A. Additional Proofs

A.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 assumptions 1 and 2 from Eldan & Shamir (2016):

Assumption 1. *Given the activation function σ, there is a constant $c_\sigma \geq 1$ (depending only on σ) 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 δ, 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.

Assumption 2. *The activation function σ is (Lebesgue) measurable and satisfies*

$$|\sigma(x)| \leq C\left(1 + |x|^\alpha\right)$$

for all $x \in \mathbb{R}$ and for some constants $C, \alpha > 0$.

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 $a + \sum_{i=1}^w \alpha_i \sigma\left(\beta_i x - \gamma_i\right)$ 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 $\frac{8 c_\sigma d}{\delta}$. Consequently, the function

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

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

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

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 μ 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

$$\begin{aligned}
&\int_{\mathbb{R}^d} \left(\tilde{f}\left(c_\mu \cdot \tilde{\ell}(\mathbf{x}) - c_\mu\right) - \mathbf{1}\left(\|\mathbf{x}\|_2 \leq 1\right)\right)^2 \mu(\mathbf{x})\, d\mathbf{x} \\
=&\int_{R_\epsilon} \left(\tilde{f}\left(c_\mu \cdot \tilde{\ell}(\mathbf{x}) - c_\mu\right) - \mathbf{1}\left(\|\mathbf{x}\|_2 \leq 1\right)\right)^2 \mu(\mathbf{x})\, d\mathbf{x} \\
&\quad + \int_{\mathbb{R}^d \setminus R_\epsilon} \left(\tilde{f}\left(c_\mu \cdot \tilde{\ell}(\mathbf{x}) - c_\mu\right) - \mathbf{1}\left(\|\mathbf{x}\|_2 \leq 1\right)\right)^2 \mu(\mathbf{x})\, d\mathbf{x} \\
\leq &\int_{R_\epsilon} 2\mu(\mathbf{x})\, d\mathbf{x} + \int_{\mathbb{R}^d \setminus R_\epsilon} \frac{\delta}{4}\mu(\mathbf{x})\, d\mathbf{x} \\
\leq &\frac{\delta}{2} + \frac{\delta}{2} = \delta,
\end{aligned}$$

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 $1 + \sqrt{\delta/2} \leq \sqrt{2}$, and the second summand justified due to Equations (7) and (8), concluding the proof of the lemma. □

A.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,

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

By Markov's inequality, this implies

$$\Pr_{\mathbf{v}}\left(\mathbb{E}_s\left[f(sr) - F(sr\mathbf{v})^2 \middle| \mathbf{v}\right] \leq \delta_\epsilon\right) \;\geq\; \frac{1}{2}\,.$$

By the assumption on f, and the fact that s is uniform on $[1, 1+\epsilon]$, we have that $\mathbb{E}_s\left[f(sr) - F(sr\mathbf{v})^2 \middle| \mathbf{v}\right] \leq \delta_\epsilon$ 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

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

The left hand side equals

$$\begin{aligned}
&\Pr_{\mathbf{v}}\left(\exists j \in \{1\ldots N\}, s \in [1, 1+\epsilon] \text{ s.t. } \langle \mathbf{w}_j, \mathbf{v}\rangle = -b_j/sr\right) \\
&= \Pr_{\mathbf{v}}\left(\exists j \in \{1\ldots N\} \text{ s.t. } \langle \mathbf{w}_j, \mathbf{v}\rangle \text{ between } -\frac{b_j}{r} \text{ and } -\frac{b_j}{(1+\epsilon)r}\right) \\
&\leq \Pr_{\mathbf{v}}\left(\exists j \in \{1\ldots N\} \text{ s.t. } \langle \mathbf{w}_j, \mathbf{v}\rangle \text{ betwen } -\frac{b_j}{r} \text{ and } -(1-\epsilon)\frac{b_j}{r}\right) \\
&\leq \sum_{j=1}^{N} \Pr_{\mathbf{v}}\left(\langle \mathbf{w}_j, \mathbf{v}\rangle \text{ between } -\frac{b_j}{r} \text{ and } -(1-\epsilon)\frac{b_j}{r}\right) \\
&\leq N \cdot \sup_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \Pr_{\mathbf{v}}\left(\langle \mathbf{w}, \mathbf{v}\rangle \text{ between } -\frac{b}{r} \text{ and } -(1-\epsilon)\frac{b}{r}\right) \\
&= N \cdot \sup_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \Pr_{\mathbf{v}}\left(\left\langle \frac{-r\mathbf{w}}{b}, \mathbf{v}\right\rangle \in [1-\epsilon, 1]\right) \;=\; N \cdot \sup_{\mathbf{w}\in\mathbb{R}^d} \Pr_{\mathbf{v}}\left(\langle \mathbf{w}, \mathbf{v}\rangle \in [1-\epsilon, 1]\right),
\end{aligned}$$

