Depth Separation in Norm-Bounded Infinite-Width Neural Networks

Suzanna Parkinson[*], Greg Ongie[†], Rebecca Willett[‡], Ohad Shamir[§], Nathan Srebro[¶]

February 15, 2024

Abstract

We study depth separation in infinite-width neural networks, where complexity is controlled by the overall squared ℓ_2-norm of the weights (sum of squares of all weights in the network). Whereas previous depth separation results focused on separation in terms of width, such results do not give insight into whether depth determines if it is possible to learn a network that generalizes well even when the network width is unbounded. Here, we study separation in terms of the sample complexity required for learnability. Specifically, we show that there are functions that are learnable with sample complexity polynomial in the input dimension by norm-controlled depth-3 ReLU networks, yet are not learnable with sub-exponential sample complexity by norm-controlled depth-2 ReLU networks (with any value for the norm). We also show that a similar statement in the reverse direction is not possible: any function learnable with polynomial sample complexity by a norm-controlled depth-2 ReLU network with infinite width is also learnable with polynomial sample complexity by a norm-controlled depth-3 ReLU network.

1 Introduction

It has long been postulated that in training neural networks, "the size of the weights is more important than the size of the network" (Bartlett, 1996). That is, the inductive bias and generalization properties of learning neural networks come from seeking networks with small weights (in terms of magnitude or some norm of the weights), rather than constraining the number of weights. Small weight norm is sufficient to ensure generalization (e.g. Bartlett and Mendelson, 2002; Neyshabur et al., 2015; Golowich et al., 2018; Du and Lee, 2018; Daniely and Granot, 2019), and may be induced either through explicit regularization (e.g., via weight decay Hanson and Pratt, 1988) or implicitly through the optimization algorithm (e.g. Neyshabur et al., 2014, 2017; Chizat and Bach, 2020; Vardi, 2023). The reliance on weight-norm-based complexity control is particularly relevant with modern, heavily overparameterized networks, which have more weights than training examples. These networks can shatter the training set, and hence the size of the network alone does not lead to meaningful generalization guarantees (Zhang et al., 2017; Neyshabur et al., 2014). Indeed, over the years there has been increasing interest in the theoretical study of learning with *infinite width* networks, where the number of units per layer is unbounded or even infinite, while controlling the *norm* of the weights (Cho and Saul, 2009; Neyshabur et al., 2015; Bach, 2017; Bengio et al., 2005; Mei et al., 2019; Chizat and Bach, 2018; Jacot et al., 2018; Savarese et al., 2019; Ongie et al., 2019; Chizat and Bach, 2020; Pilanci and Ergen, 2020; Parhi and Nowak, 2021; Unser, 2023).

Considering infinite-width neural networks, and relying only on the norm of the weights for inductive bias and generalization, also requires a fresh look at the role of depth. The traditional study of the role of depth focused on how deeper networks can represent functions using fewer units. (e.g. Pinkus, 1999; Telgarsky, 2016; Eldan and Shamir, 2016; Liang and Srikant, 2016; Lu et al., 2017; Daniely, 2017; Safran and Shamir, 2017; Yarotsky, 2017, 2018; Rolnick and Tegmark, 2018; Arora et al., 2018; Safran et al., 2019; Vardi and Shamir, 2020; Chatziafratis et al., 2020; Venturi et al., 2022). Focusing on depth-2 (one hidden layer) versus depth-3 (two hidden layers) feedforward neural networks with ReLU activations (see Section 2 for precise details), traditional depth separation results tell us that there are functions that can be well-approximated using depth-3, low-width networks (number of neurons polynomial in the input

[*]Committee on Computational and Applied Mathematics, University of Chicago, Chicago, IL, USA.
[†]Department of Mathematical and Statistical Sciences, Marquette University, Milwaukee, WI, USA.
[‡]Department of Statistics and Department of Computer Science, University of Chicago, Chicago, IL, USA.
[§]Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel.
[¶]Toyota Technological Institute at Chicago, Chicago, IL, USA.

dimension), but cannot be approximated using depth-2 networks unless the width/number of neurons is exponentially high in the input dimension. However, this separation is not relevant when studying infinite-width networks.

Instead of studying depth separation in terms of the *number of weights* (i.e., width), one can study depth separation in terms of the *size of the weights*, i.e., the norm required to approximate the target function with a specific depth. This is captured by the *representation* cost $R_L(f)$, which is the minimal weight norm (sum of squares of all weights in the network) required to represent f using an unbounded-width depth-L network. One can ask whether there are functions that can be well approximated with a low R_3 representation cost, but which require a high R_2 representation cost to approximate, even if we allow unbounded or infinite width. One contribution of our paper is to show that the answer is "yes": the same function families that show depth separations in terms of width also demonstrate depth separations in terms of norm or representation cost. Specifically, with depth-3 networks, one can approximate functions in these families with norm polynomial in the input dimension, but with depth-2 networks, even with infinite width, an exponential norm is required to approximate functions in these families even within constant approximation error. At a technical level, this argument follows from explicitly accounting for the norm in the depth-3 representation, and by showing through a Barron-like unit-sampling argument that if such "hard" functions were approximable with a low norm in depth 2, they would also be approximable with a small width in depth 2, which we know from the width-based depth separation results is not true.

What does such separation between R_3 and R_2 representation cost tell us? Without further analysis of the effect of this separation on learning capabilities, it is unclear. One cannot directly compare the values of R_2 and R_3 since their comparison depends on the precise way we aggregate the norms across layers; see, e.g., Neyshabur et al. (2015) for a careful discussion. While width-separation results can be thought of as a separation in terms of the required memory costs, when discussing infinite networks we are already abstracting away the computational implementation, and working with exponentially large *weights* is not an inherent computational barrier as the number of bits is still polynomial.

Thus, instead of studying depth separation in terms of *approximation*, we directly study the separation in terms of *learning*, as captured by its effect on sample complexity. We ask the following question: If Alice is learning using norm-based inductive bias (i.e., regularization) with unbounded-width depth-2 networks, and Bob is learning using norm-based inductive bias with unbounded-width depth-3 networks, are there functions Bob can learn with a small *number of samples*, but which Alice would require a huge *number of samples* to learn? On the other hand, are there perhaps functions for which depth-2 would be better, i.e., which Alice can learn with a small number of samples with depth-2, but for which Bob would require a huge number of samples to learn by seeking a low-norm depth-3 network? As formalized in Section 4, we think of Alice and Bob as using a standard Regularized Empirical Risk Minimization or Structural Risk Minimizing (SRM) approach, where they learn by minimizing some combination of the empirical loss $\mathscr{L}_S(f)$ and weight norm, or equivalently representation cost $R_L(f)$, for depth $L = 2$ or depth $L = 3$.

Our main results are as follows (where we focus on learning functions with samples from a particular distribution chosen for technical convenience):

Theorem 1.1. *(Depth Separation, Informal) There is a family of functions $f_d : \mathbb{R}^{2d} \to \mathbb{R}$ that requires exponential (in d) sample complexity to learn to within constant error by regularizing the norm in an unbounded width depth-2 ReLU network, but which can be learned with $poly(d, 1/\varepsilon)$ samples to within any error ε by regularizing the norm in a depth-3 ReLU network.*

The next result ensures that the reverse of Theorem 1.1 does not occur.

Theorem 1.2. *(No Reverse Depth Separation, Informal) Any function learnable with $poly(d, 1/\varepsilon)$ samples by regularizing the norm in an unbounded width depth-2 ReLU network, can also be learned with $poly(d, 1/\varepsilon)$ samples by regularizing the norm in a depth-3 ReLU network.*

From these results, we conclude that functions that are "easy" to learn with depth-2 ReLU networks form a strict subset of the functions that are "easy" to learn with depth-3 ReLU networks.

At a high level, the proof of Theorem 1.1 relies on choosing a target function that is not approximable by a small norm depth-2 network. We then construct a depth-2 interpolant whose representation cost depends only mildly on the number of samples. Using the Alice-and-Bob terminology from earlier, since Alice (who utilizes depth-2 networks) tries to find a function that fits the data well and has a small representation cost, the representation cost of her function

will be at least as small as that of the interpolant. Hence, unless she has access to an enormous number of samples, her function will not be able to approximate the target and will not generalize. However, the target function is approximable by a depth-3 network with a small representation cost, so the Rademacher complexity results of Neyshabur et al. (2015) lead to sample complexity bounds that allow us to bound Bob's generalization error with many fewer samples. To prove Theorem 1.2, we show using a similar argument that Alice can only learn if the R_2 cost of approximating the target is small. We show that functions with small R_2 cost also have small R_3 cost, and so Bob must also be able to learn these target functions.

We see our contributions here on two levels:

1. Providing a detailed study of depth separation in neural networks in terms of the size of the weights rather than the number of the weights.

2. Establishing a framework and template for studying depth separation, or model separation more broadly, directly in terms of learning, with the separation being between low and high sample complexity. This is in contrast to a study solely in terms of the "complexity" needed to approximate target functions, which does not directly provide insights into sample complexities.

1.1 Outline

We define the representation cost and describe its connection to weight decay regularization in Section 2. In Section 3 we consider depth separation in the norm to approximate certain families. We more carefully describe what we mean by learning rules using a norm-based inductive bias in Section 4. The formal statements of Theorems 1.1 and 1.2 are in Section 5, and their proof sketches are in Sections 6 and 7, respectively. We conclude in Section 8 with a discussion of the implications and limitations of these results. All technical lemmas and their proofs are reserved for Appendix A.

1.2 Notation

The set of depth-L width-ω ReLU neural networks is denoted as $\mathcal{N}_{L,\omega}$, and the set of depth-L unbounded-width networks is denoted as $\mathcal{N}_L := \bigcup_{\omega\in\mathbb{N}} \mathcal{N}_{L,\omega}$. We use $\mathbb{S}^{d-1}$ for the hypersphere in $\mathbb{R}^d$, and $\mathcal{X}_d := \mathbb{S}^{d-1} \times \mathbb{S}^{d-1} \subseteq \mathbb{R}^{2d}$ to denote the Cartesian product of two hyperspheres. Given $\boldsymbol{x} \in \mathcal{X}_d$, we write $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ for the first and last d entries in $\boldsymbol{x}$, respectively. Throughout the remainder of the paper, we assume that the dimension parameter d is at least two. We use $\|\cdot\|_{L^2}$ for the L^2 norm over $\mathcal{X}_d$; that is, $\|f\|_{L^2}^2 = \mathbb{E}_{\boldsymbol{x}\sim\mathrm{Uniform}(\mathcal{X}_d)}[f(\boldsymbol{x})^2]$. Similarly, we use $\|\cdot\|_{L^\infty}$ for the L^∞ norm over $\mathcal{X}_d$. We write $\mathscr{D}_d$ for a distribution on $\mathcal{X}_d \times [-1, 1]$. We use the squared error loss and write $\mathscr{L}_{\mathscr{D}_d}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathscr{D}_d}[(f(\boldsymbol{x}) - y)^2]$ for the generalization error of a model f. Given a sample $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ of size m drawn i.i.d. from $\mathscr{D}_d$, we denote the sample loss as $\mathscr{L}_S(f) := \frac{1}{m}\sum_{i=1}^m (f(\boldsymbol{x}_i) - y_i)^2$.

2 Norm-Based Control in Infinite-Width Networks

In this work, we focus on the class of fully connected depth-L neural networks with ReLU activations, $2d$-dimensional inputs, and scalar output (or a *depth-L network*, for short). A depth-L network realizes a function $f_{\boldsymbol{\phi}} : \mathbb{R}^{2d} \to \mathbb{R}$ of the form:

$$f_{\boldsymbol{\phi}}(\boldsymbol{x}) = \boldsymbol{w}_L^\top[\boldsymbol{W}_{L-1}[\cdots[\boldsymbol{W}_2[\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1]_+ + \boldsymbol{b}_2]_+ \cdots]_+ + \boldsymbol{b}_{L-1}]_+ + b_L$$

where $\boldsymbol{\phi} := (\boldsymbol{W}_1, \boldsymbol{b}_1, \ldots, \boldsymbol{W}_{L-1}, \boldsymbol{b}_{L-1}, \boldsymbol{w}_L, b_L)$ denotes the collection of all weight matrices $\boldsymbol{W}_\ell \in \mathbb{R}^{\omega_\ell \times \omega_{\ell-1}}$, bias vectors $\boldsymbol{b}_\ell \in \mathbb{R}^{\omega_\ell}$, plus outer layer weights $\boldsymbol{w}_L \in \mathbb{R}^{\omega_{L-1}}$ and bias $b_L \in \mathbb{R}$, and $[\cdot]_+$ denotes the ReLU activation applied entrywise. Here, we allow the hidden-layer widths ω_ℓ for $\ell = 1, ..., L-1$ to be arbitrarily large.

Let Φ_L denote the collection of all parameter vectors $\boldsymbol{\phi}$ associated with a depth-L network, and define $\mathcal{N}_L = \{f_{\boldsymbol{\phi}} : \boldsymbol{\phi} \in \Phi_L\}$ to be the space of all functions realized by a depth-L network of unbounded width. Given a function $f \in \mathcal{N}_L$, we define its depth-L *representation cost* $R_L(f)$ by

$$R_L(f) = \inf_{\boldsymbol{\phi}\in\Phi_L : f = f_{\boldsymbol{\phi}}} \frac{\|\boldsymbol{\phi}\|^2}{L}. \tag{1}$$

where $\|\boldsymbol{\phi}\|^2$ denotes the sum of squares of all weights/biases in the network $f_{\boldsymbol{\phi}}$, and $f = f_{\boldsymbol{\phi}}$ indicates equality over the domain $\mathcal{X}_d$. More generally, following Savarese et al. (2019); Ongie et al. (2019), one can extend the definition of R_L

to a broader class of functions $f \in \mathcal{C}(\mathcal{X}_d)$ by

$$R_L(f) = \lim_{\epsilon \to 0} \inf \left\{ \frac{\|\boldsymbol{\phi}\|^2}{L} : \|f - f_{\boldsymbol{\phi}}\|_{L^\infty} \leq \epsilon, \boldsymbol{\phi} \in \Phi_L \right\} \tag{2}$$

where $R_L(f) = +\infty$ if the limit above does not exist. Any function with $R_L(f) < +\infty$ and $f \notin \mathcal{N}_L$ can be considered an "infinite-width" neural network, i.e., the uniform limit of a sequence depth L networks with unbounded width whose representation cost remains bounded. Since we focus on the representation cost needed to approximate functions, it suffices to consider networks whose width is unbounded but finite. In this case, the definition in (1) suffices.

The representation cost arises naturally when considering empirical risk minimization (ERM) with weight decay regularization:

$$\min_{\boldsymbol{\phi} \in \Phi_L} \mathscr{L}_S\left(f_{\boldsymbol{\phi}}\right) + \frac{\lambda}{L}\|\boldsymbol{\phi}\|^2, \tag{3}$$

where $\lambda > 0$ is a tunable regularization parameter. By fixing a function $f \in \mathcal{N}_L$ and optimizing over its parametrizations $f = f_{\boldsymbol{\phi}}$ as an L-layer network, we see that the above parameter space minimization problem is equivalent to the function space minimization problem

$$\min_{f \in \mathcal{N}_L} \mathscr{L}_S\left(f\right) + \lambda R_L(f). \tag{4}$$

In other words, the representation cost is the function space regularization penalty induced by imposing weight decay regularization in parameter space.

