# Depth-Width Tradeoffs in Approximating Natural Functions with Neural Networks

Itay Safran
Weizmann Institute of Science
itay.safran@weizmann.ac.il

Ohad Shamir
Weizmann Institute of Science
ohad.shamir@weizmann.ac.il

## Abstract

We provide several new depth-based separation results for feed-forward neural networks, proving that various types of simple and natural functions can be better approximated using deeper networks than shallower ones, even if the shallower networks are much larger. This includes indicators of balls and ellipses; non-linear functions which are radial with respect to the $L_1$ norm; and smooth non-linear functions. We also show that these gaps can be observed experimentally: Increasing the depth indeed allows better learning than increasing width, when training neural networks to learn an indicator of a unit ball.

## 1  Introduction

Deep learning, in the form of artificial neural networks, has seen a dramatic resurgence in the past recent years, achieving great performance improvements in various fields of artificial intelligence such as computer vision and speech recognition. While empirically successful, our theoretical understanding of deep learning is still limited at best.

An emerging line of recent works has studied the *expressive power* of neural networks: What functions can and cannot be represented by networks of a given architecture (see related work section below). A particular focus has been the trade-off between the network's width and depth: On the one hand, it is well-known that large enough networks of depth 2 can already approximate any continuous target function on $[0, 1]^d$ to arbitrary accuracy (Cybenko, 1989; Hornik, 1991). On the other hand, it has long been evident that deeper networks tend to perform better than shallow ones, a phenomenon supported by the intuition that depth, providing compositional expressibility, is necessary for efficiently representing some functions. Indeed, recent empirical evidence suggests that even at large depths, deeper networks can offer benefits over shallower networks (He et al., 2015).

To demonstrate the power of depth in neural networks, a clean and precise approach is to prove the existence of functions which can be expressed (or well-approximated) by moderately-sized networks of a given depth, yet cannot be approximated well by shallower networks, even if their size is much larger. However, the mere existence of such functions is not enough: Ideally, we would like to show such depth separation results using *natural*, interpretable functions, of the type we may expect neural networks to successfully train on. Proving that depth is necessary for such functions can give us a clearer and more useful insight into what various neural network architectures can and cannot express in practice.

In this paper, we provide several contributions to this emerging line of work. We focus on standard, vanilla feedforward networks (using some fixed activation function, such as the popular ReLU), and measure

expressiveness directly in terms of approximation error, defined as the expected squared loss with respect to some distribution over the input domain. In this setting, we show the following:

- We prove that the indicator of the Euclidean unit ball, $\mathbf{x} \mapsto \mathbb{1}\left(\|\mathbf{x}\| \leq 1\right)$ in $\mathbb{R}^d$, which can be easily approximated to accuracy $\epsilon$ using a 3-layer network with $\mathcal{O}(d/\epsilon)$ neurons, cannot be approximated to an accuracy higher than $\mathcal{O}(1/d^4)$ using a 2-layer network, unless its width is exponential in $d$. In fact, we show the same result more generally, for any indicator of an ellipsoid $\mathbf{x} \mapsto \mathbb{1}\left(\|A\mathbf{x} + \mathbf{b}\| \leq r\right)$ (where $A$ is a non-singular matrix and $\mathbf{b}$ is a vector). The proof is based on a reduction from the main result of Eldan & Shamir (2016), which shows a separation between 2-layer and 3-layer networks using a more complicated and less natural radial function.

- We show that this depth/width trade-off can also be observed experimentally: Specifically, that the indicator of a unit ball can be learned quite well using a 3-layer network, using standard backpropagation, but learning the same function with a 2-layer network (even if much larger) is significantly more difficult. Our theoretical result indicates that this gap in performance is due to approximation error issues. This experiment also highlights the fact that our separation result is for a natural function that is not just well-approximated by some 3-layer network, but can also be learned well from data using standard methods.

- We prove that *any* $L_1$ radial function $\mathbf{x} \mapsto f(\|\mathbf{x}\|_1)$, where $\mathbf{x} \in \mathbb{R}^d$ and $f : \mathbb{R} \to \mathbb{R}$ is piecewise-linear, cannot be approximated to accuracy $\epsilon$ by a depth 2 ReLU network of width less than $\tilde{\Omega}(\min\{1/\epsilon, \exp(\Omega(d))\})$. In contrast, such functions can be represented *exactly* by 3-layer ReLU networks.

- Finally, we prove that any member of a wide family of non-linear and twice-differentiable functions (including for instance $x \mapsto x^2$ in $[0, 1]$), which can be approximated to accuracy $\epsilon$ using ReLU networks of depth and width $\mathcal{O}(\text{poly}(\log(1/\epsilon)))$, cannot be approximated to similar accuracy by constant-depth ReLU networks, unless their width is at least $\Omega(\text{poly}(1/\epsilon))$. We note that a similar result appeared online concurrently and independently of ours in Yarotsky (2016); Liang & Srikant (2016), but the setting is a bit different (see related work below for more details).

## Related Work

The question of studying the effect of depth in neural network has received considerable attention recently, and studied under various settings. Many of these works consider a somewhat different setting than ours, and hence are not directly comparable. These include networks which are not plain-vanilla ones (e.g. Cohen et al. (2016); Delalleau & Bengio (2011); Martens & Medabalimi (2014)), measuring quantities other than approximation error (e.g. Bianchini & Scarselli (2014); Poole et al. (2016)), focusing only on approximation upper bounds (e.g. Shaham et al. (2016)), or measuring approximation error in terms of $L_\infty$-type bounds, i.e. $\sup_{\mathbf{x}} |f(\mathbf{x}) - \tilde{f}(\mathbf{x})|$ rather than $L_2$-type bounds $\mathbb{E}_{\mathbf{x}}(f(\mathbf{x}) - \tilde{f}(\mathbf{x}))^2$ (e.g. Yarotsky (2016); Liang & Srikant (2016)). We note that the latter distinction is important: Although $L_\infty$ bounds are more common in the approximation theory literature, $L_2$ bounds are more natural in the context of statistical machine learning problems (where we care about the expected loss over some distribution). Moreover, $L_2$ approximation lower bounds are stronger, in the sense that an $L_2$ lower bound easily translates to a lower bound on $L_\infty$ lower bound, but not vice versa[1].

---

[1] To give a trivial example, ReLU networks always express continuous functions, and therefore can never approximate a discontinuous function such as $x \mapsto \mathbb{1}\left(x \geq 0\right)$ in an $L_\infty$ sense, yet can easily approximate it in an $L_2$ sense given any continuous distribution.

A noteworthy paper in the same setting as ours is Telgarsky (2016), which proves a separation result between the expressivity of ReLU networks of depth $k$ and depth $o\left(k/\log\left(k\right)\right)$ (for any $k$). This holds even for one-dimensional functions, where a depth $k$ network is shown to realize a saw-tooth function with $\exp(\mathcal{O}(k))$ oscillations, whereas any network of depth $o\left(k/\log\left(k\right)\right)$ would require a width super-polynomial in $k$ to approximate it by more than a constant. In fact, we ourselves rely on this construction in the proofs of our results in Sec. 5. On the flip side, in our paper we focus on separation in terms of the accuracy or dimension, rather than a parameter $k$. Moreover, the construction there relies on a highly oscillatory function, with Lipschitz constant exponential in $k$ almost everywhere. In contrast, in our paper we focus on simpler functions, of the type that are likely to be learnable from data using standard methods.

Our separation results in Sec. 5 (for smooth non-linear functions) are closely related to those of Yarotsky (2016); Liang & Srikant (2016), which appeared online concurrently and independently of our work, and the proof ideas are quite similar. However, these papers focused on $L_\infty$ bounds rather than $L_2$ bounds. Moreover, Yarotsky (2016) considers a class of functions different than ours in their positive results, and Liang & Srikant (2016) consider networks employing a mix of ReLU and threshold activations, whereas we consider a purely ReLU network.

Another relevant and insightful work is Poggio et al. (2016), which considers width vs. depth and provide general results on expressibility of functions with a compositional nature. However, the focus there is on worse-case approximation over general classes of functions, rather than separation results in terms of specific functions as we do here, and the details and setting is somewhat orthogonal to ours.

## 2  Preliminaries

In general, we let bold-faced letters such as $\mathbf{x} = (x_1, \ldots, x_d)$ denote vectors, and capital letters denote matrices or probabilistic events. $\|\cdot\|$ denotes the Euclidean norm, and $\|\cdot\|_1$ the 1-norm. $\mathbb{1}\left(\cdot\right)$ denotes the indicator function. We use the standard asymptotic notation $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ to hide constants, and $\tilde{\mathcal{O}}(\cdot)$ and $\tilde{\Omega}(\cdot)$ to hide constants and factors logarithmic in the problem parameters.

**Neural Networks.** We consider feed-forward neural networks, computing functions from $\mathbb{R}^d$ to $\mathbb{R}$. The network is composed of layers of neurons, where each neuron computes a function of the form $\mathbf{x} \mapsto \sigma(\mathbf{w}^\top \mathbf{x} + b)$, where $\mathbf{w}$ is a weight vector, $b$ is a bias term and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is a non-linear activation function, such as the ReLU function $\sigma(z) = [z]_+ = \max\{0, z\}$. Letting $\sigma(W\mathbf{x} + \mathbf{b})$ be a shorthand for $\left(\sigma(\mathbf{w}_1^\top \mathbf{x} + b_1), \ldots, \sigma(\mathbf{w}_n^\top \mathbf{x} + b_n)\right)$, we define a layer of $n$ neurons as $\mathbf{x} \mapsto \sigma(W\mathbf{x} + \mathbf{b})$. By denoting the output of the $i^{\text{th}}$ layer as $O_i$, we can define a network of arbitrary depth recursively by $O_{i+1} = \sigma(W_{i+1}O_i + \mathbf{b}_{i+1})$, where $W_i, \mathbf{b}_i$ represent the matrix of weights and bias of the $i^{\text{th}}$ layer, respectively. Following a standard convention for multi-layer networks, the final layer $h$ is a purely linear function with no bias, i.e. $O_h = W_h \cdot O_{h-1}$. We define the *depth* of the network as the number of layers $l$, and denote the number of neurons $n_i$ in the $i^{\text{th}}$ layer as the *size* of the layer. We define the *width* of a network as $\max_{i \in \{1, \ldots, l\}} n_i$. Finally, a *ReLU network* is a neural network where all the non-linear activations are the ReLU function. We use "2-layer" and "3-layer" to denote networks of depth 2 and 3. In particular, in our notation a 2-layer ReLU network has the form

$$\mathbf{x} \mapsto \sum_{i=1}^{n_1} v_i \cdot [\mathbf{w}_i^\top \mathbf{x} + b_i]_+$$

for some parameters $v_1, b_1, \ldots, v_{n_1}, b_{n_1}$ and $d$-dimensional vectors $\mathbf{w}_1, \ldots, \mathbf{w}_{n_1}$. Similarly, a 3-layer ReLU

network has the form

$$\sum_{i=1}^{n_2} u_i \left[ \sum_{j=1}^{n_1} v_{i,j} \left[ \mathbf{w}_{i,j}^\top \mathbf{x} + b_{i,j} \right]_+ + c_i \right]_+$$

for some parameters $\{u_i, v_{i,j}, b_{i,j}, c_i, \mathbf{w}_{i,j}\}$.

**Approximation error.** Given some function $f$ on a domain $\mathcal{X}$ endowed with some probability distribution (with density function $\mu$), we define the quality of its approximation by some other function $\tilde{f}$ as $\int_{\mathcal{X}} (f(\mathbf{x}) - \tilde{f}(\mathbf{x}))^2 \mu(\mathbf{x}) d\mathbf{x} = \mathbb{E}_{\mathbf{x} \sim \mu}[(f(\mathbf{x}) - \tilde{f}(\mathbf{x}))^2]$. We refer to this as approximation in the $L_2$-norm sense. In one of our results (Thm. 6), we also consider approximation in the $L_\infty$-norm sense, defined as $\sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x}) - \tilde{f}(\mathbf{x})|$. Clearly, this upper-bounds the (square root of the) $L_2$ approximation error defined above, so as discussed in the introduction, lower bounds on the $L_2$ approximation error (w.r.t. any distribution) are stronger than lower bounds on the $L_\infty$ approximation error.