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

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

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

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

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

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

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 μ 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 μ 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

$$\begin{aligned}
&\frac{1}{\sqrt{2\pi v^2}} \int_{1-\epsilon}^{1} \exp\left(-\frac{(x-\mu)^2}{v^2}\right) dx \;\leq\; \frac{\epsilon}{\sqrt{2\pi v^2}} \cdot \max_{x \in [1-\epsilon, 1]} \exp\left(-\frac{(x-\mu)^2}{v^2}\right) \\
&\quad\leq\; \frac{\epsilon}{\sqrt{2\pi v^2}} \cdot \max_{x \in [1-\epsilon, 1]} \min\left\{1, \frac{v}{|x-\mu|}\right\} \;=\; \frac{\epsilon}{\sqrt{2\pi}} \cdot \max_{x \in [1-\epsilon, 1]} \min\left\{\frac{1}{v}, \frac{1}{|x-\mu|}\right\} \\
&\quad=\; \frac{\epsilon}{\sqrt{2\pi}} \cdot \max_{x \in [1-\epsilon, 1]} \frac{1}{\max\{v, |x-\mu|\}} \;=\; \frac{\epsilon}{\sqrt{2\pi}} \cdot \max_{x \in [1-\epsilon, 1]} \frac{\max\{1, \frac{\mu}{v}\}}{\max\{1, \frac{\mu}{v}\} \cdot \max\{v, |x-\mu|\}} \\
&\quad\leq\; \frac{\epsilon}{\sqrt{2\pi}} \cdot \max_{x \in [1-\epsilon, 1]} \frac{\max\{1, \frac{\mu}{v}\}}{\max\{\mu, |x-\mu|\}} \;=\; \frac{\epsilon}{\sqrt{2\pi}} \cdot \frac{\max\{1, \frac{\mu}{v}\}}{\max\{\mu, \min_{x \in [1-\epsilon, 1]} |x-\mu|\}}
\end{aligned}$$

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

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 $\boldsymbol{\sigma}$,

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

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 $\max\{\mu, 1-\epsilon-\mu\} \geq (1-\epsilon)/2$.

A.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.

A.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* λ-strongly convex *if for all $\mathbf{w}, \mathbf{u}$ and $\alpha \in (0,1)$,*

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

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

A.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 $\|f - Pf\|_{L_2}^2 \geq c\ell^5$ 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

$$\begin{aligned}\int_0^1 (f-g)^2 \, d\mu &= \sum_{j=1}^{n} \int_{a_{j-1}}^{a_j} (f-g_j)^2 \, d\mu \\ &\geq \sum_{j=1}^{n} c\ell_j^5 \\ &= c\sum_{j=1}^{n} \ell_j^5. \end{aligned} \tag{16}$$

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

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

concluding the proof of the theorem. □

Computing a lower bound for quadratic functions is made easy since the bound on any interval $[a, a+\ell]$ depends on ℓ 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 λ-strongly convex or λ-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 λ-strong convexity would be that f is λ-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 ℓ 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

$$\begin{aligned}\tilde{a}_2 &= \frac{5}{\ell}\int_a^{a+\ell} \tilde{P}_2(x) f(x)\, dx \\ &= \frac{5}{\ell}\int_a^{a+\ell} P_2\left(\frac{2}{\ell}x - \frac{2}{\ell}a - 1\right) f(x)\, dx,\end{aligned}$$