Remark 2.1. In Remark 5.4, we consider generalizations of our results to bounded-width networks. In that case, it is useful to consider the bounded-width version of the representation cost, which is the natural analog of the weight decay penalty in the function space $\mathcal{N}_{L,\omega}$ of functions realized by an L-layer network with the hidden-layer widths bounded by ω. In this case we write the representation cost as $R_L(f;\omega)$, and we formally define

$$R_L(f;\omega) := \inf_{\substack{\boldsymbol{\phi} \in \Phi_L : f = f_{\boldsymbol{\phi}} \\ \omega_1, \ldots, \omega_{L-1} \leq \omega}} \frac{\|\boldsymbol{\phi}\|^2}{L}. \tag{5}$$

To better understand the inductive bias of learning with weight decay regularization, several recent works have sought to give explicit function space characterizations of the representation cost $R_L(f)$. First, Savarese et al. (2019) showed that, for univariate functions, and assuming unregularized bias terms, $R_2(f)$ coincides with the L^1-norm of the second derivative of the function. This was generalized to multidimensional inputs ($d > 1$) by Ongie et al. (2019), where it is shown that $R_2(f)$ is equal to the L^1-norm of the Radon transform of a $(d+1)$-order derivative operator applied f. Related works have studied the impact of other activation functions (Parhi and Nowak, 2020), multi-dimensional outputs (Shenouda et al., 2023) and regularizing bias terms (Boursier and Flammarion, 2023). An ongoing effort is to characterize $R_L(f)$ with depth $L > 2$. For networks with multi-dimensional outputs, the limit as depth $L \to \infty$ is studied in (Jacot, 2022), where it is conjectured that the limiting representation cost coincides with the so-called "bottleneck rank" of a function, defined as the minimum r such that $f = g \circ h$ with $h : \mathbb{R}^{d_{\text{in}}} \to \mathbb{R}^r$ and $g : \mathbb{R}^r \to \mathbb{R}^{d_{\text{out}}}$. Finite depth modifications to this characterization are also studied by Jacot (2023).

3 Norm-Based Depth Separation in Approximation

Most previous depth separation results focus on separation in terms of the size of the network (i.e., the number of neurons) needed to represent or well-approximate a given target function. Specifically, Eldan and Shamir (2016); Daniely (2017); Safran and Shamir (2017) showed there are families of target functions parameterized by input dimension d that are well-approximated by a depth-3 network whose number of neurons is polynomial in d, but require width exponential in d to approximate within constant accuracy using a depth-2 network. For concreteness, we highlight the result from Daniely (2017):

Lemma 3.1 (Daniely (2017))**.** *There exists a family of functions $\{f_d\}_{d=1}^\infty \subset L^2(\mathcal{X}_d)$ such that any depth-two ReLU network $f_{\boldsymbol{\phi}} \in \mathcal{N}_2$ with $\|\boldsymbol{\phi}\|_\infty \leq 2^d$ satisfying $\|f_d - f_{\boldsymbol{\phi}}\|_{L^2} < 10^{-4}$ must have width $\omega = 2^{\Omega(d \log(d))}$. Conversely, for any $\epsilon > 0$, there exist a depth-three ReLU network $\tilde{f}_{\boldsymbol{\phi}} \in \mathcal{N}_3$ with $O(\text{poly}(d)/\epsilon)$ neurons and $\|\boldsymbol{\phi}\|_\infty = O(\text{poly}(d))$, such that $\|f_d - \tilde{f}_{\boldsymbol{\phi}}\|_{L^\infty} < \epsilon$.*

However, a width-based depth separation like the one above is not meaningful in the infinite-width setting. Instead, we consider whether a similar depth separation occurs in terms of the norm of the network (i.e., its representation cost). As a first result in this direction, Ongie et al. (2019) showed that there are functions in any input dimension d with finite R_3 representation cost but infinite R_2 representation cost, in the sense that any sequence of depth-2 networks converging pointwise to the target function on all of $\mathbb{R}^d$ must have unbounded representation cost. Yet, this left open whether there is still a depth separation in the representation costs required to *approximate* the target to a given accuracy on a bounded domain, and if so, its dependence on input dimension d. Here, we settle the question. In particular, we show the same function families that show depth separations in terms of width to approximate also demonstrate depth separations in terms of representation cost to approximate.

A key tool in moving from separation in terms of width to separation in terms of representation cost is the following lemma, which says that depth-2 neural networks of any width can be well approximated by narrow networks having essentially the same representation cost (i.e., up to a small constant factor). The proof follows essentially the same sampling argument as in Barron's universal approximation theorem for depth-2 networks (Barron, 1993); the details are given in Appendix A.2.

Lemma 3.2. *For any $f \in \mathcal{N}_2$, $\varepsilon > 0$, and width $\omega > \frac{3R_2(f)^2}{\varepsilon^2}$, there exists $f_{\boldsymbol{\phi}} \in \mathcal{N}_2$ having width ω and $\|\boldsymbol{\phi}\|_\infty^2 \leq \|\boldsymbol{\phi}\|_2^2 \leq 4R_2(f)$ such that $\|f - f_{\boldsymbol{\phi}}\|_{L^2} \leq \varepsilon$.*

Consider function families that we know require large widths to approximate with depth-2 networks, but can be well approximated with small width depth-3 networks with bounded weights. Functions in this family *must* have large R_2 cost; otherwise, Lemma 3.2 would imply they can be approximated with a small width. On the other hand, small width depth-3 networks with bounded weights must have low R_3 cost. In particular, a family of depth-3 networks whose width is $\mathrm{poly}(d)$ and weight magnitudes are $\mathrm{poly}(d)$ must have R_3 cost at most $\mathrm{poly}(d)$. Therefore, a depth separation in width to approximate should also imply a depth separation in representation cost to approximate.

Applying the above argument to the family of functions identified Lemma 3.1, we arrive at the following result, which is proved in Appendix A.2:

Corollary 3.3. *There exists a family of functions $\{f_d\}_{d=1}^\infty \subset L^2(\mathcal{X}_d)$ such that each f_d can be ε-approximated in L^∞-norm by a depth-three network $\tilde{f}_d \in \mathcal{N}_3$ with $R_3(\tilde{f}_d) = O(\mathrm{poly}(d)/\varepsilon)$, yet to approximate f_d by a depth-two network $\hat{f}_d \in \mathcal{N}_2$ to constant accuracy in L^2-norm requires $R_2(\hat{f}_d) = 2^{\Omega(d \log(d))}$.*

While mathematically interesting, this type of norm-based depth separation in approximation does not immediately imply anything about learning with norm-controlled networks, e.g., whether there is also a depth separation in the sample complexity needed for good generalization. In the remainder of this paper, we close this gap and show that a norm-based depth separation in approximation also implies a depth separation in sample complexity for norm-based learning rules.

4 Infinite-Width Norm-Based Learning Rules

We consider learning using the representational cost $R_L(f)$ as an inductive bias (i.e., complexity measure). Following the Structural Risk Minimization principle, we consider learning rules minimizing some combination of the empirical risk $\mathscr{L}_S(f)$ and the representational cost $R_L(f)$:

$$\min_{f \in \mathcal{N}_L} \left(\mathscr{L}_S(f), R_L(f) \right). \tag{6}$$

More specifically, we consider any learning rule returning a Pareto optimal point for the bi-criteria problem (6). This includes any minimizer of the regularized risk

$$\min_{f \in \mathcal{N}_L} \mathscr{L}_S(f) + \lambda R_L(f) \tag{7}$$

for any $\lambda > 0$, where recall that (7) is equivalent to seeking an unbounded width network and regularizing the norm of the weights, as in (3). We denote the set of all Pareto optimal points of (6) (i.e. the "Pareto frontier" or "regularization path", and including all minimizers of (7)—see Figure 1 for a visualization of the Pareto frontier and the learning rules

considered) by $\mathcal{P}_L(S)$. Similarly, we use $\mathcal{P}_{L,\omega}(S)$ to denote the Pareto frontier of the bounded-width version of this problem:

$$\min_{f \in \mathcal{N}_{L,\omega}} \left(\mathscr{L}_S\left(f\right), R_L(f;\omega)\right). \tag{8}$$

Our goal is to separate between learning rules returning depth-2 Pareto optimal points in $\mathcal{P}_2(S)$ and those returning depth-3 Pareto optional points in $\mathcal{P}_3(S)$. To make such a rule concrete, one still needs to choose *which* Pareto optimal point to return, e.g. choosing a value of λ in (7). In order to show separation, we compare the best possible depth-2 rule with a concrete depth-3 rule, showing that a concrete depth-3 rule "succeeds", but even the best possible depth-2 rule, and hence any rule returning a depth-2 Pareto optimal point, will "fail".

To obtain upper bounds (i.e., show learning is easy) we consider the following concrete rule, where the point on the frontier is specified by a threshold θ, as well as its finite-precision relaxations:

Definition 4.1. Given $\theta \geq 0$, define $\mathcal{A}_L^\theta$ to be a learning rule which, given training samples S, selects an L-layer network such that $\mathscr{L}_S\left(\mathcal{A}_L^\theta(S)\right) \leq \theta$ and

$$R_L(\mathcal{A}_L^\theta(S)) = \inf_{\substack{f \in \mathcal{N}_L \\ \mathscr{L}_S(f) \leq \theta}} R_L(f). \tag{9}$$

Given $\alpha \geq 1$, define $\mathcal{A}_L^{\theta,\alpha}$ to be a learning rule which selects an L-layer network such that $\mathscr{L}_S\left(\mathcal{A}_L^{\theta,\alpha}(S)\right) \leq \alpha\theta$ and

$$R_L(\mathcal{A}_L^{\theta,\alpha}(S)) \leq \alpha \inf_{\substack{f \in \mathcal{N}_L \\ \mathscr{L}_S(f) \leq \theta}} R_L(f). \tag{10}$$

Similarly, define a bounded width version $\mathcal{A}_{L,\omega}^{\theta,\alpha}$, to be a learning rule that selects an L-layer network of hidden width at most ω such that $\mathscr{L}_S\left(\mathcal{A}_{L,\omega}^{\theta,\alpha}(S)\right) \leq \alpha\theta$ and

$$R_L(\mathcal{A}_{L,\omega}^{\theta,\alpha}(S);\omega) \leq \alpha \inf_{\substack{f \in \mathcal{N}_{L,\omega} \\ \mathscr{L}_S(f) \leq \theta}} R_L(f;\omega). \tag{11}$$

The output of $\mathcal{A}_L^{\theta,\alpha}$ is α-close to $\mathcal{A}_L^\theta$, which lies on the Pareto frontier. However, we do not require exact Pareto optimality for $\mathcal{A}_L^{\theta,\alpha}$. See Figure 1 for a visualization of possible outputs of $\mathcal{A}_L^\theta$ and $\mathcal{A}_L^{\theta,\alpha}$ in relation to the Pareto frontier.

On the other hand, to prove lower bounds (i.e., argue learning is hard) we consider the following "ideal" rule, which "cheats" and chooses the Pareto optimal point minimizing the population error, and is thus better than any other rule returning Pareto optimal points:

Definition 4.2. We define $\mathcal{A}_L^*$ to be the learning rule which, given training samples S, selects the L-layer network that minimizes the population loss $\mathscr{L}_{\mathscr{D}_d}$ over the set $\mathcal{P}_L(S)$ of all Pareto optimal functions for the bicriterion minimization problem in Equation (6). That is, given training samples S,

$$\mathcal{A}_L^*(S) \in \operatorname*{arg\,min}_{f \in \mathcal{P}_L(S)} \mathscr{L}_{\mathscr{D}_d}(f). \tag{12}$$

Similarly, we define $\mathcal{A}_{L,\omega}^*$ to be the bounded-width version of this idealized rule;

$$\mathcal{A}_{L,\omega}^*(S) \in \operatorname*{arg\,min}_{f \in \mathcal{P}_{L,\omega}(S)} \mathscr{L}_{\mathscr{D}_d}(f). \tag{13}$$

Strictly speaking, $\mathcal{A}_L^*$ is not a learning rule because it depends on knowledge of the true target distribution instead of just samples from that distribution. It instead can be thought of as an oracle learning rule, based on side knowledge, and thus a lower bound on any learning rule returning Pareto optimal points in $\mathcal{P}_L(S)$.

Remark 4.3. For $L = 2$, Parhi and Nowak (2021); Unser (2023) show the infimum in Equation (9) is attained. For $L > 2$, it is an open question whether this infimum is attained. If it is not, one can choose a value of α arbitrarily close to 1 and consider $\mathcal{A}_L^{\theta,\alpha}$ instead of $\mathcal{A}_L^{\theta}$, for which our results still hold. For $\mathcal{A}_{L,\omega}^{\theta,\alpha}(S)$ to exist, we also need ω to be sufficiently large. For example, it suffices that ω is large enough for interpolation of the samples to be possible (see, e.g., Yun et al. (2019)). It is also possible that the argmins in Definition 4.2 are not attained. While we state the definition in terms of minimizing the population loss, our results hold even if $\mathcal{A}_L^*$ is replaced by any rule that outputs a function on the Pareto frontier.

In our main results, we equip Alice with $\mathcal{A}_2^*$ to give her the best possible choice of learning with a depth-2 network. However, we allow Bob to use the weaker learning rule $\mathcal{A}_3^{\theta}$ or even $\mathcal{A}_3^{\theta,\alpha}$.

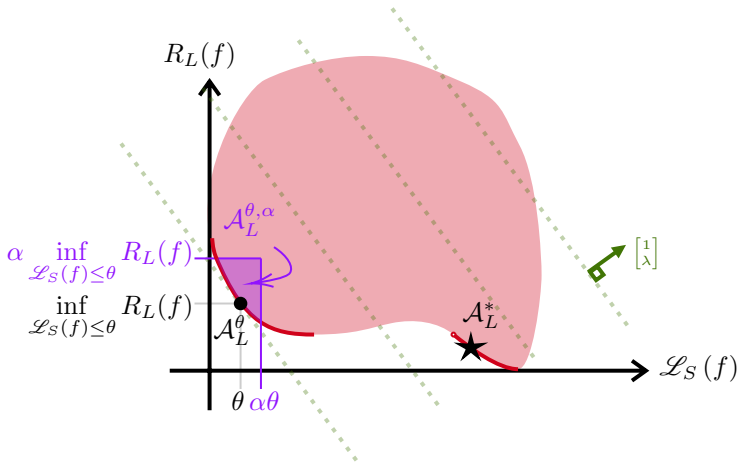

Figure 1: *Visualization of* $\mathcal{A}_L^{\theta}(S)$, $\mathcal{A}_L^{\theta,\alpha}(S)$, *and* $\mathcal{A}_L^*(S)$. The red shaded area represents the set of possible values of $(\mathscr{L}_S(f), R_L(f))$ where f is represented by an L-layer network. The red curves form the Pareto frontier $\mathcal{P}_L(S)$. Minimizing the population loss $\mathscr{L}_{\mathscr{D}_d}$ over the Pareto frontier yields $\mathcal{A}_L^*(S)$, represented by the star. In green is the vector $[1, \lambda]^\top$ and lines normal to it. These normal lines form level sets of $\mathscr{L}_S(f) + \lambda R_L(f)$. Notice the black dot on the Pareto frontier, which represents $\mathcal{A}_L^{\theta}(S)$. The output of $\mathcal{A}_L^{\theta}(S)$ corresponds to $\min_{f\in\mathcal{N}_L} \mathscr{L}_S(f) + \lambda R_L(f)$. The purple shaded region shows the possible outputs of $\mathcal{A}_L^{\theta,\alpha}(S)$, which are all α-close to $\mathcal{A}_L^{\theta}(S)$.

