Deep Residual Networks
Preserve Expected Length

Kenny Peng
Senior Thesis, Advised by Boris Hanin

April 25, 2022

Abstract

We prove upper bounds on the expected length and volume distortion of deep residual neural networks, extending the work of [HJR21] for deep fully-connected neural networks. Our results confirm experimental findings suggesting that length distortion decreases slightly with the number of residual modules.

1 Introduction

Empirical findings suggest that deep neural networks are more powerful than shallow ones. Recent theoretical work seeks to understand why. One approach studies the complexity of neural networks as they vary in depth. For example, it has been shown that deeper neural networks can express more complex functions (e.g., [BS14, CSS15, MPCB14]), an intuitive explanation for why deeper networks are preferable. At the same time, several theoretical findings have demonstrated that deeper networks do not necessarily express more complex functions on average (e.g., [HR19, HJR21, PKL19]), suggesting that the situation is more nuanced.

But even as we gain a more precise understanding of the complexity of neural networks, existing analysis focuses primarily on fully-connected neural networks. However, as deep neural networks used in practice often have more complex architectures, it is important to build a theoretical understanding of these other settings.

In this thesis, we study the complexity of residual neural networks (ResNets), a type of architecture widely used in practice. The measure of complexity we focus on is *length distortion*: how much does a neural network typically distort the length of an input curve? (See e.g., [PT21, RPK$^+$17, HJR21].) Recent work demonstrates that, at initialization, length distortion in fully-connected networks does not increase with depth, and actually slightly decreases [HJR21].

We extend the results of [HJR21] for length distortion to residual networks. We provide theoretical bounds that closely match experimental findings in [HJR21] suggesting that length distortion in residual networks also decreases slightly with depth. We also obtain bounds on the distortion of higher-dimensional manifolds.

We begin, in this section, by stating bounds on expected length distortion both in the case of fully-connected neural networks (Theorem 1) and residual neural networks (Theorem 2). We prove these theorems in Section 2. In Section 3, we state and prove generalizations of these results for higher-dimensional manifolds.

1.1 Definitions and main results

We now define both the fully-connected neural network and the residual neural network. We then state the existing result of [HJR21] for the expected length distortion of fully-connected networks, and then state our new corresponding result for residual networks.

Fix $L \geq 1, n_0, \cdots, n_L \geq 1, \sigma : \mathbb{R} \to \mathbb{R}$. Then a depth L *fully-connected neural network* $\mathcal{N}$ with input dimension n_0, output dimension n_L, hidden layer widths $n_1, \cdots, n_{L-1}$, and activation function σ is any function of the form

$$x \in \mathbb{R}^{n_0} \mapsto \mathcal{N}(x) \in \mathbb{R}^{n_L},$$

where

$$\mathcal{N}^{(\ell+1)}(x) := \begin{cases} W^{(1)}x + b^{(1)} & \ell = 0 \\ W^{(\ell+1)}\sigma(\mathcal{N}^{(\ell)}(x)) + b^{(\ell+1)} & l \geq 1 \end{cases}$$

and $\mathcal{N} = \mathcal{N}^{(L)}$. For our purposes, we will always take σ to be the ReLU activation function $\mathrm{ReLU}(x) = \max(0, x)$.

The weights and biases are given by *standard He initialization* if $W_{ij}^{(\ell)}, b_j^{(\ell)}$ are independent Gaussian variables satisfying

$$W_{ij}^{(\ell)} \sim G(0, 2/n_{\ell-1}), \quad b_j^{(\ell)} \sim G(0, C_b) \tag{1}$$

for any fixed constant $C_b > 0$ and where $G(\mu, \sigma^2)$ denotes a Gaussian with mean 0 and variance σ^2.

For a curve M, we let $\mathcal{N}(M)$ denote the image of M under $\mathcal{N}$. Let $\mathrm{len}(\cdot)$ denote the length of a curve. We may now state a simplified version of the upper bound given in [HJR21] for standard neural networks.

Theorem 1 (Length distortion in deep fully-connected networks)**.** *Consider a fully-connected network $\mathcal{N}$ of depth L, input dimension n_0, and output dimension n_L, with weights given by He normal initialization. Consider a curve M of unit length. Then*

$$\mathbb{E}[\mathrm{len}(\mathcal{N}(M))^2] \leq \frac{n_L}{n_0}. \tag{2}$$

This gives upper bounds on both $\mathbb{E}[\mathrm{len}(\mathcal{N}(M))]$ and $\mathrm{Var}[\mathrm{len}(\mathcal{N}(M))]$ since

$$\mathbb{E}[X] \leq \left(\mathbb{E}[X^2]\right)^{1/2} \qquad \mathrm{Var}[X] \leq \mathbb{E}[X^2] \tag{3}$$

for any random variable X.

Define a *residual neural network* $\mathcal{N}_L^{res}$ with residual modules $\mathcal{N}_1, \cdots, \mathcal{N}_L$ and scales $\eta_1, \cdots, \eta_L$ by the recursion

$$\mathcal{N}_0^{res}(x) = x, \quad \mathcal{N}_\ell^{res}(x) = \mathcal{N}_{\ell-1}^{res}(x) + \eta_\ell \mathcal{N}_\ell(\mathcal{N}_{\ell-1}^{res}(x)), \quad \ell = 1, \cdots, L. \tag{4}$$

For our purposes, $\mathcal{N}_1, \cdots, \mathcal{N}_L$, and $\mathcal{N}_L^{res}$ each have a shared input and output dimension n.