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

$$\begin{aligned}&\frac{5}{2}\int_{-1}^1 P_2(t) f\left(\frac{\ell}{2}t + \frac{\ell}{2} + a\right) dt \\ =&\frac{5}{4}\int_{-1}^1 \left(3t^2-1\right) f\left(\frac{\ell}{2}t + \frac{\ell}{2} + a\right) dt.\end{aligned}$$

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

$$\begin{aligned}
&\frac{5}{4}\int_{-1}^{1}\left(3t^2-1\right)f\left(\frac{\ell}{2}t+\frac{\ell}{2}+a\right)dt\\
=&\frac{5}{4}\left[\left(t^3-t\right)f\left(\frac{\ell}{2}t+\frac{\ell}{2}+a\right)\right]_{-1}^{1}-\frac{5\ell}{8}\int_{-1}^{1}\left(t^3-t\right)f'\left(\frac{\ell}{2}t+\frac{\ell}{2}+a\right)dt\\
=&\frac{5\ell}{8}\int_{-1}^{1}\left(t-t^3\right)f'\left(\frac{\ell}{2}t+\frac{\ell}{2}+a\right)dt\\
=&\frac{5\ell}{8}\left[\left(\frac{t^2}{2}-\frac{t^4}{4}\right)f'\left(\frac{\ell}{2}t+\frac{\ell}{2}+a\right)\right]_{-1}^{1}\\
&-\frac{5\ell^2}{16}\int_{-1}^{1}\left(\frac{t^2}{2}-\frac{t^4}{4}\right)f''\left(\frac{\ell}{2}t+\frac{\ell}{2}+a\right)dt\\
=&\frac{5\ell}{32}\left(f'\left(a+\ell\right)-f'\left(a\right)\right)-\frac{5\ell^2}{16}\int_{-1}^{1}\left(\frac{t^2}{2}-\frac{t^4}{4}\right)f''\left(\frac{\ell}{2}t+\frac{\ell}{2}+a\right)dt. \qquad (20)
\end{aligned}$$

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

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

Plugging this inequality in Eq. (20) yields

$$\begin{aligned}
\tilde{a}_2&\ge\frac{5\ell}{32}\left(f'\left(a+\ell\right)-f'\left(a\right)\right)-\frac{5\ell^2}{64}\int_{-1}^{1}f''\left(\frac{\ell}{2}t+\frac{\ell}{2}+a\right)dt\\
&=\frac{5\ell}{32}\left(f'\left(a+\ell\right)-f'\left(a\right)\right)-\frac{5\ell^2}{64}\left(f'\left(a+\ell\right)-f'\left(a\right)\right)\\
&=\left(1-\frac{\ell}{2}\right)\frac{5\ell}{32}\left(f'\left(a+\ell\right)-f'\left(a\right)\right),
\end{aligned}$$

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

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

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

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

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

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

Lemma 2 now gives

$$\begin{aligned}
\|f-Pf\|^2&=\ell\sum_{i=2}^{\infty}\frac{\tilde{a}_i^2}{2i+1}\\
&\ge\ell\frac{\tilde{a}_2^2}{5}\\
&\ge\frac{5\lambda^2\ell^5}{4096}.
\end{aligned}$$

Finally, by using Lemma 3 we conclude

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

□

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 λ-strongly convex on $[a,b]$, then $f\left((b-a)\,x + a\right)$ is $\lambda\,(b-a)^2$-strongly convex on $[0,1]$ since $\forall x \in [0,1]$,

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

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

$$\begin{aligned}
&\inf_{g \in \mathcal{G}_n} \int_a^b (f(x) - g(x))^2 dx \\
=& \inf_{g \in \mathcal{G}_n} (b-a) \int_0^1 \left(f\left((b-a)\,t + a\right) - g\left((b-a)\,t + a\right)\right)^2 dt \\
=& \inf_{g \in \mathcal{G}_n} (b-a) \int_0^1 \left(f\left((b-a)\,t + a\right) - g\left(t\right)\right)^2 dt \\
\geq& \frac{c \cdot \lambda^2 \cdot (b-a)^5}{n^4},
\end{aligned} \tag{22}$$