5 Main Results: Norm-Based Depth Separation in Learning

We now state our two main theorems. Theorem 5.1 says that there is a family of functions that $\mathcal{A}_3^{\theta}$ (i.e., Bob) can learn with sample complexity that is polynomial in d but $\mathcal{A}_2^*$ (i.e., Alice) needs the number of samples to grow exponentially with d in order to learn. Theorem 5.2 ensures that the reverse does not occur; families of distributions that $\mathcal{A}_2^*$ can learn with polynomial sample complexity can also be learned with polynomial sample complexity using $\mathcal{A}_3^{\theta}$. Both results still hold even when we relax Bob's depth-3 learning rule from $\mathcal{A}_3^{\theta}$ to $\mathcal{A}_3^{\theta,\alpha}$. For ease of presentation, we consider α to be a small constant, e.g., $\alpha = 2$.

Theorem 5.1 (Depth Separation in Learning)**.** *There is a family of distributions* $(\mathscr{D}_d)_{d=2}^{\infty}$ *on* $\mathcal{X}_d \times [-1, 1]$ *defined as* $\boldsymbol{x} \sim \mathrm{Uniform}(\mathcal{X}_d)$ *and* $y|\boldsymbol{x} = f_d(\boldsymbol{x})$ *for some function* $f_d : \mathcal{X}_d \to [-1, 1]$ *such that the following holds.*

1. *There are real numbers* $d_0 > 0$ *and* $C_1 > 0$*, such that if* $d > d_0$ *and* $|S| < 2^{C_1 d}$*, then* $\mathbb{E}_S[\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_2^*(S))] \ge 0.0001$.
2. *For all* $\varepsilon, \delta > 0$*, if* $\theta = \frac{\varepsilon}{4}$ *and* $|S| > O\left(\frac{d^{15}\log(1/\delta)}{\varepsilon^2}\right)$*, then* $\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_3^{\theta}(S)) \le \varepsilon$ *with probability at least* $1 - \delta$*. Furthermore, with a fixed constant* $\alpha \ge 1$*,* $\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_3^{\theta,\alpha}(S)) \le \varepsilon$ *with probability at least* $1 - \delta$ *where now the big-O suppresses a constant that depends on* α*.*

Theorem 5.2 (No Reverse Depth Separation in Learning)**.** *Consider a distribution $\mathscr{D}_d$ on $\mathcal{X}_d \times [-1,1]$ defined as $\boldsymbol{x} \sim \text{Uniform}(\mathcal{X}_d)$ and $y|\boldsymbol{x} = f_d(\boldsymbol{x})$ for some function $f_d : \mathcal{X}_d \to [-1,1]$. Assume that there is some sample complexity function $m_2(\varepsilon)$ such that $\mathbb{E}_S[\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_2^*(S))] \leq \varepsilon$ whenever $|S| \geq m_2(\varepsilon)$.*

For all $\varepsilon, \delta > 0$, if $\theta = \frac{\varepsilon}{4}$ and $|S| \geq m_3(\varepsilon, \delta)$, then $\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_3^{\theta}(S)) \leq \varepsilon$ with probability at least $1-\delta$, where the sample complexity m_3 is

$$m_3(\varepsilon, \delta) = O\left(\varepsilon^{-2}\left(d + m_2\left(\frac{\varepsilon}{64}\right)^{\frac{d+3}{d-1}}\right)^6 \log 1/\delta\right). \tag{14}$$

Furthermore, with a fixed constant $\alpha \geq 1$, $\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_3^{\theta,\alpha}(S)) \leq \varepsilon$ with probability at least $1-\delta$ where now the big-O suppresses a constant that depends on α.

In particular, if we have a family of such distributions $(\mathscr{D}_d)_{d=2}^{\infty}$ and m_2 grows polynomially with d, then m_3 also grows polynomially with d.

Remark 5.3. Theorems 5.1 and 5.2 are based on loose bounds. We conjecture that smaller sample complexities for depth-3 learning are possible in both results. Additionally, larger lower bounds on generalization for depth-2 learning are possible in Theorem 5.1.

Remark 5.4. We can generalize these results to networks of bounded widths. In Theorem 5.1, Part 1 holds for $\mathcal{A}_{2,\omega}^*$ as long as the width is at least three times the sample size, i.e., $\omega > 3|S|$. Thus, if the sample size is polynomial in d, then in sufficiently high dimensions, $\mathcal{A}_{2,\omega}^*$ cannot generalize without width that is super-polynomial in d. Part 2 holds for $\mathcal{A}_{3,\omega}^{\theta,\alpha}$ as long as $\omega \geq O\left(\varepsilon^{-1/2}d^{7/2}\right)$. That is, for depth-3 learning, we only require a width that is polynomial in dimension. To generalize Theorem 5.2 to bounded-width networks, we can modify the premise to the assumption that there is some minimal width function $\omega_0(\varepsilon)$ such that $\mathbb{E}_S[\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_{2,\omega}^*(S))] \leq \varepsilon$ whenever $|S| \geq m_2(\varepsilon)$ and $\omega \geq \omega_0(\varepsilon)$. The width ω required for $\mathcal{A}_{3,\omega}^{\theta,\alpha}$ to learn is then $\omega \geq O\left(\varepsilon^{-1}m_2\left(\frac{\varepsilon}{64}\right)^{\frac{2(d+3)}{d-1}} + d\right)$. If m_2 grows polynomially with d, then the width required for depth-3 learning is only polynomial in d.

Remark 5.5. The relatively restrictive assumptions on the distribution of $\boldsymbol{x}$ in Theorem 5.2 can be relaxed. We use these assumptions to bound the R_2 cost of interpolating samples. Our particular construction would be straightforward to generalize to other smooth distributions on $\mathcal{X}_d$ or $\mathbb{S}^{d-1}$. Other constructions could yield bounds on the R_2 cost of interpolating samples from other smooth distributions, which would allow for generalizations of this result.

6 Proof of Depth Separation in Learning

For the proof of Theorem 5.1, we use a slight modification of the construction from Daniely (2017). We choose $f_d(\boldsymbol{x}) := \psi_{3d}\left(\sqrt{d}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\rangle\right)$ where $\psi_n : \mathbb{R} \to [-1,1]$ denotes the sawtooth function that has n cycles in $[-1,1]$ and is equal to zero outside $[-1,1]$. See Figure 2 for a depiction of ψ_n. This target function is convenient for studying depth separation in norm because the sawtooth function can be represented exactly with one hidden ReLU layer, while the inner product can be approximated with another hidden ReLU layer. Thus, f_d lends itself well to explicit bounds on the R_3 representation cost needed to approximate it. Since f_d is a composition with an inner product, the framework in Daniely (2017) allows us to get a bound on the R_2 representation cost needed to approximate it as well. See Lemmas A.4 and A.11.

As with other depth-separation constructions, the Lipschitz constant of f_d is unbounded as d goes to infinity. Obtaining depth separation as d goes to infinity but with a bounded Lipschitz constant is a yet unsolved challenge; see Safran et al. (2019) for a discussion and evidence that current techniques cannot be used to show depth separation with a bounded Lipschitz constant.

In the following two subsections, we sketch the proofs of Parts 1 and 2 of Theorem 5.1.

6.1 Proof of Theorem 5.1, Part 1

Proof. Using Corollary 3.3, the construction of Daniely (2017) requires the R_2 cost to grow exponentially in d to approximate the target in the L_2 norm. We adapt this construction to f_d in Lemma A.4 to get more explicit bounds, from which we conclude that there exist real numbers $d_0, C > 0$ such that $d > d_0$ and $R_2(f) < 2^{Cd}$ implies that $\mathscr{L}_{\mathscr{D}_d}(f) \geq \frac{1}{50e^2\pi^2}$.

If $d \geq 3$ then by Lemma A.15, with probability at least $\frac{1}{2}$ there exists an interpolant $\hat{f} \in \mathcal{N}_2$ of the samples S with representation cost bounded as $R_2(\hat{f}) \leq 32\sqrt{2}|S|^{\frac{d+3}{d-1}}$. The proof of Lemma A.15 relies on the fact that with high probability the samples are sufficiently separated, and separated samples on $\mathcal{X}_d$ can be interpolated by a depth-2 neural network with small norm parameters. Similar ways to construct interpolants exist in other settings; see for example Section 5.2 in Ongie et al. (2019). Since $\mathcal{A}_2^*(S) \in \mathcal{P}_2(S)$ is Pareto optimal, we must have that $R_2(\mathcal{A}_2^*(S)) \leq R_2(\hat{f})$. Otherwise, $\mathcal{A}_2^*(S)$ would fail to be Pareto optimal because $\hat{f}$ would have a smaller sample loss and a smaller representation cost. It follows that $R_2(\mathcal{A}_2^*(S)) \leq 32\sqrt{2}|S|^{\frac{d+3}{d-1}}$ with probability at least $\frac{1}{2}$.

Choose $C_1 = C/4$. Assume $d > \max(d_0, 3, \frac{11}{2C_1})$ and $|S| < 2^{C_1 d}$. With probability at least $\frac{1}{2}$, we must have

$$R_2(\mathcal{A}_2^*(S)) \leq 32\sqrt{2}|S|^{\frac{d+3}{d-1}} < 2^{C_1 d} 2^{\frac{C_1 d(d+3)}{d-1}} \leq 2^{Cd}. \tag{15}$$

Thus, $\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_2^*(S)) \geq \frac{1}{50e^2\pi^2}$ with probability at least $\frac{1}{2}$. Therefore, by Markov's inequality, $\mathbb{E}_S[\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_2^*(S))] \geq \frac{1}{50e^2\pi^2} \cdot \frac{1}{2} \geq 10^{-4}$ □

6.2 Proof of Theorem 5.1, Part 2

We prove a slightly more general version of Part 2 in Theorem 5.1. Instead of just proving the result for $\mathcal{A}_3^\theta$ or $\mathcal{A}_3^{\theta,\alpha}$, we prove the result for the relaxed, bounded width learning rule $\mathcal{A}_{3,\omega}^{\theta,\alpha}$ for any $\alpha \geq 1$. This illuminates how the sample complexity and width we require to guarantee learning depends on α and ω.

Proof. Fix $\varepsilon, \delta > 0$ and $\alpha \geq 1$, and let $\theta = \frac{\varepsilon}{2\alpha}$. In Lemma A.11 we show that for all $K \in \mathbb{N}$ there is a depth-3 neural network $f_{d,K}$ of width $\omega_{d,K} := \max(6d+2, 2Kd)$ such that $\|f_d - f_{d,K}\|_{L^\infty} = O\left(\frac{d^{5/2}}{K}\right)$ and $R_3(f_{d,K}; \omega_{d,K}) = O(d^{5/2})$. Hence $\mathscr{L}_S(f_{d,K}) = O\left(\frac{d^5}{K^2}\right)$. We choose $K \geq O\left(\frac{d^{5/2}}{\sqrt{\theta}}\right)$ so that $\mathscr{L}_S(f_{d,K}) \leq \theta$. Now suppose that $\omega \geq \omega_{d,K}$. Then

$$R_3(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S)) \leq R_3(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S); \omega) \leq \alpha \inf_{\substack{g \in \mathcal{N}_{3,\omega} \\ \mathscr{L}_S(g) \leq \theta}} R_3(g; \omega) \tag{16}$$

$$\leq \alpha R_3(f_{d,K}; \omega) = O(\alpha d^{5/2}). \tag{17}$$

In Lemma A.19 we use the Rademacher complexity bounds from Neyshabur et al. (2015) to get the following estimation error bound on $f \in \mathcal{N}_3$ with respect to the target distribution $\mathscr{D}_d$; if $R_3(f) \leq M$, then $|\mathscr{L}_{\mathscr{D}_d}(f) - \mathscr{L}_S(f)| \leq O\left(M^3\sqrt{\frac{\log 1/\delta}{|S|}}\right)$ with probability at least $1-\delta$. Applying this, we get

$$\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S)) \leq \mathscr{L}_S\left(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S)\right) + |\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S)) - \mathscr{L}_S\left(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S)\right)| \tag{18}$$

$$\leq \alpha\theta + O\left(\sqrt{\frac{\alpha^6 d^{15} \log 1/\delta}{|S|}}\right) \tag{19}$$

with probability at least $1-\delta$. Therefore, with

$$|S| > O\left(\frac{\alpha^6 d^{15} \log 1/\delta}{\varepsilon^2}\right) \text{ and } \omega \geq \omega_{d,K} = O\left(\sqrt{\frac{\alpha d^7}{\varepsilon}}\right), \tag{20}$$

we get $\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S)) \leq \alpha\theta + \frac{\varepsilon}{2} = \varepsilon$ with probability at least $1-\delta$. □

7 Proof of No Reverse Depth Separation in Learning

To prove Theorem 5.2, we need the following lemma. Roughly speaking, this lemma says that if $\mathcal{A}_2^*(S)$ can learn with m_2 samples, then there is a good approximation of the target distribution that can be expressed as a depth-2 network with parameters whose norm is at most polynomial in m_2. The proof is a straightforward probabilistic argument, shown in Appendix A.7.1.

Lemma 7.1. *Consider a distribution $\mathscr{D}_d$ on $\mathcal{X}_d \times [-1,1]$ defined as $\boldsymbol{x} \sim \mathrm{Uniform}(\mathcal{X}_d)$ and $y|\boldsymbol{x} = f_d(\boldsymbol{x})$ for some function $f_d : \mathcal{X}_d \to [-1,1]$. Assume that there is some sample complexity function $m_2(\varepsilon)$ such that $\mathbb{E}_S[\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_2^*(S))] \leq \varepsilon$ whenever $|S| \geq m_2(\varepsilon)$. Then for any $\varepsilon > 0$, there is a function $f_\varepsilon \in \mathcal{N}_2$ such that $R_2(f_\varepsilon) \leq 100\sqrt{2} m_2\left(\frac{\varepsilon}{2}\right)^{\frac{d+3}{d-1}}$ and $\mathscr{L}_{\mathscr{D}_d}(f_\varepsilon) \leq \varepsilon$.*

Using the previous lemma and Lemma 3.2, the rest of the proof of Theorem 5.2 follows from the estimation error bound in Lemma A.19 derived from Rademacher complexity bounds and the fact that any function with small R_2-cost also has small R_3-cost. This fact is shown in Lemma A.2 by adding an identity layer. As in Section 6.2, we prove Theorem 5.2 for the relaxed, bounded width learning rule $\mathcal{A}_{3,\omega}^{\theta,\alpha}$ for any $\alpha \geq 1$ to showcase the role of α and ω, but this proof also applies to $\mathcal{A}_3^{\theta,\alpha}$ and $\mathcal{A}_3^{\theta}$. Full details are in Appendix A.7.2.

8 Conclusion

This paper demonstrates that there are functions that can be learned with depth-3 networks when the number of samples is polynomial in the input dimension d, but which cannot be learned with depth-2 networks unless the number of samples is exponential in d. Furthermore, we establish that in our setting, there are *no* functions that can easily be learned with depth-2 networks but which are difficult to learn with depth-3 networks. These results constitute the first depth separation result in terms of *learnability*, as opposed to network width.

In addition, the analysis framework we develop in this paper establishes a connection between width-based depth separation and learnability-based depth separation. As a result, our approach may be applied to other works on width-based depth separation to establish new learnability-based depth separation results.

