\text{-balanced})+1-(1-\beta)^m\end{aligned}$$

Taking $\beta=1/m^2$ we conclude Lemma 10 as

$$\Pr(f\text{ is }1/\log(d)\text{-balanced})\ge\Pr(f\text{ is }1/m^2\text{-balanced})\ge\Pr_{\mathbf{x}^1,\ldots,\mathbf{x}^m,f}(A_m)-1+(1-1/m^2)^m=1-o_d(1)$$

Hence, it is enough to prove Lemma 11. We have that $f=F_{W_l}\circ\ldots\circ F_{W_1}$ where $F_{W_1},\ldots,F_{W_l}$ are independent Xavier convolutional layers with ReLU activation in all layers, except the last one (F_{W_l}) whose activation is the identity function. Assume that $F_{W_v}:\left(\mathbb{R}^{d_{v-1}}\right)^{n_{v-1}}\to\left(\mathbb{R}^{d_v}\right)^{n_v}$ is a convolutional layer of width w_v, stride s_v and d_v channels. Note that since the range of f is $\mathbb{R}$ we have $n_l=d_l=1$. Denote $\Psi=\sqrt{\frac{2^{l-1}}{n_{l-1}d_{l-1}}}F_{W_{l-1}}\circ\ldots\circ F_{W_1}$. We will prove Lemma 11 in three steps corresponding to the following three Lemmas

Lemma 12 *W.p. $1-o_d(1)$, for any $1\le i<j\le m$ and $1\le t\le n$, $\left\langle\mathbf{x}_t^i,\mathbf{x}_t^j\right\rangle\le d/2$.*

Lemma 13 *Given that for any $1\le i<j\le m$ and $1\le t\le n$, $\left\langle\mathbf{x}_t^i,\mathbf{x}_t^j\right\rangle\le d/2$. We have w.p. $1-o_d(1)$ that for any $1\le i<j\le m$ it holds that $\|\Psi(\mathbf{x}^i)-\Psi(\mathbf{x}^j)\|\ge\beta$ and $\|\Psi(\mathbf{x}^i)\|\le 2$, where $\beta>0$ is a constant depending only on the depth l*

Lemma 14 *Fix a constant $\beta>0$. Given that for any $1\le i<j\le m$ it holds that $\|\Psi(\mathbf{x}^i)-\Psi(\mathbf{x}^j)\|\ge\beta$ and $\|\Psi(\mathbf{x}^i)\|\le 2$ we have w.p. $1-o_d(1)$ that f separates $\mathbf{x}^1,\ldots\mathbf{x}^m$*

Lemmas 12, 13 and 14 clearly implies Lemma 11, so it remains to prove them. Lemma 12 is a simple consequence of the fact that if $\mathbf{x}, \mathbf{y} \in \sqrt{d}\mathbb{S}^{d-1}$ are uniform and independent then $\Pr(\langle \mathbf{x}, \mathbf{y} \rangle \geq t) \leq e^{-\frac{t^2}{2d}}$ (e.g. Vershynin (2018) chapter 3). Hence, the probability that for some $1 \leq i < j \leq m$ and $1 \leq t \leq n$, $\left\langle \mathbf{x}_t^i, \mathbf{x}_t^j \right\rangle > d/2$ is at most $2\binom{m}{2} n e^{-\frac{d}{8}} = o_d(1)$ where the last inequity is correct as we assume that $d = \omega(\log(n))$. It remains to prove Lemmas 13 and 14, which we do next

6.1.1. PROOF OF LEMMA 13

In order to prove Lemma 13 we will use a result of Daniely et al. (2016), which shows that w.h.p. $\langle \Psi(\mathbf{x}), \Psi(\mathbf{y}) \rangle \approx k(\mathbf{x}, \mathbf{y})$ where $k : \left(\sqrt{d} \cdot \mathbb{S}^{d-1}\right)^n \times \left(\sqrt{d} \cdot \mathbb{S}^{d-1}\right)^n \to \mathbb{R}$ is a kernel function with an explicit formula, which we define next. Let

$$\hat{\sigma}(u) = \frac{1}{\pi}\left(u(\pi - \arccos(u)) + \sqrt{1-u^2}\right)$$