In this thesis, we show the corresponding version of Theorem 1 for residual networks.

Theorem 2 (Length distortion in deep residual networks)**.** *Consider a residual ReLU network $\mathcal{N}_L^{res}$ with residual modules $\mathcal{N}_1, \cdots, \mathcal{N}_L$ (each with weights given by He normal initialization) and scales $\eta_1, \cdots, \eta_L$. Consider a curve M of unit length. Then*

$$\mathbb{E}[\mathrm{len}(\mathcal{N}(M))^2] \leq \prod_{\ell=1}^{L} 1 + \eta_\ell^2. \tag{5}$$

2 Length distortion

In this section, we prove Theorem 1 and Theorem 2. We begin by introducing a general setup used in [HJR21] for studying length distortion. The setup applies for both fully-connected and residual networks (and indeed other network architectures), and relies on understanding the network's Jacobian.

Consider a path M of unit length, and suppose we have a smooth unit speed parametrization $M = \gamma([0,1])$ with

$$\gamma : \mathbb{R} \to \mathbb{R}^n, \quad \gamma(t) = (\gamma_1(t), \cdots, \gamma_n(t)). \tag{6}$$

Then the mapping

$$\Gamma := \mathcal{N} \circ \gamma, \quad \Gamma : \mathbb{R} \to \mathbb{R}^n \tag{7}$$

gives a parametrization of the curve $\mathcal{N}(M)$, and we have

$$\operatorname{len}(\mathcal{N}(M)) = \int_0^1 \|\Gamma'(t)\| \, dt. \tag{8}$$

Now let J_x denote the Jacobian of the map $x \mapsto \mathcal{N}(M)$. We recall the following lemma from [HJR21].

Lemma 3 (Lemma C.1 of [HJR21])**.** *We have*

$$\mathbb{E}[\operatorname{len}(\mathcal{N}(M))^2] \leq \int_0^1 \mathbb{E}[\|J_{\gamma(t)}\gamma'(t)\|^2] \, dt. \tag{9}$$

We are therefore interested in understanding the distribution of

$$\|J_x u\| \tag{10}$$

for $x \in \mathbb{R}^n$ and a unit vector $u \in \mathbb{R}^n$. Of course, this distribution depends on the network architecture. In the next section, we present [HJR21]'s proof of Theorem 1, which utilizes the explicit distribution of (10) for fully-connected neural networks. In Section 2.2, we extend this analysis to prove Theorem 2.

2.1 Length distortion in deep fully-connected networks

In this section, we use the analysis in [HJR21] to prove Theorem 1. [HJR21] go further, providing bounds on all moments of $\operatorname{len}(\mathcal{N}(M))$, but we focus here on the second moment.

Consider a fully-connected network $\mathcal{N}$ of depth L, input dimension n_0, and output dimension n_L, with weights given by standard He initialization. Continuing with the introduced notation, let J_x be the Jacobian of the map $x \to \mathcal{N}(x)$. The following result gives us the exact distribution of $\|J_x u\|$.

Proposition 4. *For any $x \in \mathbb{R}^{n_0}$ and any unit vector $u \in \mathbb{R}^{n_0}$, $\|J_x u\|^2$ is equal in distribution to a product of independent scaled chi-squared random variables:*

$$\|J_x u\|^2 \overset{d}{=} \frac{n_L}{n_0} \left(\prod_{\ell=1}^{L-1} \frac{2}{n_\ell} \chi_{\mathrm{Bin}(n_\ell, 1/2)} \right) \cdot \frac{1}{n_L} \chi^2_{n_L}, \tag{11}$$

where $\mathrm{Bin}(n_\ell, 1/2)$ *are independent binomial distributions determining the number of degrees of freedom.*

Theorem 1 follows shortly: We have

$$\mathbb{E}[\mathrm{len}(\mathcal{N}(M))^2] \leq \int_0^1 \mathbb{E}[\|J_{\gamma(t)}\gamma'(t)\|^2]\, dt = \mathbb{E}[\|J_x u\|^2], \tag{12}$$

where we applied Lemma 3 and then observed from Proposition 4 that $\mathbb{E}[\|J_x u\|^2]$ is the same for all choices of x and unit vector u. Then, recalling that $\mathbb{E}[\chi_k^2] = k$, Theorem 1 follows immediately, as

$$\mathbb{E}[\|J_x u\|^2] = \frac{n_L}{n_0}\left(\prod_{\ell=1}^{L-1} \frac{2}{n_\ell}\mathbb{E}[\chi_{\mathrm{Bin}(n_\ell,1/2)}]\right) \cdot \frac{1}{n_L}\mathbb{E}[\chi_{n_L}^2] = \frac{n_L}{n_0}. \tag{13}$$

We spend the remainder of the section deriving Proposition 4. The observations in this derivation will also be useful for our analysis of residual networks.

