

Theoretical Lower Bounds on the Robustness of Deep ReLU Networks

Věra Kůrková

*Institute of Computer Science of the Czech Academy of Sciences
Pod Vodárenskou věží 2, 18207 Prague, Czech Republic*



Theoretical Lower Bounds on the Robustness of Deep ReLU Networks

Věra Kůrková

*Institute of Computer Science of the Czech Academy of Sciences
Pod Vodárenskou věží 2, 18207 Prague, Czech Republic*

Abstract

We present a theoretical study of the robustness of parameterized neural networks to random input perturbations. Specifically, we analyze local robustness by quantifying the probability that a random $\mathcal{L}_2$ -perturbation of a given input results in a correct classification. For deep ReLU networks, we derive lower bounds on local robustness by combining tools from high-dimensional geometry, in particular concentration of measure, with a new characterization of the geometric structure induced by their input-output functions. We prove that each convex polyhedral region in the partition of the input space induced by a ReLU network has at most as many faces as there are network units, regardless of the network depth or architecture. This geometric property serves as the key ingredient in our robustness analysis. Finally, we analyze how local robustness scales with the input dimension and characterize the sets of inputs whose neighborhoods are most likely to contain adversarial examples. We show that the width of neighborhoods of the decision boundaries containing vulnerable points shrinks rapidly with the increasing dimension and increases only logarithmically with the number of network units. We also discuss the volume of a set of vulnerable points in terms of approximately space-filling shapes of decision boundaries.

Keywords: Robustness of networks to input perturbations; Probability of misclassification; Adversarial patterns; Deep ReLU networks; High-dimensional geometry.

1. Introduction

The reliability of AI systems based on neural networks is crucial for their practical applicability. In particular, it is desirable that networks maintain their intended functionality even when exposed to changes or variations in their inputs. The susceptibility of parameterized networks to input perturbations has been observed in various cases caused either by corruption of data (due to

Email address: vera@cs.cas.cz (Věra Kůrková)

errors or noise) or by intentionally crafted perturbations. Szegedy et al. [1] initiated a study of imperceptible perturbations to test images leading to their misclassification. They called such images *adversarial* and conjectured that

"the set of adversarial negatives is of extremely low probability, and thus is never (or rarely) observed in the test set, yet it is dense (much like the rational numbers), and so it is found near virtually every test case."

The discovery of neural network vulnerability led to extensive research on algorithms constructing adversarial patterns that networks misclassify despite their indistinguishability from the training patterns (see [2] and references therein), finding minimal adversarial distortion for ReLU networks [3], and methods of defense against adversarial attacks (see, e.g., [4, 5]) and robustness of ReLU training for simultaneous robustness to all L_p perturbations [6]. Theoretical studies of adversarial patterns have mostly focused on characterization of the smallest additive perturbations that can change the outputs [7], on comparing the influence of the worst-case and average perturbations [5], and influence of Lipschitz continuity on robustness [8].

In this paper, we study the $\mathcal{L}_2$ -robustness of parameterized neural networks (either trained or randomly initialized) from a probabilistic perspective. We introduce the concept of *local robustness* at an input x with respect to a precision parameter r defined as the probability that the network classifies correctly an input obtained by applying an additive random perturbation of magnitude r to x . Such perturbed images remain within a Euclidean distance r from the original. For sufficiently small values of r , perturbed images are likely to be indistinguishable to the human eye from the training image.

Neural networks typically process data represented as high-dimensional vectors. Geometry of high-dimensional spaces exhibits some counterintuitive properties, called the "blessing of dimensionality" [9, 10, 11, 12], which make some tasks easier or predictable. These benefits of high-dimension are based on the "concentration of measure phenomenon" [13, 14, 15], which implies, e.g., the concentration of the volume and the surface of high-dimensional spheres [16], the reduction of dimensionality by random projections [17], and the exponential growth of the quasi-orthogonal dimension [18]. High-dimensional geometry also affects functionality and accuracy of neural networks (see e.g., [19, 20, 21, 22, 23]).

In this work, we investigate how input dimensionality influences the local robustness of parameterized deep ReLU networks. We derive probabilistic lower bounds on their local robustness in $x \in \mathbb{R}^d$ in terms of the input dimension, the total number of units, and the distance of x from the boundary of the partition induced by a ReLU network on the input space. We show that the lower bounds on robustness at inputs in $\mathbb{R}^d$ that are not in a close neighborhood of the decision boundary grow rapidly with increasing input dimensionality d , while exhibiting only a mild decrease with respect to the total number of units, independently of how these units are distributed across layers.

Our theoretical analysis is based on characterizing the geometric complexity of the sets of input-output functions of deep ReLU networks, exploiting their piecewise-linear, plane-wave structure. We prove that a key structural property

of the partitions induced by ReLU networks is that each region has only a “small” number of faces, bounded above by the total number of network units. By combining this characterization with concentration phenomena in high-dimensional geometry, we derive probabilistic lower bounds on local robustness. Our estimates characterize neighborhoods of decision boundaries in which inputs are more likely to be vulnerable to random $\mathcal{L}_2$ perturbations. We further investigate how the size of these neighborhoods scales with the input dimension and the total number of network units.