We note that while the bounds developed in this paper are sufficient to establish our main results on depth separation, they may not be tight. For instance, the sample complexity bounds for depth-3 networks grow polynomially in d, but the polynomial order is quite large. Alternative constructions might lead to tighter bounds. Furthermore, the family of functions we use to establish our depth separation results does not have bounded Lipschitz constants; as d grows, our functions become highly oscillatory. Since highly oscillatory functions may not be representative of many practical predictors, it would be interesting to see whether there are families of functions *with bounded Lipschitz constants* leading to depth separation in terms of sample complexity (we note that (Safran and Shamir, 2017) studied this question but in the different context of width). A final potential limitation of our work is that it focuses on the output of learning rules seeking (approximately) Pareto optimal solutions, but neglects optimization dynamics. A major open question is how optimization dynamics affect depth separation.

9 Acknowledgements

S.P. was supported by the NSF Graduate Research Fellowship Program under Grant No. 2140001. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF. G.O. was supported by NSF CRII award CCF-2153371. R.W. and S.P. were supported by NSF DMS-2023109; R.W. was also supported by AFOSR FA9550-18-1-0166. O.S. was supported in part by European Research Council (ERC) grant 754705. N.S. was supported by the NSF/TRIPOD funded Institute of Data, Econometrics, Algorithms, and Learning (IDEAL), by the NSF/Simons funded Collaboration on the Theoretical Foundations of Deep Learning, and by NSF/IIS award 1764032.

References

Raman Arora, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. Understanding deep neural networks with rectified linear units. In *International Conference on Learning Representations*, 2018.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Peter Bartlett. For valid generalization the size of the weights is more important than the size of the network. *Advances in neural information processing systems*, 9, 1996.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. In *International Conference on Computational Learning Theory*, pages 224–240. Springer, 2001.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Yoshua Bengio, Nicolas Roux, Pascal Vincent, Olivier Delalleau, and Patrice Marcotte. Convex neural networks. *Advances in neural information processing systems*, 18, 2005.

Etienne Boursier and Nicolas Flammarion. Penalising the biases in norm regularisation enforces sparsity. 2023.

Vaggos Chatziafratis, Sai Ganesh Nagarajan, and Ioannis Panageas. Better depth-width trade-offs for neural networks through the lens of dynamical systems. In *International Conference on Machine Learning*, pages 1469–1478. PMLR, 2020.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Lenaic Chizat and Francis Bach. Implicit bias of gradient descent for wide two-layer neural networks trained with the logistic loss. In *Conference on Learning Theory*, pages 1305–1338. PMLR, 2020.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in neural information processing systems*, 22, 2009.

Amit Daniely. Depth separation for neural networks. In *Conference on Learning Theory*, pages 690–696. PMLR, 2017.

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

Simon Du and Jason Lee. On the power of over-parametrization in neural networks with quadratic activation. In *International conference on machine learning*, pages 1329–1338. PMLR, 2018.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pages 907–940. PMLR, 2016.

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

Stephen Hanson and Lorien Pratt. Comparing biases for minimal network construction with back-propagation. *Advances in neural information processing systems*, 1, 1988.

Arthur Jacot. Implicit bias of large depth networks: a notion of rank for nonlinear functions. In *The Eleventh International Conference on Learning Representations*, 2022.

Arthur Jacot. Bottleneck structure in learned features: Low-dimension vs regularity tradeoff. *arXiv preprint arXiv:2305.19008*, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Shiyu Liang and R Srikant. Why deep neural networks for function approximation? In *International Conference on Learning Representations*, 2016.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30, 2017.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. Norm-based capacity control in neural networks. In *Conference on learning theory*, pages 1376–1401. PMLR, 2015.

Behnam Neyshabur, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*, 2017.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

Rahul Parhi and Robert D Nowak. The role of neural network activation functions. *IEEE Signal Processing Letters*, 27: 1779–1783, 2020.

Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *The Journal of Machine Learning Research*, 22(1):1960–1999, 2021.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *International Conference on Machine Learning*, pages 7695–7705. PMLR, 2020.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta numerica*, 8:143–195, 1999.

David Rolnick and Max Tegmark. The power of deeper networks for expressing natural functions. In *International Conference on Learning Representations*, 2018.

Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *International conference on machine learning*, pages 2979–2987. PMLR, 2017.

Itay Safran, Ronen Eldan, and Ohad Shamir. Depth separations in neural networks: what is actually being separated? In *Conference on Learning Theory*, pages 2664–2666. PMLR, 2019.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

Joseph Shenouda, Rahul Parhi, Kangwook Lee, and Robert D Nowak. Vector-valued variation spaces and width bounds for dnns: Insights on weight decay regularization. *arXiv preprint arXiv:2305.16534*, 2023.

Panagiotis Sidiropoulos. N-sphere chord length distribution. *arXiv preprint arXiv:1411.5639*, 2014.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pages 1517–1539. PMLR, 2016.

Michael Unser. Ridges, neural networks, and the radon transform. *Journal of Machine Learning Research*, 24(37): 1–33, 2023.

Gal Vardi. On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6):86–93, 2023.

Gal Vardi and Ohad Shamir. Neural networks with small weights and depth-separation barriers. *arXiv preprint arXiv:2006.00625*, 2020.

Luca Venturi, Samy Jelassi, Tristan Ozuch, and Joan Bruna. Depth separation beyond radial functions. *The Journal of Machine Learning Research*, 23(1):5309–5364, 2022.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep relu networks. In *Conference on learning theory*, pages 639–649. PMLR, 2018.

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Small relu networks are powerful memorizers: a tight analysis of memorization capacity. *Advances in Neural Information Processing Systems*, 32, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. iclr 2017. *arXiv preprint arXiv:1611.03530*, 2017.

A Technical Lemmas & Proofs

Here, we present the technical details of the results in the main text.

A.1 Characterizing and bounding the representation cost

In this section, characterizations of and bounds on the representation cost that we use elsewhere in the appendix. To ease notation in this section, we re-label parameters defining a depth-2 network $f_{\boldsymbol{\phi}}$ as $\boldsymbol{\phi} = (\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{a}, c)$ so that $f_{\boldsymbol{\phi}}(\boldsymbol{x}) = \sum_{k=1}^{\omega_1} a_k [\boldsymbol{w}_k^\top \boldsymbol{x} + b_k]_+ + c$, where $\boldsymbol{w}_k^\top$ is the kth row of $\boldsymbol{W}$ and ω_1 is the width of the hidden layer in the parameterization.

The first result shows that the depth-2 representation cost reduces to the ℓ^1-norm of the outer-layer weights (plus half the squared outer-layer bias) assuming the first-layer weights/biases are normalized:

Lemma A.1. *Let $f \in \mathcal{N}_2$. Then*

$$R_2(f) = \inf_{\boldsymbol{\phi} \in \Phi_2 : f = f_{\boldsymbol{\phi}}} \sum_{k=1}^{\omega_1} |a_k| + \frac{c^2}{2} \;\; s.t. \;\; \|\boldsymbol{w}_k\|^2 + |b_k|^2 = 1 \; \forall k \in [\omega_1] \tag{21}$$

$$= \inf_{\boldsymbol{\phi} \in \Phi_2 : f = f_{\boldsymbol{\phi}}} \sum_{k=1}^{\omega_1} |a_k| \sqrt{\|\boldsymbol{w}_k\|^2 + |b_k|^2} + \frac{c^2}{2}. \tag{22}$$

Similarly, given $f \in \mathcal{N}_{2,\omega}$, we have a bounded width version of this:

$$R_2(f; \omega) = \inf_{\substack{\boldsymbol{\phi} \in \Phi_2 : f = f_{\boldsymbol{\phi}} \\ \omega_1 \leq \omega}} \sum_{k=1}^{\omega_1} |a_k| + \frac{c^2}{2} \;\; s.t. \;\; \|\boldsymbol{w}_k\|^2 + |b_k|^2 = 1 \; \forall k \in [\omega_1] \tag{23}$$

$$= \inf_{\substack{\boldsymbol{\phi} \in \Phi_2 : f = f_{\boldsymbol{\phi}} \\ \omega_1 \leq \omega}} \sum_{k=1}^{\omega_1} |a_k| \sqrt{\|\boldsymbol{w}_k\|^2 + |b_k|^2} + \frac{c^2}{2}. \tag{24}$$

We omit a full proof for brevity, but the result is a trivial modification of Lemma 1 in Appendix A of Savarese et al. (2019), extended to the case of regularized bias terms considered in this work. See also Boursier and Flammarion (2023).

The next result says that functions that have small representation costs with depth-2 networks also have small representation costs with depth-3 networks. The proof adds an identity layer to a depth-2 network to turn it into a depth-3 network.

Lemma A.2. *Given $f \in \mathcal{N}_{2,\omega}$, we have $f \in \mathcal{N}_{3,\max(\omega, 4d)}$ and*

$$R_3(f; \max(\omega, 4d)) \leq \frac{4d}{3} + \frac{4}{3} R_2(f; \omega). \tag{25}$$

Proof. Assume that $f \in \mathcal{N}_{2,\omega}$. Fix a particular parameterization $\boldsymbol{\phi} = (\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{a}, c)$ of f of width ω. Since $[\boldsymbol{x}]_+ - [-\boldsymbol{x}]_+ = \boldsymbol{x}$, we can rewrite f as a depth-3 neural network with an identity layer:

$$f(\boldsymbol{x}) = f_{\boldsymbol{\phi}}(\boldsymbol{x}) = \boldsymbol{a}^\top [\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}]_+ + c \tag{26}$$

$$= \boldsymbol{a}^\top \left[\begin{bmatrix} \boldsymbol{W} & -\boldsymbol{W} \end{bmatrix} \left[\begin{bmatrix} \boldsymbol{I}_{2d} \\ -\boldsymbol{I}_{2d} \end{bmatrix} \boldsymbol{x} \right]_+ + \boldsymbol{b} \right]_+ + c \tag{27}$$

$$= f_{\boldsymbol{\phi}'}(\boldsymbol{x}), \tag{28}$$

where $\boldsymbol{\phi}'$ is this new parameterization:

$$\boldsymbol{\phi}' = \left(\begin{bmatrix} \boldsymbol{I}_{2d} \\ -\boldsymbol{I}_{2d} \end{bmatrix}, \boldsymbol{0}, \begin{bmatrix} \boldsymbol{W} & -\boldsymbol{W} \end{bmatrix}, \boldsymbol{b}, \boldsymbol{a}, c \right). \tag{29}$$

Notice that $\boldsymbol{\phi}'$ has one hidden layer of width $4d$ and one hidden layer of width ω, so $f \in \mathcal{N}_{3,\max(\omega,4d)}$. Further,

$$\|\boldsymbol{\phi}'\|^2 = \left\| \begin{bmatrix} \boldsymbol{I}_{2d} \\ -\boldsymbol{I}_{2d} \end{bmatrix} \right\|_F^2 + \left\| \begin{bmatrix} \boldsymbol{W} & -\boldsymbol{W} \end{bmatrix} \right\|_F^2 + \|\boldsymbol{b}\|_2^2 + \|\boldsymbol{a}\|_2^2 + |c|^2 \tag{30}$$

$$= 4d + 2\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{b}\|_2^2 + \|\boldsymbol{a}\|_2^2 + |c|^2 \tag{31}$$

$$\leq 4d + 2\|\boldsymbol{\phi}\|^2. \tag{32}$$

Therefore,

$$R_3(f; \max(\omega, 4d)) = \inf_{\boldsymbol{\phi} \in \Phi_3 : f = f_{\boldsymbol{\phi}}} \frac{\|\boldsymbol{\phi}\|^2}{3} \tag{33}$$

$$\leq \inf_{\boldsymbol{\phi} \in \Phi_2 : f = f_{\boldsymbol{\phi}}} \frac{4d + 2\|\boldsymbol{\phi}\|^2}{3} \tag{34}$$

$$= \frac{4d}{3} + \frac{4}{3} R_2(f; \omega). \tag{35}$$

□

A.2 Approximating wide depth-2 networks by narrow networks with the same representation cost

A.2.1 Proof of Lemma 3.2

Before proving Lemma 3.2 we give an auxiliary result needed for the proof. The following is a simplified version of Lemma 1 from Barron (1993), originally credited to Maurey:

Lemma A.3 (Maurey's Lemma)**.** *Let H be a Hilbert space with norm $\|\cdot\|_H$. Assume $\mathcal{G} \subset H$ is such that $\|g\|_H \leq B$ for all $g \in \mathcal{G}$. Suppose f is a non-zero function belonging to the closed convex hull of $\mathcal{G}$. Then for any $m \in \mathbb{N}$ and there exists elements $g_1, ..., g_m \in \mathcal{G}$ such that*

$$\left\| f - \frac{1}{m} \sum_{k=1}^{m} g_k \right\|_H \leq \frac{B}{\sqrt{m}}.$$

We specialize this result to the Hilbert space $H = L^2(\mathcal{X}_d)$, and the subset $\mathcal{G} \subset L^2(\mathcal{X}_d)$ of all functions consisting of a single normalized ReLU unit. In particular, for any $\boldsymbol{w} \in \mathbb{R}^{2d}$ and $b \in \mathbb{R}$, define $u_{\boldsymbol{w},b}(\boldsymbol{x}) = [\boldsymbol{w}^\top \boldsymbol{x} + b]_+$ and let $\mathcal{G} \subset L^2(\mathcal{X}_d)$ be the set of functions

$$\mathcal{G} = \{\pm u_{\boldsymbol{w},b} : \boldsymbol{w} \in \mathbb{R}^{2d}, b \in \mathbb{R}, \|\boldsymbol{w}\|^2 + |b|^2 = 1\}.$$

Let μ_d denote the uniform probability measure on $\mathcal{X}_d = \mathbb{S}^{d-1} \times \mathbb{S}^{d-1}$. Note that for any $g = \pm u_{\boldsymbol{w},b} \in \mathcal{G}$ we have

$$\|g\|_{L^2}^2 = \int_{\mathcal{X}_d} [\boldsymbol{w}^\top \boldsymbol{x} + b]_+^2 d\mu_d(\boldsymbol{x}) \tag{36}$$

$$\leq \int_{\mathcal{X}_d} |\boldsymbol{w}^\top \boldsymbol{x} + b|^2 d\mu_d(\boldsymbol{x}) \tag{37}$$

$$= \int_{\mathcal{X}_d} \left| \begin{bmatrix} \boldsymbol{w} \\ b \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix} \right|^2 d\mu_d(\boldsymbol{x}) \tag{38}$$

$$\leq \int_{\mathcal{X}_d} (\|\boldsymbol{w}\|^2 + |b|^2)(1 + \|\boldsymbol{x}\|^2) d\mu_d(\boldsymbol{x}) \tag{39}$$

$$= 3 \int_{\mathcal{X}_d} d\mu_d(\boldsymbol{x}) = 3, \tag{40}$$

where we used the fact that $\|\boldsymbol{x}\|^2 = 2$ for all $\boldsymbol{x} \in \mathcal{X}_d$. Therefore, for $B = \sqrt{3}$ we have $\|g\|_{L^2} \leq B$ for all $g \in \mathcal{G}$.

