Robustness of Overparameterized Deep Learning

Alex Zhang [1]
[1]Department of Computer Science, Princeton University
alzhang@princeton.edu

Contents

1 Introduction

Deep learning has become the ubiquitous technique for data-driven applications. However, neural networks are notoriously treated as blackbox objects, so a key question when using these models is how to characterize their *robustness* — i.e. can we ensure that (potentially adversarial) perturbations do not drastically influence the output of our learned models?

A natural characterization of robustness is the Lipschitz constant of a neural network. More formally, given an input domain $\mathcal{D}$ in some Euclidean space (e.g. $\mathbb{R}^d$), the Lipschitz constant of a neural network $f : \mathcal{D} \to \mathbb{R}$ is defined as

$$\text{Lip}\left(f(x)\right) \triangleq \sup_{x,y\in\mathcal{D},x\neq y} \frac{|f(x) - f(y)|}{\|x - y\|_2} \tag{1}$$

A suite of recent works have attempted to characterize the relationship between the *number of parameters* and the Lipschitz constant/robustness of general classes of neural networks [BS22]. A desirable property to have is to $O(1)$-Lipschitz networks because they are inherently *smooth* and provide guarantees against adversarial perturbations data. The relationship between the Lipschitz constant of a learned network and the number of learnable parameters it has is therefore of great interest. We look into lower bounds, which motivate the necessity of overparameterized models for a reasonable Lipschitz constant, and upper bounds, which derive the actual robustness of a network with respect to its number of parameters.

2 Background

2.1 Neural networks

We describe deep neural networks with d input neurons, 1 output neurons, and L hidden layers of width N as follows:

$$\Phi : \mathbb{R}^d \to \mathbb{R}, \quad \Phi(x) \triangleq \left(V^{(L)} \circ \sigma \circ V^{(L-1)} \circ ... \circ \sigma \circ V^{(0)}\right)(x) \tag{2}$$

where σ is some non-linear activation function and

$$V^{(\ell)}(x) \triangleq W^{(\ell)}x + b^{(\ell)} \tag{3}$$

where $W^{(0)} \in \mathbb{R}^{N\times d}, W^{(\ell)} \in \mathbb{R}^{N\times N}, b^{(\ell)} \in \mathbb{R}^{1\times N}, b^L \in \mathbb{R}$ are learnable components.

2.1.1 Activation functions

There are several non-linear activation functions used in neural networks. A common function is

$$\text{ReLU(x)} \triangleq \max\{0, x\}, \quad x \in \mathbb{R} \tag{4}$$

or

$$\text{Sigmoid(x)} \triangleq \frac{1}{1+e^{-x}}, \quad x \in \mathbb{R} \tag{5}$$

For the remaining background, we recall theorems and definitions described in [Han24] and [Ver18] and refer the reader to those sources for proofs unless stated otherwise.

2.2 ϵ-nets, Covering numbers

Definition 2.1 (ϵ-net)**.** A set N is an ϵ-net for a metric space (T, d) if for every $t \in T$, there exists $\pi(t) \in N$ such that $d(t, \pi(t)) \leq \epsilon$.

Definition 2.2 (Covering number)**.** For a metric space (T, d), the ϵ-net

$$\mathcal{N}(T, D, \epsilon) \triangleq \inf\{|N|\text{: } N \text{ is an } \epsilon\text{-net for } (T, d)\}$$

is called a covering set, and $|\mathcal{N}|$ is called the covering number.

Lemma 2.1 (Bounds on covering number over unit ball)**.** *On the unit Euclidean ball B_2^n, for any $\epsilon > 0$.*

$$\left(\frac{1}{\epsilon}\right)^n \leq \mathcal{N}(B_2^n, \|\cdot\|, \epsilon) \leq \left(1 + \frac{2}{\epsilon}\right)^n$$

The following proposition is found in [[Geu+24], D.4] and is a special case of Dudley's Entropy Integral ([Han24], Corollary 6.23).

Lemma 2.2 (Dudley's Entropy Inequality). *There exists a constant $C > 0$ such that for any $T \subseteq \mathbb{R}^d$ with $0 \in T$ and a random vector $X \in \mathbb{R}^d$ that is i.i.d. coordinate-wise distributed as $\sim \mathcal{N}(0,1)$, then for any $u \geq 0$, with probability $\geq 1 - 2exp(-u^2)$,*

$$\sup_{t \in T} \langle X, t \rangle \leq C \left(\int_0^\infty \sqrt{\ln(\mathcal{N}(T, \|\cdot\|_2, \epsilon))} d\epsilon + u \cdot diam(T) \right)$$

2.3 Subgaussian Random Variables

Definition 2.3. A random variable X is called σ^2 subgaussian if

$$\mathbb{P}[|X| \geq t] \leq 2 \exp\left(-t^2/\sigma^2\right)$$

Lemma 2.3. *Suppose $X_1, ..., X_n$ are independent σ^2-subgaussian random variables with mean 0. Then $Z = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ is $18\sigma^2$-subgaussian.*