For fixed x, J_x is equal in distribution to a product of independent random matrices

$$J_x \overset{d}{=} AW^{(L)}D^{(L-1)}W^{(L-1)}\cdots D^{(1)}W^{(1)}, \tag{14}$$

where A is a diagonal matrix with independent diagonal entries that are ± 1 with equal probability, $D^{(\ell)}$ are diagonal matrices with diagonal entries that are independent Bernoulli(1/2) random variables, and each $W^{(\ell)}$ is determined by standard He initialization

$$W_{ij}^{(\ell)} \sim G(0, 2/n_{\ell-1}). \tag{15}$$

Then we have

$$\|J_x u\| \overset{d}{=} \|W^{(L)}D^{(L-1)}W^{(L-1)}\cdots D^{(1)}W^{(1)}u\| \tag{16}$$

$$\overset{d}{=} \|W^{(L)}D^{(L-1)}W^{(L-1)}\cdots D^{(2)}W^{(2)}u^{(1)}\|\|D^{(1)}W^{(1)}u\|, \tag{17}$$

where $u^{(1)} = \frac{D^{(1)}W^{(1)}u}{\|D^{(1)}W^{(1)}u\|}$. It is a standard fact that for a matrix W with i.i.d. Gaussian entries, Wu is independent of u for any unit vector u and equal in distribution to Wv for any unit vector v. Successively applying this fact, we find

$$\|J_x u\| \overset{d}{=} \|W^{(L)}u\|\|D^{(L-1)}W^{(L-1)}u\|\cdots\|D^{(1)}W^{(1)}u\|. \tag{18}$$

The result follows by observing

$$\|D^{(\ell)}W^{(\ell)}u\| \overset{d}{=} \frac{2}{n_{\ell-1}}\chi^2_{\mathrm{Bin}(n_\ell,1/2)} \tag{19}$$

$$\|W^{(L)}u\| \overset{d}{=} \frac{2}{n_{L-1}}\chi^2_{\mathrm{Bin}(n_L,1/2)}. \tag{20}$$

2.2 Length distortion in deep residual networks

We now prove Theorem 2, again using the bound in Lemma 3.

Recall that a residual network $\mathcal{N}_L^{res}$ with residual modules $\mathcal{N}_1, \cdots, \mathcal{N}_L$ and scales $\eta_1, \cdots, \eta_L$ is recursively defined by

$$\mathcal{N}_0^{res}(x) = x, \quad \mathcal{N}_\ell^{res}(x) = \mathcal{N}_{\ell-1}^{res}(x) + \eta_\ell \mathcal{N}_\ell(\mathcal{N}_{\ell-1}^{res}(x)), \quad \ell = 1, \cdots, L. \tag{21}$$

For our purposes, $\mathcal{N}_1, \cdots, \mathcal{N}_L$, and $\mathcal{N}_L^{res}$ each have a shared input and output dimension n. Also suppose that $\mathcal{N}_1, \cdots, \mathcal{N}_L$ each have weights given by standard He initialization.

Once again, consider a path M of unit length parametrized by $M = \gamma([0,1])$ with

$$\gamma : \mathbb{R} \to \mathbb{R}^n, \quad \gamma(t) = (\gamma_1(t), \cdots, \gamma_n(t)) \tag{22}$$

so that the mapping

$$\Gamma := \mathcal{N}_L^{res} \circ \gamma, \quad \Gamma : \mathbb{R} \to \mathbb{R}^n \tag{23}$$

gives a parametrization of the curve $\mathcal{N}_L^{res}(M)$. Let $J_{L,x}^{res}$ denote the Jacobian of the map $x \mapsto \mathcal{N}_L^{res}(M)$. Again, restating Lemma 3,

$$\mathbb{E}[\text{len}(\mathcal{N}_L^{res}(M))^2] \leq \int_0^1 \mathbb{E}[\|J_{L,\gamma(t)}^{res}\gamma'(t)\|^2]\, dt. \tag{24}$$

We begin by rewriting $J_{L,x}^{res}$ in terms of the Jacobians of fully-connected neural networks, which we examined in the previous section.

Define $J_{\ell,x}^{res}$ to be the Jacobian of the map $x \mapsto \mathcal{N}_\ell^{res}(x)$ and $J_{\ell,x}$ to be the Jacobian of the map $x \mapsto \mathcal{N}_\ell(x)$. Then by the chain rule,

$$J_{\ell,x}^{res} = (I + \eta_\ell J_{\ell,\mathcal{N}_{\ell-1}^{res}(x)})J_{\ell-1,x}^{res}. \tag{25}$$

Thus,

$$J_{L,x}^{res} = (I + \eta_L J_{L,\mathcal{N}_{L-1}^{res}(x)})(I + \eta_{L-1} J_{L-1,\mathcal{N}_{L-2}^{res}(x)}) \cdots (I + \eta_1 J_{1,\mathcal{N}_0^{res}(x)}). \tag{26}$$

In the remainder of the section, we consider a fixed x; so for convenience, we abbreviate $J_{\ell,\mathcal{N}_{\ell-1}^{res}(x)}$ to just J_ℓ. Using this notation, we expand (26) to get

$$J_{L,x}^{res} = \sum_{k=0}^{L} \sum_{L \geq \ell_1 > \cdots > \ell_k \geq 1} \eta_{\ell_1} \cdots \eta_{\ell_k} J_{\ell_1} \cdots J_{\ell_k}. \tag{27}$$

