On the Sample Complexity of Two-Layer Networks: Lipschitz Vs. Element-Wise Lipschitz Activation

Amit Daniely AMIT.DANIELY@MAIL.HUJI.AC.IL
School of Computer Science and Engineering, The Hebrew University and Google Research Tel-Aviv

Elad Granot ELAD.GRANOT@MAIL.HUJI.AC.IL
School of Computer Science and Engineering, The Hebrew University

Editors: Claire Vernade and Daniel Hsu

Abstract

This study delves into the sample complexity of two-layer neural networks. For a given reference matrix $W^0 \in \mathbb{R}^{\mathcal{T}\times d}$ (typically representing initial training weights) and an $O(1)$-Lipschitz activation function $\sigma : \mathbb{R} \to \mathbb{R}$, we examine the class

$$\mathcal{H}^{\sigma}_{W^0,B,R,r} = \left\{\mathbf{x} \mapsto \langle \mathbf{v}, \sigma((W + W^0)\mathbf{x})\rangle : \|W\|_{\text{Frobenius}} \le R, \|\mathbf{v}\| \le r, \|\mathbf{x}\| \le B\right\}.$$

We demonstrate that the sample complexity of $\mathcal{H}^{\sigma}_{W^0,B,R,r}$ is bounded by

$$\tilde{O}\left(\frac{L^2B^2r^2\left(R^2 + \|W^0\|^2_{\text{Spectral}}\right)}{\epsilon^2}\right).$$

This bound is optimal, barring logarithmic factors, and depends logarithmically on the width $\mathcal{T}$. This finding improves on Vardi et al. (2022), who established a similar outcome for $W^0 = 0$. Our motivation stems from the real-world observation that trained weights often remain close to their initial counterparts, implying that $\|W\|_{\text{Frobenius}} \ll \|W + W^0\|_{\text{Frobenius}}$. To arrive at our conclusion, we employed and enhanced a recently new norm-based bounds method, the Approximate Description Length (ADL), as proposed by Daniely and Granot (2019).

Finally, our results underline the crucial role of the element-wise nature of σ for achieving a logarithmic width-dependent bound. We prove that there exists an $O(1)$-Lipschitz (non-element-wise) activation function $\Psi : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}^{\mathcal{T}}$ where the sample complexity of $\mathcal{H}^{\Psi}_{W^0,B,R,r}$ increases linearly with the width.

Keywords: Sample Complexity, Approximate Description Length, Lipschitz Activation Functions

1. Introduction

The remarkable capability of Neural Networks (NN) to generalize, even with more parameters than examples, remains a foundational enigma in contemporary NN practice (Zhang et al. (2021)). A recent line of works seek to address this phenomenon through bounds based on the norms of weight vectors, with notable contributions from Neyshabur et al. (2015); Bartlett et al. (2017); Golowich et al. (2018); Nagarajan and Kolter (2019); Daniely and Granot (2019); Vardi et al. (2022).

First bounds on generalization performance were based on Rademacher Complexity and Covering Numbers, often involving implicit or explicit weight regularization. A breakthrough came with the introduction of the Approximate Description Length (ADL), which proposed a bound

that is sub-linear with respect to the number of parameters Daniely and Granot (2019). This research posited a constraint on the deviation of weights from their initialization, suggesting that for constant-depth feed-forward neural networks with a wide set of activation functions, substituting the parameter count with input dimension multiplied by the deviation could yield a more concise asymptotic bound. However, this finding did not accommodate the commonly employed ReLU function, represented by $\max\{\cdot, 0\}$, thus leaving an unresolved gap.

Vardi et al. (2022) made significant strides by addressing this lacuna for two-layer networks. Their results, obtained via Rademacher Complexity, are tight up to logarithmic factors. Notably, their bound is based on the absolute norm of weights, as opposed to the deviation from their initialization.

The primary contribution of our study is to augment the findings of Vardi et al. (2022), keeping a similar bound however obtained from the distance from initialization. This challenge, cited as an open question by Vardi et al. (2022), originates from observations that weight deviations from initialization are often significantly smaller than those from the origin (as evidenced by Nagarajan and Kolter (2019); Bartlett et al. (2017); Daniely and Granot (2019)). Our analysis confirms the existence of such a bound for any element-wise $O(1)$-Lipschitz activation function.

To substantiate our conclusions, we harness the recent ADL tool introduced by Daniely and Granot (2019). Expanding on this approach, we introduce new methodologies, incorporating a chaining-based strategy tailored for the ReLU activation. We anticipate that these enhanced methods will be instrumental in future research, showcasing the potential power of the ADL framework and catalyzing novel insights.

In the subsequent section, we examine the limits of our assumptions, questioning the extensibility of these bounds to non-point-wise Lipschitz activation functions. Our concluding contribution illustrates the essential role of the element-wise property: we design a non-element-wise Lipschitz activation function and prove lower bounds on the generalization which scale **linearly** with width.

2. Preliminaries

2.1. Notations

We denote vectors using bold letters and matrices using upper letters. We shall add a hat sign $\left[\hat{\square}\right]$ or a tilde sign $\left[\tilde{\square}\right]$ above letters to mark them as random variables whose expectation equals the letters, e.g., $\mathbb{E}\left[\hat{\mathbf{x}}\right] = \mathbf{x}$.

We denote the Frobenius norm of a matrix W by $\|W\|_F^2 = \langle W, W\rangle = \sum_{ij} W_{ij}^2$, while the spectral norm is denoted by $\|W\| = \max_{\|\mathbf{x}\|=1} \|W\mathbf{x}\|$. We will define $\|\mathbf{v}\|_\infty$ as the L^∞ norm of the vector $\mathbf{v}$. We will use $\log$ with a base of 2 and $\ln$ with the natural base.

We denote by $\{0,1\}^k$ a sequence of k bits, and by $\{0,1\}^* = \bigcup_{k\in\mathbb{N}} \{0,1\}^k$ a sequence of bits of any length.

For any number $a \in \mathbb{R}$, we will denote by $\lfloor a \rfloor$ and $\lceil a \rceil$ the floor and ceiling of a, respectively. We will denote by $\lceil a \rceil_+ = \min\{n \in \mathbb{N} \cup \{0\} : a \le n\}$. Note that if $a < 0$ then $\lceil a \rceil_+ = 0$.

We will use the asymptotic notations O, Θ, and Ω to ignore constants and $\tilde{O}$ to ignore logarithmic terms. We will use $\left[\lesssim\right]$ in equations to denote an upper bound up to constant factors.