2.4 Concentration Inequalities

Theorem 2.1 (Azuma-Hoeffding). *Let $\{\mathbb{F}_k\}_{k \leq n}$ be a filtration and Δ_k, A_k, B_k be such that*

1. *Δ_k is $\mathbb{F}_k$-measurable and $\mathbb{E}[\Delta_K | \mathbb{F}_{k-1}] = 0$.*
2. *A_k, B_k are $\mathbb{F}_{k-1}$ measurable and $A_k \leq \Delta_k \leq B_k$ almost surely.*

Then $\sum_{k=1}^n \Delta_k$ is $\frac{1}{4} \sum_{k=1}^n \|B_k - A_k\|_\infty^2$-subgaussian. In particular, combining with Definition 2.3, for all $t \geq 0$,

$$\mathbb{P}\left[\sum_{k=1}^n \Delta_k \geq t\right] \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^n \|B_k - A_k\|_\infty^2}\right)$$

2.5 Isoperimetry

Theorem 2.2 (c-isoperimetry). *A probability measure μ on a Euclidean space (say $\mathbb{R}^d$) is c-isoperimetric if for any bounded L-Lipschitz function f, we have that for $t \geq 0$,*

$$\mathbb{P}[|f - \mathbb{E}f| \geq t] \leq 2 \exp(-dt^2/2cL^2)$$

Put simply, the isoperimetry condition describes a class of data distributions where Lipschitz functions concentrate well. One may also notice that the above expression looks suspiciously similar to the subgaussian condition, and indeed under re-scaling, isoperimetry implies any Lipschitz function is $O(1)$-subgaussian. It turns out that a wide range of common distributions satisfy isoperimetry, such as high-dimensional Gaussian distributions and strongly log-concave measures in a normed space.

3 Preliminary bounds on shallow neural networks

A preliminary conjecture on the relationship between dataset size, data dimensionality, model size, and robustness was first introduced in [BLN20]. Here, they assume the data $x_1, ..., x_n$ are i.i.d uniform on the

sphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ with corresponding labels $y_1, ..., y_n$ in $\{1, -1\}$. Their analysis focuses on the case of two-layer *shallow* neural networks with p neurons of the form

$$f(\mathbf{x}) = \sum_{\ell=1}^{p} \alpha_\ell \cdot \sigma_\ell(\langle \mathbf{w}_\ell, \mathbf{x} \rangle + b_\ell) \tag{6}$$

where α_ℓ, β_ℓ are constants, $\mathbf{w}_1, ..., \mathbf{w}_p \in \mathbb{R}^d$ are learnable weights, and σ_ℓ is an $O(1)$-Lipschitz *activation* function. We denote the class of functions where $\sigma_\ell(x) = \max\{0, x\}$ as $\mathcal{F}_p$.

Theorem 3.1 (Existence of robust classifier)**.** *Let n denote the number of data points, d denote the dimension of the data, and p denote the number of parameters. For arbitrary constant C, suppose*

$$C \cdot \frac{n \log n}{d} \leq p \leq C \cdot n$$

Then over data points in $\mathbb{S}^{d-1}$, with probability $\geq 1 - \frac{1}{n^C}$, there exists an $f \in \mathcal{F}_p$ such that

$$f(x_i) = y_i \quad \textit{for all } i \in [n] \qquad \textit{(perfect classifier)}$$

$$\textit{Lip}(f) \leq C \cdot \frac{n \log d}{p} \qquad \textit{(robustness condition)}$$

Proof sketch: The full proof of Theorem 3.1 found in [BLN20] is quite dense, but the high-level idea is quite simple. For a sufficiently sized p, we can assign each neuron $w_\ell, \ell \in [p]$ to a specific cap of the sphere $\mathbb{S}^{d-1}$ that share the same label. We can ensure with high probability that this neuron ignores all other data points using the fact that $\sigma_\ell(x) = \max\{0, x\}$, so summing these neurons gives us a classifier in $\mathcal{F}_p$ that perfectly classifies the dataset. The proof concludes with a probabilistic bound over which neurons are non-zero with respect to an input point followed by a union bound over the points to show the Lipschitz condition (with high probability). □

However, there are many restrictive assumptions in Theorem 3.1 that can be peeled away. Firstly, the assumption of data that lies on $\mathbb{S}^{d-1}$ (the authors also argue that this extends well to Gaussian data) is quite restrictive and not necessarily representative of distributions we encounter in real life applications. Secondly, the restriction of analysis to shallow neural networks does not lend itself well to modern neural networks. We divide the remaining sections based on works that attempt to lower bound the Lipschitz constant (i.e. show that more parameters are **necessary but not sufficient** for robustness) and upper bound the Lipschitz constant (i.e. show that more parameters are **sufficient but not necessary** for robustness). Finally, the existence of a **perfect** classifier is often unnecessary, as most data-driven applications assume some level of noise.