The remainder of the paper is organized as follows. Section 2 introduces the necessary background, notation, and preliminary concepts. Section 3 develops a probabilistic framework for analyzing robustness to misclassification caused by random additive perturbations. Within this framework, we first illustrate the fundamental role of data dimensionality using the example of classification by a single Heaviside perceptron, and then extend the probabilistic misclassification bounds to shallow perceptron networks. Section 4 includes structural analysis of sets of input-output functions of deep ReLU networks and exploits it to derive our main results on their robustness and analyze them. Finally, Section 5 concludes with a summary and discussion of the consequences of our results for the hypothesis of Szegedy et. al [1] that motivated our work.

2. Preliminaries

Feedforward networks compute parameterized families of input-output (I/O) functions determined by directed acyclic graphs, where nodes represent computational units and edges represent connections between them. In *multilayer networks*, units are arranged in layers, inputs in $X \subset \mathbb{R}^d$ are viewed as the layer 0, layers $l = 1, \dots, L - 1$ are called *hidden layers*, and the number of layers L is called the *network depth*. In this paper, we focus on networks where *connections to any layer are only from the preceding layer*, and where the *L -th layer contains only one output unit*.

Typical computational units apply a fixed *activation function* $\psi : \mathbb{R} \rightarrow \mathbb{R}$ to *affine functions with varying parameters*. The units compute functions of the form

$$\psi(v \cdot . + b) : \mathbb{R}^d \rightarrow \mathbb{R},$$

where $v \in \mathbb{R}^d$ is called a *weight vector*, $b \in \mathbb{R}$ a *bias*, and

$$v \cdot x = \sum_{i=1}^d v_i x_i$$

denotes the *scalar product*. These units have been called *perceptrons* because their form was inspired by a naive model of a neuron with hard (Heaviside, signum) or soft (hyperbolic tangent, logistic sigmoid) threshold functions as activations. Currently, the most popular activation function is the rectified linear function (ReLU) function due to its suitability for gradient-based learning.

We denote by $\vartheta : \mathbb{R} \rightarrow \{0, 1\}$ the *Heaviside function*

$$\vartheta(t) = 0 \quad \text{for } t \leq 0 \quad \vartheta(t) = 1 \quad \text{for } t > 0,$$

by $\rho : \mathbb{R} \rightarrow \mathbb{R}$ the *rectified linear function*

$$\rho(t) = 0 \quad \text{for } t \leq 0 \quad \rho(t) = t \quad \text{for } t > 0,$$

and for $v \in S^{d-1}$ and $b \in \mathbb{R}$, by

$$H_{v,b} := \{x \in \mathbb{R}^d \mid v \cdot x + b = 0\}$$

the *hyperplane* determined by v and b .

For our study, it is essential that functions computed by perceptrons with any activation function have the form of *plane waves*. They are constant on hyperplanes parallel to $H_{v,b}$. We denote

$$H_{v,b}^- := \{x \in \mathbb{R}^d \mid v \cdot x + b \leq 0\}$$

$$H_{v,b}^+ := \{x \in \mathbb{R}^d \mid v \cdot x + b > 0\}$$

the two *half-spaces* separated by $H_{v,b}$. Note that we define the half-spaces in such a way that the hyperplane $H_{v,b}$ is contained in the half-space $H_{v,b}^-$, so $H_{v,b}^-$ is closed and $H_{v,b}^+$ is open. The reason is that the function ρ is equal to zero on the half-line $(\infty, 0]$. Thus $\mathbb{R}^d = H_{v,b}^+ \cup H_{v,b}^-$ is a genuine partition.

For $x \in \mathbb{R}^d$, we denote by $p_{H_{v,b}}(x)$ the *orthogonal projection* of x onto $H_{v,b}$ and for $H \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$,

$$\|x - H\|_2 := \inf_{z \in H} \|x - z\|_2$$

the *Euclidean distance* of x from H , by $S_r^{d-1}(x)$ the *sphere* in $\mathbb{R}^d$ of radius r centered at x , and by $S^{d-1} := S_1^{d-1}(0)$ the *unit sphere* centered at 0.

We denote by

$$\mathcal{N} := \mathcal{N}(v_{l,j}, b_{l,j}, j = 1, \dots, n_l, l = 1, \dots, L-1, v_L, b_L) \quad (1)$$

the *parameterized L -layer network*, where each of the layers $l = 1, \dots, L-1$ contains n_l perceptrons with ReLU activations and parameters $v_{l,j}, b_{l,j}$, and the layer L contains one Heaviside perceptron computing $\vartheta(v_L \cdot y_{L-1} + b_L)$, where $y_{L-1} = (y_{L-1,1}, \dots, y_{L-1,n_{L-1}}) \in \mathbb{R}^{n_{L-1}}$ is the vector of outputs of n_{L-1} units in the $(L-1)$ -st layer, and by

$$f_{\mathcal{N}} : \mathbb{R}^d \rightarrow \{0, 1\}$$

the *input-output (I/O) function* of the network $\mathcal{N}$.

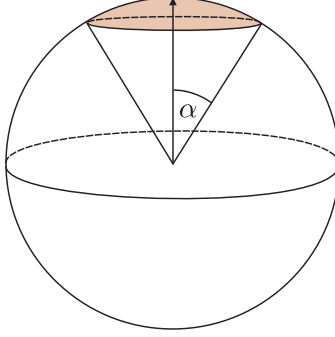


Figure 1: Spherical polar cap.