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). □

A.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 $H\left(f\right)\left(\mathbf{x}_0\right) \neq 0$. Since $H\left(f\right)$ is continuous, we have that the function $h_{\mathbf{v}}\left(\mathbf{x}\right) = \mathbf{v}^\top H\left(f\right)\left(\mathbf{x}\right)\mathbf{v}$ is continuous and there exists a direction $\mathbf{v} \in \mathbb{S}^{d-1}$ where without loss of generality $h_{\mathbf{v}}\left(\mathbf{x}_0\right) > 0$. 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

$$
\begin{aligned}
&\inf_{g\in\mathcal{G}_n^d}\int_{[0,1]^d}(f-g)^2d\mu_d \\
=&\inf_{g\in\mathcal{G}_n^d}\int_{\mathbf{u}:\langle\mathbf{u},\mathbf{v}\rangle=0}\int_{\beta:(\mathbf{u}+\beta\mathbf{v})\in[0,1]^d}(f-g)^2\,d\mu_1 d\mu_{d-1} \\
\geq&\inf_{g\in\mathcal{G}_n^d}\int_{\mathbf{u}:\langle\mathbf{u},\mathbf{v}\rangle=0}\int_{\beta:(\mathbf{u}+\beta\mathbf{v})\in U}(f-g)^2\,d\mu_1 d\mu_{d-1} \\
\geq&\int_{\mathbf{u}:\langle\mathbf{u},\mathbf{v}\rangle=0}\left(\mu_1\left(\{\beta:(\mathbf{u}+\beta\mathbf{v})\in U\}\right)\right)^5\frac{5\lambda^2}{4096n^4}d\mu_{d-1} \\
=&\frac{5\lambda^2}{4096n^4}\int_{\mathbf{u}:\langle\mathbf{u},\mathbf{v}\rangle=0}\left|\mu_1\left(\{\beta:(\mathbf{u}+\beta\mathbf{v})\in U\}\right)\right|^5d\mu_{d-1} \\
\geq&\frac{5\lambda^2}{4096n^4}\left(\int_{\mathbf{u}:\langle\mathbf{u},\mathbf{v}\rangle=0}\mu_1\left(\{\beta:(\mathbf{u}+\beta\mathbf{v})\in U\}\right)d\mu_{d-1}\right)^5 \\
=&\frac{5\lambda^2\sigma_\lambda^5}{4096n^4},
\end{aligned}
$$

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 $x\mapsto|x|^5$. □

A.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

$$
\begin{aligned}
\left|\tilde{a}\cdot\tilde{b}-a\cdot b\right|&\leq|(a+\delta)\cdot(b+\delta)-a\cdot b| \\
&=\left|\delta\left(a+b\right)+\delta^2\right|. \qquad (23)
\end{aligned}
$$

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 δ (assuming $\Theta\left(\log_2\left(M/\delta\right)\right)$ bits are used for the product, since each intermediate computation is bounded in the interval $[-M,M]$). Overall, we get an error bound of

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

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 δ, we have that the error is bounded by $\left(3M\right)^{t-1}\delta$. Choosing $\delta\leq\left(3M\right)^{1-t}\epsilon$ to guarantee approximation ϵ, we have from Thm. 5 that each operation will require at most

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

width and

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

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

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

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

B. L_1 Ball Indicator Experiment

In this section, we run a similar experiment to the one presented in 3.1, this time with respect to indicators of L_1 balls.

For this experiment, we sampled $5 \cdot 10^5$ data instances uniformly at random from the 14-dimensional L_1 unit sphere (i.e. each instance is of dimension 15). We then scaled the norm of each instance independently by a scaler chosen uniformly from the interval $[0, 2]$. To each instance, we associated a target value computed according to the target function $f(\mathbf{x}) = \mathbf{1}\left(\|\mathbf{x}\|_1 \leq 1\right)$. 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. 3. Like the L_2 ball experiment, we see that adding depth when learning such functions is much more helpful than increasing width. In fact, here the improvement by increased width is hardly noticeable, and the width 400 network actually obtained a slightly better error than the width 800 network. In contrast, the 3-layer network converged to a significantly better solution.