Now we give the proof of Lemma 3.2:

Proof. Let $f \in \mathcal{N}_2$ and $\epsilon > 0$ be given, and suppose $\omega \in \mathbb{N}$ is such that $\omega > 3R_2(f)^2/\epsilon^2$. Choose δ with $0 < \delta \leq 1$ to be any constant satisfying $\omega \geq (1+\delta)^2 3R_2(f)^2/\epsilon^2$, and let $f(\boldsymbol{x}) = \sum_{k=1}^K a_k[\boldsymbol{w}_k^\top \boldsymbol{x} + b_k]_+ + c$ with $\|\boldsymbol{w}_k\|^2 + |b_k|^2 = 1$ for all $k \in [K]$ be any realization of f whose parameter cost is within a factor of $(1+\delta)$ of the infimum in Lemma A.1, i.e., $(1+\delta)R_2(f) \geq \sum_{k=1}^K |a_k| + \frac{c^2}{2}$. Let $A = \sum_{k=1}^K |a_k|$, and define $f_0 = (f-c)/A$. Then we can write $f_0(\boldsymbol{x}) = \sum_k \gamma_k s_k [\boldsymbol{w}_k^\top \boldsymbol{x} + b_k]_+$ where $s_k = \text{sign}(a_k)$ and $\gamma_k = |a_k|/A$ for all k. This shows f_0 is in the convex hull of $\mathcal{G}$, since $f_0 = \sum_k \gamma_k g_k$ with $g_k = s_k\, u_{\boldsymbol{w}_k, b_k} \in \mathcal{G}$ and $\gamma_k \geq 0$, $\sum_k \gamma_k = 1$.

Therefore, by Lemma A.3, there exists a function $\tilde{f}_0$ of the form $\tilde{f}_0(\boldsymbol{x}) = \frac{1}{\omega}\sum_{k=1}^\omega \tilde{s}_k[\tilde{\boldsymbol{w}}_k^\top \boldsymbol{x} + \tilde{b}_k]_+$ where $\|\tilde{\boldsymbol{w}}_k\|^2 + |\tilde{b}_k|^2 = 1$ and $\tilde{s}_k \in \{-1, 1\}$, such that

$$\|f_0 - \tilde{f}_0\|_{L^2} \leq \frac{\sqrt{3}}{\sqrt{\omega}} \leq \frac{\epsilon}{(1+\delta)R_2(f)}.$$

Multiplying both sides above by A gives

$$\|(f-c) - A\tilde{f}_0\|_{L^2} \leq \frac{A\epsilon}{(1+\delta)R_2(f)} \leq \frac{A\epsilon}{A + \frac{c^2}{2}} \leq \epsilon.$$

Defining $\tilde{f} = A\tilde{f}_0 + c$, we have

$$\|f - \tilde{f}\|_{L^2} \leq \epsilon,$$

where $\tilde{f}(\boldsymbol{x}) = \frac{1}{\omega}\sum_{k=1}^\omega s_k A[\tilde{\boldsymbol{w}}_k^\top \boldsymbol{x} + \tilde{b}_k]_+ + c$ is realizable as a depth-two ReLU network with width at most ω. In particular, with the choice of weights $\boldsymbol{\phi} = (\hat{\boldsymbol{W}}, \hat{\boldsymbol{b}}, \hat{\boldsymbol{a}}, c)$ with $\hat{\boldsymbol{w}}_k := \sqrt{A/\omega}\,\tilde{\boldsymbol{w}}_k$, $\hat{b}_k := b_k\sqrt{A/\omega}$, $\hat{a}_k := s_k\sqrt{A/\omega}$, for all $k \in [\omega]$, we have $\tilde{f} = f_{\boldsymbol{\phi}}$ with $\frac{\|\boldsymbol{\phi}\|_2^2}{2} = A + \frac{c^2}{2} \leq (1+\delta)R_2(f) \leq 2R_2(f)$, and so $\|\boldsymbol{\phi}\|_2^2 \leq 4R_2(f)$. Finally, the inequality $\|\boldsymbol{\phi}\|_\infty^2 \leq \|\boldsymbol{\phi}\|_2^2$ holds for any vector $\boldsymbol{\phi}$, which proves the claim. □

A.2.2 Proof of Corollary 3.3

Proof. Let f_d be the family of functions described in Lemma 3.1. First, to prove the depth-three result, set $\tilde{f}_d$ to be equal to the approximating function $\tilde{f}_{\boldsymbol{\phi}} \in \mathcal{N}_3$ described in Lemma 3.1. By a simple parameter count and the bounds on the magnitudes of weights, we are guaranteed that $R_3(\tilde{f}_{\boldsymbol{\phi}}) \leq \frac{\|\boldsymbol{\phi}\|^2}{3} = O(\text{poly}(d)/\varepsilon)$.

Now, we prove the depth-two result. Set $\varepsilon = 10^{-4}$. By way of contradiction, assume f_d can be $\varepsilon/2$-approximated in L^2-norm by a depth-two network $\hat{f}_d$ such that $R_2(\hat{f}_d)$ is subexponential in d. Then Lemma 3.2 implies $\hat{f}_d$ can be $\varepsilon/2$-approximated by a depth-two network $\tilde{f}_d \in \mathcal{N}_2$ with $R_2(\tilde{f}_d)$ subexponential in d, width ω subexponential in d, and weights uniformly bounded by 2^d for sufficiently large d. Hence, by the triangle inequality, f_d can be ε-approximated in L^2-norm by the depth-two network $\tilde{f}_d$ for all d. But by the width-based depth separation result Lemma 3.1, we know this is impossible since $\tilde{f}_d$ has width subexponential in d. Therefore, contrary to our assumption, it must be the case that $R_2(\hat{f}_d)$ is exponential in d. □

A.3 Approximating f_d in the L^2-norm requires exponential R_2 cost

In this section we adapt the construction of Daniely (2017) to the target function

$$f_d(\boldsymbol{x}) = \psi_{3d}\left(\sqrt{d}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\rangle\right) \tag{41}$$

to prove that approximating f_d in the L^2-norm to even constant error requires R_2 cost that is exponential in dimension:

Lemma A.4. *There exist real numbers $d_0, C > 0$ such that $d > d_0$ and $R_2(f) < 2^{Cd}$ implies that $\|f - f_d\|_{L^2}^2 \geq \frac{1}{50e^2\pi^2}$.*

After outlining the proof of this result, the remainder of this section establishes several auxiliary lemmas used in the proof.

Proof. Similar to Daniely (2017), we let μ_d denote the probability distribution obtained by pushing forward the uniform measure on $\mathbb{S}^{d-1}$ via the mapping $\boldsymbol{x} \mapsto x_1$, and we use $N_{d,n}$ for the dimension of the set of spherical harmonics of order n in d dimensions. Lemma A.5 adapts Theorem 4 in Daniely (2017) to show that for any $n \in \mathbb{N}$ and any $f \in \mathcal{N}_2$,

$$4\sqrt{3}R_2(f) + 2\|f\|_{L^2} \geq \sqrt{N_{d,n}} \left(A_{d,n}(\psi_{3d}(\sqrt{d}\cdot)) - \frac{\|f - f_d\|_{L^2}^2}{A_{d,n}(\psi_{3d}(\sqrt{d}\cdot))} \right) \tag{42}$$

where $A_{d,n}(\psi_{3d}(\sqrt{d}\cdot))$ is the distance in the $L^2(\mu_d)$-norm of the function $t \mapsto \psi_{3d}(\sqrt{d}t)$ to the closest polynomial of degree less than n.

We choose $n = 2d$. In Lemma A.6, we show that if d is sufficiently large, then $A_{d,2d}(\psi_{3d}(\sqrt{d}\cdot) \geq \frac{1}{5e\pi}$; that is, the sawtooth function is bounded away from being a polynomial of degree $2d-1$. If $\|f - f_d\|_{L^2}^2 < \frac{1}{50e^2\pi^2}$, then by the reverse triangle inequality

$$\|f\|_{L^2} < \|f_d\|_{L^2} + \|f - f_d\|_{L^2} \leq 1 + \frac{1}{5\sqrt{2}e\pi}. \tag{43}$$

Plugging this all into Equation (42), we get

$$4\sqrt{3}R_2(f) + 2 + \frac{2}{5\sqrt{2}e\pi} \geq \frac{\sqrt{N_{d,2d}}}{10e\pi} \tag{44}$$

whenever $\|f - f_d\|_{L^2}^2 < \frac{1}{50e^2\pi^2}$. As shown in Lemma A.7, $N_{d,2d} > 2^d$ for sufficiently large d. We conclude that there exist real numbers $d_0, C > 0$ such that $d > d_0$ and $R_2(f) < 2^{Cd}$ implies that $\|f - f_d\|_{L^2}^2 \geq \frac{1}{50e^2\pi^2}$. □

Lemma A.5. *Consider a distribution $\mathscr{D}_d$ on $\mathcal{X}_d \times [-1, 1]$ defined as*

$$\boldsymbol{x} \sim \text{Uniform}(\mathcal{X}_d) \tag{45}$$

$$y|\boldsymbol{x} = f_d(\boldsymbol{x}) \tag{46}$$

for some function inner-product $f_d : \mathcal{X}_d \to [-1, 1]$ defined as $f_d(\boldsymbol{x}) = g_d\left(\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle\right)$. Then for all $f \in \mathcal{N}_2$ and $n \in \mathbb{N}$,

$$\|f - f_d\|_{L^2}^2 \geq A_{d,n}(g_d) \left(A_{d,n}(g_d) - \frac{4\sqrt{3}R_2(f) + 2\|f\|_{L^2}}{\sqrt{N_{d,n}}} \right). \tag{47}$$

where $A_{d,n}(g_d)$ is the distance in the $L^2(\mu_d)$-norm of the function $t \mapsto g_d(t)$ to the closest polynomial of degree less than n.

Proof. Let $\boldsymbol{\phi} = (\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{a}, c)$ be an arbitrary parameterization of f with $\|\boldsymbol{w}_k\|_2^2 + |b_k|^2 = 1$ for each unit k. That is, $f(\boldsymbol{x}) = \sum_k a_k \left[\boldsymbol{w}_k^\top \boldsymbol{x} + b_k\right]_+ + c$. We now upper bound the L^2-norm of each ReLU unit in $\boldsymbol{\phi}$. By Cauchy-Schwarz, for all $\boldsymbol{x} \in \mathcal{X}_d$ we have

$$|\boldsymbol{w}_k^\top \boldsymbol{x} + b_k| \leq \sqrt{\|\boldsymbol{w}_k\|_2^2 + |b_k|^2}\sqrt{\|\boldsymbol{x}\|_2^2 + 1} = \sqrt{3}. \tag{48}$$

Thus

$$\left\| a_k \left[\boldsymbol{w}_k^\top \cdot + b_k\right]_+ \right\|_{L^2} = \sqrt{\mathbb{E}_{\boldsymbol{x} \sim \text{Uniform}(\mathcal{X}_d)} \left[a_k^2 \left[\boldsymbol{w}_k^\top \boldsymbol{x} + b_k\right]_+^2 \right]} \leq \sqrt{3}|a_k|. \tag{49}$$

Additionally,

$$\|c\|_{L^2} = \left\| f - \sum_k a_k \left[\boldsymbol{w}_k^\top \cdot + b_k\right]_+ \right\|_{L^2} \tag{50}$$

$$\leq \|f\|_{L^2} + \sum_k \left\| a_k \left[\boldsymbol{w}_k^\top \cdot + b_k\right]_+ \right\|_{L^2} \tag{51}$$

$$\leq \|f\|_{L^2} + \sqrt{3}\sum_k |a_k|. \tag{52}$$

By Theorem 4 in Daniely (2017),

$$\|f - f_d\|_{L^2}^2 \geq A_{d,n}(g_d) \left(A_{d,n}(g_d) - \frac{2\sum_k \left\| a_k \left[\boldsymbol{w}_k^\top \cdot + b_k \right]_+ \right\|_{L^2} + 2\|c\|_{L^2}}{\sqrt{N_{d,n}}} \right) \tag{53}$$

$$\geq A_{d,n}(g_d) \left(A_{d,n}(g_d) - \frac{2\|f\|_{L^2} + 4\sqrt{3}\sum_k |a_k|}{\sqrt{N_{d,n}}} \right) \tag{54}$$

We now take the supremum of the right-hand side of (54) over all such parameterizations $\boldsymbol{\phi}$. By Lemma A.1, this gives the desired result. □

The next lemma is analogous to (Daniely, 2017, Lemma 5) but for the sawtooth function instead of a sinusoid.

Lemma A.6. *If d is sufficiently large, then $A_{d,2d}(\psi_{3d}(\sqrt{d}\cdot) \geq \frac{1}{5e\pi}$.*

Proof. By definition,

$$A_{d,2d}(\psi_{3d}(\sqrt{d}\cdot) := \min_{p \in \mathbb{R}[x;2d-1]} \|\psi_{3d}(\sqrt{d}\,\cdot) - p\|_{L^2(\mu_d)} \tag{55}$$

where $\mathbb{R}[x;2d-1]$ denotes the set of polynomials of degree less than $2d$ and $d\mu_d(t) := \frac{\Gamma(\frac{d}{2})}{\sqrt{\pi}\Gamma(\frac{d-1}{2})}(1-t^2)^{\frac{d-3}{2}}$. As shown in the proof of (Daniely, 2017, Lemma 5), for $|t| \leq \frac{1}{\sqrt{d}}$ and d sufficiently large, we have $d\mu_d(t) \geq \frac{\sqrt{d}}{2e\pi}$, and for all $p \in \mathbb{R}[x;2d-1]$ and $n \geq 2d-1$,

$$\|\psi_n(\sqrt{d}\,\cdot) - p\|_{L^2(\mu_d)}^2 = \int_{-1}^{1} (\psi_n(\sqrt{d}t) - p(t))^2 d\mu_d(t) \tag{56}$$

$$\geq \frac{\sqrt{d}}{2e\pi} \int_{-d^{-1/2}}^{d^{-1/2}} (\psi_n(\sqrt{d}t) - p(t))^2 dt \tag{57}$$

$$= \frac{1}{2e\pi} \int_{-1}^{1} (\psi_n(t) - p(t/\sqrt{d}))^2 dt. \tag{58}$$

Consider the intervals $I_i = (-1 + \frac{2i-2}{n}, -1 + \frac{2i}{n})$, $i = 1, \ldots n$, of width $2/n$. Each interval contains a full cycle of the sawtooth function. Observe that $p(t/\sqrt{d})$ is a polynomial of degree at most $2d-1$, and so it has at most $2d-1$ roots in $[-1,1]$. On at least $n - 2d + 1$ of the intervals I_i, the polynomial $p(t/\sqrt{d})$ does not change signs. On each interval I_i where $p(t/\sqrt{d})$ does not change signs, ψ_n is positive on half of I_i and negative on the other half of I_i. Thus, on at least one subinterval of I_i of width $1/n$, $\psi_n(t)$ has the same sign as $p(t/\sqrt{d})$. It follows that

$$\int_{-1}^{1} (\psi_n(t) - p(t/\sqrt{d}))^2 dt \geq (n - 2d + 1) \int_0^{1/n} \psi_n^2(t) dt \tag{59}$$

$$= 2(n - 2d + 1) \int_0^{1/2n} (-2nt)^2 dt \tag{60}$$

$$= 2(n - 2d + 1)(2n)^2 \frac{1}{3(2n)^3} \tag{61}$$

$$= \frac{n - 2d + 1}{3n} \tag{62}$$

where the first equality comes from the symmetry in ψ_n. Thus $\|\psi_n(\sqrt{d}\,\cdot) - p\|_{L^2(\mu_d)}^2 \geq \frac{n-2d+1}{6ne\pi}$. In particular, choosing $n = 3d$ gives

$$A_{d,2d}(\psi_{3d}(\sqrt{d}\cdot)^2 \geq \frac{d+1}{18de\pi} \geq \frac{1}{18e\pi} \geq \frac{1}{25e^2\pi^2}. \tag{63}$$