# 3 Indicators of $L_2$ Balls and Ellipsoids

We begin by considering one of the simplest possible function classes on $\mathbb{R}^d$, namely indicators of $L_2$ balls (and more generally, ellipsoids). The ability to compute such functions is necessary for many useful primitives, for example determining if the distance between two points in Euclidean space is below or above some threshold (either with respect to the Euclidean distance, or a more general Mahalanobis distance). In this section, we show a depth separation result for such functions: Although they can be easily approximated with 3-layer networks, no 2-layer network can approximate it to high accuracy w.r.t. any distribution, unless its width is exponential in the dimension. This is formally stated in the following theorem:

**Theorem 1** (Inapproximability with 2-layer networks). *The following holds for some positive universal constants $c_1, c_2, c_3, c_4$, and any network employing an activation function satisfying Assumptions 1 and 2 in Eldan & Shamir (2016): For any $d > c_1$, and any non-singular matrix $A \in \mathbb{R}^{d \times d}$, $\mathbf{b} \in \mathbb{R}^d$ and $r \in (0, \infty)$, there exists a continuous probability distribution $\gamma$ on $\mathbb{R}^d$, such that for any function $g$ computed by a 2-layer network of width at most $c_3 \exp(c_4 d)$, and for the function $f(\mathbf{x}) = \mathbb{1}(\|A\mathbf{x} + \mathbf{b}\| \leq r)$, we have*

$$\int_{\mathbb{R}^d} (f(\mathbf{x}) - g(\mathbf{x}))^2 \cdot \gamma(\mathbf{x}) d\mathbf{x} \geq \frac{c_2}{d^4}.$$

We note that the assumptions from Eldan & Shamir (2016) are very mild, and apply to all standard activation functions, including ReLU, sigmoid and threshold.

The formal proof of Thm. 1 (provided below) is based on a reduction from the main result of Eldan & Shamir (2016), which shows the existence of a certain radial function (depending on the input $\mathbf{x}$ only through its norm) and a probability distribution which cannot be expressed by a 2-layer network, whose width is less than exponential in the dimension $d$ to more than constant accuracy. A closer look at the proof reveals that this function (denoted as $\tilde{g}$) can be expressed as a sum of $\Theta(d^2)$ indicators of $L_2$ balls of various radii. We argue that if we could have accurately approximated a given $L_2$ ball indicator with respect to all distributions, then we could have approximated all the indicators whose sum add up to $\tilde{g}$, and hence reach a contradiction. By a linear transformation argument, we show the same contradiction would have occured if we could have approximated the indicators of an non-degenerate ellipse with respect to any distribution. The formal proof is provided below:

*Proof of Thm. 1.* Assume by contradiction that for $f$ as described in the theorem, and for any distribution $\gamma$, there exists a two-layer network $\tilde{f}_\gamma$ of width at most $c_3 \exp(c_4 d)$, such that

$$\int_{\mathbf{x} \in \mathbb{R}^d} \left( f(\mathbf{x}) - \tilde{f}_\gamma(\mathbf{x}) \right)^2 \gamma(\mathbf{x}) d\mathbf{x} \leq \epsilon \leq \frac{c_2}{d^4}.$$

Let $\hat{A}$ and $\hat{\mathbf{b}}$ be a $d \times d$ non-singular matrix and vector respectively, to be determined later. We begin by performing a change of variables, $\mathbf{y} = \hat{A}\mathbf{x} + \hat{\mathbf{b}} \iff \mathbf{x} = \hat{A}^{-1}(\mathbf{y} - \hat{\mathbf{b}})$, $d\mathbf{x} = \left| \det\left( \hat{A}^{-1} \right) \right| \cdot d\mathbf{y}$, which yields

$$\int_{\mathbf{y} \in \mathbb{R}^d} \left( f\left( \hat{A}^{-1} \left( \mathbf{y} - \hat{\mathbf{b}} \right) \right) - \tilde{f}_\gamma \left( \hat{A}^{-1} \left( \mathbf{y} - \hat{\mathbf{b}} \right) \right) \right)^2$$
$$\cdot \gamma \left( \hat{A}^{-1} \left( \mathbf{y} - \hat{\mathbf{b}} \right) \right) \cdot \left| \det \left( \hat{A}^{-1} \right) \right| \cdot d\mathbf{y} \leq \epsilon. \tag{1}$$

In particular, let us choose the distribution $\gamma$ defined as $\gamma(\mathbf{z}) = |\det(\hat{A})| \cdot \mu(\hat{A}\mathbf{z} + \hat{\mathbf{b}})$, where $\mu$ is the (continuous) distribution used in the main result of Eldan & Shamir (2016) (note that $\gamma$ is indeed a distribution, since $\int_{\mathbf{z}} \gamma(\mathbf{z}) = (\det(\hat{A})) \int_{\mathbf{z}} \mu(\hat{A}\mathbf{z} + \hat{\mathbf{b}}) d\mathbf{z}$, which by the change of variables $\mathbf{x} = \hat{A}\mathbf{z} + \hat{\mathbf{b}}$, $d\mathbf{x} = |\det(\hat{A})| d\mathbf{z}$ equals $\int_{\mathbf{x}} \mu(\mathbf{x}) d\mathbf{x} = 1$). Plugging the definition of $\gamma$ in Eq. (1), and using the fact that $|\det(\hat{A}^{-1})| \cdot |\det(\hat{A})| = 1$, we get

$$\int_{\mathbf{y} \in \mathbb{R}^d} \left( f\left( \hat{A}^{-1} \left( \mathbf{y} - \hat{\mathbf{b}} \right) \right) - \tilde{f}_\gamma \left( \hat{A}^{-1} \left( \mathbf{y} - \hat{\mathbf{b}} \right) \right) \right)^2$$
$$\cdot \mu(\mathbf{y}) d\mathbf{y} \leq \epsilon. \tag{2}$$

Letting $z > 0$ be an arbitrary parameter, we now pick $\hat{A} = \frac{z}{r} A$ and $\hat{\mathbf{b}} = \frac{z}{r} \mathbf{b}$. Recalling the definition of $f$ as $\mathbf{x} \mapsto \mathbb{1}(\|A\mathbf{x} + \mathbf{b}\| \leq r)$, we get that

$$\int_{\mathbf{y} \in \mathbb{R}^d} \left( \mathbb{1}(\|\mathbf{y}\| \leq z) - \tilde{f}_\gamma \left( \frac{r}{z} A^{-1} \left( \mathbf{y} - \frac{z}{r} \mathbf{b} \right) \right) \right)^2$$
$$\cdot \mu(\mathbf{y}) d\mathbf{y} \leq \epsilon. \tag{3}$$

Note that $\tilde{f}_\gamma \left( \frac{r}{z} A^{-1} \left( y - \frac{z}{r} \mathbf{b} \right) \right)$ expresses a 2-layer network composed with a linear transformation of the input, and hence can be expressed in turn by a 2-layer network (as we can absorb the linear transformation into the parameters of each neuron in the first layer). Therefore, letting $\|f\|_{L_2(\mu)} = \sqrt{\int_{\mathbf{y}} f^2(\mathbf{y}) d\mathbf{y}}$ denote the norm in $L_2(\mu)$ function space, we showed the following: For any $z > 0$, there exists a 2-layer network $\tilde{f}_z$ such that

$$\left\| \left( \mathbb{1}(\|\cdot\| \leq z) - \tilde{f}_z(\cdot) \right) \right\|_{L_2(\mu)} \leq \sqrt{\epsilon}. \tag{4}$$

With this key result in hand, we now turn to complete the proof. We consider the function $\tilde{g}$ from Eldan & Shamir (2016), for which it was proven that no 2-layer network can approximate it w.r.t. $\mu$ to better than constant accuracy, unless its width is exponential in the dimension $d$. In particular $\tilde{g}$ can be written as

$$\tilde{g}(\mathbf{x}) = \sum_{i=1}^{n} \epsilon_i \cdot \mathbb{1}(\|\mathbf{x}\| \in [a_i, b_i]),$$

where $[a_i, b_i]$ are disjoint intervals, $\epsilon_i \in \{-1, +1\}$, and $n = \Theta(d^2)$ where $d$ is the dimension. Since $\tilde{g}$ can also be written as

$$\sum_{i=1}^{n} \epsilon_i \left( \mathbb{1} \left( \|\mathbf{x}\| \leq b_i \right) - \mathbb{1} \left( \|\mathbf{x}\| \leq a_i \right) \right),$$

we get by Eq. (4) and the triangle inequality that

$$\left\| \tilde{g}(\cdot) - \sum_{i=1}^{n} \epsilon_i \cdot (\tilde{f}_{b_i}(\cdot) - \tilde{f}_{a_i}(\cdot)) \right\|_{L_2(\mu)}$$

$$\leq \sum_{i=1}^{n} |\epsilon_i| \left( \left\| \left( \mathbb{1} \left( \|\cdot\| \leq b_i \right) - \tilde{f}_{b_i} \right) \right\|_{L_2(\mu)} \right.$$

$$\left. + \left\| \mathbb{1} \left( \|\cdot\| \leq a_i \right) - \tilde{f}_{a_i}(\cdot) \right\|_{L_2(\mu)} \right)$$

$$\leq 2n\sqrt{\epsilon}.$$

However, since a linear combination of $2n$ 2-layer neural networks of width at most $w$ is still a 2-layer network, of width at most $2nw$, we get that $\sum_{i=1}^{n} \epsilon_i \cdot (\tilde{f}_{b_i}(\cdot) - \tilde{f}_{a_i}(\cdot))$ is a 2-layer network, of width at most $\Theta(d^2) \cdot c_3 \exp(c_4 d)$, which approximates $\tilde{g}$ to an accuracy of less than $2n\sqrt{\epsilon} = \Theta(d^2) \cdot \sqrt{c_2/d^4} = \Theta(1) \cdot \sqrt{c_2}$. Hence, by picking $c_2, c_3, c_4$ sufficiently small, we get a contradiction to the result of Eldan & Shamir (2016), that no 2-layer network of width smaller than $c \exp(cd)$ (for some constant $c$) can approximate $\tilde{g}$ to more than constant accuracy, for a sufficiently large dimension $d$. $\qquad\square$
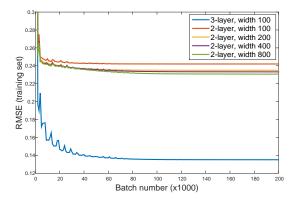
To complement Thm. 1, we also show that such indicator functions can be easily approximated with 3-layer networks. The argument is quite simple: Using an activation such as ReLU or Sigmoid, we can use one layer to approximate any Lipschitz continuous function on any bounded interval, and in particular $x \mapsto x^2$. Given a vector $\mathbf{x} \in \mathbb{R}^d$, we can apply this construction on each coordinate $x_i$ seperately, hence approximating $\mathbf{x} \mapsto \|\mathbf{x}\|^2 = \sum_{i=1}^{d} x_i^2$. Similarly, we can approximate $\mathbf{x} \mapsto \|A\mathbf{x} + \mathbf{b}\|$ for arbitrary fixed matrices $A$ and vectors $\mathbf{b}$. Finally, with a 3-layer network, we can use the second layer to compute a continuous approximation to the threshold function $z \mapsto \mathbb{1}(z \leq r)$. Composing these two layers, we get an arbitrarily good approximation to the function $\mathbf{x} \mapsto \mathbb{1}(\|A\mathbf{x} + \mathbf{b}\| \leq r)$ w.r.t. any continuous distribution, with the network size scaling polynomially with the dimension $d$ and the required accuracy. In the theorem below, we formalize this intuition, where for simplicity we focus on approximating the indicator of the unit ball:

**Theorem 2** (Approximability with 3-layer networks). *Given $\delta > 0$, for any activation function $\sigma$ satisfying Assumption 1 in Eldan & Shamir (2016) and any continuous probability distribution $\mu$ on $\mathbb{R}^d$, there exists a constant $c_\sigma$ dependent only on $\sigma$, and a function $g$ expressible by a 3-layer network of width at most $\max \left\{ 8c_\sigma d^2 / \delta, c_\sigma \sqrt{1/2\delta} \right\}$, such that the following holds:*

$$\int_{\mathbb{R}^d} \left( g(\mathbf{x}) - \mathbb{1}(\|\mathbf{x}\|_2 \leq 1) \right)^2 \mu(\mathbf{x}) \, d\mathbf{x} \leq \delta,$$

*where $c_\sigma$ is a constant depending solely on $\sigma$.*

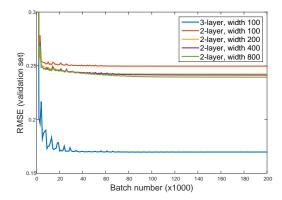The proof of the theorem appears in Subsection 6.1.

6

Figure 1: The experiment results, depicting the network's root mean square error over the training set (left) and validation set (right), as a function of the number of batches processed. Best viewed in color.