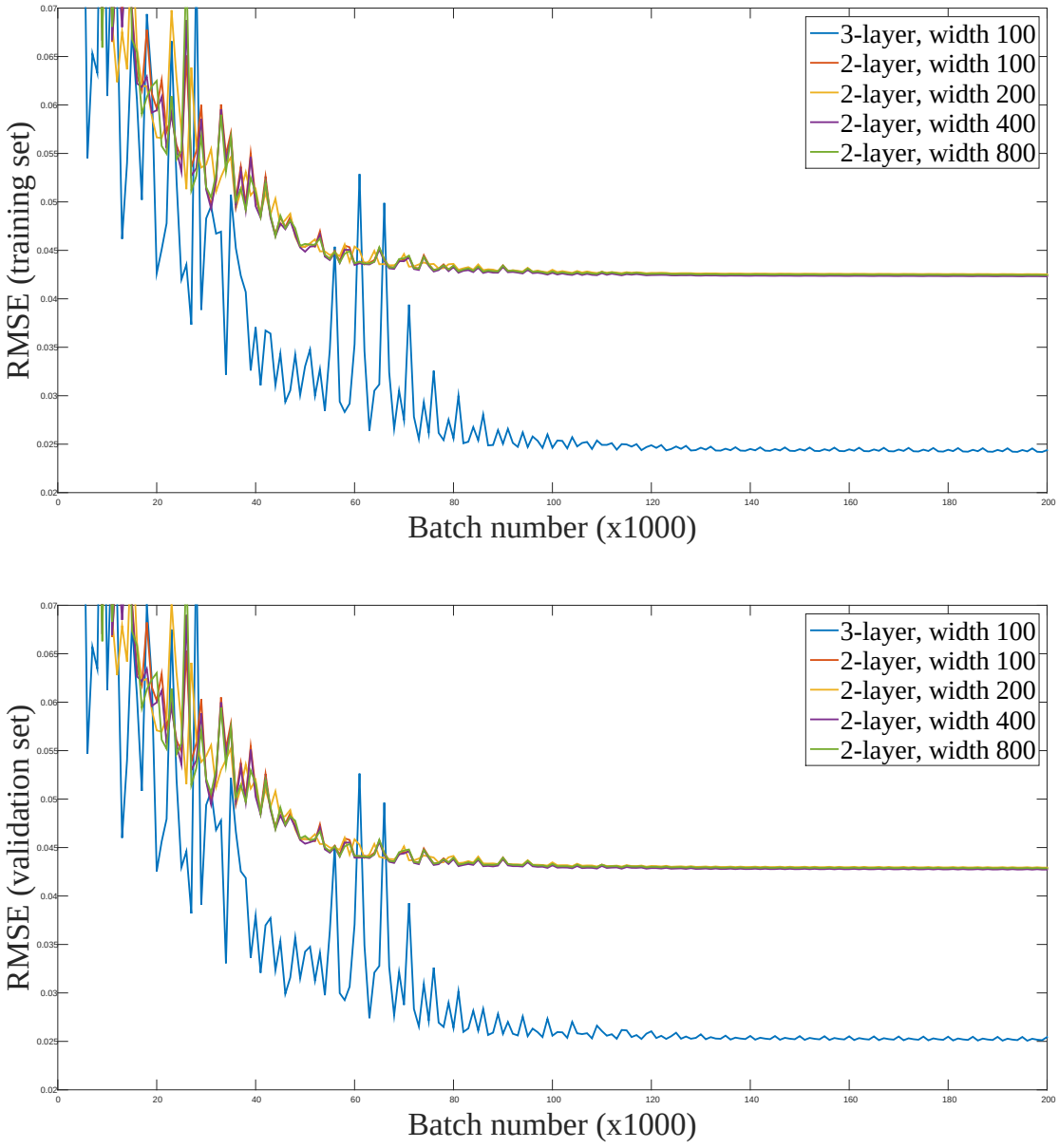

Figure 3. The L_1 experiment results, depicting the network's root mean square error over the training set (top) and validation set (bottom), as a function of the number of batches processed. Best viewed in color.