2.2. The Two-Layer Model

Let $\mathcal{X}_B \subseteq \mathbb{R}^d$ be a bounded set, s.t. $\forall \mathbf{x} \in \mathcal{X}_B, \|\mathbf{x}\| \leq B$. Let $\mathcal{T} \in \mathbb{N}$ be the width and $W^0 \in \mathbb{R}^{\mathcal{T} \times d}$ be some matrix. Let $\sigma : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}^{\mathcal{T}}$ be an L-Lipschitz activation function. For $W \in \mathbb{R}^{\mathcal{T} \times d}$ and $\mathbf{v} \in \mathbb{R}^{\mathcal{T}}$ define $h_{W,\mathbf{v}} : \mathcal{X}_B \to \mathbb{R}$ by $h_{W,\mathbf{v}}(\mathbf{x}) = \langle \mathbf{v}, \sigma(W\mathbf{x}) \rangle$ Finally, given $R, r > 0$, consider the following hypothesis class:

$$\mathcal{H}^{\sigma}_{\mathcal{T},L,B,R,r} = \left\{ h_{W,\mathbf{v}} : \left\|W - W^0\right\|_F \leq R, \|\mathbf{v}\| \leq r \right\} \tag{1}$$

which uses a total of $\mathcal{T}d$ parameters.

Note that while the above definitions do not explicitly mention a bias term in the linear operations, such cases are included in the model, e.g., by forcing the last element of $\mathbf{x}$ to be 1.

2.3. Approximate Description Length

Fix a domain $\mathcal{X}$. We say that a random function $\hat{f} : \mathcal{X} \to \mathbb{R}$ is an *ϵ-estimator* of $f : \mathcal{X} \to \mathbb{R}^{\mathcal{T}}$ if for every $x \in \mathcal{X}$, $\mathbb{E}\left[\hat{f}(x)\right] = f(x)$ and $\mathrm{Var}\left(\hat{f}(x)\right) \leq \epsilon^2$. Fix a hypothesis class $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$. We say that $\mathcal{F}$ is *ϵ-compressible using n bits* if there is a randomized mapping $f \in \mathcal{F} \mapsto \hat{f}$ such that for any $f \in \mathcal{F}$, $\hat{f}$ is an ϵ-estimator of f and there is a protocol that given f, Alice can randomly encode $\hat{f}$ using $\leq n$ bits in expectation. That is, Alice can send Bob a random string s (that depends on f and Alice's randomness) whose expected length is $\leq n$, and then Bob can generate a function $\hat{f} = f(s)$ such that $\hat{f}$ is an ϵ-estimator of f. In this case, we will say that $\hat{f}$ is an ϵ-compression of f that uses n bits. In some cases, we will allow the number of bits to depend on f or the parameters defining f. We note that Alice can send the empty string, whose length is 0.

Finally, we will say that $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$ has an *approximate description length (ADL)* of $n(m)$ if for any $A \subset \mathcal{X}$ of size m, $\mathcal{F}|_A$ is 1-compressible using $n(m)$ bits. In Daniely and Granot (2019), it is shown that the ADL bounds the sample complexity:

Theorem 1 *Fix a class $\mathcal{H}$ of functions from $\mathcal{X}$ to $\mathbb{R}$ with ADL $n(m)$ and a label space $\mathcal{Y}$. Fix L-Lipschitz and B-bounded loss function $\ell : \mathbb{R} \times \mathcal{Y} \to [0, \infty)$. Then, for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over a choice of a sample set $S \sim \mathcal{D}^m$,*

$$\sup_{h \in \mathcal{H}} \ell_{\mathcal{D}}(h) - \ell_S(h) \lesssim \frac{(L+B)\sqrt{n(m)}}{\sqrt{m}} \log(m) + B\sqrt{\frac{2\ln(2/\delta)}{m}}$$

where $\ell_{\mathcal{D}}(h) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \ell(h(x), y)$ and $\ell_S(h) = \frac{1}{m}\sum_{i=1}^m \ell(h(x_i), y_i)$

We will use the following results from Daniely and Granot (2019):

Lemma 2 *Suppose that $\hat{f}_1, \ldots, \hat{f}_k$ are i.i.d. ϵ-compressions of f that uses n bits each. Then $\frac{\sum_{i=1}^k \hat{f}_i}{k}$ is an $(\epsilon/\sqrt{k})$-compression of f that uses $O(kn)$ bits.*

Lemma 3 *Suppose that for any $1 \leq i \leq k$, $\hat{f}_i$ is an ϵ_i-compression of f_i that uses n_i bits. Assume furthermore that the $\hat{f}_i$'s are independent. Then $\sum_{i=1}^k \hat{f}_i$ is a $\sqrt{\sum_{i=1}^k \epsilon_i^2}$-compression of $\sum_{i=1}^k f_i$ that uses $O\left(\log(k) \cdot \sum_{i=1}^k n_i\right)$ bits.*

Lemma 4 *Suppose that $\mathcal{H}$ has ADL of $n(m)$ then $C \cdot \mathcal{H}$ has an ADL of $O((C^2+1)n(m))$*

Lemma 5 *Suppose the linear class*

$$\mathcal{H} = \left\{\lambda_{\boldsymbol{w}} : \boldsymbol{x} \mapsto \langle \boldsymbol{w}, \boldsymbol{x}\rangle : \left\|\boldsymbol{w} - \boldsymbol{w}^0\right\| \leq R\right\}$$

for some initialization $\boldsymbol{w}^0 \in \mathbb{R}^d$. Given $\epsilon > 1$, it is possible to ϵ-compress any $\lambda_{\boldsymbol{w}}$ defined over the set $A \subset \mathcal{X}_B$ using $O\left(\frac{Z^2 \log(dZ)}{\epsilon^2}\right)$ bits where $Z = O(B\|\boldsymbol{w} - \boldsymbol{w}^0\|)$.

We will also use the following lemmas:

Lemma 6 *Fix $A \subset \mathcal{X}_B$ of size m and $\mathcal{H}, \mathcal{H}' \subset \mathbb{R}^{\mathcal{X}_B}$ such that*

1. *For any $h \in \mathcal{H}$ there is $h' \in \mathcal{H}'$ with $\|h - h'\|_\infty \leq \delta \leq 1$*
2. *Assume that $\mathcal{H}'$ is 1-compressible using n bits.*

Then, $\mathcal{H}$ is 1-compressible using $O(n + \delta m \log(m))$ bits.