## 3.1 An Experiment

In this subsection, we empirically demonstrate that indicator functions of balls are indeed easier to learn with a 3-layer network, compared to a 2-layer network (even if the 2-layer network is significantly larger). This indicates that the depth/width trade-off for indicators of balls, predicted by our theory, can indeed be observed experimentally. Moreover, it highlights the fact that our separation result is for simple natural functions, that can be learned reasonably well from data using standard methods.

For our experiment, we sampled $5 \cdot 10^5$ data instances in $\mathbb{R}^{100}$, with a direction chosen uniformly at random and a norm drawn uniformly at random from the interval $[0, 2]$. To each instance, we associated a target value computed according to the target function $f(\mathbf{x}) = \mathbb{1}(\|\mathbf{x}\|_2 \leq 1)$. Another $5 \cdot 10^4$ examples were generated in a similar manner and used as a validation set.

We trained 5 ReLU networks on this dataset:

- One 3-layer network, with a first hidden layer of size $100$, a second hidden layer of size $20$, and a linear output neuron.

- Four 2-layer networks, with hidden layer of sizes $100, 200, 400$ and $800$, and a linear output neuron.

Training was performed with backpropagation, using the TensorFlow library. We used the squared loss $\ell(y, y') = (y - y')^2$ and batches of size 100. For all networks, we chose a momentum parameter of 0.95, and a learning rate starting at 0.1, decaying by a multiplicative factor of 0.95 every 1000 batches, and stopping at $10^{-4}$.

The results are presented in Fig. 1. As can be clearly seen, the 3-layer network achieves significantly better performance than the 2-layer networks. This is true even though some of these networks are significantly larger and with more parameters (for example, the 2-layer, width 800 network has ˜80K parameters, vs. ˜10K parameters for the 3-layer network). This gap in performance is the exact opposite of what might be expected based on parameter counting alone. Moreover, increasing the width of the 2-layer networks exhibits diminishing returns: The performance improvement in doubling the width from 100 to 200 is much larger than doubling the width from 200 to 400 or 400 to 800. This indicates that one would need a much larger 2-layer network to match the 3-layer, width 100 network's performance. Thus, we conclude that the network's depth indeed plays a crucial role, and that 3-layer networks are inherently more suitable to express indicator functions of the type we studied.

7

# 4  $L_1$ Radial Functions; ReLU Networks

Having considered functions depending on the $L_2$ norm, we now turn to consider functions depending on the $L_1$ norm. Focusing on ReLU networks, we will show a certain separation result holding for *any* non-linear function, which depends on the input $\mathbf{x}$ only via its 1-norm $\|\mathbf{x}\|_1$.

**Theorem 3.** *Let $f : [0, \infty) \mapsto \mathbb{R}$ be a function such that for some $r, \delta > 0$ and $\epsilon \in (0, 1/2)$,*

$$\inf_{a,b \in \mathbb{R}} \mathbb{E}_{x \text{ uniform on } [r,(1+\epsilon)r]}[(f(x) - (ax - b))^2] > \delta .$$

*Then there exists a distribution $\gamma$ over $\{\mathbf{x} : \|\mathbf{x}\|_1 \leq (1 + \epsilon)r\}$, such that if a 2-layer ReLU network $F(\mathbf{x})$ satisfies*

$$\int_{\mathbf{x}} (f(\|\mathbf{x}\|_1) - F(\mathbf{x}))^2 \gamma(\mathbf{x})d\mathbf{x} \leq \delta/2,$$

*then its width must be at least $\tilde{\Omega}(\min\{1/\epsilon, \exp(\Omega(d))\})$ (where the $\tilde{\Omega}$ notation hides constants and factors logarithmic in $\epsilon, d$).*

To give a concrete example, suppose that $f(z) = [z - 1]_+$, which cannot be approximated by a linear function better than $\mathcal{O}(\epsilon^2)$ in an $\epsilon$-neighborhood of 1. By taking $r = 1 - \frac{\epsilon}{2}$ and $\delta = \mathcal{O}(\epsilon^2)$, we get that no 2-layer network can approximate the function $[\|\mathbf{x}\|_1 - 1]_+$ (at least with respect to some distribution), unless its width is $\tilde{\Omega}(\min\{1/\epsilon, \exp(\Omega(d))\})$. On the flip side, $f(\|\mathbf{x}\|_1)$ can be expressed *exactly* by a 3-layer, width $2d$ ReLU network: $\mathbf{x} \mapsto [\sum_{i=1}^{d}([x_i]_+ + [-x_i]_+) - 1]_+$, where the output neuron is simply the identity function. The same argument would work for any piecewise-linear $f$. More generally, the same kind of argument would work for any function $f$ exhibiting a non-linear behavior at some points: Such functions can be well-approximated by 3-layer networks (by approximating $f$ with a piecewise-linear function), yet any approximating 2-layer network will have a lower bound on its size as specified in the theorem.

Intuitively, the proof relies on showing that any good 2-layer approximation of $f(\|\mathbf{x}\|_1)$ must capture the non-linear behavior of $f$ close to "most" points $\mathbf{x}$ satisfying $\|\mathbf{x}\|_1 \approx r$. However, a 2-layer ReLU network $\mathbf{x} \mapsto \sum_{j=1}^{N} a_j [\langle \mathbf{w}_j, \mathbf{x}\rangle + b_j]_+$ is piecewise linear, with non-linearities only at the union of the $N$ hyperplanes $\cup_j \{\mathbf{x} : \langle \mathbf{w}_j, \mathbf{x}\rangle + b_j = 0\}$. This implies that "most" points $\mathbf{x}$ s.t. $\|\mathbf{x}\|_1 \approx r$ must be $\epsilon$-close to a hyperplane $\{\mathbf{x} : \langle \mathbf{w}_j, \mathbf{x}\rangle + b_j = 0\}$. However, the geometry of the $L_1$ ball $\{\mathbf{x} : \|\mathbf{x}\| = r\}$ is such that the $\epsilon$ neighborhood of any single hyperplane can only cover a "small" portion of that ball, yet we need to cover most of the $L_1$ ball. Using this and an appropriate construction, we show that required number of hyperplanes is at least $1/\epsilon$, as long as $\epsilon > \exp(-\mathcal{O}(d))$ (and if $\epsilon$ is smaller than that, we can simply use one neuron/hyperplane for each of the $2^d$ facets of the $L_1$ ball, and get a covering using $2^d$ neurons/hyperplanes). The formal proof appears in Subsection 6.2.

We note that the bound in Thm. 3 is of a weaker nature than the bound in the previous section, in that the lower bound is only polynomial rather than exponential (albeit w.r.t. different problem parameters: $\epsilon$ vs. $d$). Nevertheless, we believe this does point out that $L_1$ balls also pose a geometric difficulty for 2-layer networks, and conjecture that our lower bound can be considerably improved: Indeed, at the moment we do not know how to approximate a function such as $\mathbf{x} \mapsto [\|\mathbf{x}\|_1 - 1]_+$ with 2-layer networks to better than constant accuracy, using less than $\Omega(2^d)$ neurons.

# 5  $C^2$ Nonlinear Functions; ReLU Networks

In this section, we establish a depth separation result for approximating continuously twice-differentiable ($C^2$) functions using ReLU neural networks. Unlike the previous results in this paper, the separation is for

depths which can be larger than 3, depending on the required approximation error. Also, the results will all be with respect to the uniform distribution $\mu_d$ over $[0,1]^d$. As mentioned earlier, the results and techniques in this section are closely related to the independent results of Yarotsky (2016); Liang & Srikant (2016), but our emphasis is on $L_2$ rather than $L_\infty$ approximation bounds, and we focus on somewhat different network architectures and function classes.

Clearly, not all $C^2$ functions are difficult to approximate (e.g. a linear function can be expressed exactly with a 2-layer network). Instead, we consider functions which have a certain degree of non-linearity, in the sense that its Hessians are non-zero along some direction, on a significant portion of the domain. Formally, we make the following definition:

**Definition 1.** *Let $\mu_d$ denote the uniform distribution on $[0,1]^d$. For a function $f : [0,1]^d \to \mathbb{R}$ and some $\lambda > 0$, denote*

$$\sigma_\lambda(f) = \sup_{\mathbf{v} \in \mathbb{S}^{d-1}, \ U \in \mathcal{U} \ s.t. \ \mathbf{v}^\top H(f)(\mathbf{x})\mathbf{v} \geq \lambda \ \forall \mathbf{x} \in U} \mu_d(U) ,$$

*where $\mathbb{S}^{d-1} = \{\mathbf{x} : \|\mathbf{x}\|_2 = 1\}$ is the d-dimensional unit hypersphere, and $\mathcal{U}$ is the set of all connected and measurable subsets of $[0,1]^d$.*

In words, $\sigma_\lambda(f)$ is the measure (w.r.t. the uniform distribution on $[0,1]^d$) of the largest connected set in the domain of $f$, where at any point, $f$ has curvature at least $\lambda$ along some fixed direction $\mathbf{v}$. The "prototypical" functions $f$ we are interested in is when $\sigma_\lambda(f)$ is lower bounded by a constant (e.g. it is 1 if $f$ is strongly convex). We stress that our results in this section will hold equally well by considering the condition $\mathbf{v}^\top H(f)(\mathbf{x})\mathbf{v} \leq -\lambda$ as well, however for the sake of simplicity we focus on the former condition appearing in Def. 1. Our goal is to show a depth separation result *individually* for any such function (that is, for any such function, there is a gap in the attainable error between deeper and shallower networks, even if the shallow network is considerably larger).

As usual, we start with an inapproximability result. Specifically, we prove the following lower bound on the attainable approximation error of $f$, using a ReLU neural network of a given depth and width:

**Theorem 4.** *For any $C^2$ function $f : [0,1]^d \to \mathbb{R}$, any $\lambda > 0$, and any function $g$ on $[0,1]^d$ expressible by a ReLU network of depth $l$ and maximal width $m$, it holds that*

$$\int_{[0,1]^d} (f(\mathbf{x}) - g(\mathbf{x})^2 \mu_d(\mathbf{x}) \, d\mathbf{x} \geq \frac{c \cdot \lambda^2 \cdot \sigma_\lambda^5}{(2m)^{4l}} ,$$

*where $c > 0$ is a universal constant.*

The theorem conveys a key tradeoff between depth and width when approximating a $C^2$ function using ReLU networks: The error cannot decay faster than polynomially in the width $m$, yet the bound deteriorates exponentially in the depth $l$. As we show later on, this deterioration does not stem from the looseness in the bound: For well-behaved $f$, it is indeed possible to construct ReLU networks, where the approximation error decays exponentially with depth.

The proof of Thm. 4 appears in Subsection 6.3, and is based on a series of intermediate results. First, we show that any strictly curved function (in a sense similar to Definition 1) cannot be well-approximated in an $L_2$ sense by piecewise linear functions, unless the number of linear regions is large. To that end, we first establish some necessary tools based on Legendre polynomials. We then prove a result specific to the one-dimensional case, including an explicit lower bound if the target function is quadratic (Thm. 9) or strongly convex or concave (Thm. 10). We then expand the construction to get an error lower bound
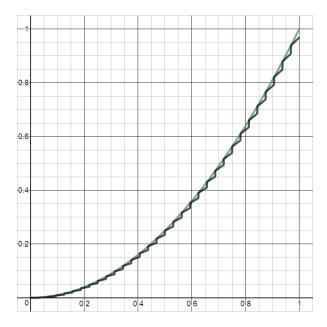
9

**Figure 2:** ReLU approximation of the function $x \mapsto x^2$ obtained by extracting 5 bits. The number of linear segments grows exponentially with the number of bits and the approximating network size.

in general dimension $d$, depending on the number of linear regions in the approximating piecewise-linear function. Finally, we note that any ReLU network induces a piecewise-linear function, and bound the number of linear regions induced by a ReLU network of a given width and depth (using a lemma borrowed from Telgarsky (2016)). Combining this with the previous lower bound yields Thm. 4.

We now turn to complement this lower bound with an approximability result, showing that with more depth, a wide family of functions to which Thm. 4 applies *can* be approximated with exponentially high accuracy. Specifically, we consider functions which can be approximated using a moderate number of multiplications and additions, where the values of intermediate computations are bounded (for example, a special case is any function approximable by a moderately-sized Boolean circuit, or a polynomial).