3. Probabilistic approach to robustness

For a parameterized network $\mathcal{N}$ (with parameters obtained by training or set randomly) that computes an I/O function $f_{\mathcal{N}}$, and for an input $x \in \mathbb{R}^d$, we study the local robustness of $\mathcal{N}$ at x in probabilistic terms. We assume that x is modified by an additive random perturbation ξ of a given size r , formally, by ξ drawn from the sphere $S_r^{d-1}(x)$ according to the uniform probability P . The Euclidean distance between x and $x + \xi$ is r , and therefore for small r , in the case of images, these two inputs might be indistinguishable.

We measure robustness of a parameterized network $\mathcal{N}$ by the probability that $f_{\mathcal{N}}$ maps $x + \xi$ to the same class as x . Formally, we define the *local robustness of $\mathcal{N}$ at x to perturbations of the size r* as

$$\mathcal{R}(f_{\mathcal{N}}, x, r) := P[f_{\mathcal{N}}(x + \xi) = f_{\mathcal{N}}(x)]. \quad (2)$$

To gain insight into the robustness of networks computing plane waves, we begin with the simplest case: a single Heaviside perceptron. Let

$$f(x) := \vartheta(v \cdot x + b)$$

be the I/O function of a parameterized perceptron with parameters $v \in \mathbb{R}^d$ and $b \in \mathbb{R}$, and $X \subset \mathbb{R}^d$ be the domain of its inputs.

Our probabilistic lower bound on its robustness relies on a counterintuitive geometric phenomenon in high dimensions: the fraction of the d -dimensional sphere occupied by a polar cap of fixed angular radius decreases exponentially fast with d (see Fig. 1).

Theorem 3.1. *Let $f : \mathbb{R}^d \rightarrow \{0, 1\}$, defined as $f(x) = \vartheta(v \cdot x + b)$, be an I/O function of a perceptron with the Heaviside activation ϑ , $r > 0$ be a size of a perturbation. Then for every $x \in \mathbb{R}^d$, $a(x) = \|p(x) - x\|_2$, where $p(x)$ is the orthogonal projection of x on the hyperplane $H_{v,b}$, and ξ drawn uniformly*

randomly from $S_r^{d-1}(x)$, we have:

(i) for $0 < a(x) \leq r$,

$$\mathcal{R}(f, x, r) := \mathbb{P}[f(x) = f(x + \xi)] > 1 - e^{-\frac{a(x)^2 d}{2r^2}};$$

(ii) for $a(x) = 0$,

$$\mathcal{R}(f, x, r) := \mathbb{P}[f(x) = f(x + \xi)] = 1/2;$$

(iii) for $r < a(x)$,

$$\mathcal{R}(f, x, r) := \mathbb{P}[f(x) = f(x + \xi)] = 1.$$

Proof. For $y \in S^{d-1}$ and $\alpha \in (0, \pi/2)$, let

$$C(y, \alpha) := \{z \in S^{d-1} \mid z \cdot y \geq \cos \alpha\}$$

denotes the *spherical polar cap* formed by the vectors from the sphere S^{d-1} with the angular distance to y at most α (see Fig. 1). Let μ denote the normalized Lebesgue measure on S^{d-1} , then for every $y \in S^{d-1}$ and every $\alpha \in (0, \pi/2)$, the polar cap $C(y, \alpha)$ satisfies

$$\mu(C(y, \alpha)) \leq e^{-\frac{(\cos \alpha)^2 d}{2}} \quad (3)$$

(see, e.g., [16, Lemma 2.2., p.11]).

Rescaling and translating $S_r^{d-1}(x)$ to S^{d-1} , we obtain from (3) for every $y \in S_r^{d-1}(x)$,

$$\mu_r\{\xi \in S_r^{d-1}(x) \mid |(x + \xi) \cdot y| \geq \cos \alpha\} \leq e^{-\frac{(\cos \alpha)^2 d}{2}}, \quad (4)$$

where μ_r denotes the normalized Lebesgue measure on $S_r^{d-1}(x)$.

(i) If $0 < a(x) < r$, then the perturbed input $x + \xi$ is separated from x by the hyperplane $H_{v,b}$ if it is in the polar cap on $S_r^{d-1}(x)$ centered at $y = x + r \frac{p(x) - x}{a(x)}$ with the angle $\alpha(x)$ such that $\cos \alpha(x) = \frac{a(x)}{r}$ (see Fig. 2). Thus from (4), we get for ξ drawn uniformly from $S_r^{d-1}(x)$

$$\mathbb{P}[\xi \in S_r^{d-1}(x) \mid f(x + \xi) \neq f(x)] \leq e^{-\frac{a(x)^2 d}{2r^2}}.$$

(ii) If $a(x) = 0$, then $x \in H_{v,b}$ and thus the probability $\mathbb{P}[x + \xi \in H_{v,b}^-] = \mathbb{P}[x + \xi \in H_{v,b}^+] = 1/2$.