Proof Denote $A = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$. To compress $h \in \mathcal{H}$ we will choose $h' \in \mathcal{H}'$ with $\|h - h'\|_\infty \leq \delta$, and generate 1-compression $\hat{h}'$ of h' using n bits. Likewise, for any $i \in [m]$ independently choose i w.p. $|h(\mathbf{x}_i) - h'(\mathbf{x}_i)|$, and let 1_i be the indicator of the event that i was chosen. Define $\hat{h} : A \to [-1, 1]$ by $\hat{h}(\mathbf{x}_i) = \hat{h}'(\mathbf{x}_i) + \text{sign}(h(\mathbf{x}_i) - h'(\mathbf{x}_i))1_i$. Clearly, for any $i \in [m]$, $\mathbb{E}\hat{h}(\mathbf{x}_i) = h(\mathbf{x}_i)$. Furthermore

$$\text{Var}(\hat{h}(\mathbf{x}_i)) = \text{Var}(\hat{h}'(\mathbf{x}_i)) + \text{Var}(1_i) \leq 1 + \delta^2$$

Finally, $\hat{h}$ can be described using $O(n + \delta m \log(m))$ bits in expectation by concatenating the description of $\hat{h}'$ and a pair $(i, \text{sign}(h(\mathbf{x}_i) - h'(\mathbf{x}_i)))$ for any i such that $1_i = 1$. ■

Corollary 7 *(Single Parameter Compression) Let $\alpha \in \mathbb{R}$. For every $\epsilon \in (0, 1)$ there is an $\sqrt{\epsilon}$-compression for α that uses $O(\log(\lceil|\epsilon\alpha|\rceil))$ bits.*

Proof We will decompose $\alpha = \epsilon \left\lceil \frac{\alpha}{\epsilon} \right\rceil + \delta$ where $\delta = \alpha - \epsilon \left\lceil \frac{\alpha}{\epsilon} \right\rceil \in (0, 1)$. We need $O(\log(\lceil|\alpha|/\epsilon\rceil))$ to describe $\lceil\alpha/\epsilon\rceil$ (remember that ϵ is given and known), and from lemma 6, an additional $O(1)$ bits for describing δ. ■

2.4. Strong Shattering

For the lower bound, we will use the notion of **Strong Shattering**, as defined by Simon (1997):

Definition 8 *A class $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ **strongly-shatters** $x_1, \ldots, x_m \in \mathcal{X}$, if there exists $\boldsymbol{s} \in [0, 1]^m$ such that, for every $\boldsymbol{b} \in \{\pm 1\}^m$, there is $h \in \mathcal{H}$ such that for each $i \in [m]$*

$$\begin{aligned} h(x_i) &\geq s_i + 1, \quad \text{if } b_i = 1 \\ h(x_i) &\leq s_i - 1, \quad \text{if } b_i = 0. \end{aligned}$$

We further define $Sdim$ as:

$$Sdim(\mathcal{H}) = \max\left\{m : \exists x_1, \ldots, x_m \in \mathcal{X}, \text{ s.t. } \mathcal{H} \text{ strongly-shatters } x_1, \ldots, x_m\right\}.$$

Informally, the $Sdim$ for real-valued functions is like the VC-dimension for $\{0, 1\}$-valued functions. Previous results (as in Bartlett et al. (1994)) showed the lower bound of the sample complexity scales linearly with $Sdim$.

3. Results and Contributions

Our first result gives an upper norm-based generalization bound for any element-wise Lipschitz activation function using ADL.

Theorem 9 *Let $A \subset \mathcal{X}$ of size m, and assume σ is an element-wise L-Lipschitz activation function. Then $\mathcal{H}^{\sigma}_{\mathcal{T},L,B,R,r}|_A$ as defined in Eq. 1 has an ADL of*[1] *$\tilde{O}\left(L^2B^2r^2(R^2 + \|W^0\|^2)\right)$. As a result, $\mathcal{H}^{\sigma}_{\mathcal{T},L,B,R,r}$ has a sample complexity of $\tilde{O}\left(\frac{L^2B^2r^2(R^2+\|W^0\|^2)}{\epsilon^2}\right)$.*

Few remarks about the result: First, we note that the bound in Theorem 9 is tight, up to a logarithmic factor. Indeed, if σ was the identity function times L, then $\mathcal{H}^{L\cdot Id}_{\mathcal{T},L,B,R,r}$ would be the hypothesis class of bounded linear functions, which has a known sample complexity of $\tilde{\Theta}\left(\frac{L^2B^2R^2r^2}{\epsilon^2}\right)$ (Shalev-Shwartz and Ben-David (2014)).

Second, we note that this bound is similar to the upper bound of Vardi et al. (2022), which showed a bound of $O\left(\frac{L^2B^2R^2r^2\log^3(m)}{\epsilon^2}\right)$, up to logarithmic factors. The main improvement over their work is that our bound considers the distance of the weights from the initialization W^0, which is a more challenging task yet more relevant to the behavior of neural networks in practice.

Third, the proof for the above theorem is based on a new chaining-based argument that extends the ADL approach of Daniely and Granot (2019). As stated above, Daniely and Granot (2019) used this tool to prove a first tight bound up to logarithmic factors for many families of neural networks. We hope that the techniques in our proof will inspire future works to achieve bounds for deeper networks.

Last, we note that this bound has only logarithmic dependency in the width $\mathcal{T}$. This raises a natural question: can the element-wise property of σ be ignored and still yield the same bounds? Our second result shows that the answer is negative in general. Specifically, there is an $O(1)$-Lipschitz function $\hat{\sigma} : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}^{\mathcal{T}}$ for which the class of Eq. 1 can be strongly-shattered using $\Theta(\mathcal{T})$ samples for $\mathcal{T}$ that is up to exponential in d. This brings us to the second result of this paper:

Theorem 10 *For any dimension $d \geq 20$ and any width $d \leq \mathcal{T} \leq O(e^d)$, there is an L-Lipschitz activation function $\bar{\sigma}$ with $L = 32$, and a set of $\Theta(\mathcal{T})$ samples that strongly shatters the class $\mathcal{H}^{\bar{\sigma}}_{\mathcal{T},L,B=1,R=\sqrt{2d},r=1}$.*

Note that the parameters L, B, R, and r are independent of $\mathcal{T}$, and yet, by increasing only the width of the hidden layer, the sample-complexity increases similarly. Specifically, when $\mathcal{T}$ is exponential in d, the sample complexity of this two-layer network defined by Theorem 10 is $\Omega\left(e^d\right)$, whereas if only $\bar{\sigma}$ would have been element-wise, the same network would be linear in d (i.e., $O(d)$), according to Theorem 9.