4 Lower bounds on Lipschitz constant

Lower bound analysis of the Lipschitz constant attempts to formalize the **necessity of large neural networks** for robustness. In other words, they prove that neural networks of a small enough size **cannot** deal with adversarial perturbations. In [BS22], they prove a lower bound on the Lipschitz constant of a function class where the data distribution **is a convex combination of distributions** that obey c-isoperimetry. We first prove a weaker theorem that assumes the distribution satisfies c-isoperimetry before generalizing and include a few more steps from the original proof for clarity.

Lemma 4.1. *Suppose we have a dataset of i.i.d. points $x_1, ..., x_n \in \mathbb{R}^d$ with corresponding labels $y_1, ..., y_n \in \{1, -1\}$. For an arbitrary **finite** function class $\mathcal{F}$ and data distribution such that $\sigma^2 \triangleq \mathbb{E}_\mu [Var[y|x]] > 0$, it follows that*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8^3}\right) + \mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} f(x_i)(y_i - \mathbb{E}[y|x]) \geq \frac{\epsilon}{4}\right)$$

Proof. For notational convenience, define

$$g(x) \triangleq \mathbb{E}[y|x]$$
$$z_i \triangleq y_i - g(x_i)$$

Observe that

$$\mathbb{E}[z_i^2] = \mathbb{E}[(y_i - \mathbb{E}[y_i|x_i])^2] = \mathbb{E}[\mathbb{E}[(y_i - \mathbb{E}[y_i|x_i])^2|x_i]] = \mathbb{E}[\text{Var}(y_i|x_i)] = \sigma^2$$

Furthermore, because $y_i \in \{-1, 1\}$, by [[Han24], Lemma 2.1] we know $\sigma^2 \leq 1$, and therefore $|z_i|^2 \leq 4$. We can therefore apply Azuma-Hoeffding's inequality (Theorem 2.1) on the sequence $(\frac{1}{n} z_i^2)_{i \in [n]}$:

$$\mathbb{P}\left(\sigma^2 - \frac{1}{n}\sum_{i=1}^{n} z_i^2 \geq \frac{\epsilon}{6}\right) \leq \exp\left(-\frac{n\epsilon^2}{8 \cdot 6^2}\right) \leq \exp\left(-\frac{n\epsilon^2}{8^3}\right)$$

Re-writing, we get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} z_i^2 \leq \sigma^2 - \frac{\epsilon}{6}\right) \leq \exp\left(-\frac{n\epsilon^2}{8^3}\right) \tag{7}$$

Furthermore, observe that $\mathbb{E}[z_i g(x_i)] = g(x_i)\mathbb{E}[z_i] = 0$ and similar to the above reasoning, $|z_i g(x_i)| \leq 4$ (this is a loose bound, but it suffices for the proof). Thus, we obtain the same Azuma-Hoeffding's bound on the sequence $(\frac{1}{n} z_i g(x_i))_{i \in [n]}$

$$\mathbb{P}\left(-\frac{1}{n}\sum_{i=1}^{n} z_i g(x_i) \geq \frac{\epsilon}{6}\right) \leq \exp\left(-\frac{n\epsilon^2}{8 \cdot 6^2}\right) \leq \exp\left(-\frac{n\epsilon^2}{8^3}\right)$$

Re-writing, we get

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} z_i g(x_i) \leq -\frac{\epsilon}{6}\right) \leq \exp\left(-\frac{n\epsilon^2}{8^3}\right) \tag{8}$$

We relate these quantities to any function $f \in \mathcal{F}$ as follows. Consider the vectors

$$F = \left(\frac{1}{\sqrt{n}} f(x_1), \frac{1}{\sqrt{n}} f(x_2), ..., \frac{1}{\sqrt{n}} f(x_n)\right) \quad \text{for some } f \in \mathcal{F}$$
$$G = \left(\frac{1}{\sqrt{n}} g(x_1), \frac{1}{\sqrt{n}} g(x_2), ..., \frac{1}{\sqrt{n}} g(x_n)\right)$$
$$Z = \left(\frac{1}{\sqrt{n}} z(x_1), \frac{1}{\sqrt{n}} z(x_2), ..., \frac{1}{\sqrt{n}} z(x_n)\right)$$

We can therefore rewrite Equation 7 and Equation 8 as

$$\mathbb{P}\left(\langle Z, Z\rangle \leq \sigma^2 - \frac{\epsilon}{6}\right) \leq \exp\left(-\frac{n\epsilon^2}{8^3}\right)$$
$$\mathbb{P}\left(\langle Z, G\rangle \leq -\frac{\epsilon}{6}\right) \leq \exp\left(-\frac{n\epsilon^2}{8^3}\right)$$