We now state the distribution of $J_{\ell_1} \cdots J_{\ell_k}$. Note that the Jacobians J_ℓ are independent. Therefore, we get the following extension of (14):

$$J_{\ell_1} \cdots J_{\ell_k} \overset{d}{=} A \prod_{i=1}^{k} W_{\ell_i}^{(L_{\ell_i})} D_{\ell_i}^{(L_{\ell_i}-1)} W_{\ell_i}^{(L_{\ell_i}-1)} \cdots D_{\ell_i}^{(1)} W_{\ell_i}^{(1)}, \tag{28}$$

where all of the matrices in the product are independent, and where A is a diagonal matrix with independent diagonal entries that are ±1 with equal probability, $D_{\ell_i}^{(\ell)}$ are diagonal matrices with diagonal entries that are independent Bernoulli(1/2) random variables, and we recall that each $W_{\ell_i}^{(\ell)}$ is determined by standard He initialization. As in the previous section, we see that

$$\mathbb{E}[\text{len}(\mathcal{N}_L^{res}(M))^2] \leq \int_0^1 \mathbb{E}[\|J_{\gamma(t)}^{res}\gamma'(t)\|^2]\, dt = \mathbb{E}[\|J_{\ell,x}^{res}u\|^2], \tag{29}$$

where we applied Lemma 3, and noted that (27) and (28) imply $\mathbb{E}[\|J_{\ell,x}^{res}u\|^2]$ is the same for all choices of x and unit vector u.

We now turn to calculating $\mathbb{E}[\|J_{\ell,x}^{res}u\|^2]$. We use the expansion in (27) to obtain

$$\begin{aligned}\mathbb{E}\left[\|J_{\ell,x}^{res}u\|^2\right] &= \mathbb{E}\left[\sum \left\langle \eta_{\ell_1} \cdots \eta_{\ell_k} J_{\ell_1} \cdots J_{\ell_k} u, \eta_{\ell'_1} \cdots \eta_{\ell'_k} J_{\ell'_1} \cdots J_{\ell'_{k'}} u \right\rangle\right] &(30)\\ &= \sum (\eta_{\ell_1} \cdots \eta_{\ell_k})(\eta_{\ell'_1} \cdots \eta_{\ell'_k}) \mathbb{E}\left[\left\langle J_{\ell_1} \cdots J_{\ell_k} u, J_{\ell'_1} \cdots J_{\ell'_{k'}} u \right\rangle\right], &(31)\end{aligned}$$

where the sums each range over $0 \leq k, k' \leq L$ and $L \geq \ell_1 > \cdots > \ell_k \geq 1, L \geq \ell'_1 > \cdots, > \ell'_{k'} \geq 1$. In Proposition 5 and Proposition 6, we calculate the expectations of the above inner products in two cases: $(\ell_1, \cdots, \ell_k) = (\ell'_1, \cdots, \ell'_{k'})$ and $(\ell_1, \cdots, \ell_k) \neq (\ell'_1, \cdots, \ell'_{k'})$.

Proposition 5. *For $0 \leq k \leq L$ and $L \geq \ell_1 > \cdots > \ell_k \geq 1$,*

$$\mathbb{E}\left[\langle J_{\ell_1} \cdots J_{\ell_k} u, J_{\ell_1} \cdots J_{\ell_k} u\rangle\right] = 1. \tag{32}$$

Proof. We have from (28) that

$$J_{\ell_1} \cdots J_{\ell_k} u \overset{d}{=} A \left(\prod_{i=1}^{k} W_{\ell_i}^{(L_{\ell_i})} D_{\ell_i}^{(L_{\ell_i}-1)} W_{\ell_i}^{(L_{\ell_i}-1)} \cdots D_{\ell_i}^{(1)} W_{\ell_i}^{(1)} \right) u.$$

Then by the same logic used to obtain (18), we may use the rotational invariance of the matrices $W_{\ell_i}^{(\ell)}$ to find that $\|J_{\ell_1} \cdots J_{\ell_k} u\|$ is equal to the product of independent random variables

$$\begin{aligned} \|J_{\ell_1} \cdots J_{\ell_k} u\| &\overset{d}{=} \prod_{i=1}^{k} \left\| W_{\ell_i}^{(L_{\ell_i})} u \right\| \left\| D_{\ell_i}^{(L_{\ell_i}-1)} W_{\ell_i}^{(L_{\ell_i}-1)} u \right\| \cdots \left\| D_{\ell_i}^{(1)} W_{\ell_i}^{(1)} u \right\| \\ &\overset{d}{=} \prod_{i=1}^{k} \|J_{\ell_i} u\|. \end{aligned}$$

Thus,

$$\mathbb{E}\left[\|J_{\ell_1} \cdots J_{\ell_k} u\|^2\right] = \prod_{i=1}^{k} \mathbb{E}[\|J_{\ell_i} u\|^2]. \tag{33}$$

Finally, recalling Proposition 4 and noting that each residual module has equal input and output dimension, we have

$$\mathbb{E}[\|J_{\ell_i} u\|^2] = 1, \tag{34}$$

from which the result follows. □

Proposition 6. *For $0 \leq k, k' \leq L$ and $L \geq \ell_1 > \cdots > \ell_k \geq 1, L \geq \ell'_1 > \cdots, > \ell'_{k'} \geq 1$, if $(\ell_1, \cdots, \ell_k) \neq (\ell'_1, \cdots, \ell'_{k'})$, then*

$$\mathbb{E}\left[\left\langle J_{\ell_1} \cdots J_{\ell_k} u, J_{\ell'_1} \cdots J_{\ell'_{k'}} u \right\rangle\right] = 0. \tag{35}$$

Proof. There exists some index i that is among either $\{\ell_1, \cdots, \ell_k\}$ or $\{\ell'_1, \cdots, \ell'_{k'}\}$ but not both. Assume without loss of generality that $i \in \{\ell_1, \cdots, \ell_k\}$. Then $J_i \overset{d}{=} -J_i$. (This can be seen, for example, by observing that $W_i^{(L_i)} \overset{d}{=} -W_i^{(L_i)}$.)