1. The hidden poly-log factor is $O(\log^3(P))$, where P is the sum of all problem's parameters. See section 4 for more details.

4. Proof of Theorem 9

Let $h_{W,\mathbf{v}}$ as in equation 1, that is $h_{W,\mathbf{v}}(\mathbf{x}) = \langle \mathbf{v}, \sigma(W\mathbf{x}) \rangle$. By lemma 4, we can assume w.l.o.g. that σ is 1-Lipschitz and that $r = 1$. We can decompose $W = \begin{pmatrix} \mathbf{w}_1^T \\ \vdots \\ \mathbf{w}_{\mathcal{T}}^T \end{pmatrix}$, where each $\mathbf{w}_i \in \mathbb{R}^d$, and rewrite

$$h_{W,\mathbf{v}} = \sum_{i=1}^{\mathcal{T}} v_i \sigma(\mathbf{w}_i^T \mathbf{x}).$$

As σ is an element-wise function, one can create statistically independent estimators for each expression $v_i\sigma(\mathbf{w}_i^T\mathbf{x})$ in the sum. Moreover, if each estimator is ϵ_i-compressible using n_i bits, then using lemma 3 we get an $\sqrt{\sum_{i=1}^{\mathcal{T}} \epsilon_i^2}$-compression for $h_{W,\mathbf{v}}$ using $log\left(\mathcal{T}\right) \sum_{i=1}^{\mathcal{T}} n_i$ bits. The following proof shows how to construct such compressors with ϵ_i^2 that scales as $\frac{v_i^2}{r^2}$ and n_i that scales as $\frac{\left\|\mathbf{w}_i - \mathbf{w}_i^0\right\|^2}{R^2}$, hence ommitting the need for $\mathcal{T}$ up to logarithmic factor.

Fix a set $A \subset \mathcal{X}_B$ of size m and a vector $\mathbf{w}^0 \in \mathbb{R}^d$. For $\mathbf{w} \in \mathbb{R}^d$ and $v \in \mathbb{R}$ we define $h_{\mathbf{w},v} : A \to \mathbb{R}$ by $h_{\mathbf{w},v}(\mathbf{x}) = v(\sigma(\langle \mathbf{w}, \mathbf{x} \rangle) - \sigma(\langle \mathbf{w}^0, \mathbf{x} \rangle))$ and consider the class of single-neuron networks:

$$\mathcal{H} = \left\{ h_{\mathbf{w},v} : \mathbf{w} \in \mathbb{R}^d, v \in \mathbb{R} \right\}.$$

We will show how to $|v|$-compress any $h_{\mathbf{w},v}$ using $\tilde{O}\left(B^2 \left\|\mathbf{w} - \mathbf{w}^0\right\|^2\right)$ bits.

From claim 5 we can get an ϵ-compression for $\lambda_{\mathbf{w}}$. Define this compression as $\hat{\mathbf{w}}(\epsilon)$. We seem to be on a good track to compress $\sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$. However, this is misleading. Indeed, if we use $\hat{\mathbf{w}}(\epsilon)$ to create the random variable $\sigma(\langle \hat{\mathbf{w}}(\epsilon), \mathbf{x} \rangle)$, we will not get an ϵ-estimation, as $\mathbb{E}\left[\sigma(\langle \hat{\mathbf{w}}(\epsilon), \mathbf{x} \rangle)\right] \neq \sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$ for many choices of σ. We will, therefore, need a different approach.

Let us move to a non-efficient yet straight-forward approach: Recall that $A \subset \mathbb{R}^{\mathcal{X}_B}$ is fixed, with $|A| = m$. Hence, the function $\sigma(\langle \mathbf{w}, \mathbf{x} \rangle)$ can get up to m different results. Using corollary 7, we can ϵ-compress each such value using at most $Z = O(\log(BR/\epsilon))$ bits, and a total of mZ bits to ϵ-compress the entire function. However, this does not seem like an optimal compression, as the number of bits is linear in m, which we want to avoid. Yet, we will still use this approach in our construction: Let $k \in \mathbb{N}$ that will be defined later, and set $\epsilon_k = 2^{-k/2} B \left\|\mathbf{w} - \mathbf{w}^0\right\|$. Based on the above, we'll construct an ϵ_k-compression $\hat{g}$ of the function $g : \mathbf{x} \mapsto \sigma(\langle \mathbf{w}, \mathbf{x} \rangle) - \mathbb{E}\left[\sigma\left(\langle \hat{\mathbf{w}}(\epsilon_k), \mathbf{x} \rangle\right)\right]$ using mZ_k bits, were $Z_k = O(\log(2^k B \left\|\mathbf{w} - \mathbf{w}^0\right\|)) = O(k \log(B \left\|\mathbf{w} - \mathbf{w}^0\right\|))$.

With this compression at hand, we proceed with the following scheme:

- Let $\hat{v}$ a $|v|$-compression for v. From lemma 7 exists such a compression that uses $O(\log(r))$ bits.
- Given $k \in \mathbb{N}$, choose $i \in \{1, \ldots, k+1\}$ such that the probability to choose $1 \leq i \leq k$ is 2^{-i}, and the probability to choose $k+1$ is 2^{-k}.
- If $i = 1$, create the random variable $\hat{\mathbf{w}}\left(\epsilon_1\right)$ with $\epsilon_1 = 2^{-1/2} B \left\|\mathbf{w} - \mathbf{w}^0\right\|$. Set $\hat{f}$ as the function $\hat{f}(\mathbf{x}) = 2(\sigma(\langle \hat{\mathbf{w}}(\epsilon_1), \mathbf{x} \rangle) - \sigma(\langle \mathbf{w}^0, \mathbf{x} \rangle))$.

- If $2 \leq i \leq k$, create two independent random variables, $\hat{\mathbf{w}}(\epsilon_{i-1})$ and $\hat{\mathbf{w}}(\epsilon_i)$ where $\epsilon_i = 2^{-i/2}B\left\|\mathbf{w}-\mathbf{w}^0\right\|$. Set $\hat{f}$ as the function $\hat{f}(\mathbf{x}) = 2^i(\sigma(\langle\hat{\mathbf{w}}(\epsilon_i),\mathbf{x}\rangle) - \sigma(\langle\hat{\mathbf{w}}(\epsilon_{i-1}),\mathbf{x}\rangle))$.