□

Lemma A.7. *$N_{d,2d} > 2^d$ for sufficiently large d.*

Proof. The quantity $N_{d,n}$ is defined to be the dimension of the set of spherical harmonics of order n in d dimensions:

$$N_{d,n} := \frac{(2n+d-2)(n+d-3)!}{n!(d-2)!}. \tag{64}$$

Using Stirling's approximation,

$$\begin{aligned}
\lim_{d\to\infty} \frac{\log_2(N_{d,2d})}{d} &= \lim_{d\to\infty} \frac{\log_2(4d+d-2) + \log_2(2d+d-3)! - \log_2(2d)! - \log_2(d-2)!}{d} \\
&= \lim_{d\to\infty} \frac{(3d-3)\log_2(3d-3) - (2d)\log_2(2d) - (d-2)\log_2(d-2)}{d} \\
&> \lim_{d\to\infty} \frac{(3d-3)\log_2(2d) - (2d)\log_2(2d) - (d-2)\log_2(d)}{d} \\
&= \lim_{d\to\infty} \frac{d\log_2(2d) - d\log_2(d)}{d} = 1.
\end{aligned}$$

Therefore there exists a d_0 such that $d \geq d_0$ implies $\frac{\log_2(N_{d,2d})}{d} > 1$. □

A.4 Approximating f_d in the L^∞-norm with polynomial R_3 Cost

In this section, we show that there is a depth-3 network $f_{d,K}$ that well approximates $f_d(\boldsymbol{x}) = \psi_{3d}\left(\sqrt{d}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}\rangle\right)$ and bound its R_3 cost. The sawtooth function ψ_n can be expressed as a depth-2 network of width $2n+2$ as follows:

$$\psi_n(t) = -2n[t+1]_+ + 2n[t-1]_+ + 4n\sum_{j=1}^{n}(-1)^{j+n+1}\left[t - \frac{2j-1}{2n}\right]_+ + (-1)^{j+n}\left[t + \frac{2j-1}{2n}\right]_+. \tag{65}$$

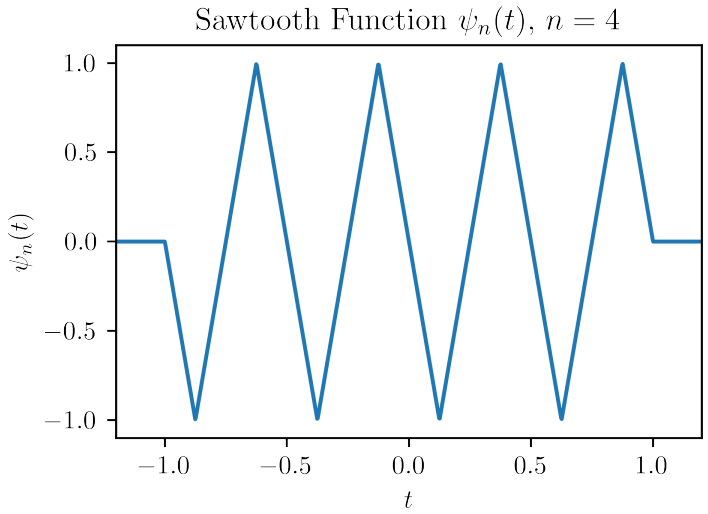

Figure 2: The sawtooth function $\psi_n : \mathbb{R} \to [-1, 1]$ with $n = 4$. The function ψ_n has n cycles in $[-1, 1]$ and is equal to zero outside $[-1, 1]$.

Lemma A.8. *For all scalars $\beta > 0$ and $n \in \mathbb{N}$, there are vectors $\boldsymbol{a}, \boldsymbol{u}, \boldsymbol{q} \in \mathbb{R}^{2n+2}$ such that $\psi_n(\beta t) = \boldsymbol{a}^\top[\boldsymbol{u}t + \boldsymbol{q}]_+$, $\boldsymbol{u}^\top\boldsymbol{q} = 0$, $\|\boldsymbol{u}\| = 1$, and $\|\boldsymbol{a}\|^2 + \|\boldsymbol{q}\|^2 = O(n^4\beta^2 + \beta^{-2})$.*

Proof. Denote the vector of all ones by $\mathbf{1}$. Using Equation (65), define vectors $\boldsymbol{a}_0, \boldsymbol{u}_0, \boldsymbol{q}_0 \in \mathbb{R}^{2n+2}$ so that $\psi_n(\beta t) = \boldsymbol{a}_0^\top [\boldsymbol{u}_0 t + \boldsymbol{q}_0]_+$ where $\boldsymbol{u}_0 = \beta \mathbf{1}$,

$$\boldsymbol{q}_0 = \begin{bmatrix} 1, & -1, & -\frac{1}{2n}, & \frac{1}{2n}, & -\frac{3}{2n}, & \frac{3}{2n}, & \cdots, & -\frac{2n-1}{2n}, & \frac{2n-1}{2n} \end{bmatrix}^\top, \tag{66}$$

and

$$\boldsymbol{a}_0 = \begin{bmatrix} -2n, & 2n, & \pm 4n, & \pm 4n, & \cdots, & \pm 4n, & \pm 4n \end{bmatrix}^\top. \tag{67}$$

Observe that

$$\|\boldsymbol{u}_0\|^2 = (2n+2)\beta^2 \tag{68}$$

$$\|\boldsymbol{q}_0\|^2 \leq (2n+2) \tag{69}$$

$$\|\boldsymbol{a}_0\|^2 \leq (2n+2)16n^2. \tag{70}$$

Let $\boldsymbol{u} = \frac{\boldsymbol{u}_0}{\|\boldsymbol{u}_0\|}$, $\boldsymbol{q} = \frac{\boldsymbol{q}_0}{\|\boldsymbol{u}_0\|}$, and $\boldsymbol{a} = \|\boldsymbol{u}_0\| \boldsymbol{a}_0$. Then $\psi_n(\beta t) = \boldsymbol{a}^\top [\boldsymbol{u} t + \boldsymbol{q}]_+$ by the homogeneity of ReLU. We also observe that $\|\boldsymbol{u}\| = 1$ and $\boldsymbol{u}^\top \boldsymbol{q} = 0$. Finally,

$$\|\boldsymbol{a}\|^2 + \|\boldsymbol{q}\|^2 \leq (2n+2)^2 16n^2\beta^2 + \beta^{-2} = O(n^4\beta^2 + \beta^{-2}). \tag{71}$$

□

The next two lemmas allow us to get an approximation of the inner product by approximating the square function.

Lemma A.9. *For all $s > 0$ and $K \in \mathbb{N}$, the function $f^{s,K}_{square}(t) := \frac{2s}{K} \sum_{k=1}^{K} [t - \frac{sk}{K}]_+ + [-t - \frac{sk}{K}]_+$ with $2K$ ReLU units satisfies*

$$\sup_{t \in [-s,s]} |f^{s,K}_{square}(t) - t^2| \leq s^2 \left(\frac{1}{K} + \frac{1}{K^2} \right).$$

Proof. Observe that $f^{s,K}_{square}(-t) = f^{s,K}_{square}(t)$, so it suffices to consider $t \in [0, s]$. Given $t \in [0, s]$, all of the $[-t - \frac{sk}{K}]_+$ terms in $f^{s,K}_{square}$ are equal to zero, and the $[t - \frac{sk}{K}]_+$ terms are nonzero if and only if $k < \frac{Kt}{s}$. That is,

$$f^{s,K}_{square}(t) = \frac{2s}{K} \sum_{k=1}^{\lfloor \frac{Kt}{s} \rfloor} \left(t - \frac{sk}{K} \right).$$

We use the summation formula $\sum_{j=1}^{n} j = \frac{n(n+1)}{2}$ and the notation $\{x\} := x - \lfloor x \rfloor \in [0, 1)$ to show that this quantity is approximately t^2; it is straightforward to verify that

$$f^{s,K}_{square}(t) = t^2 - \frac{st}{K} - \frac{s^2}{K^2} \left\{ \frac{Kt}{s} \right\} \left(\left\{ \frac{Kt}{s} \right\} - 1 \right).$$

Thus,

$$\sup_{t \in [-s,s]} |f^{s,K}_{square}(t) - t^2| = \sup_{t \in [0,s]} \left| \frac{st}{K} + \frac{s^2}{K^2} \left\{ \frac{Kt}{s} \right\} \left(\left\{ \frac{Kt}{s} \right\} - 1 \right) \right| \leq \frac{s^2}{K} + \frac{s^2}{K^2}. \tag{72}$$

□

Lemma A.10. *The function*

$$f^K_{inner}(\boldsymbol{x}) := \sum_{i=1}^{d} f^{\sqrt{2},K}_{square} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{e}_i \end{bmatrix}^\top \boldsymbol{x} \right) - 1 \tag{73}$$

satisfies

$$\sup_{\boldsymbol{x} \in \mathcal{X}_d} |f^K_{inner}(\boldsymbol{x}) - \langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle| \leq 2d \left(\frac{1}{K} + \frac{1}{K^2} \right). \tag{74}$$

Further, for any scalar $\beta > 0$, the function $\beta^{-1} f^K_{inner}(\boldsymbol{x})$ is in $\mathcal{N}_{2,2Kd}$ and

$$R_2(\beta^{-1} f^K_{inner}(\boldsymbol{x}); 2Kd) = O(d\beta^{-1} + \beta^{-2}). \tag{75}$$

Proof. Fix $\boldsymbol{x} \in \mathcal{X}_d$. Similarly to Corollary 7 in Daniely (2017), observe that

$$\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle = \sum_{i=1}^{d} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{e}_i \end{bmatrix}^\top \boldsymbol{x} \right)^2 - 1.$$

Additionally,

$$\left| \frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{e}_i \end{bmatrix}^\top \boldsymbol{x} \right| \leq \|\boldsymbol{x}\|_2 = \sqrt{2}.$$

Then

$$\begin{aligned} \sup_{\boldsymbol{x} \in \mathcal{X}_d} |f^K_{inner}(\boldsymbol{x}) - \langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle| &\leq \sup_{\boldsymbol{x} \in \mathcal{X}_d} \sum_{i=1}^{d} \left| f^{\sqrt{2},K}_{square} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{e}_i \end{bmatrix}^\top \boldsymbol{x} \right) - \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{e}_i \end{bmatrix}^\top \boldsymbol{x} \right)^2 \right| \\ &\leq d \sup_{|t| \leq \sqrt{2}} \left| f^{\sqrt{2},K}_{square}(t) - t^2 \right| \\ &\leq 2d \left(\frac{1}{K} + \frac{1}{K^2} \right). \end{aligned}$$

Now fix $\beta > 0$. Since

$$\frac{1}{\beta} f^K_{inner}(\boldsymbol{x}) = \frac{1}{\beta} \sum_{i=1}^{d} f^{\sqrt{2},K}_{square} \left(\frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{e}_i \end{bmatrix}^\top \boldsymbol{x} \right) - \frac{1}{\beta} \tag{76}$$

$$= \frac{2\sqrt{2}}{\beta K} \sum_{i=1}^{d} \sum_{k=1}^{K} \left[\frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{e}_i \end{bmatrix}^\top \boldsymbol{x} - \frac{\sqrt{2}k}{K} \right]_+ + \left[-\frac{1}{\sqrt{2}} \begin{bmatrix} \boldsymbol{e}_i \\ \boldsymbol{e}_i \end{bmatrix}^\top \boldsymbol{x} - \frac{\sqrt{2}k}{K} \right]_+ - \frac{1}{\beta} \tag{77}$$

we see that $\beta^{-1} f^K_{inner}(\boldsymbol{x}) \in \mathcal{N}_{2,2Kd}$. Finally, we apply Lemma A.1 to get

$$R_2(\beta^{-1} f^K_{inner}(\boldsymbol{x}); 2Kd) \leq \sum_{i=1}^{d} \sum_{k=1}^{K} \frac{4\sqrt{2}}{\beta K} \sqrt{1 + \frac{2k^2}{K^2}} + \frac{1}{2\beta^2} \tag{78}$$

$$\leq \frac{4\sqrt{3}d}{\beta} + \frac{1}{2\beta^2}. \tag{79}$$

□

Finally, we use f^K_{inner} to construct $f_{d,K}$ and bound its R_3 cost.

Lemma A.11. *Let $f_d(\boldsymbol{x}) = \psi_{3d}\left(\sqrt{d}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle\right)$. For all $K \in \mathbb{N}$, there is a depth-3 neural network $f_{d,K}$ of width $\omega_{d,K} := \max(6d + 2, 2Kd)$ such that $\|f_d - f_{d,K}\|_{L^\infty} = O\left(\frac{d^{5/2}}{K}\right)$ and $R_3(f_{d,K}; \omega_{d,K}) = O(d^{5/2})$.*

Proof. Choose $f_{d,K}(\boldsymbol{x}) := \psi_{3d}(\sqrt{d} f^K_{inner}(\boldsymbol{x}))$, which can be expressed as a depth-3 network with hidden widths $2Kd$ and $6d + 2$. For all $\boldsymbol{x} \in \mathcal{X}_d$, we use the fact that ψ_n is $2n$-Lipschitz to see that

$$\begin{aligned} \|f_d - f_{d,K}\|_{L^\infty} &= \sup_{\boldsymbol{x} \in \mathcal{X}_d} |\psi_{3d}(\sqrt{d} f^K_{inner}(\boldsymbol{x})) - \psi_{3d}(\sqrt{d}\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle)| \\ &\leq 6d\sqrt{d} \sup_{\boldsymbol{x} \in \mathcal{X}_d} |f^K_{inner}(\boldsymbol{x}) - \langle \boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \rangle| \\ &\leq 12 d^{5/2} \left(\frac{1}{K} + \frac{1}{K^2} \right). \end{aligned}$$

We now bound $R_3(f_{d,K};\omega_{d,K})$. Notice that $f_{d,K}$ can be expressed as $f_{d,K} = h \circ g$ where we set $h : \mathbb{R} \to \mathbb{R}$ to be $h(t) = \psi_{3d}(\sqrt{d}\beta t)$ and $g : \mathcal{X}_d \to \mathbb{R}$ to be $g(\boldsymbol{x}) = \beta^{-1} f^K_{inner}(\boldsymbol{x})$ where $\beta > 0$ is a value we will optimize over later. By Lemma A.8 there are vectors $\boldsymbol{a}, \boldsymbol{u}, \boldsymbol{q} \in \mathbb{R}^{2n+2}$ such that $h(t) = \boldsymbol{a}^\top[\boldsymbol{u}t + \boldsymbol{q}]_+$, $\boldsymbol{u}^\top \boldsymbol{q} = 0$, $\|\boldsymbol{u}\|_2 = 1$, and $\|\boldsymbol{a}\|_2^2 + \|\boldsymbol{q}\|_2^2 = O(d^5\beta^2 + \beta^{-2}d^{-1})$.