Observe that if $\langle Z, Z\rangle > \sigma^2 - \frac{\epsilon}{6}$ and $\langle Z, G\rangle > -\frac{\epsilon}{6}$ (i.e. the events described in Equation 7 and Equation 8 do not hold), then for any $f \in \mathcal{F}$, if $\|Z + G - F\|^2 \geq \sigma^2 - \epsilon$, we have

$$\begin{aligned}
\sigma^2 - \epsilon &\geq \|Z + G - F\|^2 \\
&= \|Z\|^2 + 2\langle Z, G - F\rangle + \|G - F\|^2 \\
&= \|Z\|^2 + 2\langle Z, G\rangle - 2\langle Z, F\rangle + \|G - F\|^2 \\
&> \sigma^2 - \frac{\epsilon}{6} - \frac{\epsilon}{3} - 2\langle Z, F\rangle + \|G - F\|^2 && \text{(Replace with assumptions)} \\
&\geq \sigma^2 - \frac{\epsilon}{2} - 2\langle Z, F\rangle && \text{(Norms are positive)}
\end{aligned}$$

which implies the necessary condition that $\langle Z, F\rangle \geq \frac{\epsilon}{4}$. Since f was arbitrary, we only care about the existence of $f \in \mathcal{F}$ that satisfies the above conditions.

In other words, either the event $\langle Z, Z\rangle \leq \sigma^2 - \frac{\epsilon}{6}$ or $\langle Z, G\rangle \leq -\frac{\epsilon}{6}$ occurs, or a necessary condition for the event $\|Z + G - F\|^2 \geq \sigma^2 - \epsilon$ is $\langle F, Z\rangle \geq \frac{\epsilon}{4}$. We conclude using the union bound that

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \|Z + G - F\|^2 \leq \sigma^2 - \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8^3}\right) + \mathbb{P}\left(\exists f \in \mathcal{F} : \langle Z, F\rangle \geq \frac{\epsilon}{4}\right)$$

which expands to

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon\right) \leq 2\exp\left(-\frac{n\epsilon^2}{8^3}\right) + \mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} f(x_i)(y_i - \mathbb{E}[y|x]) \geq \frac{\epsilon}{4}\right)$$

as desired. □

The lemma above says nothing about the function class $\mathcal{F}$ itself or the data distribution and is therefore hard to apply in practice, so we move to a more specific condition that we can mold into the robustness condition.

Theorem 4.1 (Finite Lipschitz Class Error Bound)**.** *Suppose we have a dataset of i.i.d. points $x_1, ..., x_n \in \mathbb{R}^d$ with corresponding labels $y_1, ..., y_n \in \{1, -1\}$. If*

1. *The distribution μ of the x_i's satisfies c-isoperimetry.*
2. *$\sigma^2 \triangleq \mathbb{E}_\mu\left[Var[y|x]\right] > 0$*

*Then for a **finite** function class $\mathcal{F}$ of L-Lipschitz functions and for all $\epsilon > 0$,*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon\right) \leq 4\exp\left(-\frac{n\epsilon^2}{8^3}\right) + 2\exp\left(\log(|\mathcal{F}|) - \frac{\epsilon^2 nd}{10^4 cL^2}\right)$$

Proof. Without loss of generality, assume functions in $\mathcal{F}$ are in $[-1, 1]$ (if not, we can simply clip the values, improving both the Lipschitz constant bound and the fit to the labels).

First, consider the random variable

$$h(x_i) = \sqrt{\frac{d}{c}}\frac{f(x_i) - \mathbb{E}[f(x_i)]}{L}$$

which has mean 0 and is now a $\sqrt{\frac{d}{c}}$-Lipschitz function. By the isoperimetry condition of the x_i's (Theorem 2.2),

$$\mathbb{P}\left(|h(x_i)|\geq t\right) \leq 2\exp\left(-\frac{dt^2}{2c\left(\frac{d}{c}\right)}\right) = 2\exp\left(-t^2/2\right)$$

so $h(x_i)$ is 2-subgaussian. Furthermore, because $|z_i|\leq 2$,

$$\begin{aligned}\mathbb{P}\left(|h(x_i)z_i|\geq t\right) &\leq \mathbb{P}\left(2|h(x_i)|\geq t\right)\\ &= \mathbb{P}\left(|h(x_i)|\geq \frac{t}{2}\right)\\ &\leq 2\exp\left(-t^2/8\right)\end{aligned}$$

so $h(x_i)z_i$ is 8-subgaussian. Furthermore, by the tower property ([Dem21], Proposition 2.3.5),

$$\mathbb{E}[h(x_i)z_i] = \mathbb{E}[\mathbb{E}[h(x_i)z_i|x_i]] = 0$$

So by the subgaussian lemma (2.3), $\frac{1}{\sqrt{n}}\sum_{i=1}^n h(x_i)z_i$ is $8\cdot 18 = 144$-subgaussian. Thus,

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n h(x_i)z_i \geq t\right) \leq 2\exp(-t^2/12^2)$$

$$\mathbb{P}\left(\sqrt{\frac{d}{ncL^2}}\sum_{i=1}^n \left(f(x_i)-\mathbb{E}[f(x_i)]\right)z_i \geq t\right) \leq 2\exp(-t^2/12^2) \qquad \text{Plug in for h}$$

Now let $t = \sqrt{\frac{nd\epsilon^2}{64cL^2}}$.