Because J_i and J_j are independent for all $i \neq j$, we have that

$$\begin{aligned} \left\langle J_{\ell_1} \cdots J_i \cdots J_{\ell_k} u, J_{\ell'_1} \cdots J_{\ell'_{k'}} u \right\rangle &\overset{d}{=} - \left\langle J_{\ell_1} \cdots (-J_i) \cdots J_{\ell_k} u, J_{\ell'_1} \cdots J_{\ell'_{k'}} \right\rangle \\ &\overset{d}{=} - \left\langle J_{\ell_1} \cdots J_i \cdots J_{\ell_k} u, J_{\ell'_1} \cdots J_{\ell'_{k'}} u \right\rangle. \end{aligned}$$

The result follows. □

Applying Proposition 5 and Proposition 6 to (31), we have

$$\mathbb{E}\left[\|J_{\ell,x}^{res} u\|^2\right] = \sum (\eta_{\ell_1} \cdots \eta_{\ell_k})(\eta_{\ell'_1} \cdots \eta_{\ell'_k}) \mathbb{E}\left[\left\langle J_{\ell_1} \cdots J_{\ell_k} u, J_{\ell'_1} \cdots J_{\ell'_{k'}} u \right\rangle\right] \tag{36}$$

$$= \sum_{k=0}^{L} \sum_{L \geq \ell_1 > \cdots > \ell_k \geq 1} \eta_{\ell_1}^2 \cdots \eta_{\ell_k}^2 \tag{37}$$

$$= \prod_{\ell=1}^{L} (1 + \eta_\ell^2), \tag{38}$$

which completes the proof of Theorem 2.

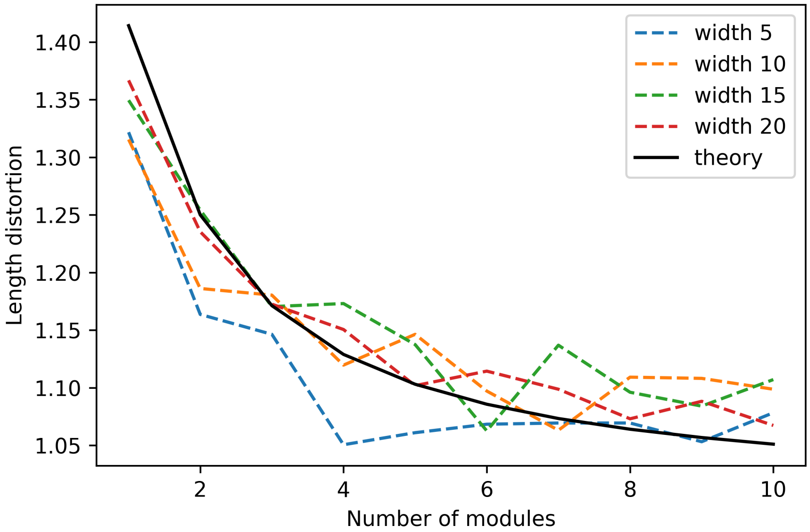

Figure 1: We compare our theoretical upper bound and empirical estimates (from [HJR21]) for expected length distortion $\mathbb{E}[\text{len}(\mathcal{N}(M))]$ when $\eta_1 = \cdots = \eta_L = \frac{1}{L}$. Our upper bound appears to be fairly tight.

2.3 Comparison with experimental results

We consider the special case when $\eta_1 = \cdots = \eta_L = \frac{1}{L}$ and compare our theoretical bounds with empirical findings given in [HJR21]. In this case, Theorem 2 gives

$$\mathbb{E}[\text{len}(\mathcal{N}(M))] \leq \left(\mathbb{E}[\text{len}(\mathcal{N}(M))^2]\right)^{1/2} \leq \left(1 + \frac{1}{L^2}\right)^{L/2}. \tag{39}$$

We plot our theoretical upper bounds alongside the empirical results in Figure 1.

3 Volume distortion

In the previous chapter, we considered the expected length distortion of fully-connected and residual networks. It is natural to further study the effect of neural networks on higher-dimensional manifolds. In this section, we consider unit volume smooth manifolds M of any dimension and bound the expected volume of its image. This gives the expected *volume distortion*. Below, we state a result of [HJR21] for fully-connected networks, as well as a new bound we obtain for residual networks.

Theorem 7 (Volume distortion in deep fully-connected networks)**.** *Consider a fully-connected neural network $\mathcal{N}$ of depth L, input dimension n_0, output dimension n_L, and hidden layer widths $n_1, \cdots, n_{L-1}$, with weights given by standard He initialization. Consider a $d-$dimensional* *input M of unit volume. Then*

$$\mathbb{E}[\mathrm{vol}_d(\mathcal{N}(M))^2] \leq \left(\frac{n_L}{n_0}\right)^d \exp\left[-\binom{d}{2}\sum_{\ell=1}^{L} n_\ell^{-1}\right]. \tag{40}$$