Let $\boldsymbol{\phi}_g = (\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{v}, c)$ be an arbitrary parameterization of g of width $2Kd$, so that $g(\boldsymbol{x}) = \boldsymbol{v}^\top[\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}]_+ + c$. This gives a parameterization $\boldsymbol{\phi}_f$ of $f_{d,K}$ as

$$f_{d,K}(\boldsymbol{x}) = \boldsymbol{a}^\top[\boldsymbol{u}\boldsymbol{v}^\top[\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}]_+ + (c\boldsymbol{u} + \boldsymbol{q})]_+.$$

Using the properties of $\boldsymbol{a}$, $\boldsymbol{u}$ and $\boldsymbol{q}$, we see that

$$\begin{aligned}\|\boldsymbol{\phi}_f\|^2 &= \|\boldsymbol{a}\|_2^2 + \|\boldsymbol{u}\|_2^2\|\boldsymbol{v}\|_2^2 + \|\boldsymbol{W}\|_F^2 + \|\boldsymbol{b}\|_2^2 + c^2\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{q}\|_2^2 &(80)\\ &= O(d^5\beta^2 + \beta^{-2}d^{-1}) + \|\boldsymbol{\phi}_g\|^2. &(81)\end{aligned}$$

Minimizing over parameterizations and using Lemma A.10, we get

$$\begin{aligned}R_3(f_{d,K};\omega_{d,K}) &\le O(d^5\beta^2 + \beta^{-2}d^{-1}) + \frac{2}{3}R_2(g; 2Kd) &(82)\\ &= O(d^5\beta^2 + \beta^{-2}d^{-1} + d\beta^{-1} + \beta^{-2}) &(83)\end{aligned}$$

Choosing $\beta = d^{-5/4}$ gives $R_3(f_{d,K};\omega_{d,K}) = O(d^{5/2})$. None of the constants hidden in the big-O depend on K. □

A.5 Existence of interpolants with mild R_2 cost

In this section, we will prove that with high probability over the samples, an interpolant exists with R_2 cost that depends only mildly on the number of samples (Lemma A.15). To do this, we show that with high probability the samples are sufficiently separated (Lemma A.13), and then show that separated samples on $\mathcal{X}_d$ can each be assigned a hyperplane that is sufficiently far away from any other sample (Lemma A.14). We start with the following simple bound on the Beta function.

Lemma A.12. *For all $d \ge 3$,*

$$B\left(\frac{d-1}{2}, \frac{1}{2}\right) \ge \frac{2\sqrt{\pi}}{d-1}. \tag{84}$$

Proof. Using the identity $z\Gamma(z) = \Gamma(z+1)$ and the fact that $\Gamma(z)$ is an increasing function on the domain $z \ge \frac{3}{2}$, we see that

$$(d-1)B\left(\frac{d-1}{2}, \frac{1}{2}\right) = 2\frac{\frac{d-1}{2}\Gamma(\frac{d-1}{2})\Gamma(\frac{1}{2})}{\Gamma(\frac{d}{2})} = 2\frac{\Gamma(\frac{d+1}{2})\sqrt{\pi}}{\Gamma(\frac{d}{2})} \ge 2\sqrt{\pi}. \tag{85}$$

□

Lemma A.13. *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ be i.i.d. samples from* $\mathrm{Uniform}(\mathcal{X}_d)$*. Then for $\eta < 1$,*

$$\mathbb{P}(\min_{i \ne j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \le \eta) < m^2\eta^{d-1}$$

Proof. We first consider the distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Since $\|\boldsymbol{x}_1^{(1)} - \boldsymbol{x}_2^{(1)}\|_2 \le \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$ it follows that

$$\mathbb{P}(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \le \eta) \le \mathbb{P}(\|\boldsymbol{x}_1^{(1)} - \boldsymbol{x}_2^{(1)}\|_2 \le \eta). \tag{86}$$

As shown in Sidiropoulos (2014)), the probability density function of $\|\boldsymbol{x}_1^{(1)} - \boldsymbol{x}_2^{(1)}\|_2$ is

$$\mathbb{P}(\|\boldsymbol{x}_1^{(1)} - \boldsymbol{x}_2^{(1)}\|_2 = \eta) = \frac{\eta\left(\eta^2 - \frac{\eta^4}{4}\right)^{\frac{d-3}{2}}}{B\left(\frac{d-1}{2}, \frac{1}{2}\right)}. \tag{87}$$

Integrating and using the bound on the Beta function from Lemma A.12, we get

$$\mathbb{P}(\|\boldsymbol{x}_1^{(1)} - \boldsymbol{x}_2^{(1)}\|_2 \leq \eta) = \frac{1}{B\left(\frac{d-1}{2}, \frac{1}{2}\right)} \int_0^\eta t \left(t^2 - \frac{t^4}{4}\right)^{\frac{d-3}{2}} dt \tag{88}$$

$$\leq \frac{d-1}{2\sqrt{\pi}} \int_0^\eta t \left(t^2\right)^{\frac{d-3}{2}} dt \tag{89}$$

$$= \frac{d-1}{2\sqrt{\pi}} \int_0^\eta t^{d-2} dt \tag{90}$$

$$< \eta^{d-1}. \tag{91}$$

Finally, there are $\binom{m}{2}$ pairwise distances between the samples, so we can use the union bound to get

$$\mathbb{P}(\min_{i \neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \leq \eta) < \binom{m}{2} \mathbb{P}(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \leq \eta) < m^2 \eta^{d-1}. \tag{92}$$

□

Lemma A.14. *For any finite set of points $\{\boldsymbol{x}_j\}_{j=1}^m \subseteq \mathcal{X}_d$ that are η-separated, there exists a unit vector $\boldsymbol{v}_j \in \mathbb{R}^{2d}$ for all $j \in [m]$ such that $\boldsymbol{x}_j$ is contained in the hyperplane $\{\boldsymbol{x} \in \mathbb{R}^{2d} : \boldsymbol{v}_j^\top \boldsymbol{x} = \sqrt{2}\}$ and $\boldsymbol{x}_j$ is the only point contained in the set $T_j := \{\boldsymbol{x} \in \mathbb{R}^{2d} : |\boldsymbol{v}_j^\top \boldsymbol{x} - \sqrt{2}| < \frac{\eta^2}{2\sqrt{2}}\}$.*

Proof. Assume $\{\boldsymbol{x}_j\}_{j=1}^m \subseteq \mathcal{X}_d$ and $\min_{i \neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 \geq \eta$. Choose $\boldsymbol{v}_j = \frac{1}{\sqrt{2}} \boldsymbol{x}_j$. Clearly $\|\boldsymbol{v}_j\|_2 = 1$, and

$$\boldsymbol{v}_j^\top \boldsymbol{x}_j = \frac{1}{\sqrt{2}} \|\boldsymbol{x}_j\|^2 = \sqrt{2}. \tag{93}$$

If $i \neq j$, then observe that

$$\eta^2 \leq \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 = \|\boldsymbol{x}_i\|^2 + \|\boldsymbol{x}_j\|^2 - 2\boldsymbol{x}_i^\top \boldsymbol{x}_j = 4 - 2\boldsymbol{x}_i^\top \boldsymbol{x}_j. \tag{94}$$

Hence,

$$|\boldsymbol{v}_j^\top \boldsymbol{x}_i - \sqrt{2}| = \left| \frac{1}{\sqrt{2}} \boldsymbol{x}_j^\top \boldsymbol{x}_i - \sqrt{2} \right| \geq \frac{\eta^2}{2\sqrt{2}}. \tag{95}$$

□

We now have the pieces we need for the proof of Lemma A.15.

Lemma A.15. *Consider a distribution $\mathscr{D}_d$ on $\mathcal{X}_d \times [-1, 1]$ defined as*

$$\boldsymbol{x} \sim \text{Uniform}(\mathcal{X}_d) \tag{96}$$

$$y|\boldsymbol{x} = f_d(\boldsymbol{x}) \tag{97}$$

for some function $f_d : \mathcal{X}_d \to [-1, 1]$. Given a sample $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$ of size m drawn i.i.d. from $\mathscr{D}_d$, with probability at least $1 - \delta$ there exists an interpolant $\hat{f}$ of S such that $R_2(\hat{f}) \leq 16\sqrt{2}|S|^{\frac{d+3}{d-1}} \delta^{-\frac{2}{d-1}}$.

Proof. By Lemma A.13, the data is $\delta^{\frac{1}{d-1}}|S|^{\frac{-2}{d-1}}$ separated with probability at least $1 - \delta$. For convenience, let $\eta = \delta^{\frac{1}{d-1}}|S|^{\frac{-2}{d-1}}$ and $\eta_0 = \frac{\eta^2}{2\sqrt{2}}$. Note that $\eta, \eta_0 \in (0, 1)$.

Consider the function $z_{\eta_0} : \mathbb{R} \to \mathbb{R}$ defined by $z_{\eta_0}(t) = \eta_0^{-1}([t - \eta_0]_+ - 2[t]_+ + [t + \eta_0]_+)$, which vanishes for $|t| > \eta_0$, and is such that $z_{\eta_0}(0) = 1$. By Lemma A.14, for all $j \in [n]$ there exists a unit vector $\boldsymbol{v}_j \in \mathbb{R}^{2d}$ for all $j \in [n]$ such that $\boldsymbol{x}_j$ is contained in the hyperplane $\{\boldsymbol{x} \in \mathbb{R}^{2d} : \boldsymbol{v}_j^\top \boldsymbol{x} = \sqrt{2}\}$ and $\boldsymbol{x}_j$ is the only training point contained in the

set $T_j := \{\boldsymbol{x} \in \mathbb{R}^{2d} : |\boldsymbol{v}_j^\top \boldsymbol{x} - \sqrt{2}| < \eta_0\}$. Define the ridge function $r_j : \mathbb{R}^{2d} \to \mathbb{R}$ by the depth-2 network of width 3 as follows:

$$r_j(\boldsymbol{x}) = z_{\eta_0}(\boldsymbol{v}_j^\top \boldsymbol{x} - \sqrt{2}) = \eta_0^{-1}([\boldsymbol{v}_j^\top \boldsymbol{x} - \sqrt{2} - \eta_0]_+ - 2[\boldsymbol{v}_j^\top \boldsymbol{x} - \sqrt{2}]_+ + [\boldsymbol{v}_j^\top \boldsymbol{x} - \sqrt{2} + \eta_0]_+). \tag{98}$$

Since the support of r_j coincides with T_j, and $\boldsymbol{v}_j^\top \boldsymbol{x}_j - \sqrt{2} = 0$, we see that $r_j(\boldsymbol{x}_i) = \delta_{ij}$. Therefore, the width $3|S|$, depth-2 network $\hat{f}(\boldsymbol{x}) = \sum_{j=1}^{|S|} y_j r_j(\boldsymbol{x})$ interpolates the samples.

Using Lemma A.1,

$$\begin{aligned} R_2\left(\hat{f}; 3|S|\right) &\le \sum_{j=1}^{|S|} |y_j| \eta_0^{-1} \left(\sqrt{1 + (\sqrt{2} + \eta_0)^2} + 2\sqrt{3} + \sqrt{1 + (-\sqrt{2} + \eta_0)^2}\right) && (99) \\ &\le 8|S|\eta_0^{-1} && (100) \\ &= 16\sqrt{2}|S|^{\frac{d+3}{d-1}} \delta^{-\frac{2}{d-1}}. && (101) \end{aligned}$$

□

A.6 Estimation error bound for depth-3 networks

In this section, we present an estimation error bound (Lemma A.19) derived from the Rademacher complexity bounds in Neyshabur et al. (2015). We begin with several auxiliary lemmas. Given a depth-3 network $f_{\boldsymbol{\phi}} \in \mathcal{N}_3$, this first lemma rewrites $f_{\boldsymbol{\phi}}$ so that it will be compatible with the framework in Neyshabur et al. (2015).

Lemma A.16. *If $\boldsymbol{\phi} = (\boldsymbol{W}_1, \boldsymbol{b}_1, \boldsymbol{W}_2, \boldsymbol{b}_2, \boldsymbol{w}_3, b_3)$ and $\frac{1}{3}\|\boldsymbol{\phi}\|^2 \le M$, then*

$$f_{\boldsymbol{\phi}}(\boldsymbol{x}) = \begin{bmatrix} \boldsymbol{w}_3^\top & b_3 \end{bmatrix} \left[\begin{bmatrix} \boldsymbol{W}_2^\top & \boldsymbol{b}_2 \\ \boldsymbol{0} & 1 \end{bmatrix} \left[\begin{bmatrix} \boldsymbol{W}_1^\top & \boldsymbol{b}_1 \\ \boldsymbol{0} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix} \right]_+ \right]_+ \tag{102}$$

with

$$\left\| \begin{bmatrix} \boldsymbol{w}_3^\top & b_3 \end{bmatrix} \right\|_2 \left\| \begin{bmatrix} \boldsymbol{W}_2^\top & \boldsymbol{b}_2 \\ \boldsymbol{0} & 1 \end{bmatrix} \right\|_F \left\| \begin{bmatrix} \boldsymbol{W}_1^\top & \boldsymbol{b}_1 \\ \boldsymbol{0} & 1 \end{bmatrix} \right\|_F \le \left(M + \frac{2}{3}\right)^{3/2}.$$

Proof. It is straightforward to verify Equation (102). Observe that

$$\begin{aligned} M \ge \frac{1}{3}\|\boldsymbol{\phi}\|^2 &= \frac{1}{3}\left(\left\| \begin{bmatrix} \boldsymbol{w}_3^\top & b_3 \end{bmatrix} \right\|_2^2 + \left\| \begin{bmatrix} \boldsymbol{W}_2^\top & \boldsymbol{b}_2 \\ \boldsymbol{0} & 1 \end{bmatrix} \right\|_F^2 + \left\| \begin{bmatrix} \boldsymbol{W}_1^\top & \boldsymbol{b}_1 \\ \boldsymbol{0} & 1 \end{bmatrix} \right\|_F^2 - 2 \right) \\ &\ge -\frac{2}{3} + \left(\left\| \begin{bmatrix} \boldsymbol{w}_3^\top & b_3 \end{bmatrix} \right\|_2 \left\| \begin{bmatrix} \boldsymbol{W}_2^\top & \boldsymbol{b}_2 \\ \boldsymbol{0} & 1 \end{bmatrix} \right\|_F \left\| \begin{bmatrix} \boldsymbol{W}_1^\top & \boldsymbol{b}_1 \\ \boldsymbol{0} & 1 \end{bmatrix} \right\|_F \right)^{2/3}. \end{aligned}$$

where the second inequality comes from the AM-GM inequality. □

We now apply Theorem 1 in Neyshabur et al. (2015) to get a bound on the Rademacher complexity of the set of depth-3 networks with representation cost bounded by M with respect to $\mathrm{Uniform}(\mathcal{X}_d)$. We use $\mathcal{N}_3^M$ to denote this set:

$$\mathcal{N}_3^M := \{f \in \mathcal{N}_3 : R_3(f) \le M\}. \tag{103}$$

Given a function class $\mathcal{H}$, we write $\mathscr{R}_m(\mathcal{H}; (\boldsymbol{x}_i)_{i=1}^m)$ for the empirical Rademacher complexity with respect to samples $(\boldsymbol{x}_i)_{i=1}^m$. That is,

$$\mathscr{R}_m(\mathcal{H}; (\boldsymbol{x}_i)_{i=1}^m) := \mathbb{E}_{\xi \sim \{\pm 1\}^m} \left[\sup_{h \in \mathcal{H}} \frac{1}{m} \left| \sum_{i=1}^m \xi_i h(\boldsymbol{x}_i) \right| \right] \tag{104}$$

where $\xi \sim \{\pm 1\}^m$ denotes that each entry in ξ is an iid draw from $\mathrm{Uniform}\{\pm 1\}$. We write $\mathscr{R}_{\mathcal{X}_d^m}(\mathcal{H})$ for the Rademacher complexity of $\mathcal{H}$ with respect to m i.i.d. samples from $\mathrm{Uniform}(\mathcal{X}_d)$:

$$\mathscr{R}_{\mathcal{X}_d^m}(\mathcal{H}) := \mathbb{E}_{(\boldsymbol{x}_i)_{i=1}^m \overset{iid}{\sim} \mathrm{Uniform}(\mathcal{X}_d)}[\mathscr{R}_m(\mathcal{H}; (\boldsymbol{x}_i)_{i=1}^m]. \tag{105}$$

Lemma A.17 (Rademacher Complexity Bound)**.** $\mathscr{R}_{\mathcal{X}_d^m}(\mathcal{N}_3^M) = O\left(\frac{M^{3/2}}{m^{1/2}}\right)$.