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n \left(f(x_i)-\mathbb{E}[f(x_i)]\right)z_i \geq \frac{\epsilon}{8}\right) &\leq 2\exp\left(-\frac{nd\epsilon^2}{8^2\cdot 12^2 cL^2}\right)\\ &\leq 2\exp\left(-\frac{nd\epsilon^2}{10^4 cL^2}\right) \qquad (9)\end{aligned}$$

Similar to the previous analysis, we again consider two disjoint events that cover the entire sample space. Consider the disjoint events $E_1 \triangleq \{\exists f\in\mathcal{F} : \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(x_i)]z_i \geq \frac{\epsilon}{8}\}$ and $E_2 \triangleq \{\nexists f\in\mathcal{F} : \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(x_i)]z_i \geq \frac{\epsilon}{8}\}$. To relate our solution to Lemma 4.1, notice by the law of total probability that

$$\begin{aligned}\mathbb{P}\left[\exists f\in\mathcal{F} : \sum_{i=1}^n f(x_i)z_i \geq \frac{\epsilon}{4}\right] &\leq \alpha\mathbb{P}[E_1] + \mathbb{P}\left[\exists f\in\mathcal{F} : \frac{1}{n}\sum_{i=1}^n \left(f(x_i)-\mathbb{E}[f(x_i)]\right)z_i \geq \frac{\epsilon}{8}\right]\mathbb{P}[E_2]\\ &\leq \mathbb{P}[E_1] + \mathbb{P}\left[\exists f\in\mathcal{F} : \frac{1}{n}\sum_{i=1}^n \left(f(x_i)-\mathbb{E}[f(x_i)]\right)z_i \geq \frac{\epsilon}{8}\right]\end{aligned}$$

for some $0<\alpha<1$ (it represents the probability of an event that we do not specify because we throw it away). We can bound $\mathbb{P}[E_1]$ with Azuma-Hoeffding's inequality (Theorem 2.1) by noticing that $\mathbb{E}[f]\in$

$[-1, 1]$ and $|z_i| \leq 2$, so

$$\begin{aligned}\mathbb{P}[E_1] &= \mathbb{P}\left[\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[f(x_i)]z_i \geq \frac{\epsilon}{8}\right] \\ &\leq \mathbb{P}\left[\frac{1}{n}\left|\sum_{i=1}^{n} z_i\right| \geq \frac{\epsilon}{8}\right] \\ &\leq 2\exp\left(-n\epsilon^2/8^3\right)\end{aligned}$$

Finally, we can bound the second term using a simple union bound over each $f \in \mathcal{F}$, giving us

$$\begin{aligned}\mathbb{P}\left[\exists f \in \mathcal{F} : \sum_{i=1}^{n} f(x_i)z_i \geq \frac{\epsilon}{4}\right] &\leq 2\exp\left(-n\epsilon^2/8^3\right) + |\mathcal{F}|\cdot\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\left(f(x_i) - \mathbb{E}[f(x_i)]\right)z_i \geq \frac{\epsilon}{8}\right] \\ &\leq 2\exp\left(-n\epsilon^2/8^3\right) + 2|\mathcal{F}|\cdot\exp\left(-\frac{nd\epsilon^2}{10^4 cL^2}\right)\end{aligned}$$

where we plugged in Equation 9 at the last step. Finally, we replace the relevant term in Lemma 4.1 to obtain the desired result. □

A stronger condition for data drawn from arbitrary *convex combinations of isoperimetric distributions* rather than an isoperimetric distribution is proved in [BS22], which we state without proof. The main difference is the original subgaussian bound we obtain using the isoperimetric inequality and controlling the mixed distribution to obtain nicer bounds.

Theorem 4.2 (Generalization of Theorem 4.1 to Convex Combinations of Isoperimetric Distributions)**.** *Suppose we have a dataset of i.i.d. points $x_1, ..., x_n \in \mathbb{R}^d$ with corresponding labels $y_1, ..., y_n \in \{1, -1\}$. There exists absolute constants C_1, C_2 such that if*

1. *The distribution μ of the x_i's can be written as a convex sum of distributions μ_j that satisfy c-isoperimetry.*
2. *$\sigma^2 \triangleq \mathbb{E}_\mu\left[Var[y|x]\right] > 0$*
3. *The dimension $d \geq C_1 \cdot \left(\frac{cL^2\sigma^2}{\epsilon^2}\right)$.*

*Then for a **finite** function class $\mathcal{F}$ of L-Lipschitz functions and for all $\epsilon > 0$,*

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon\right) \leq (4k+1)\exp\left(-\frac{n\epsilon^2}{8^3 k}\right) + \exp\left(\log(|\mathcal{F}|) - \frac{\epsilon^2 nd}{C_2 cL^2\sigma^2}\right)$$

The above gives us enough tooling to prove the main theorem.