- If $i = k+1$ then generate $\hat{\mathbf{w}}(\epsilon_k)$ and define $\hat{f}(\mathbf{x}) = 2^k\hat{g}(\mathbf{x})$.

- Output $\hat{h} = \hat{v}\hat{f}$.

The idea behind the structure above is to create a chain of events with increasing accuracy and cost (in number of bits) but with a decreasing probability of occurring. The following claim, together with lemma 2 shows that it is possible to $|v|$-compress $h_{\mathbf{w},v}$ using $\tilde{O}(B^2\|\mathbf{w}-\mathbf{w}^0\|^2)$ bits.

Claim 1 *For $k = \log_2(m)$ we have that $\hat{h}$ is a $O(|v|\,B\|\boldsymbol{w}-\boldsymbol{w}^0\|\sqrt{\log(m)})$-compression for $h_{\boldsymbol{w},v}$ that uses $O(\log(dZ)\log(m))$ bits, for Z as defined in lemma 5.*

Proof Fix $\mathbf{x} \in A$. We need to show that $\mathbb{E}\left[\hat{h}(\mathbf{x})\right] = h_{\mathbf{w},v}(\mathbf{x}) = v(\sigma(\langle\mathbf{w},\mathbf{x}\rangle) - v\sigma(\langle\mathbf{w}^0,\mathbf{x}\rangle)$, $\mathrm{Var}(\hat{f}(\mathbf{x})) \leq O(v^2B^2\left\|\mathbf{w}-\mathbf{w}^0\right\|\log(m))$, and that the number of bits that are used is $O(\log(dZ)\log(m))$. Indeed, since $\hat{v}$ is independent from the rest of the random variables and $\mathbb{E}[\hat{v}] = v$, we get:

$$\begin{aligned}
\frac{1}{v}\mathbb{E}\left[\hat{h}(\mathbf{x})\right] &= \frac{1}{v}\mathbb{E}[\hat{v}]\,\mathbb{E}\left[\hat{f}\right] \\
&= \mathbb{E}\left[\sigma(\langle\hat{\mathbf{w}}(\epsilon_1),\mathbf{x}\rangle) - \sigma(\langle\mathbf{w}^0,\mathbf{x}\rangle)\right] \\
&\quad + \sum_{i=2}^{k} 2^{-i}\mathbb{E}\left[2^i(\sigma(\langle\hat{\mathbf{w}}(\epsilon_i),\mathbf{x}\rangle) - \sigma(\langle\hat{\mathbf{w}}(\epsilon_{i-1}),\mathbf{x}\rangle))\right] \\
&\quad + 2^{-k}\mathbb{E}\left[2^k\hat{g}(\mathbf{x})\right] \\
&= \sigma(\langle\mathbf{w},\mathbf{x}\rangle) - \sigma(\langle\mathbf{w}^0,\mathbf{x}\rangle).
\end{aligned}$$

Likewise,

$$\begin{aligned}
\mathrm{Var}(\hat{h}(\mathbf{x})) &\overset{\mathrm{Var}(X)\leq\mathbb{E}X^2}{\leq} \mathbb{E}\left[\hat{v}^2(\hat{f}(\mathbf{x}))^2\right] \\
&\overset{\hat{v}\text{ independent of }\hat{f}}{=} \left(\mathrm{Var}(\hat{v}) + \mathbb{E}[\hat{v}]^2\right)\mathbb{E}\left[\left(\hat{f}(\mathbf{x})\right)^2\right] \\
&= 2v^2\mathbb{E}\left[\left(\hat{f}(\mathbf{x})\right)^2\right]
\end{aligned}$$

and as $\mathbb{E}\left[\hat{g}(\mathbf{x})\right] = g(\mathbf{x})$ and $\mathrm{Var}(\hat{g}(\mathbf{x})) \leq \epsilon_k$, we get:

$$\begin{aligned}
\mathbb{E}\left[\left(\hat{f}(\mathbf{x})\right)^2\right] &= 2^{-1}\mathbb{E}\left[4(\sigma(\langle\hat{\mathbf{w}}(\epsilon_1),\mathbf{x}\rangle) - \sigma(\langle\mathbf{w}^0,\mathbf{x}\rangle)^2\right] \\
&\quad + \sum_{i=2}^{k} 2^{-i}\mathbb{E}\left[2^{2i}\left(\sigma(\langle\hat{\mathbf{w}}(\epsilon_i),\mathbf{x}\rangle) - \sigma(\langle\hat{\mathbf{w}}(\epsilon_{i-1}),\mathbf{x}\rangle)\right)^2\right] \\
&\quad + 2^{-k}\left(\left(\mathbb{E}\left[2^k\hat{g}(\mathbf{x})\right]\right)^2 + \mathrm{Var}\left(2^k\hat{g}(\mathbf{x})\right)\right) \\
&\overset{\sigma \text{ is 1-Lipschitz}}{\leq} 2\mathbb{E}\left[(\langle\hat{\mathbf{w}}(\epsilon_1),\mathbf{x}\rangle - \langle\mathbf{w}^0,\mathbf{x}\rangle)^2\right] \\
&\quad + \sum_{i=2}^{k} 2^i\mathbb{E}\left[(\langle\hat{\mathbf{w}}(\epsilon_i),\mathbf{x}\rangle - \langle\hat{\mathbf{w}}(\epsilon_{i-1}),\mathbf{x}\rangle)^2\right] \\
&\quad + 2^k\left(\left(\mathbb{E}\left[|\langle\mathbf{w},\mathbf{x}\rangle - \langle\hat{\mathbf{w}}(\epsilon_k),\mathbf{x}\rangle|\right]\right)^2 + 2^{-k}B^2\left\|\mathbf{w}-\mathbf{w}^0\right\|^2\right) \\
&\overset{\text{Jensen's Inequality}}{\leq} 2\mathbb{E}\left[(\langle\hat{\mathbf{w}}(\epsilon_1),\mathbf{x}\rangle - \langle\mathbf{w},\mathbf{x}\rangle + \langle\mathbf{w},\mathbf{x}\rangle - \langle\mathbf{w}^0,\mathbf{x}\rangle)^2\right] \\
&\quad + \sum_{i=2}^{k} 2^i\mathbb{E}\left[(\langle\hat{\mathbf{w}}(\epsilon_i),\mathbf{x}\rangle - \langle\mathbf{w},\mathbf{x}\rangle + \langle\mathbf{w},\mathbf{x}\rangle - \langle\hat{\mathbf{w}}(\epsilon_{i-1}),\mathbf{x}\rangle)^2\right] \\
&\quad + 2^k\mathbb{E}\left[(\langle\mathbf{w},\mathbf{x}\rangle - \langle\hat{\mathbf{w}}(\epsilon_k),\mathbf{x}\rangle)^2\right] + B^2\left\|\mathbf{w}-\mathbf{w}^0\right\|^2 \\
&\overset{(*)}{\leq} 2\left\langle\mathbf{w}-\mathbf{w}^0,\mathbf{x}\right\rangle^2 + B^2\left\|\mathbf{w}-\mathbf{w}^0\right\|^2 \\
&\quad + 3\sum_{i=1}^{k} 2^i\mathrm{Var}\left(\langle\hat{\mathbf{w}}(\epsilon_i),\mathbf{x}\rangle\right) \\
&\overset{\epsilon_i\text{-compressors}}{\leq} 3B^2\left\|\mathbf{w}-\mathbf{w}^0\right\|^2 \\
&\quad + 3\sum_{i=2}^{k} 2^i 2^{-i}B^2\left\|\mathbf{w}-\mathbf{w}^0\right\|^2 \\
&\leq 6kB^2\left\|\mathbf{w}-\mathbf{w}^0\right\|^2.
\end{aligned}$$