(iii) If $r < a(x)$, then both x and $x + \xi$ are in the same open half-space separated by $H_{v,b}$. $\square$

Theorem 3.1 analyses the local robustness at inputs in dependence on their distance from the hyperplane $H_{v,b}$, the input dimension d , and perturbation

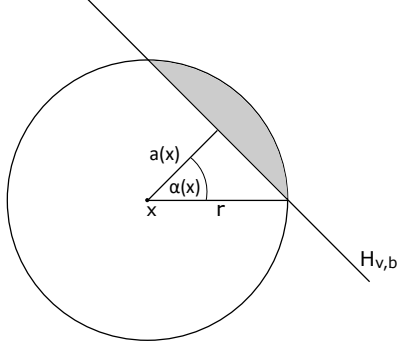


Figure 2: Area of perturbed patterns separated from x by a hyperplane.

size r . The cases (ii), when the original input x is farther from $H_{v,b}$ than the size of perturbation r , and (iii), when its distance from $H_{v,b}$ is equal to r , are obvious. In these two cases, the local robustness at x does not depend on the input dimension d .

Only the case (i), when the distance from the hyperplane $H_{v,b}$ is smaller than r , is substantive. The probability that a Heaviside perceptron correctly classifies a slightly perturbed pattern - lying at Euclidean distance r from the original pattern - is at least

$$1 - e^{-\frac{a(x)^2 d}{2r^2}}. \quad (5)$$

Thus the larger the quantity $a(x)^2 d$ relative to $2r^2$, the more robust the network is at x . The lower bound (5) demonstrates the beneficial effect of input dimensionality on the robustness of a Heaviside perceptron.

As a straightforward corollary of Theorem 3.1, we obtain the following probabilistic bound on the robustness of a shallow network with n Heaviside perceptrons.

Corollary 3.2. *Let $f(x) = \sum_{i=1}^n w_i \vartheta(v_i \cdot x + b_i) : \mathbb{R}^d \rightarrow \mathbb{R}$, where $v_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$, $i = 1, \dots, n$, be an I/O function of a parameterized shallow network with n Heaviside perceptrons, $H = \bigcup_{i=1}^n H_{v_i, b_i}$, and $r > 0$ be a size of a perturbation. Then for every $x \in \mathbb{R}^d \setminus H$ and ξ drawn uniformly randomly from $S_r^{d-1}(x)$, we have*

$$\mathcal{R}(f, x, r) := \mathbb{P}[f(x) = f(x + \xi)] \geq 1 - n e^{-\frac{a(x)^2 d}{2r^2}},$$

where $a(x) = \min\{a_i(x) \mid i = 1, \dots, n\}$, $a_i(x) = \|x - p_i(x)\|_2$, and $p_i(x)$ is the orthogonal projection of x on H_{v_i, b_i} for $i = 1, \dots, n$

Corollary 3.2 shows that the local robustness at almost all inputs in $\mathbb{R}^d$ (up to a set H of the Lebesgue measure zero) of a shallow Heaviside perceptron network increases rapidly with the input dimension, provided that the number of network units n does not dominate the term $e^{-\frac{a(x)^2 d}{2r^2}}$ (see Fig. 3). For

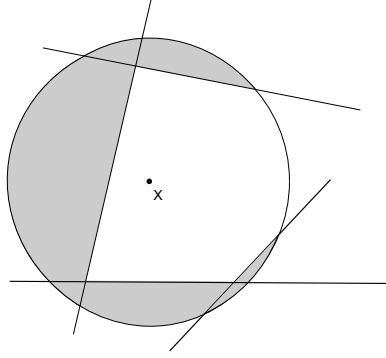


Figure 3: Robustness of a shallow Heaviside perceptron network

example, this condition is satisfied when the number of network units grows polynomially with d .

4. Robustness of deep ReLU networks

To extend our probabilistic estimates of local robustness to deep ReLU networks, we first show that their I/O functions partition the input domain into polyhedral regions (or convex in case of unbounded ones), each with only a “small” number of faces. The following theorem establishes that the number of faces bounding any individual region is at most the total number of network units. In contrast to the number of regions, which may grow exponentially with network depth, this upper bound holds regardless of network depth and unit arrangement across layers.

Theorem 4.1. *Let $\mathcal{N}(v_{l,j}, b_{l,j}, j = 1, \dots, n_l, l = 1, \dots, L-1, v_L, b_L)$ be a parameterized network with L layers, where the layers $l = 1, \dots, L-1$ contain n_l ReLU units with parameters $v_{l,j}, b_{l,j}$, $j = 1, \dots, n_l$, $n_L = 1$, the single unit in the L -th layer is the Heaviside perceptron with parameters v_L, b_L , and $n = \sum_{l=1}^L n_l$ denotes the total number of the network units. Then the I/O function $f_{\mathcal{N}} : \mathbb{R}^d \rightarrow \{0, 1\}$ of $\mathcal{N}$ is piecewise constant on a partition of $\mathbb{R}^d$ into convex subsets, such that each of them has at most n faces, and each face is contained in a hyperplane.*

Proof. For each layer $l = 1, \dots, L-1$, let $f_l : \mathbb{R}^d \rightarrow \mathbb{R}^{n_l}$ denote the function $f_l : \mathbb{R}^d \rightarrow \mathbb{R}^{n_l}$ defined as

$$f_l(x) = y_l(x) = (y_{l,1}(x), \dots, y_{l,n_l}(x)) \in \mathbb{R}^{n_l},$$

where $y_l(x)$ is the output vector of the l -th layer.

To obtain the partition, for each $x^* \in \mathbb{R}^d$, we construct a convex set $D_L(x^*)$, on which the I/O function f_N is constant, and then identify those sets $D_L(x^*)$ and $D_L(\bar{x}^*)$ that coincide.