Theorem 8 (Volume distortion in deep residual networks)**.** *Consider a residual neural network $\mathcal{N}_L^{res}$ with residual modules $\mathcal{N}_1, \cdots, \mathcal{N}_L$ (each with weights given by standard He initialization) and scales $\eta_1, \cdots, \eta_L$ such that $\mathcal{N}_L^{res}$ has a total of $\tilde{L}$ layers (i.e., the sum of the depths of $\mathcal{N}_1, \cdots, \mathcal{N}_L$ equals $\tilde{L}$). Consider a $d-$dimensional input M of unit volume. Then there exists a universal constant $c > 0$ such that for any $0 < \epsilon < \frac{1}{2}$, if the width of every layer of $\mathcal{N}_L^{res}$ is at least $\frac{cd^2\tilde{L}}{\epsilon}$, then*

$$\mathbb{E}[\mathrm{vol}_d(\mathcal{N}_L^{res}(M))^2] \leq (1+\epsilon)\prod_{\ell=1}^{L}\frac{(1+\eta_\ell)^{2d} + (1-\eta_\ell)^{2d}}{2}. \tag{41}$$

We observe that the bound in Theorem 8 depends on $\mathcal{N}_L^{res}$ having a width that is large in terms of the squared dimension of the manifold and the total number of layers. In Section 3.2, we discuss how this dependency arises and whether it is necessary.

As for length distortion, we begin by introducing a general strategy for studying volume distortion. Consider a smooth $d-$dimensional manifold M where d is at most the minimum width of a neural network $\mathcal{N}$. (Otherwise, $\mathcal{N}(M)$ would have fewer than d dimensions.) Let J_x be the Jacobian of the map $x \to \mathcal{N}(x)$. We employ the following generalization of Lemma 3.

$$\mathbb{E}[(\mathcal{N}(M))^2] \leq \int_M \mathbb{E}[\det(\Pi_{T_xM} J_x^T J_x \Pi_{T_xM})] \,\mathrm{vol}_d(dx), \tag{42}$$

where $\Pi_{T_xM} : \mathbb{R}^{n_0} \to T_xM$ is the orthogonal projection onto the tangent space of x with respect to M.

From the Gram identity,

$$\det(\Pi_{T_xM} J_x^T J_x \Pi_{T_xM}) = \|J_x e_1 \wedge \cdots \wedge J_x e_d\|^2, \tag{43}$$

where $e_1, \cdots, e_d$ is an orthonormal basis of the tangent space of M with respect to x. We recall that $\wedge$ denotes the exterior product and that $\|v_1 \wedge \cdots \wedge v_d\|$ gives the $d-$dimensional volume of the parallelepiped formed by $v_1, \cdots, v_d$. It suffices to bound

$$\mathbb{E}\left[\|J_x e_1 \wedge \cdots \wedge J_x e_d\|^2\right]. \tag{44}$$

In Section 3.1, we sketch the proof of Theorem 7. In Section 3.2, we prove Theorem 8.

3.1 Volume distortion in deep fully-connected networks

We present the key ideas to prove Theorem 7, which are almost identical to that of Theorem 1.

Consider a fully-connected network $\mathcal{N}$ of depth L, input dimension n_0, and output dimension n_L, with weights given by standard He initialization. Let J_x denote the Jacobian of the map $x \to \mathcal{N}(x)$. We would like to bound

$$\mathbb{E}\left[\|J_x e_1 \wedge \cdots \wedge J_x e_d\|^2\right] \tag{45}$$

where $e_1, \cdots, e_d$ is an orthonormal basis of the tangent space of M with respect to x.

As for length distortion, we may determine the distribution of $\mathbb{E}\left[\|J_x e_1 \wedge \cdots \wedge J_x e_d\|^2\right]$.

Proposition 9. *For any x and orthonormal unit vectors $e_1, \cdots, e_d$,*

$$\|J_x e_1 \wedge \cdots \wedge J_x e_d\|^2$$

is equal in distribution to the product of independent chi-squared random variables

$$\left(\frac{n_L}{n_0}\right)^d \left(\prod_{\ell=1}^{L-1} \prod_{j=1}^{d} \frac{2}{n_\ell} \chi_{\mathrm{Bin}(n_\ell - j+1, 1/2)}\right) \cdot \prod_{j=1}^{d} \frac{1}{n_L} \chi^2_{n_L - j+1} \tag{46}$$

where $\mathrm{Bin}(n_\ell, 1/2)$ *are independent binomial distributions determining the number of degrees of freedom.*

Theorem 7 follows directly. The key idea to proving Proposition 9 is the same as in the two-dimensional case (see (18)). We may write

$$\|J_x e_1 \wedge \cdots \wedge J_x e_d\| \overset{d}{=} \left\|W^{(L)} e_1 \wedge \cdots \wedge W^{(L)} e_d\right\| \prod_{\ell=1}^{L-1} \left\|D^{(\ell)} W^{(\ell)} e_1 \wedge \cdots \wedge D^{(\ell)} W^{(\ell)} e_d\right\|, \tag{47}$$

the product of independent random variables. This is possible because for W with i.i.d. Gaussian entries, $Wu_1 \wedge \cdots \wedge Wu_d$ is independent of $u_1, \cdots, u_d$ for any orthonormal unit vectors $u_1, \cdots, u_d$ and equal in distribution to $Wv_1 \wedge \cdots \wedge Wv_d$ for any orthonormal unit vectors $v_1, \cdots, v_d$.