Note that the step marked with $(*)$ follows from the independence between $\{\hat{\mathbf{w}}(\epsilon_i)\}$, as:

$$\begin{aligned}
\mathbb{E}\left[(\langle\hat{\mathbf{w}}(\epsilon_i)-\mathbf{w},\mathbf{x}\rangle - \langle\hat{\mathbf{w}}(\epsilon_{i-1})-\mathbf{w},\mathbf{x}\rangle)^2\right] &= \mathbb{E}\left[\langle\hat{\mathbf{w}}(\epsilon_i)-\mathbf{w},\mathbf{x}\rangle^2\right] + \mathbb{E}\left[\langle\hat{\mathbf{w}}(\epsilon_{i-1})-\mathbf{w},\mathbf{x}\rangle^2\right] \\
&\quad -2\mathbb{E}\left[\langle\hat{\mathbf{w}}(\epsilon_i)-\mathbf{w},\mathbf{x}\rangle\right]\mathbb{E}\left[\langle\hat{\mathbf{w}}(\epsilon_{i-1})-\mathbf{w},\mathbf{x}\rangle\right] \\
&= \mathrm{Var}\left(\langle\hat{\mathbf{w}}(\epsilon_i),\mathbf{x}\rangle\right) + \mathrm{Var}\left(\langle\hat{\mathbf{w}}(\epsilon_{i-1}),\mathbf{x}\rangle\right)
\end{aligned}$$

where the last equality follows since $\mathbb{E}\left[\langle\hat{\mathbf{w}}(\epsilon_i),\mathbf{x}\rangle\right] = \langle\mathbf{w},\mathbf{x}\rangle$.

Finally, from lemma 5, the expected number of bits that are required, up to a constant factor, is

$$\sum_{i=1}^{k} 2^{-}i\frac{Z^2\log(dZ)}{\epsilon_i^2} + 2^{-k}mZ_k = k\log(dZ) + 2^{-k}mZ_k.$$

When setting $k = \log_2(m)$ we get an $O(|v|\, B\|\mathbf{w} - \mathbf{w}^0\|\sqrt{\log(m)})$-compression for $h_{\mathbf{w},v}$ that uses $O(\log(dZ)\log(m))$ bits. ■

Corollary 11 *Using the above claim and lemma 2, we can construct a $|v|$-compression for $h_{\boldsymbol{w},v}$ that uses $O(B^2\left\|\boldsymbol{w} - \boldsymbol{w}^0\right\|^2 log(dZ)log^2(m))$ bits. Then, using lemma 3 we can compose a 1-compression for $h_{W,\boldsymbol{v}} - \boldsymbol{v}\sigma(W^0, \boldsymbol{x})$ that uses $O(B^2R^2r^2log(dZ)log^2(m))$ bits. Finally, we can use lemma 5 to create a 1-compression for $\boldsymbol{v}\sigma(W^0, \boldsymbol{x})$ with an addition of $O(B^2\left\|W^0\right\|^2 r^2log(dZ))$ bits.*

5. Proof of Theorem 10

Theorem 9 shows a generalization bound of the class $\mathcal{H}^{\sigma}_{\mathcal{T},L,B,R,r}$ as in Eq. 1 with a neglectible logarithmic dependency in the width, $\mathcal{T}$. The above is true, however, when σ is L-Lipschitz element-wise function. What if σ was not element-wise?

In this section, we'll proof Theorem 10 that shows that removing the element-wise property results in a bound that is **linearly** dependent on the width. We will show that there is a $\Theta(1)$-Lipschitz function, $\bar{\sigma}$, such the class $\mathcal{H}^{\bar{\sigma}}_{\mathcal{T},L,B,R,r}$ can be strongly shattered using $\Theta(\mathcal{T})$ samples, when B, R, r depends only at the input dimension, d. We then conclude that the sample complexity of Theorem 9 cannot be achievable in the non-element-wise case.

The proof is constructive and shows that by picking $m = \Theta(\mathcal{T})$ samples $\mathbf{x}^1, \ldots, \mathbf{x}^m \in \mathcal{X}_B$ and 2^m matrices $W^1, \ldots, W^{2^m} \in \mathbb{R}^{\mathcal{T}\times d}$ at random, the the set of points $P := \left\{W^k\mathbf{x}^i : i \in [m], k \in [2^m]\right\}$ are far enough from each other with a positive probability (The details are presented in Lemma 13).

Hence, we can construct an activation function, described in Lemma 14, that can move every desire point in P to a vector of our choice, while maintaining the Lipschitzness property.

Finally, we conclude that there is a set of samples $\mathbf{x}^1, \ldots, \mathbf{x}^m \in \mathcal{X}_B$ that are able to strongly-shatter $\mathcal{H}^{\bar{\sigma}}_{\mathcal{T},L,\Theta(1),\Theta(d),\Theta(1)}$. Note that the width $\mathcal{T}$ can be exponentially big with respect to d, and the number of shattered samples, m, grows linearly with it.