We built the set $D_L(x^*)$ recursively from a sequence of nested convex subsets $\{D_l(x^*) \mid l = 1, \dots, L\}$ of $\mathbb{R}^d$, that all contain x^* . We construct them together with L families of hyperplanes

$$\mathcal{H}_l = \{H_{w_{k,j}, c_{k,j}} \mid k = 1, \dots, l, j = 1, \dots, n_k\}, \quad l = 1, \dots, L,$$

such that the faces of each of $D_l(x^*)$, $l = 1, \dots, L$, are formed by subsets of the hyperplanes $\{H_{w_{k,j}, c_{k,j}} \mid k = 1, \dots, l, j = 1, \dots, n_l\}$.

We start with the first layer $l = 1$. We set $w_{1,j} = v_{1,j}$ and $c_{1,j} = b_{1,j}$ for all $j = 1, \dots, n_1$ and define

$$D_1(x^*) = \bigcap_{j=1}^{n_1} H_{w_{1,j}, c_{1,j}}^{\iota_1},$$

where ι_1 is either $+$ or $-$ depending on whether $x^* \in H_{w_{1,j}, c_{1,j}}^+$ or $x^* \in H_{w_{1,j}, c_{1,j}}^-$. The function $f_1 : \mathbb{R}^d \rightarrow \mathbb{R}^{n_1}$, that maps input $x \in \mathbb{R}^d$ to the output vector of the 1-st hidden layer, satisfies for every $j = 1, \dots, n_1$,

$$f_1(x)_j = y_{1,j}(x) = \rho(w_{1,j} \cdot x + c_{1,j}).$$

Denoting

$$J_1^+ = \{j \in \{1, \dots, n_1\} \mid x^* \in H_{w_{1,j}, c_{1,j}}^+\}$$

and

$$J_1^- = \{j \in \{1, \dots, n_1\} \mid x^* \in H_{w_{1,j}, c_{1,j}}^-\},$$

we get for every $x \in D_1(x^*)$

$$f_1(x)_j = y_{1,j}(x) = w_{1,j} \cdot x + c_{1,j} \quad \text{when } j \in J_1^+$$

$$\text{and } f_1(x)_j = y_{1,j}(x) = 0 \quad \text{when } j \in J_1^-.$$

So, $D_1(x^*)$ is convex, its boundaries are formed by subsets of at most n_1 hyperplanes $H_{w_{1,j}, c_{1,j}}$, $j = 1, \dots, n_1$, and on $D_1(x^*)$, the functions $f_1(x)_j = y_{1,j}(x)$, $j = 1, \dots, n_1$, are affine or equal to zero.

Assuming that we have $D_{l-1}(x^*)$ together with the family of hyperplanes $\mathcal{H}_{l-1} = \{H_{w_{l-1,j}, c_{l-1,j}} \mid j = 1, \dots, n_{l-1}\}$, we construct a family of hyperplanes $\mathcal{H}_l = \{H_{w_{l,k}, c_{l,k}} \mid k = 1, \dots, n_l\}$ and $D_l(x^*)$. The function $f_l : \mathbb{R}^d \rightarrow \mathbb{R}^{n_l}$, that maps $x \in \mathbb{R}^d$ to the output vector

$$f_l(x) = y_l(x) = (y_{l,1}(x), \dots, y_{l,n_l}(x)) \in \mathbb{R}^{n_l}$$

of the l -th hidden layer, satisfies for every $k = 1, \dots, n_l$

$$f_l(x)_k = y_{l,k}(x) = \rho\left(v_{l,k} \cdot y_{l-1}(x) + b_{l,k}\right) = \rho\left(\sum_{j=1}^{n_{l-1}} v_{l,k,j} y_{l-1,j} + b_{l,k}\right) =$$

$$\rho\left(\sum_{j \in J_{l-1}^+} (v_{2,k,j} w_{l-1,j} \cdot x + c_{l-1,j}) + \sum_{j \in J_{l-1}^-} c_{l-1,j} + b_{2,k}\right).$$

Setting $w_{l,k} = \sum_{j \in J_{l-1}^+} v_{l,k,j} w_{l-1,j}$ and $c_{l,k} = \sum_{j \in J_{l-1}^-} c_{l-1,j} + b_{l,k}$, we get

$$y_{l,k}(x) = \rho(w_{l,k} \cdot x + c_{l,k}).$$

We define

$$D_l(x^*) = D_{l-1}(x^*) \cap \bigcap_{j=1}^{n_l} H_{w_{l,k}, c_{l,k}}^{\iota_1},$$

where ι_1 is either $+$ or $-$ depending on whether $x^* \in H_{w_{l,k}, c_{l,k}}^+$ or $x^* \in H_{w_{l,k}, c_{l,k}}^-$. Denoting

$$J_l^+ = \{k \in \{1, \dots, n_l\} \mid x^* \in H_{w_{l,k}, c_{l,k}}^+\}$$

and

$$J_l^- = \{k \in \{1, \dots, n_l\} \mid x^* \in H_{w_{l,k}, c_{l,k}}^-\},$$

we get for all $x \in D_l(x^*)$,

$$f_l(x)_k = y_{l,k}(x) = w_{l,k} \cdot x + c_{l,k} \quad \text{when } k \in J_l^+$$

$$f_l(x)_k = y_{l,k}(x) = 0 \quad \text{when } k \in J_l^-.$$

So, $D_l(x^*)$ is convex, its faces are formed by subsets of at most $n_1 + \dots + n_l$ hyperplanes $H_{w_{p,k}, b_{p,k}}$, $p = 1, \dots, l$, $k = 1, \dots, n_l$, and the components $f_l(\cdot)_k = y_{l,k}(x) : \mathbb{R}^d \rightarrow \mathbb{R}$ of the function $f_l : \mathbb{R}^d \rightarrow \mathbb{R}^{n_l}$ that maps $x \in \mathbb{R}^d$ to the output vector $y_l = (y_{l,1}, \dots, y_{l,n_l})$ are affine or constant zero.