We note that $\hat{\sigma}$ is non-negative and monotone in $[-1, 1]$, and satisfies $\hat{\sigma}(1) = 1$. For $\mathbf{x}, \mathbf{y} \in \left(\sqrt{d} \cdot \mathbb{S}^{d-1}\right)^n$, $0 \leq v \leq l-1$ and $1 \leq t \leq n_i$ we define recursively

$$k_{0,t}(\mathbf{x}, \mathbf{y}) = \frac{\langle \mathbf{x}_t, \mathbf{y}_t \rangle}{d} \text{ and } k_{v,t}(\mathbf{x}, \mathbf{y}) = \hat{\sigma}\left(\frac{1}{w_v}\sum_{r=1}^{w_v} k_{v-1, s_v(t-1)+r}(\mathbf{x}, \mathbf{y})\right)$$

Finally, let $k(\mathbf{x}, \mathbf{y}) = k_{l,1}(\mathbf{x}, \mathbf{y})$. The following Lemma shows that w.p. $1 - o_d(1)$, for any $1 \leq i, j \leq m$, $\left\langle \Psi(\mathbf{x}^i), \Psi(\mathbf{x}^i) \right\rangle = k(\mathbf{x}^i, \mathbf{x}^j) + o_d(1)$

Lemma 15 *Daniely et al. (2016) Suppose that for any $1 \leq v \leq l-1$ we have $d_i \gtrsim \frac{l^2 \log(ln/\delta)}{\epsilon^2}$ and that $\epsilon \lesssim \frac{1}{l}$. Then,*

$$\Pr\left(|k(\mathbf{x}, \mathbf{y}) - \langle \Psi(\mathbf{x}), \Psi(\mathbf{y}) \rangle| > \epsilon\right) < \delta$$

Next, we show by induction on v that if for any $1 \leq i < j \leq m$ and $1 \leq t \leq n$, $\left\langle \mathbf{x}_t^i, \mathbf{x}_t^j \right\rangle \leq d/2$ then $k_{v,t}(\mathbf{x}^i, \mathbf{x}^j) \leq \hat{\sigma}^{\circ v}(1/2) < 1$ for $i \neq j$ and $k_{v,t}(\mathbf{x}^i, \mathbf{x}^i) = 1$. Indeed,

$$\begin{aligned} k_{v,t}(\mathbf{x}^i, \mathbf{x}^j) &= \hat{\sigma}\left(\frac{1}{w_v}\sum_{r=1}^{w_v} k_{v-1, s_v(t-1)+r}(\mathbf{x}^i, \mathbf{x}^j)\right) \\ &\overset{\hat{\sigma} \text{ monotonicity and induction hypothesis}}{\leq} \hat{\sigma}(\hat{\sigma}^{\circ(v-1)}(1/2)) \\ &= \hat{\sigma}^{\circ v}(1/2) \end{aligned}$$

and

$$k_{v,t}(\mathbf{x}^i, \mathbf{x}^i) = \hat{\sigma}\left(\frac{1}{w_v}\sum_{r=1}^{w_v} k_{v-1, s_v(t-1)+r}(\mathbf{x}^i, \mathbf{x}^i)\right) \overset{\text{induction hypothesis}}{=} \hat{\sigma}(1) = 1$$