The key result to show this is the following, which implies that the multiplication of two (bounded-size) numbers can be approximated by a ReLU network, with error decaying exponentially with depth:

**Theorem 5.** *Let $f : [-M, M]^2 \to \mathbb{R}$, $f(x, y) = x \cdot y$ and let $\epsilon > 0$ be arbitrary. Then exists a ReLU neural network $g$ of width $4 \left\lceil \log\left(\frac{M}{\epsilon}\right) \right\rceil + 13$ and depth $\left\lceil 2 \log\left(\frac{M}{\epsilon}\right) \right\rceil + 9$ satisfying*

$$\sup_{(x,y)\in[-M,M]^2} |f(x,y) - g(x,y)| \leq \epsilon.$$

The idea of the construction is that depth allows us to compute highly-oscillating functions, which can extract high-order bits from the binary representation of the inputs. Given these bits, one can compute the product by a procedure resembling long multiplication, as shown in Fig. 2, and formally proven as follows:

*Proof of Thm. 5.* We begin by observing that by using a simple linear change of variables on $x$, we may assume without loss of generality that $x \in [0, 1]$, as we can just rescale $x$ to the interval $[0, 1]$, and then map it back to its original domain $[-M, M]$, where the error will multiply by a factor of $2M$. Then by requiring accuracy $\frac{\epsilon}{2M}$ instead of $\epsilon$, the result will follow.

10

The key behind the proof is that performing bit-wise operations on the first $k$ bits of $x \in [0, 1]$ yields an estimation of the product to accuracy $2^{1-k} M$. Let $x = \sum_{i=1}^{\infty} 2^{-i} x_i$ be the binary representation of $x$ where $x_i$ is the $i^{\text{th}}$ bit of $x$, then

$$
\begin{aligned}
x \cdot y &= \sum_{i=1}^{\infty} 2^{-i} x_i \cdot y \\
&= \sum_{i=1}^{k} 2^{-i} x_i \cdot y + \sum_{i=k+1}^{\infty} 2^{-i} x_i \cdot y.
\end{aligned}
\tag{5}
$$

But since

$$
\left| \sum_{i=k+1}^{\infty} 2^{-i} x_i \cdot y \right| \leq \left| \sum_{i=k+1}^{\infty} 2^{-i} \cdot y \right| = 2^{-k} |y| \leq 2^{1-k} M,
$$

Eq. (5) implies

$$
\left| x \cdot y - \sum_{i=1}^{k} 2^{-i} x_i \cdot y \right| \leq 2^{1-k} M.
$$

Requiring that $2^{2-k} M \leq \frac{\epsilon}{2M}$, it suffices to show the existence of a network which approximates the function $\sum_{i=1}^{k} 2^{-i} x_i \cdot y$ to accuracy $\frac{\epsilon}{2}$, where $k = 2 \left\lceil \log \left( \frac{8M}{\epsilon} \right) \right\rceil$. This way both approximations will be at most $\frac{\epsilon}{2}$, resulting in the desired accuracy of $\epsilon$.

Before specifying the architecture which extracts the $i^{\text{th}}$ bit of $x$, we first describe the last 2 layers of the network. Let the penultimate layer comprise of $k$ neurons, each receiving both $y$ and $x_i$ as input, and having the set of weights $(2^{-i}, 1, -1)$. Thus, the output of the $i^{\text{th}}$ neuron in the penultimate layer is

$$
\left[ 2^{-i} y + x_i - 1 \right]_{+} = 2^{-i} x_i y.
$$

Let the final single output neuron have the set of weights $(1, \dots, 1, 0) \in \mathbb{R}^{k+1}$, this way, the output of the network will be $\sum_{i=1}^{k} 2^{-i} x_i \cdot y$ as required.

We now specify the architecture which extracts the first most significant $k$ bits of $x$. In Telgarsky (2016), the author demonstrates how the composition of the function

$$
\varphi(x) = [2x]_{+} - [4x - 2]_{+}
$$

with itself $i$ times, $\varphi^i$, yields a highly oscillatory triangle wave function in the domain $[0, 1]$. Furthermore, we observe that $\varphi(x) = 0 \ \forall x \leq 0$, and thus $\varphi^i(x) = 0 \ \forall x \leq 0$. Now, a linear shift of the input of $\varphi^i$ by $2^{-i-1}$, and composing the output with

$$
\sigma_\delta(x) = \left[ \frac{1}{2\delta} x - \frac{1}{4\delta} + \frac{1}{2} \right]_{+} - \left[ \frac{1}{2\delta} x - \frac{1}{4\delta} - \frac{1}{2} \right]_{+},
$$

which converges to $\mathbb{1}_{[x \geq 0.5]}(x)$ as $\delta \to 0$, results in an approximation of $x \mapsto x_i$:

$$
\sigma_\delta \left( \varphi^i \left( x - 2^{-i-1} \right) \right).
$$

We stress that choosing $\delta$ such that the network approximates the bit-wise product to accuracy $\frac{\epsilon}{2}$ will require $\delta$ to be of magnitude $\frac{1}{\epsilon}$, but this poses no problem as representing such a number requires $\log \left( \frac{1}{\epsilon} \right)$ bits, which is also the magnitude of the size of the network, as suggested by the following analysis.

Next, we compute the size of the network required to implement the above approximation. To compute $\varphi$ only two neurons are required, therefore $\varphi^i$ can be computed using $i$ layers with 2 neurons in each, and finally composing this with $\sigma_\delta$ requires a subsequent layer with 2 more neurons. To implement the $i^{\text{th}}$ bit extractor we therefore require a network of size $2 \times (i+1)$. Using dummy neurons to propagate the $i^{\text{th}}$ bit for $i < k$, the architecture extracting the $k$ most significant bits of $x$ will be of size $2k \times (k+1)$. Adding the final component performing the multiplication estimation will require 2 more layers of width $k$ and 1 respectively, and an increase of the width by 1 to propagate $y$ to the penultimate layer, resulting in a network of size $(2k+1) \times (k+1)$. $\qquad\square$

Thm. 5 shows that multiplication can be performed very accurately by deep networks. Moreover, additions can be computed by ReLU networks exactly, using only a single layer with 4 neurons: Let $\alpha, \beta \in \mathbb{R}$ be arbitrary, then $(x, y) \mapsto \alpha \cdot x + \beta \cdot y$ is given in terms of ReLU summation by

$$\alpha \left[x\right]_+ - \alpha \left[-x\right]_+ + \beta \left[y\right]_+ - \beta \left[-y\right]_+ .$$

Repeating these arguments, we see that any function which can be approximated by a bounded number of operations involving additions and multiplications, can also be approximated well by moderately-sized networks. This is formalized in the following theorem, which provides an approximation error upper bound (in the $L_\infty$ sense, which is stronger than $L_2$ for upper bounds):

**Theorem 6.** *Let $\mathcal{F}_{t,M,\epsilon}$ be the family of functions on the domain $[0,1]^d$ with the property that $f \in \mathcal{F}_{t,M,\epsilon}$ is approximable to accuracy $\epsilon$ with respect to the infinity norm, using at most $t$ operations involving weighted addition, $(x, y) \mapsto \alpha \cdot x + \beta \cdot y$, where $\alpha, \beta \in \mathbb{R}$ are fixed; and multiplication, $(x, y) \mapsto x \cdot y$, where each intermediate computation stage is bounded in the interval $[-M, M]$. Then there exists a universal constant $c$, and a ReLU network $g$ of width and depth at most $c \left(t \log \left(\frac{1}{\epsilon}\right) + t^2 \log (M)\right)$, such that*

$$\sup_{\mathbf{x} \in [0,1]^d} \left|f\left(\mathbf{x}\right) - g\left(\mathbf{x}\right)\right| \leq 2\epsilon.$$

As discussed in Sec. 2, this type of $L_\infty$ approximation bound implies an $L_2$ approximation bound with respect to any distribution. The proof of the theorem appears in Subsection 6.4.

Combining Thm. 4 and Thm. 6, we can state the following corollary, which formally shows how depth can be exponentially more valuable than width as a function of the target accuracy $\epsilon$:

**Corollary 1.** *Suppose $f \in C^2 \cap \mathcal{F}_{t(\epsilon), M(\epsilon), \epsilon}$, where $t\left(\epsilon\right) = \mathcal{O}\left(\text{poly}\left(\log\left(1/\epsilon\right)\right)\right)$ and $M\left(\epsilon\right) = \mathcal{O}\left(\text{poly}\left(1/\epsilon\right)\right)$. Then approximating $f$ to accuracy $\epsilon$ in the $L_2$ norm using a fixed depth ReLU network requires width at least $\text{poly}(1/\epsilon)$, whereas there exists a ReLU network of depth and width at most $p\left(\log\left(1/\epsilon\right)\right)$ which approximates $f$ to accuracy $\epsilon$ in the infinity norm, where $p$ is a polynomial depending solely on $f$.*

*Proof.* The lower bound follows immediately from Thm. 4. For the upper bound, observe that Thm. 6 implies an $\epsilon$ approximation by a network of width and depth at most

$$c \left(t\left(\epsilon/2\right) \log\left(2/\epsilon\right) + \left(t\left(\epsilon/2\right)\right)^2 \log\left(M\left(\epsilon/2\right)\right)\right),$$

which by the assumption of Corollary 1, can be bounded by

$$p\left(\log\left(1/\epsilon\right)\right)$$

for some polynomial $p$ which depends solely on $f$. $\qquad\square$

## Acknowledgements

# References

Bianchini, M. and Scarselli, F. On the complexity of shallow and deep neural network classifiers. In *ESANN*, 2014.

Cohen, Nadav, Sharir, Or, and Shashua, Amnon. On the expressive power of deep learning: A tensor analysis. In *29th Annual Conference on Learning Theory*, pp. 698–728, 2016.

Cybenko, George. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Delalleau, O. and Bengio, Y. Shallow vs. deep sum-product networks. In *NIPS*, pp. 666–674, 2011.

Eldan, Ronen and Shamir, Ohad. The power of depth for feedforward neural networks. In *29th Annual Conference on Learning Theory*, pp. 907–940, 2016.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.

Hornik, Kurt. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

Liang, Shiyu and Srikant, R. Why deep neural networks? *arXiv preprint arXiv:1610.04161*, 2016.

Martens, J. and Medabalimi, V. On the expressive efficiency of sum product networks. *arXiv preprint arXiv:1411.7717*, 2014.

Poggio, Tomaso, Mhaskar, Hrushikesh, Rosasco, Lorenzo, Miranda, Brando, and Liao, Qianli. Why and when can deep–but not shallow–networks avoid the curse of dimensionality: a review. *arXiv preprint arXiv:1611.00740*, 2016.

Poole, Ben, Lahiri, Subhaneil, Raghu, Maithreyi, Sohl-Dickstein, Jascha, and Ganguli, Surya. Exponential expressivity in deep neural networks through transient chaos. In *Advances In Neural Information Processing Systems*, pp. 3360–3368, 2016.

Shaham, Uri, Cloninger, Alexander, and Coifman, Ronald R. Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis*, 2016.

Telgarsky, Matus. Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485*, 2016.

Yarotsky, Dmitry. Error bounds for approximations with deep relu networks. *arXiv preprint arXiv:1610.01145*, 2016.

# 6 Proofs

## 6.1 Proof of Thm. 2

This proof bears resemblance to the proof provided in Eldan & Shamir (2016)[Lemma 10], albeit once approximating $\|\mathbf{x}\|_2^2$, the following construction takes a slightly different route. For completeness, we also state assumption 1 from Eldan & Shamir (2016):

**Assumption 1.** *Given the activation function $\sigma$, there is a constant $c_\sigma \geq 1$ (depending only on $\sigma$) such that the following holds: For any $L$-Lipschitz function $f : \mathbb{R} \to \mathbb{R}$ which is constant outside a bounded interval $[-R, R]$, and for any $\delta$, there exist scalars $a, \{\alpha_i, \beta_i, \gamma_i\}_{i=1}^w$, where $w \leq c_\sigma \frac{RL}{\delta}$, such that the function*

$$h(x) = a + \sum_{i=1}^w \alpha_i \cdot \sigma(\beta_i x - \gamma_i)$$

*satisfies*

$$\sup_{x \in \mathbb{R}} |f(x) - h(x)| \leq \delta.$$

As discussed in Eldan & Shamir (2016), this assumption is satisfied by ReLU, sigmoid, threshold, and more generally all standard activation functions we are familiar with.