3.2 Volume distortion in deep residual networks

We now prove Theorem 8. Again, consider a residual network $\mathcal{N}_L^{res}$ with residual modules $\mathcal{N}_1, \cdots, \mathcal{N}_L$ and scales $\eta_1, \cdots, \eta_L$, and let $J_{\ell,x}^{res}$ denote the Jacobian of the map $x \to \mathcal{N}_L^{res}(x)$ and J_ℓ the Jacobian of the map $x \to \mathcal{N}_\ell(x)$.

Recall from (27) that

$$J_{L,x}^{res} = \sum_{k=0}^{L} \sum_{L \geq \ell_1 > \cdots > \ell_k \geq 1} \eta_{\ell_1} \cdots \eta_{\ell_k} J_{\ell_1} \cdots J_{\ell_k}. \tag{48}$$

We introduce the following additional notation for this section. Let $2^{[L]}$ denote the power set of $[L] = \{1, 2, \cdots, L\}$. Then for $\boldsymbol{\ell} = \{\ell_1, \cdots, \ell_k\} \in 2^{[L]}$ with $\ell_1 > \cdots > \ell_k$, we let $\eta_{\boldsymbol{\ell}} = \eta_{\ell_1} \cdots \eta_{\ell_k}$ and $J_{\boldsymbol{\ell}} = J_{\ell_1} \cdots J_{\ell_k}$. With this notation, we rewrite (48) as

$$J_{L,x}^{res} = \sum_{\boldsymbol{\ell} \in 2^{[L]}} \eta_{\boldsymbol{\ell}} J_{\boldsymbol{\ell}}. \tag{49}$$

We now bound, for $x \in M$ and $e_1, \cdots, e_d$ an orthonormal basis of the tangent space of M at x,

$$\mathbb{E}\left[\|J_{L,x}^{res} e_1 \wedge \cdots \wedge J_{L,x}^{res} e_d\|^2\right] = \mathbb{E}\left[\left\|\bigwedge_{i=1}^{d} \left(\sum_{\boldsymbol{\ell} \in 2^{[L]}} \eta_{\boldsymbol{\ell}} J_{\boldsymbol{\ell}}\right) e_i\right\|^2\right]. \tag{50}$$

We expand this to get

$$\sum_{\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_d \in 2^{[L]}} \sum_{\boldsymbol{\ell}'_1, \cdots, \boldsymbol{\ell}'_d \in 2^{[L]}} \left(\prod_{i=1}^{d} \eta_{\boldsymbol{\ell}_i}\right) \left(\prod_{i=1}^{d} \eta_{\boldsymbol{\ell}'_i}\right) \mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle\right] \tag{51}$$

The following three results address the expectations of the inner products in (51).

Proposition 10. *There exists a universal constant c such that for any $0 < \epsilon < \frac{1}{2}$, if the width of every layer of $\mathcal{N}_L^{res}$ is at least $\frac{cd^2\tilde{L}}{\epsilon}$, then for any $\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_d \in 2^{[L]}$,*

$$\mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i \right\rangle\right] \leq 1 + \epsilon. \tag{52}$$

We recall Theorem 5.1 in [HJR21] as a lemma.

Lemma 11. *Let J_x be the Jacobian of a fully-connected network $\mathcal{N}$ with input dimension n_0, output dimension n_L, and hidden layer widths $n_1, \cdots, n_{L-1}$, with weights and biases given by standard He intialization. Then there exist univeral constants $c_1, c_2 > 0$ such that if $m < c_1 \min\{n_1, \cdots, n_{L-1}\}$, then*

$$\mathbb{E}\left[\|J_x u\|^m\right] \leq \left(\frac{n_L}{n_0}\right)^{m/2} \exp\left[c_2 m^2 \sum_{\ell=1}^{L} n_\ell^{-1}\right]. \tag{53}$$

Proof of Proposition 10. We use the naïve bound

$$\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i \right\rangle \leq \prod_{i=1}^{d} \|J_{\boldsymbol{\ell}_i} e_i\|^2 \tag{54}$$

where u is any unit vector. Using Lemma 11,

$$\mathbb{E}\left[\prod_{i=1}^{d} \|J_{\boldsymbol{\ell}_i} e_i\|^2\right] \leq \mathbb{E}\left[\prod_{\ell=1}^{L} \|J_\ell u\|^{2d}\right] \leq \exp\left[\sum_{\ell=1}^{L} \frac{4c_2 d^2 L_\ell}{cd^2 \tilde{L} \frac{1}{\epsilon}}\right] \tag{55}$$

where L_ℓ is the number of layers of $\mathcal{N}_\ell$. Taking $c = 8c_2$,

$$\exp\left[\sum_{\ell=1}^{L} \frac{4c_2 d^2 L_\ell}{cd^2 \tilde{L} \frac{1}{\epsilon}}\right] \leq \exp\left[\frac{\epsilon}{2}\right] \leq 1 + \epsilon, \tag{56}$$

as desired, where the last inequality follows because $\epsilon < 1/2$. □

Proposition 10 may be extended to more general inner products.

Corollary 12. *There exists a universal constant c such that for any $0 < \epsilon < \frac{1}{2}$, if the width of every layer of $\mathcal{N}_L^{res}$ is at least $\frac{cd^2\tilde{L}}{\epsilon}$, then for any $\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_d \in 2^{[L]}$ and $\boldsymbol{\ell}'_1, \cdots, \boldsymbol{\ell}'_d \in 2^{[L]}$,*

$$\mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle\right] \leq 1 + \epsilon. \tag{57}$$

Proof. Taking $c = 8c_2$ as in Proposition 10, we have

$$\begin{aligned}
\mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle\right] &\leq \mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i \right\rangle^{1/2} \left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle^{1/2}\right] && (58)\\
&\leq \left(\mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i \right\rangle\right] \mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle\right]\right)^{1/2} && (59)\\
&\leq 1 + \epsilon && (60)
\end{aligned}$$

where (58) and (59) each follow from Cauchy-Schwarz and (60) follows from Proposition 10. □

We now show how in certain cases, the expected inner product is zero.