The I/O function $f_{\mathcal{N}} : \mathbb{R}^d \rightarrow \{0, 1\}$ satisfies

$$f_{\mathcal{N}}(x) = f_L(x) = \vartheta(v_L \cdot y_{L-1}(x) + c_L),$$

where $y_{L-1}(x) = (y_{L-1,1}(x), \dots, y_{L-1,n_{L-1}}(x)) \in \mathbb{R}^{n_{L-1}}$. $D_{L-1}(x^*)$ was constructed in such a way, that for all $x \in D_{L-1}(x^*)$, we have

$$y_{L-1,s} = w_{L-1,s} \cdot x + c_{L-1,s} \quad \text{when } s \in J_{L-1}^+$$

$$\text{and } y_{L-1,s}(x) = 0 \quad \text{when } s \in J_{L-1}^-.$$

Thus for $x \in D_{L-1}(x^*)$

$$f_{\mathcal{N}}(x) = \vartheta\left(\sum_{s=1}^{n_{L-1}} v_{L,s} y_{L-1,s}(x) + b_L\right) =$$

$$\vartheta \left(\sum_{s \in J_{L-1}^+} (v_{L,s} w_{L-1,s} \cdot x) + \sum_{s \in J_{L-1}^-} c_{L-1,s} + b_L \right).$$

Setting $w_L = \sum_{s \in J_{L-1}^+} v_{L,s} w_{L-1,s}$ and $c_L = \sum_{s \in J_{L-1}^-} c_{L-1,s} + b_L$, we define

$$D_L(x^*) = D_{L-1}(x^*) \cap H_{w_L, c_L}^\iota,$$

where ι is either $+$ or $-$ depending on whether $x^* \in H_{w_L, c_L}^+$ or $x^* \in H_{w_L, c_L}^-$. As $D_L(x^*)$ is either a subset of the positive or the negative half-space induced by the hyperplane H_{w_L, c_L} , the I/O function f_N is constant on $D_L(x^*)$ (it is either 0 or 1 depending on ι).

It follows from the construction that for each pair x^* and $\bar{x}^*$, $D_L(x^*)$ and $D_L(\bar{x}^*)$ are either disjoint or coincide. Identifying coincident regions, we get the partition of $\mathbb{R}^d$. $\square$

Theorem 4.1 characterizes an I/O function of a parameterized deep ReLU network with a single Heaviside output as a piecewise constant function defined on a partition of $\mathbb{R}^d$ into convex regions, where each region has at most as many faces as the total number of units in the network. This result on bounded complexity property of each individual region measured by its number of faces is the key feature exploited in the next theorem.

Theorem 4.2. *Let $\mathcal{N}(v_{l,j}, b_{l,j}, j = 1, \dots, n_l, l = 1, \dots, L-1, v_L, b_L)$ be a parameterized network with L layers, where the layers $l = 1, \dots, L-1$ contain n_l ReLU units with parameters $v_{l,j}, b_{l,j}$, $j = 1, \dots, n_l$, $n_L = 1$ and the single unit in the L -th layer is the Heaviside perceptron with parameters v_L, b_L . Then there exists a subset H of $\mathbb{R}^d$ in the form of a union of subsets of hyperplanes, such that the I/O function $f_N : \mathbb{R}^d \rightarrow \{0, 1\}$ of $\mathcal{N}$ satisfies for every $x \in \mathbb{R}^d \setminus H$, every size of perturbation $r > 0$, and every ξ drawn from $S_r^{d-1}(x)$ according to the uniform probability P ,*

$$\mathcal{R}(f_N, x, r) := P[f_N(x) = f_N(x + \xi)] \geq 1 - n e^{-\frac{a(x)^2 d}{2r^2}},$$

where $n := \sum_{l=1}^L n_l$ is the total number of network units and $a(x) := \|x - H\|_2$.

Proof. By Theorem 4.1, there exists a partition of $\mathbb{R}^d$ into convex sets, each of which has at most n faces formed by subsets of hyperplanes. Let H be the union of faces of these convex sets. For each $x^* \in \mathbb{R}^d \setminus H$, f_N is constant on the set $D_L(x^*)$ from the partition constructed in the proof of Theorem 4.1 (see Fig. 4). On each of these sets, f_N is equal either to 0 or 1. For a fixed x^* , let $H_1^*, \dots, H_m^*$ be the hyperplanes, the parts of which form faces of $D_L(x^*)$. By Theorem 4.1, $m \leq n$.