Proof. Theorem 1 in Neyshabur et al. (2015) bounds the empirical Rademacher complexity of

$$\mathcal{N}_{\gamma_{22}\leq\gamma}^{3,\dim=D} := \{f : \mathbb{R}^D \to \mathbb{R} | f(\boldsymbol{x}) = \boldsymbol{w}_3^\top \left[\boldsymbol{W}_2 \left[\boldsymbol{W}_1 \boldsymbol{x}\right]_+\right]_+, \|\boldsymbol{w}_3\|_2 \|\boldsymbol{W}_2\|_F \|\boldsymbol{W}_1\|_F \leq \gamma\} \tag{106}$$

as

$$\mathscr{R}_m(\mathcal{N}_{\gamma_{22}\leq\gamma}^{3,\dim=D}; (\boldsymbol{x}_i)_{i=1}^m) \leq \frac{4\sqrt{2}\gamma \max_i \|\boldsymbol{x}_i\|_2}{\sqrt{m}}. \tag{107}$$

By Lemma A.16, $\mathcal{N}_3^M \subseteq \mathcal{N}_{\gamma_{22}\leq\gamma}^{3,\dim=D}$ with $D = 2d+1$ and $\gamma = \left(M + \frac{2}{3}\right)^{3/2}$. Therefore,

$$\mathscr{R}_m(\mathcal{N}_3^M; (\boldsymbol{x}_i)_{i=1}^m) \leq \frac{4\sqrt{2}\left(M + \frac{2}{3}\right)^{3/2} \max_i \sqrt{1 + \|\boldsymbol{x}_i\|_2^2}}{\sqrt{m}}.$$

where we have replaced $\|\boldsymbol{x}_i\|_2$ with $\sqrt{1 + \|\boldsymbol{x}_i\|_2^2}$ because $\mathcal{N}_3^M$ is embedded in $\mathcal{N}_{\gamma_{22}\leq\gamma}^{3,\dim=D}$ by extending in the input $\boldsymbol{x} \in \mathbb{R}^{2d}$ to $\begin{bmatrix}\boldsymbol{x}^\top & 1\end{bmatrix}^\top \in \mathbb{R}^{2d+1}$. Since all samples $\boldsymbol{x}_i \sim \mathrm{Uniform}(\mathcal{X}_d)$ have norm $\sqrt{2}$, we get

$$\mathscr{R}_{\mathcal{X}_d^m}(\mathcal{N}_3^M) \leq \frac{4\sqrt{2}\left(M + \frac{2}{3}\right)^{3/2}\sqrt{3}}{\sqrt{m}} = O\left(\frac{M^{3/2}}{m^{1/2}}\right).$$

□

The other piece we need for an estimation error bound is to uniformly bound $\|f_d - h\|_{L^\infty}$ over $\mathcal{N}_3^M$.

Lemma A.18. *If $f_d : \mathcal{X}_d \to [-1, 1]$, then* $\sup_{h\in\mathcal{N}_3^M} \|f_d - h\|_{L^\infty} = O(M^{3/2})$.

Proof. If $h \in \mathcal{N}_3^M$, then by Lemma A.16,

$$h(\boldsymbol{x}) = \begin{bmatrix}\boldsymbol{w}_3^\top & b_3\end{bmatrix}\left[\begin{bmatrix}\boldsymbol{W}_2^\top & \boldsymbol{b}_2 \\ \boldsymbol{0} & 1\end{bmatrix}\left[\begin{bmatrix}\boldsymbol{W}_1^\top & \boldsymbol{b}_1 \\ \boldsymbol{0} & 1\end{bmatrix}\begin{bmatrix}\boldsymbol{x} \\ 1\end{bmatrix}\right]_+\right]_+$$

for some parameterization $\boldsymbol{\phi} = (\boldsymbol{W}_1, \boldsymbol{b}_1, \boldsymbol{W}_2, \boldsymbol{b}_2, \boldsymbol{w}_3, b_3)$ with

$$\left\|\begin{bmatrix}\boldsymbol{w}_3^\top & b_3\end{bmatrix}\right\|_2 \left\|\begin{bmatrix}\boldsymbol{W}_2^\top & \boldsymbol{b}_2 \\ \boldsymbol{0} & 1\end{bmatrix}\right\|_F \left\|\begin{bmatrix}\boldsymbol{W}_1^\top & \boldsymbol{b}_1 \\ \boldsymbol{0} & 1\end{bmatrix}\right\|_F \leq \left(M + \frac{2}{3}\right)^{3/2}.$$

Because $\|\boldsymbol{AB}\|_F \leq \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_F$ and $\|[\boldsymbol{A}]_+\|_F \leq \|\boldsymbol{A}\|_F$, we see that for $\boldsymbol{x} \in \mathcal{X}_d$,

$$|h(\boldsymbol{x})| \leq \left\|\begin{bmatrix}\boldsymbol{w}_3^\top & b_3\end{bmatrix}\right\|_2 \left\|\begin{bmatrix}\boldsymbol{W}_2^\top & \boldsymbol{b}_2 \\ \boldsymbol{0} & 1\end{bmatrix}\right\|_F \left\|\begin{bmatrix}\boldsymbol{W}_1^\top & \boldsymbol{b}_1 \\ \boldsymbol{0} & 1\end{bmatrix}\right\|_F \left\|\begin{bmatrix}\boldsymbol{x} \\ 1\end{bmatrix}\right\|_2 \leq \sqrt{3}\left(M + \frac{2}{3}\right)^{3/2}.$$

This shows that

$$\sup_{h\in\mathcal{N}_3^M} \|f_d - h\|_{L^\infty} \leq \|f_d\|_{L^\infty} + \sup_{h\in\mathcal{N}_3^M} \|h\|_{L^\infty} \leq 1 + \sqrt{3}\left(M + \frac{2}{3}\right)^{3/2} = O(M^{3/2}).$$

□

Using Lemmas A.17 and A.18, standard Rademacher complexity arguments yield an estimation error bound over $\mathcal{N}_3^M$, as shown in the following lemma.

Lemma A.19. *Consider a distribution $\mathscr{D}_d$ on $\mathcal{X}_d \times [-1,1]$ defined as*

$$\boldsymbol{x} \sim \text{Uniform}(\mathcal{X}_d) \tag{108}$$

$$y|\boldsymbol{x} = f_d(\boldsymbol{x}) \tag{109}$$

for some function $f_d : \mathcal{X}_d \to [-1,1]$. If $f \in \mathcal{N}_3$ with $R_3(f) \leq M$, then

$$|\mathscr{L}_{\mathscr{D}_d}(f) - \mathscr{L}_S(f)| \leq O\left(M^3\sqrt{\frac{\log 1/\delta}{|S|}}\right) \tag{110}$$

with probability at least $1-\delta$ over samples S drawn i.i.d. from $\mathscr{D}_d$.

Proof. We apply the properties of Rademacher complexity (see for example Theorem 12 in Bartlett and Mendelson (2001) and Theorem 4.10 in Wainwright (2019)) to give an estimation error bound over $\mathcal{N}_3^M$ as follows. Define the loss class $\mathcal{L}_{\mathcal{N}_3^M, f_d} := \{(h - f_d)^2 : h \in \mathcal{N}_3^M\}$. With probability at least $1-\delta$,

$$\sup_{h\in\mathcal{N}_3^M} |\mathscr{L}_{\mathscr{D}_d}(h) - \mathscr{L}_S(h)| \tag{111}$$

$$\leq O\left(\mathscr{R}_{\mathcal{X}_d^m}(\mathcal{L}_{\mathcal{N}_3^M, f_d}) + \sqrt{\frac{\log(1/\delta)}{m}} \sup_{h\in\mathcal{N}_3^M} \|f_d - h\|_{L^\infty}^2\right) \tag{112}$$

$$\leq O\left(\sup_{h\in\mathcal{N}_3^M} (\|f_d - h\|_{L^\infty}) \left(\mathscr{R}_{\mathcal{X}_d^m}(\mathcal{N}_3^M) + 1/\sqrt{m}\right) + \sqrt{\frac{\log(1/\delta)}{m}} \sup_{h\in\mathcal{N}_3^M} \|f_d - h\|_{L^\infty}^2\right). \tag{113}$$

Plugging in the bounds from Lemmas A.17 and A.18, this becomes

$$\sup_{h\in\mathcal{N}_3^M} |\mathscr{L}_{\mathscr{D}_d}(h) - \mathscr{L}_S(h)| = O\left(M^3\sqrt{\frac{\log 1/\delta}{m}}\right). \tag{114}$$

□

A.7 Full Proof of No Reverse Depth Separation

A.7.1 Proof of Lemma 7.1

Proof. Fix $\varepsilon > 0$. Let S be a sample from $\mathscr{D}_d$ of size $m_2\left(\frac{\varepsilon}{2}\right)$. As in the proof of Theorem 5.1 Part 1, we rely on the existence of an interpolant. By Lemma A.15, with probability at least 0.6 there is an interpolant $\hat{f}_S \in \mathcal{N}_2$ of the samples S with $R_2\left(\hat{f}_S\right) \leq 100\sqrt{2}m_2\left(\frac{\varepsilon}{2}\right)^{\frac{d+3}{d-1}}$. Because $\mathcal{A}_2^*(S) \in \mathcal{P}_2(S)$ is Pareto optimal, it follows that $R_2(\mathcal{A}_2^*(S)) \leq R_2(\hat{f}_S)$. We conclude that

$$\mathbb{P}\left(R_2(\mathcal{A}_2^*(S)) > 100\sqrt{2}m_2\left(\frac{\varepsilon}{2}\right)^{\frac{d+3}{d-1}}\right) \leq 0.4.$$

On the other hand, since $\mathbb{E}_S[\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_2^*(S))] \leq \frac{\varepsilon}{2}$ whenever $|S| \geq m_2\left(\frac{\varepsilon}{2}\right)$, it follows from Markov's inequality that

$$\mathbb{P}\left(\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_2^*(S)) > \varepsilon\right) \leq 0.5. \tag{115}$$

Therefore,

$$\mathbb{P}\left(\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_2^*(S)) > \varepsilon \text{ or } R_2(\mathcal{A}_2^*(S)) > 100\sqrt{2}m_2\left(\frac{\varepsilon}{2}\right)^{\frac{d+3}{d-1}}\right) \leq 0.9 < 1. \tag{116}$$

We conclude that there is some sample S_ε from $\mathscr{D}_d$ of size $m_2\left(\frac{\varepsilon}{2}\right)$ such that

$$\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_2^*(S_\varepsilon)) \leq \varepsilon \text{ and } R_2(\mathcal{A}_2^*(S_\varepsilon)) \leq 100\sqrt{2}m_2\left(\frac{\varepsilon}{2}\right)^{\frac{d+3}{d-1}}. \tag{117}$$

We choose $f_\varepsilon = \mathcal{A}_2^*(S_\varepsilon)$. □

A.7.2 Proof of Theorem 5.2 (No Reverse Depth Separation)

Proof. Fix $\varepsilon, \delta > 0$ and $\alpha \geq 1$. Let $\theta = \frac{\varepsilon}{2\alpha}$. Under the assumptions of the theorem, Lemma 7.1 tells us there is a function $f_\theta \in \mathcal{N}_2$ such that $\mathscr{L}_{\mathscr{D}_d}(f_\theta) \leq \theta/8$ and $R_2(f_\theta) \leq O\left(m_2\left(\frac{\varepsilon}{32\alpha}\right)^{\frac{d+3}{d-1}}\right)$. Let

$$\omega_2 = \frac{24 R_2(f_\theta)^2}{\theta} = O\left(\frac{m_2\left(\frac{\varepsilon}{32\alpha}\right)^{\frac{2(d+3)}{d-1}}\alpha}{\varepsilon}\right). \tag{118}$$

Lemma 3.2 allows us to approximate f_θ — and thus $\mathscr{D}_d$ — with width ω_2; there is some $\tilde{f}_\theta \in \mathcal{N}_{2,\omega_2}$ such that $R_2(\tilde{f}_\theta; \omega_2) \leq R_2(f_\theta)$ and $\|f_\theta - \tilde{f}_\theta\|_{L^2} < \sqrt{\theta/8}$. Thus,

$$\mathscr{L}_{\mathscr{D}_d}(\tilde{f}_\theta) \leq 2\left(\mathscr{L}_{\mathscr{D}_d}(f_\theta) + \|f_\theta - \tilde{f}_\theta\|_{L^2}^2\right) \leq \theta/2. \tag{119}$$

If $\omega \geq \max(\omega_2, 4d)$, then Lemma A.2 tells us that $\tilde{f}_\theta \in \mathcal{N}_{3,\omega}$ and

$$\begin{aligned} R_3(\tilde{f}_\theta; \omega) &\leq \frac{4d}{3} + \frac{4}{3} R_2(\tilde{f}_\theta; \omega_2) && (120)\\ &\leq \frac{4d}{3} + R_2(f_\theta) && (121)\\ &= O\left(d + m_2\left(\frac{\varepsilon}{32\alpha}\right)^{\frac{d+3}{d-1}}\right). && (122) \end{aligned}$$

By the estimation error bound in Lemma A.19 and the union bound, with probability at least $1 - \delta$ we have that

$$\left|\mathscr{L}_S\left(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S)\right) - \mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S))\right| = O\left(\sqrt{\frac{R_3(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S); \omega)^6 \log(1/\delta)}{|S|}}\right) \tag{123}$$

and

$$\left|\mathscr{L}_S\left(\tilde{f}_\theta\right) - \mathscr{L}_{\mathscr{D}_d}(\tilde{f}_\theta)\right| = O\left(\sqrt{\frac{R_3(\tilde{f}_\theta; \omega)^6 \log(1/\delta)}{|S|}}\right). \tag{124}$$

If $|S| \geq m_3(\varepsilon, \delta, \alpha)$, where

$$m_3(\varepsilon, \delta, \alpha) = O\left(\frac{\alpha^6\left(d + m_2\left(\frac{\varepsilon}{64}\right)^{\frac{d+3}{d-1}}\right)^6 \log 1/\delta}{\varepsilon^2}\right), \tag{125}$$

then Equations (122) and (124) imply that $\left|\mathscr{L}_S\left(\tilde{f}_\theta\right) - \mathscr{L}_{\mathscr{D}_d}(\tilde{f}_\theta)\right| \leq \theta/2$, and so $\mathscr{L}_S\left(\tilde{f}_\theta\right) \leq \theta$. Hence

$$\begin{aligned} R_3(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S); \omega) &\leq \alpha \inf_{\substack{f \in \mathcal{N}_{3,\omega} \\ \mathscr{L}_S(f) \leq \theta}} R_3(f; \omega) && (126)\\ &\leq \alpha R_3(\tilde{f}_\theta; \omega) && (127)\\ &= O\left(\alpha\left(d + m_2\left(\frac{\varepsilon}{32\alpha}\right)^{\frac{d+3}{d-1}}\right)\right). && (128) \end{aligned}$$

By Equations (123) and (128), if $|S| \geq m_3(\varepsilon, \delta, \alpha)$ then $\left|\mathscr{L}_S\left(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S)\right) - \mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S))\right| \leq \frac{\varepsilon}{2}$. Therefore $\mathscr{L}_{\mathscr{D}_d}(\mathcal{A}_{3,\omega}^{\theta,\alpha}(S)) \leq \alpha\theta + \frac{\varepsilon}{2} = \varepsilon$. □