*Proof.* Consider the $4$-Lipschitz function

$$l(x) = \min\left\{x^2, 4\right\},$$

which is constant outside $[-2, 2]$, as well as the function

$$\ell(x) = \sum_{i=1}^d l(x_i) = \sum_{i=1}^d \min\left\{x_i^2, 4\right\}$$

on $\mathbb{R}^d$. Applying assumption 1, we obtain a function $\tilde{l}(x)$ having the form $a + \sum_{i=1}^w \alpha_i \sigma(\beta_i x - \gamma_i)$ so that

$$\sup_{x \in \mathbb{R}} \left|\tilde{l}(x) - l(x)\right| \leq \frac{\delta}{d},$$

and where the width parameter $w$ is at most $\frac{8c_\sigma d}{\delta}$. Consequently, the function

$$\tilde{\ell}(\mathbf{x}) = \sum_{i=1}^d \tilde{l}(x_i)$$

can be expressed in the form $a + \sum_{i=1}^w \alpha_i \sigma(\beta_i x - \gamma_i)$ where $w \leq \frac{8c_\sigma d^2}{\delta}$, yielding an approximation satisfying

$$\sup_{\mathbf{x} \in \mathbb{R}^d} \left|\tilde{\ell}(\mathbf{x}) - \ell(\mathbf{x})\right| \leq \delta.$$

We now invoke assumption 1 again to approximate the $1$-Lipschitz function

$$f(x) = \begin{cases} 0 & x < -0.5 \\ x + 0.5 & x \in [-0.5, 0.5] \\ 1 & x > 0.5 \end{cases}$$

14

and obtain an approximation $\tilde{f}(x) = \tilde{a} + \sum_{i=1}^{\tilde{w}} \tilde{\alpha}_i \sigma\left(\tilde{\beta}_i x - \tilde{\gamma}_i\right)$ satisfying

$$\sup_{x \in \mathbb{R}} \left|\tilde{f}(x) - f(x)\right| \leq \sqrt{\frac{\delta}{2}} \tag{6}$$

where $\tilde{w} \leq c_\sigma \sqrt{1/2\delta}$.

Now consider the composition $\tilde{f} \circ \left(c_\mu \cdot \tilde{\ell} - c_\mu\right)$, where $c_\mu > 0$ is to be determined later. This composition has the form

$$a + \sum_{i=1}^{w} u_i \sigma\left(\sum_{j=1}^{w} v_{i,j} \sigma\left(\langle \mathbf{w}_{i,j}, \mathbf{x}\rangle + b_{i,j}\right) + c_i\right)$$

for appropriate scalars $a$, $u_i$, $c_i$, $v_{i,j}$, $b_{i,j}$ and vectors $\mathbf{w}_{i,j}$, and where $w$ is at most $\max\left\{8c_\sigma d^2/\delta, c_\sigma \sqrt{1/2\delta}\right\}$. It is now left to bound the approximation error obtained by $\tilde{f} \circ \left(c_\mu \cdot \tilde{\ell} - c_\mu\right)$. Define for any $\epsilon > 0$,

$$R_\epsilon = \left\{\mathbf{x} \in \mathbb{R}^d : 1 - \epsilon \leq \|\mathbf{x}\|_2^2 \leq 1 + \epsilon\right\}.$$

Since $\mu$ is continuous, there exists $\epsilon > 0$ such that

$$\int_{R_\epsilon} \mu(\mathbf{x})\, d\mathbf{x} \leq \frac{\delta}{4}.$$

Now, for any $\mathbf{x} \in \mathbb{R}^d$ such that $1 + \epsilon \leq \|\mathbf{x}\|_2^2$ we have

$$\tilde{\ell}(\mathbf{x}) \geq \ell(\mathbf{x}) - \delta = \min\left\{\|\mathbf{x}\|_2^2, 4\right\} - \delta \geq 1 + \epsilon - \delta.$$

Assuming $\delta < \epsilon/2$, we have the above is at least

$$1 + \epsilon/2.$$

Taking $c_\mu = 1/\epsilon$, we get

$$c_\mu \cdot \tilde{\ell}(\mathbf{x}) - c_\mu = c_\mu \cdot \left(\tilde{\ell}(\mathbf{x}) - 1\right) \geq \frac{c_\mu \epsilon}{2} = 0.5,$$

and thus

$$\tilde{f}\left(c_\mu \cdot \tilde{\ell}(\mathbf{x}) - c_\mu\right) \in \left(1 - \sqrt{\frac{\delta}{2}}, 1 + \sqrt{\frac{\delta}{2}}\right), \tag{7}$$

For any $\mathbf{x} \in \mathbb{R}^d$ satisfying $1 + \epsilon \leq \|\mathbf{x}\|_2^2$. A similar argument shows that for any $\mathbf{x} \in \mathbb{R}$ satisfying $\|\mathbf{x}\|_2^2 \leq 1 - \epsilon$ we have

$$\tilde{f}\left(c_\mu \cdot \tilde{\ell}(\mathbf{x}) - c_\mu\right) \in \left(-\sqrt{\frac{\delta}{2}}, \sqrt{\frac{\delta}{2}}\right). \tag{8}$$

Combining both Eq. (7) and Eq. (8) we obtain

$$\int_{\mathbb{R}^d} \left( \tilde{f}\left(c_\mu \cdot \tilde{\ell}(\mathbf{x}) - c_\mu\right) - \mathbb{1}\left(\|\mathbf{x}\|_2 \le 1\right)\right)^2 \mu(\mathbf{x})\, d\mathbf{x}$$

$$= \int_{R_\epsilon} \left( \tilde{f}\left(c_\mu \cdot \tilde{\ell}(\mathbf{x}) - c_\mu\right) - \mathbb{1}\left(\|\mathbf{x}\|_2 \le 1\right)\right)^2 \mu(\mathbf{x})\, d\mathbf{x}$$

$$\qquad + \int_{\mathbb{R}^d \setminus R_\epsilon} \left( \tilde{f}\left(c_\mu \cdot \tilde{\ell}(\mathbf{x}) - c_\mu\right) - \mathbb{1}\left(\|\mathbf{x}\|_2 \le 1\right)\right)^2 \mu(\mathbf{x})\, d\mathbf{x}$$

$$\le \int_{R_\epsilon} 2\mu(\mathbf{x})\, d\mathbf{x} + \int_{\mathbb{R}^d \setminus R_\epsilon} \frac{\delta}{4}\mu(\mathbf{x})\, d\mathbf{x}$$

$$\le \frac{\delta}{2} + \frac{\delta}{2} = \delta,$$

Where the first summand in the penultimate inequality is justified due to $\tilde{f}$ being bounded in the interval $\left[-\sqrt{\delta/2}, 1 + \sqrt{\delta/2}\right]$ by Eq. (6), and assuming $1 + \sqrt{\delta/2} \le \sqrt{2}$, and the second summand justified due to Equations (7) and (8), concluding the proof of the lemma. $\qquad\square$

## 6.2   Proof of Thm. 3

Consider an input distribution of the form

$$\mathbf{x} = sr\mathbf{v},$$

where $\mathbf{v}$ is drawn from a certain distribution on the unit $L_1$ sphere $\{\mathbf{x} : \|\mathbf{x}\|_1 = 1\}$ to be specified later, and $s$ is uniformly distributed on $[1, 1 + \epsilon]$.

Let

$$F(\mathbf{x}) = \sum_{j=1}^{N} a_j \left[\langle \mathbf{w}_j, \mathbf{x}\rangle + b_j\right]_+$$

be a 2-layer ReLU network of width $N$, such that with respect to the distribution above,

$$\mathbb{E}_{\mathbf{x}}\left[(f(\|\mathbf{x}\|_1) - F(\mathbf{x}))^2\right] \;=\; \mathbb{E}_{\mathbf{v}}\left[\mathbb{E}_s\left[(f(sr) - F(sr\mathbf{v}))^2 \big| \mathbf{v}\right]\right] \;\le\; \frac{\delta_\epsilon}{2}\,.$$

By Markov's inequality, this implies

$$\mathrm{Pr}_{\mathbf{v}}\left(\mathbb{E}_s\left[f(sr) - F(sr\mathbf{v})^2 \big| \mathbf{v}\right] \le \delta_\epsilon\right) \;\ge\; \frac{1}{2}\,.$$

By the assumption on $f$, and the fact that $s$ is uniform on $[1, 1+\epsilon]$, we have that $\mathbb{E}_s\left[f(sr) - F(sr\mathbf{v})^2 \big| \mathbf{v}\right] \le \delta_\epsilon$ only if $\tilde{f}_N$ is not a linear function on the line between $r\mathbf{v}$ and $(1+\epsilon)r\mathbf{v}$. In other words, this line must be crossed by the hyperplane $\{\mathbf{x} : \langle \mathbf{w}_j, \mathbf{x}\rangle + b_j = 0\}$ for some neuron $j$. Thus, we must have

$$\mathrm{Pr}_{\mathbf{v}}\left(\exists j \in \{1, \ldots, N\}, s \in [1, 1 + \epsilon] \text{ s.t. } \langle \mathbf{w}_j, sr\mathbf{v}\rangle + b_j = 0\right) \;\ge\; \frac{1}{2}. \qquad (9)$$

The left hand side equals

$$\Pr_{\mathbf{v}}\left(\exists j \in \{1 \ldots N\}, s \in [1, 1+\epsilon] \text{ s.t. } \langle \mathbf{w}_j, \mathbf{v} \rangle = -b_j/sr\right)$$

$$= \Pr_{\mathbf{v}}\left(\exists j \in \{1 \ldots N\} \text{ s.t. } \langle \mathbf{w}_j, \mathbf{v} \rangle \text{ between } -\frac{b_j}{r} \text{ and } -\frac{b_j}{(1+\epsilon)r}\right)$$

$$\leq \Pr_{\mathbf{v}}\left(\exists j \in \{1 \ldots N\} \text{ s.t. } \langle \mathbf{w}_j, \mathbf{v} \rangle \text{ betwen } -\frac{b_j}{r} \text{ and } -(1-\epsilon)\frac{b_j}{r}\right)$$

$$\leq \sum_{j=1}^{N} \Pr_{\mathbf{v}}\left(\langle \mathbf{w}_j, \mathbf{v} \rangle \text{ between } -\frac{b_j}{r} \text{ and } -(1-\epsilon)\frac{b_j}{r}\right)$$

$$\leq N \cdot \sup_{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}} \Pr_{\mathbf{v}}\left(\langle \mathbf{w}, \mathbf{v} \rangle \text{ between } -\frac{b}{r} \text{ and } -(1-\epsilon)\frac{b}{r}\right)$$

$$= N \cdot \sup_{\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}} \Pr_{\mathbf{v}}\left(\left\langle \frac{-r\mathbf{w}}{b}, \mathbf{v} \right\rangle \in [1-\epsilon, 1]\right) \;\; = \;\; N \cdot \sup_{\mathbf{w} \in \mathbb{R}^d} \Pr_{\mathbf{v}}\left(\langle \mathbf{w}, \mathbf{v} \rangle \in [1-\epsilon, 1]\right),$$

where in the first inequality we used the fact that $\frac{1}{1+\epsilon} \geq 1 - \epsilon$ for all $\epsilon \in (0,1)$, and in the second inequality we used a union bound. Combining these inequalities with Eq. (9), we get that

$$N \geq \frac{1}{\sup_{\mathbf{w}} \Pr_{\mathbf{v}}\left(\langle \mathbf{w}, \mathbf{v} \rangle \in [1-\epsilon, 1]\right)} \; .$$

As a result, to prove the theorem, it is enough to construct a distribution for $\mathbf{v}$ on the on the unit $L_1$ ball, such that for any $\mathbf{w}$,

$$\Pr_{\mathbf{v}}\left(\langle \mathbf{w}, \mathbf{v} \rangle \in [1-\epsilon, 1]\right) \leq \tilde{\mathcal{O}}(\epsilon + \exp(-\Omega(d))) \tag{10}$$

By the inequality above, we would then get that $N = \tilde{\Omega}(\min\{1/\epsilon, \exp(\Omega(d))\})$.

Specifically, consider a distribution over $\mathbf{v}$ defined as follows: First, we sample $\boldsymbol{\sigma} \in \{-1, +1\}^d$ uniformly at random, and $\mathbf{n} \in \mathbb{R}^d$ from a standard Gaussian distribution, and define

$$\hat{\mathbf{v}} = \frac{1}{d}\left(\boldsymbol{\sigma} + c_d\left(I - \frac{1}{d}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\right)\mathbf{n}\right),$$

where $c_d > 0$ is a parameter dependent on $d$ to be determined later. It is easily verified that $\langle \boldsymbol{\sigma}/d, \hat{\mathbf{v}} \rangle = \langle \boldsymbol{\sigma}/d, \boldsymbol{\sigma}/d \rangle$ independent of $\mathbf{n}$, hence $\hat{\mathbf{v}}$ lies on the hyperplane containing the facet of the $L_1$ ball on which $\boldsymbol{\sigma}/d$ resides. Calling this facet $F_{\boldsymbol{\sigma}}$, we define $\mathbf{v}$ to have the same distribution as $\hat{\mathbf{v}}$, conditioned on $\hat{\mathbf{v}} \in F_{\boldsymbol{\sigma}}$.