Hence, we have w.p. $1 - o_d(1)$ that $\|\Psi(\mathbf{x}_i)\|^2 = 1 + o_d(1)$ and that

$$\|\Psi(\mathbf{x}_i) - \Psi(\mathbf{x}_j)\|^2 = 1 - 2k(\mathbf{x}^i, \mathbf{x}^j) + 1 + o_d(1) \geq 2(1 - \hat{\sigma}^{\circ(l-1)}(1/2)) + o_d(1)$$

which proves Lemma 13

6.1.2. PROOF OF LEMMA 14

Theorem 16 (Sudakov e.g. Van Handel (2014) chapter 6.1) *Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$ be vectors such that $\|\mathbf{x}_i - \mathbf{x}_j\| \geq \alpha$ for any $1 \leq i < j \leq m$. Let $\mathbf{w} \in \mathbb{R}^d$ be standard Gaussian. Then, $\mathbb{E} \max_i \langle \mathbf{w}, \mathbf{x}_i \rangle \gtrsim \alpha \sqrt{\log(m)}$*

Let $Z = \max_{1 \leq i \leq m} f(\mathbf{x}^i)$. Given that for any $1 \leq i < j \leq m$ it holds that $\|\Psi(\mathbf{x}^i) - \Psi(\mathbf{x}^j)\| \geq \beta$ and $\|\Psi(\mathbf{x}^i)\| \leq 2$ we have that $\mathbb{E}Z = \omega(1)$ by Sudakov Lemma. Since $W_l \mapsto \max_i \left\langle W_l, \sqrt{d_{l-1} n_{l-1}} \Psi(\mathbf{x}_i) \right\rangle$ is 2-Lipchitz, and W_l is a matrix with i.i.d. centered Gaussians with variance $\frac{1}{d_{l-1} n_{l-1}}$, Gaussian concentration (e.g. Vershynin (2018) section 5.2.1) implies that $\mathrm{Var}(Z) \leq 4$. Hence, w.p. $1 - o_d(1)$, $Z > 0$, implying that there is i such that $f(\mathbf{x}^i) > 0$. Similarly, w.p. $1 - o_d(1)$ there is also i such that $f(\mathbf{x}^i) < 0$. This proves Lemma 14

Acknowledgments

The research described in this paper was funded by the European Research Council (ERC) under the European Union's Horizon 2022 research and innovation program (grant agreement No. 101041711), the Israel Science Foundation (grant number 2258/19), and the Simons Foundation (as part of the Collaboration on the Mathematical and Scientific Foundations of Deep Learning).

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Peter Bartlett, Sébastien Bubeck, and Yeshwanth Cherapanamjeri. Adversarial examples in multi-layer random relu networks. *Advances in Neural Information Processing Systems*, 34:9241–9252, 2021.

Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840, 2019.

Sébastien Bubeck, Yeshwanth Cherapanamjeri, Gauthier Gidel, and Remi Tachet des Combes. A single gradient step finds adversarial examples on random two-layers neural networks. *Advances in Neural Information Processing Systems*, 34:10081–10091, 2021.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017.

Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE, 2018.

Amit Daniely and Hadas Shacham. Most relu networks suffer from ℓ^2 adversarial perturbations. *Advances in Neural Information Processing Systems*, 33:6629–6636, 2020.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in neural information processing systems*, 29, 2016.

Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. In *Advances in Neural Information Processing Systems*, pages 1178–1187, 2018.

Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner. Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. 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, pages 4536–4543, 2019.

Elizabeth S Meckes. *The random matrix theory of the classical compact groups*, volume 218. Cambridge University Press, 2019.

Odelia Melamed, Gilad Yehudai, and Gal Vardi. Adversarial examples exist in two-layer relu networks for low dimensional linear subspaces. *Advances in Neural Information Processing Systems*, 36:5028–5049, 2023.

Andrea Montanari and Yuchen Wu. Adversarial examples in random neural networks with general activations. *Mathematical Statistics and Learning*, 6(1):143–200, 2023.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems*, pages 5014–5026, 2018.

Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance. *arXiv preprint arXiv:1901.10861*, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna Estrach, Dumitru Erhan, Ian Goodfellow, and Robert Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014*, 2014.

Ramon Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2(3):2–3, 2014.

Gal Vardi, Gilad Yehudai, and Ohad Shamir. Gradient methods provably converge to non-robust networks. *Advances in Neural Information Processing Systems*, 35:20921–20932, 2022.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295, 2018.

€