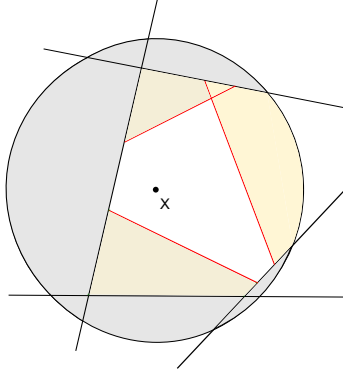


Figure 4: Probability of a misclassification by a deep ReLU network

Using properties of high-dimensional spheres as in the proof of Theorem 3.1, we get for each x in the interior of $D_L(x^*)$ and for each $i = 1, \dots, m$ and $a_i := \|x - H_i^*\|_2$ the upper bound

$$e^{-\frac{a_i(x)^2 d}{2r^2}},$$

on the probability that $x + \xi$ is separated from x by H_i^* . As for all $i = 1, \dots, m$, $a(x) = \|x - H\|_2 \leq \|x - H_i^*\|_2$, we get

$$\mathbb{P}[x + \xi \notin D_L(x^*)] \leq n e^{-\frac{a(x)^2 d}{2r^2}}.$$

□

Note that the bound from Theorem 4.2 is conservative in two ways: adjacent regions can have the same label, and $a(x)$ only roughly estimates a_i . This bound is non-vacuous for inputs $x \in \mathbb{R}^d$ satisfying

$$\frac{a(x)^2 d}{2r^2} > \ln n \iff a(x) > r \sqrt{\frac{2 \ln n}{d}}.$$

So the bound on robustness improves as the quantity

$$\frac{2r^2 \ln n}{d} - a(x)^2$$

becomes more negative. Note that the dependence of this robustness bound on the number of network units is only $\ln n$, whereas the dependence d is stronger, scaling as $\frac{1}{d}$.

In practical applications, networks are typically of moderate size, making it reasonable to assume that the total number of units grows at most polynomially

with d . Assuming that $n = d^k$, the condition characterizing the vulnerable neighborhood of the decision boundary becomes

$$a(x)^2 < \frac{2r^2 k \ln d}{d}.$$

So the neighborhood of H , in which inputs may be susceptible to misclassification, is of order $\sqrt{\frac{\ln d}{d}}$ which shrinks as the dimension d grows. For a fixed d and a perturbation radius r , one can derive a lower bound on the distance $a(x)$ of an input $x \in \mathbb{R}^d$ from H that guarantees that the probability of misclassification remains below a prescribed threshold. Equivalently, for any desired upper bound on the misclassification probability and any fixed perturbation radius r , one can determine a minimum distance from H such that the parameterized ReLU network achieves the required level of robustness for all inputs x with $a(x)$ above this threshold.

Although the set H has Lebesgue measure zero and can be described explicitly by computing the parameters $w_{l,j}$ and $c_{l,j}$ of the hyperplanes $H_{w_{l,j}, c_{l,j}}$ constructed in the proof of Theorem 4.1, the volume of the neighbourhood of $H \cap X$ (where $X \subset \mathbb{R}^d$ is bounded), consisting of all points within the distance

$$\sqrt{\frac{2r^2 \ln n}{d}},$$

depends on the geometry of H , which can be highly convoluted. ReLU networks realize shatterings on finite domains of size m characterized by the growth function $\Pi(m)$ satisfying

$$\Pi(m) \leq (4eLm)^{LW},$$

which grows exponentially with the network depth L and the total number of parameters W (see the proof of Theorem 7 in [24] and the discussion on bound tightness). Nevertheless, in practical applications, the decision boundary H is unlikely to exhibit a space-filling geometry. Consequently, as the input dimension d increases, the proportion of vulnerable points can be expected to decrease.

5. Conclusion

We theoretically investigated the susceptibility of neural networks to average $\mathcal{L}_2$ -perturbations by analyzing local robustness, quantified as the probability that a randomly perturbed input is not misclassified. We derived lower bounds on local robustness in terms of the input dimension, the perturbation size, and the distance to the boundaries of the convex cells in the partition of the input space induced by the I/O functions of ReLU networks. Our analysis combines concentration phenomena on high-dimensional spheres with a new estimate of the number of faces of each polyhedral region in this partition. In particular, we proved that although the number of regions may grow exponentially with network depth, each individual region is bounded by at most as many faces

as there are network units. Our results imply that the local robustness at an input $x \in \mathbb{R}^d$ depends on the quantity $\frac{2r^2 \ln n}{d} - a(x)^2$, where n denotes the total number of network units, r the perturbation radius, and $a(x)$ the distance of input x to the boundary set H .

Our proofs rely on the piecewise-linear structure induced by ReLU activations combined with tools from high-dimensional geometry. Consequently, our techniques do not directly extend to kernel methods or convolutional neural networks, whose robustness properties would require substantially different analytical tools.

We were motivated by the hypothesis of Szegedy et al. [1], quoted in the Introduction. Concerning the second part of their hypothesis, namely that adversarial examples may form a dense set (analogous to the rational numbers), our analysis shows that this cannot hold for ReLU networks. As H is closed and has Lebesgue measure zero, there are abundantly many points x and distances r , for which the balls of radius r centered at x do not intersect H . For such x , there are no misclassified inputs arbitrarily close to x . Our results also provide partial support for the first part of their hypothesis, namely that adversarial examples occur with low probability. We proved that if the total number of ReLU units grows at most polynomially with the input dimension d , then the width of the neighborhood of the boundary set H containing inputs susceptible to misclassification decreases of the order of $\frac{\ln d}{d}$. Consequently, unless H exhibits an approximately space-filling geometry, the probability that a randomly sampled input lies outside this narrow neighborhood is expected to increase with the input dimension.