Denote by $\mathrm{Vol}_k(A)$ the k-dimensional volume of a set $A \subset \mathbb{R}^d$ normalized such that the volume $\mathrm{Vol}_{\mathcal{T}-1}\left(\mathbb{S}^{\mathcal{T}-1}\right) = 1$. Denote also $B^d(\mathbf{x}, R) = \{\mathbf{x}' \in \mathbb{R}^d : \|\mathbf{x}-\mathbf{x}'\| \le R\}$. We will use the following fact

Lemma 12 *For any $\boldsymbol{x} \in \mathbb{S}^{\mathcal{T}-1}$ and sufficiently large $\mathcal{T}$ we have*

$$\mathrm{Vol}_{\mathcal{T}-1}\left(\mathbb{S}^{\mathcal{T}-1} \cap B^{\mathcal{T}}(\boldsymbol{x}, 1/2)\right) \le e^{-\frac{\mathcal{T}}{3}} < \mathcal{T}^{-2}2^{-\mathcal{T}/4}$$

Proof Denote $\epsilon = \frac{1}{2}$. We have

$$\|\mathbf{x} - \mathbf{w}\|^2 < \epsilon^2 \Leftrightarrow 2 - 2\langle\mathbf{w}, \mathbf{x}\rangle < \epsilon^2 \Leftrightarrow \langle\mathbf{w}, \mathbf{x}\rangle > 1 - \epsilon^2/2$$

Hence,

$$\mathbb{S}^{\mathcal{T}-1} \cap B^{\mathcal{T}}(\mathbf{x}, \epsilon) = \{\mathbf{w} \in \mathbb{S}^{\mathcal{T}-1} : \langle\mathbf{w}, \mathbf{x}\rangle > 1 - \epsilon^2/2\}$$

Let $\mathbf{w} \in \mathbb{S}^{\mathcal{T}-1}$ be a uniform vector. For any $a > 0$ we have $\Pr(\langle\mathbf{w}, \mathbf{x}\rangle > a) \le 2e^{-\frac{\mathcal{T}a^2}{2}}$ (e.g. chapter 14 in Matousek (2013)). Hence,

$$\begin{aligned}\mathrm{Vol}_{\mathcal{T}-1}\left(\mathbb{S}^{\mathcal{T}-1} \cap B^{\mathcal{T}}(\mathbf{x}, \epsilon)\right) &= \Pr(\langle\mathbf{w}, \mathbf{x}\rangle > 1 - \epsilon^2/2) \\ &\le 2e^{-\frac{\mathcal{T}(1-\epsilon^2/2)^2}{2}} = 2e^{-\frac{\mathcal{T}49}{128}} \le (e/2)^{-\mathcal{T}/4}2^{-\mathcal{T}/4}\end{aligned}$$

this concludes the proof as $(e/2)^{-\mathcal{T}/4} \leq \mathcal{T}^{-2}$ for sufficiently large $\mathcal{T}$. ■

Lemma 13 *For $e^{d/3} \geq \mathcal{T} \geq d \geq 20$, there exists a set of vectors $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^m \in \mathbb{S}^{d-1}$ and a set of matrices $A^1, \ldots A^{2^m} \in \mathbb{R}^{\mathcal{T} \times d}$ that have the following properties:*

1. *$m = \lfloor \mathcal{T}/4 \rfloor$*
2. *A^s in an isometry (and hence $\|A^s\|_F^2 = d$), for each $s \in [2^m]$*
3. *$\left\|A^s \boldsymbol{x}^i - A^t \boldsymbol{x}^j\right\| \geq \frac{1}{2}$, for each $i, j \in [m]$ and $s, t \in [2^m]$ such that $(s, i) \neq (t, j)$*

Proof Choose m vectors $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{S}^{d-1}$ such that $\|\mathbf{x}_i - \mathbf{x}_j\| \geq \frac{1}{2}$ if $i \neq j$. By lemma 12 this is possible as long as $\mathcal{T}/4 \leq e^{d/3}$. Let $A^1, \ldots, A^k$ be the maximal set of matrices that satisfy items 2. and 3. We need to show that $k \geq 2^m$.

Let $A \in \mathbb{R}^{\mathcal{T} \times d}$ be a random matrix chosen uniformly from the set of matrices with unit norm columns that are orthogonal to one another. We have A is an isometry with $\|A\|_F = \sqrt{d}$. Furthermore, adding A to $A^1, \ldots, A^k$ will violate item 2. or 3. only if $\left\|A\mathbf{x}^i - A^t \mathbf{x}^j\right\| < \frac{1}{2}$ for some $i, j \in [m]$ and $t \in [k]$. Since $A\mathbf{x}^i$ is a uniform vector in $\mathbb{S}^{\mathcal{T}-1}$, the probability of violation is bounded by $km^2\mathcal{T}^{-2}2^{-\mathcal{T}/4} \leq k2^{-\mathcal{T}/4}$. On the other hand, by the maximality of k, this probability is 1. This implies that $k \geq 2^{\mathcal{T}/4} \geq 2^m$. ■

Lemma 14 *Let $x_1, \ldots, x_m$ be a finite set of different points in some metric space $(\mathcal{X}, d)$, such that for each $i \neq j \in [m]$, $d(x_i, x_j) \geq \alpha$. Let further be $p_1, \ldots, p_m \in \mathbb{R}$ any set of points. Then there exists an L-Lipschitz function, $f : \mathcal{X} \to \mathbb{R}$, where*

$$L = \frac{2}{\alpha} \min_{C \in \mathbb{R}} \max_{i \in [m]} (|p_i - C|).$$

such that for each $i \in [m]$, $f(x_i) = p_i$.

Proof Choose C such that $L = \frac{2}{\alpha} \max_{i \in [m]} (|p_i - C|)$ and define

$$f(x) = \max_{i \in [m]} \{p_i - Ld(x, x_i)\}$$

f is L-Lipschitz as a maximum of L-Lipschitz functions. Fix x_j. It is enough to show that $f(x_j) = p_j$. First, $f(x_j) \geq p_j - Ld(x_j, x_j) = p_j$. Thus, it remain to show that $f(x_j) \leq p_j$. Fix some $i \in [m] \setminus \{j\}$ it is enough to show that $p_i - Ld(x_j, x_i) \leq p_j$. Indeed,

$$\begin{aligned}
p_i - Ld(x_j, x_i) &\overset{d(x_j, x_i) \geq \alpha}{\leq} && p_i - L\alpha \\
&\overset{\text{definition of } L}{=} && p_i - 2 \max_{i \in [m]} (|p_i - C|) \\
&\leq && p_i - (|p_i - C| + |p_j - C|) \\
&= && C + (p_i - C) - (|p_i - C| + |p_j - C|) \\
&\leq && C - |p_j - C| \\
&\leq && C - (C - p_j) = p_j
\end{aligned}$$

■

We are now ready to prove the main theorem.