Remark. The theorem below proved in [BS22] seems rather loose. I tried to keep the constants as tight as possible in the proof below, which is why the proofs/results may slightly differ.

Theorem 4.3 (Robustness-Lipschitz Lower Bound)**.** *Suppose we have a dataset of i.i.d. points $x_1, ..., x_n \in \mathbb{R}^d$ with corresponding labels $y_1, ..., y_n \in \{1, -1\}$. Fix $\epsilon, \delta \in (0, 1)$. There exists absolute constants C_1, C_2 such that if*

1. *The function class can be written as $\mathcal{F} = \{f_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}\}$ where $\mathcal{W} \subset \mathbb{R}^p, diam\,(\mathcal{W}) \leq W$. Furthermore, for any $\mathbf{w}_1, \mathbf{w}_2 \in \mathcal{W}$,*

$$\|f_{\mathbf{w}_1} - f_{\mathbf{w}_2}\|_\infty \leq J\|\mathbf{w}_1 - \mathbf{w}_2\|$$

2. *The distribution μ of the x_i's can be written as a convex sum of k distributions μ_j that satisfy c-isoperimetry. Furthermore,*
$$2 \cdot 8^4 k \log(8k/\delta) \leq n\epsilon^2 \tag{10}$$
3. $\sigma^2 \triangleq \mathbb{E}_\mu \left[Var[y|x] \right] > 0$
4. *The dimension $d \geq C_1 \cdot \left(\frac{cL^2\sigma^2}{\epsilon^2} \right)$.*

Then with probability at least $1 - \delta$ with respect to sampling the data, for all $f \in \mathbb{F}$, if

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon$$

then we have

$$Lip(f) \geq \frac{\epsilon}{\sigma\sqrt{C_2 c}} \sqrt{\frac{nd}{p \log\left(1 + 32JW\epsilon^{-1} + \log(4/\delta)\right)}}$$

Proof. Let $\mathcal{W}_L \subset \mathcal{W}$ be defined as

$$\mathcal{W}_L \triangleq \{\mathbf{w} \in \mathcal{W} : \text{Lip}(f_{\mathbf{w}} \leq L\}$$

We define the covering $\frac{\epsilon}{8J}$-net $\mathcal{W}_{L,\epsilon}$ of $\mathcal{W}_L$. Then, applying Lemma 2.1 over a normalized space (by a factor of $2W$ to get a subset of the unit ball), we have

$$|\mathcal{W}_{L,\epsilon}| \leq \left(1 + 32JW\epsilon^{-1}\right)^p$$

By Theorem 4.2 over our ϵ-net (now a finite function class $\mathcal{F}_{L,\epsilon} \triangleq \{f_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}_{L,\epsilon}\}$),

$$\mathbb{P}\left(\exists f \in \mathcal{F}_{L,\epsilon} : \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon/2 \right) \leq (4k+1) \exp\left(-\frac{n\epsilon^2}{4 \cdot 8^3 k} \right) + \exp\left(p \log(1 + 32JW\epsilon^{-1}) - \frac{\epsilon^2 nd}{C_2 c L^2 \sigma^2} \right)$$

Observe that to go from functions in $f_{L,\epsilon} \in \mathcal{F}_{L,\epsilon}$ to functions in $f_L \in \mathcal{F}_L$, if

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - f_{L,\epsilon})^2 \leq \frac{\epsilon}{2} + \frac{1}{n} \sum_{i=1}^{n} (y_i - f_L)^2 \tag{11}$$

then we incur a cost $\|f_{L,\epsilon} - f_L\| \leq \frac{\epsilon}{8}$ because $f_{L,\epsilon}, f_L, y_i$ all satisfy $\|\cdot\|_\infty \leq 1$. Therefore, for some constant $C > 0$ and $L > 0$,

$$\mathbb{P}\left(\exists f \in \mathcal{F}_L : \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \right) \leq (4k+1) \exp\left(-\frac{n\epsilon^2}{2 \cdot 8^4 k} \right) + \exp\left(p \log(1 + 32JW\epsilon^{-1}) - \frac{\epsilon^2 nd}{CcL^2\sigma^2} \right)$$

Finally, by Equation 10,

$$(4k+1) \exp\left(-\frac{n\epsilon^2}{4 \cdot 8^3 k} \right) \leq \frac{(4k+1)\delta}{8k} \leq \frac{3\delta}{4}$$

and for some large enough $C_2 > 0$,

$$\exp\left(p\log(1+32JW\epsilon^{-1}) - \frac{\epsilon^2 nd}{C_2 cL^2\sigma^2}\right) \leq e^{-\log(4/\delta)} = \frac{\delta}{4}$$

allowing us to conclude that

$$L \leq \frac{\epsilon}{C_2\sigma\sqrt{c}}\sqrt{\frac{nd}{p\log(1+32JW\epsilon^{-1}) + \log(4/\delta)}}$$

implies that

$$\mathbb{P}\left(\exists f \in \mathcal{F}_L : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon\right) \leq \delta$$

giving us the desired result.