We begin by arguing that

$$\Pr_{\mathbf{v}}\left(\langle \mathbf{w}, \mathbf{v} \rangle \in [1-\epsilon, 1]\right) \leq 2 \cdot \Pr_{\hat{\mathbf{v}}}\left(\langle \mathbf{w}, \hat{\mathbf{v}} \rangle \in [1-\epsilon, 1]\right). \tag{11}$$

To see this, let $A = \{\mathbf{x} : \langle \mathbf{w}, \mathbf{x} \rangle \in [1-\epsilon, 1]\}$, and note that the left hand side equals

$$\Pr(\mathbf{v} \in A) = \mathbb{E}_{\boldsymbol{\sigma}}[\Pr(\mathbf{v} \in A | \boldsymbol{\sigma})] = \mathbb{E}_{\boldsymbol{\sigma}}\left[\Pr\left(\hat{\mathbf{v}} \in A | \boldsymbol{\sigma}, \hat{\mathbf{v}} \in F_{\boldsymbol{\sigma}}\right)\right]$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\frac{\Pr(\hat{\mathbf{v}} \in A \cap F_{\boldsymbol{\sigma}} | \boldsymbol{\sigma})}{\Pr(\hat{\mathbf{v}} \in F_{\boldsymbol{\sigma}} | \boldsymbol{\sigma})}\right] \leq \frac{1}{\min_{\boldsymbol{\sigma}} \Pr(\hat{\mathbf{v}} \in F_{\boldsymbol{\sigma}} | \boldsymbol{\sigma})} \mathbb{E}_{\boldsymbol{\sigma}}\left[\Pr(\hat{\mathbf{v}} \in A | \boldsymbol{\sigma})\right]$$

$$= \frac{\Pr(\hat{\mathbf{v}} \in A)}{\min_{\boldsymbol{\sigma}} \Pr(\hat{\mathbf{v}} \in F_{\boldsymbol{\sigma}} | \boldsymbol{\sigma})}.$$

Therefore, to prove Eq. (11), it is enough to prove that $\Pr(\hat{\mathbf{v}} \in F_{\boldsymbol{\sigma}}|\boldsymbol{\sigma}) \geq 1/2$ for any $\boldsymbol{\sigma}$. As shown earlier, $\hat{\mathbf{v}}$ lies on the hyperplane containing $F_{\boldsymbol{\sigma}}$, the facet of the $L_1$ ball in which $\boldsymbol{\sigma}/d$ resides. Thus, $\hat{\mathbf{v}}$ can be outside $F_{\boldsymbol{\sigma}}$, only if at least one of its coordinates has a different sign than $\boldsymbol{\sigma}$. By definition of $\hat{\mathbf{v}}$, this can only happen if $\left\| c_d (I - \boldsymbol{\sigma}\boldsymbol{\sigma}^\top/d)\mathbf{n} \right\|_\infty \geq 1$. The probability of this event (over the random draw of $\mathbf{n}$) equals

$$\Pr\left(\max_{j\in\{1...d\}} \left| c_d \left( n_j - \frac{1}{d} \langle \boldsymbol{\sigma}, \mathbf{n} \rangle \sigma_j \right) \right| \geq 1 \right) \;=\; \Pr\left(\max_{j\in\{1...d\}} \left| n_j - \sigma_j \cdot \frac{1}{d} \sum_{i=1}^{d} \sigma_i n_i \right| \geq \frac{1}{c_d} \right).$$

Since $\sigma_i \in \{-1, 1\}$ for all $i$, the event on the right hand side can only occur if $|n_j| \geq 1/2c_d$ for some $j$. Recalling that each $n_j$ has a standard Gaussian distribution, this probability can be upper bounded by

$$\Pr\left(\max_{j\in\{1...d\}} |n_j| \geq \frac{1}{2c_d}\right) \;\leq\; d \cdot \Pr\left(|n_1| \geq \frac{1}{2c_d}\right) \;=\; 2d \cdot \Pr\left(n_1 \geq \frac{1}{2c_d}\right) \;\leq\; 2d \cdot \exp\left(-\frac{1}{4c_d^2}\right),$$

where we used a union bound and a standard Gaussian tail bound. Thus, by picking

$$c_d = \sqrt{\frac{1}{4\log(4d)}},$$

we can ensure that the probability is at most $1/2$, hence proving that $\Pr(\hat{\mathbf{v}} \in F_{\boldsymbol{\sigma}}|\boldsymbol{\sigma}) \geq 1/2$ and validating Eq. (11).

With Eq. (11) in hand, we now turn to upper bound

$$\Pr\left(\langle \mathbf{w}, \hat{\mathbf{v}} \rangle \in [1 - \epsilon, 1]\right) \;=\; \Pr\left(\frac{1}{d}\left\langle \mathbf{w}, \boldsymbol{\sigma} + c_d\left(I - \frac{1}{d}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\right)\mathbf{n} \right\rangle \in [1 - \epsilon, 1]\right).$$

By the equation above, we have that conditioned on $\boldsymbol{\sigma}$, the distribution of $\langle \mathbf{w}, \hat{\mathbf{v}} \rangle$ is Gaussian with mean $\langle \boldsymbol{\sigma}, \mathbf{w} \rangle /d$ and variance

$$\frac{c_d^2}{d^2} \cdot \mathbf{w}^\top \left(I - \frac{1}{d}\boldsymbol{\sigma}\boldsymbol{\sigma}^\top\right)^2 \mathbf{w} \;=\; \frac{c_d^2}{d^2} \cdot \left(\|\mathbf{w}\|^2 - \frac{\langle \mathbf{w}, \boldsymbol{\sigma} \rangle^2}{d}\right) \;=\; \left(\frac{c_d\|\mathbf{w}\|}{d}\right)^2 \cdot \left(1 - \frac{1}{d}\left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, \boldsymbol{\sigma} \right\rangle^2\right).$$

By Hoeffding's inequality, we have that for any $t > 0$,

$$\Pr_{\boldsymbol{\sigma}}\left(\left|\frac{\langle \boldsymbol{\sigma}, \mathbf{w} \rangle}{d}\right| > t \cdot \frac{\|\mathbf{w}\|}{d}\right) \;\leq\; 2\exp(-2t^2)$$

and

$$\Pr_{\boldsymbol{\sigma}}\left(\left|\left\langle \frac{\mathbf{w}}{\|\mathbf{w}\|}, \boldsymbol{\sigma} \right\rangle\right| > \sqrt{\frac{d}{2}}\right) \;\leq\; 2\exp(-d).$$

This means that with probability at least $1 - 2\exp(-d) - 2\exp(-2t^2)$ over the choice of $\boldsymbol{\sigma}$, the distribution of $\langle \mathbf{w}, \hat{\mathbf{v}} \rangle$ (conditioned on $\boldsymbol{\sigma}$) is Gaussian with mean bounded in absolute value by $t\|\mathbf{w}\|/d$, and variance of at least $\left(\frac{c_d\|\mathbf{w}\|}{d}\right)^2 \cdot \left(1 - \frac{1}{d} \cdot \frac{d}{2}\right) = \frac{1}{2}\left(\frac{c_d\|\mathbf{w}\|}{d}\right)^2$. To continue, we utilize the following lemma:

**Lemma 1.** *Let $n$ be a Gaussian random variable on $\mathbb{R}$ with mean $\mu$ and variance $v^2$ for some $v > 0$. Then for any $\epsilon \in (0, 1)$,*

$$\Pr\left(n \in [1 - \epsilon, 1]\right) \;\leq\; \sqrt{\frac{2}{\pi}} \cdot \max\left\{1, \frac{|\mu|}{v}\right\} \cdot \frac{\epsilon}{1 - \epsilon}.$$

*Proof.* Since the probability can only increase if we replace the mean $\mu$ by $|\mu|$, we will assume without loss of generality that $\mu \geq 0$.

By definition of a Gaussian distribution, and using the easily-verified fact that $\exp(-z^2) \leq \min\{1, 1/z\}$ for all $z \geq 0$, the probability equals

$$
\frac{1}{\sqrt{2\pi v^2}} \int_{1-\epsilon}^{1} \exp\left(-\frac{(x-\mu)^2}{v^2}\right) dx \leq \frac{\epsilon}{\sqrt{2\pi v^2}} \cdot \max_{x \in [1-\epsilon,1]} \exp\left(-\frac{(x-\mu)^2}{v^2}\right)
$$

$$
\leq \frac{\epsilon}{\sqrt{2\pi v^2}} \cdot \max_{x \in [1-\epsilon,1]} \min\left\{1, \frac{v}{|x-\mu|}\right\} = \frac{\epsilon}{\sqrt{2\pi}} \cdot \max_{x \in [1-\epsilon,1]} \min\left\{\frac{1}{v}, \frac{1}{|x-\mu|}\right\}
$$

$$
= \frac{\epsilon}{\sqrt{2\pi}} \cdot \max_{x \in [1-\epsilon,1]} \frac{1}{\max\{v, |x-\mu|\}} = \frac{\epsilon}{\sqrt{2\pi}} \cdot \max_{x \in [1-\epsilon,1]} \frac{\max\{1, \frac{\mu}{v}\}}{\max\{1, \frac{\mu}{v}\} \cdot \max\{v, |x-\mu|\}}
$$

$$
\leq \frac{\epsilon}{\sqrt{2\pi}} \cdot \max_{x \in [1-\epsilon,1]} \frac{\max\{1, \frac{\mu}{v}\}}{\max\{\mu, |x-\mu|\}} = \frac{\epsilon}{\sqrt{2\pi}} \cdot \frac{\max\{1, \frac{\mu}{v}\}}{\max\{\mu, \min_{x \in [1-\epsilon,1]} |x-\mu|\}}
$$

A simple case analysis reveals that the denominator is at least[2] $\frac{1-\epsilon}{2}$, from which the result follows. $\qquad\square$

Using this lemma and the previous observations, we get that with probability at least $1 - 2\exp(-d) - 2\exp(-2t^2)$ over the choice of $\boldsymbol{\sigma}$,

$$
\Pr(\langle \mathbf{w}, \hat{\mathbf{v}} \rangle \in [1-\epsilon, 1] | \boldsymbol{\sigma}) \leq \sqrt{\frac{2}{\pi}} \cdot \max\left\{1, \frac{t \|\mathbf{w}\| / d}{c_d \|\mathbf{w}\| / \sqrt{2d}}\right\} \cdot \frac{\epsilon}{1-\epsilon}
$$

$$
= \sqrt{\frac{2}{\pi}} \cdot \max\left\{1, \frac{t}{c_d\sqrt{2}}\right\} \cdot \frac{\epsilon}{1-\epsilon}.
$$

Letting $E$ be the event that $\boldsymbol{\sigma}$ is such that this inequality is satisfied (and noting that its probability of non-occurence is at most $2\exp(-d) + 2\exp(-2t^2)$), we get overall that

$$
\begin{aligned}
\Pr(\langle \mathbf{w}, \hat{\mathbf{v}} \rangle \in [1-\epsilon, 1]) &= \Pr(E) \cdot \Pr(\langle \mathbf{w}, \hat{\mathbf{v}} \rangle \in [1-\epsilon, 1] | E) + \Pr(\neg E) \cdot \Pr(\langle \mathbf{w}, \hat{\mathbf{v}} \rangle \in [1-\epsilon, 1] | \neg E) \\
&\leq 1 \cdot \Pr(\langle \mathbf{w}, \hat{\mathbf{v}} \rangle \in [1-\epsilon, 1] | E) + \Pr(\neg E) \cdot 1 \\
&\leq \sqrt{\frac{2}{\pi}} \cdot \max\left\{1, \frac{t}{c_d\sqrt{2}}\right\} \cdot \frac{\epsilon}{1-\epsilon} + 2\exp(-d) + 2\exp(-2t^2).
\end{aligned}
$$

Recalling Eq. (11) and the definition of $c_d$, we get that

$$
\Pr(\langle \mathbf{w}, \mathbf{v} \rangle \in [1-\epsilon, 1]) \leq \sqrt{\frac{8}{\pi}} \cdot \max\left\{1, t \cdot \sqrt{2\log(4d)}\right\} \cdot \frac{\epsilon}{1-\epsilon} + 2\exp(-d) + 2\exp(-2t^2).
$$

Picking $t = \sqrt{\frac{1}{2}\log\left(\frac{1-\epsilon}{\epsilon}\right)}$, we get the bound

$$
\left(\sqrt{\frac{8}{\pi}} \cdot \max\left\{1, \sqrt{\log\left(\frac{1-\epsilon}{\epsilon}\right)\log(4d)}\right\} + 2\right) \cdot \frac{\epsilon}{1-\epsilon} + 2\exp(-d) = \tilde{\mathcal{O}}\left(\epsilon + \exp(-d)\right).
$$

This justifies Eq. (10), from which the result follows.