Proof (of Theorem 10) Based on the previous lemmas, we'll strongly shatter a set of $m = \frac{\mathcal{T}}{10}$ samples.

Order the elements of the set $2^{[m]}$ as $S_1, \dots S_{2^m}$ in some arbitrary order, and define the function $f : [m] \times [2^m] \to \{\pm 1\}$ as:

$$f(k,i) = \begin{cases} 1, & i \in S_k \\ -1, & i \notin S_k \end{cases} \qquad \forall i \in [m], k \in [2^m].$$

Let $\mathbf{x}^1, \dots, \mathbf{x}^m \in \mathbb{S}^{d-1}$ and $A^1, \dots A^{2^m} \in \mathbb{R}^{\mathcal{T} \times d}$ be the sets defined in lemma 13, and note from the lemma that the set $Q = \left\{A^s \mathbf{x}^i : i \in [m], s \in [2^m]\right\}$ contains $m2^m$ different elements such that for each pair $A^s\mathbf{x}^i \neq A^t\mathbf{x}^j$ we have

$$\left\|A^s\mathbf{x}^i - A^t\mathbf{x}^j\right\| \geq \frac{1}{2}.$$

We can now apply Lemma 14 with the Euclidean metric space, to get a 4-Lipschitz function, $\hat{f} : \mathbb{R}^d \to \mathbb{R}$, such that for all $i \in [m], k \in [2^m]$,

$$\hat{f}(A^k\mathbf{x}^i) = f(k,i).$$

The activation function will therefore be $\bar{\sigma}(\mathbf{v}) = \hat{f}(\mathbf{v})\mathbf{e}_1$ (or alternatively, we can distribute $\hat{f}$ evenly over the all the $\mathcal{T}$ hidden neurons).

Finally, as each $\|A^s\|_F^2 = d$, we can create the hypothesis class (using the definition of Eq. 1):

$$\mathcal{H}^{\bar{\sigma}}_{\mathcal{T},4,1,\sqrt{d},1} \supset \left\{\mathbf{x} \mapsto \mathbf{e}_1^T \bar{\sigma}(Ax) : A \in \mathbb{R}^{\mathcal{T} \times d}, \|A\|_F \leq d\right\}$$

and note that it defines a neural network that can 1-shatter the m points, and $m = \Theta(\mathcal{T})$. ■

6. Discussion and Open Questions

This work aims to understand the sample complexity of depth-two neural networks and the effect of element-wise activation functions (i.e., functions that work on each neuron independently) on the sample complexity of neural networks. Using the ADL approach, we have shown that this property is sufficient and necessary for two-layer networks to achieve logarithmic width-dependency bounds. By necessary, we mean that the set of general non-pointwise Lipschitz contains activations under which the sample complexity is larger than any element-wise Lipschitz activation functions. One can view a non-element-wise Lipschitz function as a set of neurons that can share knowledge. Our work shows that this ability amplifies the sample complexity of the network.

We note that the upper bound presented this work is tight w.r.t. to all parameters (i.e., $\mathcal{T}$, L, B, d, R, r, $\left\|W^0\right\|$ and ϵ). To the best of our knowledge, such a tightness is not implied by previous results. The optimally of the dependence on L, B, R, r and ϵ is true already for non relative bound, as discussed in Vardi et al. (2022). As for the spectral norm of W^0, note that even if $R = 0$, $\mathcal{H}^{\sigma}_{\mathcal{T},L,B,0,r}$ contains linear classifiers of norm r over examples of norm $B\|W_0\|$, which yields a

sample complexity at least $(LBr\|W_0\|)^2$. Finally, the tightness of $\mathcal{T}$ is shown by the upper and lower bounds of this work.

Additional to the above, in this work we have developed a new technique that extends ADL and creates a chain of events with increasing accuracy but with a decreasing probability of occurring. This provides better control over both competing values: the variance and the number of bits. We hope that this idea will spark following works.

We are still left with two open questions, one for sufficiency and one for necessity. Regarding sufficiency, a natural question is whether the results in the paper can be extended to deeper networks. Daniely and Granot (2019) gave a hint for this question, showing a sample complexity for deep neural networks that require only the sum of the widths (which is sublinear in the number of parameters). Yet, their result does not hold for any element-wise Lipschitz activation function. We believe achieving similar bounds for any element-wise Lipschitz activation function is possible.

As for the necessity, we note that our lower bound is not valid for *any* non-element-wise Lipschitz activation function. Indeed, if we take some permutation of an element-wise activation function, we do not expect to get width-dependent bounds, although we lost the element-wise property. Instead, we want to ask whether there exists a (non-element-wise) Lipschitz activation function that guarantees a linear lower bound in the number of parameters, hence matching the upper bound obtained via the "parameters counting" approach. In our result, the lower bound is still sublinear in the number of parameters.

References

Peter L Bartlett, Philip M Long, and Robert C Williamson. Fat-shattering and the learnability of real-valued functions. In *Proceedings of the seventh annual conference on Computational learning theory*, pages 299–310, 1994.

Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *Advances in neural information processing systems*, 30, 2017.

Amit Daniely and Elad Granot. Generalization bounds for neural networks via approximate description length. *Advances in Neural Information Processing Systems*, 32, 2019.

Noah Golowich, Alexander Rakhlin, and Ohad Shamir. Size-independent sample complexity of neural networks. In *Conference On Learning Theory*, pages 297–299. PMLR, 2018.

Jiri Matousek. *Lectures on discrete geometry*, volume 212. Springer Science & Business Media, 2013.

Vaishnavh Nagarajan and J Zico Kolter. Generalization in deep networks: The role of distance from initialization. *arXiv preprint arXiv:1901.01672*, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-Based Capacity Control in Neural Networks. *Proceedings of The 28th Conference on Learning Theory*, 40:1376–1401, 2015. ISSN 15337928. URL http://jmlr.csail.mit.edu/proceedings/papers/v40/Neyshabur15.html.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *SIAM Journal on Computing*, 26(3):751–763, jul 1997. ISSN 00975397. doi: 10.1137/S0097539793259185.

Gal Vardi, Ohad Shamir, and Nati Srebro. The sample complexity of one-hidden-layer neural networks. *Advances in Neural Information Processing Systems*, 35:9139–9150, 2022.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.