Remark. The application of the above lower bound to deep learning is quite simple. The isoperimetry condition used to control the class of Lipschitz functions is a property of the *data distribution*, so it is independent of the choice of model used. We can explicitly enforce a bound on the Lipschitz constant of a neural network by enforcing a bound on the weights, allowing us to apply this analysis to deep learning neural networks. Of course, it applies to a much wider class of models, but the observation that overparameterization is necessary to find a robust classifier for isoperimetric distributions is especially interesting as an explanation to why overparameterized regimes have found empirical success in deep learning.

□

5 Upper bounds on Lipschitz constant

We focus on results from [Geu+24], which consider networks of the form Equation 2 with ReLU activation functions. While the aforementioned lower bounds give us a necessary overparameterization condition for robustness, they do not bound the control itself, meaning overparameterized networks may also be non-robust. This section is a different flavor of analysis that looks into upper bounding the Lipschitz constant of deep neural networks independent of their data distribution or number of samples. These proofs are quite dense, so for brevity we only include full proofs when the intuition is useful.

Theorem 5.1 (Lipschitz upper bound for deep neural networks)**.** *For random deep neural networks of the form Equation 2 with ReLU activations such that for all $0 \leq \ell < L$,*

$$\left(W^{(\ell)}\right)_{i,j} \sim \mathcal{N}(0, \frac{2}{N}), \left(W^{(L)}\right)_{1,j} \sim \mathcal{N}(0,1)$$

while the biases are i.i.d. from some arbitrary bounded symmetric distribution about 0. Then, there exists constants $C, c_1 > 0$ such that if $N > d+2$ and $f : \mathbb{R}^d \to \mathbb{R}$, then

$$Lip(f) \leq C \cdot (3\sqrt{2})^L \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot \sqrt{d} \tag{12}$$

We begin with a recursive definition of the gradient of a ReLU network (which is not everywhere differentiable). For $0 \leq \ell < L$

$$\begin{aligned} D^{(\ell)}(x) &\triangleq \Delta\left(W^{(\ell)}x^{(\ell)} + b^{(\ell)}\right) \\ x^{(\ell+1)} &\triangleq D^{(\ell)}(x)\left(W^{(\ell)}x^{(\ell)} + b^{(\ell)}\right) = \text{ReLU}\left(W^{(\ell)}x^{(\ell)} + b^{(\ell)}\right) \end{aligned} \tag{13}$$

We make of the following two lemmas from [Geu+24], which we use without proof.

Lemma 5.1. *Assume $d+2<N$ and fix $W^{(0)},...,W^{(L-1)},b^{(0)},...,b^{(L-1)}$. Then for*

$$\mathcal{D} \triangleq \left\{\left(D^{(L-1)}(x),...,D^{(0)}(x)\right) : x \in \mathbb{R}^d\right\}$$

It follows that

$$|\mathcal{D}| \leq \left(\frac{eN}{d+1}\right)^{L(d+1)}$$

The proof by induction is found in [Geu+24], Lemma 5.7.

Lemma 5.2. *Assume $d+2<N$ and fix $W^{(0)},...,W^{(L-1)},b^{(0)},...,b^{(L-1)}$. Define*

$$\Lambda \triangleq \|W^{(L-1)}\|_2 \cdots \|W^{(0)}\|_2$$

For $x \in \mathbb{R}^d$ and $z \in \bar{B}_d(0,1)$, let

$$Y_{z,x} \triangleq D^{(L-1)}(x)W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}z \in \mathbb{R}^d$$

$$\mathcal{L} \triangleq \left\{Y_{z,x} : x \in \mathbb{R}^d, z \in \bar{B}_d(0,1)\right\} \subseteq \mathbb{R}^d$$

Then, for any $\epsilon \in (0,\Lambda)$, we can bound the covering number

$$\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \epsilon) \leq \left(\frac{3\Lambda}{\epsilon}\right)^d \cdot \left(\frac{eN}{d+1}\right)^{L(d+1)}$$

We prove the next lemma, which uses Dudley's inequality (Lemma 2.2) as the core tool for eventually deriving our Lipschitz bound.

Lemma 5.3. *Under the same assumptions as Lemma 5.2, there exist an absolute constant C such that given any $u \geq 0$, with probability $\geq 1-2\exp(-u^2)$ with respect to the choice of $W^{(L)}$,*

$$\sup_{x \in R^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}\|_2 \leq C \cdot \int_0^\Lambda \sqrt{\ln\left(\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \epsilon)\right)} d\epsilon$$