---

[2] If $\mu \in [1-\epsilon, 1]$, then we get $\max\{\mu, 0\} = \mu \geq 1-\epsilon$. If $\mu > 1$, we get $\max\{\mu, \mu - 1\} > 1 \geq 1-\epsilon$. If $\mu < 1-\epsilon$, we get $\max\{\mu, 1-\epsilon-\mu\} \geq (1-\epsilon)/2$.

## 6.3 Proof of Thm. 4

The proof rests largely on the following key result:

**Theorem 7.** *Let $\mathcal{G}_n$ be the family of piece-wise linear functions on the domain $[0,1]$ comprised of at most $n$ linear segments. Let $\mathcal{G}_n^d$ be the family of piece-wise linear functions on the domain $[0,1]^d$, with the property that for any $g \in \mathcal{G}_n^d$ and any affine transformation $h : \mathbb{R} \to \mathbb{R}^d$, $g \circ h \in \mathcal{G}_n$. Suppose $f : [0,1]^d \to \mathbb{R}$ is $C^2$. Then for all $\lambda > 0$*

$$\inf_{g \in \mathcal{G}_n^d} \int_{[0,1]^d} (f-g)^2 d\mu_d \geq \frac{c \cdot \lambda^2 \cdot \sigma_\lambda(f)^5}{n^4},$$

*where $c = \frac{5}{4096}$.*

Thm. 7 establishes that the error of a piece-wise linear approximation of a $C^2$ function cannot decay faster than quartically in the number of linear segments of any *one-dimensional* projection of the approximating function. Note that this result is stronger than a bound in terms of the total number of linear regions in $\mathbb{R}^d$, since that number can be exponentially higher (in the dimension) than $n$ as defined in the theorem.

Before proving Thm. 7, let us explain how we can use it to prove Thm. 4. To that end, we use the result in Telgarsky (2016, Lemma 3.2), of which the following is an immediate corollary:

**Corollary 2.** *Let $\mathcal{N}_{m,l}^d$ denote the family of ReLU neural networks receiving input of dimension $d$ and having depth $l$ and maximal width $m$. Then*

$$\mathcal{N}_{m,l}^d \subseteq \mathcal{G}_{(2m)^l}^d.$$

Combining this corollary with Thm. 7, the result follows. The remainder of this subsection will be devoted to proving Thm. 7.

### 6.3.1 Some Technical Tools

**Definition 2.** *Let $P_i$ denote the $i^{th}$ Legendre Polynomial given by Rodrigues' formula:*

$$P_i(x) = \frac{1}{2^i i!} \frac{d^i}{dx^i} \left[ \left( x^2 - 1 \right)^i \right].$$

These polynomials are useful for the following analysis since they obey the orthogonality relationship

$$\int_{-1}^{1} P_i(x) P_j(x) \, dx = \frac{2}{2i+1} \delta_{ij}.$$

Since we are interested in approximations on small intervals where the approximating function is linear, we use the change of variables $x = \frac{2}{\ell} t - \frac{2}{\ell} a - 1$ to obtain an orthogonal family $\left\{ \tilde{P}_i \right\}_{i=1}^{\infty}$ of shifted Legendre polynomials on the interval $[a, a+\ell]$ with respect to the $L_2$ norm. The first few polynomials of this family are given by

$$\begin{aligned}
\tilde{P}_0(x) &= 1 \\
\tilde{P}_1(x) &= \frac{2}{\ell} x - \left( \frac{2}{\ell} a + 1 \right) \\
\tilde{P}_2(x) &= \frac{6}{\ell^2} x^2 - \left( \frac{12a}{\ell^2} + \frac{6}{\ell} \right) x + \left( \frac{6a^2}{\ell^2} + \frac{6a}{\ell} + 1 \right).
\end{aligned} \tag{12}$$

The shifted Legendre polynomials obey the orthogonality relationship

$$\int_a^{a+\ell} \tilde{P}_i(x) \tilde{P}_j(x)\, dx = \frac{\ell}{2i+1}\delta_{ij}. \tag{13}$$

**Definition 3.** *We define the Fourier-Legendre series of a function $f : [a, a+\ell] \to \mathbb{R}$ to be*

$$f(x) = \sum_{i=0}^{\infty} \tilde{a}_i \tilde{P}_i(x),$$

*where the Fourier-Legendre Coefficients $\tilde{a}_i$ are given by*

$$\tilde{a}_i = \frac{2i+1}{\ell} \int_a^{a+\ell} \tilde{P}_i(x) f(x)\, dx.$$

**Theorem 8.** *A generalization of Parseval's identity yields*

$$\|f\|_{L_2}^2 = \ell \sum_{i=0}^{\infty} \frac{\tilde{a}_i^2}{2i+1}. \tag{14}$$

**Definition 4.** *A function $f$ is $\lambda$-strongly convex if for all $\mathbf{w}, \mathbf{u}$ and $\alpha \in (0,1)$,*

$$f(\alpha\mathbf{w} + (1-\alpha)\mathbf{u}) \le \alpha f(\mathbf{w}) + (1-\alpha)f(\mathbf{u}) - \frac{\lambda}{2}\alpha(1-\alpha)\|\mathbf{w} - \mathbf{u}\|_2^2.$$

*A function is $\lambda$-strongly concave, if $-f$ is $\lambda$-strongly convex.*

### 6.3.2 One-dimensional Lower Bounds

We begin by proving two useful lemmas; the first will allow us to compute the error of a linear approximation of one-dimensional functions on arbitrary intervals, and the second will allow us to infer bounds on the entire domain of approximation, from the lower bounds we have on small intervals where the approximating function is linear.

**Lemma 2.** *Let $f \in C^2$. Then the error of the optimal linear approximation of $f$ denoted $Pf$ on the interval $[a, a+\ell]$ satisfies*

$$\|f - Pf\|_{L_2}^2 = \ell \sum_{i=2}^{\infty} \frac{\tilde{a}_i^2}{2i+1}. \tag{15}$$

*Proof.* A standard result on Legendre polynomials is that given any function $f$ on the interval $[a, a+\ell]$, the best linear approximation (w.r.t. the $L_2$ norm) is given by

$$Pf = \tilde{a}_0 \tilde{P}_0(x) + \tilde{a}_1 \tilde{P}_1(x),$$

where $\tilde{P}_0, \tilde{P}_1$ are the shifted Legendre polynomials of degree $0$ and $1$ respectively, and $\tilde{a}_0, \tilde{a}_1$ are the first two Fourier-Legendre coefficients of $f$ as defined in Eq. (3). The square of the error obtained by this approximation is therefore

$$\begin{aligned}
\|f - Pf\|^2 &= \|f\|^2 - 2\langle f, Pf \rangle + \|Pf\|^2 \\
&= \ell \left( \sum_{i=0}^{\infty} \frac{\tilde{a}_i^2}{2i+1} - 2\left(\tilde{a}_0^2 + \frac{\tilde{a}_1^2}{3}\right) + \tilde{a}_0^2 + \frac{\tilde{a}_1^2}{3} \right) \\
&= \ell \sum_{i=2}^{\infty} \frac{\tilde{a}_i^2}{2i+1}.
\end{aligned}$$

Where in the second equality we used the orthogonality relationship from Eq. (13), and the generalized Parseval's identity from Eq. (14). □

**Lemma 3.** *Suppose* $f : [0,1] \to \mathbb{R}$ *satisfies* $\|f - Pf\|^2_{L_2} \geq c\ell^5$ *for some constant* $c > 0$, *and on any interval* $[a, a+\ell] \subseteq [0,1]$. *Then*

$$\inf_{g \in \mathcal{G}_n} \int_0^1 (f-g)^2 d\mu \geq \frac{c}{n^4}.$$

*Proof.* Let $g \in \mathcal{G}_n$ be some function, let $a_0 = 0, a_1, \ldots, a_{n-1}, a_n = 1$ denote its partition into segments of length $\ell_j = a_j - a_{j-1}$, where $g$ is linear when restricted to any interval $[a_{j-1}, a_j]$, and let $g_j$, $j = 1, \ldots, n$ denote the linear restriction of $g$ to the interval $[a_{j-1}, a_j]$. Then

$$\int_0^1 (f-g)^2 \, d\mu = \sum_{j=1}^n \int_{a_{j-1}}^{a_j} (f - g_j)^2 \, d\mu$$

$$\geq \sum_{j=1}^n c\ell_j^5$$

$$= c \sum_{j=1}^n \ell_j^5. \tag{16}$$

Now, recall Hölder's sum inequality which states that for any $p, q$ satisfying $\frac{1}{p} + \frac{1}{q} = 1$ we have

$$\sum_{j=1}^n |x_j y_j| \leq \left( \sum_{j=1}^n |x_j|^p \right)^{1/p} \left( \sum_{j=1}^n |y_j|^q \right)^{1/q}.$$

Plugging in $x_j = \ell_j, y_j = 1 \ \forall j \in \{1, \ldots, n\}$ we have

$$\sum_{j=1}^n |\ell_j| \leq \left( \sum_{j=1}^n |\ell_j|^p \right)^{1/p} n^{1/q},$$

and using the equalities $\sum_{j=1}^n |\ell_j| = 1$ and $\frac{p}{q} = p - 1$ we get that

$$\frac{1}{n^{p-1}} \leq \sum_{j=1}^n |\ell_j|^p. \tag{17}$$

Plugging the inequality from Eq. (17) with $p = 5$ in Eq. (16) yields

$$\int_0^1 (p-g)^2 \, d\mu \geq \frac{c}{n^4},$$

concluding the proof of the lemma. □

Our first lower bound for approximation using piece-wise linear functions is for non-linear target functions of the simplest kind. Namely, we obtain lower bounds on quadratic functions.

**Theorem 9.** *If $\mathcal{G}_n$ is the family of piece-wise linear functions with at most $n$ linear segments in the interval $[0, 1]$, then for any quadratic function $p(x) = p_2 x^2 + p_1 x + p_0$, we have*

$$\inf_{g \in \mathcal{G}_n} \int_0^1 (p - g)^2 d\mu \geq \frac{p_2^2}{180 n^4}. \tag{18}$$

*Proof.* Observe that since $p$ is a degree 2 polynomial, we have that its coefficients satisfy $\tilde{a}_i = 0 \ \forall i \geq 3$, so from Lemma 2 its optimal approximation error equals $\frac{\tilde{a}_2^2 \ell}{5}$. Computing $\tilde{a}_2$ can be done directly from the equation

$$p(x) = \sum_{i=0}^{2} \tilde{a}_i \tilde{P}_i(x),$$

Which gives $\tilde{a}_2 = \frac{p_2 \ell^2}{6}$ due to Eq. (12). This implies that

$$\|p - Pp\|^2 = \frac{p_2^2 \ell^5}{180}.$$

Note that for quadratic functions, the optimal error is dependent solely on the length of the interval. Using Lemma 3 with $c = \frac{p_2^2}{180}$ we get

$$\int_0^1 (p - g)^2 \, d\mu \geq \frac{p_2^2}{180 n^4},$$

concluding the proof of the theorem. $\qquad \square$

Computing a lower bound for quadratic functions is made easy since the bound on any interval $[a, a + \ell]$ depends on $\ell$ but not on $a$. This is not the case in general, as can be seen by observing monomials of high degree $k$. As $k$ grows, $x^k$ on the interval $[0, 0.5]$ converges rapidly to 0, whereas on $\left[1 - \frac{1}{k}, 1\right]$ its second derivative is lower bounded by $\frac{k(k-1)}{4}$, which indicates that indeed a lower bound for $x^k$ will depend on $a$.

For non-quadratic functions, however, we now show that a lower bound can be derived under the assumption of strong convexity (or strong concavity) in $[0, 1]$.