In this paper, we focused on average-case adversarial robustness in the binary classification setting. Extending our analysis to worst-case robustness and multi-class classification with softmax outputs is a direction for our future work.

Acknowledgments

This work was partially supported by the Czech Science Foundation grant 25-15490S and the institutional support of the Institute of Computer Science RVO 67985807.

References

- [1] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks (2014). arXiv:1312.6199.
- [2] B. Wu, Z. Zhu, L. Liu, Q. Liu, Z. He, S. Lyu, Attacks in adversarial machine learning: A systematic survey from the life-cycle perspective (2024). arXiv:2302.09457.
- [3] T.-W. Weng, H. Zhang, H. Chen, Z. Song, C.-J. Hsieh, L. Daniel, D. Boning, I. S. Dhillon, Towards fast computation of certified robustness for relu networks, in: Proceedings of the 35th International Conference on Machine

- Learning, Vol. 80 of Proceedings of Machine Learning Research, 2018, pp. 5276–5285.
- [4] X. Yuan, P. He, Q. Zhu, X. Li, Adversarial examples: Attacks and defenses for deep learning, *IEEE Transactions on Neural Networks and Learning Systems* 30 (9) (2019) 2805–2824. doi:10.1109/TNNLS.2018.2886017.
 - [5] J. Gao, Z. Lu, R. Mudumbai, X. Wu, J. Yi, M. Cho, C. Xu, H. Xie, W. Xu, Feature compression is the root cause of adversarial fragility in neural network classifiers (2025). arXiv:2406.16200.
 - [6] F. Croce, M. Andriushchenko, M. Hein, Provable robustness of ReLU networks via maximization of linear regions, in: Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, Vol. 89 of Proceedings of Machine Learning Research, 2019, pp. 2057–2066.
 - [7] I. J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples (2015). arXiv:1412.6572.
 - [8] M. Zühlke, D. Kudenko, Adversarial robustness of neural networks from the perspective of Lipschitz calculus: A survey, *ACM Computing Surveys* 57 (6). doi:https://doi.org/10.1145/3648351.
 - [9] D. L. Donoho, High-dimensional data analysis: the curses and blessings of dimensionality, *AMS Math Challenges of the 21st Century* (2000) 1–32.
 - [10] A. N. Gorban, I. Tyukin, Blessing of dimensionality: Mathematical foundations of the statistical physics of data, *Philosophical Transactions of Royal Society A* 376 (2018) 2017–2037.
 - [11] A. N. Gorban, V. Makarov, I. Tyukin, The unreasonable effectiveness of small neural ensembles in high-dimensional brain, *Physics of Life Review* 9 (2019) 55–58.
 - [12] V. Kůrková, Some insights from high-dimensional spheres, *Physics of Life Review* 29 (2019) 98–100.
 - [13] P. Lévy, F. Pellegrino, *Problèmes concrets d’analyse fonctionnelle*, Gauthier-Villars, Paris, 1951.
 - [14] V. Milman, G. Schechtman, *Asymptotic theory of finite dimensional normed spaces*, Vol. 1200 of Lecture Notes in Mathematics, Springer-Verlag, 1986.
 - [15] O. Guédon, J. Prochno, Talagrand’s mathematical journey to the Abel prize 2024 (2024). arXiv:2410.07945.
 - [16] K. Ball, An elementary introduction to modern convex geometry, in: S. Levy (Ed.), *Flavors of Geometry*, Cambridge University Press, 1997, pp. 1–58.

- [17] W. B. Johnson, J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space, *Contemporary Mathematics* 26 (1984) 189–206.
- [18] P. C. Kainen, V. Kůrková, Quasiorthogonal dimension of Euclidean spaces, *Applied Mathematical Letters* 6 (3) (1993) 7–10.
- [19] V. Kůrková, M. Sanguineti, Model complexities of shallow networks representing highly varying functions, *Neurocomputing* 171 (2016) 598–604.
- [20] V. Kůrková, M. Sanguineti, Classification by sparse neural networks, *IEEE Transactions on Neural Networks and Learning Systems* 30(9) (2019) 2746–2754.
- [21] O. J. Sutton, Q. Zhou, A. N. Gorban, I. Y. Tyukin, Relative intrinsic dimensionality is intrinsic to learning, in: *Artificial Neural Networks and Machine Learning – ICANN 2023*, Springer Nature Switzerland, 2023, pp. 516–529.
- [22] A. Bastounis, A. N. Gorban, A. C. Hansen, D. J. Higham, D. V. Prokhorov, O. J. Sutton, I. Y. Tyukin, Q. Zhou, The boundaries of verifiable accuracy, robustness, and generalisation in deep learning, in: *Artificial Neural Networks and Machine Learning – ICANN 2023*, Springer Nature Switzerland, 2023, p. 530–541.
- [23] V. Kůrková, M. Sanguineti, Approximation of classifiers by deep perceptron networks, *Neural Networks* 165 (2023) 654–661. doi:<https://doi.org/10.1016/j.neunet.2023.06.004>.
- [24] P. L. Bartlett, N. Harvey, C. Liaw, A. Mehrabian, Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks, *Journal of Machine Learning Research* 20 (2019) 1–17.