Proof. For some $y \in \mathcal{L}$,

$$\begin{aligned}
\|y\|_2 &= \|D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}z\|_2 \\
&\leq \|D^{(L-1)}(x)\|_2 \|W^{(L-1)}\|_2 \cdots \|D^{(0)}(x)\|_2 \cdot \|W^{(0)}\|_2 \cdot \|z\|_2 \\
&\leq \|W^{(L-1)}\|_2 \cdots \|W^{(0)}\|_2 = \Lambda
\end{aligned}$$

where by construction, each $\|D^{(\ell)}(x)\|_2 \leq 1$ and $\|z\|_2 \leq 1$, giving us the last inequality. Finally, we can replace our desired left-hand side as

$$\begin{aligned}
\sup_{x \in R^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}\|_2 &= \sup_{x \in R^d, z \in \bar{B}_d(0,1)} \left\langle \left(W^{(L)}\right)^T, Y_{z,x} \right\rangle \\
&= \sup_{Y \in L} \left\langle \left(W^{(L)}\right)^T, Y \right\rangle
\end{aligned}$$

Since $0 \in \mathcal{L}$ implies we can cover $\mathcal{L}$ with one ball when $\epsilon \geq \Lambda$ and $\mathcal{L} \subseteq \bar{B}_N(0,\Lambda) \implies \text{diam}(\mathcal{L}) \leq 2\Lambda$, we can apply Dudley's inequality (Lemma 2.2) to obtain the desired bound. □

Lemma 5.4. *Under the same assumptions as Lemma 5.2, there exist an absolute constant C such that given any $u \geq 0$, with probability $\geq 1 - 2\exp(-u^2)$ with respect to the choice of $W^{(L)}$,*

$$\sup_{x \in R^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}\|_2 \leq C \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot (\sqrt{d} + u)$$

Proof. From Lemma 5.2, we get

$$\begin{aligned}\sqrt{\ln(\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \epsilon))} &\leq \sqrt{d \ln\left(\frac{3\Lambda}{\epsilon}\right) + L(d+1)\ln\left(\frac{eN}{d+1}\right)} \\ &\leq \sqrt{d \ln\left(\frac{3\Lambda}{\epsilon}\right)} + \sqrt{L(d+1)\ln\left(\frac{eN}{d+1}\right)}\end{aligned}$$

This conveniently allows us to derive an upper bound for Dudley's integral

$$\begin{aligned}\int_0^\Lambda \sqrt{\ln(\mathcal{N}(\mathcal{L}, \|\cdot\|_2, \epsilon))} d\epsilon &\leq \sqrt{d} \int_0^\Lambda \sqrt{\ln\left(\frac{3\Lambda}{\epsilon}\right)} d\epsilon + \Lambda \cdot \sqrt{L(d+1)\ln\left(\frac{eN}{d+1}\right)} \\ &\leq \Lambda \cdot \left(C_1 \cdot \sqrt{d} + \sqrt{L(d+1)\ln\left(\frac{eN}{d+1}\right)}\right) \\ &\leq C_2 \cdot \Lambda \cdot \sqrt{Ld} \cdot \ln \sqrt{\left(\frac{eN}{d+1}\right)}\end{aligned}$$

for constants C_1, C_2 (we use the fact that $N > d+2$). Finally, we plug into the right-hand side of Lemma 5.3 to get that for a constant $C_3 > 0$ and any $u > 0$, with probability $\geq 1 - 2\exp(-u^2)$ with respect to the choice of $W^{(L)}$,

$$\sup_{x \in R^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}\|_2 \leq C_2 C_3 \cdot \Lambda \cdot \sqrt{L} \cdot \sqrt{\ln\left(\frac{eN}{d+1}\right)} \cdot (\sqrt{d} + u)$$

as desired. □

Proof sketch of Theorem 5.1: We first observe that the ReLU network f is a composition of Lipschitz continuous functions such that

$$\text{Lip}(f) \leq \sup_{x \in R^d} \|W^{(L)} \cdot D^{(L-1)}(x) \cdot W^{(L-1)} \cdots D^{(0)}(x) \cdot W^{(0)}\|_2$$

which is what the previous lemmas were upper bounding with high probability. The remainder of the proof is to bound each $W^{(\ell)}$ to massage the constants and obtain the desired bound, but we leave that for [Geu+24].

Remark. The tools above are rather specific to deep learning methods, but are still restrictive with respect to modern practical models that use mechanisms like attention [Vas+23]. Furthermore, a universal law for upper bounding Lipschitz constants with respect to model size has not been proven — hence characterizing robustness is still an unsolved problem.

References

[BLN20] Sébastien Bubeck, Yuanzhi Li, and Dheeraj Nagaraj. *A law of robustness for two-layers neural networks*. 2020. arXiv: 2009.14444 [cs.LG].

[BS22] Sébastien Bubeck and Mark Sellke. *A Universal Law of Robustness via Isoperimetry*. 2022. arXiv: 2105.12806 [cs.LG].

[Dem21] Amir Dembo. *Stochastic Processes*. 2021.

[Geu+24] Paul Geuchen et al. *Upper and lower bounds for the Lipschitz constant of random neural networks*. 2024. arXiv: 2311.01356 [stat.ML].

[Han24] Ramon van Handel. *Probability in high dimension. Lecture notes for ORF 550.* 2024.

[Vas+23] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[Ver18] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science.* 2018.