**Theorem 10.** *Suppose $f : [0, 1] \to \mathbb{R}$ is $C^2$ and either $\lambda$-strongly convex or $\lambda$-strongly concave. Then*

$$\inf_{g \in \mathcal{G}_n} \int_0^1 (f - g)^2 d\mu \geq c \lambda^2 n^{-4}, \tag{19}$$

*where $c > 0$ is a universal constant.*

*Proof.* We first stress that an analogous assumption to $\lambda$-strong convexity would be that $f$ is $\lambda$-strongly concave, since the same bound can be derived under concavity by simply applying the theorem to the additive inverse of $f$, and observing that the additive inverse of any piece-wise linear approximation of $f$ is in itself, of course, a piece-wise linear function. For this reason from now on we shall use the convexity assumption, but will also refer without loss of generality to concave functions.

As in the previous proof, we first prove a bound on intervals of length $\ell$ and then generalize for the unit interval. From Lemma 2, it suffices that we lower bound $\tilde{a}_2$ (although this might not give the tightest lower

bound in terms of constants, it is possible to show that it does give a tight bound over all $C^2$ functions). We compute

$$\tilde{a}_2 = \frac{5}{\ell} \int_a^{a+\ell} \tilde{P}_2(x) f(x) \, dx$$

$$= \frac{5}{\ell} \int_a^{a+\ell} P_2 \left( \frac{2}{\ell} x - \frac{2}{\ell} a - 1 \right) f(x) \, dx,$$

using the change of variables $t = \frac{2}{\ell} x - \frac{2}{\ell} a - 1$, $dt = \frac{2}{\ell} dx$, we get the above equals

$$\frac{5}{2} \int_{-1}^1 P_2(t) f \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt$$

$$= \frac{5}{4} \int_{-1}^1 (3t^2 - 1) f \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt.$$

We now integrate by parts twice, taking the anti-derivative of the polynomial to obtain

$$\frac{5}{4} \int_{-1}^1 (3t^2 - 1) f \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt$$

$$= \frac{5}{4} \left[ (t^3 - t) f \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) \right]_{-1}^1 - \frac{5\ell}{8} \int_{-1}^1 (t^3 - t) f' \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt$$

$$= \frac{5\ell}{8} \int_{-1}^1 (t - t^3) f' \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt$$

$$= \frac{5\ell}{8} \left[ \left( \frac{t^2}{2} - \frac{t^4}{4} \right) f' \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) \right]_{-1}^1$$

$$- \frac{5\ell^2}{16} \int_{-1}^1 \left( \frac{t^2}{2} - \frac{t^4}{4} \right) f'' \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt$$

$$= \frac{5\ell}{32} \left( f'(a + \ell) - f'(a) \right) - \frac{5\ell^2}{16} \int_{-1}^1 \left( \frac{t^2}{2} - \frac{t^4}{4} \right) f'' \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt. \qquad (20)$$

But since $\frac{t^2}{2} - \frac{t^4}{4} \in \left[ 0, \frac{1}{4} \right] \ \forall t \in [-1, 1]$ and since $f'' > 0$ due to strong convexity, we have that

$$\int_{-1}^1 \left( \frac{t^2}{2} - \frac{t^4}{4} \right) f'' \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt \leq \frac{1}{4} \int_{-1}^1 f'' \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt.$$

Plugging this inequality in Eq. (20) yields

$$\tilde{a}_2 \geq \frac{5\ell}{32} \left( f'(a + \ell) - f'(a) \right) - \frac{5\ell^2}{64} \int_{-1}^1 f'' \left( \frac{\ell}{2} t + \frac{\ell}{2} + a \right) dt$$

$$= \frac{5\ell}{32} \left( f'(a + \ell) - f'(a) \right) - \frac{5\ell^2}{64} \left( f'(a + \ell) - f'(a) \right)$$

$$= \left( 1 - \frac{\ell}{2} \right) \frac{5\ell}{32} \left( f'(a + \ell) - f'(a) \right),$$

but $\ell \leq 1$, so the above is at least

$$\frac{5\ell}{64} \left( f'(a + \ell) - f'(a) \right). \qquad (21)$$

24

By Lagrange's intermediate value theorem, there exists some $\xi \in [a, a + \ell]$ such that $f'(a + \ell) - f'(a) = \ell f''(\xi)$, so Eq. (21) is at least

$$\frac{5\ell^2}{64} f''(\xi),$$

and by using the strong convexity of $f$ again, we get that

$$\tilde{a}_2 \geq \frac{5\lambda \ell^2}{64}.$$

Lemma 2 now gives

$$\|f - Pf\|^2 = \ell \sum_{i=2}^{\infty} \frac{\tilde{a}_i^2}{2i + 1}$$

$$\geq \ell \frac{\tilde{a}_2^2}{5}$$

$$\geq \frac{5\lambda^2 \ell^5}{4096}.$$

Finally, by using Lemma 3 we conclude

$$\inf_{g \in \mathcal{G}_n} \int_0^1 (f - g)^2 d\mu \geq \frac{5\lambda^2}{4096 n^4}.$$

$\square$

We now derive a general lower bound for functions $f : [0, 1] \to \mathbb{R}$.

**Theorem 11.** *Suppose $f : [0, 1] \to \mathbb{R}$ is $C^2$. Then for any $\lambda > 0$*

$$\inf_{g \in \mathcal{G}_n} \int_0^1 (f - g)^2 d\mu \geq \frac{c \cdot \lambda^2 \cdot \sigma_\lambda(f)^5}{n^4}.$$

*Proof.* First, observe that if $f$ is $\lambda$-strongly convex on $[a, b]$, then $f((b - a) x + a)$ is $\lambda (b - a)^2$-strongly convex on $[0, 1]$ since $\forall x \in [0, 1]$,

$$\frac{\partial}{\partial x^2} f((b - a) x + a) = (b - a)^2 f''((b - a) x + a) \geq \lambda (b - a)^2.$$

Now, we use the change of variables $x = (b - a) t + a$, $dx = (b - a) dt$

$$\inf_{g \in \mathcal{G}_n} \int_a^b (f(x) - g(x))^2 dx$$

$$= \inf_{g \in \mathcal{G}_n} (b - a) \int_0^1 (f((b - a) t + a) - g((b - a) t + a))^2 dt$$

$$= \inf_{g \in \mathcal{G}_n} (b - a) \int_0^1 (f((b - a) t + a) - g(t))^2 dt$$

$$\geq \frac{c \cdot \lambda^2 \cdot (b - a)^5}{n^4}, \tag{22}$$

where the inequality follows from an application of Thm. 10. Back to the theorem statement, if $\sigma_\lambda = 0$ then the bound trivially holds, therefore assume $\lambda > 0$ such that $\sigma_\lambda > 0$. Since $f$ is strongly convex on a set of measure $\sigma_\lambda > 0$, the theorem follows by applying the inequality from Eq. (22). $\square$

25

### 6.3.3 Multi-dimensional Lower Bounds

We now move to generalize the bounds in the previous subsection to general dimension $d$. Namely, we can now turn to proving Thm. 7.

*Proof of Thm. 7.* Analogously to the proof of Thm. 11, we identify a neighborhood of $f$ in which the restriction of $f$ to a line in a certain direction is non-linear. We then integrate along all lines in that direction and use the result of Thm. 11 to establish the lower bound.

Before we can prove the theorem, we need to assert that indeed there exists a set having a strictly positive measure where $f$ has strong curvature along a certain direction. Assuming $f$ is not piece-wise linear; namely, we have some $\mathbf{x}_0 \in [0,1]^d$ such that $H(f)(\mathbf{x}_0) \neq 0$. Since $H(f)$ is continuous, we have that the function $h_{\mathbf{v}}(\mathbf{x}) = \mathbf{v}^\top H(f)(\mathbf{x})\mathbf{v}$ is continuous and there exists a direction $\mathbf{v} \in \mathbb{S}^{d-1}$ where without loss of generality $h_{\mathbf{v}}(\mathbf{x}_0) > 0$. Thus, we have an open neighborhood containing $\mathbf{x}_0$ where restricting $f$ to the direction $\mathbf{v}$ forms a strongly convex function, which implies that indeed $\sigma_\lambda > 0$ for small enough $\lambda > 0$.

We now integrate the approximation error on $f$ in the neighborhood $U$ along the direction $\mathbf{v}$. Compute

$$\inf_{g \in \mathcal{G}_n^d} \int_{[0,1]^d} (f-g)^2 d\mu_d$$

$$= \inf_{g \in \mathcal{G}_n^d} \int_{\mathbf{u}:\langle \mathbf{u},\mathbf{v}\rangle=0} \int_{\beta:(\mathbf{u}+\beta\mathbf{v})\in[0,1]^d} (f-g)^2 \, d\mu_1 d\mu_{d-1}$$

$$\geq \inf_{g \in \mathcal{G}_n^d} \int_{\mathbf{u}:\langle \mathbf{u},\mathbf{v}\rangle=0} \int_{\beta:(\mathbf{u}+\beta\mathbf{v})\in U} (f-g)^2 \, d\mu_1 d\mu_{d-1}$$

$$\geq \int_{\mathbf{u}:\langle \mathbf{u},\mathbf{v}\rangle=0} (\mu_1(\{\beta:(\mathbf{u}+\beta\mathbf{v})\in U\}))^5 \frac{5\lambda^2}{4096n^4} d\mu_{d-1}$$

$$= \frac{5\lambda^2}{4096n^4} \int_{\mathbf{u}:\langle \mathbf{u},\mathbf{v}\rangle=0} |\mu_1(\{\beta:(\mathbf{u}+\beta\mathbf{v})\in U\})|^5 \, d\mu_{d-1}$$

$$\geq \frac{5\lambda^2}{4096n^4} \left( \int_{\mathbf{u}:\langle \mathbf{u},\mathbf{v}\rangle=0} \mu_1(\{\beta:(\mathbf{u}+\beta\mathbf{v})\in U\}) \, d\mu_{d-1} \right)^5$$

$$= \frac{5\lambda^2 \sigma_\lambda^5}{4096n^4},$$

where in the second inequality we used Thm. 11 and in the third inequality we used Jensen's inequality with respect to the convex function $x \mapsto |x|^5$. $\qquad\square$

## 6.4 Proof of Thm. 6

We begin by monitoring the rate of growth of the error when performing either an addition or a multiplication. Suppose that the given input $\tilde{a}, \tilde{b}$ is of distance at most $\delta > 0$ from the desired target values $a, b$, i.e., $|a - \tilde{a}| \leq \delta, |b - \tilde{b}| \leq \delta$. Then for addition we have

$$\left| (a+b) - \left( \tilde{a} + \tilde{b} \right) \right| \leq |a - \tilde{a}| + \left| b - \tilde{b} \right| \leq 2\delta,$$

and for multiplication we compute the product error estimation

$$\left| \tilde{a} \cdot \tilde{b} - a \cdot b \right| \leq \left| (a + \delta) \cdot (b + \delta) - a \cdot b \right|$$
$$= \left| \delta (a + b) + \delta^2 \right|. \tag{23}$$

Now, we have bounded the error of approximating the product of two numbers which we only have approximations of, but since the computation of the product itself cannot be done with perfect accuracy using ReLU networks, we need to suffer the error of approximating a product, as shown in Thm. 5. We add the error of approximating the product of $\tilde{a} \cdot \tilde{b}$, which we may assume is at most $\delta$ (assuming $\Theta \left( \log_2 (M/\delta) \right)$ bits are used for the product, since each intermediate computation is bounded in the interval $[-M, M]$). Overall, we get an error bound of

$$\left| \delta (a + b) + \delta^2 + \delta \right| \leq 3M\delta.$$

From this, we see that at each stage the error grows by at most a multiplicative factor of $3M$. After $t$ operations, and with an initial estimation error of $\delta$, we have that the error is bounded by $(3M)^{t-1} \delta$. Choosing $\delta \leq (3M)^{1-t} \epsilon$ to guarantee approximation $\epsilon$, we have from Thm. 5 that each operation will require at most

$$4 \left\lceil \log \left( \frac{M (3M)^{t-1}}{\epsilon} \right) \right\rceil + 13 \leq c \left( \log \left( \frac{1}{\epsilon} \right) + t \log (M) \right)$$

width and

$$2 \left\lceil \log \left( \frac{M (3M)^{t-1}}{\epsilon} \right) \right\rceil + 9 \leq c \left( \log \left( \frac{1}{\epsilon} \right) + t \log (M) \right)$$

depth for some universal $c > 0$. Composing the networks performing each operation, we arrive at a total network width and depth of at most

$$c \left( t \log \left( \frac{1}{\epsilon} \right) + t^2 \log (M) \right).$$

Now, our target function is approximated to accuracy $\epsilon$ by a function which our network approximates to the same accuracy $\epsilon$, for a total approximation error of the target function by our network of $2\epsilon$.