Proposition 13. *For $\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_d \in 2^{[L]}$ and $\boldsymbol{\ell}'_1, \cdots, \boldsymbol{\ell}'_d \in 2^{[L]}$, if there exists an index i such that i appears in an odd number of $\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_d$ and $\boldsymbol{\ell}'_1, \cdots, \boldsymbol{\ell}'_d \in 2^{[L]}$, then*

$$\mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle\right] = 0. \tag{61}$$

Proof. The proof is nearly identical to that of Proposition 6. Because $J_i \overset{d}{=} -J_i$ and is independent of all J_j where $i \neq j$, we see that

$$\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle \overset{d}{=} -\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle, \tag{62}$$

where we used that J_i appears an odd number of times among $J_{\boldsymbol{\ell}_1}, \cdots, J_{\boldsymbol{\ell}_d}, J_{\boldsymbol{\ell}'_1}, \cdots, J_{\boldsymbol{\ell}'_d}$. The result follows. □

Theorem 7 follows shortly. Applying Proposition 13 to (51),

$$\sum_{\boldsymbol{\ell}_1, \cdots, \boldsymbol{\ell}_d \in 2^{[L]}} \sum_{\boldsymbol{\ell}'_1, \cdots, \boldsymbol{\ell}'_d \in 2^{[L]}} \left(\prod_{i=1}^{d} \eta_{\boldsymbol{\ell}_i}\right) \left(\prod_{i=1}^{d} \eta_{\boldsymbol{\ell}'_i}\right) \mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle\right], \tag{63}$$

viewed as a polynomial in $\eta_1, \cdots, \eta_L$, vanishes at all terms where some η_ℓ is raised to an odd power. The remaining terms are found in

$$\prod_{\ell=1}^{L} \left(\binom{2d}{0} + \binom{2d}{2} \eta_\ell^2 + \binom{2d}{4} \eta_\ell^4 \cdots + \binom{2d}{2d} \eta_\ell^{2d} \right) = \prod_{\ell=1}^{L} \frac{(1+\eta_\ell)^{2d} + (1-\eta_\ell)^{2d}}{2}. \tag{64}$$

The result follows after applying the bound in Corollary 12.

Remark 14. *The bounds used to obtain Proposition 10 and Corollary 12 are very naïve. Future work can study*

$$\mathbb{E}\left[\left\langle \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}_i} e_i, \bigwedge_{i=1}^{d} J_{\boldsymbol{\ell}'_i} e_i \right\rangle\right] \tag{65}$$

more carefully. We do, however, expect there to remain some dependency on d, the number of layers, and the layer widths. This is because the inner products of interest (65) *can depend on the higher moments of $\|J_\ell u\|$, requiring the use of bounds like Lemma 11 that contain these dependencies.*

Let us consider one example, briefly. We have

$$\mathbb{E}\left[\langle J_2 J_1 e_1 \wedge J_3 J_1 e_2, J_2 J_1 e_1 \wedge J_3 J_1 e_2 \rangle\right] \tag{66}$$

$$= \mathbb{E}\left[\langle J_2 u_1 \wedge J_3 u_2, J_2 u_1 \wedge J_3 u_2 \rangle\right] \mathbb{E}[\|J_1 e_1\| \|J_1 e_2\|], \tag{67}$$

where $u_1 = \frac{J_1 e_1}{\|J_1 e_1\|}$ and $u_2 = \frac{J_1 e_2}{\|J_1 e_2\|}$. The term $\mathbb{E}[\|J_1 e_1\| \|J_1 e_2\|]$ has some dependency on the second moment of $\|J_1 u\|$.

References

[BS14] Monica Bianchini and Franco Scarselli. On the complexity of neural network classifiers: A comparison between shallow and deep architectures. *IEEE Transactions on Neural Networks and Learning Systems*, 25:1553–1565, 2014.

[CSS15] Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. *arXiv: Neural and Evolutionary Computing*, 2015.

[HJR21] Boris Hanin, Ryan Jeong, and David Rolnick. Deep relu networks preserve expected length. *ArXiv*, abs/2102.10492, 2021.

[HR19] Boris Hanin and David Rolnick. Deep relu networks have surprisingly few activation patterns. *ArXiv*, abs/1906.00904, 2019.

[MPCB14] Guido Montúfar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. On the number of linear regions of deep neural networks. In *NIPS*, 2014.

[PKL19] Giacomo De Palma, Bobak Toussi Kiani, and Seth Lloyd. Deep neural networks are biased towards simple functions. In *NeurIPS*, 2019.

[PT21] Ilan Price and Jared Tanner. Trajectory growth lower bounds for random sparse deep relu networks. *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1004–1009, 2021.

[RPK$^+$17] Maithra Raghu, Ben Poole, Jon M. Kleinberg, Surya Ganguli, and Jascha Sohl-Dickstein. On the expressive power of deep neural networks. *ArXiv*, abs/1606.05336, 2017.